Economics 360 Data Analysis Project

For this project, students will apply the methods from class to a real set of data. Below are the milestones at which students are expected to have tangible progress towards completion.

Critical Due Dates:
September 18, 2015: Summary of topic and first 2 FAQs due.
November 13, 2015: Due date to present data set and "working" regression model.
December 14, 2015 (last day of class): Final project due.

1. Pose a question. What interests you? Your data set and hypotheses do not have to have obvious Economics overtones, so if you want to study sports or entertainment, that's okay. Just make sure you can find data on the topic of interest. For example:

- Your friend says that "free throw shooting percentage isn't important to winning in the NBA playoffs." What is the causal relationship of interest? Winning playoff games is caused by making a larger proportion of your free throws, *ceteris paribus*. You should be able to find data on the FT% of teams that win a lot of NBA playoff games and compare them to the FT% of teams that don't. If you find a significantly higher % for teams that go deep in the playoffs, you can go back to your friend and say, "Aha! You don't know diddily about basketball, and I've got the data to prove it!"

Think of some claim that has been made in one of your other classes or by a friend/co-worker/family member that you want to test with data. Then try to find a sample that contains observations you can use to test the claim.

Remember: a good question is specific, capable of being answered empirically, and interesting (non-obvious, non-trivial, original).

By **Friday 18 September**, students must submit a 1 page typed summary of their topic and responses to the first 2 FAQs (from Angrist and Pischke, first day of class) due. The summary must include:

- A testable causal relationship between observable variables,
- A compelling explanation why this relationship interests the student,
- The unit of observation, e.g., individual, country, football team, that will be used in the test.

This is scored pass/fail and counts as 10% of the points on the project. Do not to wait until the last day to submit your topic so you have the opportunity to re-submit in the event that the first proposal is rejected.[1]

2. Data collection. Go find data! Data are all around you, waiting to be organized and analyzed. All one has to do is observe the phenomenon of interest and systematically record observations.

Keep in mind your first 2 FAQs. How will you operationalize them into a regression? Where can you go to observe the "x" and "y" variables in the causal relationship of interest?

[1] This can happen. The instructor will also veto topics he considers too similar to other students' that have already been approved. If you want to "stake out" your research area, get started early!

Constraints:

- Data consist of observations (rows) and variables (columns) and should have a "spreadsheet" layout. A data set must observe multiple variables for multiple (n) elements.
- I'm not asking you to formulate your own survey or anything like that; if you're really ambitious, you can certainly do it, but there are plenty of suitable sample data sets already collected that you can use (see below).
- You need at least two (ratio level) variables, and it is strongly preferred that you have a ratio level dependent variable like wage, price, population, et al., because regression is better suited to analyzing these.
- You need enough information to make meaningful statistical inferences, i.e., large enough sample size and variation in your variables. E.g., it would be hard to infer much about a small Indiana town that enacts a zoning regulation, based on a comparison with 5 neighboring towns that didn't ($n = 6$ and $x = 1$ for only 1 observation!).

Where should you look?

- As with most things these days, the internet is a good place to start: bls.gov, bea.gov, nber.org/data, quandl.com are good general places to find data sets.
- If you have difficulty deciding on a set of data or finding a set that you can use to test your hypotheses, please consult me, and I will help get you going.

Students will sign up for a small group meeting with the instructor or TA to present the following during the week of **9 November**:

- "Working" regression specification,
- Data set in Stata format, and
- Codebook (printed hand-out for the grader) explaining variable definitions.

"Present" means a 5-10 minute demonstration in which you open the data set, explain what variables and observations you have, and answer a couple practical questions that will help make the rest of the project easier. This is scored "pass" (15), "low pass" (10), or "fail" (0) and counts as 15% of the grade on the project.

3. Econometric Analysis. Students will document all of the following in a word-processed report and submit it on the last day of class. All tables and figures should be "self-contained" by including a caption and intuitive labels for the rows, columns and axes.

3a. Give a sense of how your variables are distributed. Your write-up should include a professional and easily understandable table of the descriptive statistics on your variables. Label this one "Table 1: Descriptive Statistics" in your write-up. This means sample size, sample mean, a measure of variability such as standard deviation, and skewness. For categorical or binary variables, make it clear how you have made them quantitative and that the means represent proportions, e.g., the proportion that is male, lives in Tippecanoe county, or the proportion of the songs on your mp3 player that is a particular genre.

In the write-up:

- Carefully explain the units (weekly income? monthly? annual?) and the unit of observation (county? state? occupation-state?)

- Do all the descriptive statistics seem plausible? If they do not, what are some explanations for their bias?
- Are there missing observations or outliers for any variables? If so offer an explanation.
- Does the size of your sample present any concerns about the normality of the sampling distribution? Speculate about whether the dependent variable's distribution (skewness, outliers) presents any problems for the Central Limit Theorem. Would taking logs help?[2]

3b. Use Stata to estimate a simple linear regression for the relationship between the (hypothesized) causally related variables:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Also use Stata to produce a scatterplot showing the mapping of *x* to *y*, and include the estimated regression line on the plot to summarize their co-movement.

In the write-up:

- State the null and alternative hypotheses in terms of parameters (βs) that will test the relationship of interest.
- Discuss the sign (+/-) on $\hat{\beta}_1$? Does it confirm your original prediction?
- Discuss the default and robust standard errors of $\hat{\beta}_1$ and how statistically different from the null (usually but not always zero) hypothesized value the estimate is. In practical terms, is there a "wide" confidence interval around the point estimate?
- In terms a non-economist could understand, interpret the coefficient estimate: ". . . a one unit change in . . . is associated with a . . . ." Is this a practically large effect?
- Discuss how well the linear trend "fits" the data. What is the coefficient of determination (R^2)?

3c. Robustness part I. Build up your regression specification with explanatory variables that either: i) shrink the error variance and improve precision of the estimates, or ii) control for omitted factors in the error term (in the simple OLS specification). Create a table like the one in Ben's lecture notes, showing between 4 and 6 different specifications (1 estimate per column) and enabling the reader to compare the $\hat{\beta}$s of interest by reading across one row.

The lower portion of the table should have a row that enables the reader to differentiate the estimates according to what else is included in the model, e.g.,

Table 2: Label this one "Table 2: Regression Estimates" in your Write-Up

	a	**b**	**c**	**d**
Coefficient estimate	$\hat{\beta}_{1\|simple\ OLS}$ $(s.e.)$	$\hat{\beta}_{1\|spec.\ b}$ $(s.e.)$	$\hat{\beta}_{1\|c}$ $(s.e.)$	$\hat{\beta}_{1\|d}$ $(s.e.)$
Controls	None	Age	Age and state	All
Adjusted R^2				

[2] I recommend, before proceeding to write up your results from 3 b-e, that students get their functional specifications (especially of *y*) right: logs or levels, scaling by 1000 or 1/1000.

As usual the table should have a caption that explains the cells, e.g., what is included in "All"?

In the write-up:

- Devote at least 1 paragraph (each) to discussing variables in the error term that could create omitted variable bias. State specifically what in the error term (think education and omitted ability) is related to *x* (and why, theoretically, you should worry about this) and whether it would bias $\hat{\beta}$ upward or downward. Do this for 2 different potential sources of bias.
    - This might seem challenging if you haven't taken a lot of other Econ. theory classes, but feel free to talk to your instructor or TA about this to pitch your ideas.
- Discuss the *set* of estimates. How does $\hat{\beta}$ change with the addition of controls? Is this consistent with controlling for omitted variables and reducing bias (see above)?
- Comment on what's going on with $\bar{R}^2$ and standard errors as you add controls.
- Assess your level of satisfaction with how the multiple regression tackles omitted variable bias.
    - It's okay if you are critical. Often the omitted factors are very difficult to observe and control for in cross sectional samples.

3d. Robustness part II. Extend your causal hypothesis to groups within the sample. For example: "stricter parental ratings will have a negative effect on video game sales. But it will have a *bigger* negative effect on 'first person shooter' style video games." Report on a table the results of a specification that involves interacting the *x* variable of interest with 1 or more other regressors. Report the marginal effect of *x* for each group separately and a standard error for it. Label this one "Table 3: Interaction Estimates".

In the write-up:

- Explain why you think this interaction is a relevant test of the robustness of your hypothesis. "Why should 1st person shooters be more adversely affected by ratings guidelines?" "Oh yeah, because they tend to be more violent than other genres of games."
- Does the group with the biggest (absolute value) effect match your hypothesis?
- Are the marginal effects statistically different between/among multiple groups? State a null hypothesis you can test to verify this and report the results.

3e. Diagnostics. Run the B-P and White tests for heteroskedasticity and report the results. They don't necessarily have to be on a table, because the code will be in your *do* file. Report (and explain in the caption) on the table in part (c) robust standard errors if warranted.

Run the RESET to detect functional form misspecification and speculate about a likely cause if it does not pass. Your most saturated specification in part (3c) should include polynomial terms and interactions that, if omitted, would significantly reduce R^2. Your *do* file and your summary of the results should include *F* statistics to confirm the joint significance of these regressors.

Produce and include in the write-up the leverage-residuals plot from the *full-sample* specification with the highest adjusted R squared. Are there any outliers or influential observations that concern you? If so your tables in parts (b-d) should probably exclude this observation and

contain a note in the caption explaining your treatment of outliers. If you decide that the observation(s) should be in the sample, explain your reasoning in the caption.

4. Overall instructions for the write-up. Organize your written summary as follows.
   - 1 page containing: a statement of the causal relationship of interest, answers to the first 2 FAQs, and a summary of the (observational) data source you attempt to use to answer FAQ #3.[3]
   - 1 page containing: the regression model specification in equation form and a written explanation of the variables you will use in your analysis and the units, e.g., individuals or countries, that are observed. This is where you state hypotheses about parameters you will test, too.
   - The descriptive stats table and supporting text. Depending on the size, about 1 page.
   - A figure containing the 2-D scatterplot and simple OLS line.
   - The multiple regression tables (simple OLS as 1st column) and supporting text, statistics, and diagnostics.
   - A brief summary of your results. Have you accurately measured the causal relationship of interest (again it's okay if you're skeptical)?
     - What kind of "natural experiment"[4] would you seek out if you could spend another semester (doesn't that sound fun?) studying this and improving your methods?

As a minimum for a good grade, **the caliber of written communication will befit a college graduate**. A paper that is incomprehensible (because of poor sentence structure, grammar, using words out of context, or subject-verb disagreement, et al.) will earn you no points. I will not (nor will any reader) waste time trying to decipher poorly written paragraphs. I have to read over 100 papers from the class, and **I reserve the right to award a failing grade to any paper that is too hard to read for grammatical or mechanical reasons**.

- If you are concerned about your writing ability, visit the writing center.[5] Get a friend, sibling, or co-worker to read your paper and proofread it. Run spellcheck (!) and search your paper for incorrect homonyms (spellcheck won't find these). Do whatever it takes to avoid handing in a poorly written paper.
- Cite any sources in the text, (Author year) and include a *works cited* page.
- Use active voice.
- Avoid the following phrases: "I think", "I believe", "I feel". You're writing the thing; you wouldn't be writing it if you didn't think it.
- Double space your text.
- Do all the other good things you learned in English composition classes.

Remember it's your job to communicate your thoughts to the reader—not the reader's job to divine what you are trying to say.

[3] "1 page" is a target you should hit if you are being concise and clear, rather than an absolute limit.

[4] An event that is exogenous to the individuals and induces randomness in the *x* variable of interest. E.g., some people live in states that pass laws banning electronic "e-cigs" cigarettes; this alters their calculus of whether to use e-cigs, tobacco cigarettes, or none at all, in a way that has nothing to do with their individual preferences. So some people who would likely continue using e-cigs are induced to stop and can be compared to people in other states that are left to their preferences.

[5] http://owl.english.purdue.edu/writinglab/servicesoverview

On (or before!) **December 14**, students will turn in the following, by uploading 3 files to the Semester Project folder on Blackboard.

1. The 7-8 page (mostly tables and figures) write-up of the project. Has its own folder on Bb and checks for plagiarism; upload in Word (**.doc** or **.docx**) or **.pdf** format.

2. The (cleaned, **.dta** format) data set you used to produce the results.
3. The Stata *do* file containing the commands, in the order they appear in your write-up, that you used to produce the regression estimates, test hypotheses, and run other tests. I should be able to open the data set in Stata and run your *do* file from start to finish without any errors and re-produce your results.

#s 2 and 3 go in the same folder, which allows multiple files per student.

Project Grading Rubric

The instructor evaluates students' papers on the following criteria. Each criterion will receive a "pass" (20 points) "low pass" (10 points) or "fail" (0 points) score, and the grade on the paper is determined by how many criteria pass the evaluation (see next page).

1. Introduction:
    a. Describes a novel and interesting empirical question
    b. Adequately addresses the first 2 "FAQs" in empirical analysis
    c. Clearly explains the data source and unit of observation
2. Description of methods:
    a. Includes a regression model
    b. Clearly explains the variables and units in the model
    c. Clearly states hypotheses that will be tested statistically
3. Tables and figures:
    a. All assigned parts are present
    b. Are well-labeled, well-formatted, easy-to-read
    c. Are self-contained with informative captions
4. Empirical methods/results are:
    a. Correct and applied consistently with in-class examples
    b. Explained clearly and concisely in text form
    c. Accompanied by Stata code, enabling the reader to reproduce the findings
5. Conclusion(s) drawn:
    a. Are supported by appropriate testing
    b. Are consistent with the quantitative results and principles of statistical inference studied in class
    c. Include the practical significance of the results, e.g., elasticity of y with respect to x, when using a log-log model
6. Data set:
    a. Has value added, e.g., intuitive variable names and/or labels, redundant variables dropped, nonnumeric characters (like %) removed
    b. Is cited and enables the reader to locate its original source(s)
7. Written communication (only "pass" or "fail"):
    a. Is coherently organized (as described in the instructions)
    b. Transitions from each idea to next smoothly
    c. Contains minimal proofreading/formatting/grammatical errors
    d. Data and empirical results/methods are described in comprehensible language

Item(s)	Score	
1&2		/40
3-6		/80
7	1	0
1 page write-up (by 9/18/15)		/16
Data set presentation (by 11/13/15)		/24

$$\text{Overall Score} = 1[\text{Item 7 "Pass"}](\text{Score on Items 1\&2}) + (\text{Score on Items } 3-6) + (\text{Points on intermediate steps})$$

The maximum score is 160 points.