PROBLEM SET 3: BIVARIATE REGRESSION

PAUL SCHRIMPF
DUE: JANUARY 27, 2015 AT THE START OF LECTURE
UNIVERSITY OF BRITISH COLUMBIA
ECONOMICS 326

Problem 1 (Wooldridge (2013) Ch2, Q4)**:** The data set BWGHT.RAW contains data on births to women in the United States. Two variables of interest are infant birth weight in ounces (*bwght*) and average number of cigarettes smoked per day by the mother during pregnancy (*cigs*). The following regression was estimated using data on $n = 1388$ births:

$$\widehat{bwght} = 119.77 - 0.514 cigs$$

(i) What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$? Comment on the difference.
(ii) To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.
(iii) The proportion of women in the sample who do not smoke while pregnant is 0.85. Does this help reconcile your finding from (ii)?
(iv) Does this simple regression necessarily capture a causal relationship between the child's birth weight and mother's smoking habits? Explain.

Problem 2 (Wooldridge (2013) Ch2, Q6)**:** Using data from 1988 for houses sold in Andover, MA, from Kiel and McCain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$

(i) Interpret the coefficient on $\log(dist)$. Is the sign of this estimate what you expect?
(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision of where to put the incinerator.)
(iii) What other factors affect houses' prices? Might these be correlated with distance from the incinerator?

Problem 3 (Based on Wooldridge (2013) Ch2, Q9)**:** Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the intercept and slope from the regression of y_i on x_i, using n observations. Let c_1 and c_2 be constants.

(i) Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$ with $c_2 \neq 0$. Show that $\tilde{\beta}_0 = c_1 \hat{\beta}_0$ and $\tilde{\beta}_1 = \frac{c_1}{c_2} \hat{\beta}_1$.
(ii) Now let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$. Show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1$.
(iii) Now let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $(c_1 y_i + a_1)$ on $(c_2 x_i + a_2)$ with $c_2 \neq 0$. Express $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in terms $\hat{\beta}_0$ and $\hat{\beta}_1$

Problem 4 (Angrist (2009) Problem Set 2, Question 3)**:** The Angrist data archive (`http://economics.mit.edu/faculty/angrist/data1/data/anglavy99`) contains data from Angrist and Lavy (1999). This article uses the fact that Israeli class size is capped at 40 to estimate the effects of class size on test scores with an Instrumental Variables/regression Discontinuity research design. But for now, well just use the data to explore regression basics.

R-code to complete much of this problem is posted at `https://bitbucket.org/paulschrimpf/econ326/src/master/problemSets/03/angristLavy.R?at=master` I may update this code in response to questions.
Stata code to complete much of this problem is posted at `https://bitbucket.org/paulschrimpf/econ326/src/master/problemSets/03/angristLavy.do?at=master`

(i) Read the article through Section I (at least), download the data, and construct the descriptive stats in Table 1 for 5th graders. From here you should be able to mostly tell whats what as far as variable names go (note that the unit of observation is the class average). Note that enrollment is called c_size and percent disadvantaged in called tipuach. To exactly reproduce the numbers in table 1, you must follow footnote 11 and restrict the sample to schools with enrollment of at least 5 and classes of size less than 45. There are also a couple of non-obvious data corrections. There is an avgmath score and an avgverb score greater than 100 due to a data entry error. The correct values of these scores are 87.606 and 81.246 (not 187.606 and 181.246). Finally, there is a non-missing math score for an observation with mathsize==0 (i.e. no math test takers). This is impossible. Replace avgmath=NA if mathsize==0.

(ii) Economists and educators have long debated whether it's worth paying the extra labor costs required to reduce class size. What should the sign of the achievement/class-size relationship be if the investment is worthwhile? Regress average math and verbal scores on class size. What is the sign of this relationship? Is it significantly different from zero? How does it look so far for the class size optimists?

References

Angrist, J.D. and V. Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* 114 (2):533–575. URL `http://qje.oxfordjournals.org/content/114/2/533.short`.

Angrist, Joshua. 2009. "14.32 Econometrics." Unpublished course material.

Angrist, Joshua, James Berry, and Emre Kocatulum. 2007. "14.32 Econometrics." Massachusetts Institute of Technology: MIT OpenCourseWare. URL `http://ocw.mit.edu/courses/economics/14-32-econometrics-spring-2007/`.

Wooldridge, J.M. 2013. *Introductory econometrics: A modern approach*. South-Western.