STUDY GUIDE

THE CHICAGO GUIDE TO WRITING ABOUT MULTIVARIATE ANALYSIS

SECOND EDITION

JANE E. MILLER

The University of Chicago Press   Chicago & London

CONTENTS

PREFACE

This study guide was designed to provide practice applying the principles and tools covered in *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*, with a problem set of suggested course extensions for each chapter. A series of podcasts, spreadsheet templates, and other supplemental learning materials are available on the website.

The problem sets reinforce the concepts and skills from each chapter, usually working with data or written examples provided as part of the question. Some require calculations; others involve creating or critiquing tables, charts, or sentences. They can be used as homework assignments for courses on regression analysis, research methods, or research writing, or independently by readers who are trained in regression methods. Solutions for the odd-numbered problems can be downloaded separately. See also podcasts and online appendix on teaching how to write about multivariate analysis.

The suggested course extensions apply the skills and concepts from *Writing about Multivariate Analysis, 2nd Edition* to the actual writing process. They involve reviewing existing work, applying statistics, writing, and revising—using either your own work in progress or published materials (books, articles, reports, or web pages) in your field or that of your intended audience.

The "applying statistics" questions require access to a computerized database that includes several nominal, ordinal, and interval or ratio variables for at least several hundred cases. Ideally these variables should be related to a research question involving application of multivariate regression that you can use for the exercises throughout the study guide, yielding a comprehensive analysis for a complete research paper. These exercises also require access to the accompanying documentation describing the study design, data collection, coding, use of sampling weights, and related methodological issues for the data set from which your variables are taken. If you do not have a data set and documentation that fits these criteria, you can download data sets from the supplemental online materials provided on the website that accompanies this study guide. Alternatively, you can often find suitable data sets on CD-ROMS that accompany research methods or statistics textbooks, or you can download data from sites such as the Inter-University Consortium for Political and Social Research (ICPSR).

Many of the suggested exercises for writing or revision entail peer-editing. They are most effective if done with one or more other people, whether as part of a course in which class time is devoted to these exercises, or working with a colleague. These exercises often involve writing or revising work to meet the instructions for authors for a leading journal in your field. Identify one or two such journals before you begin these tasks, allowing you to generate a coherent finished product for submission to that journal.

1. *Introduction*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article from your field that involves an application of multivariate analysis. Identify the audience for that journal in terms of

    a. their discipline(s).
    b. their expected level of familiarity with the type of multivariate model used in the article. E.g., is that method widely used in the field, new to the field or topic but well-established elsewhere, or new to all fields?
    c. their expected use of the results (e.g., research, policy, education).

2. In that article

    a. Circle one numeric fact or comparison each in the introduction, results section, and concluding section. For each
        i. Identify its purpose. Does the author explicitly or implicitly convey the purpose, or is it left unclear?
        ii. Evaluate the ease of understanding that fact or comparison. Does the author convey its meaning and interpretation?
    b. Are there other places in the article where a number or comparison would be helpful? Identify the purpose of the number for each such situation.
    c. What tools are used to present numbers? Do they suit the objective and audience for the article?

3. Find an article in the popular press that refers to an application of a multivariate analysis. (The science and health sections of newspapers, magazines, and websites are good resources.)

    a. Who is the audience for the article (e.g., what is their expected reading level and amount of statistical training)?
    b. What is the objective of the article?
    c. Is the article written with appropriate vocabulary and examples for that audience?
    d. What tools (tables, charts, prose) are used to present numbers in the article? Do they suit the objective and audience?

2. *Seven Basic Principles*

PROBLEM SET

1. Use complete sentences to describe the relative sizes of the cities shown in table 2A.

TABLE 2A. Population of three largest cities worldwide, 1995

City	Population (millions)
Sao Paulo	16.5
Mexico City	16.6
Tokyo	27.0

Source: Population Reference Bureau, "World Population: More than Just Numbers," Washington DC: Population Reference Bureau, 1999.

2. One of the W's is missing from each of the following descriptions of table 2B. Rewrite each sentence to include that information.

TABLE 2B. Final medal standings of the top four countries, 2002 Winter Olympic Games

Country	Gold	Silver	Bronze	Total
Germany	12	16	7	35
United States	10	13	11	34
Norway	11	7	6	24
Canada	6	3	8	17

a. "Germany did the best at the 2002 Winter Olympics, with 35 medals, compared to 34 for the United States, 24 for Norway, and 17 for Canada."
b. "Gold, silver, and bronze medals each accounted for about one-third of the medal total."
c. "At the 2002 Winter Olympics, the United States won more medals than all other countries, followed by Canada, Germany, and Norway."

3. For each of the following situations, specify whether you would use prose, a table, or a chart.

a. Statistics on five types of air pollutants in the 10 largest US cities for a government report
b. Trends in the value of three stock market indices over a one-year period for a web page

c. Notification to other employees in your corporation of a change in shipping fees

d. Distribution of voter preferences for grade-level composition of a new middle school (grades 5–8, grades 6–8, or grades 6–9) for a presentation at a local school board meeting

e. National estimates of the number of uninsured among part-time and full-time workers for an introductory section of an article analyzing effects of employment on insurance coverage in New York City

4. For each of the situations in the previous question, state whether you would use and define technical terms or avoid jargon.

5. Identify terms that need to be defined or restated for a nontechnical audience.

a. "The Williams family's income of $25,000 falls below 185% of the Federal Poverty Threshold for a family of four, qualifying them for food stamps."

b. "A population that is increasing at 2% per year has a doubling time of 35 years."

6. Rewrite the sentences in the previous question for an audience with a fifth-grade education. Convey the main point, not the calculation or the jargon.

7. Read the sentences below. What additional information would someone need in order to answer the associated question?

a. "Brand X costs twice as much as Brand Q. Can I afford Brand X?"

b. "My uncle is 6'6" tall? Will he fit in my new car?"

c. "New Diet Limelite has 25% fewer calories than Diet Fizzjuice. How much faster will I lose weight on Diet Limelite?"

d. "It has been above 25 degrees every day. We're really having a warm month, aren't we?"

8. Rewrite each of these sentences to specify the direction and magnitude of the association.

a. "In the United States, race is correlated with income." See table 2C.

TABLE 2C. Median income by race and Hispanic origin, United States, 1999

Race/Hispanic origin	Median income
White	$42,504
Black	$27,910
Asian/Pacific Islander	$51,205
Hispanic (can be of any race)	$30,735

Source: US Bureau of the Census, *Statistical Abstract of the United States*, 2001, table 662.

b. “There is an association between average speed and distance traveled.” (Pick two speeds to compare.)

c. Write a hypothesis about the relationship between amount of exercise and weight gain.

9. Use the GEE approach on pp. 30–32 of *Writing about Multivariate Analysis, 2nd Edition* to describe the patterns in figure 2A. Include an introductory sentence about the purpose of the chart before summarizing the patterns.

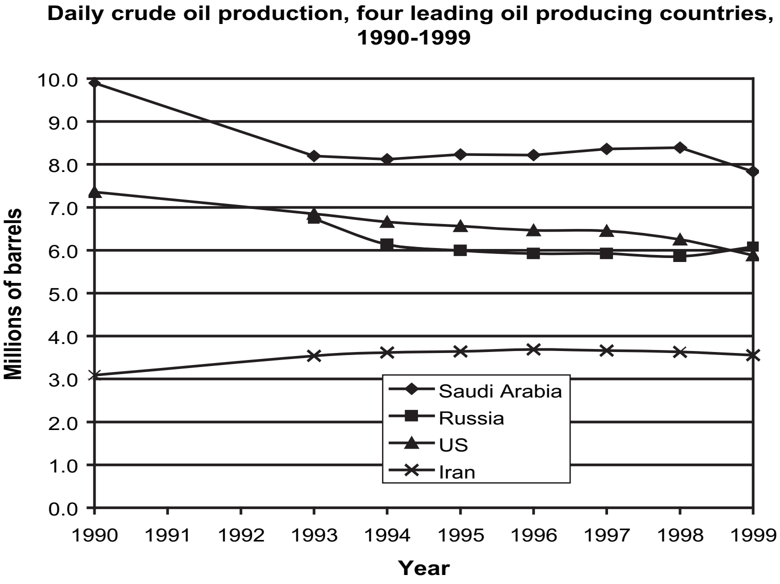


Figure 2A.

2. *Seven Basic Principles*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

Find a journal article about an application of a multivariate analysis. Use it to answer the following questions.

1. Is the context (W's) of the study specified? If not, which W's are missing or poorly defined?

2. Evaluate the technical language.

   a. Are definitions provided for all technical and statistical terms that might be unfamiliar to the audience?
   b. Are all acronyms used in the paper spelled out and defined?
   c. Are pertinent synonyms for methods or concepts familiar to the intended audience mentioned?

3. Circle all analogies or metaphors used in the paper. Are they likely to be familiar to the intended audience? If not, replace them with more suitable analogies or metaphors.

4. Identify the major tools (text, tables, charts) used to present numbers in the article.

   a. For one example of each type of tool, identify its intended purpose or task in that context (e.g., presenting detailed numeric values; conveying a general pattern).
   b. Use the criteria in chapter 2 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate whether it is an appropriate choice for that task. If so, explain why. If not, suggest a more effective tool for that context.

5. Find a numeric fact or comparison in the introduction or conclusion to the article.

   a. Is it clear what question that fact or comparison is intended to answer?

b. Are the raw data for that fact or comparison presented in the text, a table, or chart?
c. Are the values interpreted in the text?
d. Revise the paragraph to address any shortcomings you identified in parts a through c.

6. Find a description of an association between two variables. Are the direction and magnitude of the association specified? If not, rewrite the description.

7. Find a description of a pattern involving more than three values, subgroups, or results of models that are presented in a table or a chart.

   a. Is the purpose of the chart or table explained?
   b. Is the pattern generalized, or is it described piecemeal?
   c. Are representative values reported to illustrate the pattern?
   d. Are exceptions to the general pattern identified?
   e. Rewrite the description of the table or chart using the "Generalization, example, exception" (GEE) approach on pp. 30–32 of *Writing about Multivariate Analysis, 2nd Edition* to address any shortcomings you identified in parts a through d.

B. Writing Papers

1. For a bivariate association among variables in your data,

   a. Specify which tool you would use to present the findings in a paper for a scientific audience in your field.
   b. Write one to two sentences to describe that association, including the W's, units, direction, magnitude, and statistical significance.
   c. Redo parts a and b to present the same association in a talk to a lay audience.

2. Begin with the introduction.

   a. Write an introduction that integrates the concepts and methods used in your study.
   b. Use the criteria in chapter 2 of *Writing about Multivariate Analysis, 2nd Edition* to assess use of technical language in your introduction.
   c. Revise your introduction to address any shortcomings you identified in part b.

3. Graph the distribution of a continuous variable in your data set. Describe it using an analogy.

4. Design a chart to portray a three-way association among variables in your data set. Use the GEE approach to describe the pattern.

C. Revising Papers

1. Repeat questions A.1 through A.7 for a paper you have written previously about a multivariate analysis.

2. Have someone who is unfamiliar with your research question peer-edit your answers to question C.1, using the checklist from chapter 2 of *Writing about Multivariate Analysis, 2nd Edition*. “Editors” should suggest specific sentences, examples, or other changes (e.g., “replace a table with a chart”) to replace the material needing revision. Revise according to the feedback you receive.

2. *Seven Basic Principles*

SOLUTIONS

1. “In 1995, the world’s largest city, Tokyo, had a population of 27 million people. With populations of roughly 16.5 million apiece, the next two largest cities, Mexico City and Sao Paolo, were only about 60% as large as Tokyo.”

3. Choice of prose, a table, or a chart for specific situations.

   a. Table to show detailed values and organize the 50 numbers
   b. Multiple-line chart to illustrate approximate pattern
   c. Prose (memo)
   d. Pie chart
   e. Prose (few sentences)

5. Terms that need to be defined or restated for a nontechnical audience are shown in bold.

   a. “The Williams family’s income of $25,000 falls below **185% of the Federal Poverty Threshold for** a family of four, qualifying them for food stamps.”
   b. “A population that is increasing at 2% per year has a **doubling time** of 35 years.”

7. Additional information needed to answer the associated question:

   a. How much does Brand Q (or Brand X) cost? How much money do you have?
   b. How big is the door opening to your car? The headroom and legroom?
   c. How many calories does Diet Fizzjuice (or Diet Limelite) have?
   d. Where are you located? What month is it? Is temperature being measured in degrees Fahrenheit or degrees Celsius?

9. “Figure 2A shows trends in daily crude oil production in the world’s four leading oil-producing countries during the 1990s. Over the course of that decade, Saudi Arabia consistently had the highest crude oil production, followed by Russia, the United States, and Iran. However, downward trends in production in the top three oil-producing

countries, coupled with steady production in Iran, led to a narrowing of the gap between those countries between 1990 and 1999. In 1990, Saudi Arabia produced 30% more oil than the United States and more than three times as much as Iran (10 million, 7 million, and 3 million barrels per day, respectively). By 1999, Saudi Arabia's advantage had decreased to 25% more than the United States or Russia, and about twice as much as Iran."

3. *Causality, Statistical Significance, and Substantive Significance*

PROBLEM SET

1. Evaluate whether each of these statements correctly conveys statistical significance. If not, rewrite the sentence so that the verbal description about statistical significance matches the numbers; leave the numeric values unchanged.

   a. There was a statistically significant increase in average salaries over the past three years ($p = .04$).
   b. The p-value for the t-test for difference in mean ozone levels equals 0.95, so we can be 95% certain that the observed difference is not due to chance.
   c. The difference in voter participation between men and women was not statistically significant ($p = 0.35$).
   d. The p-value for the t-test for difference in mean ozone levels equals 0.95. This test shows we can be 95% certain that the difference in ozone levels can be explained by random chance; hence the difference is not statistically significant.
   e. The price of gas increased by $0.05 over the past three months, meaning that the p-value $= 0.05$.
   f. The p-value comparing trends in gas prices $= 0.05$, hence the price of gas increased by $0.05.
   g. Voter participation was 20% higher among Democrats than among Republicans in the recent local election. Statistical tests show $p < .01$, so we can be 99% certain that the observed difference is not due to chance.
   h. The average processor speed was slightly higher for Brand A than for Brand B; however $p = .09$, so the effect was not statistically significant. If the sample size were increased from 40 to 400, the difference in processor speeds between the two brands would increase, so it might become statistically significant.
   i. The average processor speed was slightly higher for Brand A than for Brand B; however $p = .09$, so the effect was not statistically significant. If the sample size were increased from 40 to 400, the standard error would decrease, so the difference might become statistically significant.

2. For each of the following findings, identify background facts that could help decide whether the effect is big enough to matter. Look up your suggested facts for one of the research questions. What do you conclude about the substantive significance of the finding?

    a. Jo's IQ score increased 2 points in one year.
    b. The average response on a political opinion poll for two adjacent counties differed by 2 points. The question was scaled "agree strongly," "agree," "neither agree nor disagree," "disagree," and "disagree strongly."
    c. The Dow Jones Industrial Index dropped 2 points since this morning.
    d. Bed rest is expected to prolong Mrs. Peterson's pregnancy to 36 weeks from 34 weeks gestation.

3. Discuss whether each of the following research questions involves a causal relationship. If the relationship is causal, describe one or more plausible mechanisms by which one variable could cause the other. If the relationship is not causal, give alternative explanations or mechanisms for the association.

    a. April showers bring May flowers.
    b. People with blue eyes are more likely to have blond hair.
    c. Pollen allergies increase rapidly with longer daylight hours.
    d. Eating spicy foods is negatively correlated with heartburn.
    e. Prices and sales volumes are inversely related.
    f. Fair-skinned people sunburn faster than do those with dark skin.
    g. Average reading ability increases dramatically with height between 4' and 5'.

4. For each of the studies summarized in table 3A

    a. explain how would you describe the findings in the results section of a scientific paper;
    b. identify the criteria you used to decide how to discuss the findings for that study.

TABLE 3A. Hypothetical study results

Topic I: Effect of new math curriculum on test scores*	Effect size	Statistical significance (*p*-value)	Sample size
Study 1	+½ point	$p < .01$	1 million
Study 2	+½ point	$p = .45$	1 million
Study 3	+5 points	$p < .01$	1 million
Study 4	+5 points	$p = .07$	1 hundred
Study 5	+5 points	$p = .45$	1 million

(*continued*)

TABLE 3A. (continued)

Topic II: Effect of white hair on mortality**	Effect size	Statistical significance (*p*-value)	Sample size
Study 1	+5%	$p < .01$	1 million
Study 2	+5%	$p = .45$	1 million
Study 3	+50%	$p < .01$	1 million
Study 4	+50%	$p = .07$	1 hundred
Study 5	+50%	$p = .45$	1 million

* Effect size for math curriculum studies = scores under new curriculum – scores under old curriculum.

** Effect size for hair color studies = death rate for white-haired people – death rate for people with other hair colors.

5. For each of the topics in table 3A, indicate whether you would recommend a policy or intervention based on the results, and explain the logic behind your decision.

3. *Causality, Statistical Significance, and Substantive Significance*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article in your field, find an example of a highly correlated association.

    a. Is that association causal? Why or why not?
    b. List facts or comparisons that could be used to evaluate the substantive meaning of the association:
        i. that the authors report and interpret in the article;
        ii. other facts or comparisons that could be used to improve the explanation in the article.

2. In a journal article in your field, find an association with a low correlation or nonstatistically significant association.

    a. Is that association causal? Why or why not?
    b. List facts or comparisons that could be used to evaluate whether the association is substantively meaningful:
        i. that the authors report and interpret in the article;
        ii. other facts or comparisons that could be used to improve the explanation in the article.

3. Find a journal article that uses multivariate regression to analyze a policy problem and proposes one or more solutions to that problem.

    a. Evaluate how well the article addresses each of these aspects of "importance." Does the article
        i. specify a cause-and-effect type of relationship?
        ii. provide a plausible argument for a causal association?
        iii. discuss bias, confounding, or reverse causation?
        iv. report results of statistical tests for that association?
        v. assess whether the expected benefits of the proposed solution are big enough to outweigh costs or otherwise matter in a larger social context?
    b. Given your answers to part a, write a short critique of the appropriateness of the proposed solution.

4. Repeat question A.3 with an article in the popular press about a scientific or policy problem and solution that is currently being touted for implementation.

B. Writing and Revising

1. Identify an aspect of your research question that involves the association between an independent and dependent variable. Do you hypothesize that that association is causal?

    a. If so, describe the mechanisms through which the hypothesized causal variable affects the hypothesized outcome variable.
    b. If not, explain how those variables could be correlated. Identify possible bias, confounding factors, or reverse causation.
    c. Rewrite your research question as a hypothesis, making it clear whether the association you are studying is expected to be causal.
    d. What background facts could help assess the substantive meaning of that association? Look them up and write a short description to make that assessment.
    e. Write a description of the substantive importance of the association for a discussion section of a scientific paper.
    f. Write a statement for a lay audience, explaining the nature of the association between the variables.

2. For one or two key statistical results pertaining to the main research question in your paper, identify ways to quantify the broad social or scientific impact of that finding, following the guidelines in chapter 3 of *Writing about Multivariate Analysis, 2nd Edition*.

    a. Locate statistics on the prevalence of the phenomenon you are studying.
    b. Find information on the consequences of the issue. For example, what will it cost in money, time, or other resources? What are its benefits? What does it translate into in terms of reduced side effects, improved skills, or other dimensions suited to your topic?
    c. Use information from parts a and b in conjunction with measures of effect size and statistical significance to make a compelling case for or against the importance of the topic.

3. Repeat question B.2 for a paper you have already written about an application of a multivariate regression model.

3. *Causality, Statistical Significance, and Substantive Significance*

SOLUTIONS

1. Evaluation of whether the statements correctly convey statistical significance.

   a. Correct.
   b. Incorrect. A *p*-value of 0.95 corresponds to a 5% probability that the observed difference is *not* due to chance (e.g., a 95% probability that the observed difference *is* due to chance.) "The *p*-value for the *t*-test for difference in mean ozone levels equals 0.95, so we can be 95% certain that the observed difference is due to chance."
   c. Correct.
   d. Correct.
   e. Incorrect. This sentence doesn't reveal anything about statistical significance of that change. The most we can say from the information given is "The price of gas increased by $0.05 over the past three months."
   f. Incorrect. Test statistics and *p*-values are indicators of statistical significance. They do not measure the size of the association, in this case, difference between two values, which cannot be calculated from the information given. The most we can say is "The *p*-value comparing trends in gas prices = 0.05."
   g. Correct.
   h. Incorrect. Sample size does not affect size of a difference between values, in this case, difference in average processor speeds. See part i of this question for correct wording.
   i. Correct.

3. Discuss the causal or noncausal relationships in the presented research questions.

   a. Causal (partly). The flowers will bloom in May whether or not it rains in April, but will bloom more nicely if it rains.
   b. Noncausal association. In many populations, blue eyes and blond hair co-occur but neither causes the other.
   c. Spurious. Positive correlations between both pollen allergies and daylight with more flowers blooming cause a spurious association between allergies and daylight. In other words, if you could have

more daylight without more blooming plants, there wouldn't be an association of daylight hours with pollen allergies.

d. Could be causal or reverse causal. For example, people with heartburn might stop eating spicy foods if they think those foods irritate their heartburn.

e. Reverse causal. Low prices probably induced greater sales. Could be causal in the long run if greater sales allow economies of scale in production, which in turn could lower prices.

f. Causal. Lack of protective pigment in fair-skinned people allows them to sunburn faster.

g. Spurious. Both reading ability and height increase dramatically with children's age, which in turn is positively related to number of years of education. Education is the cause of improved reading ability. Comparing kids of the same age or years of education but different heights would likely show much less difference in reading abilities than if age isn't taken into account.

5. For both topics I and II in table 3A, the findings of studies 1 and 3 are statistically significant, studies 2 and 5 are not, and study 4 is borderline because the *p*-value is slightly above 0.05 and the sample size is small. However, the white hair/mortality association in topic II is spurious, so substantive and statistical significance are irrelevant. For topic I (curriculum change and test scores) where there is a plausible causal explanation, only the findings of study 3 are likely to be of substantive interest because the effect size in study 1 is so small.

4. *Five More Technical Principles*

PROBLEM SET

1. For each of the following topics, indicate whether the variable or variables used to measure it are continuous or categorical, and single or multiple response.

   a. Respondent's current marital status
   b. Respondent's current number of siblings
   c. Siblings' current heights
   d. Current marital status of siblings
   e. Temperature at 9 a.m. today
   f. The forms of today's precipitation

2. A new school is being considered in your hometown. Several possible grade configurations are being considered (Plan A: grades K–3, 4–5, 6–8, 9–12; Plan B: grades K, 1–4, 5–7, 8–12). The current configuration is K–5, 6–9, and 10–12. Design a question to collect information from school principals on the age distribution of students, making sure the data collection format provides the detail and flexibility needed to compare the different scenarios for the district now and in five years.

3. In a health examination survey, several hundred girls aged five to ten years were measured with a metric measuring tape marked in increments of millimeters. The estimated coefficient on age (years) from an OLS model of height was reported as 5.06666667 centimeters. Write a sentence to report that coefficient.

4. In a microbiology lab exercise, the size of viral cells being compared ranged from 0.000000018 meters (m) in diameter for Parovirus to 0.000001 m in length for Filoviridae (American Society for Microbiology 1999). What scale would you use to report those data in a table? In the text?

5. Write one or two sentences to compare the four specimens in table 4A. Which specimen is the heaviest? The lightest? By how much do

they differ? What steps do you need to take before you can make the comparison?

TABLE 4A.1. Mass of four specimens

Specimen	Mass
1	1.2 pounds
2	500 grams
3	0.7 kilograms
4	12 ounces

6. For each of the figures 4.3a through 4.3e (pp. 62–63 of *Writing about Multivariate Analysis, 2nd Edition*), choose

   a. a typical value;
   b. an atypical value;
   c. a plausible contrast (two values to compare).

   Explain your choices, with reference to range, central tendency, variation, and skewness.

7. Identify pertinent standards or cutoffs and other information needed to answer each of the following questions.

   a. Does Mr. Jones deserve a speeding ticket?
   b. Is the new alloy strong enough to be used for the library renovations?
   c. How tall is five-year-old Susie expected to be next year?
   d. Does Vioxx increase the odds of a heart attack?
   e. Is this year's projected tuition increase at Public U unexpected?
   f. Should we issue an ozone warning today?

8. Indicate whether each of the following sentences correctly reflects table 4B. If not, rewrite the sentence so that it is correct. Check both correctness and completeness of the data.

   a. Between 1964 and 1996, there was a steady decline in voter participation, from 95.8% in 1964 to 63.4% in 1996.
   b. Voter turnout was better in 1996 (63.4%) than in 1964 (61.9%).
   c. Almost all registered voters participated in the 1964 US presidential election.
   d. The best year for voter turnout was 1992, with 104,600 people voting.
   e. Less than half of the voting age population voted in the 1996 presidential election.
   f. A higher percentage of the voting-age population was registered to vote in 1996 than in 1964.

TABLE 4B. Voter turnout, US presidential elections, 1964 through 1996

Year	Total Vote (1,000s)	Registered Voters (RV) (1,000s)	Vote/RV (%)	Voting Age Pop. (VAP) (1,000s)	Vote/VAP (%)
1964	70,645	73,716	95.8	114,090	61.9
1968	73,212	81,658	89.7	120,328	60.8
1972	77,719	97,329	79.9	140,776	55.2
1976	81,556	105,038	77.6	152,309	53.5
1980	86,515	113,044	76.5	164,597	52.6
1984	92,653	124,151	74.6	174,466	53.1
1988	91,595	126,380	72.5	182,778	50.1
1992	104,600	133,821	78.2	189,529	55.2
1996	92,713	146,212	63.4	196,511	47.2

Source: Institute for Democracy and Electoral Assistance 1999.

9. A billboard reads "1 in 250 Americans is HIV positive. 1 in 500 of them knows it."

   a. According to the two statements above, what share of Americans are HIV positive and know it? Does that seem realistic?
   b. Rewrite the second statement to clarify the intended meaning
      i. as a fraction of HIV-positive Americans;
      ii. as a fraction of all Americans.

10. An advertisement for a health education program included figure 4A to show the prevalence of two common health behavior problems among teenage girls. What is wrong with the graph?

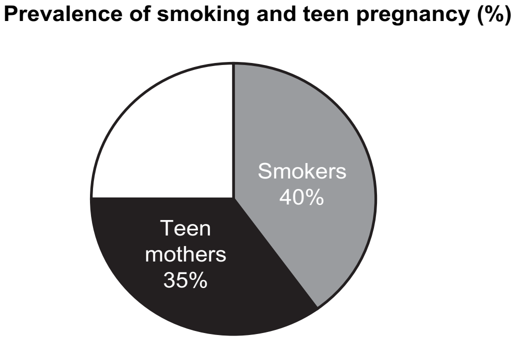


Figure 4A.

11. You are involved in a research team that is conducting a study of commuting. One of the team members submits the following question to be included on the questionnaire:

"How do you commute to work?

Car__ Public transportation__ Train__ Carpool__ Walk__"

   a. Critique the wording of the question using the guidelines in chapter 4 of *Writing about Multivariate Analysis, 2nd Edition*.

b. Revise the question to correct the problems you identified in part a.

12. What is wrong with the following fictitious set of instructions for authors from a scientific journal that frequently publishes results of multivariate regression analyses? "In the interest of saving space, round all numeric results to the nearest single decimal place."

13. Each of these statements contains an error. Identify the problem and rewrite the statement to correct the error. If additional information would be needed to make the correction, indicate what kind of information is needed.

a. The proportionate increase in income during the 1990s was 20%.
b. Male infants outnumbered females (sex ratio at birth = 0.95).
c. A majority of respondents (0.67) agreed that there should be a waiting period before buying a gun.
d. Cancer accounted for two out of every ten deaths, equivalent to a death rate of 20%.

4. *Five More Technical Principles*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article in your field, find a discussion of an association between two or three variables. For each of those variables, identify

   a. the type of variable (nominal, ordinal, interval, or ratio);
   b. whether it is single or multiple response.
   c. For continuous variables, identify
      i. the system of measurement;
      ii. the unit of analysis;
      iii. the scale of measurement;
      iv. the appropriate number of digits and decimal places for reporting the mean value in the text and a table.
   d. For categorical variables, list the categories for each variable.
   e. If the items requested in c and d aren't described in the article, list plausible versions of that information. For example, if you are studying family income in the United States, you would expect the system of measurement to be US dollars, the unit of analysis to be the family, and the scale of measurement to be either dollars or thousands of dollars.

2. Read the article's description of the variables you listed in question A.1. Does it provide the information about the distribution of that type of variable that is recommended in chapter 4 of *Writing about Multivariate Analysis, 2nd Edition?* If not, what additional information is needed?

3. Read the literature in your field to determine whether standard cutoffs or standard patterns are used to assess one of the variables in the association you listed in question A.1. Find a reference source that explains its application and interpretation.

B. Applying Statistics

1. Repeat question A.1 using variables available in your database.

2. Using the same data,

    a. calculate the frequency distribution for each variable;
    b. create a simple chart of the distribution;
    c. select and calculate the appropriate measure of central tendency for that type of variable;
    d. determine whether the measure of central tendency calculated in part c typifies the overall distribution. Why or why not? If not, what is a more typical value?
    e. for continuous variables, identify the minimum and maximum values and the cutoffs for the quartiles of the distribution.

3. For one of the variables in your database, repeat question A.3. Use the standard or cutoff to classify or evaluate your data (e.g., what percentage of cases falls below the cutoff? Does the distribution of that variable in your data follow the expected pattern based on the published literature on that topic for a similar pattern?)

4. Compare the eligibility thresholds for your state's State Children's Health Insurance (SCHIP, or CHIPRA) for the most recent year available against the Federal Poverty Thresholds (see websites for your state's SCHIP program and the "Poverty" page on the US Census website). What is the highest income that would qualify for free SCHIP benefits for a family of one adult and one child? A family of one adult and two children? A family of two adults and two children?

4. *Five More Technical Principles*

SOLUTIONS

1. Identify the variable(s) as continuous or categorical, and single or multiple response.

    a. Categorical, single response
    b. Continuous, single response
    c. Continuous, multiple response
    d. Categorical, multiple response
    e. Continuous, single response
    f. Categorical, multiple response

3. "The model suggests that on average, girls grow approximately 5.07 centimeters per year between the ages of five and ten."

5. All measurements must be converted into consistent units (scale and system of measurement). I chose to convert all measurements to kilograms (see revised table 4A), using the conversion factor 2.2 pounds/kilogram.

    "Of the four specimens compared here, specimen 3 is the heaviest (0.70 kilograms). It is about twice as heavy as the lightest (specimen 4, 0.34 kg). The other two specimens were each about 70% as heavy as specimen 3."

TABLE 4A.2. Mass of four specimens

Specimen	Weight (original units)	Weight (kg)
1	1.2 pounds	0.54
2	500 grams	0.50
3	0.7 kilograms	0.70
4	12 ounces	0.34

7. Identify pertinent standards or cutoffs.

    a. The speed limit where he was driving and his actual speed
    b. The weight-bearing capacity of the alloy (in weight per unit area) and the expected weight load (again, in weight per area) in the library

c. Her current height and a growth chart (height for age) for girls
d. The odds ratio of a heart attack for Vioxx users versus non-Vioxx users, compared to an odds ratio of 1.0 (the null hypothesis of equal odds in both groups)
e. The rate of inflation, current tuition, and rates of tuition increase at Public U over the past few years
f. Today's ozone measurement and the cutoff for an ozone warning

9. a. Taken together, the two statements imply that 1 in 125,000 Americans are HIV positive and know it, clearly a misstatement of the facts.
b. Rewrite the statement to clarify.
i. "Half of HIV-positive Americans know they are infected."
ii. "One in 500 Americans is HIV positive and knows it."

11. Critique the commuting questionnaire question.

a. First, the responses are not mutually exclusive. For example, "car" and "carpool" overlap, as do "public transportation" and "train." Second, the responses aren't exhaustive, excluding bus and bicycle, among other possibilities, and omitting an "other (specify)" response. Third, they don't provide a way for people to record more than one mode of transportation. Fourth, there is no appropriate response for people who don't work or those who work at home. And finally, there are no instructions given about how many responses are allowed.
b. "How do you commute to work? (Mark all that apply.)
Car__ Train__ Bus__ Bicycle__ Walk__
Other (specify) ______________________________________
I work at home__ I do not work___"

13. Identify the errors and rewrite.

a. Proportion and percentage are not consistent units. Write "The proportionate increase in income during the 1990s was 0.20." or "Income increased by 20% during the 1990s."
b. The reported sex ratio indicates a lower number in the numerator than the denominator. Either write "Male infants outnumbered females (sex ratio at birth = 1.05 males per female)" (flipping over the ratio to be consistent with the wording, and reporting units as males per female) or "There were slightly fewer male than female infants (sex ratio at birth = 0.95 males per female)" (revising the wording to be consistent with the numeric value, and reporting units as males per female).
c. The value 0.67 does not indicate a majority unless labeled as a proportion. Better to express the value as a percentage. Write "A majority of respondents (67%) agreed that there should be a waiting period before buying a gun."

d. A death rate is expressed relative to the population (e.g., number of living people), not as a percentage of deaths (e.g., relative to the total number of deaths). Unless the total population and number of deaths are known, the first half of the sentence doesn't include enough information to calculate the death rate. Write "Cancer accounted for two out of every ten deaths."

5. *Creating Effective Tables*

PROBLEM SET

1. Write a title for table 5A.

TABLE 5A.

Year	Median age (years)
1960	
1970	
1980	
1990	
2000	

Source: US Census of Population, various dates.

2. Answer the following questions for tables 5.2 through 5.7 in *Writing about Multivariate Analysis, 2nd Edition*.

   a. Who is described by the data?
   b. To what date or dates do the data pertain?
   c. Where were the data collected?
   d. What are the units of measurement? Are they the same for all cells in the table?
   e. Where in the table are the units of measurement defined?
   f. Does the table use footnotes? If so, why? If not, are any needed?
   g. Are panels used within the table? If so, why? If not, would the addition of panels improve the clarity of the table?

3. Table 5B needs several footnotes to be complete. What information would those footnotes provide?

TABLE 5B. Estimated OLS coefficients and standard errors from a model of BMI by demographic factors and health behaviors, Dietville, 2003

	Coefficient	Standard error
Intercept	19.03**	1.27
Age (years)		
Female		
Income level		
Poor		
Near poor		
Nonpoor		

	Coefficient	Standard error
Smoking		
None		
<1 pack/day		
1+ packs/day		
Exercise (days/week)		
<1		
1–2		
3+		
R^2	0.28	
F-statistic	4.21*	

4. What is missing from table 5C?

TABLE 5C. Results of an OLS model of log(poverty rate)

State median wage	−0.174	0.043
State median wage, squared	0.006	0.002
Log(state – federal EITC)	0.023	0.015
Log(state – federal minimum wage)	−0.015	0.011
Log(max state AFDC/FSP benefit)	0.543	0.194

5. Design a table for each of the following topics. Provide complete labeling and notes, show column spanner and panels if pertinent, and indicate what principle(s) you would use to organize items within the rows and/or columns, following the guidelines in chapters 5 and 6 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. Age (years), gender, race, and educational attainment composition of a study sample.
   b. Bivariate measures of association between height (cm), weight (kg), percentage body fat, systolic blood pressure (millimeters of mercury [mm Hg]), and resting pulse (beats per minute).
   c. Results of logistic regression models of chances of high school graduation in the United States in 1998, stratified by gender and residence (urban versus rural). The key independent variables are mother's and father's educational attainment and occupation. Other control variables include race, family income, and number of siblings. Report effect size as odds ratios; statistical significance with *z*-statistics and symbols.
   d. Projected number of people receiving college degrees by region of the country from 2010 to 2025 under three different scenarios about rates of college attendance and completion.
   e. Net effects of an interaction between tercile of a student's own high school class rank and their mother's educational attainment (<HS, =HS, >HS) on the student's first-year college grade point average (GPA). Results are based on an OLS regression controlling for gender, race, and family income, using data from the high school classes of 1995 through 2000. Report results of inferential statistical tests using symbols, with the highest tercile of each independent variable as the reference category.

6. A journal for which you are writing an article allows no more than two tables, but your current draft has three. Combine tables 5D and 5E below into one table of 18 or fewer rows.

TABLE 5D. Number of wildfires by month, United States, 1998–2000

Month	1998	1999	2000	30-year average[a]
January				
February				
March				
April				
May				
June				
July				
August				
September				
October				
November				
December				
Total				

[a] 1970–1999.

TABLE 5E. Number of acres consumed by wildfire, by month, United States, 1998–2000

Month	1998	1999	2000	30-year average[a]
January				
February				
March				
April				
May				
June				
July				
August				
September				
October				
November				
December				
Total				

[a] 1970–1999.

7. There are at least seven things wrong with the labeling of table 5F. Identify and suggest ways to correct each error. Note: All numbers are correct.

TABLE 5F.1. Results of a logistic regression of political party preference, US, 2004

Variable	Odds ratio	Confidence interval	Wald chi-square
Age group 2	1.82	−0.015–3.83	4.13
Age group 3	2.01	−0.25–5.19	3.67
Race	0.53	−1.31–1.03	5.99
Proportion poor			
<10	1.26	−0.51–2.64	0.67
10–19	2.36	0.04–5.36	7.25
20–29			
>29	0.35	−2.02–0.93	7.69

5. *Creating Effective Tables*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a simple table in a newspaper or magazine article. Evaluate whether it can stand alone without the text. Suggest ways to improve labeling and layout, using the guidelines in chapter 5 of *Writing about Multivariate Analysis, 2nd Edition*.

2. In a journal article from your field, find a table that presents the relationship between a nominal independent variable with more than two categories, and a dependent variable.

   a. Identify the principle used to organize the categories of the nominal variable in the rows or columns of the table, referring to the criteria in chapters 5 and 6.
   b. Critique whether that organization coordinates with the associated narrative.
   c. Sketch a revised version of the table that addresses any shortcomings you identified in part b.

3. In a journal article from your field, find a table of regression results.

   a. Evaluate whether you can interpret all the numbers in the table without reference to the text. Suggest ways to improve labeling and layout.
   b. Using information in the article, revise the table to correct those errors.
   c. Consider whether a different table layout would work more effectively.
   d. Assess whether additional tables are needed in the paper, to present net effects of an interaction, convey nonlinear specifications, or illustrate effects of multiunit changes in an independent variable, for example (see chapters 9, 10, and 16 of *Writing about Multivariate Analysis, 2nd Edition*).
   e. Pick a chart from the article. Draw a rough draft of a table to present the same information. Show what would go into the rows and columns, whether the table would have spanners or panels, and write complete title, labels, and notes.

B. Applying Statistics

1. Create a table to display univariate statistics for your main dependent variable and three or more independent variables that you later use in your multivariate model (see question B.3).

2. Create a table to show bivariate associations (e.g., correlations, cross-tabulations, or a difference in means) between each of the independent variables and the dependent variables you selected for question B.1.

3. Create a table to show coefficients, standard errors, and model goodness-of-fit statistics from three nested models of the association between the variables you selected for question B.1.

4. Make a list of two or three simple tables to show two-way or three-way associations that pertain to your research question. Write individualized titles for each table.

5. Obtain a copy of the instructions for authors for a leading journal in your field. Revise the tables you created in questions B.1 through B.3 to satisfy their criteria.

C. Writing and Revising

1. Design a table to report results of a bivariate analysis involving a nominal independent variable with more than two categories. Specify which organizing principle(s) you would use to display values of the independent variable in the rows, referring to the criteria in chapters 5 and 6 of *Writing about Multivariate Analysis, 2nd Edition*. Justify your choice, with reference to the specific objectives of your analysis.

2. Design a table to report the results of a multivariate analysis. Specify which organizing principle(s) you would use to organize those items in the rows of the table. Explain your choice.

3. Evaluate a table of bivariate statistics that you created previously for a paper, using the checklist in chapter 5, the criteria for organizing data in charts (chapter 6), and the instructions for authors for a leading journal in your field.

4. Evaluate a table of regression results that you created previously for that paper, again using the checklist from chapter 5 and the instructions for authors for your selected journal.

5. Exchange drafts of the bivariate and multivariate tables from questions C.1 through C.4 with a peer. Evaluate them, using the checklist in chapter 5 and the instructions for authors for their selected journal. Revise according to the feedback you receive.

6. Read through the results section of a paper you have written previously. Identify topics or statistics for which to create additional tables to present net effects of interactions, nonlinear specifications, or multiunit changes related to your multivariate model. Draft them with pencil and paper, including complete title, labels, and notes.

5. *Creating Effective Tables*

SOLUTIONS

1. Title for table 5A: “Median age of the US population, 1960 to 2000.”

3. Notes to table 5B.

   Spell out BMI (body mass index), show the formula, and provide a citation.
   Specify numeric cutoffs for income or the income-to-poverty ratio to define “poor,” “near poor,” and “nonpoor.”
   Define what “*” and “**” denote.
   Cite the data sources.

5. Design tables for the given topics.

   a. Title: “Age, gender, race, and educational attainment composition of [fill in who, when, and where for study sample].” Table structure: Demographic variables in the rows, with units specified in row header for age, subgroups for the categorical variables shown with indented row headings. Columns for number of cases and percentage of cases. Note citing data source.
   b. Title: “Pearson correlation coefficients between height, weight, percentage body fat, systolic blood pressure, and resting pulse, [W’s].” Table structure: one row and one column for each variable, with label indicating units or footnote callout for abbreviated units. Correlations reported in the below-diagonal cells (see *Writing about Multivariate Analysis, 2nd Edition*, table 5.7, for an example). Symbols in the table cells to identify $p < 0.05$, with a note to explain the meaning of the symbol. Another note to define unit abbreviations
   c. Title: “Estimated odds ratios and z-statistics from a logistic regression of high school graduation, by gender and residence, United States, 1998.” Mother’s and father’s educational attainment and occupation in the top rows, followed by other independent variables. Column spanner for each gender over columns for urban and rural (total of four models), with z-statistics in parentheses below odds ratios for each independent variable with symbols denoting $p < 0.01$ and $p < 0.05$. Goodness of fit statistics and degrees of

freedom for each model in rows at bottom of the table. Footnotes to cite data sources and to define symbols.

d. Title: “Low, medium, and high projections of number of college degrees earned (thousands), by region, United States, 2010 to 2025.” Columns for low, medium, and high with a spanner labeled “scenario,” rows for years. Notes about data sources, assumptions used in each scenario.

e. Title: “Net effects of an interaction between student’s high school class rank and mother’s educational attainment on student’s first-year college grade point average, high school classes of 1995 to 2000.” One column each for bottom, middle, and top tercile of class rank with a column spanner labeled “class rank,” one row for each level of mother’s education (<HS, =HS, >HS). Interior cells include estimated values of first-year college GPA to nearest two decimal places with symbols denoting statistical significance. Notes specifying data source and other variables controlled in the model (or naming a table in which those estimates are shown), identifying the top terciles as the reference category, and defining symbols used to denote statistical significance.

7. Errors are labeled in the table using lettered superscripts keyed to the comments below.

TABLE 5F.2. Results of a logistic regression of political party preference,[a] US, 2004

Variable	Odds ratio	Confidence interval[b, c, d]	Wald chi-square
Age group 2[e]	1.82	−0.015–3.83	4.13
Age group 3	2.01	−0.25–5.19	3.67
Race[f]	0.53	−1.31–1.03	5.99
Proportion poor[g]			
<10	1.26	−0.51–2.64	0.67
10–19	2.36	0.04–5.36	7.25
20–29[h]			
>29	0.35	−2.02–0.93	7.69

Comments on errors in table 5F:

a. The category of the dependent variable being modeled is not specified, so it is unclear whether the regression is estimating relative odds of a Democratic party preference or a Republican party preference.

b. The width of the confidence interval isn’t specified. (The correct value is 99% CI.)

c. The confidence intervals are specified in terms of log-odds, not odds ratios. (You can tell because odds ratios can never be below 0, but the corresponding log-odds will be <0.0 whenever the OR

< 1.0.) Either report log-odds instead of odds ratios and keep the current CI, or calculate the CI in terms of odds ratios.

d. Using a dash ("–") to separate confidence limits that include negative values is confusing. Replace the dash with a comma, e.g.,–0.015, 3.83

e. The reference category for age group isn't included in the table, and the labels for the other age groups don't provide enough information for readers to infer the identity of the reference category.

f. The identities of the included and reference categories of the race dummy variable cannot be determined by the row label "Race."

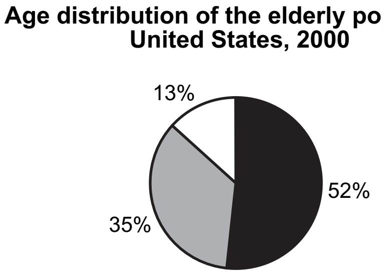
g. Proportions must be between 0.0 and 1.0, therefore the reported values are probably percentages. Either change the label to read "Percentage poor," or convert the values to proportions and label accordingly (e.g., < 0.10, 0.10–0.19).

h. The reference category could be more clearly marked using one of the conventions described in chapter 5 of *Writing about Multivariate Analysis, 2nd Edition*. Identify the convention with a note to the table.

6. *Creating Effective Charts*

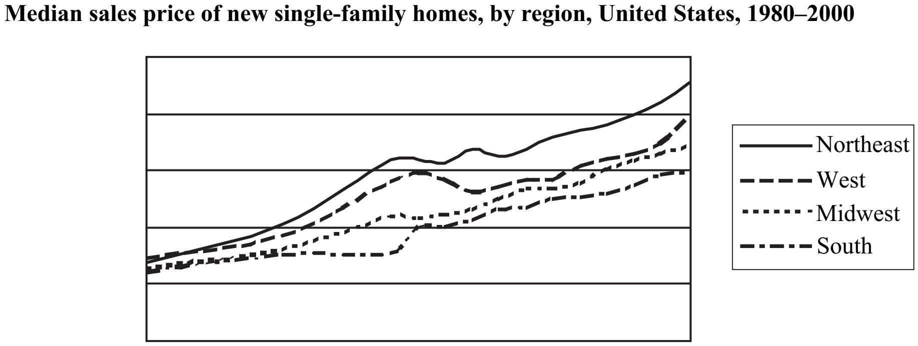
PROBLEM SET

1. List what is missing from the charts in figures 6A and 6B.



Age distribution of the elderly population
United States, 2000

Figure 6A.



Median sales price of new single-family homes, by region, United States, 1980–2000

Figure 6B.

2. Answer the following questions for figures 6.4 and 6.5 in *Writing about Multivariate Analysis, 2nd Edition*.

   a. Who is described by the data?
   b. To what date or dates do the data pertain?
   c. Where were the data collected?
   d. What criteria were used to organize the values of the variables on chart axes? (Hint: Consider type of variable.)
   e. What are the units of measurement? Are they the same for all numbers shown in the chart?
   f. Are there footnotes to the chart? If so, why? If not, are any needed?

3. For each of the following topics, identify the type of task (e.g., univariate distribution, bivariate association, or relationship among three variables), and types of variables to be presented (nominal, ordinal, interval, or ratio), then state which type of chart would be most suitable, using the guidelines in table 6.1 on pp. 140–41 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. Projected number of people receiving college degrees by region of the country from 2010 to 2025 under three different scenarios about rates of college attendance and completion
   b. Average commuting costs per month, by mode of transportation (bicycle, bus, car, train, walk, other); one number per type of transportation
   c. Number of cases in a study sample from rural, suburban, and urban areas
   d. Educational attainment distribution (<HS, =HS, >HS) for native-born US residents and immigrants from other North American countries, Africa, Asia, Australia and New Zealand, Europe, and Latin America in the year 2000
   e. Estimated odds ratios and 95% confidence intervals for gender, major occupation category (blue collar, white collar, service, other), and region (four major census regions) from a logistic regression of being laid off in the past year
   f. Overall effect of a quadratic specification of percentage body fat in an OLS model of systolic blood pressure (millimeters of mercury [mm Hg])
   g. Overall effects of an interaction between tercile of a student's own high school class rank and their mother's educational attainment (<HS, =HS, >HS) on the student's first-year college grade point average (GPA). Results are based on an OLS regression controlling for gender, race, and family income, using data from the high school classes of 1995 through 2000. The top tercile of each variable in the interaction is the reference category.

4. Use the data in table 5.5 (p. 89 of *Writing about Multivariate Analysis, 2nd Edition*) to create a chart comparing the racial composition of the NHANES III study sample to that of all US births. Include a complete title, labels, legend, and notes.

5. Draft one or more charts to present the findings shown in table 6A.

   a. Use the criteria in table 6.1 on pp. 140–41 of *Writing about Multivariate Analysis, 2nd Edition* to determine which type of chart matches the number and types of variables.
   b. Indicate which variables would go on the axes and which would go in the legend. Hint: Consider whether panels are needed, and if so, which portions of the table go into each panel.

TABLE 6A. Means and standard deviations of psychiatric symptoms by gender and pubertal timing, African American children, 1997 Family and Community Health Study

Psychiatric symptoms	Early maturers (E)	On-time maturers (O)	Late maturers (L)	*F*-statistic	*p*-value	Post-hoc comparisons[a]
Girls	**(*N* = 88)**	**(*N* = 286)**	**(*N* 111)**			
Attention deficit hyperactivity disorder	7.08 (5.62)	5.66 (5.42)	4.78 (4.97)	4.46	0.01	E>O; E>L
Conduct disorder	3.64 (2.11)	3.62 (1.93)	3.49 (2.18)	1.82	0.16	
Generalized anxiety disorder	4.20 (2.59)	3.64 (2.73)	3.37 (2.66)	2.34	0.10	E>O
Major depression	7.19 (4.51)	6.16 (4.83)	5.04 (4.85)	4.96	0.01	E>L; O>L
Oppositional defiance disorder	3.37 (3.16)	2.68 (2.85)	2.07 (2.44)	5.06	0.01	E>O; E>L
Social anxiety	4.62 (2.71)	4.40 (2.78)	4.18 (2.88)	0.60	0.55	
Boys	**(*N* = 77)**	**(*N* = 256)**	**(*N* = 78)**			
Attention deficit hyperactivity disorder	7.41 (5.53)	6.24 (5.38)	4.72 (4.85)	4.81	0.01	E>O; E>L
Conduct disorder	2.39 (2.91)	2.08 (2.95)	1.88 (2.56)	0.61	0.55	
Generalized anxiety disorder	4.27 (2.79)	3.43 (2.76)	2.76 (2.47)	5.81	<0.05	E>O; E>L
Major depression	7.53 (4.62)	5.84 (4.64)	4.97 (4.72)	5.98	0.00	E>O; E>L
Oppositional defiance disorder	3.57 (3.03)	2.92 (3.01)	2.39 (2.86)	2.93	<0.05	E>O
Social anxiety	4.55 (2.47)	3.59 (2.69)	3.28 (2.44)	5.17	0.01	E>O; E>L

Adapted from Ge et al., "Pubertal Maturation and African American Children's Internalizing and Externalizing Symptoms," *Journal of Youth and Adolescence* 35, no. 4 (2006):528–537, table IV.

Notes: Symptoms based on Diagnostic Interview Schedule for Children, Version 4 (DISC-IV). I intentionally changed the order in which the symptoms are listed in the table from that used by the authors. All numbers are unchanged from the article. Post-hoc comparisons are based on $p < 0.05$. "E" = "early maturer," "O" = "on-time maturer," "L" = "late maturer."

c. Include complete titles, axis labels, and footnotes to define terms and indicate statistical significance.
d. In table 6A, the types of psychiatric symptoms are arranged in alphabetical order. What principle(s) would you use to reorganize the order of those symptoms to improve the coordination of the chart with the associated prose? Explain why you chose those criteria, with reference to the guidelines in chapter 6.

6. Use the criteria in chapter 3 to assess the findings in table 6A in terms of

a. The statistically significant findings
b. Substantively meaningful findings
c. The additional information you would need to evaluate causality of the associations

7. Create a stacked bar chart to present the data shown in table 6B, allowing the bar height to vary to show total number of ozone days. To help you plan your chart, answer the following questions, then draw an approximate stacked bar chart, allowing the level to vary by county.

a. Which variable goes on the *x* axis, and what principle would you use to organize its values?
b. Which variable goes in the slices (and legend)?
c. Which variable goes on the *y* axis, and in what units is it measured?
d. What is the title for the chart?

TABLE 6B. Number of unhealthy ozone days by level of warning for selected counties in Indiana, 1996–1998

	Level of warning[a]		
	Unhealthy for sensitive groups	Unhealthy	Very unhealthy
Allen	25	0	0
Clark	29	3	1
Elkhart	15	0	0
Floyd	27	6	0
Hamilton	31	3	0
Hancock	28	2	0
Lake	29	2	0
La Porte	26	6	1
Madison	27	3	0
Marion	32	3	0
Porter	25	3	0
Posey	14	1	0
St. Joseph	21	1	0
Vanderburgh	32	2	0
Vigo	25	1	0
Warrick	40	3	0

[a] Unhealthy for sensitive groups = 0.085–0.104 parts per million (ppm); Unhealthy = 0.105–0.124 ppm; Very unhealthy = 0.125–0.374 ppm.
Source: American Lung Association.

8. Revise your chart from the previous question to illustrate the relative importance (share) of different levels of ozone warning in each county.

    a. What aspects of each chart remain the same as in the previous question? What aspects change?
    b. What are the advantages and disadvantages of the two versions of the chart with reference to this topic and data?

9. Fussell and Massey (2004) used data from the Mexican Migration Project to study relationships among demographic factors, human capital, social capital in the family and community, and migration from Mexico to the United States (table 6C). Use that information to create charts showing the following patterns. Hint: Use a spreadsheet, following the guidelines in appendix D of *Writing about Multivariate Analysis, 2nd Edition*.

    a. The association between age in years and relative odds of first trip to the United States, compared to 15-year-olds. Allow age to vary from 15 to 64 years.
    b. The association between migration prevalence ratio and relative odds of first trip to the United States, with 95% confidence intervals.

TABLE 6C. Estimated log-odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

	Log-odds	Standard error
Demographic background		
Age (years)	−0.003	0.02
Age-squared	−0.001	0.0002
Ever married	−0.09	0.06
Number of minor children in household	0.01	0.01
Human capital		
Years of education	−0.04	0.006
Months of labor-force experience	−0.002	0.0007
Social capital in the family		
Parent a prior US migrant	0.51	0.05
Siblings prior US migrants	0.36	0.02
Social capital in the community		
Migration prevalence ratio[a]		
0–4	−0.99	0.15
5–9	−0.09	0.12
(10–14)		
15–19	0.35	0.10
20–29	0.57	0.13
30–39	0.95	0.15
40–59	0.74	0.19
60 or more	0.34	0.15
Intercept	−3.31	0.26
−2 log likelihood	23,369.2	
df	26	

Source: Adapted from Elizabeth Fussell and Douglas S. Massey, "The Limits to Cumulative Causation: International Migration from Mexican Urban Areas," *Demography* 41, no. 1 (2004): 151–71, table 2. http://muse.jhu.edu/journals/demography/v041/41.1fussell.pdf.

Note: Model also includes controls for occupational sector, internal migratory experience, community characteristics, and Mexican economic and US policy context.

[a] The migration prevalence ratio = (the number of people aged 15+ years who had ever been to the US/the number of people aged 15+ years) × 100.

10. In a study of sexual behavior among youths in Kenya, Mensch and colleagues (2003) evaluated whether audio computer-assisted self-interviewing (ACASI) produces more valid reporting of sexual activity and related sensitive behaviors than face-to-face interviews or self-administered written interviews. Their results are reported in table 6D. Use that information to create charts

   a. to accompany a "Generalization, example, exception" (GEE) description of whether reporting a sensitive behavior differs by mode of interview among boys;
   b. to accompany a GEE description of whether the association between mode of interview and reporting having had more than one sexual partner differs by gender.

TABLE 6D. Odds ratios from logistic regressions of reporting sensitive behaviors, by mode of interview and gender, Kisumu District, Kenya, 2002

Behavior	Boys	Girls
Ever had a boyfriend or girlfriend		
Interviewer-administered	1.00	1.00
Self-administered	0.78	0.82
ACASI[a]	0.43***	0.69*
Ever had more than one sexual partner		
Interviewer-administered	1.00	1.00
Self-administered	1.02	0.72
ACASI[a]	1.28	2.35***
Ever had sex with a stranger		
Interviewer-administered	1.00	1.00
Self-administered	1.43	1.24
ACASI[a]	2.42**	4.25***
Ever tricked/coerced/forced into sex		
Interviewer-administered	1.00	1.00
Self-administered	2.33***	1.89**
ACASI[a]	2.40***	3.35***

Source: Adapted from Barbara S. Mensch, Paul C. Hewett, and Annabel S. Erulkar, "The Reporting of Sensitive Behavior by Adolescents: A Methodological Experiment in Kenya," *Demography* 40, no. 2 (2003): 247–68, table 2. http://muse.jhu.edu/journals/demography/v040/40.2mensch.pdf.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

[a] ACASI = audio computer-assisted self-interviewing.

6. *Creating Effective Charts*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article from your field,

   a. Find a chart that presents the relationship between two variables. Use table 6.1 on pp. 140–41 of *Writing about Multivariate Analysis, 2nd Edition* to assess whether that type of chart is appropriate for the types of variables involved.
   b. Evaluate whether you can understand the meaning of the numbers in the chart based only on the information in the chart. Suggest ways to improve labeling and layout.
   c. Using information in the article, revise the chart to correct those errors.
   d. Consider whether a different chart format would be more effective.
   e. Pick a table from the article. Draft a chart to present the same information, including complete title, axis labels, legend, and notes.

2. Repeat question A.1 with a chart that portrays the relationship among three variables (e.g., two independent variables and a dependent variable).

3. In a journal article from your field, find a chart that presents the relationship between a nominal independent variable with more than two categories, and a dependent variable.

   a. Identify the principle used to organize the categories of the nominal variable on the axis of the chart, with reference to the criteria in chapter 6.
   b. Critique whether that organization coordinates with the associated narrative.
   c. Sketch a revised version of the chart that addresses any shortcomings you identified in part b.

B. Applying Statistics

1. Create a chart to show the frequency distribution of a variable from your data set. See table 6.1 on pp. 140–41 of *Writing about Multivariate Analysis, 2nd Edition* to decide on the most suitable type of chart for that variable's level of measurement.

2. Estimate a difference in means for a continuous dependent variable according to values of a categorical independent variable. Create a chart to present the results, using the checklist in chapter 6.

3. Estimate a logistic regression model of a binary dependent variable as a function of three or four dummy variables. Using the criteria in chapter 6, create a chart to show the 95% confidence intervals around the log-odds estimate for each of the independent variables, including a reference line to convey the null hypothesis.

4. Obtain a copy of the instructions for authors for a leading journal in your field. Revise the charts you created in questions B.1 through B.3 to satisfy their criteria.

C. Writing and Revising

1. Design a chart to portray results of a bivariate analysis involving a nominal independent variable with more than two categories. Specify which principle you would use to decide in what order to display values of the independent variable the on the *x* axis, referring to the criteria in chapter 6 of *Writing about Multivariate Analysis, 2nd Edition*. Explain your choice of organizing principle, with reference to the specific objectives of your analysis.

2. Design a chart to portray the frequencies or mean values of a series of related items (e.g., symptoms, sources of income) in your data set. Specify which of the organizing principle(s) in chapter 6 you would use to organize those items on the *x* axis, and explain your choice:

   a. For a description in the results section of an academic paper;
   b. For a chart to be used as a source of secondary data for other users.

3. Evaluate a chart you created previously for a paper about a multivariate analysis, using the checklist for chapter 6 and the instructions for authors for your selected journal.

4. Peer-edit another student's charts after he or she has revised them, again using the checklist and the instructions for authors for their selected journal.

5. Read through a results section you have written previously. Identify topics or statistics for which to create additional charts such as net effects of interactions or multiterm specifications from your multivariate model. Draft them using pencil and paper, including complete title, labels, legend, and notes.

6. Identify a table or portion of a table in your paper that would be more effective as a chart. Draft that chart, including complete title, labels, legend, and notes.

6. *Creating Effective Charts*

SOLUTIONS

1. Figure 6A is missing a legend; 6B is missing axis titles, axis labels, and units of measurement.

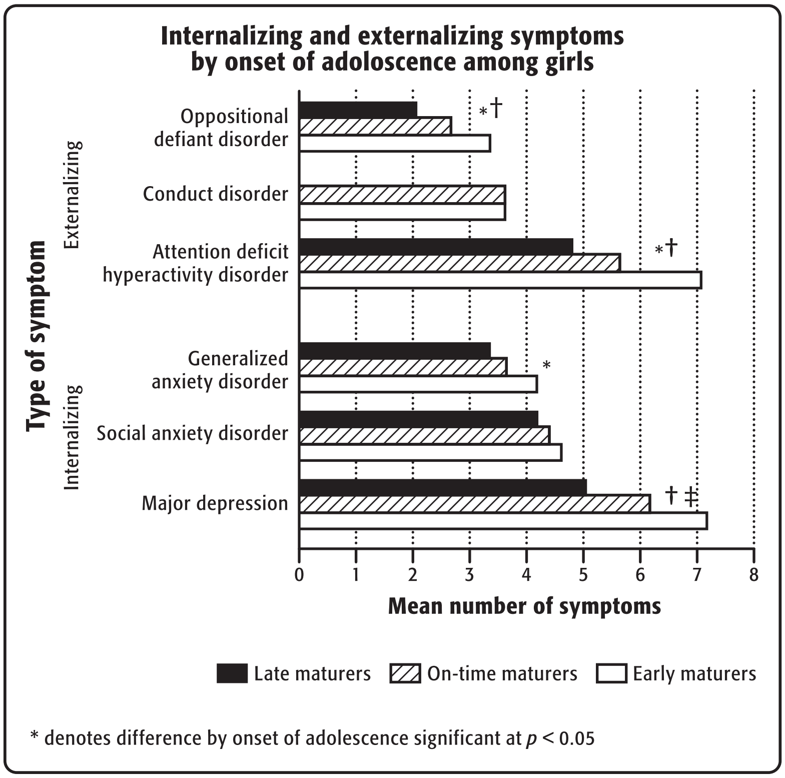
3. Identify the task and types of variables, then state the appropriate type of chart.

   a. Three-way association between one continuous and one ordinal predictor (date and type of scenario, respectively), and a continuous outcome (number of people receiving degrees). Multiple-line chart, to show projected number by date (on the *x* axis) in the number of people receiving college degrees (on the *y* axis), with different lines and line styles for low, medium, and high scenarios (identified in the legend). Notes about data sources and assumptions used in each scenario.
   b. Two-way (bivariate) association between transportation mode (nominal) and cost (continuous). Simple bar chart, with one bar for each transportation mode on the *x* axis and cost on the *y* axis.
   c. Composition (univariate) of a nominal variable. Pie chart to illustrate the percentage (or number of cases) from rural, suburban, and urban areas.
   d. Distribution of one categorical variable (educational attainment) within another categorical variable (continent). Stacked bar chart, with separate bars for US native-born people and each continent of origin, and one slice for each educational attainment level (in the legend). Each bar totals 100% of that continent's immigrants (on the *y* axis) to illustrate composition while correcting for different numbers of immigrants across continents.
   e. Association between several nominal independent variables (gender, occupation, and region) and a continuous dependent variable (relative odds of being laid off in the past year). High/low/close chart ("high" and "low" show the upper and lower 95% confidence limits), with the independent variables on the *x* axis, the odds ratios on the *y* axis, and a reference line at $y = 1.0$.
   f. Association between a continuous independent variable (percentage body fat) and a continuous dependent variable (systolic blood pressure). Single-line chart calculated from the regression coefficients and input values of percentage body fat, with the percentage body fat on the *x* axis and blood pressure on the *y* axis, each labeled with its respective units.

g. Overall effects of an interaction between two ordinal independent variables (tercile of student's class rank and mother's educational attainment) and a continuous independent variable (first-year college GPA). Clustered bar chart with one cluster for each category of mother's education on the *x* axis and a different bar color for each tercile of class rank (in the legend). The *y* axis shows predicted mean first-year college GPA. Notes specifying data source and other variables controlled in the model (or naming a table in which those estimates are shown), identifying the reference categories for class rank and mother's education, and defining symbols used to denote statistical significance.

5. Charts to portray the relationships shown in table 6A.

a. Clustered bar chart with two panels to display the association among type of disorder (six nominal independent variables, one for each of six types), pubertal timing (ordinal categorical independent variable), mean number of symptoms (continuous dependent variable), and gender (nominal independent variable).
b. The type of disorder goes on the *x* axis variables, pubertal timing in the legend, and mean number of symptoms on the *y* axis, with one panel for boys and one for girls.
c. Figures 6C1 and 6C2.
d. In figures 6C1 and 6C2, the types of psychiatric disorders are arranged into one set of clusters for the three internalizing disorders


Internalizing and externalizing symptoms by onset of adoloscence among girls

* denotes difference by onset of adolescence significant at $p < 0.05$

Figure 6C1.

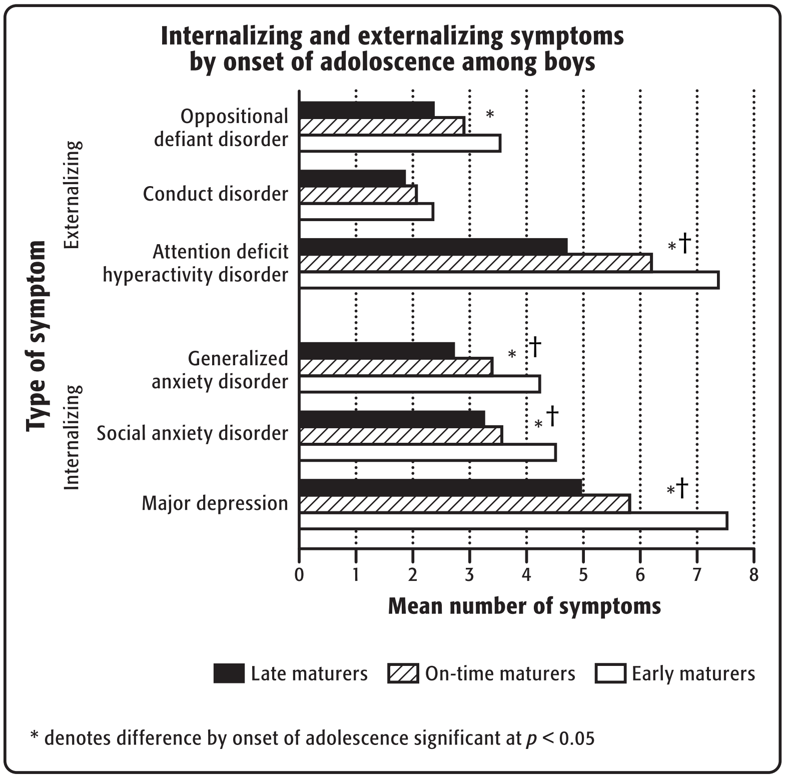


Figure 6C1. and C.2. Mean number of psychiatric symptoms by type of disorder, timing of pubertal maturation, and gender among African American children, 1997 Family and Community Health Study.
Panel 1: Girls
Panel 2: Boys
Source: Ge et al. 2006. "Pubertal Maturation and African American Children's Internalizing and Externalizing Symptoms." *Journal of Youth and Adolescence.* 35(4):528–537. Table IV.
* denotes early maturer > on-time maturer ; † denotes early maturer > late maturer; ‡ denotes on-time maturer > late mature at p < 0.05 based on post-hoc tests.

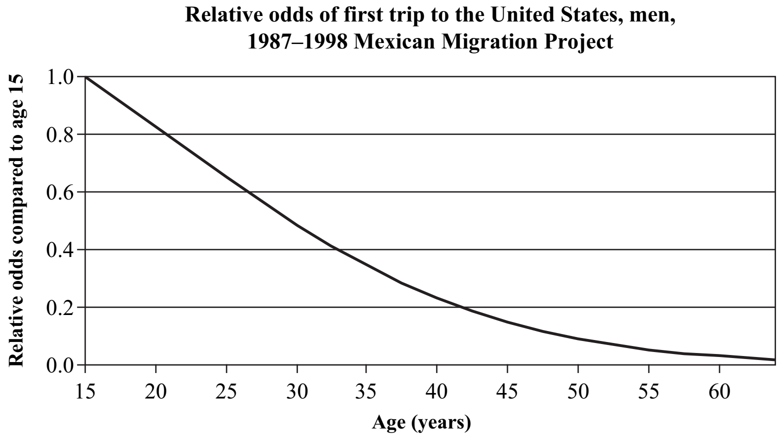
and another set of clusters for the three externalizing disorders, following those conceptual groupings which are mentioned in the title to the article (see footnote to table 6A). The categories of pubertal timing are retained in ordinal sequence, fitting the conceptual meaning of that variable. There is one panel for each gender with one cluster for each type of symptom because that is consistent with the statistical tests, which test whether the mean number of symptoms differ across pubertal timing groups within each gender. Separately within the sets of internalizing and externalizing disorders, the conditions are arranged in descending order of mean number of symptoms.

7. Stacked bar charts, based on the given answers.

   a. Counties arranged on the *x* axis in descending order of total number of unhealthy ozone days
   b. A different color slice for each level of ozone warning, identified in the legend

c. Number of unhealthy ozone days goes on the *y* axis
d. Same title as table 6B: "Number of unhealthy ozone days by level of warning for selected counties in Indiana, 1996–1998"

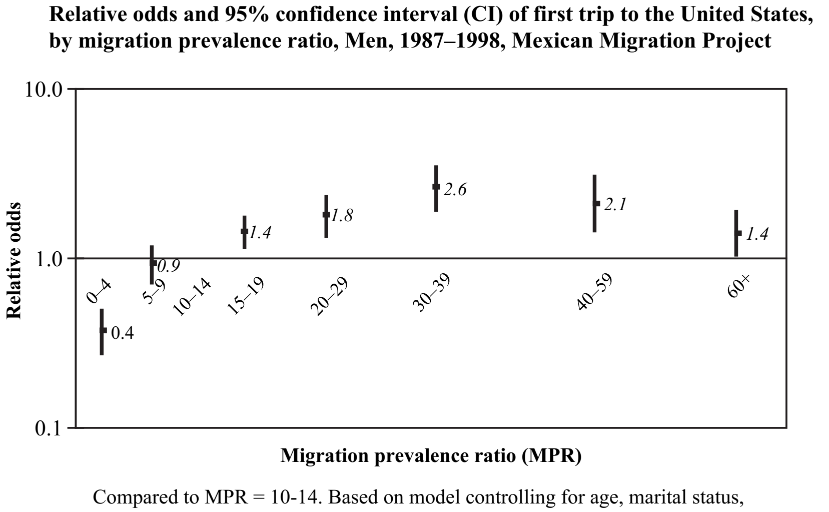
9. Create charts showing the specified patterns from analysis by Fussell and Massey (2004).


Relative odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

Based on model controlling for marital status, number of children, education, labor force experience, family migrant history, and migration prevalence ratio. Reference category = 15 year olds.

Figure 6D.

a. Figure 6D portrays the association between age in years and relative odds of first trip to the United States, compared to 15-year-olds.


Relative odds and 95% confidence interval (CI) of first trip to the United States, by migration prevalence ratio, Men, 1987–1998, Mexican Migration Project

Compared to MPR = 10-14. Based on model controlling for age, marital status, number of children, education, labor force experience, and family migrant history.

Figure 6E.

b. Figure 6E portrays the association between the migration prevalence ratio and relative odds of first trip to the United States, with 95% confidence intervals.

Comments: A logarithmic scale was used to preserve symmetry in apparent sizes of odds ratios above and below 1.0; see "Charts to Display Logistic Regression Results" on pp. 147–49 of *Writing about Multivariate Analysis, 2nd Edition* for an explanation. Spacing of categories on the *x* axis is proportional to actual width of the Migration Prevalence Ratio (MPR) categories: 5-year-wide MPR categories (e.g., 0–4, 15–19) appear half as wide as 10-year-wide MPR categories (e.g., 30–39), which are half as wide as the 20-year-wide MPR category (40–59).

7. *Choosing Effective Examples and Analogies*

PROBLEM SET

1. For each of the following topics, give an analogy to suit a general audience.

   a. A 12-inch snowfall
   b. Two numbers at opposite ends of a distribution
   c. An erratic pattern of change
   d. Something moving rapidly
   e. A few things
   f. Something very heavy
   g. Prices that are rising rapidly
   h. Something that has been level for a long time and then declines suddenly and substantially
   i. A repetitive pattern

2. Repeat the previous question but for a scientific audience in your field.

3. Devise short phrases to convey the concept of small size to the people listed below.

   a. A cooking aficionado
   b. A gardening nut
   c. An artist
   d. A sports fanatic

4. Each of the following analogies would work better for some audiences than others. Name a suitable audience, an unsuitable audience, and an improved analogy for the latter group.

   a. "The size of a Blackberry"
   b. "The gasoline shortage of the early 1970s"

5. For each of the following topics, state whether information from Illinois in 1990 would be useful as a numeric example. If so, give an example of a type of contrast in which that information could be used.

   a. Chicago in 1990
   b. Illinois in 2000

c. Illinois schoolchildren in 1990
d. Iowa voters in 2004

6. Your state is considering three alternative income tax scenarios: a stable tax rate (at 5%), an increase of 0.5 percentage points, and an increase of 1.0 percentage points. Your local representative wants to know how each scenario would affect low-, moderate-, and high-income residents.

   a. What criteria could you use to define "low," "moderate," and "high" income?
   b. What kinds of numeric contrasts would you use to compare the different scenarios?
   c. Create a table to present those effects to the government budget agency.
   d. Create a chart to illustrate the effects to citizens of the state.

7. *Choosing Effective Examples and Analogies*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article in your field,

   a. Circle all analogies or metaphors used to illustrate quantitative patterns or relationships.
      i. Does the author explicitly or implicitly convey the purpose of each analogy or metaphor, or is it left unclear?
      ii. Is it easy to understand the analogy and the pattern or relationship it is intended to illustrate?

   b. Choose one unclear analogy from the paper and revise it, using the principles in chapter 7 of *Writing about Multivariate Analysis, 2nd Edition*.

   c. Are there other places in the article where an analogy or metaphor would be helpful? Identify the purpose of the analogy or metaphor for each such situation.

   d. Design an analogy or metaphor to suit one instance where you have suggested adding one (from part c), using the principles in chapter 7.

   e. Identify the intended audience for the article. Choose a different audience (e.g., more quantitatively sophisticated; younger) and rewrite one analogy to suit them.

2. In the same article, circle all numeric examples where a single number is reported (e.g., not a comparison of two or more numbers). For each, indicate whether the author conveys the purpose of the example (e.g., whether it is a typical or unusual value).

3. In the same article, circle all numeric contrasts.

   a. Indicate whether in each instance the author provides enough information for you to assess whether it is a realistic difference or change for the research question and context.

   b. Evaluate whether different or additional size contrasts would be useful for the intended audience, considering
      i. plausibility;
      ii. real-world application;
      iii. measurement issues.

c. Identify an audience that would be interested in different applications than the audience for whom the article is currently written. Describe how you would select numeric contrasts to meet their interests.

B. Writing and Revising

1. For each of the following audiences, devise an analogy to describe one of the main numeric patterns or relationships in the results section of your paper, using the criteria in chapter 7 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. Readers of a leading journal in your field
   b. Undergraduate students in an intermediate-level substantive course in your field
   c. Readers of the popular press, assuming an eighth-grade reading level
   d. Exchange your answers to parts a through c with someone studying writing about a different topic or data. Peer-edit each other's work and revise according to the feedback you receive.

2. Repeat questions A.1 through A.3 for a paper you have written previously.

7. *Choosing Effective Examples and Analogies*

SOLUTIONS

1. Provide analogies for the given topics.

    a. "Knee deep"
    b. "Polar opposites"
    c. "All over the map"
    d. "Faster than a speeding bullet"
    e. "A handful"
    f. "As heavy as an elephant"
    g. "Going through the roof"
    h. "Like it fell off a cliff"
    i. "Like a broken record"

3. Devise short phrases conveying the concept of small size to the given audience.

    a. "Pea-sized"
    b. "Like a grain of sand or a seed"
    c. "Like a speck of paint"
    d. "Like a drop of water in an Olympic-sized swimming pool"

5. Consider whether information from Illinois in 1990 would be useful for the specified comparison.

    a. Useful for a comparison of the state and its largest city in the same year
    b. Useful for analysis of trends over time in the entire state
    c. Useful for comparison of one age group to the total population
    d. A poor choice, as too many dimensions differ (time, place, and age)

8. *Basic Types of Quantitative Comparisons*

PROBLEM SET

1. Identify the type of quantitative comparison used in each of the following statements:

   a. "Yesterday, New York City received 5.5 inches of snow."
   b. "Ian Thorpe's margin of victory in the 400-meter freestyle was 0.74 seconds."
   c. "A 30-year-old man has 0.59 times the odds of migrating as a 20-year-old man."
   d. "The Dow Jones Industrial Average dropped 0.6% since this morning's opening."
   e. "Women's GPAs are on average 0.26 points higher than men's GPAs."
   f. "Cornstarch has twice the thickening power of flour; for each teaspoon of flour called for in a recipe, substitute one-half teaspoon of cornstarch."
   g. "Median income for the metro region was $31,750."
   h. "Among males, self-esteem averages nearly half a standard deviation unit lower among widowers than among nonwidowers."
   i. "Sixty-eight percent of registered voters turned out for the primary election."
   j. "State U was seeded first in the tournament."

2. In the 2000 presidential election, Al Gore received 50,996,116 votes while George W. Bush received 50,456,169 votes.

   a. Write a sentence to describe the ranks of the two candidates.
   b. Calculate the difference between the numbers of votes each candidate received. What impression does that information alone convey?
   c. Calculate the percentage difference between the numbers of votes each candidate received. What impression does that information give?

3. Indicate whether each of the following statements is correct. If not, rewrite the second part of the sentence to agree with the first.

   a. "Brand X lasts longer than Brand T, with an average lifetime 40% as high as Brand T's."

b. “The unemployment rate increased 25% since last year, from 4.0% to 5.0%.”
c. “The ratio of flour to butter in shortbread is 2:1; the recipe uses twice as much butter as flour.”
d. “At this time of year, reservoirs are usually 90% full. Currently, with reservoirs at 49% of capacity, water levels are only about 54% of normal.”
e. “Nadia’s test score was higher than 68% of students nationwide ($Z = 1.0$).”
f. “A panel of 200 consumers rated ISP A four to one over ISP B. In other words, four more panelists preferred Company A as their Internet service provider.”
g. “Matt is in the top decile for height. He is among the tallest 10% of boys his age.”
h. “The coefficient dropped 15% between the unadjusted and adjusted models, decreasing from 2.0 to 1.7.”
i. “The value of mutual fund ABCD tripled since last year, going from 100 to 33.”

4. In the 1999 Diallo case in New York City, 41 bullets hit the victim. Write down the criteria that you would intuitively use to interpret that number: against what are you comparing the number of bullets?

5. Each of the following statements correctly describes part of table 8A, but each description is incomplete. Fill in the missing information.

TABLE 8A. Median income by race and Hispanic origin, United States, 1999

Race/Hispanic origin	Median income
White	$42,504
Black	$27,910
Asian/Pacific Islander	$51,205
Hispanic (can be of any race)	$30,735

Source: US Bureau of the Census, *Statistical Abstract of the United States*, 2001, table 662.

a. “Asians make about twice as much income.”
b. “Hispanics earn $2,825 more.”
c. “Whites rank second.”
d. “The percentage difference for Asians was 20%.”

6. Use table 8B to perform the tasks listed below.

TABLE 8B. Price per gallon for regular unleaded gasoline at selected gas stations, June 2011 and June 2012

Gas station	June 2011	June 2012
AAA	$1.45	$1.71
Bosco	$1.37	$1.75
Cargo	$1.48	$1.68
Dart	$1.30	$1.66
Essow	$1.46	$1.74

a. Rank the stations from highest to lowest gas price for each of the two dates.
b. Write a description of the distribution of prices in each year. Use difference and ratio in your description to compare the two distributions.
c. Describe how you might use rank in conjunction with difference or ratio in deciding where to buy gas.

7. For each of the phrases listed below, identify other phrases on the list that have the same meaning; write the equivalent dollar value, assuming comparison against a price of $200; and write the corresponding ratio. For statement a, for example, the equivalent dollar value would be $50 and the corresponding ratio would be 0.25.

a. "25% of the original price"
b. "costs 25% less than . . ."
c. "costs 25% more than . . ."
d. "priced 25% off"
e. "125% of the original price"
f. "marked down 75%"
g. "75% of the original price"
h. "costs 75% as much as . . ."

8. The homicide rate in Texas dropped from 16 homicides per 100,000 persons in 1990 to 10 per 100,000 in 1995. Calculate and write sentences to describe

a. the differences between the homicide rates in the two periods;
b. the ratio of the homicide rates in the two periods;
c. the percentage change between the two periods using
   i. the 1990 rate as the denominator;
   ii. the average of the two rates as the denominator.

9. In table 8C, fill in the *z*-score for height for each boy in the sample.

TABLE 8C.1. Heights of a sample of six-year-old boys

Name	Height (cm)	*Z*-score
David	117.51	
Jamal	113.90	
Ryan	124.81	
Luis	115.45	
JC	112.73	

SD = standard deviation (standard population: mean = 115.12 cm; SD = 4.78 cm)

a. Describe how Ryan's, Luis's, and JC's heights compare to the national norms for boys their age based on their *z*-scores. (See table 8.3 in *Writing about Multivariate Analysis*, *2nd Edition* for ways to avoid using the phrase "*z*-scores" as you write).

b. Two boys have heights about equidistant from the mean—one above and one below average. Who are they and about how far are their heights from those of average six-year-old boys? Report the difference in terms of standard deviation units.

c. A new boy, Mike, joins the class. He is one standard deviation taller than the average six-year-old boy. How tall is Mike?

10. One thousand people lived in Peopleland in 2000 and the population was growing at an annual rate (r) of 2.0% per year.

TABLE 8D. Population of Peopleland, 2000–2010

Year	Population	Increase from previous year	Cumulative increase since 2000	Percentage change since 2000
2000	1,000			
2001				
2002				
2003				
2004				
2005				
2006				
2007				
2008				
2009				
2010				

a. Use the formula $P_t = P_0 \times e^{rt}$ to fill the population for each year into table 8D. The year 2000 is year 0, t is the number of years since 2000, r (the annual growth rate, expressed as a proportion) is 0.02, and e is the base of the natural logarithms (2.718).

b. Calculate the increase in population from the preceding year. Write a sentence explaining the pattern of annual population increase across the 10-year period.

c. The cumulative increase is the total number of people added to the population since 2000. How many more people live in Peopleland in 2010 than in 2000?

d. Calculate the percentage change relative to 2000 for each year. Write a sentence to describe the percentage change in population between 2000 and 2010.

e. What is the ratio of the population size for 2010 compared to 2000? How does that ratio relate to the percentage change over that 10-year period?

f. How do the annual rate of growth and the percentage change between 2000 and 2010 relate?

11. Suppose the adjusted odds ratio of hospital admission for diabetics compared to nondiabetics is 3.5.

a. If 5% of the population is diabetic, calculate the attributable risk of hospital admission associated with diabetes.

b. Write a sentence explaining that result without using the term "attributable risk."

8. *Basic Types of Quantitative Comparisons*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a report about recent patterns in mortality, fertility (National Center for Health Statistics website), or unemployment (Bureau of Labor Statistics website).

   a. Identify an example of each of the following: rank, difference, ratio, and percentage difference or change.
   b. For each example, identify the reference value. Does it come from within their data or some other source (e.g., a historic value or a reference population)?
   c. Read the explanations of those examples. Is each one clear? If not, use the criteria outlined in chapter 8 of *Writing about Multivariate Analysis, 2nd Edition* to improve the explanation.
   d. Identify at least one instance where a different (or additional) comparison would be useful. Perform the calculations and write a sentence to present the results.

2. Find a journal article about an application of a multivariate model.

   a. Identify which kinds of basic quantitative comparisons are used to contrast and interpret numeric findings.
   b. Repeat questions A.1b through A.1d for the quantitative comparisons in that article.

B. Applying Statistics

1. For a continuous independent variable from your data set

   a. Identify a pair of values to contrast.
   b. Choose two ways to compare the numbers. Explain your choice of types of quantitative comparisons, with reference to common usage in your field.
   c. Calculate the pertinent comparisons.
   d. Write a paragraph to explain the results of your calculations from part c.
   e. Use the checklist at the end of chapter 8 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate the completeness and clarity of your explanation.

2. List all of the categorical variables used in your multivariate model, either as a dependent or independent variable. For each,

    a. Identify the modal value.
    b. Read the literature to see which value of that variable is most commonly used as the reference category.
    c. Consider the role of that variable in your research question and whether that affects your choice of a reference category.
    d. Cross-tabulate the independent variables to identify the modal categories of the variables in bivariate combination with one another.
    e. Using the information in parts a through d and the criteria in chapter 8, specify which category you will use as the reference category and explain the basis for your choice.

3. Calculate attributable risk for a risk factor and outcome in your data.

    a. Use logistic regression to estimate the relative odds (odds ratio) of a categorical dependent variable for a dichotomous risk factor (independent variable).
    b. In conjunction with information on the prevalence of that risk factor, calculate the attributable risk.
    c. Write a sentence interpreting the results of the attributable risk calculation with reference to the specific variables involved.

C. Writing and Revising

1. Identify a numeric background fact to compare with information for other time periods or cases as part of the introductory section of a research paper.

    a. Select two pertinent types of quantitative comparisons for that fact. Explain your choice, with reference to the topic of your paper.
    b. Look up the relevant data, and calculate the comparisons.
    c. Write a paragraph that integrates those quantitative comparisons, including citations.
    d. Use the checklist at the end of chapter 8 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate the completeness and clarity of your description.

2. Repeat question C.1 for the results section of your paper.

8. *Basic Types of Quantitative Comparisons*

SOLUTIONS

1. Identify the type of quantitative comparison in the given statements.

   a. Value
   b. Difference
   c. Ratio
   d. Percentage change
   e. Difference
   f. Ratio
   g. Rank (median is the 50th percentile)
   h. *z*-score (standardized value)
   i. Value (in this case, the units of measurement are percentage points)
   j. Rank

3. Identify the correct statements; rewrite the incorrect statements to correct them. Bold denotes corrected portion of the sentence.

   a. "Brand X lasts longer than Brand T, with an average lifetime 40% **higher than** Brand T's."
   b. Correct as written.
   c. "The ratio of flour to butter in shortbread is 2:1; the recipe uses twice as much **flour as butter**."
   d. Correct as written.
   e. "Nadia's test score was higher than **84%** of students nationwide ($Z = 1.0$)." (Sixty-six percent are within one standard deviation of the mean [e.g., ± 1 standard deviation], but you must also include those for which $z < -1.0$ to answer this question correctly.)
   f. "A panel of 200 consumers rated ISP A four to one over ISP B. In other words, **four times as many** panelists preferred Company A as their Internet service provider."
   g. Correct as written.
   h. Correct as written.
   i. "The value of mutual fund ABCD tripled since last year, going from **33 to 100**."

5. Fill in the missing information, shown in bold.

   a. “Asians make about twice as much income **as blacks**.”
   b. “Hispanics earn $2,825 more **than blacks**.”
   c. “Whites rank second **in terms of median income, below only Asians and Pacific Islanders**.”
   d. “Asians **earn** 20% **more than whites**.”

7. With a comparison value of $200

   i. The two phrases “25% of the original price” (item a) and “marked down 75%” (f) have the same meaning. Each of those phrases corresponds to a price of $50, equivalent to a ratio of 0.25.
   ii. The phrases “costs 25% less than . . .” (item b), “priced 25% off” (d),”75% of the original price” (g), and “costs 75% as much as . . .” (h) are equivalent. They correspond to a price of $150, equivalent to a ratio of 0.75.
   iii. The two phrases “costs 25% more than . . .” (item c) and “125% of the original price” (e) have the same meaning. They correspond to a price of $250 and a ratio of 1.25.

9. Fill in the *z*-score for height for each boy in the sample.

TABLE 8C.2. Heights of a sample of six-year-old boys

Name	Height (cm)	*Z*-score
David	117.51	0.50
Jamal	113.90	−0.26
Ryan	124.81	2.03
Luis	115.45	0.07
JC	112.73	−0.50

SD = standard deviation (standard population: mean = 115.12 cm; SD = 4.78 cm)

   a. Ryan is approximately two standard deviations above the average height for a six-year-old boy, while Luis is just about average and JC is half a standard deviation below average for his age.
   b. David and JC are half a standard deviation taller and shorter than the average six-year-old boy, respectively.
   c. Mike stands 119.90 cm tall.

11. Answer the questions about attributable risk from the information given.

   a. The attributable risk of hospital admission associated with diabetes is calculated $[0.05(3.5 - 1)]/[(0.05[3.5 - 1]) + 1] \times 100 = 11.1\%$. Prevalence is expressed as a proportion in the calculation.
   b. If diabetes could be eliminated, hospital admissions would decline by 11%.

9. *Quantitative Comparisons for Multivariate Models*

PROBLEM SET

1. Indicate whether each of the following statements is correct. If not, rewrite the second part of the sentence to agree with the first.

   a. "The odds ratio of passing the test was 0.60 for students in School A compared to School B, meaning that students in School A were 60% more likely to pass than those in School B."
   b. "Log-odds of migration for men whose siblings had migrated were 0.51, reflecting higher chances of migration for them than for men whose siblings had not migrated."
   c. "Relative odds of migration for ever-married men were 0.91, reflecting higher chances of migration for ever-married than never-married men."
   d. "The relative risk of divorce for teens compared to older adults was 2.50, corresponding to an excess risk of 150% for teens."
   e. "The relative risk dropped from 2.50 to 2.00 between the unadjusted and adjusted models, corresponding to a 50% reduction in excess risk."

2. For each of the following research questions, indicate whether you would specify an OLS model or a logit model, and identify the units or omitted category of the dependent variable.

   a. Whether income is associated with chances of being arrested.
   b. Whether a new medication decreases average cholesterol levels.
   c. Whether child's IQ varies by parents' IQs.
   d. Whether cohabitation prior to marriage is associated with risk of divorce.

In a 2003 article in the journal *Review of Economics and Statistics*, Zimmerman uses data from Williams College on individual students' grades, their SAT scores, and their roommates' SAT scores to estimate models of peer effects on academic performance (table 9A). Use that information to answer questions 3 through 7 below.

TABLE 9A. Regression of cumulative grade point average by own SAT scores and roommate's SAT scores, Williams College classes of 1999–2001

	Coeff. (s.e.)
Own verbal SAT score/100	0.195
	(0.011)
Own math SAT score/100	0.092
	(0.011)
Race (ref. = white)	
Black	−0.264
	(0.033)
Hispanic	−0.160
	(0.035)
Native American	0.098
	(0.175)
Not a US citizen	0.099
	(0.043)
Asian	−0.085
	(0.022)
Female	0.128
	(0.013)
Roommate's verbal SAT score/100	0.027
	(0.010)
Roommate's math SAT score/100	−0.016
	(0.010)
Sample size	3,151
R^2	0.378

Source: Adapted from David A. Zimmerman, "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *Review of Economics and Statistics* 85, no. 1 (2003): 9–23, table 3. Also available to subscribers at http://weblinks2.epnet.com.

Notes: GPA is on a scale from 0 to 4 points; scores for *each* SAT test (math and verbal) are on a scale from 200 to 800 points in increments of 10 points.

3. For the model shown in table 9A,

   a. Identify the dependent variable, the type of variable (continuous or categorical), its units or coding, and theoretically possible range.
   b. State whether an OLS model or logit model is more suitable for this analysis; explain.
   c. Identify the continuous independent variables, their units as specified in the model, and their theoretically possible ranges.
   d. Identify the categorical independent variables and their reference categories.

4. What is the estimated difference between male and female GPAs? Is that difference statistically significant?

5. What is the difference in predicted GPAs if a student's own verbal SAT score was 720 instead of 680? (Assume the student is in the reference category for all categorical variables in the model and that the other SAT scores are held constant.)

6. What is the difference in predicted GPAs if a student's roommate's math SAT score was 720 instead of 680? (Assume the student is in the

reference category for all categorical variables in the model and that the other SAT scores are held constant.)

7. If the intercept term is 0.780, what would the predicted GPA be for a white male student with a verbal SAT of 720, a math SAT of 700, and a roommate with a verbal SAT of 680 and a math SAT of 650? (Actual intercept terms could not be reported due to confidentiality of students' information.)

Fussell and Massey (2004) used data from the Mexican Migration Project to study relationships among demographic factors, human capital, social capital in the family and community, and migration from Mexico to the United States. Use the information in table 9B to answer questions 8 through 11.

TABLE 9B. Estimated log-odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

	Log-odds	Standard error
Demographic background		
Age (years)	−0.003	0.02
Age-squared	−0.001	0.0002
Ever married	−0.09	0.06
Number of minor children in household	0.01	0.01
Human capital		
Years of education	−0.04	0.006
Months of labor-force experience	−0.002	0.0007
Social capital in the family		
Parent a prior US migrant	0.51	0.05
Siblings prior US migrants	0.36	0.02
Social capital in the community		
Migration prevalence ratio[a]		
0–4	−0.99	0.15
5–9	−0.09	0.12
(10–14)		
15–19	0.35	0.10
20–29	0.57	0.13
30–39	0.95	0.15
40–59	0.74	0.19
60 or more	0.34	0.15
Intercept	−3.31	0.26
−2 log likelihood	23,369.2	
Df	26	

Source: Adapted from Elizabeth Fussell and Douglas S. Massey, "The Limits to Cumulative Causation: International Migration from Mexican Urban Areas," *Demography* 41, no. 1 (2004): 151–71, table 2. http://muse.jhu.edu/journals/demography/v041/41.1fussell.pdf.
Note: Model also includes controls for occupational sector, internal migratory experience, community characteristics, and Mexican economic and US policy context.
[a] The migration prevalence ratio = (the number of people aged 15+ years who had ever been to the US/the number of people aged 15+ years) × 100.

8. Perform these tasks using the information in table 9B.

    a. Identify the dependent variable, the type of variable (continuous or categorical), its units or coding, and theoretically possible range.

b. State whether an OLS model or logit model is more suitable for this analysis; explain.
c. Identify the continuous independent variables, their units as specified in the model, and their theoretically possible ranges.
d. Identify the categorical independent variables and their reference categories.
e. Evaluate whether the authors explained their choice of reference category, and if not, whether you agree with that choice based on the information in the article about substantive considerations and distributions of the variables involved.

9. Assuming all other variables are in the reference category or at their mean values, calculate the relative odds of first migration to the United States for

   a. an ever-married man compared to a never-married man
   b. a 30-year-old man compared to a 20-year-old man
   c. a man with a parent who is a prior US migrant compared to a man without parents who migrated there
   d. a man from a community with a migration prevalence ratio (MPR) of 0–4 compared to one from a community with an MPR of 10–14
   e. a man from a community with a migration prevalence ratio (MPR) of 0–4 compared to one from a community with an MPR of 60 or more

10. Create a table contrasting odds of first trip to the United States at 10-year age intervals from 15 through 64 years; specify the values of the other variables you used in your calculations.

11. Calculate the odds of first migration for a 20-year-old never-married man with no children, eight years of education, 24 months of labor force experience, neither parents nor sibling prior migrants, from a community with a migration prevalence ratio of 10–14.

12. Suppose a study found that the unadjusted odds ratio of hospital admission for diabetics compared to nondiabetics is 3.50.

    a. Calculate the excess risk of hospital admission for diabetics.
    b. When demographic factors and other health conditions are taken into account, the adjusted odds ratio for diabetics is 3.00. Calculate the change in excess risk of hospital admission for diabetics between the adjusted and unadjusted models.

13. Suppose a study found that 20% of nondiabetics were admitted to the hospital.

    a. Using the adjusted odds ratio from the previous question, calculate the corresponding relative risk of hospital admission for diabetics.

b. Express the discrepancy between the odds ratio and the relative risk as a percentage difference.

c. Write a sentence describing the association between diabetes and hospital admission, using the criteria under “An Aside on Relative Risk and Relative Odds” on pp. 204–6 of *Writing about Multivariate Analysis, 2nd Edition*.

9. *Quantitative Comparisons for Multivariate Models*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article in your field that presents results of an OLS model with at least one categorical independent variable and at least one continuous independent variable. Use their results and the criteria in chapter 9 of *Writing about Multivariate Analysis, 2nd Edition* to answer the following questions.

   a. Critique the description of the coefficient on a continuous independent variable in terms of direction, magnitude, statistical significance, and units.
   b. Critique the description of the coefficient on a categorical independent variable.
   c. Evaluate whether the authors explained their choice of reference category for that variable, and whether they provided enough substantive and empirical information to justify their choice.
   d. Rewrite the descriptions of the coefficients to correct any problems you identified in parts a through c of this question.

2. Find a journal article that presents results of a logistic regression of a binary dependent variable, with at least one categorical independent variable and at least one continuous independent variable. Use the results to answer the following questions.

   a. Do they report log-odds or odds ratios? If odds ratios, do they interpret them in terms of multiples of odds or multiples of risk?
   b. Critique the description of the effect size for a continuous independent variable in terms of direction, magnitude, statistical significance, and units, using the criteria in chapter 9.
   c. Critique the description of the effect size for a categorical independent variable.
   d. Rewrite the descriptions to correct any shortcomings you identified in parts b and c.

B. Applying Statistics and Writing

Notes: For the “applying statistics” questions, use variables from your own data or the data sets available with the supplemental online materials to substitute for Y_1, Y_2, X_1, *DUMMY*, and *CATEGVAR* in the models described below. For example, suppose you want to examine factors that predict income. You might use income in dollars as a continuous dependent variable (Y_1), educational attainment in years as a continuous independent variable (X_1), gender as a binary independent variable (*DUMMY*), and residence (urban/suburban/rural) as a multicategory independent variable (*CATEGVAR).* If you wanted to study factors that predict poverty, you might use poverty status (poor/nonpoor) as a categorical dependent variable (Y_2) to estimate logit models with the same set of independent variables.

If possible, choose variables that are part of an ongoing research project. Save the computer output from the models you estimate in questions B.1 through B.3 for use in the exercises for chapters 11, 15, and 16.

1. Using data on a continuous dependent variable (denoted Y_1 in the equations below) and a continuous independent variable (denoted X_1 in the equations below),

   a. Estimate an OLS model of the form $Y_1 = \beta_0 + \beta_1 X_1$ in the original, untransformed units of both the dependent and independent variables, with unstandardized coefficients.
   b. Write a sentence interpreting the value of β_1, referring to the specific independent and dependent variables you have used and specifying the units using the guidelines in chapter 9 of *Writing about Multivariate Analysis, 2nd Edition.*

2. Using data on the dependent variable used in the preceding question and a binary independent variable (denoted *DUMMY* in the equations below, coded 1 for a specified value and 0 for the reference category),

   a. Estimate an OLS model of the specification: $Y_1 = \beta_0 + \beta_1 DUMMY$.
   b. Write a sentence interpreting β_1.
   c. Using the estimated coefficients from part a, calculate predicted values of Y_1 for cases in each category of *DUMMY*. Compare these against the mean value of Y_1 for each of those categories of *DUMMY* from a bivariate calculation.

3. Using data on the same variables used in the two preceding questions, estimate an OLS model of the form $Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 DUMMY$.

   a. Write a sentence interpreting the value of β_1, making sure to specify what else was controlled in the model.
   b. Write a sentence interpreting the value of β_2.

4. Using data from your data set on a dichotomous dependent variable (Y_2), a continuous independent variable (X_1), and a categorical independent variable (*DUMMY*), estimate a logistic regression model of the form logit(Y_2) $= \beta_0 + \beta_1 X_1 + \beta_2 DUMMY$. See your software manual for instructions on how to specify which category of your

dependent variable to model. Using the guidelines in chapter 9 for writing about odds ratios,

a. Write a sentence interpreting the value of β_1.
b. Write a sentence interpreting the value of β_2.
c. Revising

C. Revising

1. Repeat question A.1 for a results section you have written previously that describes results of an OLS regression.

2. Repeat question A.2 for a results section you have written previously that describes results from a logistic regression analysis of a binary dependent variable.

9. *Quantitative Comparisons for Multivariate Models*

SOLUTIONS

1. Correct the given statements, if they are not already correct. Corrections are shown in bold.

   a. "The odds ratio of passing the test was 0.60 for students in School A compared to School B, meaning that students in School A were **only** 60% **as likely to pass** as those in School B." (Or ". . . , meaning that students in School A were 40% **less likely** to pass than those in School B.")
   b. Correct as written.
   c. "Relative odds of migration for ever-married men were 0.91, reflecting **lower** chances of migration for ever-married than never-married men."
   d. Correct as written.
   e. "The relative risk dropped from 2.50 to 2.00 between the unadjusted and adjusted models, corresponding to a **33%** reduction in excess risk."

3. Answer these questions using the information in table 9A (Zimmerman 2003).

   a. The dependent variable is cumulative GPA, a continuous variable measured in points, with a theoretical range from 0.0 to 4.0.
   b. An OLS model is suitable because the dependent variable is continuous.
   c. The continuous independent variables are own and roommate's verbal and math SAT scores, each divided by 100 (see row labels) in the model specification shown in table 9A. Because SAT scores can range from 200 to 800 points, this transformation (change of scale) means that each of these variables could range from 2.0 to 8.0.
   d. The categorical independent variables in the model are gender (ref. = male) and race (ref. = white American citizens, with five dummy variables, one for each of the other racial/citizenship groups [black, Hispanic, Native American, not a US citizen, Asian]).

5. The difference in GPA would be roughly 0.08 points if the student had a verbal SAT score of 720 instead of 680. Calculate this change by

multiplying the coefficient for own verbal SAT (0.195) by the requested difference in SAT score (40 points, divided by 100 in accordance with the model specification). $0.195 \times 0.40 = 0.078$.

7. His predicted GPA would be $2.906 = 0.780 + [(720/100) \times 0.195] + [(700/100) \times 0.092] + [(680/100) \times 0.027] + [(650/100) \times -0.016]$. No terms are needed for race or gender because they are the reference categories, which are captured in the intercept term.

9. Calculate the relative odds of first migration for the given situations using the results in table 9B (Fussell and Massey 2004).

   a. The relative odds of migrating for an ever-married man compared to a never-married man = 0.91. (Exponentiate the coefficient on ever-married; $\exp[–0.09] = 0.91$.)
   b. The relative odds of migrating for a 30-year-old man compared to a 20-year-old man = 0.59. Use the following expression, which plugs a ten-year age difference into the linear and square terms on age: $\exp[(30 \times [-.003]) + (30^2 \times [-0.001])]/\exp[(20 \times [-.003]) + (20^2 \times [-0.001])] = 0.59$.
   c. The relative odds of migrating for a man with a parent who is a prior US migrant compared to a man without parents who migrated there = 1.67. (Exponentiate the coefficient on "parent is a prior US migrant"; $\exp[0.51] = 1.67$.)
   d. The relative odds of migrating man from a community with a migration prevalence ratio (MPR) of 0–4 compared to a man from a community with an MPR of 10–14 = 0.37. (Exponentiate the coefficient on MPR = 0–4; MPR = 10–14 is the reference category; $\exp[-0.99] = 0.37$.)
   e. The relative odds of migrating for a man from a community with a migration prevalence ratio (MPR) of 0–4 compared to one from a community with an MPR of 60 or more = 0.26. (Divide the relative odds for an MPR of 0–4 by the relative odds for an MPR of 60+ to "cancel" the 10–14 MPR reference group; $0.37/1.40 = 0.26$.)

11. The odds of first migration for a 20-year-old never-married man with no children, eight years of education, 24 months of labor force participation, neither parents nor siblings who are prior migrants, from a community with an MPR of 10–14 are calculated $\exp[-3.31 + (20 \times [-0.003]) + (20^2 \times [-0.0001]) + (8 \times [-0.04]) + (24 \times [-0.002])] = 0.016$ or 1.6%. No terms are needed for MPR, marital status, children, or parent or sibling migrants, as those values are all in the reference category.

13. Calculate the odds ratio and relative risk with the following information.

   a. Assuming an odds ratio of 3.0 and a prevalence of the outcome (hospital admission) among the unexposed (nondiabetics) of 0.20, the corresponding relative risk of hospital admission for diabetics = 3.0/[(1.0 – .20) + (3.0 × .20)] = 3.0/[0.8 + 0.6] = 3.0/1.4 = 2.14

b. With an estimated odds ratio of 3.0 and a corresponding relative risk of 2.14, the percentage difference is calculated [3.00 − 2.14]/ 2.14 × 100 = 40%. In other words, the estimated odds ratio overstates the relative risk by 40%.

c. "Diabetics are more than twice as likely as nondiabetics to be admitted to the hospital."

10. *The "Goldilocks Problem" in Multivariate Regression*

PROBLEM SET

1. State whether a one-unit increase would be a useful contrast for each of the following topics. If not, suggest a more reasonable increment.

   a. Annual income (in dollars) for a family of four in the United States in 2004
   b. A Likert scale measuring extent of agreement with a gun control law
   c. Cholesterol level in milligrams per deciliter (mg/dL)
   d. Proportionate increase in the unemployment rate
   e. Hourly minimum wage (in dollars) in the United States in 2004

2. Zimmerman (2003) reports that the mean combined (verbal + math) SAT score for Williams College students in the classes of 1990–2001 was 1,396 points, with a standard deviation of 123. He estimates an OLS regression model of college GPA, with combined SAT score as an independent variable, with the results shown in table 9A. For each of the following situations, select pairs of plausible values of combined SAT scores to use as inputs for an illustration of effect size. Explain your reasoning, keeping in mind that each SAT score (math and verbal) can range from 200 to 800 points in increments of 10.

   a. A sample of students from an elite liberal arts college.
   b. A sample of all high school students nationwide who entered college.

Laditka et al. conducted a multivariate analysis of factors associated with the ambulatory care sensitive hospitalization rate in urban counties in the United States. Answer questions 3 through 8 using the guidelines in chapters 9 and 10 of *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*.

TABLE 10A. Means and standard deviations of variables used in models predicting rate of ambulatory care sensitive hospitalization in US urban counties, 2000

Variable	Mean	Standard deviation
Outcome variables		
County-level ambulatory care sensitive hospitalization rate (ACSH) per 100,000 population, by age group (years)		
Ages 18–39	7.11	3.02
Ages 40–64	20.45	8.54
County-level health system and use factors		
Number of primary care MDs per 100,000 population	71.12	40.19
Number of short-term general hospital beds per 1,000 population	2.75	2.00
Percentage of hospitals that are investor owned	9.10	23.13
Medicaid generosity[a]	1.31	0.30
Number of community health centers	0.43	0.50
Number of emergency department visits per 1,000 population	381.51	177.01

[a] $1,000s of Medicaid expenditures per person under age 65 years below 200% of the poverty threshold.

Adapted from James N. Laditka, Sarah B. Laditka, and Janice C. Probst,. "More May Be Better: Evidence of a Negative Relationship between Physician Supply and Hospitalization for Ambulatory Care Sensitive Conditions," *Health Services Research* 40, no. 4 (2005): 1148–66, tables 2 and 3.

3. Answer the following questions based on table 10A from Laditka et al. (2005):

   a. What is the unit of analysis in this study?
   b. For each of the following variables, report the requested mean value and explain how you calculated it from the information in the table. Hint: What transformation was needed to get from the scale shown in the table to the scale requested in this question?
      i. Primary care MD's per person
      ii. Short-term general hospital beds per person
      iii. Medicaid generosity in dollars
      iv. Emergency room visits per person
   c. With reference to your answers to part b and the concepts covered in chapter 10 of *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*, explain why you think the authors changed the scales of those variables for their analysis.

4. Calculate the value of the ambulatory care sensitive hospitalization rate (ACSH) one standard deviation below the mean and one standard deviation above the mean for

   a. Persons aged 18 to 39 years
   b. Persons aged 40 to 64 years

TABLE 10B. Standardized coefficients[a] from an OLS regression predicting rates of hospitalization for ambulatory care sensitive conditions in US urban counties, 2000

	Standardized coefficients	
Variable	Ages 18–39 years	Ages 40–64 years
County-level health system and use factors		
Number of primary care MDs per 100,000 population	−0.164***	−0.196***
Number of short-term general hospital beds per 1,000 population	0.227***	0.183***
Percentage of hospitals that are investor owned	0.083**	0.072*
Medicaid generosity[b]	−0.066†	−0.064†
Number of community health centers	0.044	0.037
Number of emergency department visits per 1,000 population	0.059	0.056
R^2	0.53	0.62

[a] Model also controls for county racial composition, age composition, crime rate, population density, population growth rate, household composition, household income, disability rate, death rates from heart disease, chronic obstructive pulmonary disease (COPD), diabetes, and liver disease, and for percentage of days with unhealthy air quality.

[b] $1,000s per person under age 65 years below 200% of the poverty threshold.

† $p < 0.05$; * $p < 0.01$; ** $p < 0.001$; *** $p < 0.0001$

Adapted from James N. Laditka, Sarah B. Laditka, and Janice C. Probst,. "More May Be Better: Evidence of a Negative Relationship between Physician Supply and Hospitalization for Ambulatory Care Sensitive Conditions," *Health Services Research* 40, no. 4 (2005): 1148–66, table 4.

5. Write sentences interpreting each of the following coefficients from the model for persons aged 18–39 shown in table 10B. Be sure to specify direction, magnitude, statistical significance, and units for both independent and dependent variables *as specified in the model*:

    a. Community health centers
    b. General hospital beds
    c. Primary care MD physicians
    d. Which variable had the largest effect per standard deviation unit increase?

6. Rewrite each of the sentences from the preceding question, rephrasing the results in the original units (not standardized units) of the dependent variable. Hint: Use the information in table 10A above.

7. Suppose Congress passed a law to add one community health center to every urban county. Write a sentence to predict the effect of that change on the ambulatory care sensitive hospitalization rate holding all other variables constant. Hint: Refer to table 10A to relate a one-unit increase to standard deviations of that independent variable.

8. Write a sentence reporting the effect on the ambulatory care sensitive hospitalization rate of moving from 2.75 to 4.75 short-term general hospital beds per 1,000 county residents.

Xu et al. (2006) analyzed the role of cohabitation in remarriage in the United States in the 1980s. Answer questions 9 and 10 based on the information in table 10C and the guidelines in chapters 9 and 10 of *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*. Hint: Check the form of the dependent variable in the model.

TABLE 10C. Ordinary least squares regression coefficients for a model of waiting time to remarry (years, logged), United States, 1980s

Variable	Estimated coefficient
Intercept	1.843***
Respondent's overall cohabitation	
No cohabitation	
Cohabited prior to first marriage	−0.109
Cohabited prior to remarriage	−0.214***
Cohabitated prior to both marriages	0.028
Respondent's marital history	
Duration of first marriage (years)	−0.017†
Age at first divorce (years)	−0.019*
Residential children at time of divorce	
None	
Minor	−0.035
Adolescent	−0.130
Adult	−0.049

Adapted from Xiaohe Xu, Clark D. Hudspeth, and John P. Bartkowski.. "The Role of Cohabitation in Remarriage," *Journal of Marriage and Family* 68, no. (2006): 261–74, table 2.

Model also controls for gender, race/ethnicity, religious affiliation, employment status, educational attainment, and birth cohort. *F-statistic* = 6.342***; R^2 = 0.079; N = 1, 583.
† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Mean log(waiting time to remarry, years) = 0.901; standard deviation = 1.043

9. Write sentences to interpret each of the following coefficients from table 10C:

   a. Cohabitation prior to remarriage
   b. Duration of marriage
   c. Presence of minor children at the time of first divorce

10. What is the predicted waiting time in years for respondents who did not cohabit prior to either their first marriage or remarriage, had no residential children at the time of divorce, and who were divorced at age 30 after being married for 5 years?

10. *The "Goldilocks Problem" in Multivariate Regression*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article in your field, circle all numeric contrasts.

   a. Indicate whether in each instance the author provides enough information for you to assess whether it is a realistic difference or change for the research question and context.
   b. Review the authors' description and interpretation of coefficients on each continuous variable in their multivariate model, considering whether they
      i. reported associated units for the dependent and independent variables
      ii. stated the size of the contrast used to interpret the size of the coefficient
   c. Evaluate whether different or additional size contrasts would be useful for the intended audience, considering
      i. plausibility;
      ii. real-world application;
      iii. measurement issues.
   d. Identify an audience that would be interested in different applications than the audience for whom the article is currently written, e.g., an applied rather than academic audience. Describe how you would select numeric contrasts to meet their interests.

2. Find a journal article that estimates an OLS model with some continuous and some categorical independent variables. Evaluate whether the authors explicate the coefficients in ways that differentiate those types of variables and their associated scales, using the criteria in chapter 10 of *Writing about Multivariate Analysis*, *2nd Edition*. If not, rewrite the description to rectify those errors.

3. Find a journal article in your field about an application of an OLS model with standardized coefficients for at least two continuous independent variables.

   a. Evaluate whether they discuss why estimate that specification, with reference to the distributions of the variables, the shape of their relationship, or theoretical reasons for their topic.

   b. Evaluate whether they interpret the coefficients in ways that clearly convey the scale and substantive importance of the respective variables in the model.
   c. Evaluate whether the units of the statistical test information are consistent with the units of the standardized coefficients. If not, suggest a correct alternative for presenting statistical test results, using the guidelines in chapter 11.

4. Find a journal article in which the authors estimate models with one or more logarithmic specifications (log-lin, lin-log, or log-log). Review whether the authors

   a. Discuss why they estimate that type of specification, with reference to the distributions of the variables, the shape of their relationship, or theoretical reasons for their topic.
   b. Interpret the coefficients in ways that explicate their units and the shape of the association with the dependent variable.
   c. Rewrite their description of results to address any shortcomings you identified in parts a and b, using the guidelines in chapter 10 and the associated online materials.

5. Find a journal article that presents results of an OLS model involving a quadratic specification for an independent variable.

   a. Critique the description of the coefficient for that variable, using the criteria in chapter 10.
   b. Rewrite the description to correct any shortcomings you identified in part a.

B. Applying Statistics and Writing

1. Calculate and graph the frequency distribution of a continuous independent variable using the highest possible level of detail (e.g., the smallest units for that variable available in your data).

   a. Name the shape of the distribution (e.g., normal, uniform, skewed).
   b. Mark the cutpoints for the quartiles of that variable on the chart.
   c. Mark ± 1 standard deviation (SD) and ± 2 SD on the chart.
   d. Assess the appropriate scale of numeric contrasts for that variable given the precision with which it was collected.
   e. Evaluate whether there is appreciable heaping in the reported values of that variable.
   f. Referring to your answers to parts a through e, explain the criteria you will use to select appropriate values to contrast within your data as you illustrate model findings in your results section.

2. Answer the following questions using the graph you created in the preceding question.

   a. If you wanted to use a categorical version of that independent variable in your model, what does the graph suggest might be empirically appropriate cutpoints between categories? Why?
   b. Read the literature on the relationship between that independent variable and your dependent variable. Are there standard ways to classify the independent variable?
   c. Are there policy-, program-, or other "practical" criteria related to your research question that suggest ways you might classify that variable?
   d. Do the empirical cutoffs you identified in part a match the cutoffs you found for parts b and c? If not, explain which of these criteria you will use to classify your data and why they suit your intended audience.
   e. Design a table or chart to contrast results obtained using the approaches to classifying your independent variable in parts a through c.

3. Complete the "Getting to Know Your Variables" assignment, available in the supplemental online materials.

4. Assess the appropriate scale of numeric contrasts for each of the variables in your analysis given the precision with which those data were collected.

5. Identify several continuous independent variables in your data that have different scales and distributions (e.g., one that is measured in proportions, others that range up to values in the thousands or higher). Investigate the various "Goldilocks solutions" described in chapter 10 of *Writing about Multivariate Analysis, 2nd Edition* for an OLS model involving those variables.

   a. Transforming one or more variables.
   b. Specifying a model with standardized coefficients.
   c. Specifying one of the logarithmic specifications.
   d. For the methods section of your paper, describe how you arrived at your preferred solution and how it affects the definitions of variables or model specifications.

6. Using data on the same variables as in question B.1 of the suggested course extension for chapter 9 (a continuous dependent variable, denoted Y_1 in the equations below, and a continuous independent variable denoted X_1), estimate the following variants of an OLS model. For each, write a sentence interpreting the value of β_1, referring to the variables you have used and specifying the units using

the guidelines on pp. 221–23 of *Writing about Multivariate Analysis, 2nd Edition*.

a. $Y_1 = \beta_0 + \beta_1 X_1$ (in the original, untransformed units of both the dependent and independent variables, but specifying standardized coefficients)
b. $\ln Y_1 = \beta_0 + \beta_1 X_1$ (a log-lin model)
c. $Y_1 = \beta_0 + \beta_1 \ln X_1$ (a lin-log model)
d. $\ln Y_1 = \beta_0 + \beta_1 \ln X_1$ (a double-log model)

7. Using the same variables as in the preceding question estimate an OLS model with a quadratic specification of X_1: $Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$.

a. Calculate the predicted value of Y_1 for selected values of X_1 that span its observed range in your data.
b. Consider whether increments other than a one-unit increase in X_1 are better suited to your research question and data, following the guidelines in chapter 10.
c. Create a chart to show the shape of the estimated relationship between Y_1 and X_1, using the results from part a and following the guidelines in chapter 6.
d. Write a sentence to describe the relationship between Y_1 and X_1 across the observed range of X_1 in your data, using the calculations from parts a or b.
e. Optional: Use a spreadsheet to perform parts a and c, using the quadratic spreadsheet template available online or following the instructions in appendix D.

C. Revising

1. Critique a data and methods section you have previously written, considering each of the following and using the guidelines in chapter 10 of *Writing about Multivariate Analysis, 2nd Edition*:

a. Reporting of the units of measurement for all variables in your analysis.
b. Descriptions of the distributions of all continuous variables, and how those distributions affected the ways in which you specified those variables in your statistical models;
c. Description of the precision of measurement of your variables and the implications for how you analyzed those variables;
d. Explanation of the calculations and reasons for transformations you made to any of the variables, including references to standard transformations or classifications used in your field;
e. Description of your model specification and how it was affected by Goldilocks issues, including references to standard practice in your field;

f. Revise the data and methods section to rectify any shortcomings you identified in parts a through e.

2. Critique a table of descriptive statistics you previously created, using the criteria in chapter 10 to evaluate the following elements:

 a. Labeling of units (system of measurement, units, and scale) and categories for all variables, following the guidelines in chapter 4;
 b. Labeling of all variables measured as proportions, percentages, or rates that correctly conveys their units and scale as used in the multivariate model specification;
 c. Labeling of all transformed variables that correctly conveys their units and scale or categories as used in the multivariate model specification;
 d. Pertinent measures of central tendency and distribution for each variable, given its level of measurement;
 e. Revise the table to rectify any shortcomings you identified in parts a through d.

3. Critique a table of multivariate regression results you previously created, considering each of the following as explained in chapter 10:

 a. Labeling of units and categories for all variables;
 b. Labeling of all variables measured as proportions, percentages, or rates that correctly conveys their units and scale as used in the multivariate model specification;
 c. Labeling of all transformed variables that correctly conveys their units and scale or categories as used in the multivariate model specification;
 d. Title, row or column headings, or footnotes to convey the model specification (e.g., standardized or unstandardized coefficients, logarithmic specification);
 e. Revise the table to rectify any shortcomings you identified in parts a through d.

4. Critique a description you have previously written about results of an OLS model with several continuous independent variables with different ranges and scales of values.

 a. Evaluate whether you specified the size of the contrast used to interpret the coefficients for each continuous variable.
 b. Consider whether a one-unit contrast was suited to the interpretation of the coefficients for each of those variables, based on the criteria in chapter 10.
 c. Revise the description to rectify any shortcomings you identified in parts a and b.

5. Critique a description you have previously written about results of an OLS model with at least one continuous independent variable and one categorical dependent variable, considering whether you clearly conveyed

   a. the nature of the contrast that suited each type of independent variable;
   b. for ordinal variables, the substantive meaning of a one-unit increase (from one category to the next);
   c. whether you directly compared the size of coefficients on categorical and continuous variables;
   d. Revise the description to rectify any shortcomings you identified in parts a through c.

6. Critique a description you have previously written about a quadratic association between one of your independent variables and your dependent variable, using the guidelines in chapter 10.

   a. Assess whether a chart would complement the narrative description. If so, create one, using the guidelines in chapter 6 and the spreadsheet template available online or the instructions in appendix D.
   b. Revise the description to improve the shortcomings you found.

7. Critique and rewrite a description you have previously written about estimated coefficients from one or more types of logarithmic specifications, using the guidelines in chapter 10.

8. Critique and rewrite a description you have previously written about estimated coefficients from a model with standardized coefficients, using the guidelines in chapter 10.

9. Exchange revised drafts of the materials in questions C.1 through C.8 with someone writing about a different topic or data set. Peer-edit each other's work and revise according to the feedback you receive.

10. *The "Goldilocks Problem" in Multivariate Regression*

SOLUTIONS

1. State whether a one-unit increase is a useful contrast for the specified topics and if not, give alternatives.

    a. Too low to be of substantive interest. Use increments of $1,000 instead.
    b. Reasonable.
    c. Too low to be clinically meaningful or measured precisely. Use an increment of 10 mg/dL.
    d. Too high. An increase of one unit would span the entire theoretically possible range. Use an increase of 0.05 or 0.10.
    e. Reasonable.

3. Answer the following questions based on table 10A from Laditka et al. (2005):

    a. The unit of analysis is the county, as shown in the title and row labels for the outcome and independent variables.
    b. For each of the following variables, report the requested mean value and explain how you calculated it from the information in the table. Hint: What transformation was needed to get from the scale shown in the table to the scale requested in this question? Rephrase it to show the rate or value per person.
        i. Mean primary care MDs per person = 0.000711. Divide the number shown in the table (scaled per 100,000 population) by 100,000. By taking the reciprocal of that number, we calculate that there was roughly one primary care MD for every 1,406 people in the counties studied in the year 2000—an alternative way to express the concept, e.g., in a discussion section.
        ii. Mean short-term general hospital beds per person = 0.00275. Divide the number shown in the table (scaled per 1,000 population) by 1,000. By taking the reciprocal of that number, we calculate that there was one hospital bed for every 363 people, on average, in the counties studied.
        iii. Mean Medicaid generosity in dollars = $1,310 per person under aged 65 below 200% of the poverty threshold. Multiply the number shown in the table (which is in multiples of $1,000s) by 1,000.

iv. Mean emergency room visits per person = 0.38151, which rounds to 0.38. Divide the number shown in the table (scaled per 1,000 population) by 1,000. By taking the reciprocal of that number we calculate that on average about one out of every three people visited the ER in the year 2000 in the urban counties studied.

c. Of these four continuous measures of health system capacity and use, the values per *person* range from well below zero to several thousand. For example, with means of 0.0007 and 0.002 for primary doctors per person and general hospital beds per person, respectively, a change of one unit in that original scale would be far too large, because the observed variation is detectable only in the third or fourth decimal place. When planning for health system capacity, these are the scales in which those concepts are conventionally discussed and analyzed using the scales shown in the table.

For Medicaid generosity per person with a mean value over $1,000, a $1 increase would be too small to be of interest, so it is conventionally analyzed in $1,000's or perhaps $100's. These changes were thus made to accommodate a combination of theoretical and empirical considerations and common usage.

5. Write sentences interpreting each of the following coefficients from the model for persons aged 18–39. Be sure to include direction, magnitude, statistical significance, and units for both independent and dependent variables *as specified in the model*:

a. A one standard deviation unit increase in the number of community health centers per county was associated with a 4.4 percent higher rate of ambulatory care sensitive hospitalization (ACSH) among persons ages 18 to 39 years, but the difference was not statistically significant. (Reminder: 0.044 of a standard deviation is equal to 4.4%. Multiply the standardized coefficient by 100 to convert it from multiples of a standard deviation into percentage points.)

b. For the same age group ASCH rates were approximately 23% lower for each one standard deviation unit increase in the number of short-stay general hospital beds per 1,000 county residents ($p < 0.001$).

c. A one standard deviation unit increase in the number of primary care MDs per 100,000 county residents was associated with a 16% lower rate of ACSH ($p < 0.001$).

d. Number of short-stay general hospital beds had the largest effect of those three variables, as gauged by effects of a one standard deviation unit increase in each variable on the ASCH rate in the model for 18 to 39 year olds.

7. For CHCs (community health centers), a one-unit increase is two standard deviations (1 SD = 0.50, as shown in table 10A). Multiplying the estimated standardized coefficient for CHCs by two, we have

$0.044 \times 2 = 0.088$. Based on the results of the model, the addition of one CHC per county would be expected to be associated with nearly a 9% increase in the ACSH, although that effect is not statistically significant.

9. Write sentences to interpret each of the following coefficients from the model of waiting time to remarry. The model is specified with logged income, so the percentage change in waiting time to remarry for each one unit increase in the independent variable is calculated $(e^{B}-1) \times 100$:

   a. Taking into account a range of socioeconomic and demographic factors, respondents who cohabited prior to their remarriage waited on average about 24% longer to remarry than those who did not cohabit before either marriage ($p < 0.001$). Note: By exponentiating the intercept from table 10C, we can calculate that the mean waiting time in the reference category for the overall model was 6.3 years. Multiplying that by 24% and converting to months, we can restate the finding for cohabitation as follows: "Taking into account a range of socioeconomic and demographic factors, respondents who cohabited prior to their first marriage waited on average about 18 months longer to remarry than those who did not cohabit before either marriage."
   b. For each additional year that a respondent's first marriage had lasted, waiting time to remarry was reduced by about 1.7%, or about 1.3 months ($p < 0.10$). For example, persons whose first marriage lasted for 20 years would be predicted to remarry just over a year faster than those whose first marriage lasted for 10 years, all else equal.
   c. Presence of minor children at the time of the respondent's first divorce was associated with a 3.5% shorter waiting time to remarry, or about 2.5 months less than those without residential children at the time of divorce, but the difference was not statistically significant.

11. *Choosing How to Present Statistical Results*

PROBLEM SET

Answer questions 1 through 3 using the information in table 11A .

TABLE 11A. Estimated coefficients and standard errors from a model of cumulative grade point average by own SAT scores and roommate's SAT scores, stratified by own SAT score, Williams College classes of 1999–2001

	Student's own combined math & verbal SAT score		
	Lowest 15%	Middle 70%	Top 15%
Own verbal SAT score/100	0.205	0.199	0.118
	(0.039)	(0.015)	(0.055)
Own math SAT score/100	0.065	0.112	0.045
	(0.036)	(0.017)	(0.051)
Race (ref. = white)			
Black	−0.181	−0.386	−0.800
	(0.046)	(0.053)	(0.059)
Hispanic	−0.036	−0.254	−0.050
	(0.059)	(0.046)	(0.274)
Native American	−0.238	0.212	dropped
	(0.169)	(0.168)	
Not a US citizen	0.076	0.126	0.055
	(0.091)	(0.055)	(0.066)
Asian	0.210	−0.065	−0.201
	(0.120)	(0.026)	(0.047)
Female	0.262	0.103	0.107
	(0.038)	(0.016)	(0.028)
Roommate's verbal SAT score/100	0.006	0.043	−0.013
	(0.025)	(0.012)	(0.021)
Roommate's math SAT score/100	−0.038	−0.021	0.030
	(0.028)	(0.012)	(0.022)
Sample size	450	2,072	629
R^2	0.41	0.27	0.21

Source: Adapted from David A. Zimmerman, "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *Review of Economics and Statistics* 85, no. 1 (2003): 9–23, table 4

1. For the estimated coefficient on female gender among students with combined SATs in the lowest 15%

   a. What is the t-statistic?
   b. What is the 95% confidence interval?
   c. What is the 99% confidence interval?
   d. What is the p-value based on a two-tailed test?

e. If * denotes $p < 0.05$ and ** denotes $p < 0.01$, what symbol would accompany the "female" coefficient?

2. Among students in the middle 70% of combined SAT scores, which of the following differences in GPA are statistically significant?

   a. That between black and white students
   b. That between black and Hispanic students
   c. That between Hispanic and Native American students
   d. What additional information (if any) do you need to conduct a formal statistical test for these differences?

3. Answer the following questions using the information in table 11A.

   a. Three models are shown in table 11A. How do they differ? How can you tell from the table?
   b. Is the relationship between gender and GPA statistically significantly different across categories of own combined SAT score?
   c. What additional information (if any) do you need to conduct a formal statistical test for this difference?

Answer questions 4 through 8 using the information in table 11B.1.

TABLE 11B.1. Median income (constant 1999 $) by type of household, United States, 1998 and 1999

	1998		1999	
Type of household	Median income	90% confidence interval (+/−)	Median income	90% confidence interval (+/−)
Family households	48,517	419	49,940	449
Married-couple families	55,475	541	56,827	502
Female householder, no husband present	24,932	669	26,164	594
Male householder, no wife present	40,284	1,670	41,838	1,311
Nonfamily households	23,959	477	24,566	444
Female householder	19,026	472	19,917	454
Male householder	31,086	572	30,753	568
All households	39,744	387	40,816	314

Source: US Census Bureau, *Current Population Report*s, P60–209, *Money Income in the United States: 1999* (Washington, DC: US Government Printing Office), table A.

4. What are the lower and upper 90% confidence limits for 1998 median income for all households?

5. Is the change in real household income between 1998 and 1999 statistically significant at $p < 0.10$

   a. For all households?
   b. For family households?
   c. For nonfamily households?

6. What is the standard error associated with the 1998 estimate of median income for nonfamily households with a female householder? Explain how you calculated it.

7. Calculate 95% confidence intervals around estimated median income for each household type in table 11B.1 and show the results in a new table. Hints: Use the critical value for $p < 0.10$ based on a large sample to calculate the standard error of each estimate. Then multiply the standard error by 1.96 to obtain the 95% CI. A spreadsheet vastly simplifies these calculations.

8. Create a table that shows change in median income for each household type between 1998 and 1999, denoting differences that are statistically significant at $p < 0.10$ with a dagger.

Answer questions 9 and 10 using the information in table 11C from Fussell and Massey (2004).

TABLE 11C. Estimated log-odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

	Log-odds	Standard error
Demographic background		
Age (years)	−0.003	0.02
Age-squared	−0.001	0.0002
Ever married	−0.09	0.06
Number of minor children in household	0.01	0.01
Human capital		
Years of education	−0.04	0.006
Months of labor-force experience	−0.002	0.0007
Social capital in the family		
Parent a prior US migrant	0.51	0.05
Siblings prior US migrants	0.36	0.02
Social capital in the community		
Migration prevalence ratio[a]		
0–4	−0.99	0.15
5–9	−0.09	0.12
(10–14)		
15–19	0.35	0.10
20–29	0.57	0.13
30–39	0.95	0.15
40–59	0.74	0.19
60 or more	0.34	0.15
Intercept	−3.31	0.26
−2 log likelihood	23,369.2	
Df	26	

Source: Adapted from Elizabeth Fussell and Douglas S. Massey, "The Limits to Cumulative Causation: International Migration from Mexican Urban Areas," *Demography* 41, no. 1 (2004): 151–71, table 2. http://muse.jhu.edu/journals/demography/v041/41.1fussell.pdf.

Note: Model also includes controls for occupational sector, internal migratory experience, community characteristics, and Mexican economic and US policy context.

[a] The migration prevalence ratio = (the number of people aged 15+ years who had ever been to the US/the number of people aged 15+ years) × 100.

9. For the estimated coefficient on “ever-married,” calculate

    a. The test statistic (name it)
    b. The *p*-value
    c. The 95% confidence interval for the coefficient (e.g., the 95% CI around the log-odds point estimate)

10. Revise table 11C to report odds ratios with associated 95% confidence intervals and symbols to denote statistical significance instead of log-odds and standard errors.

11. *Choosing How to Present Statistical Results*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article in your field about an application of an OLS model.

   a. Which approaches to presenting statistical significance results do the authors use?
   b. Do the authors label those approaches adequately in the text (e.g., identifying the type of test statistic)? In the tables?
   c. If the authors used more than one approach to presenting statistical significance results, are those approaches complementary or redundant with one another?
   d. Would a different or additional approach be more suitable for that intended audience based on the criteria in table 11.3 in *Writing about Multivariate Analysis, 2nd Edition*? If so, name it and, if the information in the article is sufficient, calculate it for each variable in one of their models.
   e. Do the authors mention whether their statistical tests are one-tailed or two-tailed?
   f. Do the authors specify the number of degrees of freedom for their models?

2. Does the article used in question A.1 address any hypotheses *other than* the null hypothesis (e.g., $\beta_i = \beta_j$, or tests across models)?

   a. If so, do the authors provide information such as test statistics or *p*-values to formally test those hypotheses? Are their explanations of those hypothesis test results clear?
   b. If they do not test other hypotheses, are there others that would suit their main research question? If you had access to their data, what approach would you use to present results of those hypothesis tests to the same audience?

3. Find a journal article in your field about an application of a logistic regression of a binary dependent variable.

   a. Which approaches to presenting statistical significance results do the authors use?
   b. Are the units of the statistical test information consistent with the units in which they present the effects' estimates (log-odds or odds

ratios)? If not, suggest a correct alternative for presenting statistical test results.

4. Obtain a copy of a leading journal in your field.

   a. Which approaches to presenting statistical significance results are specified in the instructions for authors for that journal?
   b. If they do not specify a particular approach to presenting statistical significance, which ones are mostly widely used in the journal?
   c. Critique those choices, given the intended audience for that journal and the guidelines in table 11.3.

5. Find a report about a survey in your field or at websites such as the Census Bureau or Bureau of Labor Statistics.

   a. Which approaches to presenting statistical significance results are used?
   b. Who is the intended audience for that report or websites?
   c. Do the approaches used to present statistical significance suit that audience?

B. Applying Statistics

Note: These questions use the regression output from the "applying statistics" questions in the suggested course extensions to chapter 9. See notes to those questions for additional information about the types of variables and notation used below.

1. Using the OLS regression output from question B.3 in the suggested course extensions for chapter 9, identify or calculate each of the following for each of the coefficients in the model. Most of these pieces of information can be requested as part of the computerized output.

   a. The standard error
   b. The test statistic (name it)
   c. The *p*-value based on a two-tailed test
   d. The *p*-value based on a one-tailed test
   e. The 95% confidence interval
   f. The 99% confidence interval
   g. The symbol denoting level of statistical significance, assuming a two-tailed test, if ** denotes $p < 0.01$ and * denotes $p < 0.05$.

2. Create tables to present results of the OLS model in the preceding question for each of the following audiences or objectives, using the criteria in chapters 5 and 11 and appendix B of *Writing about Multivariate Analysis, 2nd Edition*:

   a. A paper to be submitted to the journal you used in question A.4
   b. A 15-page report for a nonstatistical audience interested in the issues you study

c. A five-minute presentation to a lay audience interested in the issues you study

3. Estimate an OLS regression using a continuous dependent variable Y_1 and a three-category independent variable *CATEGVAR* from which you have created two dummy variables (denoted CAT_1 and CAT_2 in the equations below); label your dummy variables to reflect their actual content!

   a. Estimate a model of the form $Y_1 = \beta_0 + \beta_1 CAT_1 + \beta_2 CAT_2$, requesting the variance-covariance matrix for the model.
   b. Perform a ballpark assessment of whether $\beta_1 = \beta_2$, using the approach described on p. 246 of *Writing about Multivariate Analysis, 2nd Edition*.
   c. Use information from the variance-covariance matrix to calculate the test statistic for whether $\beta_1 = \beta_2$, following the instructions under "Differences between Coefficients from the Same Model" on p. 244.
   d. Write a sentence to report the conclusions of that test, with reference to the specific variables and concepts in your model.
   e. Reestimate the same model as in part a, requesting a formal statistical test for $\beta_1 = \beta_2$. Compare your conclusion based on this approach to your conclusion based on the method used in part c.

4. Using the logistic regression output from question B.4 in the suggested course extensions for chapter 9, identify or calculate each of the following for each of the coefficients in the model. Most of these pieces of information can be requested as part of the computerized output.

   a. The standard error
   b. The test statistic (name it)
   c. The *p*-value based on a two-tailed test
   d. The *p*-value based on a one-tailed test
   e. The 95% confidence interval for the coefficient (e.g., the 95% CI around the log-odds point estimate)
   f. The odds ratio
   g. The 95% confidence interval for the odds ratio
   h. The symbol denoting level of statistical significance, assuming a two-tailed test, if ** denotes $p < 0.01$ and * denotes $p < 0.05$

5. Create tables to present results of the logistic regression model from the preceding question for each of the following audiences or objectives, using the criteria in chapters 5, 11, and 20, and appendix B.

   a. A paper to be submitted to the journal you used in question A.4
   b. A 15-page report for a nonstatistical audience interested in the issues you study

c. A five-minute presentation to a lay audience interested in the issues you study

C. Writing and Revising

1. Repeat questions A.1 and A.2 for a results section you have written previously that describes results of an OLS regression.

2. Revise or create tables to present results of that OLS model for each of the following audiences or objectives, using the criteria in chapters 5, 11, and 20, and appendix B.

   a. A paper to be submitted to a leading journal in your field
   b. A 15-page report for a nonstatistical audience interested in the issues you study
   c. A five-minute presentation to a lay audience interested in the issues you study

3. Repeat question A.3 for a results section you have written previously that describes results of a logistic regression analysis of a binary dependent variable.

4. Repeat question C.2 with the results of that logistic regression.

11. *Choosing How to Present Statistical Results*

SOLUTIONS

1. For the estimated coefficient on female gender among students with combined SATs in the lowest 15%

   a. The t-statistic = 6.985 (= coefficient/standard error = 0.262/0.038).
   b. The 95% confidence interval is 0.188, 0.336 (= 0.262 ± [1.96 × 0.038]).
   c. The 99% confidence interval is 0.165, 0.359 (= 0.262 ± [2.56 × 0.038]).
   d. $p < 0.001$ based on the t-statistic of 6.99 and criteria for a large sample.
   e. ** would accompany the "female" coefficient.

3. Answer these questions using the information in table 11A (Zimmerman 2003).

   a. There is one model for each of three subsamples of combined own SAT score: students in the bottom 15% of the Williams College SAT range, those in the middle 70%, and those in the top 15%. This information is presented in the column spanner ("Student's own combined math & verbal SAT score") and column headers.
   b. The coefficient for "female" is statistically significantly higher in the bottom 15% of SAT scores (0.262, s.e. = 0.038) than for the other two groups ($\beta = 0.103$, s.e. = 0.016, and $\beta = 0.107$, s.e. = 0.028 for the middle 70% and top 15% of SAT scores, respectively). The difference between the lower and middle groups, for example, is calculated 0.262 − 0.103 = 0.159. The corresponding standard error of the difference = $\sqrt{(0.038)^2 + (0.016)^2}$ = 0.041. Dividing the difference between coefficients by the standard error of the difference, we obtain 0.159/0.041, or a t-statistic of 3.86, which exceeds 2.56, the critical value of the test statistic for $p < 0.01$ for a sample of this size. However, the difference between the female coefficients for the upper two SAT groups is not statistically significant because the difference (−0.004 = 0.103 − 0.017) is swamped by the standard error of the difference.
   c. No additional information is needed to conduct a formal statistical test of this difference. The estimates and their standard errors are

independent of one another because they are from separate (stratified) models. Hence we do not need to take the covariances into account, as would be necessary with interaction terms between gender and SAT group estimated within one model that pooled all SAT groups.

5. Consider real household income as reflected in table 11B.1.

   a. Yes, the change in real household income between 1998 and 1999 for all households is statistically significant at $p < 0.10$. The upper 90% CL for 1998 median income for all households ($40,131) is below the lower 90% CL for the corresponding figure for 1999 ($40,502). Hence the 90% confidence intervals for the respective years do not overlap, so the increase in median income from $39,744 to $40,816 is significant at $p < 0.10$. Because the estimates for the two years are independent, the covariance between estimates does not need to be taken into account when performing the test.
   b. Yes, the change in real household income between 1998 and 1999 for family households is statistically significant at $p < 0.10$. The upper 90% CL for 1998 median income for family households ($48,936) is below the lower 90% CL for the corresponding figure for 1999 ($49,491). Same logic as for part a.
   c. No, the change in real household income between 1998 and 1999 for nonfamily households is not statistically significant. The upper 90% CL for 1998 median income for nonfamily households ($24,436) is above the lower 90% CL for the corresponding figure for 1999 ($24,122). Hence the 90% confidence intervals for the two estimates overlap, and we cannot conclude that they are statistically significantly different at $p < 0.10$.

7. The multiplier (critical value) for $p < 0.10$ and a large sample size is 1.64, so we divide the reported ± values from the 90% CI by 1.64 to

TABLE 11B.2. Median income (constant 1999 $) with 95% CI, by type of household, United States, 1998 and 1999

	1998				1999			
Type of household	Median income	Standard error	Lower 95% CL	Upper 95% CL	Median income	Standard error	Lower 95% CL	Upper 95% CL
Family households	48,517	255	48,016	49,018	49,940	274	49,403	50,477
Married-couple families	55,475	330	54,828	56,122	56,827	306	56,227	57,427
Female householder, no husband present	24,932	408	24,132	25,732	26,164	362	25,454	26,874
Male household, no wife present	40,284	1,018	38,288	42,280	41,838	799	40,271	43,405
Nonfamily households	23,959	291	23,389	24,529	24,566	271	24,035	25,097
Female householder	19,026	288	18,462	19,590	19,917	277	19,374	20,460
Male householder	31,086	349	30,402	31,770	30,753	346	30,074	31,432
All households	39,744	236	39,281	40,207	40,816	191	40,441	41,191

obtain the standard error (s.e.) of each estimate. Then calculate the 95% CL as estimate ± (1.96 × s.e.), as shown in table 11B.2.

9. For the estimated coefficient on "ever-married,"

   a. The test statistic is the chi-square $(\chi^2) = (\beta_k/\text{s.e.}_k)^2 = (-0.09/0.06)^2 = 2.25$.
   b. $p < 0.10$.
   c. The 95% confidence interval for the coefficient (e.g., the 95% CI around the log-odds point estimate) = $-0.208, 0.028 = -0.09 \pm (1.96 \times 0.06)$.

12. *Writing Introductions, Conclusions, and Abstracts*

PROBLEM SET

Answer questions 1 through 4 for a scientific paper about AIDS knowledge for different language groups in the United States (results shown in table 5.2 on p. 85 of *Writing about Multivariate Analysis*, *2nd Edition*). Assume you are writing for a social science journal with a 5,000-word limit for research articles (e.g., several double-spaced pages apiece for the introduction, literature review, and conclusion).

1. For your scientific paper,

   a. Write an outline of the introduction, including complete topic sentences for each major paragraph.
   b. List the kinds of numeric background information you would incorporate, and suggest useful types of quantitative comparisons to highlight why the topic is interesting or important.

2. Write an outline of the literature review, including headings for the different topics for which you would summarize published literature.

3. Write an outline of the concluding section, including notes on the following issues.

   a. How would you summarize the main numeric results?
   b. How would the statistical significance of findings influence the way you discuss the results?
   c. List the types of numeric background information you would use to show how findings of that study might be applied to health education programs in the United States.

4. Write a title, abstract, and keywords for the paper.

5. Write one or two paragraphs discussing the research implications of Mensch and colleagues' (2003) findings about how mode of interview relates to reporting of sensitive behaviors among adolescents (table 12A).

TABLE 12A. Odds ratios from logistic regressions of reporting sensitive behaviors, by mode of interview and gender, Kisumu District, Kenya, 2002

Behavior	Boys	Girls
Ever had a boyfriend or girlfriend		
Interviewer-administered	1.00	1.00
Self-administered	0.78	0.82
ACASI[a]	0.43***	0.69*
Ever had more than one sexual partner		
Interviewer-administered	1.00	1.00
Self-administered	1.02	0.72
ACASI[a]	1.28	2.35***
Ever had sex with a stranger		
Interviewer-administered	1.00	1.00
Self-administered	1.43	1.24
ACASI[a]	2.42**	4.25***
Ever tricked/coerced/forced into sex		
Interviewer-administered	1.00	1.00
Self-administered	2.33***	1.89**
ACASI[a]	2.40***	3.35***

Source: Adapted from Barbara S. Mensch, Paul C. Hewett, and Annabel S. Erulkar, "The Reporting of Sensitive Behavior by Adolescents: A Methodological Experiment in Kenya," *Demography* 40, no. 2 (2003): 247–68, table 2. http://muse.jhu.edu/journals/demography/v040/40.2mensch.pdf.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

[a] ACASI = audio computer-assisted self-interviewing.

12. *Writing Introductions, Conclusions, and Abstracts*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article about an application of multivariate analysis to a topic in your field. Critique it for the following, using the guidelines in chapters 3 and 12 of *Writing about Multivariate Analysis, 2nd Edition*:

   a. A clear introduction of the main substantive issues or questions to be investigated
   b. A review of the previous literature to identify theories and existing evidence on that topic
   c. A discussion and conclusions section that summarizes numeric findings and relates them back to the research question and to previous studies
   d. Consideration of causality and substantive significance of findings in the conclusions

2. Critique the abstract, keywords, and title to the article using the guidelines in chapter 12 and the instructions for authors for a leading journal in your field. Revise them to correct any problems you identify.

B. Writing

Note: If you are writing a paper on a new multivariate analysis, complete these questions. If you have already written a draft of your paper, complete section C instead.

1. Write an introductory section for your paper following the guidelines in chapter 12 of *Writing about Multivariate Analysis, 2nd Edition*.

2. Write the discussion and conclusions section of your paper, including

   a. summaries of major numeric findings rather than repetition of detailed numbers from the results;
   b. discussion of causality, statistical significance, and substantive significance of findings; see guidelines in chapter 3;

c. explanation of how your findings relate to initial hypotheses and to findings of other studies.

3. Write an abstract for your paper following the guidelines in chapter 12 and the instructions for a journal in your field.

4. Investigate which online databases list the leading journals in your field. Write keywords to satisfy the criteria for that database, following the guidelines in chapter 12 and the instructions for that database.

5. Write a title for your paper

    a. worded as a statement;
    b. worded as a rhetorical question.

C. Revising

1. Critique the introductory section of a paper you have written previously, using the criteria in question A.1. Rewrite it to rectify any problems you have identified.

2. Critique the discussion and conclusions section of that paper using the criteria listed under question B.2. Rewrite it to rectify any problems you have identified.

3. Repeat question A.2 for your paper.

4. Exchange initial and revised drafts of the materials in questions C.1 through C.3 with someone writing about a different topic or data. Peer-edit each other's work and revise according to the feedback you receive.

12. *Writing Introductions, Conclusions, and Abstracts*

SOLUTIONS

1. For a scientific paper about AIDS knowledge

   a. Outline of introduction to study of AIDS knowledge by language in the United States.

      I. (Paragraph on why AIDS is of concern) Introductory sentence: "AIDS (Acquired Immunodeficiency Syndrome) is a leading cause of death in the United States."

      II. (Paragraph on why it is important to assess AIDS knowledge) Introductory sentences: "In the absence of a vaccine against AIDS, prevention must rely on individual behavior to avoid spread of the disease. It is unlikely that appropriate behavioral change will occur without knowledge about AIDS and how it is transmitted; hence it is important to assess levels of AIDS knowledge in the general population."

      III. (Paragraph on why language is an important possible mechanism related to AIDS knowledge) Introductory sentence: "Language can affect AIDS knowledge either through linguistic barriers or cultural differences."

   b. Kinds of numeric information to incorporate, and useful quantitative comparisons for an introduction to the AIDS knowledge paper.

      For paragraph I, statistics on levels and trends in AIDS prevalence and mortality in the United States, using percentage change to quantify trends over time in AIDS prevalence and mortality rates, and rank as a cause of death to indicate overall importance.

      For paragraph II, evidence on how knowledge about AIDS or other similar diseases such as STDs translates into changes in preventive behaviors.

      For paragraph III, statistics on how AIDS prevalence and mortality vary by language ethnic group, with supplementary evidence by race or socioeconomic status if statistics by language are not available. Use ratios or percentage difference to contrast rates across groups. Information on the number of persons, percentage share, and trends in number and share of major language groups in the United States.

3. Outline of conclusion to study of AIDS knowledge by language in the United States:

   a. Summarize the main numeric results.
      I. Summary of differences in AIDS knowledge by language group using GEE technique (English speakers did best, Spanish/Spanish did worst; example of size of differences for a representative AIDS topic)
      II. Synthesis of which AIDS topics were best understood, least well understood using GEE approach (reporting percentage of respondents who answered questions correctly within broad conceptual groupings of AIDS knowledge topics, generalized across language groups where possible)
      III. Description of how these knowledge patterns correspond to which topics are most important for people to understand (e.g., correct information about likely means of transmission is more essential than correct information about unlikely means of transmission)

   b. Discuss only statistically significant differences across language groups or AIDS knowledge topics. For topics where language differences in knowledge were not statistically significant, describe overall knowledge levels only, not differences across groups.

   c. To show how these results might be used to evaluate or influence health education programs, include statistics from other studies about
      i. the availability of education materials that emphasize the most important AIDS transmission topics
      ii. the association between AIDS knowledge and preventive behaviors
      iii. availability in the US of AIDS education materials in Spanish and other non-English languages

5. Research implications of the findings in table 6D (Mensch et al. 2003). "This study has shown that method of data collection has a substantial effect on reported levels of sensitive behaviors among adolescents. Teens were more likely to report normative behaviors such as having a boyfriend or girlfriend if questioned in person than using audio computer-aided self-interview (ACASI) techniques. For sensitive (stigmatized) behaviors such as having been coerced into sex, however, ACASI yielded higher rates than in-person interviews.

   "These results have several implications for future research on similar topics and populations. First, the method of data collection should be chosen to maximize the chances of subjects reporting their true behavior instead of responding in ways that conform to perceived social norms about acceptability of that behavior. Second, results should not be compared across sources that used different methods of data collection, because apparent differences (or lack of differences) across groups could be attributable to reporting biases rather than differences in actual behavior."

13. *Writing about Data and Methods*

PROBLEM SET

1. For each of the following scenarios, list what information you would report in a data section for a scientific paper. Hint: What additional information would you want to know?

   a. A three-year study of a six-month drug rehabilitation program that recruited 200 subjects to examine cure and relapse rates
   b. A study of calcium intake among 50 pregnant women, based on their recall over a two-week period

2. Dr. Dollar is conducting a study of poverty patterns in the United States based on annual income data from the 2000 census. She defines a categorical measure of income group comparing family income (calculated from income of individual family members, alimony, and four types of social benefits) against the federal poverty thresholds. Classifications are defined in terms of multiples of the threshold: <.50, .50–.99, 1.00–1.84, 1.85–2.99, and 3.00 or greater. Search for "poverty" on the US Census web page (http://www.census.gov) for more details. State how you would describe the poverty measure in

   a. a one-page summary of the study for a local newspaper;
   b. documentation of a new data set that has collected data on each of the income components as part of a written questionnaire;
   c. a journal article on poverty patterns, written for people who are familiar with poverty thresholds.

3. Making use of newly available data from a three-year panel study of a sample of 10,000 people drawn from the 2000 census, Dr. Dollar describes movement in and out of poverty and duration of poverty (in months) over the study period. Poverty was defined as family income below the threshold (<1.0). Data were collected annually, with retrospective recall of income in each of the previous 12 months. What information should be added to item 2.c to describe these data for this research question?

4. A researcher at the Panel Study of Income Dynamics accidentally erased a file containing information from two years' worth of data. Embarrassed, he went ahead and analyzed data for the other 30 years

in the study. What assumptions did he implicitly make about the missing data?

5. Fauth et al. (2004) studied the effects of a residential mobility experiment, comparing outcomes of low-income adults who moved to public housing in low-poverty neighborhoods with outcomes for those who stayed in public housing in their original high-poverty neighborhoods. They studied the six neighborhood and housing quality measures shown in table 13A. What information about these variables should be included in a data section for a scientific paper about this study?

TABLE 13A. Means and standard deviations of neighborhood and housing characteristics, Yonkers Residential Mobility Program, 1994–1995

Measure[a]	Mean	Standard deviation
Danger	0.72	0.91
Number of victimizations in past year	0.21	0.58
Disorder	0.72	0.74
Cohesion	0.52	0.32
Resources	2.98	0.60
Housing problems[b]	0.35	0.43

Source: Adapted from Rebecca C. Fauth, Tama Leventhal, and Jeanne Brooks-Gunn, “Short-Term Effects of Moving from Public Housing in Poor to Middle-Class Neighborhoods on Low-Income, Minority Adults’ Outcomes,” *Social Science and Medicine* 59 (2004): 2271–84, table 1. http://www.sciencedirect.com/science.

[a] Ranges of values for the neighborhood and housing quality measures are: Danger: 0 to 3; disorder: 0 to 5; cohesion: 0 to 4; resources: 0 to 5; housing problems: 0 to 5.

[b] In the published paper, this measure was termed “housing quality,” but I relabeled it “housing problems” to reduce confusion because a higher value indicates more problems, e.g., with rats and mice.

6. For each of the following data, methods, and objectives, write a short discussion of strengths and limitations for the concluding section of a scientific article.

   a. Study: 20 subjects were interviewed at the Snooty Golf Club at noon on a Friday in early April regarding their preferred color and fit of jeans. Objective: a marketing study by the Gap clothing store.
   b. Study: two classes of second graders in the same school were given a math test in September. One class was then taught with a new math curriculum, the other with the standard curriculum. The classes were tested again in June. Objective: an evaluation of the new math curriculum.
   c. Study: data on hair color and age were collected for everyone aged 25–84 in a city of 200,000 people. Deaths over a two-year period were ascertained from death certificates. Two models were estimated: one with hair color as the independent variable and mortality as the dependent variable; the second with age as another independent variable. Objective: understand the potential benefit of hair dye in improving survival.

7. In her study "Gender, Preloss Marital Dependence, and Older Adults' Adjustment to Widowhood," Carr (2004) uses data on respondents who were widowed between waves 1 and 2 of the study, matched to control subjects who remained married at wave 2. (See table 16A for more on her study.) Carr's study used data from a longitudinal study over a seven-year period. In her methods section, she describes a model of attrition (nonparticipation in wave 2) from the sample between waves 1 and 2. She found that "age and anxiety increased the risk of nonparticipation, and home ownership decreased the risk of nonparticipation at wave 2." None of the other demographic, socioeconomic, or health characteristics were associated with attrition.

    a. Write an equation to convey her final specification for the model of attrition, including the dependent variable and type of model estimated.
    b. What questions is she trying to answer with that model?
    c. Write a short discussion of the implications of her attrition findings for interpretation of her results about psychological adjustment to widowhood.

13. *Writing about Data and Methods*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a one- or two-page article in the health or science section of a newspaper or magazine, circle the information on data and methods.

   a. Critique the presentation of that information, using the guidelines in chapter 13 of *Writing about Multivariate Analysis, 2nd Edition* regarding writing about data and methods in general-interest articles for a lay audience.
   b. Assess whether additional information would be helpful for people seeking information to compare with findings from another study.
   c. Evaluate the authors' discussion of how the data and methods affect interpretation of the findings.

2. Read the data and methods section from an article about an application of OLS regression in a journal from your field.

   a. Critique it, using the guidelines in chapter 13 regarding writing about data and methods for scientific articles.
   b. List additional information needed by researchers seeking to replicate the data collection protocol.
   c. List additional information needed by researchers seeking to replicate the statistical analysis.
   d. Assess how well the article discusses how the data and methods affect interpretation of the findings.
   e. Indicate whether the authors suggest directions for future research.
   f. Rewrite the description of data and methods in the discussion to rectify problems you identified in parts d and e.

3. Read the methods section of an article about an application of logistic regression in a journal from your field.

   a. Evaluate whether the categories of the dependent variable were defined in the raw data or calculated by the authors. If the latter, indicate whether the authors specified the criteria or cutoffs used to perform the classification.
   b. Indicate whether the authors identify the omitted category of the dependent variable in the text and the tables.

4. Go to a data website such as the US Census Bureau, National Center for Health Statistics, or the Bureau of Labor Statistics and identify a topic of interest involving two or three variables. Evaluate the website in terms of how easy it is to find information about

    a. the type of study design (e.g., cross-sectional sample survey, retrospective, prospective);
    b. the data sources (e.g., vital registration forms, questionnaires, administrative records);
    c. the wording of questions used to collect the variables of interest to you;
    d. the units or coding of those variables;
    e. sampling weights, if applicable;
    f. the response rate;
    g. loss to follow-up (for longitudinal studies only).

B. Writing

1. Outline the data section for a scientific paper about a multivariate analysis you are conducting, using the checklist in chapter 13 of *Writing about Multivariate Analysis, 2nd Edition*.

2. Write an equation to convey your final model specification.

3. Write an explanation of why you chose the type of statistical model used in your analysis given your research question and data, following the guidelines in chapter 13.

4. Write an explanation of how you arrived at your final model specification, including the following topics:

    a. The criteria used to determine which variables were included in the model, with reference to your specific research question.
    b. Whether and why nonlinear specifications were used for any of the independent variables.
    c. Whether interactions were included among two or more independent variables, and if so, which ones and why; see also chapter 16.

5. Write a discussion of the strengths and limitations of your data and methods for a scientific audience.

6. Exchange your answers to questions B.1 through B.5 with someone studying a different topic or data. Peer-edit each other's work and revise according to the feedback you receive.

7. Write a short discussion for a lay audience about how strengths and limitations of your data and methods affect how your study's findings should be interpreted and applied in a real-world context, following the guidelines in chapters 13 and 20.

C. Revising

1. Critique a data and methods section of a scientific paper you have written previously, using the criteria in chapter 13 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. Identify elements you have omitted.
   b. Track down that information in data documentation or other publications on the same data.
   c. Identify material that could be organized better or explained more clearly.
   d. Revise your data and methods section to fix the problems you identified in parts a and c.

2. Critique the discussion of data and methods in the discussion section of a scientific paper you have written previously, using the guidelines in chapter 13.

   a. Identify implications of strengths or limitations of the data that were omitted or explained poorly.
   b. Identify directions for future research related to your data and methods that were omitted or explained poorly.
   c. Revise your discussion section to correct the problems you identified in parts a and b.

3. Exchange your answers to questions C.1 and C.2 with someone studying a different topic or data. Peer-edit each other's work and revise according to the feedback you receive.

4. Exchange data and methods sections with someone who is analyzing different data and a different research question. Using only the information in that section (e.g., without reference to their computer output or data documentation, and without asking them any questions),

   a. Write an equation to express their final model specification (or a selected model if several models are presented in the paper). If some of the information needed to write an equation is missing information, list it.
   b. Identify the units or coding and omitted categories for each variable in the final model specification (or a selected model) based on the data section and tables of descriptive statistics. If any of this information is missing, unclear, or inconsistent between the tables and prose, list it.
   c. Rewrite your data and methods section to correct the problems identified by your peer-editor.

13. *Writing about Data and Methods*

SOLUTIONS

1. Information you would report in a data section for a scientific paper for the specified studies.

    a. What were the demographic characteristics (when? where? who?) of those in the study? How were subjects recruited? What was the baseline response rate among recruits? What percentage of the initial sample was lost to follow-up and how? How did the sample compare demographically to all clients at that rehab center? How were "cure" and "relapse" defined and measured? By whom were these assessments made?

    b. Again, the W's. How were they recruited, what was the response rate, and how did the sample compare to all pregnant women? Were they asked specifically about calcium intake or to list foods? Were open- or closed-ended questions asked about food?

3. Loss to follow-up, how income data were collected (using what methods and data sources? total or by components? in what ranges? continuous or categorical?).

5. Information on each of the items used to comprise each of the six outcome measures (dependent variables) and the method of data collection. Information on the development, reliability, and validity of those items. For example, what was the wording of the three items included in the "danger" scale? How were they coded? From what sources were those items drawn? Are those three items the standard measure of danger in other related studies? If not, how were they developed? Were they pretested on similar populations?

7. With regard to the analysis of attrition in Carr's study on widowhood:

    a. Logit(attrition) $= \beta_0 + \beta_1$Age $+ \beta_2$Anxiety $+ \beta_3$Homeowner.

    b. Whether those who participated at wave 2 were representative of the original wave 1 sample in terms of major sociodemographic and health characteristics.

c. "An analysis of attrition showed that older respondents, those with higher anxiety, and those who did not own their homes were more likely to drop out between waves 1 and 2. As a consequence, these results about psychological adjustment to widowhood may not be generalizable to people in those groups because they were under-represented in the sample used in this analysis."

14. *Writing about Distributions and Associations*

PROBLEM SET

1. Write descriptions of the following tables from *Writing about Multivariate Analysis, 2nd Edition*:

   a. the age, gender, and racial distributions shown in table 5.3 (p. 88);
   b. the distribution of major categories of federal outlays in figure 6.2b (p. 116).

2. Write a description of the race/household type associations in table 5.1 (p. 80) using the GEE approach. Hint: To compare across racial/ethnic groups, report percentage distribution of household type within each race. Why are percentages preferred to counts in this case?

3. Use the results from Zimmerman's (2003) analysis of cumulative college grade point averages (GPAs) shown in table 11A on p. 85 of this study guide to answer the following questions.

   a. Among students in the middle 70% of SAT scores, the coefficient for "female" is 0.107 with a standard error of 0.016. Write a sentence explaining the direction, magnitude, and statistical significance of that finding.
   b. Among students in the bottom 15% of SAT scores, the coefficient for the variable "roommates' math SAT score/100" is −0.038 with a standard error of 0.028. Write a sentence interpreting that finding, assuming that roommates' math SAT scores range from 400 to 800.

4. Write a description of the age pattern of mortality shown in figure 6.10 (p. 128) in *Writing about Multivariate Analysis*, *2nd Edition*. Use descriptive phrases to convey the shape of the pattern, then document with appropriate numeric evidence. Incorporate selected quantitative comparisons to illustrate the sizes of differences in the chart.

5. In the analysis conducted by Mensch et al. (2003), the association between mode of interview and odds of boys reporting a sensitive

behavior differs by the type of behavior in question (table 12A on p. 97 of this study guide). What is such a pattern called in statistical terms? In GEE lingo? Write paragraphs to describe that pattern to

a. a group of first-year high school students;
b. a group of graduating statistics majors.

14. *Writing about Distributions and Associations*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. In a journal article in your field, find descriptions of univariate distributions for each of the following types of variables. Critique them, using the criteria in chapter 14 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. A nominal variable
   b. An ordinal variable
   c. An interval or ratio variable with many possible values

2. For each of the descriptions in question A.1

   a. Identify the criteria the author used to select which value(s) to highlight. Does that value match the research question and introductory material in the article?
   b. If all values are described with equal emphasis, assess whether one or more values should be featured and explain why.

3. In a journal article in your field, find descriptions of each of the following types of bivariate associations. Critique them, using the principles in chapter 14.

   a. An association between two categorical variables
   b. An association between a categorical independent and a continuous dependent variable
   c. Bivariate correlations among a series of continuous variables

B. Applying Statistics and Writing

1. Using variables from your data set, run frequency distributions on one nominal, one ordinal, and one interval or ratio variable.

   a. Write a brief description of each distribution, emphasizing the modal value. Summarize, then report key indicators of central tendency.
   b. Write a second description of each distribution, this time highlighting a value of interest other than the mean or mode, such as a minority group, unusual value, or most recent value.

2. Using variables from your data set, calculate one example of each of the following types of bivariate associations. Write a brief description of each pattern, using the principles in chapter 14.

   a. An association between two categorical variables
   b. An association between a categorical independent variable and a continuous dependent variable
   c. Bivariate correlations among a series of continuous variables

3. Using variables from your data set, run a three-way association among two categorical independent variables and a continuous dependent variable. Write a description of that association using the GEE approach explained in chapters 2 and 14 and appendix A.

4. Using variables from your data set, run a three-way cross-tabulation of two categorical independent variables and a categorical dependent variable. Write a description of that association using the GEE approach explained in chapters 2 and 14 and appendix A.

C. Revising

1. Critique and rewrite descriptions of univariate statistics (distributions, central tendency) from a paper you have written previously, using the criteria in chapter 14 of *Writing about Multivariate Analysis, 2nd Edition*.

2. Critique and rewrite descriptions of bivariate statistics (cross-tabulations, differences in means, or correlations) from the same paper.

3. Critique and rewrite a description of a three-way association from a results section you have written previously, using the GEE approach explained in chapters 2 and 14 and appendix A.

4. Exchange drafts of your answers to questions C1 through C3 with a colleague who is working with a different topic and data. Peer-edit each other's work and revise according to the feedback you receive.

14. *Writing about Distributions and Associations*

SOLUTIONS

1. Descriptions of the specified tables and charts.

   a. "Table 5.3 shows the demographic composition of the study sample. Just over half of the 2,058 respondents were female (51.4%). Persons aged 40 to 64 years were the largest single age group in the sample (41.4%), just edging out persons aged 18–39 (37.8%). Elderly persons (aged 65 and older) made up about one-fifth of the sample.

      "The most common racial/ethnic group was non-Hispanic whites, with 2½ times as many respondents as the second largest racial/ethnic group, non-Hispanic blacks (55.6% and 22.1%, respectively). Hispanics comprised the third-largest group (15.9%), followed by Asians (4.2%) and persons of other racial/ethnic origin (2.2%)."

   b. "In 2000 in the United States, human resources comprised by far the largest single category of federal outlays (61% of the $1.8 trillion spent that year; figure 6.2b). The second largest category—national defense—accounted for only about one-quarter as much as human resources (16% of the total). Net interest, physical resources, and other functions together comprised the remaining 23% of all outlays."

3. Use the results from table 11A (Zimmerman 2003) to answer the given questions.

   a. "Among Williams College students with SAT scores in the middle 70%, women's GPAs averaged 0.11 points higher than men's ($p < 0.01$)."

   b. "Among students in the bottom 15% of SAT scores, there was no significant association between roommate's math SAT score and student's college GPA. Although the estimated coefficient suggests a GPA 0.15 points lower if roommate's math SAT were 400 instead of 800, the finding was not statistically significant."

5. In statistical terminology, a situation where the association between one independent variable (mode of interview) and the dependent variable (odds of reporting a sensitive behavior) depends on a second

independent variable (type of behavior) is called an interaction or effect modification. In GEE lingo, it is called an exception.

a. Description of the pattern for a group of first-year high school students: “A recent study in Kenya found that the chances of reporting specific sensitive behaviors such as having had sex with a stranger or being coerced into sex differed depending on how the data were collected (table 12A). For the three most sensitive topics studied, boys were more likely to report having experienced those behaviors if they were interviewed using a self-administered computer-aided interview than if interviewed in person. On the other hand, they were more likely to report ever having had a girlfriend if interviewed in person than if they completed a computer-aided self-interview.”
b. Description of the pattern for a group of graduating statistics majors: “A study by Mensch and colleagues of teenagers in Kenya found that method of data collection and type of sensitive behavior interact in their effect on odds of reporting sensitive behaviors such as having had sex with a stranger or being coerced into sex (table 12A). For the three most sensitive topics studied, the odds of reporting those behaviors were highest among boys interviewed using a self-administered computer-aided interview and lowest among those interviewed in person. In contrast, for the topic ‘ever having had a girlfriend’ the odds were highest among boys interviewed in person and lowest among those completing a computer-aided self-interview.”

15. *Writing about Multivariate Models*

PROBLEM SET

Fauth et al. (2004) studied the effects of a residential mobility experiment, comparing outcomes of low-income adults in public housing who moved to low-poverty neighborhoods to those who stayed in their original, high-poverty neighborhoods. "Movers" were chosen by lottery from among those who applied for the program. Their results are summarized in tables 15A (bivariate statistics) and 15B (multivariate model results). Use those data to answer questions 1 through 3.

TABLE 15A. Individual background characteristics, neighborhood, and housing characteristics of movers and stayers, Yonkers Residential Mobility Program, 1994–1995

	Residential status			
	Movers ($n = 173$)	Stayers ($n = 142$)	Total ($n = 315$)	χ^2 or *F*[a]
Background characteristics				
Age (mean years)	36.69	34.07	35.51	6.45**
Female	97%	96%	97%	0.41
Latino (ref. = black)	31%	25%	28%	1.07
At least high school education	67%	53%	61%	6.62**
Female household head	76%	85%	80%	4.39*
Mean # children in household	1.72	2.01	1.85	6.04*
Neighborhood/housing[b]				
Danger	0.26	1.29	0.72	144.11***
# of victimizations in past year	0.12	0.32	0.21	9.21*
Disorder	0.15	1.41	0.72	796.17***
Cohesion	0.62	0.40	0.52	43.48***
Resources	3.05	2.89	2.98	4.90*
Housing problems[c]	0.20	0.54	0.35	54.40***

Source: Adapted from Rebecca C. Fauth, Tama Leventhal, and Jeanne Brooks-Gunn, "Short-Term Effects of Moving from Public Housing in Poor to Middle-class Neighborhoods on Low-Income, Minority Adults' Outcomes," *Social Science and Medicine* 59 (2004): 2271–84, table 1. http://www.sciencedirect.com/science.

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

[a] χ^2 statistic reported for difference in categorical variable between movers and stayers; *F*-statistic for difference in continuous variable.

[b] Ranges of values for the neighborhood and housing quality measures are as follows: Danger: 0 to 3; disorder: 0 to 5; cohesion: 0 to 4; resources: 0 to 5; housing problems: 0 to 5.

[c] In the published paper, this measure was termed "housing quality," but I relabeled it "housing problems" to reduce confusion because a higher value indicates more problems, e.g., with rats and mice.

1. Answer the following questions based on the information in table 15A:

   a. Did the random assignment succeed in equalizing the background characteristics of movers and stayers? Write a paragraph summarizing the similarities and differences in background characteristics between those two groups.
   b. Did neighborhood and housing characteristics differ according to residential status (e.g., for movers versus stayers)? Write a paragraph generalizing these findings.
   c. What do these statistics suggest about the need for multivariate models of these outcomes by residential status? Explain your reasoning.

2. Write a paragraph describing the results in table 15A, using your answers to question 1 and the principles on p. 312 of *Writing about Multivariate Analysis, 2nd Edition* for building the case for a multivariate model.

3. Write a description of the findings in table 15B, using the GEE approach to summarize findings across the six dependent variables, following the guidelines in chapters 2, 14, and 15 and appendix A.

TABLE 15B. Results from OLS models of six neighborhood characteristics and housing problems measures, Yonkers Residential Mobility Program, 1994–1995

	Dependent variable					
Independent variable	Danger	Victimization	Disorder	Cohesion	Resources	Housing problems[a]
Mover	−0.99***	−0.19**	−1.25***	0.21***	0.13	−0.30***
Age (years)	0.01	0.00	0.00	0.00	0.00	0.00
Latino	0.16	0.00	−0.02	−0.01	0.09	−0.19***
High school graduate	0.06	0.07	0.04	0.02	0.05	−0.06
Female headed HH	−0.27*	−0.01	0.02	−0.03	−0.05	0.07
# children in HH	0.05	0.07*	0.05*	−0.01	0.00	0.03
R^2	0.34	0.05	0.73	0.14	0.02	0.20

Source: Adapted from Fauth, Leventhal, and Brooks-Gunn 2004, table 3.

[a] In the published paper, this measure was termed "housing quality," but I relabeled it "housing problems" to reduce confusion because higher value indicates more problems, e.g., with rats and mice.

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4. Write a description of Zimmerman's findings (table 15C), focusing on the results for own SAT scores and roommate's SAT scores. Follow the guidelines in chapter 15 about organizing your description. Generalize across the three models to the extent possible: Which results are similar for the three groups, and which differ? Why did Zimmerman run three models?

TABLE 15C. Estimated coefficients and standard errors from a model of cumulative grade point average by own SAT scores and roommate's SAT scores, stratified by own SAT score, Williams College classes of 1999–2001

	Student's own combined math & verbal SAT score		
	Lowest 15%	Middle 70%	Top 15%
Own verbal SAT score/100	0.205	0.199	0.118
	(0.039)	(0.015)	(0.055)
Own math SAT score/100	0.065	0.112	0.045
	(0.036)	(0.017)	(0.051)
Race (ref. = white)			
Black	−0.181	−0.386	−0.800
	(0.046)	(0.053)	(0.059)
Hispanic	−0.036	−0.254	−0.050
	(0.059)	(0.046)	(0.274)
Native American	−0.238	0.212	dropped
	(0.169)	(0.168)	
Not a US citizen	0.076	0.126	0.055
	(0.091)	(0.055)	(0.066)
Asian	0.210	−0.065	−0.201
	(0.120)	(0.026)	(0.047)
Female	0.262	0.103	0.107
	(0.038)	(0.016)	(0.028)
Roommate's verbal	0.006	0.043	−0.013
SAT score/100	(0.025)	(0.012)	(0.021)
Roommate's math	−0.038	−0.021	0.030
SAT score/100	(0.028)	(0.012)	(0.022)
Sample size	450	2,072	629
R^2	0.41	0.27	0.21

Source: Adapted from David A. Zimmerman, "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *Review of Economics and Statistics* 85, no. 1 (2003): 9–23, table 4.

Answer questions 5 through 7 based on the results in table 15D from Fussell and Massey (2004).

TABLE 15D. Estimated log-odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

	Log-odds	Standard error
Demographic background		
Age (years)	−0.003	0.02
Age-squared	−0.001	0.0002
Ever married	−0.09	0.06
Number of minor children in household	0.01	0.01
Human capital		
Years of education	−0.04	0.006
Months of labor-force experience	−0.002	0.0007
Social capital in the family		
Parent a prior US migrant	0.51	0.05
Siblings prior US migrants	0.36	0.02

	Log-odds	Standard error
Social capital in the community		
Migration prevalence ratio[a]		
0–4	−0.99	0.15
5–9	−0.09	0.12
(10–14)		
15–19	0.35	0.10
20–29	0.57	0.13
30–39	0.95	0.15
40–59	0.74	0.19
60 or more	0.34	0.15
Intercept	−3.31	0.26
−2 log likelihood	23,369.2	
Df	26	

Source: Adapted from Elizabeth Fussell and Douglas S. Massey, "The Limits to Cumulative Causation: International Migration from Mexican Urban Areas," *Demography* 41, no. 1 (2004): 151–71, table 2. http://muse.jhu.edu/journals/demography/v041/41.1fussell.pdf.

Note: Model also includes controls for occupational sector, internal migratory experience, community characteristics, and Mexican economic and US policy context.

[a] The migration prevalence ratio = (the number of people aged 15+ years who had ever been to the US/the number of people aged 15+ years) × 100.

5. Write a description of the age pattern of migration to the United States, with reference to the chart you created in question 9a of the problem set for chapter 6.

6. Write a description of the relationship between human capital and migration.

7. Write one to two paragraphs describing the association between social capital in the family and community and migration from Mexico to the United States, with reference to the results in table 15D and the chart you created in question 9b of the problem set for chapter 6.

Pan et al. (2005) estimated a series of multilevel growth trajectory models of toddler vocabulary. The model specification and goodness of fit statistics are shown in table 15E.

TABLE 15E. Model specification and goodness-of-fit statistics for four multilevel growth trajectory models of toddler vocabulary development among children from low-income families

Variables	Model 1	Model 2	Model 3	Model 4
Intercept	X	X	X	X
Age and age-squared	X	X	X	X
Mother tokens (3 variables)		X		
Mother word types (3 variables)			X	
Mother points (3 variables)				X
Random effects parameters (4 variables)	X	X	X	X
−2 Log likelihood	1,932.9	1,931.6	1,928.4	1,929.1
Akaike Information Criterion (*AIC*)	1,952.9	1,957.6	1,954.4	1,955.1
Degrees of freedom	7	10	10	10

Adapted from Barbara Alexander Pan, Meredith L. Rowe, Judith D. Singer, and Catherine E. Snow, "Maternal Correlates of Growth in Toddler Vocabulary Production in Low-Income Families," *Child Development* 76, no. 4 (2005): 763–82, table 2.

8. Use the information in table 15E to answer the following questions:

   a. Which models are nested? Explain why.
   b. Which models are not nested? Explain why.
   c. Keeping in mind your answers to parts a and b, identify the parsimonious model among fit of models 1 through 4 using the guidelines on "Comparing Models using *AIC* or *BIC*" from p. 335 of *Writing about Multivariate Analysis, 2nd Edition*.

15. *Writing about Multivariate Models*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article about an application of multivariate analysis to a research question in your field. Critique the methods and results sections, using the principles in chapter 15 of *Writing about Multivariate Analysis, 2nd Edition* to check for the following:

    a. An explanation of why a multivariate model is needed for this research question and data;
    b. Topic sentences that introduce the purpose of each table, chart, or quantitative comparison;
    c. Identification of the role of each independent variable (e.g., key predictor, mediator, confounder, control variable);
    d. Descriptions of direction, magnitude, and statistical significance of the association between the key independent variable and the dependent variable;
    e. Explanations of how specific numeric findings address the questions under study;
    f. Transition sentences that explain how one paragraph follows from the previous paragraph.

2. Find a journal article about an application of multivariate analysis to a topic in your field that involves a series of nested models. Critique the description of the nested model results, using the criteria described under “Comparing a Series of Nested Models” and “GEE Revisited” on pp. 332–34 and 337–38, respectively. Rewrite it to correct the flaws you identified.

3. Find a journal article about an application of multivariate analysis to a topic in your field that involves a set of stratified models, such as the same model estimated separately by gender, region, or time period. Critique the description of the stratified model results, using the criteria described under “GEE Revisited” on pp. 337–38. Rewrite the description to correct the flaws you identified.

B. Applying Statistics and Writing

1. Using the same variables as in question B.1 of the suggested course extension for chapter 9, estimate an OLS model with a quadratic specification of X_1: $Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$.

   a. Use the *F-s*tatistic to test whether the quadratic specification of X_1 statistically significantly improves the fit of the model compared to a linear specification of X_1 (without the quadratic term; from the results of your analysis for question B.1 of chapter 9). Contrast your conclusions from that test this that based on the test statistic for β_2.
   b. Use the *BIC* statistic to test whether the quadratic specification of X_1 statistically significantly improves the fit of the model compared to a linear specification of X_1.

2. Using the output from questions B.1, B.2, and B.3 from the suggested course extensions for chapter 9

   a. Create a table to present the results of the three models, following the guidelines in chapters 5 and 15 of *Writing about Multivariate Analysis, 2nd Edition* to report the estimated coefficients, standard errors, *F*-statistics, and *BIC* statistics for each model.
      $Y_1 = \beta_0 + \beta_1 X_1$
      $Y_1 = \beta_0 + \beta_1 DUMMY$
      $Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 DUMMY$
   b. Write a sentence interpreting how the coefficient on *DUMMY* changes with the introduction of controls for X_1, following the guidelines on pp. 332–34 for writing about results of nested models.
   c. Identify which models are nested, and which are non-nested.
   d. Use the *F-s*tatistic to test whether
      i. the addition of X_1 statistically significantly improves the fit of the model compared to the model with *DUMMY* only.
      ii. the addition of *DUMMY* statistically significantly improves the fit of the model compared to the model with X_1 only.
   e. Use the *BIC* statistic to identify the best-fitting model among those three specifications, using the criteria on p. 335 for contrasting non-nested models.

3. Write a description of one or more tables of bivariate results from your own data, using the criteria on pp. 317–25.

4. Write a description of results of one multivariate model for the same research question as in the preceding question, using the criteria listed under question A.1.

5. Write a description of a series of nested models for the same research question as in question B.3, using the criteria described under “Com-

paring a Series of Nested Models" and "GEE Revisited" on pp. 332–34 and 337–38, respectively.

6. Write a description of a set of stratified models for the same research question as in question B.3, using the criteria described under "GEE Revisited" on pp. 337–38.

C. Revising

1. Evaluate a description of a single multivariate model from the results section of a paper you have written previously, using the criteria listed under question A.1. Rewrite that description to rectify any problems you identified.

2. Evaluate a description of a series of nested models from the results section of a paper you have written previously, using the criteria described under "Comparing a Series of Nested Models" and "GEE Revisited" on pp. 332–34 and 337–38 of *Writing about Multivariate Analysis, 2nd Edition*. Rewrite that description to rectify any problems you identified.

3. Evaluate a description of a set of stratified models from the results section of a paper you have written previously, using the criteria described under "GEE Revisited" on pp. 337–38. Rewrite that description to rectify any problems you identified.

15. *Writing about Multivariate Models*

SOLUTIONS

1. Answer questions based on the data in tables 15A and 15B.

   a. No, the random assignment didn't succeed in equalizing the background characteristics of movers and stayers. "Despite random assignment of treatment and control groups in the Yonkers Residential Mobility Program, there were statistically significant differences in four of the six measured background characteristics between participants who moved versus those who stayed in their original neighborhoods (table 15A). Movers were on average slightly older, more likely to have at least a high school education, less likely to be in female-headed households, and had slightly fewer children than stayers (all $p < 0.05$). No differences were observed in terms of race/ethnicity or gender."

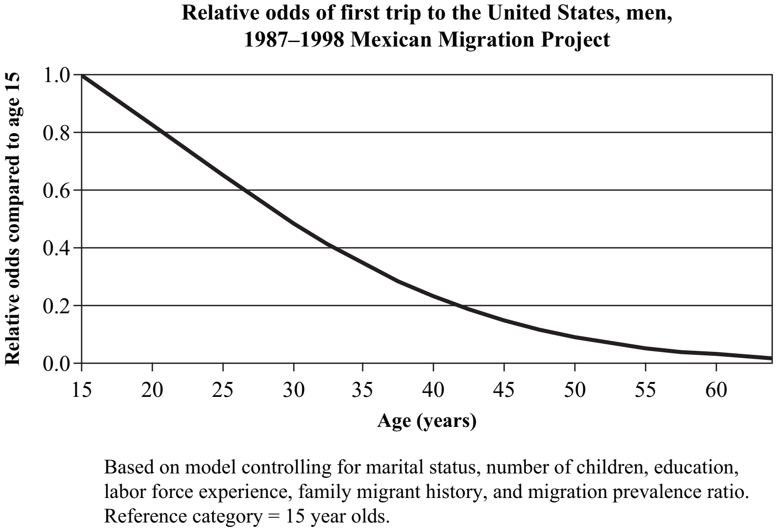
   b. Yes, neighborhood and housing characteristics differed according to residential status. "On all six dimensions studied, outcomes were statistically significantly better among movers than stayers (table 15A). Negative outcomes (danger, victimizations, disorder, and indicators of poor housing) were all lower among movers than stayers, while favorable outcomes (cohesion and resources) were higher among movers than stayers."

   c. These bivariate statistics suggest that a multivariate regression is necessary to assess the impact of residential status on the outcomes studied, net of the potentially confounding effect of the background characteristics. All of the observed differences in background characteristics would be expected to favor better outcomes among movers than stayers regardless of where they live. For example, older age, two-parent households, better education, and smaller families are often associated with better resources than younger, female-headed, less-educated, and larger families. Hence a multivariate model is needed to control for those characteristics in order to measure the net effect of moving versus staying.

3. "Table 15B presents results of multivariate models of six measures of neighborhood characteristics and housing problems from the Yonkers Residential Mobility Program. On five of the six outcomes studied, subjects who moved showed statistically significant better

outcomes than those who remained in their original neighborhoods, even when the effects of potential confounders were taken into account. The negative outcomes (danger, victimization, disorder, and housing problems) were each lower among movers than stayers, while the favorable outcomes (cohesion and resources) were higher among movers, though the difference in resources was not statistically significant. Although some of the background control variables were statistically significantly associated with one or two of the outcomes, none showed a consistent pattern of association."

5. "The odds of first migration to the United States declined rapidly between ages 15 and 40, then continued to decline with age, but at a slower rate (figure 15A). For example, the relative odds of migration were roughly 0.60 among 25-year-olds, 0.30 among 35-year-olds, and 0.15 among 45-year-olds when each was compared to 15-year-olds."


Relative odds of first trip to the United States, men, 1987–1998 Mexican Migration Project

Based on model controlling for marital status, number of children, education, labor force experience, family migrant history, and migration prevalence ratio. Reference category = 15 year olds.

Figure 15A.

7. "Social capital in the family and in the community is an important predictor of odds of migration from Mexico to the United States even when individual demographic background, human capital, and community economic and policy context are taken into account. In terms of family social capital, both having a parent and having a sibling who was a prior US migrant increased the chances of migrating (OR = 1.67 and 1.43, respectively, compared to having no family members as prior US migrants; both $p < 0.001$). In terms of community social capital, odds of migration increased with increasing migration prevalence ratio (MPR) up to an MPR of 40%, then declined slightly among communities with very high MPRs (figure 15B). For example, the odds of migration were nearly seven times as high among men from communities where 30% to 39% of people aged 15 and older had ever been to the United States as among those from communities where fewer than 5% had been there."

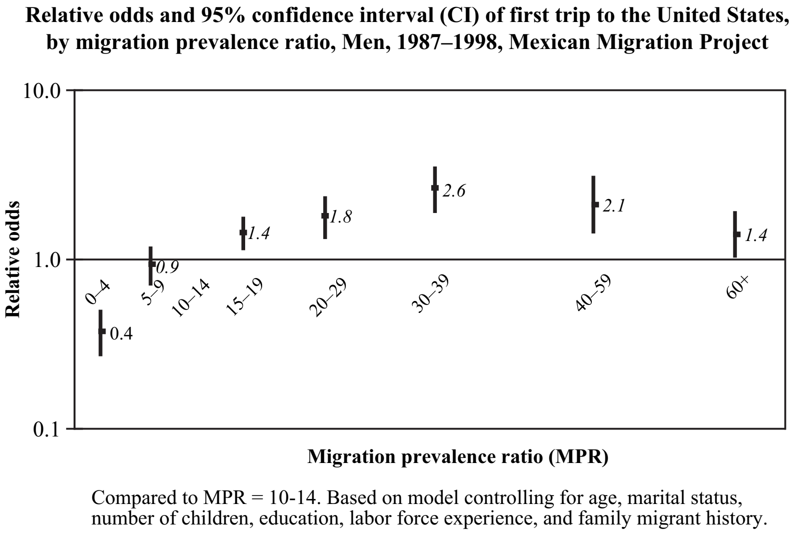

Relative odds and 95% confidence interval (CI) of first trip to the United States, by migration prevalence ratio, Men, 1987–1998, Mexican Migration Project
10.0
1.0
0.1
Relative odds
0–4
5–9
10–14
15–19
20–29
30–39
40–59
60+
0.4
0.9
1.4
1.8
2.6
2.1
1.4
Migration prevalence ratio (MPR)
Compared to MPR = 10-14. Based on model controlling for age, marital status, number of children, education, labor force experience, and family migrant history.

Figure 15B.

16. *Writing about Interactions*

PROBLEM SET

1. For each of the listed figures from *Writing about Multivariate Analysis, 2nd Edition*, (i) name the independent and dependent variables involved in the interaction, and state (ii) whether the interaction is in terms of direction or magnitude of association (or both), and (iii) whether the interaction is ordinal or disordinal.

    a. Figure 17.4 (p. 377) adapted from Pottick et al. (1999)
    b. Figure 16.1 (p. 342)
    c. Figure 16.2 (p. 343) adapted from Miller and Rodgers (2008)
    d. Figure 18.1 (p. 401) from Krivo et al. (2009)
    e. Figure 18.2 (p. 405) adapted from Phillips et al. (2004)

2. For figures 16.4d, e, and f on p. 345 in *Writing about Multivariate Analysis, 2nd Edition*, think of a topic for which that shape association makes sense. E.g., for figure 16.4f, a relationship with an upward sloping curve between an independent variable (IV) and dependent variable (DV) for one group, coupled with a downward sloping curve between the same IV and DV for another group. List the independent and dependent variables involved, and the context in which such a relationship might exist.

Table 16A summarizes results of Carr's (2004) analysis of relations among dependence on a spouse, gender, and psychological adjustment to the death of a spouse.

TABLE 16A. OLS regressions of self-esteem at wave 2, overall and by gender, changing lives of older couples (CLOC) study, 1987–1994

	Total sample		Women		Men	
Variable	Coeff.	Std. error	Coeff.	Std. error	Coeff.	Std. error
Widow	−0.51*	0.24	0.25†	0.15	1.67	1.22
Female	−0.60**	0.22				
Interaction: female_widow	0.70**	0.26				
Emotional dependence on spouse			−0.35**	0.13		
Interaction: emotional dependence on spouse_widow			0.34**	0.15		

(*continued*)

TABLE 16A. (*continued*)

	Total sample		Women		Men	
Variable	Coeff.	Std. error	Coeff.	Std. error	Coeff.	Std. error
Dependence on spouse for homemaking tasks					2.67*	1.35
Interaction: dependence on spouse for homemaking tasks_widow					−2.92*	1.39
Dependence on spouse for home maintenance and financial tasks					−1.30*	0.55
Interaction: dependence on spouse for home maintenance and financial task_ widow					1.58**	0.59
Intercept	2.13	0.76*	0.54	0.79	1.75	2.12
R^2 adjusted	0.19		.024		0.19	
Unweighted *N*	297		217		80	

Source: Adapted from Deborah Carr, "Gender, Preloss Marital Dependence, and Older Adults' Adjustment to Widowhood," *Journal of Marriage and the Family* 66 (2004): 220–35, table 2.

Models also control for wave 1 well-being, demographic characteristics, and number of months between wave 1 and 2 interviews. Dependence measures assessed at wave 1.

* $p < 0.05$; ** $p < 0.01$; † $p < 0.10$

3. Using the results for the total sample in table 16A

   a. Create a table to show predicted self-esteem for each of the four possible combinations of gender and widowhood status.
   b. Create a chart to portray that association.
   c. Write a short description of the association between gender, widowhood status, and predicted self-esteem using the GEE approach.

4. Using the results for women in table 16A

   a. Create a spreadsheet to calculate the net effect of the interaction between emotional dependence on spouse, widowhood status, and predicted self-esteem, working from the online spreadsheet template for continuous by categorical interactions, or using the guidelines in appendix D of *Writing about Multivariate Analysis, 2nd Edition*. Both self-esteem and emotional dependence are in standardized units (mean = 0, standard deviation [SD] = 1). Allow emotional dependence to vary from −1.0 to 1.0 SD in your calculations.
   b. Design a chart to portray this pattern following the guidelines in chapters 6 and 16.
   c. Write a short description of the association between emotional dependence on spouse, widowhood status, and predicted self-esteem using the GEE approach.

d. Explain why there isn't a dummy variable for "female" in the stratified models.

Miller and Rodgers (2008) estimated a model of monthly earnings with an interaction between gender and marital status in Taiwan. The estimated coefficients for variables involved in the interaction are shown in table 16B, and the associated variance-covariance matrix in table 16C.

TABLE 16B. Estimated coefficients from a model of monthly earnings NT$, Taiwan, 1992

Variable	Coefficient	Standard error
Intercept	−21,022.2	1,897.62
Man	3,204.9	201.34
Married	−1,594.7	213.30
Man_married	4,771.2	248.65

Model also controls for work experience, tenure on the job, educational attainment, urban residence, supervisory occupation, and gender composition of the respondent's occupation.

TABLE 16C. Variance-covariance matrix for the estimated coefficients in table 16B

	Man	Married	Man_married
Man	40,538.61		
Married	20,834.16	45,497.59	
Man_married	−34,094.11	−36,700.40	61,826.53

5. Perform the following tasks using the information in tables 16B and 16C and the techniques explained in chapter 16 and the associated online materials for *Writing about Multivariate Analysis, 2nd Edition* and the associated references.

   a. Calculate the difference in monthly earnings for married men compared to unmarried women (the reference category).
   b. Calculate the standard error of the compound coefficient for married men from the information in the variance-covariance matrix.
   c. Calculate the 95% confidence interval around the point estimate of the difference in earnings for each marital status/gender combination compared to the reference category (unmarried women).
   d. Conduct and write up results of statistical tests for differences in earnings between the following pairs of groups, explaining direction, magnitude, and statistical significance:
      i. Married versus unmarried women
      ii. Married versus unmarried men
   e. Optional: Create a spreadsheet to conduct steps a through c, working from the online spreadsheet template for categorical by categorical interactions, or the guidelines in appendix D of *Writing about Multivariate Analysis, 2nd Edition*.

16. *Writing about Interactions*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find an article in your field that posits an interaction between two or more independent variables. Evaluate whether they have explained the reasons for that hypothesis

   a. Based on theory
   b. Based on empirical analysis of their own data

2. Find a journal article that presents results of an OLS model with an interaction between a categorical independent variable and a continuous independent variable. Use the criteria in chapter 16 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate the following aspects of the article:

   a. The description of the variables and model specification in the data and methods section.
   b. The table of regression coefficients.
      i. Did they provide enough information to assess the statistical significance of individual main effects and interactions terms?
      ii. Did they provide enough information to assess the contribution of the interactions to overall model fit?
   c. Whether they used a chart to portray the overall shape of the interaction, and if so, whether it satisfied the criteria in chapters 6 and 16 for effective charts.
   d. Whether their prose description satisfied criteria for effective presentation of an interaction pattern.
   e. Rewrite the description of the interaction to correct any shortcomings you identified in parts a through d.

3. Repeat question A.2 for a journal article that presents results of an OLS model with an interaction between two categorical independent variables.

4. Repeat question A.2 for a journal article that presents results of an OLS model with an interaction between two continuous independent variables.

B. Applying Statistics and Writing

1. Using the same variables that you used for Y, X_1, and *DUMMY* in question B.3 in the suggested course extensions for chapter 9, estimate an OLS model with an interaction between X_1 and *DUMMY*.

    a. Write an equation to convey the model specification, including both main effects and interaction terms.
    b. Calculate predicted values of Y for cases in the reference category and those in the other category of *DUMMY* across the observed range of X_1 in your data.
    c. Create a chart showing the shape of the estimated relationship among Y, X_1, and *DUMMY*, using the results from part b, and the guidelines in chapters 6 and 16 of *Writing about Multivariate Analysis, 2nd Edition*.
    d. Calculate differences in predicted values of Y for one-unit increases in X_1 for cases in each category of *DUMMY*.
    e. Optional: Use a spreadsheet to perform parts b through d, working from the online spreadsheet template for continuous by categorical interaction, or by following the instructions in appendix D of *Writing about Multivariate Analysis, 2nd Edition*.

2. Using the same variables as in question B.3 of the suggested course extensions for chapter 9, estimate an OLS model with an interaction between *DUMMY* and a three-category independent variable (*CATEGVAR*). Request the variance-covariance matrix as part of the output.

    a. Write an equation to convey the model specification, including both main effects and interaction terms. Use this equation to help you define appropriate dummy variables to specify the interaction.
    b. Calculate the predicted values of Y for all possible combinations of the variables *DUMMY* and *CATEGVAR*.
    c. Create a chart showing the shape of the estimated relationship between Y, *DUMMY*, and *CATEGVAR*, using the results from part b and guidelines in chapters 6 and 16.
    d. Use the simple slopes technique to test the statistical significance of differences between cases that are *not* in the reference category of *either DUMMY* or *CATEGVAR*, compared to cases in the reference category for *both DUMMY* and *CATEGVAR*.
    e. Optional: Use a spreadsheet to perform parts b through d, working from the online spreadsheet template for a categorical by categorical interaction in OLS, or by following the instructions in appendix D.

3. Write the portion of the data and methods section that pertains to your interaction.

    a. Describe how you defined variables to test for an interaction between two independent variables.

b. Describe the sequence of model specifications you used to test for interactions.
   i. Using equations
   ii. In prose

4. Based on the results to question B.1 or B.2 above,

   a. create a table to present coefficients and goodness-of-fit statistics from models of main effects only, and main effects plus interactions.
   b. referring to the chart you made in part c of that question and the guidelines in chapters 2 and 16, use the GEE approach to describe the overall shape of the interaction, specifically mentioning exceptions in direction or magnitude of the association.

5. Estimate a logit model of a dichotomous dependent variable Y_2, with an interaction between *DUMMY* and a three-category independent variable (*CATEGVAR*).

   a. Calculate the odds ratio of the outcome you are modeling for all possible combinations of the variables *DUMMY* and *CATEGVAR*.
      i. Working from the logit coefficients (log relative odds) on the pertinent main effect and interaction terms;
      ii. Working from the odds ratios calculated from the pertinent main effect and interaction coefficients;
   b. Create a chart showing the shape of the estimated relationship between Y_2, *DUMMY*, and *CATEGVAR*, using the results from part b and guidelines in chapters 6 and 16.
   c. Optional: Use a spreadsheet to perform parts b and c, working from the online spreadsheet template for a categorical by categorical interaction logit interaction.

C. Revising

1. Review a data and methods section you have written previously about an interaction among variables, using the checklist for chapter 16 in *Writing about Multivariate Analysis, 2nd Edition* to evaluate how you have described the variables and the model specification. Revise the section to correct any shortcomings you find.

2. Review a results section you have written previously about an interaction among variables, using the checklist for chapter 16 to evaluate the following elements:

   a. Bivariate and multivariate tables;
   b. Charts to portray the overall shape of an interaction;
   c. Prose, including direction, magnitude, and statistical significance of the interaction.
   d. Revise those elements to correct any shortcomings you find.

3. Exchange your revised data and methods and results sections from questions C.1 and C.2 with a peer or colleague.

   a. Review them, using the checklist for chapter 16.
   b. Revise your prose, tables, and/or chart to correct the errors he or she found.

16. *Writing about Interactions*

SOLUTIONS

1. For each of these figures from *Writing about Multivariate Analysis, 2nd Edition*, (i) name the independent and dependent variables involved in the interaction, and state (ii) whether the interaction is in terms of direction or magnitude of association, and (iii) whether it is ordinal or disordinal.

   a. Figure 17.4 from Pottick et al.
      i. The independent variables involved in the interaction are time since admission (*x* axis) and type of health insurance (legend), and the dependent variable is discharge from the hospital (*y* axis).
      ii. The interaction between type of insurance and time since admission is in terms of direction of association. The slopes of the two hazard curves are in opposite directions, but of approximately equal steepness.
      iii. The interaction is disordinal because the hazard curves for the two types of insurance cross one another in the observed range of values of the independent variables.

   b. Figure 16.1
      i. The independent variables involved in the interaction are educational attainment (*x* axis) and race/ethnicity (legend), and the dependent variable is birth weight (*y* axis).
      ii. The interaction between race/ethnicity and educational attainment is in terms of magnitude, because birth weight increases with rising educational attainment for all three racial/ethnic groups (same direction of association) but with a decreasing racial gap (magnitude).
      iii. The interaction is ordinal because the rank order of birth weight by educational attainment is the same for all three racial/ethnic groups.

   c. Figure 16.2 from Miller and Rodgers (2008)
      i. The independent variables involved in the interaction are marital status (*x* axis) and gender (legend), and the dependent variable is monthly earnings (*y* axis).
      ii. The interaction is in terms of both direction and magnitude. Not only does the earnings difference by marital status work in opposite directions for men than for women, the size of

the earnings gap is larger for men than for women: NT$3,176 *more* per month for married compared to unmarried men, but NT$1,595 *less* per month for married compared to unmarried women.

iii. The interaction is disordinal because the rank order of earnings by marital status for women is the reverse of that for men. For women, married earnings < unmarried earnings; for men, married earnings > unmarried earnings.

d. Figure 18.1 from Krivo et al.

i. The independent variables involved in the interaction are neighborhood-level racial/ethnic composition (panels) and city-level segregation (legend), and the dependent variable is neighborhood crime rate (*y* axis). Neighborhood disadvantage is also plotted (on the *x* axis) to show how different the levels and ranges of that variable are for neighborhoods with different racial/ethnic compositions.

ii. The cross-level interaction between neighborhood racial/ethnic composition and segregation shows up primarily as a difference in the intercept—the level of the crime rate. The curves relating neighborhood disadvantage, city-level segregation, and neighborhood crime rate are upward sloping for all of the racial/ethnic groups.

iii. The interaction is ordinal because the curves relating disadvantage, segregation, and crime remain approximately parallel within each of the neighborhood racial/ethnic composition groups.

e. Figure 18.2 from Phillips et al. (2004)

i. The independent variables involved in the interaction are NJ KidCare Plan level (*x* axis), family race/ethnicity (legend), and county physician racial composition (legend), and the dependent variable is disenrollment in NJ KidCare (*y* axis).

ii. The interaction is in terms of magnitude, which appears as a wider gap in disenrollment rates for families in NJ KidCare Plan D than in Plans B and C.

iii. The interaction is ordinal because the rank order of disenrollment by family race/ethnicity and county physician racial composition is the same in both Plans B/C and Plan D.

3. Using the results for the total sample

a. **TABLE 16D.** Predicted self-esteem by gender and widowhood status, CLOC sample, 1987–1994

	Male	Female
Widow	1.62	1.72
Nonwidow	2.13	1.53

a. Explanation: Each of the cells includes the intercept. The "female/nonwidow" cell adds in the coefficient on the "female" dummy; the "male/widow" cell adds in the coefficient on the "widow" dummy; the "female/widow" cell adds in both of those coefficients along with the "female _ widow" interaction term. (Note: Results differ from those shown in Carr [2004] because they do not include values of other variables in her model that are excluded from table 16A.)

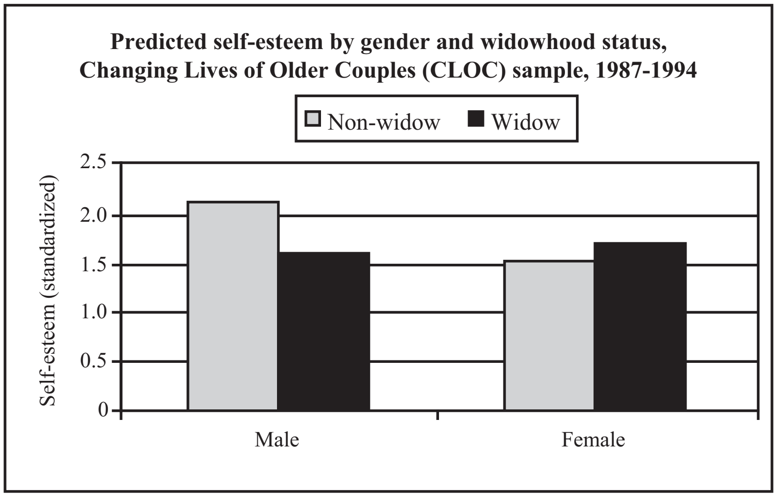


Figure 16A.

b. Figure 16A presents predicted self-esteem for each of the four possible combinations of gender and widowhood status (Carr 2004).

c. "As shown in table 16D the association between widowhood and self-esteem differs by gender. Among males, self-esteem averages nearly half a standard deviation unit lower among widows than among those whose spouses are still alive at wave 2 (1.62 versus 2.13 points, respectively). Among females, however, widows have higher self-esteem than nonwidows (1.72 and 1.53, respectively)."

5. Perform the following tasks using the information in tables 16B and 16C and the techniques explained in chapter 16 of *Writing about Multivariate Analysis, 2nd Edition* and the associated references.

a. The difference in earnings for married men compared to unmarried women (the reference category) = $\beta_{\text{Married}} + \beta_{\text{Man}} + \beta_{\text{Man\_married}}$ = –1,595 + 3,205 + 4,771 = 6,381.

b. The formula for the standard error of the compound coefficient for married males = square root [(variance (β_{Married}) + (2 × covariance (β_{Married}, $\beta_{\text{Man\_married}}$) + variance ($\beta_{\text{Man\_married}}$)]. Substituting the values from table 16D gives square root [45,497.59 + (2 × (–36,700.40)) + 61,826.53] = 184.18

c. Calculate the 95% confidence interval around the point estimate of the difference in earnings for each marital status/gender combination compared to the reference category (unmarried women).

95% confidence interval for married women = β_{Married} ± (1.96 × std error (β_{Married})) = –1,595 ± (1.96 × 213) = –1,595 ± 418

$= -2{,}103$ to $-1{,}177$. Coefficient and standard error are from table 16B; or you can use the square root of the variance from table 16C to calculate the standard error.

95% confidence interval for unmarried men $= \beta_{\text{Man}} \pm (1.96 \times$ std error $(\beta_{\text{Man}})) = 3{,}205 \pm (1.96 \times 201) = 3{,}205 \pm 395 = 2{,}810$ to 3,600. Coefficient and standard error are from table 16B.

95% confidence interval for married men $= (\beta_{\text{Man}} + \beta_{\text{Married}} + \beta_{\text{Man\_married}}) \pm (1.96 \times$ std error $(\beta_{\text{Man \& Married}})) = 6{,}381 \pm (1.96 \times 184.2) = 6{,}381 \pm 395 = 6{,}021$ to 6,743. Standard error for the compound coefficient was calculated in part b.

d. Conduct and write up results of statistical tests for differences in predicted earnings between the following pairs of groups:

i. "Married women are predicted to earn NT$1,595 less than their unmarried counterparts (95% CI: NT$–2,103 to NT$–1,177)."

Notes: The statistical test for a difference in earnings for married versus unmarried women is based solely on the coefficient and standard error for the dummy variable "Married" since the reference category is unmarried women.

ii. Married versus unmarried men. "Married men are predicted to earn NT$3,177 more than their unmarried counterparts ($p < 0.05$)."

Notes: Subtract the differences for married men and unmarried men (when each is compared to unmarried women) to obtain NT$3,177. Because the 95% confidence intervals around the differences in earnings for married men (6,021 to 6,743) and unmarried men (2,810 to 3,600) do not overlap, their values are statistically significantly different from one another.

17. *Writing about Event History Analysis*

PROBLEM SET

1. Tammemagi et al. (2005) conducted an analysis of racial disparities in breast cancer survival. Based on figure 17.1 from their study (see p. 371 of *Writing about Multivariate Analysis, 2nd Edition*), write sentences to describe

   a. Sample sizes at baseline for black and white women
   b. Direction and magnitude of racial differences in
      i. median survival time
      ii. the proportion surviving until at least 5 years after baseline

2. Based on the information in table 17.1 from Valiyeva et al. (2006) (p. 372 of *Writing about Multivariate Analysis, 2nd Edition*), write a short paragraph for the methods section, reporting the sample size, number of spells, and number of admissions, following the guidelines in chapter 17.

3. Smith et al. (2005) conducted an event history analysis of relationships between managed care and rehospitalization among stroke patients. Write the following materials based on the information from their study shown in figure 17.2 (p. 374):

   a. For the methods section, describe
      i. The types of events they modeled subsequent to discharge from the index admission.
      ii. The total number of persons "at risk" in their competing risks model of outcomes in the 30 days after discharge from index admission.
      iii. The events they modeled subsequent to discharge after first rehospitalization.
      iv. The number of persons "at risk" in their competing risks model subsequent to rehospitalization.
      v. The total number of deaths observed during the study period.
   b. For the results section, report and interpret the direction, magnitude, and statistical significance of the following associations for HMO compared to fee-for-service clients:
      i. rehospitalization following index admission
      ii. death following index admission
      iii. a second rehospitalization

4. Based on table 17.2 from Valiyeva et al. (2006) (p. 380 of *Writing about Multivariate Analysis, 2nd Edition*), use the GEE approach to summarize the direction, magnitude, and statistical significance of the associations between each of the following risk factors and nursing home admission across the two age groups studied:

   a. Systolic blood pressure of 140+ mmHg
   b. Cholesterol of 240+ mg/dL
   c. Diabetes

5. DesJardins et al. (2002) analyzed how financial aid affects chances of a first "stopout" (temporary or permanent leave from college). Write the following materials based on information from their study shown in figure 17.5 on p. 379 of *Writing about Multivariate Analysis, 2nd Edition* and table 17A:

   a. Write a paragraph for the data and methods section defining how the independent variable financial aid is measured and specified in the analysis of college stopout.
   b. Write an equation to convey the specification between the financial aid measures and college stopout, using subscripts to convey which variables and parameters are time-varying.
   c. Use the GEE technique to write a paragraph for the results section describing the time-dependent pattern of amount of financial aid by type shown in figure 17.5.

TABLE 17A. Relative risk of first stopout from college, by number of years of enrollment and type of financial aid,[a] Minnesota, 1986–1994

# years of enrollment	Loans	Earnings	Scholarship	Grants	Work/study
1	0.78	1.03	0.28	1.03	0.50
2	0.93	0.83	0.38	1.03	0.75
3	0.99	0.73	0.45	1.04	0.92
4	0.97	0.68	0.49	1.06	0.96
5	0.90	0.66	0.51	1.09	0.92
6	0.82	0.67	0.52	1.11	0.84
7	0.75	0.69	0.52	1.12	0.77

Excerpted from DesJardins et al. 2002, table 4.

[a] Per $1,000 in aid. Compared to no aid. Aid measures are time-varying. Model also controls for race/ethnicity, gender, age, disability, type of college, in- versus out-of-state residence, ACT score, high school class rank, college grade point average, transfer credits, and type of college.

6. Write sentences interpreting the effects of each of the following amounts, types, and timing of financial aid on chances of dropping out of college, based on the results in table 17A.

   a. A $1,000 increase in the amount of scholarship aid in the first year of enrollment;
   b. A $1,000 increase in the amount of scholarship aid in the fourth year of enrollment;

c. A $1,000 increase in the amount of grant aid in the first year of enrollment;
d. A $1,000 increase in the amount of grant aid in the fourth year of enrollment;
e. A $500 increase in the amount of earnings in the first year of enrollment;
f. A $2,000 increase in the amount of earnings in the fourth year of enrollment.

7. Perform the following tasks based on table 17A from DesJardins et al.:

   a. Create a chart to show how the relative risks of first stopout vary by time and type of financial aid.
   b. Use the GEE technique to write a paragraph describing the time-dependent effects of on college stopout of financial aid by type, following the guidelines in chapters 9, 14, and 17 of *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*. Mention which types of financial aid have the largest effect on risk of dropout and whether those patterns are consistent across time.

17. *Writing about Event History Analysis*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find an article in your field about an application of Cox proportional hazards models. Use the guidelines in chapter 17 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate whether they justified use of an event history analysis based on the following criteria:

   a. Theory for the topic;
   b. Previous literature on the topic;
   c. Data structure.

2. For the same article as in question A.1, evaluate the following aspects of their data and methods section:

   a. The units in which time is measured;
   b. The definition of the event(s) under study;
   c. Whether the event is repeatable;
      i. If so, whether they included all spells for each case, and what statistical corrections they made for multiple spells per case.
      ii. If they did not include all spells for a repeatable event, what criteria they used to select cases for their analysis.
   d. What comprises censoring in their data;
   e. The source(s) of data from which the event history was constructed;
   f. Whether they used dates of events, status at different time points, or respondent reports of time since event to calculate duration of each period at risk and the indicator of event or censoring;
   g. The maximum possible length of the follow-up and dates or intervals of follow-up;
   h. Whether any of the independent variables were specified as time-varying; if so
      i. the timing of those measures;
      ii. the sources of information for the values of that variable at different time points.
   i. Diagnostics for proportionality of hazards.

3. For the same article you used for question A.1, evaluate the following aspects of their results section, using the guidelines in chapter 17.

    a. Whether they include a graph or table of the unadjusted (univariate or stratified) temporal pattern of event occurrence;
    b. Whether they interpret the direction, magnitude, and statistical significance of hazards ratios for key independent variables;
    c. If they specify time-dependent effects, whether they convey how the hazards ratios change over time;
        i. in prose
        ii. in a chart
    d. If they included time-dependent covariates, whether they described how values of that variable changed over time.
    e. Rewrite the materials to rectify any shortcomings you identify in parts b, c, and d.

B. Applying Statistics and Writing

For the following questions, identify a data set that includes information needed to create an event history data set. Conduct the following steps, using the guidelines in chapter 17 of *Writing about Multivariate Analysis, 2nd Edition*.

1. Identify a single-decrement nonrepeatable event for which an event history can be created from your data set. Write the portion of the data and methods section that explains how you created the event history for that event from the original data source. Cover the following elements:

    a. The event under study (e.g., what type of transition is to be analyzed);
    b. Whether you used dates of events, status at different time points, or respondent reports of time since event to calculate duration of each period at risk and the indicator of event or censoring;
    c. What constitutes right censoring in your data and for your topic (type of event);
    d. Whether the data are affected by left censoring;
    e. The units of time used to measure duration;
    f. Whether you conduct a discrete time or continuous time event history analysis.

2. For the event you selected for the preceding question and a two- or three- category independent variable (e.g., gender or employment status) related to your research question to conduct these steps, using the guidelines in chapter 17

    a. Create a table to report the following descriptive statistical information, for the overall sample and for each subgroup defined based on that categorical independent variable.

   i. total number of spells observed in the sample or subgroup;
   ii. number of events observed in the sample or subgroup;
   iii. total person-time at risk in the sample or subgroup;
   iv. median time to event (if observed in your data);
   v. proportion of cases that were censored at the end of the observation period;
   vi. proportion of cases that experience the event by a specified time since baseline that is suited to your topic and data.

b. Create a chart to display the temporal pattern of event occurrence (hazard curves),
   i. overall;
   ii. stratified by the categorical independent variable.

c. Write a description of results of a bivariate statistical test of whether the pattern of event occurrence differs across categories of your key independent variable. Discuss whether the hazards curves in the chart are proportional (parallel), and if not, whether they converge, diverge, or are disordinal (see chapter 16 of *Writing about Multivariate Analysis, 2nd Edition)*, and the implications of that pattern for your multivariate hazards specification.

3. Estimate a Cox proportional hazards model of the event analyzed in the preceding question, including the categorical independent variable from question B.2 and one continuous independent variable.

   a. Create a table to report the hazards ratios, inferential statistical information, and model goodness-of-fit statistics, following the guidelines in chapters 11 and 17 of *Writing about Multivariate Analysis, 2nd Edition*.
   b. Write a sentence to interpret the direction, magnitude, and statistical significance of the hazards ratio on a categorical independent variable in your model.
   c. Write a sentence to interpret the direction, magnitude, and statistical significance of the hazards ratio on a continuous independent variable in your model.

4. Using the same data and variables as in question B.3, estimate a nonproportional hazards model by interacting time since baseline with the independent variable used in question B.2.

   a. Create a table to report the hazards ratios, inferential statistical information, and model goodness-of-fit statistics.
   b. Create a chart to convey the shape of the nonproportional hazards association between the independent variable and time.
   c. Write a sentence to interpret the direction, magnitude, and statistical significance of the time-dependent effect.
   d. Conduct and interpret results of a comparison in model goodness of fit for the proportional and nonproportional hazards specifications in questions B.3 and B.4, respectively, using the guidelines on pp. 334–35.

5. For a categorical time-varying covariate (independent variable) in your data

    a. Create a table of descriptive statistics to show how the distribution of that variable changes over time since baseline.
    b. Write a description of that pattern, following the guidelines in chapter 17.

6. For a continuous time-varying covariate (independent variable) in your data

    a. Create a chart to portray how the mean value of that variable changes over time since baseline.
    b. Write a description of that pattern.

7. Estimate a hazard model with the time-varying covariate from *either* question B.5 or B.6.

    a. Write a sentence to interpret the hazard ratio on the time-varying covariate.
    b. Conduct and interpret results of a comparison in model goodness of fit of the models with time-invariant and time-varying covariates from the specifications in questions B.3 and B.7, respectively and the guidelines on pp. 334–35.

C. Revising

1. For a paper you have written previously on an application of a Cox proportional hazards model, repeat question A.1 (on the introduction).

2. For that same paper, repeat question A.2 (on the data and methods section).

3. For that same paper, repeat question A.3 (on the results section).

4. Design a survival or hazards chart to convey the unadjusted pattern of event occurrence for a paper you have written previously about an application of event history analysis.

5. Have a peer evaluate a table of descriptive statistics you previously created for an event history analysis, using the guidelines in chapter 17. Revise it to rectify any shortcomings they identify.

6. Have a peer evaluate a table of multivariate hazards results you previously created. Revise it to rectify any shortcomings they identify.

7. Exchange revised drafts of the materials in questions C.1 through C.4 with someone writing about an application of event history analysis to a different topic or data set. Peer-edit each other's work and revise according to the feedback you receive.

17. *Writing about Event History Analysis*

SOLUTIONS

1. Based on figure 17.1 from the study by Tammemagi et al (2005),

   a. At baseline (time of diagnosis), the sample included 629 white women and 257 black women.

   b. Median survival time following breast cancer diagnosis was about one-third longer for white than for black women: 13.5 years and 10 years, respectively (figure 17.1).

   c. White women were approximately 13 percentage points more likely to survive at least 5 years after diagnosis as were their black counterparts ($p < 0.01$; figure 17.1).

3. Based on the information in figure 17.2 from the study by Smith and colleagues (2005)

   a. For the methods section

      i. Smith and colleagues (2005) estimated competing risks models of rehospitalization or death within 30 days of discharge, with "no event" as censoring.

      ii. They included 9,003 persons in their model of outcomes in the 30 days after index admission (see *N* in the "Index Admission" box). Persons discharged from that admission could have been readmitted more than once, or readmitted and then died, which is why the sum of the three numbers in the 30-day outcomes boxes (9,167) can exceed the 9,003. In other words, some people contributed more than one event (and hence more than one spell) to the analysis of the competing events within 30 days of discharge.

      iii. The competing risks model for the subsequent 11 months also analyzed relative hazards of rehospitalization or death.

      iv. The analysis of the subsequent 11 months included information on 1,262 persons competing risks model of events in the 11 months subsequent to rehospitalization. Again, the sum of the numbers in the three 11-month outcomes boxes (1,295) is greater than the number of persons at risk because some people contributed more than one event (and hence more than one spell) to that analysis.

      v. A total of 2,866 deaths were observed during their study period. (= 1,324 after discharge from index admission + 1,176 among those who were not rehospitalized in the 30 days after index discharge + 366 among those who were rehospitalized in the 30 days after index discharge.)

b. For the results section, report and interpret the direction, magnitude, and statistical significance of the following associations for HMO compared to fee-for-service clients:
   i. Adults who were covered by an HMO were 1.29 times as likely as those covered by fee-for-service to be rehospitalized following the discharge from their index admission ($p < 0.05$).
   ii. There was no statistically significant difference in the risk of death following the index admission for HMO versus fee-for-service clients (HR = 1.07; 95% CI 0.95–1.21).
   iii. Type of health insurance coverage was not associated with risk of a rehospitalization in the 11 months after a first rehospitalization (HR = 0.96; 95% CI 0.79–1.16 for HMO compared to fee-for-service).

5. Describe the temporal pattern of financial aid shown in figure 17.5.

   a. Amount and type of financial aid is specified as a series of time-varying covariates, with observations in each term during which a student was enrolled in college. It is classified into six types: scholarships, loans, grants, work/study, other on-campus earnings, and no aid (the reference category). A student could have more than one type of aid in each term. In the model of college stopout, measures of each type of aid (in $1,000s) were included for each term that a student was enrolled up until their first stopout or graduation.
   b. Hazard of college stopout$_t$ = $\beta_0 + \beta_{1t}$*Loan amount*$_t$ + β_{2t}*Scholarship amount*$_t$ + β_{3t}*Grant amount*$_t$ + β_{4t} *Work/study amount*$_t$ + β_{5t} *Earnings amount*$_t$, where amount of each type of financial aid is measured in $1,000s (see footnote to table 17A). Note that each of the financial aid covariates and their associated coefficients have subscript *t*, indicating that financial aid is specified as a series of time-varying covariates, each of which is allowed to have a time-varying effect on the dependent variable (stopout).
   c. Figure 17.5 in *Writing about Multivariate Analysis, 2nd Edition* portrays the average amount of each three types of financial aid offered in Minnesota colleges and universities, by duration of enrollment. In the first three terms of enrollment, average work/study offers were 2.5 times as high as loan amounts and more than three times as high as on-campus earnings ($2,000 per term for work/study, $800 per term for loans, and $600 for earnings). After the fourth term, however, work/study offers were cut in half (to less than $1,000 per term, on average), while on-campus earnings increased to about the same amount. Loan amounts remained roughly constant until the sixth term, and then rose slightly. By the eighth term, average offers from on-campus employment provided

the highest average offer ($1,300), followed by work/study ($900) and loans ($850).

7a. Create a chart from the data in table 17A.

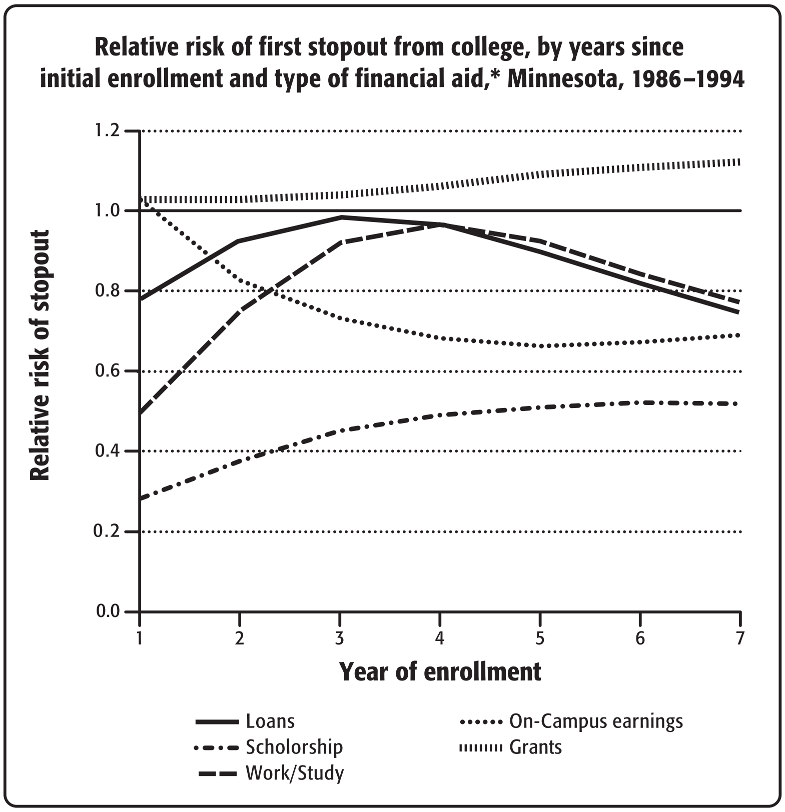


Figure 17A. See notes to table 17A.

7b. In the first year of enrollment, financial aid in the form of scholarships was associated with the lowest risk of stopout among Minnesota college students, followed by work/study funding and loans (RR = 0.28, 0.50, and 0.78 per $1,000 of the specified type of aid, respectively, when each was compared against no financial aid; figure 17A). Risks of stopout were similar for other on-campus earnings, grants (RR = 1.03 per $1,000 for either type of aid), and no aid.

However, the relative risks of stopout for scholarships, work/study, and loans each rose over time, bringing them closer to the risk among students with no financial aid. By year 4, the relative risk for scholarships rose to about 0.5 per $1,000 of aid. In contrast, relative risk of stopout associated with other on-campus earnings decreased with time since enrollment; RR by year 4 = 0.70 compared to no aid. As a consequence of these different temporal patterns of stopout for the various types of financial aid, by the fourth year of enrollment, scholarships were associated with the lowest risk of stopout (RR = 0.5 per $1,000 compared to no aid), followed by own earnings (RR = 0.68), and work/study, loans, and grants, for which risks of stopout were quite close to one another and to no aid (RR = 0.96, 0.97, and 1.06, respectively).

In other words, for equivalent dollar amounts of financial aid, scholarships consistently had the most beneficial effects on student retention in college throughout their years of enrollment. In the first year of college, work/study and loans also substantially increased retention rates, but their beneficial effects faded with time. Conversely, although non-work/study earnings had little effect on retention in the first year of college, by the third year, they were associated with a substantial improvement in retention compared to students who had financial aid other than scholarships.

18. *Writing about Hierarchical Linear Models*

PROBLEM SET

Harrington and Elliott (2009) conducted a multilevel analysis of individual and neighborhood determinants of overweight and obesity. Selected results from their analysis are shown in table 18A.

TABLE 18A. Estimated coefficients and 95% confidence intervals for a multilevel random intercept model of body mass index,[a] Ontario, Canada, 1992

Variables	Coefficient	Lower 95% confidence limit	Upper 95% confidence limit
Individual level variables			
Age (years)	21.17*	20.35	22.00
Male	0.053*	0.037	0.069
High school not complete	0.94*	0.47	1.41
Married or with partner	NS		
Regular smoker	−0.82*	−1.50	−0.15
Sedentary	0.99*	0.42	1.55
Area-level variables			
Average dwelling value (ref. = high)			
Low	1.93*	1.01	2.78
Middle	1.28*	0.70	1.86
Model statistics			
Level 1 variance (standard error)		19.13* (1.17)	
Level 2 variance (standard error)		0.90* (0.29)	
Intraclass correlation		4.45%	

Adapted from Daniel W. Harrington and Susan J. Elliott, "Weighing the Importance of Neighbourhood: A Multilevel Exploration of the Determinants of Overweight and Obesity," *Social Science and Medicine* 68 (2009) 593–600, table 4, combined model.

[a] Body mass index in kilograms/meter^2

* $p < 0.05$; NS not statistically significant

1. Answer question based on table 18A.

   a. For the methods section, write a series of equations to convey their model specification, including
      i. Level-1
      ii. Level-2

   b. Explain what you learn based on the statistical significance for the level-1 and level-2 variances.

   c. Show how to calculate the intraclass correlation and write a sentence that interprets the number.

d. Write sentences that report and interpret the coefficients for
   i. “regular smoker”
   ii. “low average dwelling value”

2. Based on table 18.1 from Krivo et al. (2009) on p. 392 of *Writing about Multivariate Analysis, 2nd Edition*, write a sentence that identifies the level-1 and level-2 units of analysis, how they relate to one another, and their respective sample sizes.

Subedi et al. (2011) use a three-level HLM to study the effect of student, teacher, and school characteristics on mathematics gain scores over a one-year period among middle school students. Selected results from their analysis are shown in table 18B. Their analysis included 6,184 students and 253 teachers from all middle schools in the Orange County Public Schools. Mathematics scores were from the Norm Referenced Test-Normal Curve Equivalent portion of the FCAT (Florida Comprehensive Assessment Test); the range for this study was from 1 to 99. Mathematics gain scores were calculated by subtracting a student’s 2004 score from his or her 2005 score. The mathematics gain scores in this sample ranged from −31.4 to 45.

TABLE 18B. Fixed effects estimates of mathematics gains scores by student, teacher, and school characteristics, Orange County Public Middle Schools, Florida, 2004–2005

Variable	Coefficient	*t*-statistic
Intercept	19.33**	27.69
Student characteristics		
Mathematics pretest score	0.026**	26.00
Low socioeconomic status[a]	−2.15**	−6.61
Teacher characteristics		
Holds mathematics teaching certification[b]	1.97**	3.21
Teaching experience (years)	0.042*	2.21
School characteristics		
School poverty[c]	−4.15**	−5.00
Cross-level interactions		
Student math pretest score _ teacher math certification	0.033**	3.30
Student math pretest score_school percent advanced math degree[d]	0.015**	3.01
Low student socioeconomic status_teacher math certification	0.91**	2.98
Low student socioeconomic status_school poverty	−0.12**	−2.93

Model also controls for main effects of school mean teacher experience and percentage of teachers in a school that have an advanced mathematics degree. * $p < 0.05$; ** $p < 0.01$.

[a] Low socioeconomic status as assessed by participation in the free and reduced lunch program.

[b] Holds mathematics content-area teaching certification for grades 5–9 or grades 6–12.

[c] Percentage of students in the school who participate in the free and reduced lunch program.

[d] Percentage of mathematics teachers in the school with a masters degree or higher in mathematics.

Adapted from Bidya Raj Subedi, Bonnie Swan, and Michael C. Hynes, “Are School Factors Important for Measuring Teacher Effectiveness? A multilevel Technique to Predict Student Gains through a Value-Added Approach,” *Education Research International* 2011, Article ID 532737, doi:10.1155/2011/532737, table 1.

Use the information in table 18B from the study by Subedi et al. (2011) to answer questions 3 through 5.

3. For the methods section,

   a. Identify the level-1, level-2, and level-3 units of analysis.
   b. Write a paragraph for the methods section that justifies the use of an HLM.
   c. Write a rationale (hypothesis) for examining the cross-level interaction between low socioeconomic status (SES) and mathematics certification.

4. Create a chart to show the effect on mathematics gains scores of the cross-level interaction between SES and mathematics certification, following the guidelines in chapters 6, 16, and 18 and appendix D of *Writing about Multivariate Analysis, 2nd Edition*.

5. Write sentences for the results section that report and interpret

   a. the coefficient on math pretest scores
   b. the coefficient on teaching experience
   c. the shape of the pattern between student SES, teacher mathematics certification, and mathematics change scores, taking into account the cross-level interaction between SES and mathematics certification

Answer questions 6 and 7 based on table 18C, adapted from Subedi et al. (2011).

TABLE 18C. Variance components, variance explained, and statistical significance at teacher and school levels, Orange County Public Middle Schools, Florida, 2004–2005.

Random effect	Variance component	% of variance explained	*p*-value
Teacher-level effect			
Unconditional model	4.50	3.6	<0.0001
Conditional model	4.65	4.6	<0.0001
School-level effect			
Unconditional model	0.47	0.4	0.04
Conditional model	0.26	0.3	0.16

Adapted from: Bidya Raj Subedi, Bonnie Swan, and Michael C. Hynes, "Are School Factors Important for Measuring Teacher Effectiveness? A multilevel Technique to Predict Student Gains through a Value-Added Approach," *Education Research International* 2011, Article ID 532737, doi:10.1155/2011/532737, table 2.

6. Write sentences for the results section that

   a. report and interpret the following aspects of the teacher-level effect:
      i. variance components
      ii. percentage of variance explained
      iii. *p*-values for the unconditional and condition models

b. report and interpret the following aspects of the school-level effect:
   i. variance components
   ii. percentage of variance explained
   iii. *p*-values for the unconditional and condition models

7. Write sentences for the methods section that explain the purpose of comparing the random effects for different levels of analysis from unconditional and conditional models.

Pan et al. (2005) used growth trajectory HLM to study maternal correlates of growth in toddler vocabulary production among children from low-income American families. Selected results from their analysis are shown in table 18D.

TABLE 18D. Estimates of fixed and random effects from a series of individual growth models of toddler vocabulary between ages 14 and 26 months by maternal input, low-income families in Early Head Start

	Unconditional means model		Unconditional growth model		Growth model with number of types[a]	
Variable	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Fixed effects						
Intercept	2.48**	0.45	1.08*	0.43	1.87*	0.09
Age (centered)[b]			2.37***	0.32	1.74*	0.80
Age^2			0.03*	0.01	0.06†	0.04
Mother types[a]					−0.006	0.009
Mother types × age					0.014*	0.007
Mother types × age^2					−0.001**	0.000
Random effects						
Level 1: Time 1	17.81***	2.55	14.76***	2.57	14.31***	2.45
Level 1: Time 2	1,784.62***	274.29	566.82***	111.03	530.89***	106.22
Level 1: Time 3	6,095.77***	977.51	335.18†	206.27	319.02†	203.97
Level 2: Slope (linear)			0.79*	0.42	0.84*	0.42
Goodness of fit						
−2 Log likelihood	2,116.0		1,932.9		1,928.4	
AIC	2,130.0		1,952.9		1,954.4	

Adapted from Barbara Alexander Pan, Meredith L. Rowe, Judith D. Singer, and Catherine E. Snow.. "Maternal Correlates of Growth in Toddler Vocabulary Production in Low-Income Families," *Child Development* 76, no. 4 (2005): 763–82, table 2.

[a] "Types" are the number of different words produced by mother.

[b] Age centered at 14 months.

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; $N = 108$ mother/child dyads

Use the information in table 18B to answer questions 8 through 10.

8. Create a chart to show the overall age pattern of vocabulary development based on the cross-level interaction between mother's word types, child's age, and child's age-squared, following the guidelines

in chapters 6, 10, 16, and 18 and appendix D of *Writing about Multivariate Analysis, 2nd Edition*. Hint: Use the online spreadsheet templates for quadratic specifications and interactions to conduct the calculations and create the chart.

9. Identify the level-1 and level-2 units of analysis in the study by Pan et al (2005).

10. Write the following aspects of the results section:

    a. Interpret the results of the unconditional means model, including the
       i. Intercept
       ii. Random effects terms for level 1
    b. Interpret the results of the unconditional growth model, including
       i. The linear slope
       ii. How the unconditional growth specification adds to the overall fit of the model compared to the unconditional means model

18. *Writing about Hierarchical Linear Models*

SUGGESTED COURSE EXTENSIONS

A. Reviewing

1. Find a journal article in your field that presents results of a two-level hierarchical linear model (OLS). Use the criteria in chapter 18 of *Writing about Multivariate Analysis, 2nd Edition* to evaluate the following aspects of the article:

    a. The authors' rationale for using a hierarchical linear model to address the research question at hand;
    b. Their definition of the units of analysis at each level, and the hierarchical structure by which they relate to one another;
    c. The sample size at each level. Discuss whether sizes meet minimum sample size requirements within levels.
    d. The description of the variables and the levels to which they pertain;
    e. The description of the model specifications, including
        i. the series of models estimated and what substantive question each model or comparison of models is intended to address;
        ii. which parameters are specified with fixed effects and which with random effects.
    f. The table of multivariate HLM results, including whether the authors provided information on
        i. the statistical significance of individual coefficients;
        ii. the overall fit of models across different HLM specifications;
        iii. the variance components.
    g. The prose description of within- and between-group variation, and the intraclass correlation coefficient.
    h. The prose interpretation of key level-1 and level-2 coefficients for their research question, including what they show about the associations among the variables at each level, for each model alone and in conjunction with the other models;
    i. Presentation of cross-level interactions, if included. Consider whether a chart would facilitate presentation of that pattern. If so, sketch the design of the chart to show what goes on each axis and in the legend, and include a complete title, labels, and footnotes.
    j. Their discussion of the strengths and limitations of HLM for their topic and data;

  k. Rewrite the description of the cross-level interaction to correct any shortcomings you identified in parts i and j.

2. Repeat question A.1 for a journal article that presents results of a growth trajectory HLM.

B. Applying Statistics and Writing

Use a data set with a hierarchical structure to perform the following tasks for a two-level HLM, using the guidelines in chapter 18 of *Writing about Multivariate Analysis, 2nd Edition*:

1. Address sample size issues for your HLM:

   a. Assess whether the minimum number of level-1 cases within level-2 units meets the standard for HLM, following the guidelines under "Sample Size" on pp. 389–91;
   b. Evaluate the extent and distribution of missing values at levels 1 and 2;
   c. Write a paragraph for the methods section describing the sample sizes at level 1 and level 2, and your conclusions from parts a and b of this question.

2. Create a table of descriptive statistics for all variables in your analysis, including the following elements:

   a. Organize the variables to convey the level at which each variable is measured, and to identify the dependent variable.
   b. Report the sample size at each level.
   c. Label each variable to convey its meaning, units, and categories, following the conventions in chapters 5 and 18.
   d. Report the pertinent descriptive statistics for that type of variable (level of measurement).

3. Conduct the following steps for an HLM with your data:

   a. Estimate the following three types of models:
      i. an unconditional (null) model;
      ii. a random intercept model;
      iii. a random intercept-and-slope model in which you permit the intercept and the slope of *one* key independent variable to vary randomly across the level-2 units.
   b. Write a series of equations to convey the statistical specifications for your HLM.
   c. Write a paragraph for the methods section describing the series of models you estimated, using the guidelines in chapter 18, including the following elements:
      i. what substantive questions each model or comparison of models is intended to address;

ii. which parameters are specified with fixed effects and which with random effects;
iii. the covariance structure and error terms you used;
iv. your assumptions about distributions of residuals at each level of analysis;
v. the type of software, estimation method, and algorithm used to estimate your models.

4. Create a table of multivariate HLM results for the series of models you estimated for the preceding question, including the following elements and following the guidelines in chapters 5 and 18:

   a. Label each model to convey the pertinent type of HLM specification (e.g., unconditional means, fixed effects, random effects, etc.);
   b. Organize the variables to convey the level at which they were measured, and to identify cross-level interactions;
   c. Report estimated coefficients and inferential statistical test information for each variable;
   d. Report variance components for each model;
   e. Report goodness-of-fit statistics for each model.

5. If your specification includes a cross-level interaction, create a chart to present that pattern, following the guidelines in chapters 6, 16, and 18 of *Writing about Multivariate Analysis, 2nd Edition*.

6. Write paragraphs to report and interpret the results of your multivariate HLM models, working from the table you created in question B.4. Be sure to address each of the following topics and to follow the guidelines in chapters 15 and 18:

   a. the direction, magnitude, and statistical significance of your key individual level-1 and level-2 coefficients;
   b. how the coefficient on a key level-1 variable compares when treated as fixed (in the random intercept model) and when treated as random (in the random intercept-and-slope model), and what you can learn about the variation in that key level-1 variable across level-2 units from comparing that coefficient from those two different HLM specifications;
   c. description of any cross-level interaction patterns included in your specification, referring to the chart created in the preceding question;
   d. how the overall fit of the model compares between the unconditional model, the random intercept model, and the random intercept-and-slope model, and what you can learn about the pattern of variation within and across levels from the comparison of model goodness of fit;
   e. the variance components results for the different specifications;
   f. within- and between-group variation, and the intraclass correlation coefficient for different models;

g. what your results show about the associations among the key independent variables at each level, for each model alone and in conjunction with the other models.

C. Revising

1. Repeat question A.1 for a paper you have written previously about an application of an HLM. Revise it to rectify any shortcomings you identified.

2. Evaluate a table of multivariate HLM results you created previously, using the guidelines in chapters 5 and 18 of *Writing about Multivariate Analysis, 2nd Edition*.

3. Repeat questions B.5 and B.6c for a paper you have written previously about an HLM involving a cross-level interaction. Revise those descriptions to rectify any shortcomings you identified.

4. Exchange drafts of your materials from questions C.1 through C.3 with someone who conducted an HLM analysis of a different research question or data. Peer-edit each other's work and revise according to the feedback you receive.

18. *Writing about Hierarchical Linear Models*

SOLUTIONS

1. a. i. $BMI_{ia} = \alpha_{0a} + \alpha_{1a}AGE + \alpha_{2a}MALE + \alpha_{3a}INC\ HIGHSCHOOL + \alpha_{4a}MARRIED + \alpha_{5a}SMOKER + \alpha_{6a}SEDENTARY + \varepsilon_{ia}$, where the subscript *i* is used to index individuals and *a* to index neighborhoods.

   ii. $\alpha_{0a} = \eta_{00} + \eta_{01}LOW\ DWELLING\ VALUE + \eta_{02}MID\ DWELLING\ VALUE + \gamma_{0a}$ The title of table 18A indicates that a random intercept model was estimated. Only the intercept in the level-1 equation is permitted to vary randomly as a function of area-level characteristics. The slopes are treated as fixed effects.

   b. Even after accounting for a variety of individual-level characteristics (e.g. age, sex, education level), there is still statistically significant random variation in the BMI between individuals. There remains random variation across areas/neighborhoods in BMIs after accounting for the average dwelling value of areas.

   c. The intraclass correlation can be calculated level-2 variance / (level-1 variance + level-2 variance). Substituting values from table 8A, we obtain = 0.90 / (19.13 + 0.90) = 4.45%. Approximately 4.5% of the total variation in BMI across individuals can be explained by differences in the neighborhoods in which they live.

   d. i. Regular smokers were estimated to have a BMI that is, on average, 0.82 kg/m^2 lower than a nonsmoker, adjusting for other individual and neighborhood characteristics ($p < 0.05$).

   ii. Average BMI was positively associated with extent of neighborhood disadvantage. Individuals living in the mid-dwelling-value areas had 1.28 kg/m^2 higher BMI, and those in the low-dwelling-value areas 1.93 kg/m^2 higher, when each was compared with those living in the high-dwelling-value areas (both $p < 0.05$).

3. a. Level-1 = student; Level-2 = teacher; Level-3 = school

   b. “Hierarchical data structures are present in education settings where students are nested within a teacher and teachers are nested within a school. The nesting form of the data structure generates a hierarchical linear model (HLM). In other words, models at different levels can be built based on a specific number of lower level units nested within upper level units, eventually forming a HLM design. . . . Thus, in such situations, students’ gain scores in mathematics from one year to the next can be predicted based on characteristics not only of the student, but also of the teacher (e.g., teacher qualifications and experience) and of the school (e.g., poverty)." (Adapted from Subedi et al. [2011], p. 4.)

c. The positive effect on mathematics gains scores from having a teacher with content certification in mathematics may be enhanced among students of lower student socioeconomic status.

5. a. When other student, teacher, and school-level characteristics were taken into account, each one-point increase in baseline (pretest) math scores was associated with approximately a quarter of a point increase in math gain scores between rounds ($p < 0.01$).

   b. Each additional year of teaching experience was associated with a four-tenths of a point increase in students' math gain scores ($p < 0.01$).

   c. Student SES and teacher mathematics content certification interacted in their effect on students' math gain scores: Low SES students (as identified by participation in the free lunch program) had lower mean math gains than high SES students. Moreover, although having a math content certified teacher was associated with higher math gain scores regardless of student SES level, the effect was amplified for low SES students. Having a math certified teacher was associated with a mean math gain score of 2.88 points among low SES students, versus a mean increase of 1.97 points among higher SES students ($p < 0.01$).

7. The unconditional models estimate the amount of total variation in students' mathematics gains scores that is found between teachers (level-2) and between schools (level-3). Statistically significant level-2 and level-3 variance components in the unconditional models would indicate that it is important to consider teacher-level and school-level factors, and that an HLM is appropriate. The conditional models, which add student, teacher, and school characteristics, are estimated in order to assess the degree to which the random variation across levels can be accounted for by the included factors.

9. In the study by Pan et al. (2005) the level-1 unit of analysis was a time point at which child's vocabulary was measured, and level-2 unit of analysis was the mother/child dyad. In the longitudinal study, time points were nested within children.

19. *Speaking about Multivariate Analyses*

PROBLEM SET

1. Adapt the material in text box 20.3 and figure 20.5 on pp. 460–61 of *Writing about Multivariate Analysis, 2nd Edition* into slides for a 10-minute presentation to a general audience, including the comments that explain how the material illustrates the principles of how to write about numbers.

2. Write the speaker's notes to accompany the slides you created for the previous question, following the guidelines in chapter 19.

3. Create one or more slides to present the following material to a scientific audience. "The Center for Epidemiological Studies-Depression Scale (CES-D) is a 20-item scale for epidemiological research that was developed by the National Institute of Mental Health. Respondents are asked to choose from four possible responses in a Likert format, where '0' is 'rarely or none of the time (less than one day per week),' and '3' is 'almost all or all of the time (five to seven days per week).' The theoretical range for the overall CES-D is from 0 to 60, with higher scores reflecting greater levels of depressive symptoms. The CES-D has four separate factors: depressive affect, somatic symptoms, positive affect, and interpersonal relations. The CES-D has very good internal consistency with alphas of 0.85 for the general population and 0.90 for a psychiatric population (Radloff 1977)."

4. Adapt the following tables into simpler tables or charts for use on slides for a speech. Aim for one concept or series of closely related concepts per chart. See table 6.1 on pp. 140–41 in *Writing about Multivariate Analysis, 2nd Edition* for guidance on which type of chart to use for each topic.

   a. Table 5.1 ("Households by type, race, and Hispanic origin" p. 80)
   b. Table 6C ("Estimated log-odds of first trip to the United States," p. 39 of the *Study Guide to Writing about Multivariate Analysis, 2nd Edition*).
   c. Table 11A. ("Effect of own SAT scores and roommate's SAT scores on cumulative grade point average, by range of own SAT score," Zimmerman [2003], p. 85 of the *Study Guide to Writing about Multivariate Analysis, 2nd Edition*). Create one chart to show how the coefficients on own and roommate's math and verbal SAT

scores vary across the models for different levels of combined own SAT score.

5. Write Vanna White notes to introduce and explain the following the tables or charts to a scientific audience. Use the GEE approach to summarize the patterns where appropriate:

    a. Figure 6.8 ("Log-odds from competing risks model of reasons for program disenrollment," p. 124 in *Writing about Multivariate Analysis, 2nd Edition*)
    b. Figure 6.2b ("Federal outlays by function, 2000," p. 116)
    c. Figure 6.12 ("Predicted birth weight by race/ethnicity and income-to-poverty ratio," p. 129)
    d. Table 7.1. ("Poverty rates [%] by group under current and proposed poverty measures, United States, 1992," p. 166)
    e. Figure 16.1 ("Predicted difference in birth weight by mother's educational attainment and race/ethnicity," p. 342)

6. Practice presenting one table and one chart from question 5, using the Vanna White notes you wrote for that exhibit. Evaluate each of those mini-presentations using the checklist in chapter 19. Revise the oral presentation of each slide to fit within two minutes.

7. Create the following materials for speeches.

    a. Adapt the material in table 15A (p. 116 of this study guide) into a series of chart slides demonstrating why a multivariate model is needed to assess the impact of the Yonkers Residential Mobility Program on neighborhood and housing outcomes. Aim for one concept or series of closely related concepts per chart. Include text annotations to describe the patterns.
    b. Adapt the multivariate model results from table 15B (p. 117 of this study guide) into one or two chart slides.
    c. Write speaker's notes for the slides you created in parts a and b, including Vanna White descriptions of charts, and transition sentences between slides, following the guidelines in the section on "Speaker's Notes" on pp. 430–34 of *Writing about Multivariate Analysis, 2nd Edition*.

19. *Speaking about Multivariate Analyses*

SUGGESTED COURSE EXTENSIONS

A. Writing

1. Create slides and speaker's notes for a 20-minute presentation of a paper involving a multivariate analysis, to be presented at a professional conference in your field. Include slides for each major section of the paper, including introduction, literature review, data and methods, results (several charts or tables; see question A.2), and conclusions.

2. Adapt charts or tables from your paper to be used on the slides. Write speaker's notes with Vanna White directions for each.

3. Exchange draft slides and speaker's notes with a colleague who is working on a different topic and data. Evaluate each other's work, using the checklist from chapter 19 of *Writing about Multivariate Analysis, 2nd Edition*. Revise your slides and speaker's notes according to the feedback you receive.

4. Ask a test audience to evaluate a live presentation of your talk for your specified audience and allotted time, using the criteria on "Dress Rehearsal" on pp. 435–36.

5. Make revisions to slides and speaker's notes based on what you learned in your rehearsal.

B. Revising

1. Critique and revise slides you have created previously for a 15-to-20-minute speech about a multivariate analysis to a scientific audience, using the criteria in chapter 19.

2. Critique and revise the speaker's notes for the same speech.

3. Revise a table of multivariate results from your paper into several simpler table slides or chart slides, with individualized titles that reflect the specific content of each slide.

4. Write Vanna White notes to introduce and explain one table and one chart from your revised presentation following the guidelines on pp. 431–34.

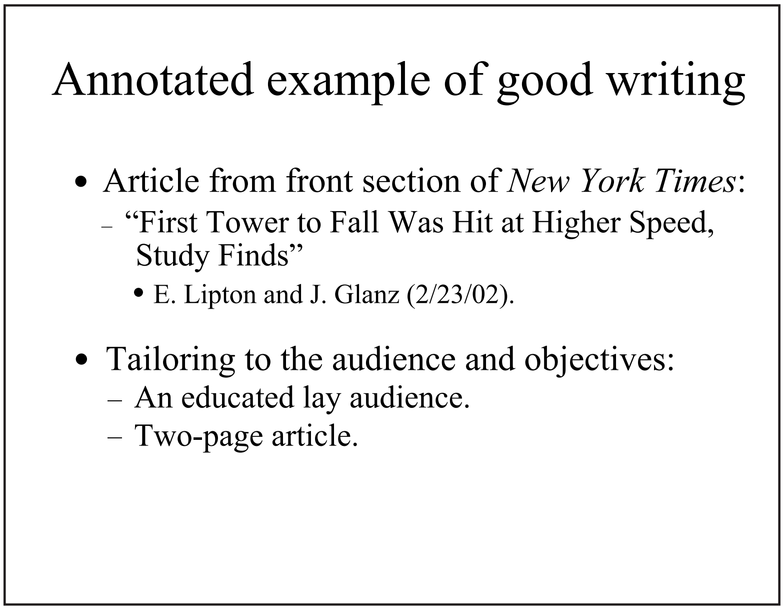
5. Exchange your revised work from questions B.1 through B.3 with someone working on a different topic and data. Peer-edit and revise your work according to the feedback you receive.

6. Revise the slides from question B.1 to create a 10-minute presentation for a lay audience.

19. *Speaking about Multivariate Analyses*

SOLUTIONS

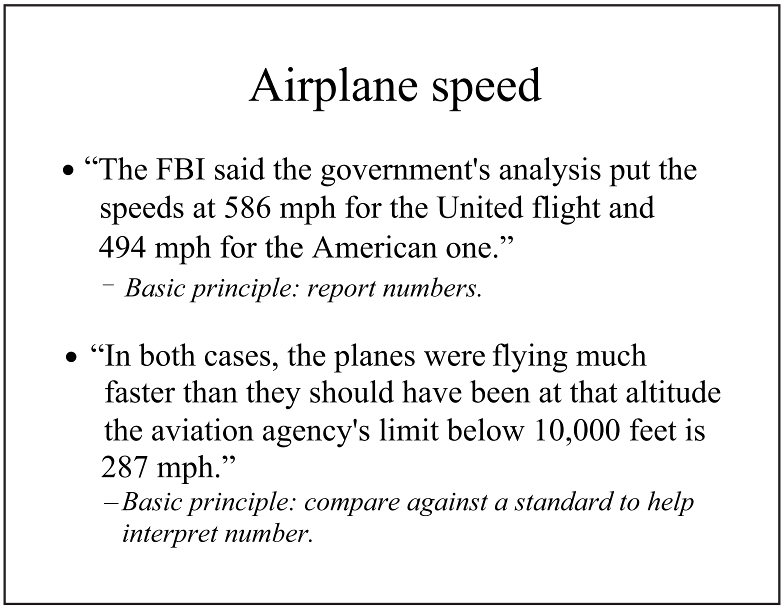
1. Figures 19A–19G are slides for a presentation about the physical impact of the planes on the Twin Towers (box 20.3 and figure 20.5).



Annotated example of good writing

- Article from front section of *New York Times*:
  - “First Tower to Fall Was Hit at Higher Speed, Study Finds”
    - E. Lipton and J. Glanz (2/23/02).
- Tailoring to the audience and objectives:
  - An educated lay audience.
  - Two-page article.

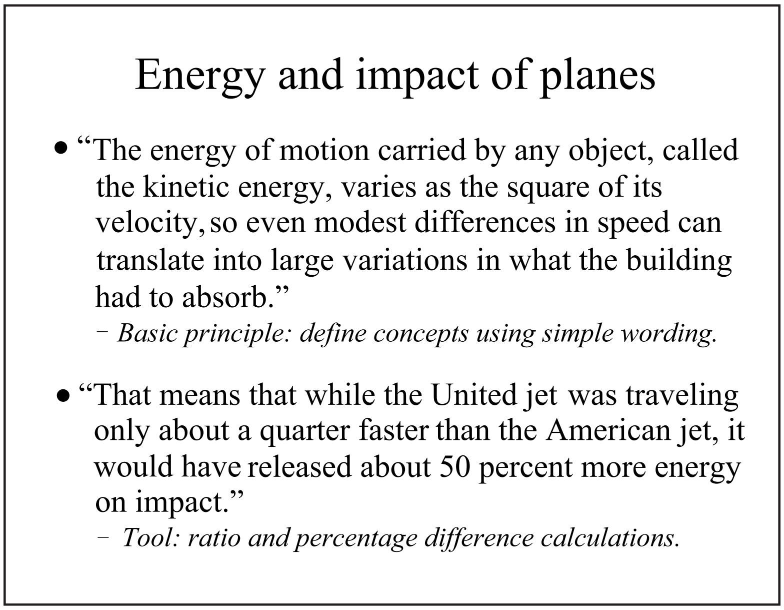
Figure 19A.



Airplane speed

- “The FBI said the government's analysis put the speeds at 586 mph for the United flight and 494 mph for the American one.”
  - *Basic principle: report numbers.*
- “In both cases, the planes were flying much faster than they should have been at that altitude the aviation agency's limit below 10,000 feet is 287 mph.”
  - *Basic principle: compare against a standard to help interpret number.*

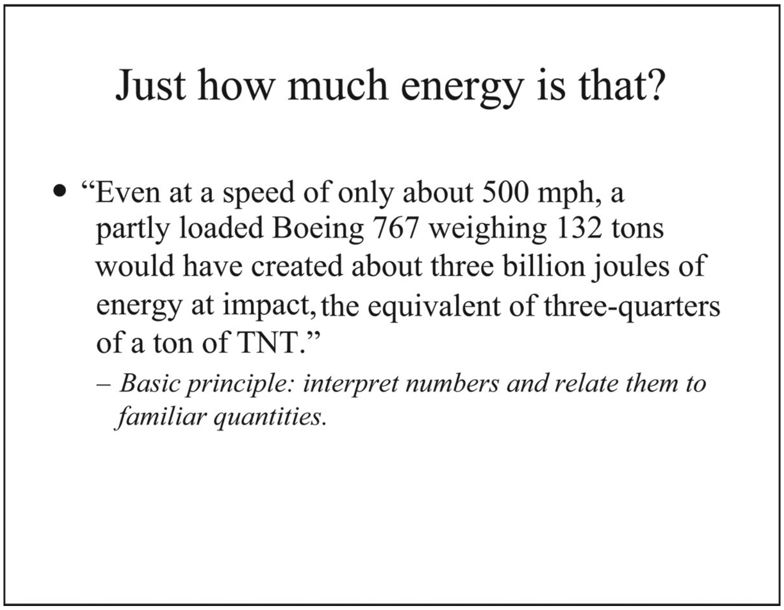
Figure 19B.



Energy and impact of planes

- “The energy of motion carried by any object, called the kinetic energy, varies as the square of its velocity, so even modest differences in speed can translate into large variations in what the building had to absorb.”
  - *Basic principle: define concepts using simple wording.*
- “That means that while the United jet was traveling only about a quarter faster than the American jet, it would have released about 50 percent more energy on impact.”
  - *Tool: ratio and percentage difference calculations.*

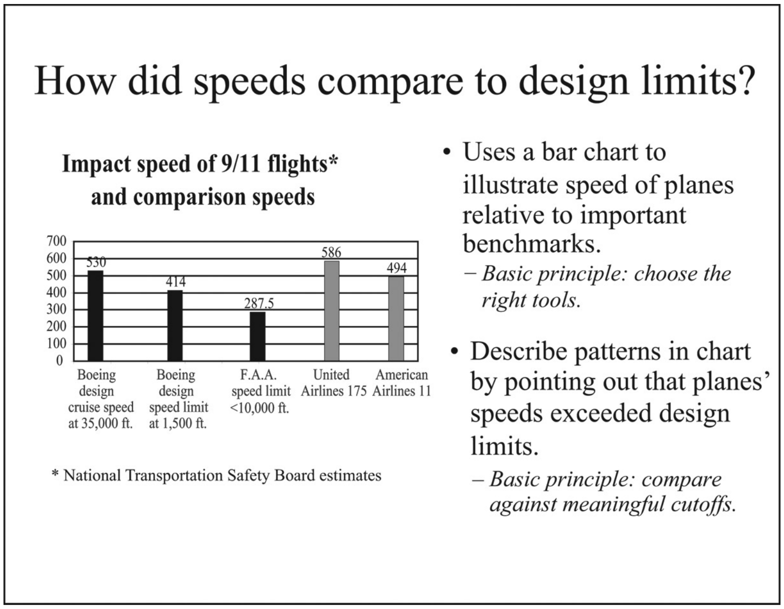
Figure 19C.



Just how much energy is that?

- "Even at a speed of only about 500 mph, a partly loaded Boeing 767 weighing 132 tons would have created about three billion joules of energy at impact, the equivalent of three-quarters of a ton of TNT."
  - *Basic principle: interpret numbers and relate them to familiar quantities.*

Figure 19D.



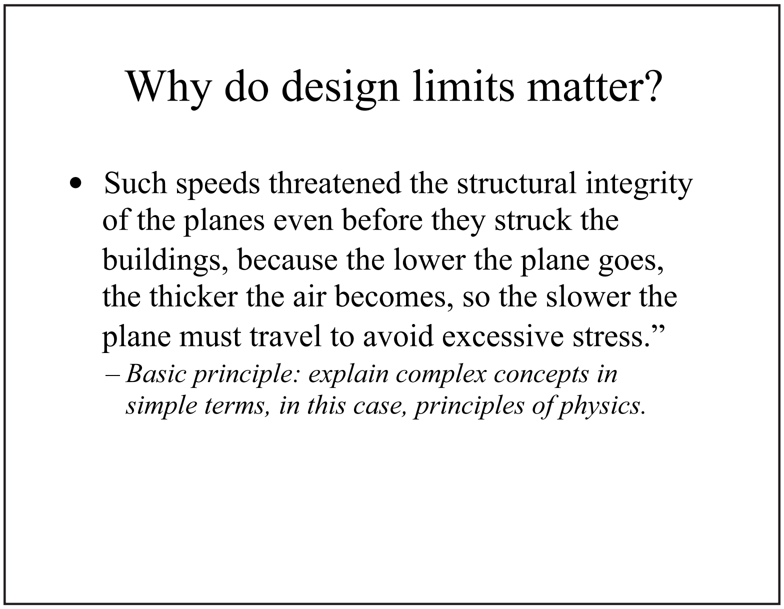
How did speeds compare to design limits?

Impact speed of 9/11 flights* and comparison speeds

* National Transportation Safety Board estimates

- Uses a bar chart to illustrate speed of planes relative to important benchmarks.
  - *Basic principle: choose the right tools.*
- Describe patterns in chart by pointing out that planes' speeds exceeded design limits.
  - *Basic principle: compare against meaningful cutoffs.*

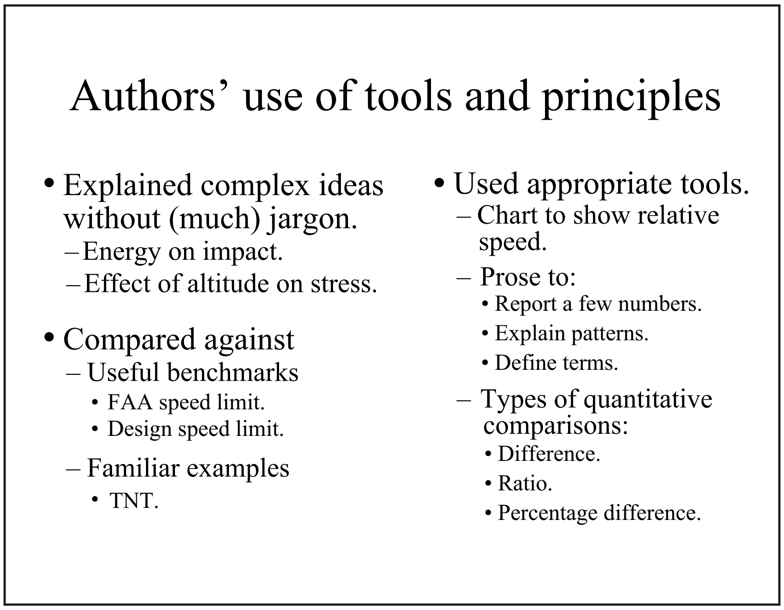
Figure 19E.



Why do design limits matter?

- Such speeds threatened the structural integrity of the planes even before they struck the buildings, because the lower the plane goes, the thicker the air becomes, so the slower the plane must travel to avoid excessive stress.”
  - *Basic principle: explain complex concepts in simple terms, in this case, principles of physics.*

Figure 19F.

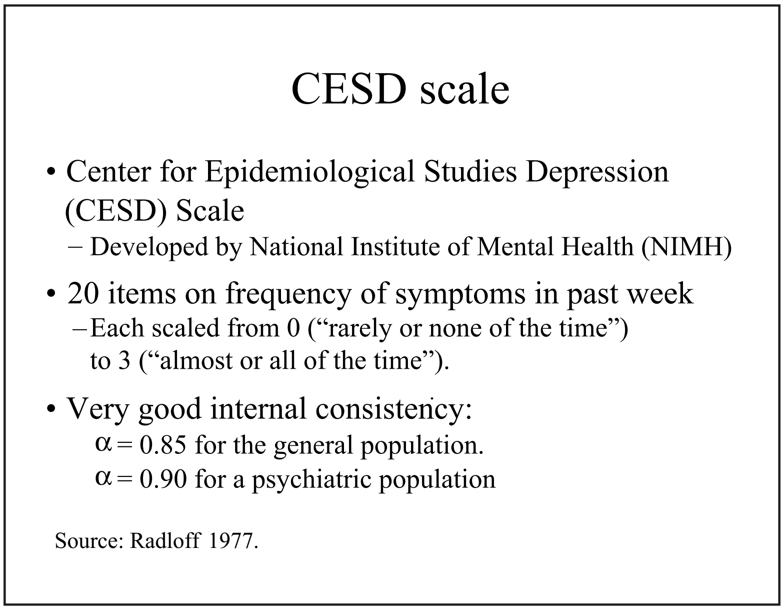


Authors’ use of tools and principles

- Explained complex ideas without (much) jargon.
  - Energy on impact.
  - Effect of altitude on stress.
- Compared against
  - Useful benchmarks
    - FAA speed limit.
    - Design speed limit.
  - Familiar examples
    - TNT.
- Used appropriate tools.
  - Chart to show relative speed.
  - Prose to:
    - Report a few numbers.
    - Explain patterns.
    - Define terms.
  - Types of quantitative comparisons:
    - Difference.
    - Ratio.
    - Percentage difference.

Figure 19G.

3. Figures 19H and 19I are slides about data and methods regarding CES-D scale for a scientific audience.



CESD scale

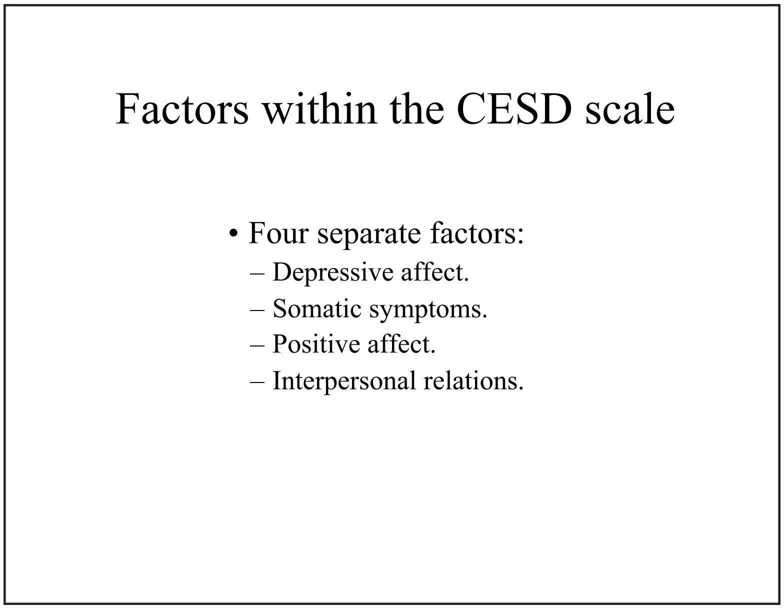
- Center for Epidemiological Studies Depression (CESD) Scale
  - Developed by National Institute of Mental Health (NIMH)
- 20 items on frequency of symptoms in past week
  - Each scaled from 0 (“rarely or none of the time”) to 3 (“almost or all of the time”).
- Very good internal consistency:

  $\alpha = 0.85$ for the general population.

  $\alpha = 0.90$ for a psychiatric population

Source: Radloff 1977.

Figure 19H.



Factors within the CESD scale

- Four separate factors:
  - Depressive affect.
  - Somatic symptoms.
  - Positive affect.
  - Interpersonal relations.

Figure 19I.

5. Vanna White notes in “[ ]”, and GEE approach to describe figures or tables

   a. “Figure 6.8 illustrates how the chances of disenrolling from the State Children’s Health Insurance Program vary by reason and demographic factors, based on a set of competing risks models controlling for all variables shown in the chart. Demographic factors are arrayed across the *x* axis [wave horizontally]. Each cluster [point to one] shows how that factor is associated with each of the three possible reasons for disenrollment, with other insurance shown in gray, other government program in white, and nonpayment in black [point at respective bars]. The log-odds of disenrollment are shown on the *y* axis [wave vertically]. Bars that drop below the line at $y = 0.0$ represent lower odds than in the reference category, while those above the line represent higher odds." (Interpret the pattern as in the description of figure 6.8 on p. 124 of *Writing about Multivariate Analysis*, *2nd Edition*.)

b. "The distribution of federal outlays by major function in the United States in 2000 is shown in figure 6.2b. Human resources (the black wedge [point]) comprised by far the largest single category of federal outlays (61% of the $1.8 trillion spent that year). The second largest category—national defense (dotted fill)—accounted for only about one-quarter as much as human resources (16% of the total). Net interest, physical resources, and other functions together amounted for the remaining 23% of total outlays [point to each wedge as you mention its category]."

c. "The predicted pattern of birth weight by race/ethnicity and income-to-poverty ratio (or 'IPR') is shown in figure 6.12. The results are based on a multivariate model with controls for gender, mother's age, educational attainment, and smoking status. The *x* axis shows the IPR, ranging from 0 to 4 times the poverty line [wave across horizontal axis]. There are separate lines for each of the racial/ethnic groups—the solid line for non-Hispanic whites, the dotted line with triangles for Mexican Americans, and the dashed line with squares for non-Hispanic blacks [point at each in turn, top to bottom on the left-hand side of the *x* axis]. Predicted birth weight in grams is shown on the *y* axis [wave vertically]."

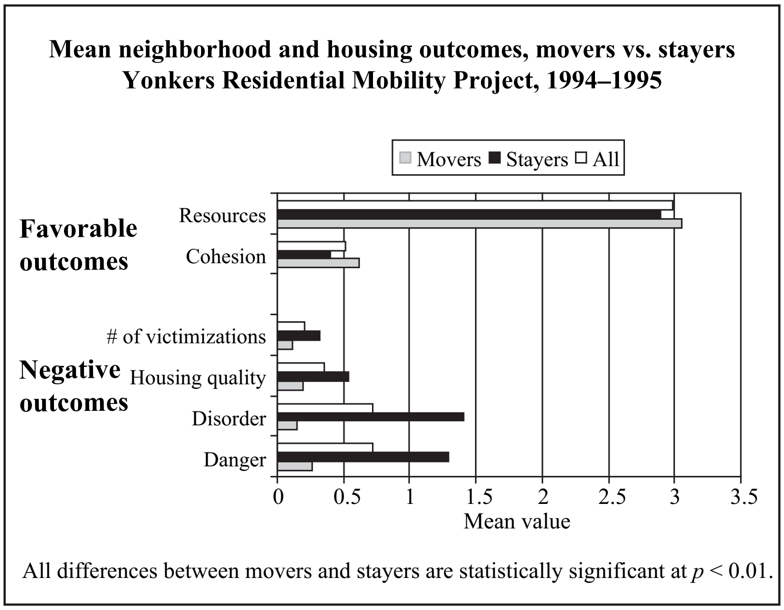
d. "Table 7.1 shows poverty rates for the United States in 1992 under different poverty definitions, for the overall population and several age and racial groups in the rows [gesture at the row labels]. The leftmost column of numbers [point to 'Current' column label] is the poverty rate under the current poverty definition, while the next two columns to the right [point to 'Proposed measure' column labels] show poverty rates under two alternative definitions. The rightmost two columns [point to 'Percentage point change' column label] show the percentage point change in the poverty rate between each of the two alternative definitions and the current definition." [Note: Explain the alternative poverty definitions on a previous slide, as viewers will focus on the results when presented with the table. Remove the footnote from the slide of this table and turn it into a text slide to precede the table slide.]

"Under either of the proposed alternative definitions, the poverty rate is several percentage points higher than under the current definition. For example, the overall poverty rate would increase by 3.6 points under alternative definition 1, and by 4.5 points under alternative definition 2 [point to pertinent cells in 'Total population' row]. Differences for some subgroups are quite small. For example, the poverty rate for the elderly would be projected to increase by only 1.7 percentage points under alternative 1. For other groups, such as Hispanics, the projected increases are considerably larger: 10.6 points [point to pertinent cell]."

e. "Figure 16.1 shows the predicted differences in birth weight by mother's educational attainment and race/ethnicity. Racial/ethnic groups are shown in the legend [name them and point to associated bar colors: white bars for non-Hispanic white infants, striped for Mexican American, and black bars for non-Hispanic blacks]. Educational attainment is shown in increasing order across the *x*-axis [name them and gesture along the horizontal axis]. The length of each bar shows the difference in predicted birth weight (grams) between the pertinent group and non-Hispanic whites with at least some college, which is the reference category from the multivariate model." (Describe the pattern as in box 16.1 on p. 363 of *Writing about Multivariate Analysis, 2nd Edition*.)

7. Slides to present results of Yonkers Residential Mobility Program evaluation (Fauth et al. 2004).

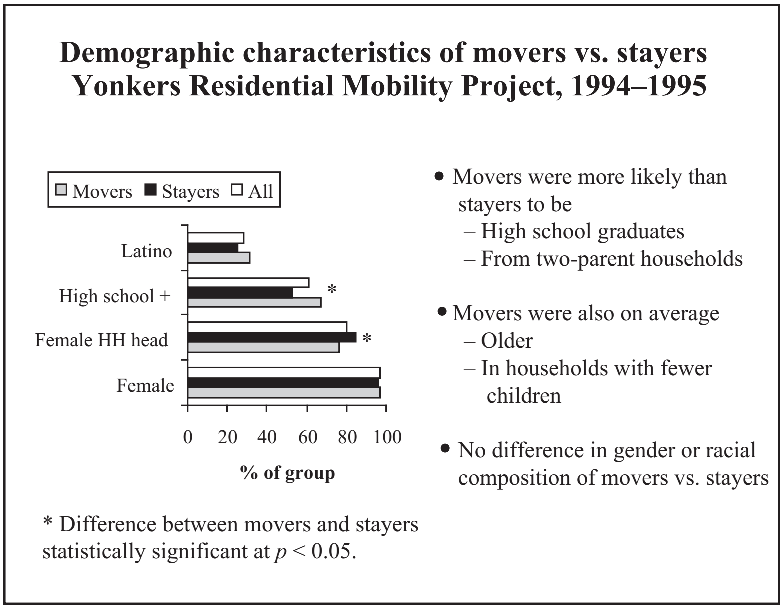
   a. Figures 19J through 19L are slides demonstrating why a multivariate model is needed.



Mean neighborhood and housing outcomes, movers vs. stayers Yonkers Residential Mobility Project, 1994–1995

All differences between movers and stayers are statistically significant at $p < 0.01$.

Figure 19J.

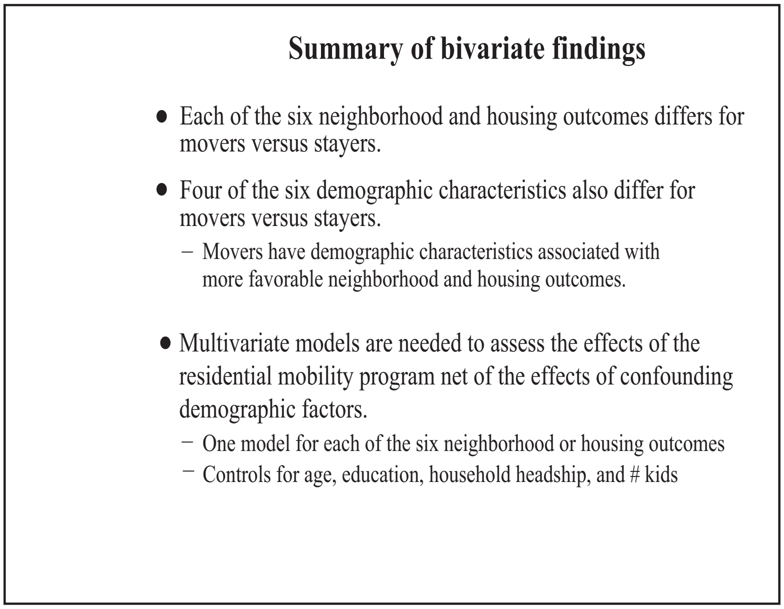


Demographic characteristics of movers vs. stayers Yonkers Residential Mobility Project, 1994–1995

- Movers were more likely than stayers to be
  - High school graduates
  - From two-parent households
- Movers were also on average
  - Older
  - In households with fewer children
- No difference in gender or racial composition of movers vs. stayers

* Difference between movers and stayers statistically significant at $p < 0.05$.

Figure 19K.

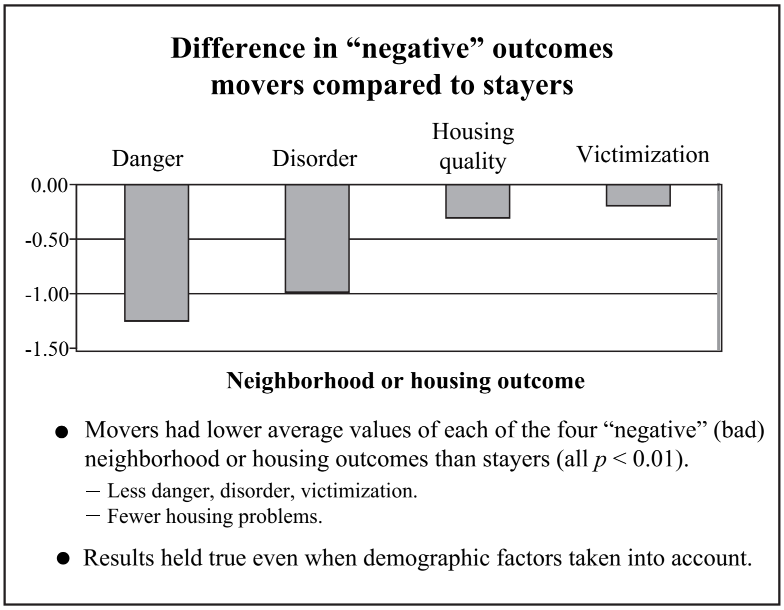


Summary of bivariate findings

- Each of the six neighborhood and housing outcomes differs for movers versus stayers.
- Four of the six demographic characteristics also differ for movers versus stayers.
  - Movers have demographic characteristics associated with more favorable neighborhood and housing outcomes.
- Multivariate models are needed to assess the effects of the residential mobility program net of the effects of confounding demographic factors.
  - One model for each of the six neighborhood or housing outcomes
  - Controls for age, education, household headship, and # kids

Figure 19L.

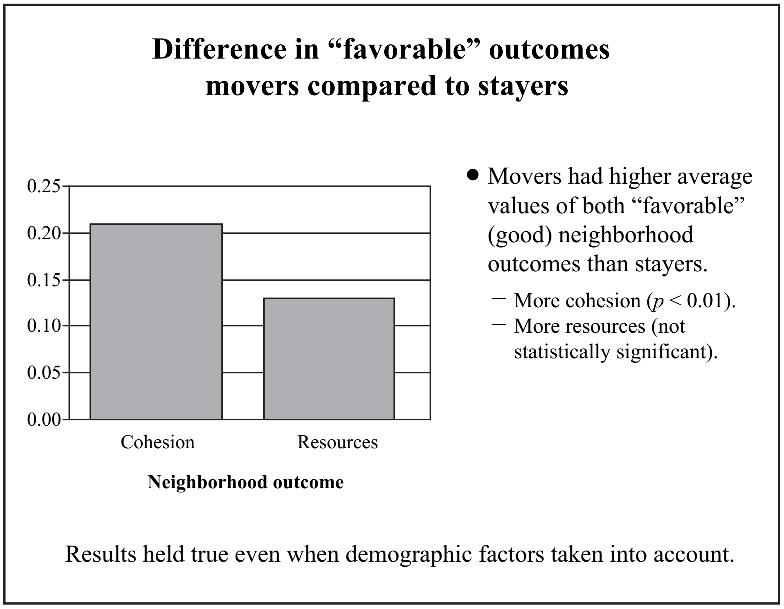
b. Figures 19M and 19N are slides presenting multivariate model results.



Difference in "negative" outcomes movers compared to stayers

- Movers had lower average values of each of the four "negative" (bad) neighborhood or housing outcomes than stayers (all $p < 0.01$).
  - Less danger, disorder, victimization.
  - Fewer housing problems.
- Results held true even when demographic factors taken into account.

Figure 19M.



Difference in "favorable" outcomes movers compared to stayers

- Movers had higher average values of both "favorable" (good) neighborhood outcomes than stayers.
  - More cohesion ($p < 0.01$).
  - More resources (not statistically significant).

Results held true even when demographic factors taken into account.

Figure 19N.

c. Vanna White notes are shown in "[ ]."

"Figure 19J shows mean values of six different measures of neighborhood and housing quality for low-income families who moved to low-poverty neighborhoods versus those who stayed in high-poverty neighborhoods. In all six dimensions studied, outcomes were statistically significantly better among movers (the gray bars) than among stayers (the black bars). Favorable outcomes (cohesion and resources) [gesture at top two clusters on the chart] were higher among movers than stayers, whereas negative outcomes (danger, victimizations, disorder, and indicators of poor housing) [gesture at four bottom clusters on the chart] were all lower among movers than stayers.

"However, it is important to consider whether differences in demographic characteristics might explain some of the observed differences in these outcomes. Although participants in the Yonkers Residential Mobility Program were randomly assigned to be movers or stayers, some differences in these characteristics are possible. In figure 19K, we see that four of the six background characteristics are more auspicious among movers than stayers. Movers were more likely to be from two-parent households and to have completed high school [point to respective clusters on chart]. They were also on average older and had fewer children in the household.

"[Transition to slide 19L] These bivariate statistics suggest that multivariate models are needed to assess the impact of residential status on each of the outcomes, net of the potentially confounding effect of the background characteristics. All of the observed differences in background characteristics would be expected to favor better outcomes among movers than stayers regardless of residence. For example, older age, two-parent households, better education, and smaller families are often associated with better resources than younger, female-headed, less-educated, and larger families. Hence multivariate models are needed to control for those characteristics.

"Figure 19M shows results of multivariate models of the four negative measures of neighborhood characteristics and housing quality studied as part of the Yonkers Residential Mobility Program (danger, victimization, disorder, and housing problems) [point to respective bars, moving left to right across *x*-axis]. Even when the effects of potential confounders were taken into account, subjects who moved had statistically significant better values of each of these four outcomes than those who remained in their original neighborhoods. Put differently, movers experienced less danger, victimization, disorder, and housing problems than stayers.

"Figure 19N shows the results of multivariate models of the two favorable outcomes (cohesion and resources). Both were higher (better) among movers [gesture along *y* axis], but the difference in resources was not statistically significant. Although some of the background control variables were statistically significantly associated with one or two of the outcomes, none showed a consistent pattern of association."

20. *Writing for Applied Audiences*

PROBLEM SET

1. Write the following components of a two-page policy brief about the study by Fauth et al. (2004), using the information in tables 15A and 15B on pp. 116–17 of this study guide and the guidelines on pp. 451–55 of *Writing about Multivariate Analysis, 2nd Edition*. It may be helpful to obtain a copy of the complete article, which is available online. (See table notes for reference.)

    a. A title.
    b. One or two simplified tables or charts to summarize their key results. Hint: Use some of the figures you created for question 7 of the problem set to chapter 19.
    c. Short descriptions of each table or chart from part b of this question.
    d. Paragraphs explaining how the findings apply to at least two sets of stakeholders.
    e. A sidebar describing the study methods.

2. Using the information in table 6C on p. 39 of this study guide and the guidelines on pp. 455–56 of *Writing about Multivariate Analysis, 2nd Edition*, design chartbook pages to present the results of the analysis by Fussell and Massey (2004) to an applied audience. Adapt the charts you created for question 9 in the problem set to chapter 6, and design other charts to illustrate the remaining results. It may be helpful to obtain a copy of the complete article, which is available online. (See table notes for reference.)

Answer questions 3 and 4 using the information in boxes 12.1, 12.2, 13.1, 13.2, 15.1b, and 15.2b (*Writing about Multivariate Analysis, 2nd Edition*).

3. Design a research poster about the birth weight study for an applied audience. Sketch the poster layout and provide notes about the contents of each page, adapting them from the tables, charts, slides, and text boxes from *Writing about Multivariate Analysis, 2nd Edition*.

4. Write a one-page general-interest article about the birth weight study.

5. Write an executive summary of the study by Zimmerman about peer effects on academic outcomes (2003). See questions 3 through 7 in the

problem set for chapter 9 (pp. 62–63 of this study guide), and associated reference.

6. Outline a descriptive report about the Zimmerman study for a lay audience.

    a. Write the section headings—one for each major question or topic covered in that study.
    b. Adapt table 11A (p. 85 of this study guide) into simplified tables or charts, each of which focuses on one finding or set of related findings. Write the titles for the charts or tables that would go in each section of the report.

20. *Writing for Applied Audiences*

SUGGESTED COURSE EXTENSIONS

Reviewing

1. Find a poster related to your interests at a professional conference in your field. Discuss the research project with the poster's author. After you return, write a critique evaluating the following, using the criteria under "Posters" on pp. 447–51 of *Writing about Multivariate Analysis, 2nd Edition*:

   a. Title of the poster
   b. Ease of understanding of data and methods description for (i) researchers in your field; (ii) nonstatisticians
   c. Accessibility of research findings to (i) researchers in your field; (ii) nonstatisticians
   d. Relevance of conclusions for an applied audience
   e. Clarity of the overall story line on the poster
   f. Poster layout
   g. Type size and other formatting
   h. Availability and quality of handouts
   g. Length and clarity of the presenter's oral description of the poster contents

2. Find an issue brief or policy brief related to a research topic in your field or at a website such as the Urban Institute (http://www.urban.org). Critique the following elements of the brief, using the guidelines under "Issue and Policy Briefs" on pp. 451–55:

   a. Ease of understanding for nonstatisticians
   b. Simplicity of tables and charts
   c. Appropriateness of vocabulary for the intended audience
   d. Layout

3. Find a chartbook about a research topic in your field or at a website such as the US Social Security Administration (http://www.ssa.gov/policy/docs/chartbooks/) or Healthy People 2020 (http://www.healthypeople.gov/). Critique it using the criteria on pp. 455–56.

4. Find a descriptive report about a topic in your field or at a website such as the Office of Human Services Policy (http://aspe.hhs.gov/topics0.cfm/). Critique it using the criteria on pp. 456–57.

5. In the popular press, find a general-interest article about a technical topic. Critique it using the criteria on p. 458.

B. Writing

1. Create a 4' by 8' poster about a research paper for a conference of your professional association following the guidelines on pp. 447–51 of *Writing about Multivariate Analysis, 2nd Edition*.

   a. Design pages for each major section of the paper, including an introduction, literature review, data and methods, results (several charts or tables; see question B.2 below), and conclusions.
   b. Draft the layout of the poster, including space for a title banner and abstract as well as the pages from part a of this question.

2. Adapt charts or tables from your paper to be used on the poster. Write titles and Vanna White notes for each table or chart following the guidelines on pp. 431–34.

3. Write a narrative to accompany your poster. Include short modules for each of the following.

   a. An introduction to your topic and project
   b. The key findings of your study
   c. The policy or program implications of your work
   d. The research implications of your work
   e. A description of the data used in your analysis
   f. An explanation of your methods for someone familiar with multivariate statistics
   g. An explanation of your methods for nonstatisticians

4. Create handouts.

   a. For a statistical audience
   b. For an applied audience

5. Critique and revise the poster, narrative, and handouts.

   a. Ask a colleague to evaluate your poster and associated narrative and handouts, given your specified audience and using the criteria under "Posters" on pp. 447–51 of *Writing about Multivariate Analysis, 2nd Edition*.
   b. Revise the poster, narrative, and handouts based on what you learned in your rehearsal.

6. Write a two-page issue brief about a multivariate analysis, following the guidelines on pp. 451–55.

7. Write a two- or three-page general-interest article about the purpose, findings, and implications of your multivariate analysis, following the guidelines on p. 458.

8. Write a chartbook about a multivariate analysis, following the guidelines on pp. 455–56.

9. Repeat questions A.1 through A.5 from the suggested course extensions to chapter 19, writing a ten-minute oral presentation to a lay audience.

C. Revising

1. Critique a poster you have created previously for an applied audience about an application of a multivariate analysis, using the criteria on pp. 447–51 of *Writing about Multivariate Analysis, 2nd Edition*. Consider

   a. The poster itself
   b. Your narrative introduction to the poster
   c. Your narrative modules about the purpose of the project, the data and methods, major findings, and implications for applications of your results
   d. Handouts to accompany the poster
   e. Revise the poster to rectify any problems you identified in parts a through d

2. Critique a report you have written previously for an applied audience about an application of a multivariate analysis, using the criteria on pp. 456–57. Revise it to rectify any problems you identified.

3. Ask a peer who is familiar with the statistical and substantive knowledge level of your intended audience to critique the revised draft of the report you used in the preceding question, using the criteria on pp. 456–57. Revise it to correct any shortcomings they identified.

4. Critique a speech you have written previously for an applied audience about an application of a multivariate analysis, using the criteria in chapters 19 and 20. Revise it to rectify any problems you identified.

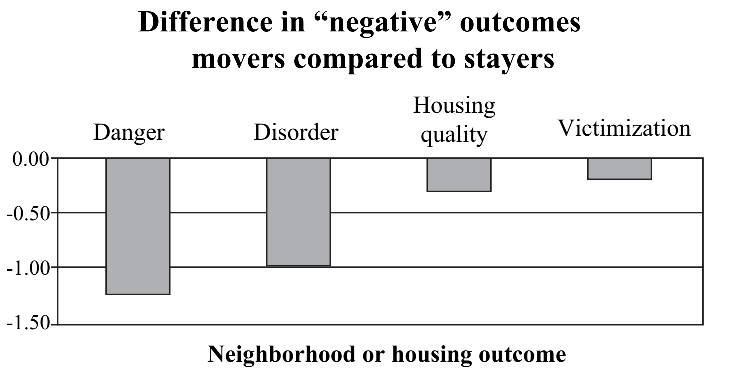
5. Ask a peer who is familiar with the statistical and substantive knowledge level of your intended audience to listen to the revised speech you used in the preceding question. Have them critique it, using the criteria in chapters 19 and 20. Revise it to correct any shortcomings they identified.

20. *Writing for Applied Audiences*

SOLUTIONS

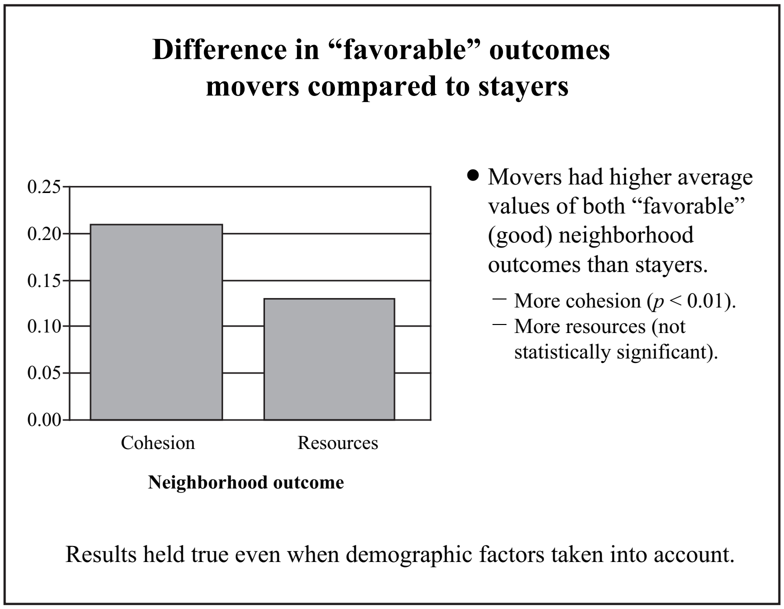
1. Write the specified components of a two-page policy brief.

   a. Title: “Moving to Low-Poverty Areas Improves Outcomes for Families in Public Housing”
   b. Charts of key results



- Movers had lower average values of each of the four “negative” (bad) neighborhood or housing outcomes than stayers (all $p < 0.01$).
  - Less danger, disorder, victimization.
  - Fewer housing problems.
- Results held true even when demographic factors taken into account.

Figure 20A.



Difference in "favorable" outcomes movers compared to stayers

- Movers had higher average values of both "favorable" (good) neighborhood outcomes than stayers.
  - More cohesion ($p < 0.01$).
  - More resources (not statistically significant).

Results held true even when demographic factors taken into account.

Figure 20B.

c. "As shown in figure 20A, low-income families who moved into low-poverty neighborhoods showed appreciably lower levels of danger, victimization, disorder, and housing problems than those who remained in their original, high-poverty neighborhoods, even when demographic characteristics were taken into account. Likewise, the favorable outcomes were better among movers than stayers, with higher levels of cohesion and resources (figure 20B)."

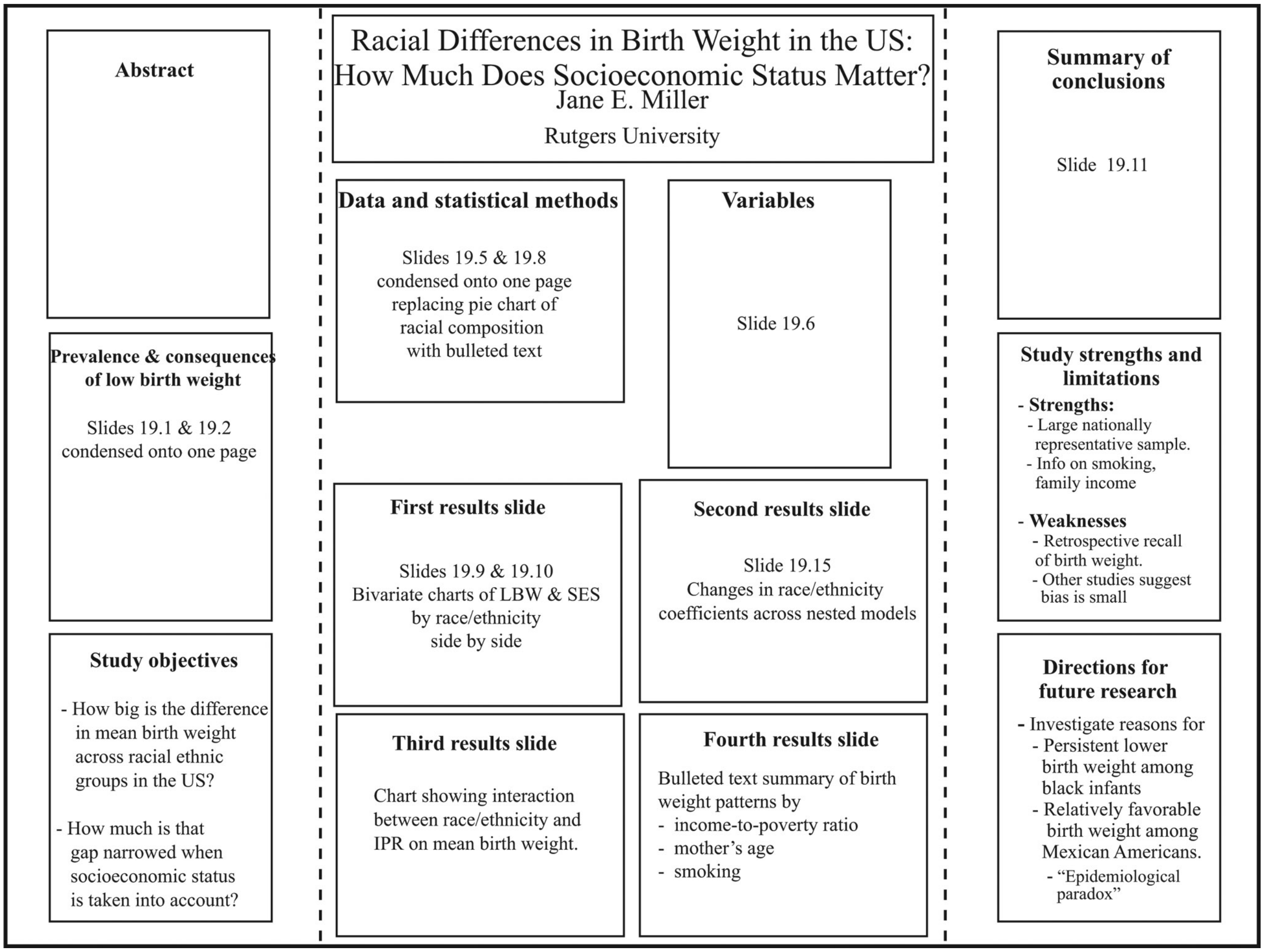
d. "Low-income residents of public housing should advocate for more public housing in low-poverty neighborhoods, and should apply for such benefits when they are available.

"Housing experts are in the best position to organize grassroots efforts to identify locations for public housing in low-poverty areas, and to enroll eligible persons in those programs. They should lobby for additional public housing in low-poverty areas and should disseminate information about available opportunities to low-income families who are eligible for such housing.

"Policy makers are in the best position to develop legislation on these topics and to seek funding to support public housing. They should support legislation to fund and maintain public housing in low-poverty areas."

e. Sidebar: In the Yonkers Residential Mobility Program, low-income residents of public housing were randomly assigned to either move to a low-poverty neighborhood or stay in their current high-poverty neighborhood. The statistical analyses shown here correct for slightly more favorable age, educational attainment, and household composition among movers than stayers.

3. Design of a research poster for the birth weight study. Slide numbers refer to figures in chapter 19 of *Writing about Multivariate Analysis, 2nd Edition*.



Racial Differences in Birth Weight in the US: How Much Does Socioeconomic Status Matter?

Jane E. Miller

Rutgers University

Abstract

Prevalence & consequences of low birth weight

Slides 19.1 & 19.2 condensed onto one page

Study objectives

- How big is the difference in mean birth weight across racial ethnic groups in the US?
- How much is that gap narrowed when socioeconomic status is taken into account?

Data and statistical methods

Slides 19.5 & 19.8 condensed onto one page replacing pie chart of racial composition with bulleted text

Variables

Slide 19.6

First results slide

Slides 19.9 & 19.10 Bivariate charts of LBW & SES by race/ethnicity side by side

Second results slide

Slide 19.15 Changes in race/ethnicity coefficients across nested models

Third results slide

Chart showing interaction between race/ethnicity and IPR on mean birth weight.

Fourth results slide

Bulleted text summary of birth weight patterns by

- income-to-poverty ratio
- mother's age
- smoking

Summary of conclusions

Slide 19.11

Study strengths and limitations

- **Strengths:**
  - Large nationally representative sample.
  - Info on smoking, family income
- **Weaknesses**
  - Retrospective recall of birth weight.
  - Other studies suggest bias is small

Directions for future research

- Investigate reasons for
  - Persistent lower birth weight among black infants
  - Relatively favorable birth weight among Mexican Americans.
    - "Epidemiological paradox"

Figure 20C.

5. Executive summary of the study by Zimmerman (2003)

Background

- Peer effects have been observed in many issues related to higher education.
- Students' attitudes, values, and academic performance may be affected by peers.

Study Objectives

- To measure peer effects on academic performance, taking into account other possible determinants such as demographic attributes.

Data and Methods

- Data are from 3,151 students from the Williams College classes of 1990 through 2001.
- Information was collected on student's own math and verbal SAT scores, roommate's math and verbal SAT scores, student's grade point averages (GPA), and roommate matching preferences for freshman year.
- Multivariate regression was used to estimate association between own and roommate's SAT scores on GPA, taking into account gender, race, class year, and type of major.
- Models were estimated for all students combined, and separately for students with combined SAT scores in the bottom 15%, middle 70%, and top 15% of the class.

Key Findings

- Mean combined (verbal + math) SAT score for the study sample was 1,396 points, with a standard deviation of 123.
- Students' own SAT scores were positively associated with cumulative GPA at all levels of combined SAT scores. Effects were smaller for math (less than one-tenth of a point increase in GPA per 100-point rise in math SAT) than verbal scores (one-tenth to two-tenths of a point increase in GPA per 100-point rise in verbal SAT).
- Roommate's SAT scores were associated with student's GPA, but the effect was statistically significant only in the middle 70% of the SAT range.
- Roommate's verbal SAT had a modest positive effect on student's GPA—equivalent to a rise of four-hundredths (0.04) of a grade point per 100-point increase in roommate's verbal SAT.
- In contrast, roommate's math SAT had a small negative effect on student's GPA—equivalent to a drop of two-hundredths (–0.02) of a grade point per 100-point increase in roommate's math SAT.

Conclusions

- Peer effects on grade point average appear to be minimal, at least in the context of an elite, four-year private college.