Rutgers-Camden Doctoral Program in Public Affairs/Community Development

Topics, Sample Questions, and Suggested Readings for the Comprehensive Methods Examination

May 15, 2015
Table of Contents

I. Logic of Social Inquiry and Research Design ................................................................. 1
   A. List of Topics .................................................................................................................. 1
   B. Sample Questions ......................................................................................................... 2
   C. Suggested Readings ...................................................................................................... 5

II. Quantitative Methods I & II ............................................................................................ 8
   A. List of topics for Research Methods Exam ................................................................. 8
   B. Sample Questions: ....................................................................................................... 9
      i) Descriptive and Inferential Statistics ....................................................................... 9
      ii) Regression Analysis ............................................................................................... 12
   C. Suggested Readings .................................................................................................... 25
I. Logic of Social Inquiry and Research Design

A. List of Topics

Approaches to Social Science Research
- Positivism
- Post-Positivism
- Critical Theory
- Interpretivism
- Correlation vs. causation
- Logic of causal inference

Measurement Issues
- Types of data
- Validity
- Reliability
- Index and scale construction

Sampling Techniques
- Simple Random sampling
- Stratified Sampling
- Multi-Stage Cluster Sampling

Survey Research
- Questionnaire Construction
- Survey Administration

Research design
- Experiments
- Quasi-Experiments
- Natural Experiments
- Qualitative Methods
  - Ethnography
  - Participant observation
  - Case studies
  - Principles of field work
  - Structured interviews
- Quantitative methods
  - Difference in differences
  - Interrupted time series
  - Regression discontinuity
  - Instrumental variables
  - Propensity score matching
  - Statistical power
B. Sample Questions

1. The publication of Theda Skocpol’s *States and Social Revolutions* initiated a heated debate within the social sciences about not only her theory of the state but also her methodology. For some, her research demonstrated all of the pitfalls of small-N comparative analysis. Describe Skocpol’s research design. What are the weaknesses of her research design? Explain. What are the strengths of her research design? Explain. Does her research design solve the *fundamental problem of causal inference*? Why or why not?

2. Social scientists are often admonished to distinguish facts from values in order to frame testable empirical theories or hypotheses. Public policy analysis depends on social scientific inquiry to formulate responses to socio-economic problems, but is explicitly concerned with values or normative issues (i.e., fairness, efficiency, etc.) Select a specific public policy problem of interest to you (e.g., affordable housing, urban crime, childhood obesity, unemployment) and explain the role that facts and values respectively, would play in an analysis of the problem.

3. Does playing violent video games at home cause boys to become more violent and aggressive in school? This question concerns many parents and educators. Assume you have just received a grant to conduct a study of this question. For a previous study, you collected data on a simple random sample of 200 ten year old boys in the Richardson Independent School District (RISD), including data on the boy’s demographics, their leisure activities including time spent playing violent video games, and disciplinary infractions at school. However, most of the boys in the previous study did not play violent video games at that time and video games were not the focus of the previous study. The boys are now 16, and you suspect many more of them now play these games and play them for longer periods. Your grant will allow you to go back and reinterview the boys as well as collect additional data from the school. Since the boys’ families previously consented to participation, you assume they will continue to participate. RISD has agreed to participate as well.

   a) State your research hypothesis, and the appropriate null hypothesis.
   b) What research design will you employ?
c) What is your dependent variable?  
d) What is your main independent variable?  
e) What additional independent variables do you intend to control if any?  
f) What additional data will you have to collect to implement your design?  
g) What specific statistical tests will you conduct relevant to your research hypothesis?  
h) What is the main threat to internal validity in this design? List only one. Explain.  
i) What is the main threat to external validity in this design? List only one. Explain.  
j) Fifteen boys in the original sample have transferred out of the district. You do not have the funding to track them and the new school districts are unlikely to cooperate. Does this present any problems for your study? Is so, what are they?

4. In *White Logic, White Methods*, sociologists Tukufu Zuberi and Eduardo Bonilla-Silva offer a series of startling claims about the relationship between race and social science research. For example, they write: “Not only are Whites a dominant population, the dominant perspective in sociology has been defined by a view of reality that privileges Whites in the United States and Europe…[S]ociology has been—and still is—a White-led and White-dominated field and, therefore, it should not surprise anyone that the logic of analysis and methods used to investigate racial matters reflect this social fact.” Do you find this argument compelling? Why or why not? Given the logic of statistical methods (e.g. the linear model), under what conditions (if any) does race shape applications of this method? Under what conditions (if any), can this method escape the racial dynamics of the society in which it is applied?

5. In his seminal work, *The Structure of Scientific Revolutions*, Thomas Kuhn characterized the common understand of science, including social science, as “the constellation of facts, theories and methods collected in current texts,” and scientific development as, “the piecemeal process by which these items have been added, singly and in combination, to the ever growing stockpile that constitutes scientific technique and knowledge.” Kuhn, of course, argues that this model of science as an incremental progression is flawed. Explain Kuhn’s critique of this historiographic model. In discussing his alternative view of paradigmatic shifts be sure to offer examples from the social as well as the natural sciences. Select a policy or approach to community development and illustrate how Kuhn’s theory, rather than the incremental progression model, applies to your case.

6. The rational choice approach has assumed a prominent position within the social sciences. While some scholars have welcomed this development, others have
fiercely opposed it. In *Pathologies of Rational Choice*, political scientists Donald Green and Ian Shapiro write: “We contend that much of the fanfare with which the rational choice approach has been heralded in political science must be seen as premature once the question is asked: What has this literature contributed to our understanding of politics? We do not dispute that theoretical models of immense and increasing sophistication have been produced by practitioners of rational choice theory, but in our view the case has yet to be made that these models have advanced our understanding of how politics works in the real world.” Do you find this argument compelling? Why or why not? Outline and discuss the foundational propositions of this theoretical approach. Then, drawing upon the work of two different interpretivist scholars, formulate a critique of the rational choice approach.

7. Cross-national comparisons can be a fruitful method of analysis. They can also present significant analytical, conceptual, and measurement challenges. Select a specific public policy problem of interest to you (for example, affordable housing, urban crime, childhood obesity, or unemployment) and formulate a specific research question related to that problem. Then outline and discuss the benefits and drawbacks of a large-N, cross-national design that attempts to answer that question.

8. Recently, three prominent scholars at Harvard University received intense criticism for approving a dissertation that found that the “average IQ of immigrants in the United States is substantially lower than that of the white native population” and argued that the lower intelligence of immigrants should be considered when drafting immigration policy.
   (a) What are the advantages and disadvantages of using IQ as a measure in social scientific studies?
   (b) How would positivists evaluate this measure?
   (c) How would interpretivists evaluate this measure?

9. Karl Popper and Thomas Kuhn offer two different perspectives on the development of scientific knowledge. Briefly outline the logic of those frameworks. Identify specific issues on which the two frameworks disagree and issues on which they agree, if any.
10. The City of Philadelphia observed that one of the most common causes of auto accidents is drivers who run red lights. Philadelphia has data on traffic accidents over many years, coded by location. In an attempt to reduce accidents, Philadelphia installed red light cameras at some of the most accident-prone intersections. The cameras take pictures of red light violators and a ticket is sent to the registered owner of the vehicle. The City expects a reduction in accidents over time, as more people learn about the cameras by getting tickets or by hearing about the program from others who did. The city has given you their monthly accident data going back to 1990, coded by location, as well as the date of red light camera installations and the number of tickets per month at each location.

(a) State the City’s hypothesis and the appropriate null hypothesis.
(b) What analytic strategy will you use to test the City’s hypothesis?
(c) What is your dependent variable? What is your main independent variable?
(d) What additional independent variables do you intend to control if any? (Assume you have the variables you need in the dataset.)
(e) What is the main threat to internal validity in this design? List only one. Explain.

C. Suggested Readings


Theda Skocpol, “State and Revolution: Old Regimes and Revolutionary Crises in France, Russia, and China,” Theory and Society 7, 1/2, Special Double Issue on State and Revolution (Jan. - Mar., 1979), pp. 7-95


II. Quantitative Methods I & II

A. List of topics for Research Methods Exam

Descriptive Statistics
- Levels of measurement (types of variables)
- Types of data (cross-sectional, longitudinal, cohort, panel)
- Measures of central tendency
- Measures of dispersion
- Correlation

Probability
- Basic rules of probability
- Random Variables
- Probability distributions
- The normal distribution

Inferential Statistics
- Sampling theory
- Sampling distributions and the properties of estimators
- Confidence Intervals and hypothesis tests, including $z$, $t$, $F$, and $\chi^2$ tests as appropriate, for:
  - Means and proportions
  - Difference of two means and difference of two proportions
  - Difference in variances
  - Association between variables
  - Goodness of fit

Linear Regression Analysis
- Regression models (bivariate and multiple)
  - Estimation by Ordinary Least Squares (OLS)
  - Hypothesis testing (single and joint hypotheses)
  - Use of indicator variables (aka dummy variables)
    - Varying intercepts
    - Varying slopes
    - Interactions
- Alternative functional forms
  - Logarithmic functions forms
  - Quadratic and polynomial functional forms
  - Piece-wise linear models
- Specification Errors
  - Omitted variable bias
  - Incorrect functional form
  - Errors in variables
- Deviations from OLS assumptions and corrections


B. Sample Questions:

i) Descriptive and Inferential Statistics

1. Five hundred students apply for a certain scholarship. Each student receives a rating based on their financial aid, references, personal essays, and so on. The ratings are normally distributed with a mean score of 50 and a standard deviation of 15.
   a) How many students received ratings higher than 70?
   b) How many students had ratings between 40 and 60?
   c) One student’s z score was -1.33; what was his rating?

2. A recent New York Times/CBS News survey of 688 persons found that 352 approved of President Obama’s handling of the economy. The sample included 356 women, of whom 196 approved, and 332 men, of whom 156 approved.
   a) What is the 95 percent confidence interval for the proportion of the population that approves of President Obama’s handling of the economy?
   b) What is the 90 percent confidence interval for the difference between men and women in the proportion who favor President Obama’s handling of the economy?


   Scientists are reporting today that they have created the first vaccine that appears able to prevent cervical cancer. The vaccine works by making people immune to a sexually transmitted virus that causes many cases of the disease. The vaccine is experimental and will not be available to the public for several years. A successful vaccine could sharply reduce rates of cervical cancer, which affects 470,000 women a year worldwide and kills 225,000. In the United States, there are 13,000 cases a year and 4,100 deaths. In a study of 2,392 young women, half of them vaccinated and half given placebo shots, the vaccine was 100 percent
Based on the information in the paragraph above, test the hypothesis that the new vaccine is effective in preventing the infections associated with cervical cancer.

4. The Social Security Disability program pays benefits to former workers who are no longer able to work due to a physical or mental disability. The hard part is deciding who is disabled enough to qualify for the program. Many people who apply for benefits are denied, and they have a right to appeal. The appeals are randomly assigned to one of several judges, who reviews the case and makes a decision. Overall, 70 percent of the appeals are denied, so the applicant does not receive any benefits. The table below shows the record on appeals for 2001. Since the cases are randomly assigned and the judges are supposed to use the same standards, there isn’t supposed to be any systematic difference between the judges in terms of their decisions. Test the hypothesis that the outcome of the case is related to the judge to whom the case is assigned.

<table>
<thead>
<tr>
<th>Judge</th>
<th>Outcome of Appeal</th>
<th>Anderson</th>
<th>Bennett</th>
<th>Cramer</th>
<th>Davis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approved</td>
<td>19</td>
<td>13</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Denied</td>
<td>20</td>
<td>43</td>
<td>41</td>
<td>28</td>
</tr>
</tbody>
</table>

5. In the 1970s, experiments were conducted to determine whether an income subsidy would lead to increased rent expenditures. Summary statistics for one of the experimental groups is shown below. “Rent1” is the rent of the households before the subsidy began. “Rent2” is the rent of those households 1 year after the subsidy began. “Diff” is the difference between rent1 and rent2 for each household. There were 362 total households in the study. Test the hypothesis that the amount households spent on rent increased. Can we conclude that the subsidy was the cause of the increase? Explain your answer.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>rent1</td>
<td>362</td>
<td>116.88</td>
<td>3.32</td>
<td>63.22</td>
</tr>
<tr>
<td>rent2</td>
<td>362</td>
<td>126.01</td>
<td>3.76</td>
<td>71.59</td>
</tr>
<tr>
<td>diff</td>
<td>362</td>
<td>9.13</td>
<td>2.05</td>
<td>39.02</td>
</tr>
</tbody>
</table>

6. Two students made two different mistakes in the final exam for descriptive and inferential statistics class. Student A used t-test for one sample mean test for a sample of more than 100; student B used a Z-test for one sample mean test for a sample of less than 10. We do not know the underlying distribution of the variable in question. Which mistake is more serious? Or maybe both are serious? Explain.
7. Why, when we test hypotheses, do we always test them in terms of population parameters, not in terms of the sample statistics? There are at least two serious reasons.

8. Is there something wrong with this statement, or is it correct? Explain.

   “When you increase your sample size, your test statistic is going up in absolute value and your p value is going up as well, hence your results are more statistically significant.”

9. A researcher wants to figure out the effectiveness of a flu vaccine. The vaccine was provided free of charge in a two-shot sequence over a period of two weeks to those wishing to avail themselves of it. Some people received the two-shot sequence, some appeared only for the first shot, and others received neither. Do the data present sufficient evidence to indicate dependence between vaccine classification and the occurrence or nonoccurrence of flu?

<table>
<thead>
<tr>
<th>Flu</th>
<th>No Vaccine</th>
<th>One Shot</th>
<th>Two Shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td>No Flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
<td>954</td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>

10. A city manager is trying to determine if her new personnel procedures are decreasing the time it takes to hire an employee. She takes a sample shown below, and calculates the average time to hire an employee in days before and after implementing the new procedures. Help her decide.

<table>
<thead>
<tr>
<th>Bureau</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>b</td>
<td>49</td>
<td>45</td>
</tr>
<tr>
<td>c</td>
<td>55</td>
<td>12</td>
</tr>
<tr>
<td>d</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>e</td>
<td>44</td>
<td>24</td>
</tr>
</tbody>
</table>

11. During the Senate impeachment hearings of President Bill Clinton in 1999, the *Dallas Morning News* published a letter to the editor from David Read of Grand Prairie that read as follows:

   I’m sick and tired of listening and reading news commentary on some poll that shows that the *American people* are tired of the Senate hearings and that two-thirds of the public want Mr. Clinton to remain in office. STOP SAYING THAT! It’s not true and any intelligent person who knows anything about how phone surveys are taken knows that. The only statement that could be honestly reported would be one which stated that “two-thirds of the people
polled” expressed a certain opinion. Let’s leave the American people out of this poll spin.

Please write a response to Mr. Read. Specifically comment on whether the underlined statement is true or false and explain your reasoning.

12. In *Differences that Matter* (Ithaca NY: Cornell, 2006), Dan Zuberi interviews low-wage hotel workers in Seattle and Vancouver to assess the importance of different social policies in the U.S. and Canada. One of the largest differences, at least at the time of the study, is that Canada had a universal health insurance program in place, whereas the U.S. did not and therefore many low wage workers were uninsured. Of the 39 workers Zuberi interviewed in Seattle, he reports that 67 percent saw a doctor regularly. Of the 38 workers he interviewed in Vancouver, he states that 85 percent saw a doctor regularly.

(a) What is the 95 percent confidence interval for the difference in the proportion that sees a doctor regularly?
(b) Test the hypothesis that low-wage workers in Vancouver are more likely to see a doctor regularly than in Seattle at the 0.05 level of significance.

**ii) Regression Analysis**

1. Does a landfill (garbage dump) affect property values? If so, how much? This question is addressed by Nelson et al. (“Price Effects of Landfills on House Values,” *Land Economics*, 1992) for a sample of over 700 home sales in Ramsey, Minnesota, within 2 miles of an active landfill. They estimated the following regression:

\[
\text{PRICE}_i = \beta_1 + \beta_2 \text{AGE}_i + \beta_3 \text{BDRMS}_i + \beta_4 \text{BTHRMS}_i + \beta_5 \text{DI-694}_i + \beta_6 \text{FF}_i + \beta_7 \text{FIREPL}_i + \beta_8 \text{SMALLLOT}_i + \beta_9 \text{SALEYR}_i + \beta_{10} \text{STYLE20}_i + \beta_{11} \text{STYLE50}_i + \beta_{12} \text{TREES}_i + \beta_{13} \text{DFILL}_i + u_i
\]

The variables are defined as follows:

- **PRICE** is the sales price,
- **AGE** is the age of the house in years,
- **BDRMS** is the number of bedrooms,
- **BTHRMS** is the number of bathrooms,
- **DI-694** is the distance to access Interstate 694,
- **FF** is the square footage of the first-floor of the house,
- **FIREPL** is the number of fireplaces,
- **SMALLLOT** is a dummy for lot size less than one acre,
- **SALEYR** is the last two digits of the year of sale (1980=80, etc.),
- **STYLE20** and **STYLE50** are dummies for 2 and 3 story houses respectively,
- **TREES** is an measure of the tree coverage of the property,
- **DFILL** is the distance to the landfill in miles.

The regression results and descriptive statistics are shown on the following page.

a) What do these results tell you about:
(i) the effect of distance to the landfill on the sales price?
(ii) the effect of the age of the house on the sales price?
(iii) the effect of the presence of a second or third story on the sales price?
(iv) the effect of tree coverage on the sales price?

b) What variable has the greater impact on sales price, distance to the landfill (DFILL) or first-floor square footage (FF)? Justify your answer based on the results.

c) The regression results reported here suggest that the effect of the independent variables on sales price is a fixed dollar amount, regardless of the level of price. What alternative model would you estimate if your hypothesis was that the independent variables caused sales price to vary in percentage terms?

d) Discuss what these results imply about the effect of time on sales price. Is this a reasonable way to model the way that house prices vary with time? If not, what is a better to take account of the affect of time on sales prices?

e) Another researcher suggested dropping the year of the sale altogether, arguing that the distance to the landfill is constant over time so that leaving it out would not cause left-out variable bias. Is this argument correct? Why or why not?
### TABLE 1
**Ramsey, Minnesota, House Sale Analysis: Linear Regression Results with Respect to Landfill Distance**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Standard Coefficient</th>
<th>Tolerance</th>
<th>T</th>
<th>P(1-tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-166194.466</td>
<td>20003.448</td>
<td>0.000</td>
<td>0.7896960</td>
<td>-8.308</td>
<td>0.000</td>
</tr>
<tr>
<td>AGE</td>
<td>-204.745</td>
<td>92.066</td>
<td>-0.061</td>
<td>0.7896960</td>
<td>-2.224</td>
<td>0.013</td>
</tr>
<tr>
<td>BDRMS</td>
<td>-2499.788</td>
<td>707.367</td>
<td>-0.107</td>
<td>0.6594665</td>
<td>-3.534</td>
<td>0.000</td>
</tr>
<tr>
<td>BTHRMS</td>
<td>14239.881</td>
<td>1025.489</td>
<td>0.428</td>
<td>0.6347619</td>
<td>13.886</td>
<td>0.000</td>
</tr>
<tr>
<td>DFILL</td>
<td>4896.297</td>
<td>1160.251</td>
<td>0.115</td>
<td>0.8145561</td>
<td>4.220</td>
<td>0.000</td>
</tr>
<tr>
<td>DI-694</td>
<td>-635.387</td>
<td>325.504</td>
<td>-0.556</td>
<td>0.7429921</td>
<td>-1.952</td>
<td>0.026</td>
</tr>
<tr>
<td>FF</td>
<td>41.314</td>
<td>3.057</td>
<td>0.398</td>
<td>0.6970723</td>
<td>13.514</td>
<td>0.000</td>
</tr>
<tr>
<td>FIREPL</td>
<td>2290.754</td>
<td>909.205</td>
<td>0.074</td>
<td>0.7051258</td>
<td>2.520</td>
<td>0.006</td>
</tr>
<tr>
<td>SMALLLOT</td>
<td>807.899</td>
<td>1066.072</td>
<td>0.021</td>
<td>0.7932588</td>
<td>0.758</td>
<td>0.225</td>
</tr>
<tr>
<td>SALEYR</td>
<td>2121.608</td>
<td>210.664</td>
<td>0.263</td>
<td>0.8823538</td>
<td>10.071</td>
<td>0.000</td>
</tr>
<tr>
<td>STYLE20</td>
<td>4186.224</td>
<td>1028.881</td>
<td>0.113</td>
<td>0.7848099</td>
<td>4.069</td>
<td>0.000</td>
</tr>
<tr>
<td>STYLE50</td>
<td>3620.194</td>
<td>2198.845</td>
<td>0.043</td>
<td>0.8792989</td>
<td>1.646</td>
<td>0.050</td>
</tr>
<tr>
<td>TREES</td>
<td>340.109</td>
<td>473.184</td>
<td>0.020</td>
<td>0.7425526</td>
<td>0.719</td>
<td>0.237</td>
</tr>
</tbody>
</table>

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>DF</th>
<th>Mean-Square</th>
<th>F-Ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.139784E+12</td>
<td>12</td>
<td>0.116487E+11</td>
<td>80.188</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>0.100960E+12</td>
<td>695</td>
<td>0.145266E+09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N: 708
Multiple R: 0.762
Squared Multiple R: 0.581
Adjusted Squared Multiple R: 0.573
Standard Error of Estimate: 12052.639

### TABLE 2
**Ramsey, Minnesota, House Sales Analysis: Descriptive Statistics**
(Total Observations: 708)

<table>
<thead>
<tr>
<th>PRICE</th>
<th>AGE</th>
<th>BDRMS</th>
<th>BTHRMS</th>
<th>DFILL</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of Cases</td>
<td>708</td>
<td>708</td>
<td>708</td>
<td>708</td>
</tr>
<tr>
<td>Minimum</td>
<td>36000.000</td>
<td>0.100</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>187400.000</td>
<td>49.000</td>
<td>9.000</td>
<td>4.000</td>
</tr>
<tr>
<td>Mean</td>
<td>74318.294</td>
<td>6.315</td>
<td>2.768</td>
<td>1.371</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>18453.037</td>
<td>5.540</td>
<td>0.789</td>
<td>0.555</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DI-694</th>
<th>FF</th>
<th>FIREPL</th>
<th>SALEYR</th>
<th>SMALLLOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of Cases</td>
<td>708</td>
<td>708</td>
<td>708</td>
<td>708</td>
</tr>
<tr>
<td>Minimum</td>
<td>6.000</td>
<td>704.000</td>
<td>0.000</td>
<td>80.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>16.000</td>
<td>2358.000</td>
<td>2.000</td>
<td>89.000</td>
</tr>
<tr>
<td>Mean</td>
<td>13.705</td>
<td>1088.064</td>
<td>0.410</td>
<td>85.542</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.616</td>
<td>177.595</td>
<td>0.594</td>
<td>2.291</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STYLE20</th>
<th>STYLE50</th>
<th>TREES</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of Cases</td>
<td>708</td>
<td>708</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Mean</td>
<td>0.555</td>
<td>0.051</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.497</td>
<td>0.220</td>
</tr>
</tbody>
</table>
2. Persico et al. (2004) investigate the relationship between height and wages in the U.S. and Britain. They have measures of both adult and teen height in inches, age, number of siblings, parental education, and dummy variables indicating whether the parents are in skilled or professional occupations. Table 3 presents several regression models in which the dependent variable is the natural logarithm of the hourly wage for white male workers.

a) Which coefficients in Model 6 are significant at the 0.05 level of significance? Show the tests.

b) Discuss the exact interpretation of the significant coefficients in Model 6.

c) Based on Model 5, what is the predicted wage for a 25 year old man who is 6 feet tall?

d) Are the four parental schooling and occupation variables in Model 6 jointly significant at the 0.05 level of significance?

e) An explanation for the wage premium for height is that employers discriminate in favor of taller people. Models 7 and 8 add teen height to the previous models. What conclusion do you draw from these models about the role of height in wage determination?

<table>
<thead>
<tr>
<th>TABLE 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Estimates ln(Wage) Equation for Adult, White Male Workers, NCDS and NLSY</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Britain: NCDS (N=1,772)</th>
<th>United States: NLSY (N=1,577)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult height (inches)</td>
<td>(.027, .022, .004, .005)</td>
<td>(.025, .018, .002, -.004)</td>
</tr>
<tr>
<td>Youth height (inches)</td>
<td>(.026, .021, .0066, .0066)</td>
<td>(.027, .026, .0095, .0090)</td>
</tr>
<tr>
<td>Age</td>
<td>...</td>
<td>(.028, .027, .024, .023)</td>
</tr>
<tr>
<td>Mother’s years of schooling</td>
<td>(.016, .016, .025, .028)</td>
<td>(.097, .097, .092, .092)</td>
</tr>
<tr>
<td>Mother skilled/professional</td>
<td>(-.080, -.074, -.019, .024)</td>
<td>(.035, .035, .060, .060)</td>
</tr>
<tr>
<td>Father’s years of schooling</td>
<td>(.008, .007, .030, .090)</td>
<td>(.0086, .0087, .0065, .0065)</td>
</tr>
<tr>
<td>Father skilled/professional</td>
<td>(.135, .130, .050, .052)</td>
<td>(.0467, .0465, .0459, .0458)</td>
</tr>
<tr>
<td>Number of siblings</td>
<td>-.033</td>
<td>(.0084, .0084, .0077, .0077)</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>.082</td>
<td>.047, .087, .049</td>
</tr>
<tr>
<td>F-statistic (K, N – K – 1)</td>
<td>9.99</td>
<td>10.25, 11.47, 10.97</td>
</tr>
</tbody>
</table>

Note – Standard errors robust to heteroskedasticity are in parentheses. See the note to table 2. The sample consists only of white male, full-time workers. Each specification includes controls for region and a constant term (results omitted).
3. Many argue that when metropolitan areas experience high levels of suburbanization, central cities become economically and socially isolated. In turn, economic and social isolation has long been suspected to increase crime. Thus, suburbanization may indirectly increase crime. To test the hypothesis, a researcher regressed metropolitan crime rates on a standard set of demographic factors and four indicators of suburbanization. Descriptive statistics are shown below, and regression models are shown on the next page.

a) What is the predicted violent crime rate for the average metropolitan area? Either show your calculations or explain your reasoning.

b) Other things equal, how does the predicted violent crime rate in a metropolitan area with 500,000 residents compare to one with 400,000 residents?

c) Are the suburbanization indicators jointly significant at the 0.05 level of significance?

d) Which of the suburbanization indicators has the largest effect on the violent crime rate?

e) Crime in the central city may encourage people to move to the suburbs. If this is true, does it pose a problem for the analysis? If so, what should the authors do?

Table 1: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent Crime Rate</td>
<td>318</td>
<td>438.78</td>
<td>236.73</td>
</tr>
<tr>
<td>Population (log)</td>
<td>318</td>
<td>12.77</td>
<td>1.05</td>
</tr>
<tr>
<td>Mean Household Income</td>
<td>318</td>
<td>41.03</td>
<td>7.97</td>
</tr>
<tr>
<td>% Non-Hispanic Black</td>
<td>318</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>% Hispanic</td>
<td>318</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>% 13-17</td>
<td>318</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>% 18-24, not in college</td>
<td>318</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>% 65 or older</td>
<td>318</td>
<td>0.13</td>
<td>0.03</td>
</tr>
<tr>
<td>Density Gradient</td>
<td>318</td>
<td>-0.19</td>
<td>0.11</td>
</tr>
<tr>
<td>Population Density</td>
<td>318</td>
<td>423.31</td>
<td>944.48</td>
</tr>
<tr>
<td>% in Central Cities</td>
<td>318</td>
<td>0.42</td>
<td>0.19</td>
</tr>
<tr>
<td>Mean Travel Time</td>
<td>318</td>
<td>22.45</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>312.197</td>
<td>134.380</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(245.645)</td>
<td>(272.748)</td>
<td></td>
</tr>
<tr>
<td><strong>Demographic Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population (log)</td>
<td>51.741</td>
<td>49.547</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.746)</td>
<td>(12.757)</td>
<td></td>
</tr>
<tr>
<td>Mean Household Income</td>
<td>-7.587</td>
<td>-7.627</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.591)</td>
<td>(1.672)</td>
<td></td>
</tr>
<tr>
<td>% Non-Hispanic Black</td>
<td>1176.275</td>
<td>1129.772</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(110.334)</td>
<td>(123.317)</td>
<td></td>
</tr>
<tr>
<td>% Hispanic</td>
<td>714.524</td>
<td>657.073</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(71.197)</td>
<td>(78.050)</td>
<td></td>
</tr>
<tr>
<td>% Age 13-17</td>
<td>-3810.414</td>
<td>-3899.739</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1676.789)</td>
<td>(1765.286)</td>
<td></td>
</tr>
<tr>
<td>% Age 18-24*</td>
<td>-3236.698</td>
<td>-3640.026</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1194.783)</td>
<td>(1186.957)</td>
<td></td>
</tr>
<tr>
<td>% Age 65+</td>
<td>78.001</td>
<td>343.789</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(444.801)</td>
<td>(455.680)</td>
<td></td>
</tr>
<tr>
<td><strong>Suburbanization Indicators</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density Gradient</td>
<td>96.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(175.200)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population Density</td>
<td>-0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% in Central City</td>
<td>199.581</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(64.777)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Travel Time</td>
<td>6.905</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.218)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>57.8</td>
<td>39.98</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.5662</td>
<td>0.5835</td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td>153.07</td>
<td>151.05</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>318</td>
<td>318</td>
<td></td>
</tr>
</tbody>
</table>

*Standard errors in parentheses
*Excluding college students
4. Some U.S. states use annual budgets, while other use multi-year budgets. For example, Texas uses a two year period. Scholars have argued that more populous states with larger budgets are more likely to have annual budgets, and that there are also regional factors and time trends. Kearns (1994) analyzes the determinants of states having an annual budget cycle. Table 2 provides estimates from a linear probability model, a logit model, and a probit model in which the dependent variable PERIOD equals 1 if the state had an annual budget and 0 otherwise. LEGIS is a dummy variables indicating whether the state's legislature meets annually. EXPEND and POP are the natual logarithms of per capita state spending and state population respectively. Also included are dummy variables for year, with 1985 as the base year. The variables NE through PAC are dummies for region; the Midwest is the omitted category. Note: all 50 states are included in the regression in each time period.

a) What is the predicted probability of having an annual budget in 1985 for a state in the Midwest with an annual legislative cycle, $10,000 per capita spending, and a population of 5,000,000? Assume TAX, VAR, AID and CONFLICT are all zero.

b) Describe the precise meaning of the coefficient on EXPEND in each of the three models.

c) What is the marginal effect of EXPEND in each of the models for a state like the one described in (a)?

d) In a linear probability model identical to the one shown except that the regional dummies were omitted, the R² was 0.5411. As a group, do the regional dummies improve the fit of the model?
Table 2. Comparison of LP, LOGIT, and PROBIT results, coefficient estimates, and (t-statistics).

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP</th>
<th>LOGIT</th>
<th>PROBIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEGIS</td>
<td>0.7923***</td>
<td>8.7239***</td>
<td>4.2914***</td>
</tr>
<tr>
<td></td>
<td>(17.73)</td>
<td>(5.5252)</td>
<td>(7.4427)</td>
</tr>
<tr>
<td>EXPEND</td>
<td>0.21483*</td>
<td>3.0181**</td>
<td>1.7390**</td>
</tr>
<tr>
<td></td>
<td>(1.695)</td>
<td>(2.2421)</td>
<td>(2.2802)</td>
</tr>
<tr>
<td>POP</td>
<td>0.08971***</td>
<td>1.0093***</td>
<td>0.5886***</td>
</tr>
<tr>
<td></td>
<td>(2.572)</td>
<td>(2.7406)</td>
<td>(2.7888)</td>
</tr>
<tr>
<td>TAX</td>
<td>-0.05388</td>
<td>-0.58503</td>
<td>-0.47672</td>
</tr>
<tr>
<td></td>
<td>(0.3673)</td>
<td>(0.34043)</td>
<td>(0.50284)</td>
</tr>
<tr>
<td>VAR</td>
<td>0.05513**</td>
<td>0.4874*</td>
<td>0.28862*</td>
</tr>
<tr>
<td></td>
<td>(2.022)</td>
<td>(1.6551)</td>
<td>(1.7429)</td>
</tr>
<tr>
<td>AID</td>
<td>-0.09785</td>
<td>-1.1760</td>
<td>-0.69357</td>
</tr>
<tr>
<td></td>
<td>(0.9715)</td>
<td>(1.0778)</td>
<td>(1.1504)</td>
</tr>
<tr>
<td>CONFLICT</td>
<td>0.02677</td>
<td>0.15797</td>
<td>0.12713</td>
</tr>
<tr>
<td></td>
<td>(0.6714)</td>
<td>(0.35721)</td>
<td>(0.51389)</td>
</tr>
<tr>
<td>1960</td>
<td>0.35940</td>
<td>5.6346</td>
<td>2.7899</td>
</tr>
<tr>
<td></td>
<td>(2.675)</td>
<td>(2.8886)</td>
<td>(2.8571)</td>
</tr>
<tr>
<td>1965</td>
<td>0.29944</td>
<td>3.9031</td>
<td>2.0737</td>
</tr>
<tr>
<td></td>
<td>(2.876)</td>
<td>(2.9842)</td>
<td>(2.9933)</td>
</tr>
<tr>
<td>1970</td>
<td>0.1956</td>
<td>1.6125</td>
<td>0.8665</td>
</tr>
<tr>
<td></td>
<td>(2.196)</td>
<td>(1.8909)</td>
<td>(1.7765)</td>
</tr>
<tr>
<td>1975</td>
<td>0.1237</td>
<td>0.94331</td>
<td>0.50043</td>
</tr>
<tr>
<td></td>
<td>(1.803)</td>
<td>(1.4026)</td>
<td>(1.3210)</td>
</tr>
<tr>
<td>1980</td>
<td>0.02872</td>
<td>0.3171</td>
<td>0.1197</td>
</tr>
<tr>
<td></td>
<td>(0.4297)</td>
<td>(0.52689)</td>
<td>(0.349)</td>
</tr>
<tr>
<td>NE</td>
<td>0.32652</td>
<td>2.6523</td>
<td>1.7056</td>
</tr>
<tr>
<td></td>
<td>(3.440)</td>
<td>(2.7846)</td>
<td>(3.1406)</td>
</tr>
<tr>
<td>MA</td>
<td>0.45262</td>
<td>14.991</td>
<td>7.8105</td>
</tr>
<tr>
<td></td>
<td>(4.571)</td>
<td>(1.8781)</td>
<td>(1.8273)</td>
</tr>
<tr>
<td>WNC</td>
<td>0.29329</td>
<td>2.5247</td>
<td>1.4973</td>
</tr>
<tr>
<td></td>
<td>(3.290)</td>
<td>(2.8781)</td>
<td>(2.9667)</td>
</tr>
<tr>
<td>SA</td>
<td>0.40395</td>
<td>3.1677</td>
<td>1.8760</td>
</tr>
<tr>
<td></td>
<td>(4.926)</td>
<td>(3.8431)</td>
<td>(4.0078)</td>
</tr>
<tr>
<td>ESC</td>
<td>0.48625</td>
<td>5.4289</td>
<td>2.9331</td>
</tr>
<tr>
<td></td>
<td>(4.785)</td>
<td>(3.5376)</td>
<td>(3.9828)</td>
</tr>
<tr>
<td>WSC</td>
<td>0.43105</td>
<td>4.7899</td>
<td>2.5409</td>
</tr>
<tr>
<td></td>
<td>(4.389)</td>
<td>(2.8273)</td>
<td>(3.20902)</td>
</tr>
<tr>
<td>MTN</td>
<td>0.46442</td>
<td>4.0321</td>
<td>2.3723</td>
</tr>
<tr>
<td></td>
<td>(4.679)</td>
<td>(3.8738)</td>
<td>(4.0013)</td>
</tr>
<tr>
<td>PAC</td>
<td>0.20726</td>
<td>0.6518</td>
<td>0.52497</td>
</tr>
<tr>
<td></td>
<td>(2.091)</td>
<td>(0.65341)</td>
<td>(0.88399)</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>-3.0839</td>
<td>-42.348</td>
<td>-23.738</td>
</tr>
<tr>
<td></td>
<td>(2.793)</td>
<td>(3.3662)</td>
<td>(3.3453)</td>
</tr>
</tbody>
</table>

\[ R^2 \quad 0.6192 \]
\[ \text{Maddala } R^2 \quad 0.5772 \quad 0.5740 \]
\[ \text{McFadden } R^2 \quad 0.6217 \quad 0.6163 \]

Correct predictions

- Number 269 270 269
- Percentage 0.89667 0.9000 0.89667

Log likelihood function

- -72.6616 -78.568 -79.706

* Significant at 95% confidence level, one-tailed test.
** Significant at 97.5% confidence level, one-tailed test.
*** Significant at 99% confidence level or higher, one-tailed test.
The Women, Infants, and Children (WIC) program provides prenatal nutrition to low-income mothers. The goal of the program is to increase the health of newborn children. One of the most important indicators of infant health is birth weight. Low birth weight babies suffer from many more physical problems and often require very expensive care. The regression shown in Table 1 regresses birth weight (in grams) on an indicator variable for the WIC program and a number of control variables.

a) Based on Model 4, what is the predicted birth weight for a male child, not part of a multiple birth, who is the first birth to an unmarried mother who is on WIC, age 16, with 12 years of education?

b) What is the effect of the WIC program according to the different models? Are the estimated effects statistically significant? Why do they vary so much? Which estimate do you think is the best?

c) Explain the meaning of the three age coefficients in Model 4.

d) Are the mother’s characteristics in Model 4 jointly significant at the 0.05 level? Show the test and explain the result.

e) The variance inflation factors (VIF) and the Breusch-Pagan tests for model 5 are shown below. What do you conclude based on this information?

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>1/VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>58.42</td>
<td>0.0171</td>
</tr>
<tr>
<td>age2</td>
<td>58.05</td>
<td>0.0172</td>
</tr>
<tr>
<td>wic</td>
<td>2.40</td>
<td>0.4173</td>
</tr>
<tr>
<td>educ</td>
<td>2.18</td>
<td>0.4594</td>
</tr>
<tr>
<td>married</td>
<td>1.52</td>
<td>0.6569</td>
</tr>
<tr>
<td>prev</td>
<td>1.01</td>
<td>0.9914</td>
</tr>
<tr>
<td>male</td>
<td>1.01</td>
<td>0.9921</td>
</tr>
<tr>
<td>multi</td>
<td>1.00</td>
<td>0.9955</td>
</tr>
</tbody>
</table>

Mean VIF | 15.70

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of bw
chi2(1) = 0.00
Prob > chi2 = 0.9964
Table 1: Models of Birth Weight (Grams) of Children Born to Low Income, WIC Eligible Mothers

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td>3395</td>
<td>3367</td>
<td>3071</td>
<td>3078</td>
<td>2571</td>
</tr>
<tr>
<td></td>
<td>(16)</td>
<td>(18)</td>
<td>(115)</td>
<td>(104)</td>
<td>(209)</td>
</tr>
<tr>
<td><strong>WIC (Yes=1, No=0)</strong></td>
<td>-64</td>
<td>-71</td>
<td>84</td>
<td>80</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>(24)</td>
<td>(22)</td>
<td>(34)</td>
<td>(31)</td>
<td>(31)</td>
</tr>
<tr>
<td><strong>Child Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>118</td>
<td></td>
<td>99</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(22)</td>
<td></td>
<td>(20)</td>
<td>(20)</td>
<td></td>
</tr>
<tr>
<td>Multiple Birth</td>
<td>-763</td>
<td>-781</td>
<td>-783</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(59)</td>
<td>(54)</td>
<td>(54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mother Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of Mother (Base: &lt;18 years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-19 years</td>
<td>9</td>
<td>-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(57)</td>
<td>(52)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-34 years</td>
<td>76</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(44)</td>
<td>(40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35+ years</td>
<td>-54</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(80)</td>
<td>(73)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmarried (yes=1, no=0)</td>
<td>176</td>
<td>192</td>
<td>191</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(27)</td>
<td>(25)</td>
<td>(24)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous Births (number)</td>
<td>-153</td>
<td>-147</td>
<td>-145</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(16)</td>
<td>(14)</td>
<td>(14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education (years)</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>(8)</td>
<td>(8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td>44</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(15)</td>
<td></td>
</tr>
<tr>
<td>Age squared</td>
<td></td>
<td></td>
<td></td>
<td>-0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.29)</td>
<td></td>
</tr>
<tr>
<td><strong>r2</strong></td>
<td>0.007</td>
<td>0.173</td>
<td>0.142</td>
<td>0.305</td>
<td>0.308</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>977</td>
<td>977</td>
<td>977</td>
<td>977</td>
<td>977</td>
</tr>
</tbody>
</table>

Note: Standard errors in parentheses.
In conducting regression analyses, researchers have a seemingly limitless tool bag of specification tricks. Variables can be transformed and interacted in a variety of ways. Suppose you have a dependent variable $Y$, several continuous independent variables $X_2$ through $X_4$, and a “dummy variable,” i.e. a binary independent variable $D$, coded as 1 if the observation is a member of a specific group and 0 otherwise. In what circumstances would you employ each of the specifications listed below? Also discuss how the interpretation of each model would differ from a standard linear model, i.e. $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + u_i$

(a) A “log-lin” model:
$$\ln(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + u_i$$

(b) A “log-log” specification for the relationship between $Y$ and $X_2$:
$$\ln(Y_i) = \beta_1 + \beta_2 \ln(X_{2i}) + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + u_i$$

(c) A quadratic specification for the effect of $X_3$.
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + \beta_6 X_{3i}^2 + u_i$$

(d) A model that includes an interaction between $X_4$ and $D$.
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + \beta_6 D_i X_{4i} + u_i$$

(e) A piece-wise linear specification for the effect of $X_2$, where $C$ is dummy indicating that $X_2$ is above a specified value ($X_2^*$).
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 D_i + \beta_6 C_i (X_{2i} - X_{2i}^*) + u_i$$
7. In an effort to understand the racial and ethnic gaps in student achievement, a researcher administered a word recognition test to a sample of kindergarten children. All of the children were between 5 and 6 years old. (The mean age in the sample was 66 months.) In Model I, the researcher regressed the word recognition score (converted to a z score) on indicator variables for race/ethnicity and gender, and a quadratic specification for age in months. This model establishes the baseline gaps between the groups and between genders, conditional on age. In Model II, the researcher controlled for several variables that measure the socio-economic status of the child’s family (family income and parents’ education levels). Finally, in Model III, the researcher controlled for the mean test score in the child’s school to test for neighborhood, school, and peer effects. The word recognition score, the socio-economic variables, and school mean scores have been converted to z scores; in other words, these variables have means of 0 and variances of 1. The results of the regressions are presented in Table 1 on the following page.

(a) Based on Model III, what is the predicted test score for a white male child who has average values on all other characteristics?

(b) Based on Model I, what is the marginal effect of age on word recognition score for a 60 month old child? How does that effect change as the child ages?

(c) Only one of the socioeconomic variables is even borderline significant. If you had the raw data, what steps would you take to test for multicollinearity? If multicollinearity is present, what would you do about it?

(d) Test the hypothesis that family socioeconomic status helps to explain word recognition score at the 0.01 level of significance.

(e) Discuss what these results tell you about the direct and indirect effects of race, ethnicity, and gender on achievement among students.
Table 1: Ordinary Least Squares Regression of Word Recognition Scores

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-8.019 ***</td>
<td>-8.917 ***</td>
<td>-8.125 ***</td>
</tr>
<tr>
<td>Black</td>
<td>-0.376 ***</td>
<td>-0.191 **</td>
<td>0.009</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.110 **</td>
<td>0.015 *</td>
<td>-0.006</td>
</tr>
<tr>
<td>Asian</td>
<td>0.349 ***</td>
<td>0.139 **</td>
<td>0.033 *</td>
</tr>
<tr>
<td>Indian</td>
<td>-0.323 ***</td>
<td>-0.043 *</td>
<td>-0.056 *</td>
</tr>
<tr>
<td>Age (months)</td>
<td>0.199 ***</td>
<td>0.211 ***</td>
<td>0.197 ***</td>
</tr>
<tr>
<td>Age squared</td>
<td>-0.001 **</td>
<td>-0.001 ***</td>
<td>-0.001 ***</td>
</tr>
<tr>
<td>Male</td>
<td>-0.468 ***</td>
<td>-0.228 ***</td>
<td>-0.192 **</td>
</tr>
<tr>
<td>Family Income</td>
<td>0.158 *</td>
<td>0.109 *</td>
<td></td>
</tr>
<tr>
<td>Father's Education</td>
<td>0.045</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>Mother's Education</td>
<td>0.055</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>School Mean Score</td>
<td>0.211 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R^2</td>
<td>0.171</td>
<td>0.256</td>
<td>0.270</td>
</tr>
<tr>
<td>N</td>
<td>597</td>
<td>597</td>
<td>597</td>
</tr>
</tbody>
</table>

Significance Level: *p<0.10, **p<0.05, ***p<0.01
C. Suggested Readings


Paul A. Jargowsky and Rebecca Yang, *Descriptive and Inferential Statistics*.

Paul A. Jargowsky, *Using the F Statistic to Test Hypotheses*.

Paul A. Jargowsky, *Omitted Variable Bias*.


