Study Questions For Exam 2

1. Your dependent variable is "deaths by famine" in developing countries. You are surprised to learn that this variable has a -0.40 correlation with how totalitarian each country's government is. (That is, more totalitarian countries have LESS deaths by famine.) After reexamining your data, you discover that most developing countries with totalitarian governments have fertile, habitable land. Thus before accepting your initial result, you adjust your independent variable for the proportion of each country's area that is uninhabitable desert. (Be sure that you have read this paragraph VERY carefully.)

   a. Describe in words and/or set up a formula that explains how you would make this adjustment.

   b. If you found the correlation between adjusted totalitarianism and your dependent variable to be 0.30, explain in words (i.e., substantively) how the three variables (deaths by famine, totalitarianism, and the proportion of uninhabitable desert) are associated. (You will not be able to calculate their exact correlations with each other.)

2. Consider the following three variables:

   Y = Respondent's agreement or disagreement with the statement, "If I were to die today, I am sure I would go to heaven." Scores on this variable are 1 = strongly agree, 2 = agree, 3 = undecided, 4 = disagree, 5 = strongly disagree. (Assume equal intervals between consecutive levels of this variable. That is, assume that this is an interval-level variable.)

   X = Respondent's church attendance in "number of visits to church each year"

   Z = Respondent's religion. Scores on this variable are 1 = Catholic, 2 = other

   You have found the following relation between church attendance (as measured by X) and the belief in the existence of heaven (as measured by Y):

   The correlation between X and Y is r = 0.43 and the partial correlation between X and Y controlling for Z equals zero.

   a. Express in words (i.e., theoretically/substantively) how these three variables might be related such that this is possible.

   b. Illustrate in a scatter plot the description you have just given in part a. This can be done by showing how Catholics and non-Catholics would be distributed according to X and Y. On the axes of this scatter plot, designate a Catholic with a star (*) and a non-Catholic with a circle (o).
3. The warden of a prison believes that his rehabilitation program does not work, since there is a zero-order correlation of $r = .21$ between how well a prisoner does in the program and the number of crimes committed after a prisoner leaves prison. However, you discover that the first-order partial correlation between the two variables equals $-.21$ when you adjust for the effects of race (Black versus white). Explain how this is possible in the following ways:

a. First of all show how it is mathematically possible to have a zero-order correlation of $0.21$ and a first-order partial correlation as small as $-0.21$. You can do this by choosing two zero-order correlations (namely, between race and subsequent crime and between race and performance in program) such that the partial correlation is smaller than $-0.21$.

b. Next, make a rough sketch of how the relation between (1) prisoners’ performance in the program and (2) their later criminal record may differ according to (3) race. (That is, draw a graph [or scatter plot] that roughly indicates how the three variables may be related.) Then explain in words a possible (that is, a theoretical) reason why Black and white prisoners may react differently to the rehabilitation program. Be sure to use an explanation that is consistent with your sketch.

4. You are doing research on political culture in the United States. Previous research has found evidence of the following two movements in the U.S.: (1) The student movement in the United States that occurred during the late 1960s was characterized by liberal political values. (2) A growth among fundamentalist religions during the early 1970s has led to an increase in the number of people with conservative political views. You are considering three variables measured on a large national probability sample:

- $Y =$ a measure of political conservatism vs. liberalism. (Assume that high scores are conservative, low scores are liberal, and that the measure is an interval-level measure.)

- $X =$ ABS(year of birth - 1946) i.e., the absolute value of the subject’s year of birth minus 1946. (You have chosen this measure because previous research has found that people born around 1946 were exposed to the student movement during their formative years and thus were most likely to have had their attitudes affected by the movement. The measure ranges from 0 = most formative exposure to the 1960s to 51 = least formative exposure to the 1960s student movement.)

- $Z =$ a measure of religious fundamentalism vs. nonfundamentalism. (Assume that high scores are fundamentalist, low scores are nonfundamentalist, and that the measure is an interval-level measure.)

Correlations among these variables are as follows:
\[ \begin{array}{ccc}
X & Y & Z \\
X & 1.0 & .10 & -.54 \\
Y & .10 & 1.0 & .62 \\
Z & -.54 & .62 & 1.0 \\
\end{array} \]

a. The correlation between X and Y is not statistically different from zero at the .05 level of significance. Given this and the other data in the correlation matrix, would you conclude that subjects' formative exposure to the 1960s student movement is unrelated to their political attitudes? Support your answer with evidence and with a plausible substantive (i.e., theoretical) argument.

b. Consider for the moment only that amount of "the variance in Y that is unexplained by Z." What proportion of this unexplained variance in Y is explained by X? What statistic does this proportion represent?

c. What is the increment in the proportion of the variance explained in Y when X is entered after Z is already in the equation?

d. Is the increment found in part c significant at the .05 level of significance? (Assume a sample size of \( n = 350 \).)

5. You are analyzing data from a national survey of China. Your objective is to understand what motivates Chinese people to increase their incomes. Your dependent variable, Y, is the respondent's annual income in 1995 U.S. dollars. Your independent variables, W and X, are Likert-type items in which respondents were asked to indicate their agreement or disagreement with two statements. The statements are "Everyone should labor for the welfare of all Chinese people." and "Everyone should labor for their own welfare and the welfare of their families." You argue that agreement with the former statement indicates "patriotism," whereas agreement with the latter statement indicates "individualism." Responses to both statements were on a five-point scale: 1=strongly agree, 2=agree, 3=undecided, 4=disagree, and 5=strongly disagree. (You may assume that W and X are interval-level measures.) In sum, your variables are

\[ W = \text{patriotism: should labor for welfare of Chinese people} \]

\[ X = \text{individualism: should labor for own/family welfare} \]

\[ Y = \text{annual income} \]

The variables' means and standard deviations and the correlations among them are as follows:
<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>4.1</td>
<td>2.8</td>
</tr>
<tr>
<td>X</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Y</td>
<td>514</td>
<td>21.2</td>
</tr>
</tbody>
</table>

\[ r_{XY} = -.64 \quad r_{YW} = -.45 \quad r_{XW} = -.35 \]

a. What proportion of the total variance in annual income does the individualism measure explain in addition to the variance in annual income explained by the patriotism measure?

b. How much more income would you estimate would be earned by Chinese who strongly agree on the individualism measure than by those who merely agree on the measure. (In your response, be sure to take into account the role that patriotism plays in determining Chinese incomes.)

c. What proportion does patriotism explain of that part of the variance in income that is not explained by individualism?

d. Something is counterintuitive about the way that W, X, and Y are related. How can you tell? What theoretical reason(s) might you give for this counterintuitive finding?

e. Illustrate in a scatter plot how the data would be related if the theoretical reason(s) given in part d were correct. (Hints: Make sure your slopes are of the right signs. In constructing the scatter plot you may wish to assume that either the patriotism or individualism measure takes only two values.)

f. Any partial slope can be expressed as the slope from a bivariate regression in which the dependent variable is regressed on an independent variable adjusted for all other independent variables from the regression in which the partial slope was estimated. Imagine that you are analyzing your data using SPSS. Your variable names are W, X, and Y as above. What SPSS command(s) would you use to obtain a new variable (call it WADJUST) that measures patriotism adjusted for individualism? (Show how any numbers [i.e., nonvariables] were derived.)

6. You wish to determine whether exposure to agent orange (a chemical defoliant) has caused cancer among U.S. soldiers who served in Vietnam. You have a sample of 200 Vietnam veterans. Your variables are as follows:

\[ Y = \text{number of hospitalizations for cancer treatment over the past five years} \]
X = amount of exposure to agent orange in grams of the chemical

Z = 1 (if either of respondent's parents have had cancer) or 0 (if neither of respondent's parents have had cancer)

Correlations among these variables are as follows:

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1.0</td>
<td>.48</td>
<td>.60</td>
</tr>
<tr>
<td>Y</td>
<td>.48</td>
<td>1.0</td>
<td>.80</td>
</tr>
<tr>
<td>Z</td>
<td>.60</td>
<td>.80</td>
<td>1.0</td>
</tr>
</tbody>
</table>

a. The way your dependent variable is defined may result in a particular pattern of heteroscedasticity. What pattern is this? What variance stabilizing transformation would be appropriate to correct for this pattern? Explain the reason for your choice. Give the SPSS command that you would use to make this transformation.

b. Do your data suggest that agent orange causes cancer? Explain in words what role parents' cancer most likely plays in the relation between these two phenomena. Use the above data to support your argument.

7. The "contagion theory of deviance" states that deviant behavior (such as drug use) is likely to be found in deviance-prone communities (such as communities with high crime rates). The rival "structural theory of deviance" argues that all deviance (including both drug use and other crimes) has a common cause in overpopulation. I.e., the more densely populated an area, the greater the likelihood that all kinds of deviance will occur.

You have data on fifty students each from 100 high schools throughout Iowa. (I.e., your data are about 5,000 high school students.) Each student was asked, "Have you ever smoked marijuana or used any other illicit drug?" You also have data on the current population density and crime rate of the community within which each high school is located. The following are your variables:

Y = percent of a high school's students who responded that they had used drugs

X = current crime rate of a high school's community in crimes per 1000 citizens

Z = current population density of a high school's community in citizens per square mile
Correlations among these variables are as follows:

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1.0</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>Y</td>
<td>.30</td>
<td>1.0</td>
<td>.60</td>
</tr>
<tr>
<td>Z</td>
<td>.50</td>
<td>.60</td>
<td>1.0</td>
</tr>
</tbody>
</table>

a. The way your dependent variable is defined may result in a particular pattern of heteroscedasticity. Draw a diagnostic plot (residual by predicted values from the regression of Y on X and Z) that depicts this pattern. (Be sure to [i] indicate scale values on the plot by placing appropriate numbers on both the horizontal and vertical axes, and [ii] disregard any hunches you might have about how close to zero the predicted values are likely to be.)

b. What variance stabilizing transformation would be appropriate to correct for the pattern found in part a? Explain the reason for your choice. Give the SPSS command (or a concise mathematical formula) that you would use to make this transformation.

c. Which theory do your data support? Explain in words what role population density most likely plays in the relation between community crime rate and students' drug use. Use the above data to support your argument.

d. Consider for the moment only that amount of "the variance in drug use (as measured by Y) that is unexplained by crime rate (as measured by X)." What proportion of this unexplained variance in drug use is explained by population density (as measured by Z)? What statistic does this proportion represent?

e. What is the increment in "the proportion of the variance explained in drug use (i.e., Y)" when population density (i.e., Z) is entered into the regression equation after crime rate (i.e., X) is already in the equation?

f. Is the increment found in part e significant at the .05 level of significance?

8. In the U.S. people tend to become more satisfied with their lives as they reach ages in their late 80s and older. Some gerontologists argue that this increase in life satisfaction results as old people learn to accept "what they cannot do" (i.e., their physical limitations resulting from declining health, etc.). Other gerontologists argue that life satisfaction results only when old people are actively involved with other people. You wish to test which theory (i.e., "acceptance theory" or "involvement theory") is correct.
You have data generated during face-to-face interviews with 63 U.S. centenarian (i.e., over 100 year-old) nursing home residents. Three of your variables are as follows:

Y = "life satisfaction" measured on a 100-point scale from 0 = no satisfaction with life to 100 = total satisfaction with life

X = "acceptance" measured as the number of hours each week that a respondent spends sitting in a rocking chair and sighing

W = "involvement" measured as the number of hours each week that a respondent spends interacting with other nursing home residents, personnel, or visitors

Means, standard deviations, and correlations on these variables are . . .

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>57</td>
<td>19</td>
<td>Y 1.0</td>
</tr>
<tr>
<td>X</td>
<td>22</td>
<td>40</td>
<td>X 0.6</td>
</tr>
<tr>
<td>W</td>
<td>30</td>
<td>24</td>
<td>W -0.7</td>
</tr>
</tbody>
</table>

a. Consider for the moment only that amount of "the variance in life satisfaction that is not explained by acceptance." What proportion of this unexplained variance is explained by involvement?

b. What is the increment in the proportion of the variance in life satisfaction explained from a regression model in which "involvement is the only independent variable" to one in which "both acceptance and involvement are independent?"

c. Is the increment computed in part b significant at the .05 level?

d. Based on the zero-order correlations among the three variables, illustrate in a scatter plot approximately how the data are related. (Be sure your slopes are of the right signs. In constructing the scatter plot you may wish to assume that either the acceptance or involvement measure takes only two values.)

e. Do the data provide support for either theory? Provide evidence consistent with your conclusion regarding each theory. (Use the .05 significance level throughout.)

f. If one assumes the above correlations and standard deviations to reflect the true relations among life satisfaction, acceptance, and involvement for the population of all U.S. centenarian nursing home residents, one's regression model would be misspecified if it only included the involvement measure as an independent variable. What would be the bias in an unstandardized slope estimated using this misspecified model? (Hints: You are being asked to calculate a number here. Please give the units associated with this number, and show how the number was
calculated. Finally, you should also assume that no additional important variables are excluded from the model.)

9. This problem tests whether you are able to recognize when one of the assumptions of multiple regression analysis is violated: One assumption underlying multiple regression analysis is that values on one’s independent variables are not correlated with error values on one’s dependent variable. In matrix and expectation notations this is expressed as follows:

\[
E(X^T \varepsilon) = E(X^T) \cdot E(\varepsilon)
\]

a. Give an illustration of a research situation in which this assumption is violated. (Hint: In your illustration you may refer to variables used in either of the previous problems, or you may make up other variables yourself.)

b. Often researchers are unaware when the \(E(X^T \varepsilon) = E(X^T) \cdot E(\varepsilon)\) assumption is violated. Assuming that you are aware of the violation described in part a, what action might you take to ensure that the assumption is met? Please answer this question in terms of the illustration in part a.

10. Health researchers commonly take patients’ diet into consideration when they estimate how often patients should obtain medical exams. Yet these researchers have different theories about how a patient’s diet influences the frequency with which she obtains medical exams. On the one hand, "health consciousness theory” states that a patient’s healthy diet reflects her general concern with health, and that this concern is likely to motivate her to obtain frequent medical exams. On the other hand, proponents of "self-assurance theory" argue that patients will not obtain medical exams when they feel healthy (presumably because they believe, "Only sick people need doctors.") making those who are healthy by virtue of their good diets less likely to obtain medical exams than those with less healthy diets. Using the medical records of three Baltimore hospitals you obtain a stratified random sample of their adult women patients. Fifty women were randomly sampled from among patients with breast cancer, and fifty women were randomly sampled from among patients without breast cancer. The following are variable information and data on three variables for each of these 100 women:

\[T = \text{the time (in number of days) between medical exams (Note: For patients without breast cancer, "time" is between their two most recent exams. For patients with breast cancer, time is between the exam during which the patient's breast cancer was diagnosed and the patient's next previous medical exam.)}\]
D = the patient's diet (measured on a 100-point scale from 1 = junk food only to 100 = an all fiber/no fat diet)

S = the patient's stratum (0 = no breast cancer; 1 = breast cancer)

<table>
<thead>
<tr>
<th>Variable</th>
<th>T</th>
<th>D</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-0.2</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0.8</td>
<td>-0.7</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(Be sure that you correctly read the correlation matrix. For example, the correlation between T and D is \( r_{TD} = -0.2 \).)

a. The dichotomous stratifying variable (whether or not the patient has breast cancer) is highly collinear with the variable that indicates how healthy the patient's diet is. Is this collinearity because women with breast cancer have more (or because they have less) healthy diets than women who do not have breast cancer?

b. Sketch a plot (in the space provided) that depicts how the three variables must be interrelated. In the plot use the plus symbol (+) to designate women diagnosed as having breast cancer, and use the minus symbol (-) to designate women without breast cancer.

```
long ─┬─
│   │
│   │
│   │
│   │
Time between medical exams
short ─┴─────────────────────────────────────
unhealthy                     healthy
Diet
```

c. Taking into account the collinearity mentioned in part a, do the data provide evidence consistent with "health consciousness theory" and/or with "self-assurance theory"? Justify your answer in words, basing it entirely on the data pattern depicted in the sketch drawn in part b. (I.e., do NOT perform a significance test.)
11. Proponents of "product loyalty theory" argue that residents of rural communities do not innovate. Unlike in urban areas life in these communities does not change much, and rural folks like it that way. They enjoy their routines, their traditions, and the products (e.g., leaded gasoline) they have loyally consumed since they were teenagers. This is disputed by proponents of "resonance theory," who argue that people in rural areas will buy products that "fit in" (or resonate) with their values. For example, in comparison to urban people's values, the values of rural people are more likely to include a "reverence for nature" (e.g., a desire to preserve the natural environment from airborne pollutants such as lead from the combustion of leaded gasoline). Your research is intended to determine which theory is more useful in accounting for community residents' use (or nonuse) of ethanol (a less-polluting alternative to traditional leaded gasoline). In this research your unit of analysis is "the community." You have data on three variables for each of 99 randomly sampled U.S. communities:

E = the proportion of the community's residents, who purchase ethanol on a regular basis

D = the community's population density in thousands of residents per square mile (a measure of how rural or urban the community is)

P = whether the price of ethanol is greater or less than the price of leaded gasoline (0 = ethanol is less expensive than gasoline; 1 = ethanol is more expensive than gasoline)

a. What is the range of possible values on the dependent variable? Given that the dependent variable is defined in this way, what pattern of heteroscedasticity is most likely to be found in the data? Sketch this pattern below (in the space provided). In this sketch be sure to specify what each axis measures and to give approximate values at the "tick points" (namely, at the + and ─ points) along each axis.

Range of values:

Pattern of heteroscedasticity:
b. What variance stabilizing transformation would be appropriate to correct for the pattern of heteroscedasticity depicted in the sketch drawn in part a? Answer this question by giving a formula for T, the transformed dependent variable.

c. After making the transformation indicated in part b, you find that the zero-order correlation between the transformed dependent variable (i.e., "T") and the population density variable (i.e., "D") equals \( r_{TD} = -0.6 \), whereas the partial correlation, \( r_{TD, P} = -0.2 \). Do these findings provide empirical support (at the .05 level of significance) for "product loyalty theory" and/or for "resonance theory"? Explain how you arrived at your answer, basing this explanation on the two correlations just provided. (Hint: You may assume that ethanol is less expensive in rural communities than in urban ones.)

12. Prior to changing its name from "Home Economics" to "Family and Consumer Sciences," ISU’s College of Home Economics had two promotional videos made regarding their academic and research programs. These videos are identical in every way except in that in the numerous places where the one video refers to "Home Economics" (HE), the other video refers instead to "Family and Consumer Sciences" (FCS). The first showing of these videos is to be as part of an experiment that you are conducting. You begin collecting data by administering a two-question pretest to all students taking Psych 101 (ISU’s introductory psychology class) this semester. The pretest asks (1) whether the student is aware of any colleges at ISU that have recently changed their names, and, if the answer to the first question is “yes,” (2) which colleges these are? You randomly assign to a viewing of each video 25 students who indicate an awareness of the HE to FCS name change and 25 students who indicate no awareness of the change. That is, your total sample size is 100 (25 see the HE video and are aware of the change, 25 see the HE video and are unaware of the change, 25 see the FCS video and are aware of the change, 25 see the HE video and are unaware of the change).

After seeing one of the two videos, each student is asked to fill out a questionnaire. Responses to the questionnaire yield values on two measures—a measure of how much the student likes the college, and a measure of how much the student knows about the college. Thus the experiment yields four measures:

\[
\begin{align*}
V &= \text{Video seen (taking either a value of 1=HE or -1=FCS)} \\
A &= \text{Awareness of the HE to FSC name change (taking either a value of 1=aware or -1=unaware)} \\
L &= \text{Liking of the college (ranging from zero [no liking] to 100 points)} \\
K &= \text{Knowledge about the college (ranging from zero [no knowledge] to 100 points)}
\end{align*}
\]
The following table lists these variables’ means and standard deviations plus their intercorrelations:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Deviation</th>
<th>Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.0</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>1.0</td>
<td>-0.71 0.15 1.00</td>
</tr>
<tr>
<td>L</td>
<td>35</td>
<td>15</td>
<td>-0.05 -0.84 -0.01 1.00</td>
</tr>
<tr>
<td>K</td>
<td>45</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

a. Missing from the above table is the correlation between video seen (V) and awareness of name change (A). Calculate this correlation using information about the design of your experiment given in the first paragraph of this problem. (Hints: Be sure to show how you obtained your answer. It may help to begin by putting your data on V and A into a table.)

b. In the space below sketch the relations among Liking (L), Knowledge (K), and Awareness (A). When doing this let each of the sketch’s points represent data on 5 students (i.e., only sketch a total of 20 points). In the plot depict as a plus (+) points for students aware of the college’s name change and depict as a circle (o) points for students unaware of the change.

c. Give a theoretical explanation why L, K, and A might be interrelated as they are in the sketch in part b.

d. Compute the unstandardized partial slope between Liking (L) and Awareness (A) from the regression of Liking on Awareness and Video seen (V). Is there statistically significant evidence (at $\alpha = .05$) that after watching one of the videos, students aware of the college’s name change like the college more than students unaware of the name change?

e. Assuming the means, standard deviations, and correlation coefficients at the bottom of the first page to be parameters of the population you are studying, what is the bias of the partial slope estimated in part d that results from omitting Video
seen (V) from the regression of L on A and V? (Hint: Again be sure to show how you obtained your answer.)

13. Create an interview question that would yield data distributed as a Poisson random variable when asked as part of a survey. What in the wording of this question makes it likely that the data will be distributed this way? If this variable were the dependent variable in a regression analysis, what pattern of heteroscedasticity would you be likely to find? How would you transform the variable to reduce this heteroscedasticity? (Hint: Be sure to answer all 4 parts of this question.)

14. Science teachers at Ames High School wish to ensure that students come to understand science as a process of inquiry. In prior years, biology was taught without involving students in experiments for testing how well plants thrive under various conditions (e.g., different amounts of water, sunlight, fertilizer, etc.). This year two alternative approaches to teaching biology are being evaluated in the school. The first alternative approach has students perform experiments on plants by following procedures set out by their teachers. The second involves students in developing procedures for performing experiments on plants. To evaluate these approaches, students in their sophomore (i.e., second) year of high school were randomly assigned to one of three classes: a class in which the old teaching method was used (i.e., the class with no student involvement), a class in which students followed teacher’s procedures (i.e., the class with assigned procedures), and a class in which students were involved in developing experimental procedures (i.e., the class with self-developed procedures). At the end of the school year, all students in the classes were administered a questionnaire in which their understanding of science as a process of inquiry was measured on a 100-point scale. Your total sample size is 100 students, who are distributed as follows by class and gender:

Table 1: Female versus Male Ames High School Sophomores Enrolled in One of 3 Biology Classes.

<table>
<thead>
<tr>
<th>Gender</th>
<th>no involvement</th>
<th>assigned procedures</th>
<th>self-developed procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>10</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Male</td>
<td>15</td>
<td>21</td>
<td>24</td>
</tr>
</tbody>
</table>

a. Because the only rooms available for this research were of different sizes, you will note that the three classes have different numbers of students in them. Nonetheless, students were assigned to classes in a way that preserved the same ratio of females to males (namely, 2 females for every 3 males) in each of the 3 classes. Given this method of assigning students to classes, how much of the variance in gender is explained by which class a student was assigned to?
b. After performing a preliminary analysis of data on students’ understanding of science as a process of inquiry, a colleague concludes, “Overall, students’ average understanding score was 59 points. Students in the class with assigned procedures had understanding scores 2 points less than this, whereas students in the class with self-developed procedures had understanding scores 5 points more than the overall average.” Assume that the three numbers in the colleague’s sentences are regression coefficients. What were the colleague’s independent variables in the OLS regression model that she used to obtain these coefficients? (Hint: Be sure that your answer spells out the value [i.e., a specific number] that the colleague assigned to each category of student on each independent variable.

c. Given the colleague’s findings in part b, how much below or above the overall average was the average understanding score among students in the “no involvement class” (i.e., the class in which the old teaching method was used)? Be sure to show how you obtained your answer.

d. In a separate analysis you note that among the 100 students, the average understanding score among female students is 20 points less than among male students. Would you be justified in informing your colleague that her findings (i.e., those reported in part b) are biased, because they do not take this gender difference into account? Explain your answer. (Hint: Only use data provided in Table 1 when answering this question.)

It turns out that the differences in understanding scores among classes (i.e., the differences reported in part b) are NOT statistically significant. As a result, you decide to ignore your original research objective (i.e., you decide to no longer consider the variable that differentiates students according to which biology class they attended), and to explore other variables obtained from the questionnaire that might also influence the students’ understanding scores. Below is a table that lists correlation coefficients among the following five variables (including understanding scores and gender) obtained from your questionnaire:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>The student’s understanding score on a 100-point scale on which 0=no understanding and 100=complete understanding</td>
</tr>
<tr>
<td>G</td>
<td>A dummy variable for gender for which 1=female and 0=male</td>
</tr>
<tr>
<td>I</td>
<td>The student’s intelligence on a 100-point scale from 0=“the intelligence of a water opossum” to 100=“an Einstein clone”</td>
</tr>
<tr>
<td>H</td>
<td>The average number of hours each week that the student spends watching television</td>
</tr>
<tr>
<td>M</td>
<td>The student’s motivation to learn biology on a 100-point scale from 0=“no motivation to learn biology” to 100=“obsessed with learning biology”</td>
</tr>
</tbody>
</table>
Table 2: Correlations among Understanding, Gender, Intelligence, Hours, and Motivation for Ames High School Sophomores Enrolled in One of 3 Biology Classes.

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>G</th>
<th>I</th>
<th>H</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>1.00</td>
<td>-.80</td>
<td>-.40</td>
<td>.00</td>
<td>.40</td>
</tr>
<tr>
<td>G</td>
<td>-.80</td>
<td>1.00</td>
<td>.75</td>
<td>-.55</td>
<td>-.50</td>
</tr>
<tr>
<td>I</td>
<td>-.40</td>
<td>.75</td>
<td>1.00</td>
<td>-.85</td>
<td>-.95</td>
</tr>
<tr>
<td>H</td>
<td>.00</td>
<td>-.55</td>
<td>-.85</td>
<td>1.00</td>
<td>.90</td>
</tr>
<tr>
<td>M</td>
<td>.40</td>
<td>-.50</td>
<td>-.95</td>
<td>.90</td>
<td>1.00</td>
</tr>
</tbody>
</table>

In each of the following questions you are asked to consider the influences that intelligence (I), hours watching television (H), and motivation (M) respectively have on students’ understanding of science as a process of inquiry (U), after adjusting for gender-related (G) differences in the values taken by each of these variables (i.e., I, H, and M).

e. In the space below sketch the relations among Understanding (U), Intelligence (I), and Gender (G) in a way that reflects the correlations among these variables given in Table 2. When doing this let each of the sketch’s points represent data on 5 students (i.e., only sketch a total of 20 points). In the plot depict as a plus (+) points for female students and depict as a circle (o) points for male students. (Hint: Be sure to plot U on the vertical axis and I on the horizontal one.)

f. In the space below sketch the relations among Understanding (U), Hours watching television (H), and Gender (G) in a way that reflects the correlations among these variables given in Table 2. When doing this let each of the sketch’s points represent data on 5 students (i.e., only sketch a total of 20 points). In the plot depict as a plus (+) points for female students and depict as a circle (o) points for male students. (Hint: Be sure to plot U on the vertical axis and H on the horizontal one.)
g. In the space below sketch the relations among Understanding (U), Motivation (M), and Gender (G) in a way that reflects the correlations among these variables given in Table 2. When doing this let each of the sketch’s points represent data on 5 students (i.e., only sketch a total of 20 points). In the plot depict as a plus (+) points for female students and depict as a circle (o) points for male students. (Hint: Be sure to plot U on the vertical axis and M on the horizontal one.)

h. Give a theoretical explanation why U, M, and G might be interrelated as they are in the sketch in part g. (Hint: In your response be sure to explain whether or not your data provide evidence that motivated students are more likely than unmotivated ones to understand science as a process of inquiry.)
S = “On a scale from zero (for no success) to ten (total success), how successful do you believe your movement has been in attaining its environmental goals?”

M = “How many years and months has your movement been working on its environmental goals?” (coded as “number of months”)

N = “How many churches or other nonprofit organizations have cooperated with you in attaining your movement's environmental goals?”

Your data on these variables are as follows:

<table>
<thead>
<tr>
<th>Correlation Coefficients</th>
<th>Success (S)</th>
<th>Months (M)</th>
<th>NPOs (N)</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success (S)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.7</td>
<td>7.2</td>
<td>1.6</td>
</tr>
<tr>
<td>Months (M)</td>
<td>0.5</td>
<td>1.0</td>
<td>0.6</td>
<td>101.3</td>
<td>32.0</td>
</tr>
<tr>
<td>NPOs (N)</td>
<td>0.7</td>
<td>0.6</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
</tr>
</tbody>
</table>

a. Find the unstandardized regression slope for the bivariate regression in which S is regressed on N.

b. Give a theoretical reason why a specification error may have occurred by not having included M as an independent variable in the regression model estimated in part a. (Hint: No calculations are required here. You are being asked to give a nonmathematical reason for why one of the assumptions of regression may have been violated.)

c. Estimate the bias in the slope obtained in part a that resulted from excluding M from the regression model.

d. In calculating the bias in part c, you needed to compute the partial slope associated with M from the regression in which S is regressed on both N and M. Does this slope illustrate a case of explanation, interpretation, suppression, or distortion? Explain your answer.

e. Imagine that no social movement had more than 2 NPOs cooperating with it. More specifically, imagine that 30 leaders indicated zero NPOs (i.e., N=0), 30 indicated one NPO (i.e., N=1), and 30 indicated two NPOs (i.e., N=2). In the space provided below, draw a sketch of the relation between S and M that is consistent with the data in the table at the beginning of this problem. In particular, depict this relation by placing six 0s, six 1s, and six 2s between the vertical and horizontal axes below (i.e., such that each number represents the responses of five leaders).
f. Given the pattern sketched in part e (and thus the data provided in the table at the beginning of this problem), does Months (M) appear to have a causal effect on Success (S)? Explain your answer.

When you check your sketch in part e against an actual plot of your data, you find that (ignoring the number of NPOs [i.e., N] for the moment) the plot looks as follows:
In words: When a social movement has not been working very long on its environmental goals, it is unlikely that it will have had enough time to be successful in attaining these goals.

g. What variable transformation might you use to correct for the pattern of heteroscedasticity depicted in this plot of your data? In your answer please specify the regression model that you would use in estimating the effect of M on S (i.e., the effect of months working on environmental goals on a social movement’s success in attaining these goals). (Hint: Do NOT include N in this regression model.)

h. Your research assistant performs an analysis of the relation between S and N, and describes the three parameters estimated in her model as follows: “On average the movement leaders gave a score of 7.2 points on how successful they believed their movement had been in attaining its environmental goals. Among the leaders who indicated that their movement had cooperated with two (2) “churches or other nonprofit organizations” in attaining the movement’s environmental goals, the average success score was 1.3 points higher than this overall average. Among the leaders who indicated that their movement had cooperated with only one (1) such organization, the average success score was 0.4 points higher than the overall average.” Assuming that your research assistant computed and interpreted her findings correctly, write down her regression model (including each of the three numbers underlined in this paragraph) and give a clear explanation of the independent variables used in this model. (Hint: Be sure to specify the conditions under which each value of each independent variable is taken.)

16. Did you know that most organic food you buy in a grocery store is produced on huge factory farms owned by large corporations (e.g., Dole, Kraft, General Mills, Unilever, Coca Cola) that are more interested in earning profits than producing safe, sustainably grown food? Over time the USDA’s Organic label has become less and less meaningful—a situation that Certified Naturally Grown (CNG), Inc., was established to remedy. CNG lists its member-farmers online, and conducts random checks of these members’ exclusive use of allowable materials and practices in their farming operations. The objective is to encourage consumers to buy their meat and produce from small local farmers, whose market niche is one of quality (rather than cheap) food.

As a long-time member of CNG’s evaluation team, your specialty involves testing the quality of farmers’ irrigation water. If a farmer uses water from a river or a well, you test it for contaminants (e.g., nitrogen). If contaminant levels have been increasing, you usually counsel farmers on alternative technologies for harvesting rainwater. If a farmer
does not change her or his water source before contaminant levels reach intolerable limits, she or he will lose CNG membership (and may no longer sell food under the CNG label).

During the past decade you have counseled thousands of US farmers on the advantages of rainwater-harvesting (RH) technologies like cisterns and catch-basins. Given your expertise in RH technologies, you decide to investigate the extent to which farmers have been adopting these technologies. Your unit of analysis is the “US farm census tract” (i.e., a statistical subdivision established by the US Census Bureau, and reported in the 2000 census as having mostly “rural farm” housing units). To investigate the tracts where RH technologies have been adopted, you consider two measures of adoption:

\[ Y_1 = \text{The increase from 1990 to 2000 in a tract’s number of farms using RH technologies} \]  
\[ \text{(For example, } Y_1 = 10 \text{ means that this tract had 10 more farmers using RH technologies in 2000 than in 1990.)} \]

\[ Y_2 = \text{The increase from 1990 to 2000 in a tract’s proportion of farms using RH technologies} \]  
\[ \text{(For example, } Y_2 = .10 \text{ means that this tract had 10% more farmers using RH technologies in 2000 than in 1990.)} \]

a. No matter which of these two measures you decide to use as your dependent variable, your choice will leave you with heteroscedastic data. In each case (i.e., for \( Y_1 \) and \( Y_2 \) separately), sketch a diagnostic plot that shows the pattern of heteroscedasticity you would expect each to have, and (b) explain what variance-stabilizing transformation you would use to correct for this heteroscedasticity. (Hint: Be sure to label your axes!)

Workspace for part a:

Diagnostic plot for \( Y_1 \):

[Diagram not provided]
Transformation of $Y_1$:

Diagnostic plot for $Y_2$:

Transformation of $Y_2$:

**More data**: You decide to use the transformed values of $Y_1$ as your dependent variable. (From now on we shall refer to this transformed variable as “$Y$”.) You use the following two independent variables to estimate variations in $Y$:

$W = \text{Government subsidies available between 1980 and 2000 for farmers’ adoption of RH technologies (taking either a value of 1 = subsidies were available or 0 = subsidies were not available)}$

$X = \text{A tract’s total number of farms using RH technologies in 1990 (For example, X=100 means that this tract had 100 farmers using RH technologies in 1990.)}$

Some hypotheses:

*Incentive*: Regarding the variable, $W$, your thinking is that more adoption is likely in tracts in which government subsidies are available (and thus are incentives) for it.
Approval: Regarding the variable, X, your thinking is that adoption of a technology is more likely in tracts in which others have already adopted (and thus approve use of) the technology.

Means and standard deviations on these variables are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adoption (Y)</td>
<td>4.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Incentive (W)</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Approval (X)</td>
<td>150</td>
<td>50</td>
</tr>
</tbody>
</table>

b. Given these data plus the fact that the correlation between Y and W (i.e., \( r_{YW} \)) is .75, find values of the constant and slope from the regression of Y on W. State in words what these two coefficients mean.

For the next three parts (namely, parts c, d, and e) of this problem, imagine that you run two regression models. You first regress Y on X, and then you regress Y on both X and W. Parts of your SPSS output would look something like the following:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>AAAA</td>
</tr>
<tr>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>BBBB</td>
</tr>
<tr>
<td>X</td>
<td>CCCG</td>
</tr>
<tr>
<td>W</td>
<td></td>
</tr>
</tbody>
</table>

c. If the value of AAAA were 0.5 and the value of BBBB were -0.5, display on the below sketch what the relation among Y, X, and W would look like. (Hint: Indicate data points as “1” if W=1, and as “0” if W=0.)
d. If the value of AAAA were 0.5 and the value of BBBB were 0.0, display on the below sketch what the relation among Y, X, and W would look like. (Hint: Indicate data points as “1” if W=1, and as “0” if W=0.)

![Diagram](image1)

e. If the value of AAAA were 0.0 and the value of BBBB were 0.5, display on the below sketch what the relation among Y, X, and W would look like. (Hint: Indicate data points as “1” if W=1, and as “0” if W=0.)

![Diagram](image2)
f. Only one of the above three sketches depicts empirical support for your hypothesis that “adoption of a technology is more likely in tracts in which others have already adopted the technology.” Which sketch is this? State in words the meaning of the number, BBBB, from the output associated with this sketch.

g. After all your hard work, a colleague points out to you that Y is an imperfect measure of “adoption behavior among a tract’s farmers.” The colleague tells you, “Values on the measure may simply reflect increases or decreases in the total number of the tract’s farmers during the decade from 1990 to 2000. This looks to me like a specification problem that you might solve using two-stage least squares (2SLS).” The colleague then suggests that you use the following instrumental variable (I):

I = “the increase or decrease in the number of a tract’s farmers (For example, I=-2 means that this tract had 2 fewer farmers in 2000 than in 1990)”

Explain what your colleague is suggesting. That is, how would one perform the two stages of 2SLS by using the variable, I, as the instrumental variable in an analysis of the effects of W and X on Y? Would 2SLS be appropriate in your analysis of Y? Explain why or why not.