How to Prepare Appendix B

One of the skills that the term paper demonstrates is the ability to think through a statistical model (i.e., the dependent variable in an equation and the independent variable used to explain the dependent variable) and estimate the statistical results. The first step is to examine the datasets listed ahead and see what variable would be useful to explain given your particular policy area. This is the dependent variable. This is a critical step. Without a dependent variable there is nothing to analyze. Next, you need to select independent variables from the same dataset as the dependent variable (i.e., you can't use variables from two different datasets) that would logically be related to your dependent variable. Unless you have a firm theory to indicate why a particular independent variable should be used, DON'T use it. In other words, DON'T just include an independent variable to "see if matters." Only use independent variables that have a firm theory as to why they should influence your dependent variable. Also, what should the direction of the relationship between the independent variable and the dependent variable be? Thus, should a higher score on the independent variable lead to a higher, or lower score, on the dependent variable? What theory would explain the anticipated direction of this relationship? Make sure you bring this reading assignment to the Horn Center when you estimate your statistical results. DON'T expect to be able to do the assignment without consulting this reading assignment as you estimate the statistical results.

Paragraphs 1-4 of Appendix B in the sample term paper contain a discussion of why attitudes toward the Kyoto Protocol may tell us something useful about attitudes toward high speed internet access. You need to have a similar pattern of reasoning. Thus, why should the reader be concerned about the analysis you will subsequently provide? Both the version of Appendix B you submit for the take-home quiz and the version that appears in the term paper itself need to have a firm rationale for the model used in the statistical analysis. As the data sets ahead are not likely to directly examine the policy area of your term paper, you need to think "creatively."

Since the data sets are surveys of people (i.e., a person is the "unit of analysis"), let me mention several "fundamental" relationships between a person's income and education and their likely positions on political issues/policies that might be useful in formulating your model. In economic issues (or issues argued on an economic basis, e.g., the minimum wage, health care) income is typically a good predictor of someone's opinion. The "basic" relationship is that the higher a person's income the less supportive they will be of liberal economics (using the government to reduce economic inequality and maintain economic security). Alternatively, we could say that the relationship between income and support for economic liberalism is negative. Thus, those with higher incomes are less likely to support universal health insurance, increasing the minimum wage and having the wealthy bare a higher proportion of the tax burden than are those with lower incomes. In economic issues, self-interest is a good, but far from perfect, predictor of opinion.

In noneconomic issues education is a better predictor of a person's opinion than income. The "basic" relationship is that the higher the level of education an individual has the more liberal their opinions on noneconomic issues (i.e., supporting the freedom to differ on noneconomic issues – e.g., support for gay marriage - civil rights, the right of dissent, rights of the accused, etc.). Alternatively, we could say that the relationship between education and noneconomic liberalism is positive. The probable reason for this relationship is that education exposes a person to different ideas and cultures. While this process does not mean a person will change their views, it does typically lead to a greater appreciation and understanding for why others may hold different opinions. Since tolerance and equality are the underpinnings of liberal positions on noneconomic issues, increasing education often translates into more liberal thinking on noneconomic questions.

Variable List for Datasets for Appendix B

This file contains the variable names and descriptions for the various datasets available for the statistical analysis in Appendix B. The datasets are in Excel and are available at my website: www.csulb.edu/~cdennis (click on "Courses"). Later instructions will show how to read Excel files into STATA 11 (the statistical package we'll use). After the variable descriptions for each dataset, I will explain how to estimate the statistical results which appear in Appendix B of the term paper.

300Cigarette

This dataset examines cigarette consumption. The data are annual by state (i.e., a state is the unit of analysis) over the 1985-95 period. The data were supplied by Professor Jonathan Gruber (MIT) and was taken from:
http://econpapers.repec.org/paper/bocbocins/
http://fmwww.bc.edu/ec-p/data/stockwatson/cig_ch10.dta

packpc – packs of cigarettes consumed per person

educ90 – percentage of a state's who are 25, or older, who have at least a bachelor's degree (as of 1990)

incpc – income per capita (i.e., per person)

avgprs – average price of a pack of cigarettes including excise taxes

taxs – average excise taxes for fiscal year, including sales taxes

cpi – consumer price index

pop – state population

300Environmental1

This dataset contains respondents' attitudes toward the Kyoto Treaty on Global Warming, international trade as well as other variables. The variables were selected from the Global Climate Change Data Project. The data were provided by Professor David Weimer of the University of Wisconsin.

Potential Dependent Variables

kyoto - The U.S. Senate has not yet voted on whether to ratify the Kyoto Protocol. If the U.S. does not ratify the treaty, it is very unlikely that the Protocol can be successfully implemented.

Suppose that a national vote or referendum were held today in which U.S. residents could vote to advise their Senators whether to support or oppose ratifying the Kyoto Protocol. If U.S. compliance with the treaty would cost your household (randomly selected price for a gallon of gasoline – e.g., $2.75) dollars per year in increased energy and gasoline prices, would you vote for or against having your Senators support ratification of the Kyoto Protocol? Keep in mind that the (the dollar figure used previously is repeated here) dollars spent on increased energy and gasoline prices could not be spent on other things, such as other household expenses, charities, groceries, or car payments. Note: The form of this question is especially interesting because respondents were told that ratification would result in higher gasoline prices (amount randomly chosen for each respondent). Therefore, respondents knew that ratification of the Kyoto Protocol would not be "costless."

Numerical Label	
0	against
1	for

intagree - Government officials in the US are currently considering a proposed international treaty that concerns global climate change, called the Kyoto Protocol. In 1997 representatives from the U.S. and approximately 150 other nations developed and signed the Kyoto Protocol, which calls for reducing the production of greenhouse gasses.
The U.S. has negotiated similar treaties with other nations to try to deal with other environmental problems, such as acid rain and ozone depletion. On a scale from zero to ten where zero means it is a very bad idea and ten means it is a very good idea, how do you view international treaties as a way to deal with environmental problems?

Numerical label	
0	very bad idea
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	very good idea

trade - Where tradeoffs must be made between environmental protection and property rights, the emphasis should be on protecting property rights.

Numeric	Label
0	strongly disagree
1	disagree
2	agree
3	strongly agree

brink - On a scale from zero to ten where zero means that there is no real environmental threat to civilization and ten means that human civilization is on the brink of collapse due to environmental threats, what do you think about the current environmental situation?

Numeric	Label
0	no real threat
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	brink of collapse

Potential Independent Variables

educ - education level

Numeric	Label
0	less than high school
1	some high school
2	high school graduate
3	some college
4	college graduate
5	some graduate school
6	graduate degree

income - income in dollars—midpoint of ordinal income categories (e.g. if the respondent selected an income category of between $25,000 to $30,000 – see categories in "Incord" above - the computer would read a score of $27,500 (i.e., the midpoint between $25,000 and $30,000).

age – respondent's age in years

gender

Numeric	Label
0 = male; 1 = female	

ideology – Which of the following best describes your political ideology?

Numeric	Label
0	strongly liberal
1	liberal
2	slightly liberal
3	middle of the road
4	slightly conservative
5	conservative
6	strongly conservative

party - With what political party do you identify?

Numeric	Label
0 =	Republican party
1 =	Independent/No party
2 =	Democratic party
3 =	Green party

Thus, a continuum from most conservative to most liberal

hear- Has the respondent heard about the proposed international treaty called the Kyoto Protocol? Note: I listed "Hear" as a potential independent variable. It is also a potential dependent variable. Thus, it could useful to explain the information level the respondent has (e.g., How well is the respondent's information level explained by the education, income, age, etc.?)

Numeric	Label
0	no
1	yes

comph - Do you have regular access to a computer at your residence?

Numeric	Label
0	no
1	yes

compo - Do you have regular access to a computer outside home--like at work or school?

Numeric	Label
0	no
1	yes

300Fatalities

This dataset examines highway fatalities. The data are for 48 U.S. states (excluding Alaska and Hawaii) annually for 1982 through 1988. The data were provided by Professor Christopher J. Ruhm of the Department of Economics at the University of North Carolina.

mrall - Vehicle Fatality Rate - the number of traffic deaths in a given state in a given year, per 10,000 people living inthat state in that year. Traffic fatality data were obtained from the U.S. Department of Transportation Fatal Accident Reporting System.

spircons - Spirits Consumption

beertax - Tax on Case of Beer - The beer tax is the tax on a case of beer, which is an available measure of state alcohol taxes more generally.

yngdrv - % of Drivers Aged 15-24

jaild - Mandatory Jail Sentence – Coded "1" if the state requires jail time for an initial drunk driving conviction and "0" otherwise.

comserd - Mandatory Community Service – "Coded" if the state requires community for an initial drunk driving conviction and "0" otherwise.

unrate - Unemployment Rate

perinc - Per Capita Personal Income

educ90 – As of 1990 the percentage of those 25, and older, who have at least a bachelor's degree.

300 Hibbs

This is a portion of the dataset used by Douglas A. Hibbs and Violeta Piculescu in "Tax Toleration and Tax Compliance: How Government Affects the Propensity of Firms to Enter the Unofficial Economy" (American Journal of Political Science, January, 2010, pp. 18-33). The data were provided by Professor Douglas A. Hibbs. The "unofficial" economy refers to the production and sale of goods that evade official taxation and regulation. This data could be useful for examining factors impacting the degree to which business complies with tax policy/regulation and the perception by business managers of the impact of various government policies.

Most of the data are based on interviews obtained from managers of 3,686 enterprises distributed over 55 countries by the World Bank's World Business Environment Surveys in 2000. The following variables are responses by the business managers surveyed to the following type of question: "Please judge on a four point scale how problematic are these different regulatory areas for the operation and growth of your business? (0 = no obstacle, 1 = minor obstacle, 2 = moderate obstacle and 3 = major obstacle)

q17lic – business licensing (i.e., Please judge on a four point scale how problematic business licensing is for the operation and growth of your business? - 0 = no obstacle, 1 = minor obstacle, 2 = moderate obstacle and 3 = major obstacle)

q17cus – customs/foreign trade regulations (same setup as q17lic)

q17lab – labor regulations (same setup as q17lic)

q17for – foreign currency/exchange regulations (same setup as q17lic)

q17env – environmental regulations (same setup as q17lic)

q17fir – fire and safety regulations (same setup as q17lic)

q17hit – high taxes (same setup as q17lic)

q49fin – financing (same setup as q17lic)

q49jud – functioning of the judiciary (same setup as q17lic)

paytax99 - payroll taxes (i.e., social insurance taxes) as a percentage of a nation's gross domestic product in 1999.

cinctax - highest marginal tax rate on corporate profits in 2000

assets – managers' estimates of value of their firm's fixed assets (land, buildings and equipment) in U.S. dollars – 10 categories (1-11 ranging from $250,000 to $500,000 or more).

taxcomp – tax compliance – percentage of a firm's total sales which are reported for tax purposes (broken into seven categories of responses - 0= <50%, 1=50%-59%, 2=60%-69%, 3=70%-79%, 4=80%-89%,5=90-99% and 6=100%). This is the central dependent variable in this study.

300California1

The variables in 300California1 were selected from the Public Policy Institute of California's November, 2008 survey on Californian's attitudes toward California's public colleges and universities.

Q36. How about spending more state government money to keep down tuition and fee costs, even if it means less money for other state programs? (Do you favor or oppose this proposal?)

0 favor
1 oppose

Q37. How about having a sliding scale for tuition and fee costs, so that students pay according to their income status? (Do you favor or oppose this proposal?)

0 favor
1 oppose

Next, California Community College enrollment fees are currently $20 dollars per unit, which is a decrease from $26 dollars per unit two years ago.

[ROTATE Q38 AND Q39 - This is to minimize "order of questions" effect.]

Q38. Do you think that enrollment fees in the California Community College system are currently about the right amount, too high or too low?

0 too low
1 about the right amount
2 too high

Q39. Do you think that enrollment fees in the California Community College system are currently about the same as, higher than, or lower than enrollment fees in other states?

0 lower than
1 about the same as
2 higher than

Changing topics,

As you may know, the state government has an annual budget of around $100 billion dollars and currently faces a multibillion dollar gap between spending and revenues.

Q40. How concerned are you that the state's budget gap will cause significant spending cuts in higher education?

0 very concerned
1 somewhat concerned
2 not too concerned
3 not at all concerned

Q41. Given the state's current budget situation, on a scale of 1 to 5--with 1 being a very low priority and 5 being a very high priority –what priority should be given to spending for California's public colleges and

universities? [INTERVIEWER: Do not read text of answers, if necessary repeat, "on a scale of 1-5 with 1 being a very low priority and 5 being a very high priority, what priority should be given to spending for California's public colleges and universities?"]

0 very low priority
1 low priority
2 medium priority
3 high priority
4 very high priority

Next, what if the state said it needed more money just to maintain current funding for public colleges and universities.

Q42. Would you be willing to pay higher taxes for this purpose, or not?

0 yes
1 no

Q43. Would you be willing to increase student fees for this purpose, or not?

0 yes
1 no

Q44. Next, in general, how important is California's higher education system to the quality of life and economic vitality of the state over the next 20 years—very important, somewhat important, not too important, or not at all important?

0 very important
1 somewhat important
2 not too important
3 not at all important

Q45. In thinking ahead 20 years, if current trends continue do you think California's economy will need [2] a higher percentage, [0] a lower percentage, [OR][1] about the same percentage of college educated workers as today?

0 lower percentage
1 about the same percentage
2 higher percentage

Q46. In thinking ahead 20 years, if current trends continue, do you think California will have (2) more than enough, (0) not enough, [OR] (1)just enough college educated residents needed for the jobs and skills likely to be in demand?

0 not enough
1 just enough
2 more than enough

Q47. In thinking ahead 20 years, how important do you think it is for the state government to be spending more public funds to increase capacity in public colleges and universities—very important, somewhat important, not too important, or not at all important?

0 very important
1 somewhat important
2 not too important
3 not at all important

Q48. How much confidence do you have in the state government's ability to plan for the future of California's higher education system—a great deal, only some, very little, or none?

0 a great deal
1 only some
2 very little
3 none

Q49. Generally speaking, how much interest would you say you have in politics—a great deal, a fair amount, only a little, or none?

0 great deal
1 fair amount
2 only a little
3 none

Q50. Would you consider yourself to be politically:

0 very liberal
1 somewhat liberal
2 middle-of-the-road
3 somewhat conservative
4 very conservative

D1. Finally, we have a few demographic questions. What is your age? [IF NECESSARY: READ LIST]

D4. What do you hope will be the highest grade level that your youngest child will achieve: some high school; high school graduate; some college; college graduate; or a graduate degree after college?

0 some high school
1 high school graduate
2 some college
3 college graduate
4 a graduate degree after college

D6. What was the last grade of school that you completed?
[IF NECESSARY: READ LIST; ENTER "ASSOCIATES DEGREE" AS PUNCH <3> SOME COLLEGE]

0 some high school or less
1 high school graduate/GED
2 some college
3 college graduate
4 post graduate

D9. Finally, which of the following categories best describes your total annual household income before taxes, from all sources?

[PROBE: your best estimate is fine AND/OR REREAD LIST BEFORE ACCEPTING DON'T KNOW OR REFUSED"]

[IF RESPONDENT REFUSES, SAY "We understand and respect that this information is confidential, we ask only for research purposes and will keep all of this information absolutely anonymous"]

0 under $20,000
1 $20,000 to under $40,000
2 $40,000 to under $60,000
3 $60,000 to under $80,000
4 $80,000 to under $100,000
5 $100,000 to under $200,000
6 $200,000 or more

300California2

The variables in 300California2 were selected from the Public Policy Institute of California's January, 2007 survey. The topics include spending levels on various budget categories (corrections, K-12 public education, colleges and universities, health and human services and roads and infrastructure), tradeoffs between spending reductions and tax increases and health care.

Q13. How about the state's corrections system, including prisons? (Do you think that the state government should spend more money than it does now, the same amount as now, or less money than now?)

0 more money
1 same amount of money
2 less money
3 [VOL] should spend no money at all

Q14. How about the K through 12 public education system? (Do you think that the state government should spend more money than it does now, the same amount as now, or less money than now?)

0 more money
1 same amount of money
2 less money
3 [VOL] should spend no money at all

Q15. How about public colleges and universities? (Do you think that the state government should spend more money than it does now, the same amount as now, or less money than now?)

0 more money
1 same amount of money
2 less money
3 [VOL] should spend no money at all

Q16. How about health and human services? (Do you think that the state government should spend more money than it does now, the same amount as now, or less money than now?)

0 more money
1 same amount of money
2 less money
3 [VOL] should spend no money at all

Q17. How about roads and other infrastructure projects? (Do you think that the state government should spend more money than it does now, the same amount as now, or less money than now?)

0 more money
1 same amount of money
2 less money
3 [VOL] should spend no money at all

Q19. And, in general, which of the following statements do you agree with more—I'd rather pay higher taxes and have a state government that provides more services, or I'd rather pay lower taxes and have a state government that provides fewer services?

0 higher taxes and more services
1 lower taxes and fewer services

Q48. Which would you prefer [0] the current health insurance system in the United States, in which most people get their health insurance from private employers, but some people have no insurance [OR] [1] a universal health insurance program, in which everyone is covered under a program like Medicare that is run by the government and financed by taxpayers?

0 current system
1 universal health insurance system

Q49. Do you favor or oppose the U.S. government guaranteeing health insurance for all citizens, even if it means raising taxes?

0 favor
1 oppose

Q53. Next, would you consider yourself to be politically: [READ LIST, ROTATE ORDER TOP TO BOTTOM]

0 very liberal
1 somewhat liberal

2 middle-of-the-road
3 somewhat conservative
4 very conservative

D1. Finally, we have a few demographic questions. What is your age?

D7. What was the last grade of school that you completed?

1 some high school or less
2 high school graduate/GED
3 some college
4 college graduate
5 post graduate

D10. Finally, which of the following **categories** best describes your total annual household income before taxes, from all sources?

1 Under $20,000
2 $20,000 to under $40,000
3 $40,000 to under $60,000
4 $60,000 to under $80,000
5 $80,000 to under $100,000
6 $100,000 to under $200,000
7 $200,000 or more

Gender: 1 Male 2 Female

300California3

This data contains the percentage of the county-wide vote in favor of some important ballot initiatives in California. One of the initiatives may concern a subject logically related to your term paper. If you use this data, you need to put a disclaimer in Appendix B of your term paper. Assuming your model is that the vote on a ballot initiative is the dependent variable and the independent variables are a group of county demographics (e.g., educational attainment, median household income, etc.), the hypotheses you can test with this data concern county-level voting while the theory underlying the hypotheses is based on the behavior of individuals, not counties. As the readings discuss, there is a potential fallacy in using aggregate measures (e.g., a county vote) to infer the behavior of individuals (300 Reader, pp. 13-14). Given that we do not have access to individual-level data on the variables we need, a county-level analysis is the best available alternative. *For this reason, among others, I'd recommend this dataset only as a "last resort." See if you can't use one of the other datasets.*

Variable Name	Description
county	**Name of county**
prop8	**Percentage of the vote cast in the county in favor of Proposition 8 in November, 2008. A "yes" vote was in favor of banning same-sex marriage.**
prop10	**Percentage of the countywide vote in favor of Proposition 10 (tobacco tax increase use used for early childhood development – brought by Rob Reiner), November 3, 1998 (California Secretary of State – Statement of the Vote).**
prop56	**Percentage of the countywide vote in favor of Proposition 56 (reduce budget threshold in both houses of the state legislature to 55%), 2004.**
prop71	**Percentage of the countywide vote in favor of Proposition 71 (stem cell research), November, 2004.**
prop75	**Percentage of the countywide vote in favor of Proposition 75 (requiring union members to give their consent for their dues to be used for political purposes). November, 2005.**
prop79	**Percentage of the countywide vote in favor of Proposition 79 (using bulk buying power of the state to obtain lower drug prices to those eligible). November, 2005 special election.**
prop128	**Percentage of the countywide vote in favor of Proposition 128 ("Big Green" – environmental), November, 1990.**
prop187	**Percentage of the countywide vote in favor of Proposition 187 (deny government benefits to illegal immigrants), November, 1994.**
prop209	**Percentage of the countywide vote in favor of Proposition 209 (prohibit the State of California from using affirmative action), November, 1996.**

coll00	**Percentage of those 25, or older having a least a Bachelor's degree in 2000.**
coll90	**Same as "coll00" except for 1990.**
medinc05	**Median household income in the county in 2005 in thousands of dollars. Thus, a score of 45.4 means that half the households in that county had an income greater than $45,400 and half the households in that same county had an income less than $45,400 in 2005.**
medinc90	**Same as "medinc05" except for 1990.**
dens06	**Persons per square mile of land area in the county in 2006. This is a measure of population density.**
white05	**Percentage of the county population who were white in 2005.**
afam05	**Percentage of the county population who were African-American in 2005.**
asian05	**Percentage of the county population who were Asian in 2005.**
hispan05	**Percentage of the county population who were either Latino or Hispanic in 2005.**
senior05	**Percentage of the county population who were 65, or older, in 2005.**

NOTE: Demographic data are from the County and City Databook: 2007 (www.census.gov/prod/2008pubs/07ccdb) and earlier editions. Most of the votes are from the California Secretary of State website. Pre-2000 demographic data was supplied by Dan Hopkins (Harvard University from geolytics – which makes available Census Bureau data.

Estimating the Statistical Results for Appendix B of the Term Paper

Now that you have selected the data set you wish to work with, you can estimate the statistical results that will appear in Appendix B of your term paper. ***Make sure read all remaining portions of this document before attempting the statistical analysis for Appendix B of your term paper. As you will read later, the type of dependent variable you are working with will dictate the appropriate statistical technique. Therefore, you need to read the entire discussion before attempting the statistical analysis. Don't just "do what I did" in Appendix B of the sample term paper (i.e., use the same statistical estimator). The estimation procedure I used may not be appropriate for your analysis.***

The statistical package we will use is called "Stata" and is available in the Horn Center and the computer lab in SPA-206. Since it is open many hours and has many computers, I'd recommend you use the Horn Center (Monday-Thursday: 7:45 a.m. – 11:00 p.m., Friday: 7:45 a.m. – 5:00 p.m., Saturday – closed, Sunday 12:30 p.m. – 9:00 p.m. – as these times may change, "double check" these hours by calling 985-2303). If you want to use the lab in SPA-206 call (to make sure they'll be open and room 206 will not have a class in it at the time you want to use it) 985-4986. Since you will not have access to someone familiar with Stata in either the Horn Center or SPA-206, try to coordinate the time you will be in the lab with my phone office hours (i.e., so you can call me at home if you have trouble - 562-597-7287 – see syllabus for the hours I'm available). ***For reasons that will become clear, bring a "flash drive" to the lab with you.***

You can save yourself much time and consternation in the computer lab by going through the variable lists for the available data sets and thinking through what model(s) you want to estimate *prior* to going to the computer lab. For example, read through Appendix B of the sample term paper and follow the reasoning. Pay particular attention to the discussion of why attitudes toward the Kyoto Protocol might be useful in understanding attitudes toward high speed internet service. Your topic may be much different than the one in the sample term paper and, hence, a different data set and/or different model may be appropriate. Just a piece of advice: write out the variable names, beginning with the dependent variable and then proceeding through the independent variables, that you want to estimate *before* going to the computer lab.

Look under "programs" or "classes" or "courses" for STATA 11 (DO NOT USE STATA 9). You can download each of the datasets from my website (www.csulb.edu/~cdennis click on "Courses" and look under POSC 300). Since the datasets are in Excel, you need to save the file as a "tab delimited textfile" in order to read the file into Stata. For example, if you are using the Excel file "300Environmental1" do the following: (1) download the file into Excel on the computer you are using; (2) save the file as 300Enviornmental1 but change the "Save as Type" to "Text (Tab delimited) to a lettered drive (e.g., a flash drive in

the "F" or "H" drives – don't save it to a non-capital lettered drive – e.g., "documents" or "my computer" – because I can't tell you how to access it in Stata); (3) make sure the file name does NOT have spaces in it – e.g., you might save a file under the title: 300termpaper but NOT: 300 term paper. WHEN YOU TRY TO SAVE YOUR FILE AS A TAB DELIMITED TEXT FILE EXCEL WILL ASK YOU QUESTIONS – ANSWER EITHER "OKAY" OR "YES" (WHICHEVER OPTION YOU ARE GIVEN). BE CAREFUL: ONE OF THE BIGGEST PROBLEMS PEOPLE HAVING DOING THIS ASSIGNMENT IS THAT THEY WERE NOT ABLE TO SAVE THE EXCEL FILE AS A "Text (Tab delimited)" FILE (i.e., they thought they saved it as such but actually didn't). You can check to see if your file has been saved as a Text (Tab delimited) file by going in to Excel and: (1) click on "open"; (2) look in the lower right corner of the box which appeared as a result of step 1 and change the header from "All Excel Files" to "Text Files" and see if you actually have a file by the correct name (i.e., the only Excel files that should now be visible will be Tab delimited text files – so if the file name doesn't appear, you need to repeat the previous steps on creating the Text (Tab delimited) file). A "Text(Tab delimited)" file has the file extension .txt (i.e., the file name must end with .txt – not .dtatxt or .dta.txt or something else, just .txt – e.g., 300California1.txt or 300Environmental1.txt).

Assuming you saved the file as a Text (Tab delimited) file to the H drive, type your version of the following command in the Stata 11 command box to retrieve the file:

insheet using H:/300Environmental1.txt (press "enter").

BE CAREFUL! (e.g., don't forget insheet using in the previous command line). Now your data should be read into Stata 11. The variable names should appear on the left side of the screen. Read through the variable list which appears in the coursepack. The variable list tells you the variable names for the variables in each data set.

Before estimating your model, read through the rest of this Appendix. You need to understand the entire discussion before continuing. **As discussed in the multivariate readings for this course, since the dependent variable in the analysis in Appendix B of the sample term paper is dichotomous (i.e., two possible responses - the respondent either favored or opposed ratification of the Kyoto Protocol) multiple regression could not be used. If the dependent variable has two categories of responses (as in Appendix B), the researcher can use either probit or logit. The choice is largely arbitrary. I used probit in Appendix B of the sample term paper. The dependent variable must be listed immediately to the right of the estimation procedure (just keep reading). For example, notice the first equation estimated in Appendix B of the sample term paper:**

probit kyoto educ income gender brink hear comph compo

The above command tells Stata 11 that: (1) "probit" is the estimation procedure to be executed; (2) the dependent variable is kyoto; and (3) there are 7 independent variables (educ income gender brink hear comph compo).

If the dependent variable has *more* than two categories of responses there are several possibilities. First, let us suppose that the dependent variable has three, or more, categories of responses and is an ordinal level measure (just keep reading). Remember from previous readings and class discussion that ordinal means that the response categories *can* be rank-ordered but that we do not know that the difference between the categories is equal. For example, suppose that a survey question asks the respondent to indicate a level of agreement/disagreement with a statement and the possible responses are: strongly agree, agree, neutral, disagree and strongly disagree. This set of possible responses *are* rank-ordered because each succeeding category indicates less agreement with the statement. However, the level of measure is ordinal, not interval, because we do not know that there is an equal distance between each response. Thus, we do not know that the difference between "strongly agree" and "agree" is the same as between "agree" and "neutral." Many of the variables in the datasets mentioned in this appendix are similarly measured.

If the dependent variable is ordinal then choose either ordinal probit (i.e., replace probit in the above command line with oprobit) or ordinal logit (replace probit with ologit in the command line). Once you use either ordinal probit or ordinal logit, the statistical output will contain what are called "cutpoints." Just forget about them. For purposes of this course "cutpoints" are not necessary.

If you use any form of either probit or logit the results are interpreted in the same manner as regression *except* that you *cannot* directly make the magnitude statements that were made on pages 68-69 and 72-73 of the 300Reader. For example, if a probit coefficient is -.677 you CAN'T say that if that independent variable increases by one unit (and we hold the level of all other independent variables constant) the score on the dependent variable will decrease, on average, by six-tenths of a unit (i.e., by .6). If you use regression you CAN make such a statement, but not in either probit or logit. However, like regression we CAN interpret both the direction of the relationship and the degree of statistical significance in probit and logit. Pay VERY careful attention to how probit coefficients are interpreted both on the next page and in Appendix B of the sample term paper. We can make magnitude statements in either probit or logit, but there are a number of "hoops" to jump through first. These "hoops" are beyond the scope of this course. However, the rest of the discussion over the aforementioned pages *is* applicable (i.e., the direction of the relationship between each independent variable and the dependent variable after removing the impact of all other independent variables and the 2.0 t statistic standard for achieving statistical significance at the .05 level).

As an example, look at the first statistical results in Appendix B of the sample term paper (reprinted ahead).

```
probit   kyoto educ income gender brink hear comph compo

Probit regression                                 Number of obs   =      12409
                                                  LR chi2(7)      =    2055.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -7463.2865                       Pseudo R2       =     0.1210
------------------------------------------------------------------------------
       kyoto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0789208   .0104864     7.53   0.000     .0583678    .0994737
      income |   1.74e-06   3.51e-07     4.95   0.000     1.05e-06    2.43e-06
      gender |   .0251577    .024829     1.01   0.311    -.0235063    .0738216
       brink |   .2381675   .0058842    40.48   0.000     .2266348    .2497003
        hear |  -.0940281   .0332897    -2.82   0.005    -.1592748   -.0287814
       comph |  -.0332803   .0366544    -0.91   0.364    -.1051216    .0385609
       compo |   .1800847   .0265302     6.79   0.000     .1280865    .2320829
       _cons |  -1.727818   .0649327   -26.61   0.000    -1.855084   -1.600553
------------------------------------------------------------------------------
```

Since the coefficient for "educ" is positive (i.e., .078 rather than -.078) we know that after removing the impact of all other independent variables in the model (i.e., income, gender, etc.) the more highly educated the respondent the more likely they are to support ratification of the Kyoto Protocol. I used "probit" as the estimation technique because the dependent variable (support or opposition to the Kyoto Protocol) had only two categories of responses (0 = against, 1 = for). A positive relationship between education and support for the Kyoto Protocol means that higher scores on "educ" (higher levels of education) are associated with a higher score on "kyoto." Since "1" means support for the Kyoto Protocol and "0" means opposition, and "1" is a higher score than "0," this means that higher levels of education are associated with a greater probability that the respondent will favor ratification of the Kyoto Protocol. If the probit coefficient for "educ" had been negative (e.g., -.078) it would have meant that the more educated the respondent the less likely they are to favor ratification of the Kyoto Protocol. Read the variable list carefully. You need to know what higher or lower scores on each variable indicate.

Since the absolute value (i.e., disregard positive or negative sign) of the t statistic for education is 7.53 (see "Z" column above) and this figure is well above the 2.0 threshold, we know that there is less than a 5% chance that a respondent's level of education has no impact on their probability of favoring ratification of the Kyoto Protocol (the actual probability is less than 1 in 1,000 – see "P>|z|" column where the entry is 0.000). Much of the analysis in Appendix B of the sample term paper is based upon these interpretations.

Keep in mind that the choice of an estimation technique is determined by the level of measurement of the *dependent variable.* For example, I could use ordinal level *independent* variables in regression, but *not* an ordinal level *dependent* variable. If the dependent variable is either an interval or ratio level measure (i.e., where we can rank-order the responses and we are sure that there is an equal interval between the categories of responses) we should use regression as the estimation procedure (i.e., replace probit with regress in the Stata 11 command line).

I can't stress enough that you need to look very carefully at how the variables are measured. For example, consider the "taxcomp" variable in the Hibbs dataset. The variable list defines "taxcomp" as follows: percentage of a firm's total sales which are reported for tax purposes (broken into seven categories of responses - 0= <50%, 1=50%-59%, 2=60%-69%, 3=70%-79%, 4=80%-89%, 5=90-99% and 6=100%). This is not either an interval or ratio level variable (i.e., the variable is not measured as a percentage). If the firm scores "0" they could report anywhere from "0%" to "50%" of their total sales for tax purposes. If the responses were individual firm percentages rather than broad categories (e.g., the computer read the actual percentage – thus such scores such as 25%, 27%, 52%, 71%, etc.) then the variable would have been a percentage. If this variable was the dependent variable (as in Hibbs' study) then regression would have been appropriate. However, since the World Bank (the data source Hibbs used) coded the responses in the previously mentioned categories, this variable is ordinal (each succeeding score on the 0 through 6 scale indicates a higher percentage of sales reported) but not either interval or ratio (e.g., all scores within a category are treated the same – thus 52% and 59% are in the same category when they are actually different scores and the difference between the broad categories is not equal – category 0 is from 0% to 49% while the next several categories cover only 10%). For these reasons Hibbs had to use either ordinal probit or ordinal logit instead of regression.

If the scores on the dependent variable *cannot* logically be rank- ordered use multinomial probit (i.e., enter "mprobit" instead of "probit" in the Stata 11 command line). For example, suppose the dependent variable is race. What would the continuum be: African-American, Asian, Latino, White, or Latino, White, African-American, Asian? Race is simply not a variable that can be rank-ordered. In such circumstance multinomial probit is the appropriate estimation procedure.

If the categories of responses of the dependent variable can be rank-ordered and each unit on the measuring continuum is equal (e.g., the computer is reading actual percentages – the difference between 32% and 33% is the same as the difference between 72% and 73%) regression is the appropriate estimation procedure (i.e., replace probit with regress in the Stata command line). For regression, use the interpretation procedure discussed on pages 68-69 and 72-73 of the 300Reader.

Let me mention an additional procedure that can enhance the discussion in Appendix B. In the analysis in Appendix B of the sample term paper, two of the independent variables, gender and comph (whether, or not, the respondent had a home computer) are statistically insignificant (i.e., had "t scores" – "Z" in probit – of less than an absolute value of 2.0). Perhaps gender and comph are statistically insignificant because they are highly related to the other independent variables in the model. This is the situation described over page 77 of the 300Reader: an independent variable that is theoretically important may be statistically insignificant due to high multicollinearity. Remember from the discussion (page 77 of the 300Reader) that we only need to be concerned about high multicollinearity for statistically insignificant independent variables. Thus, we do

not **need to be concerned about multicollinearity for the statistically significant independent variables (i.e., education, income, brink, hear and compo).**

Fortunately, Stata 11 provides a very easy way to check for how much of the variation in each independent variable is explained by all the other independent variables in the analysis. To assess the degree of multicollinearity execute the following steps: (1) run the equation of interest (e.g., the first equation that appears in Appendix B of the sample term paper:
probit kyoto educ income gender brink hear comph compo); (2) rerun this equation using regress rather than probit (i.e.,
regress kyoto educ income gender brink hear comph compo); (3) after running the regression in step 2, type vif in the command box and press "enter." Following the aforementioned three steps with the equation in Appendix B produced the following results:

```
vif

    Variable |       VIF       1/VIF
-------------+----------------------
        educ |      1.20    0.834214
      income |      1.15    0.872184
       compo |      1.11    0.899267
        hear |      1.11    0.900412
      gender |      1.10    0.905312
       brink |      1.04    0.962844
       comph |      1.03    0.967975
-------------+----------------------
    Mean VIF |      1.11
```

Remember that multicollinearity is only a concern for statistically insignificant independent variables. From the results for the first equation in Appendix B of the sample term paper, we see that "gender" and "comph" (the whether or not the respondent had a home computer) are statistically insignificant. Subtracting the number in the 1/VIF column from 1.0 indicates the percentage of variation in that particular independent variable that is explained by all the other independent variable together. Only 10% of the variation in gender is explained by all the other independent variables (1 - .90 = .10) and only 3% of the variation in the home computer variable is explained by all other independent variables together (1 - .967 = .033). Since 10% and 3% are well below the threshold for high multicollinearity of 70% (i.e., .10 and .03 are well below .70), high multicollinearity is not likely the reason either gender or having access to a home computer are statistically insignificant predictors of attitudes toward the Kyoto Protocol.

If you have a strong theory linking a statistically insignificant independent variable to the dependent variable, use the above discussion in your version of Appendix B. Just apply the three-step procedure above. Thus, don't explain Stata's approach in Appendix B. Simply mention how much of the variation in the statistically insignificant independent variable in question is explained by the other independent variables and what this indicates. See how this is done at the end of Appendix B of the sample term paper.