MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2015 **Luu Nguyen(hqn@mit.edu)**

Problem Set 1

Due: Thursday, Feb. 19

A. From Stock and Watson: 2.20, 2.23, 2.24, 3.10, 3.13, 3.15

B. Additional problems

1. Let $\bar{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider two estimators of μ: $W_1 = [(n-1)/n]\bar{Y}$ and $W_2 = \bar{Y}/2$.

   (a) Show that W_1 and W_2 are both biased estimators of μ.

   (b) Are these estimators consistent?

   (c) Compare standard errors and MSE. Which estimator is more precise? Does precision carry the day in this case?

   (d) Argue that for some values of μ, the more biased estimator has lower MSE.

2. Let Y denote a Bernoulli (θ) random variable with $0 < \theta < 1$. Suppose we are interested in estimating the odds ratio, $\gamma = \theta/(1-\theta)$, which is the probability of success over the probability of failure. Given a random sample $\{Y_1, ..., Y_n\}$, we know that an unbiased and consistent estimator of θ is $\bar{Y}$, the proportion of successes in n trials.

   (a) A natural estimator of γ is $\widehat{\gamma} = \bar{Y}/(1-\bar{Y})$. Show that $\hat{\gamma}$ is a consistent estimator of γ.

   (b) (extra credit) Is $\hat{\gamma}$ unbiased? Why or why not?

3. Table 3 in Woodbury and R.G. Spiegelman (1987) reports the results of two social experiments meant to encourage Unemployment Insurance (UI) recipients to return to work. In the Employer Experiment, any UI recipient finding employment for at least 4 months received a voucher worth $500 to his or her employer. In the Claimant Experiment, any UI recipient finding employment for at least 4 months received $500 directly. [*Note: Woodbury and Spiegelman (1987) is posted on Stellar.*]

   (a) For each experiment, test the hypothesis that bonuses decreased the proportion of UI claimants who exhausted their benefits. Compute the test statistic under two scenarios: (i) the experiment has no effect and (ii) the experiment has an effect.

   (b) For each experiment, pick a significance level and test the hypothesis that the experiment reduced weeks of insured unemployment in the first spell using a one-tailed and two-tailed test. Which test seems to make more sense in this case?

4. This problem asks you to conduct a series of sampling experiments.

   (a) Draw 500 random samples of size 8 from a random number generator for a standard normal distribution. Then increase the sample size to 32. Finally, increase the sample size to 128. Plot histograms of the sampling distributions of (i) the sample mean and (ii) the sample variance, for each of these three sample sizes. Now repeat your experiments for three samples drawn from another parametric distribution of your choice (e.g., a uniform distribution). Discuss the results of your experiments in light of the central limit theorem.

   (b) Your experiments produce "samples of sample means." Compute the mean and variance of the sample means generated by each experiment and compare them to the mean and variance predicted by statistical theory. Does the variance of the sample means (i.e., the sampling variance) decrease with sample size at the rate predicted by the theory? Does Normality matter for this?

C. Theory

1. Prove that the properties in Section 2.4 of LN1 hold in samples.
2. Prove the claims in Section 2.5 of LN1.

D. Stata exercise with NHIS Data

Table 1.1 in MM compares the health and demographic characteristics of insured and uninsured couples in the NHIS. Panel A compares the health across husbands in this sample with and without health insurance.

1. Calculate the t-statistic for the null hypothesis that there is no difference between the health of husbands with and without health insurance in this sample. Is the difference significantly different from zero?
2. Panel B of Table 1.1 shows that husbands with and without health insurance differ along many demographic dimensions. It is possible that the difference in health between the "Some HI" and "No HI" groups may be smaller if we compare across groups that are more homogeneous. To investigate this, go to http://masteringmetrics.com/resources/ and download the Stata data and .do file used to produce MM Table 1.1. Execute the Stata code in NHIS2009_hicompare.do through line 35 to make sure that you use the same selection criteria that were used to produce Table 1.1.

   Is the difference between the health of husbands with Some and No HI significantly different from zero if you restrict to men who:

   (a) are employed?
   (b) are employed and have at least 12 years of education?
   (c) are employed, have at least 12 years of education, and earn income of at least $80,000?

MIT 14.32
Spring 2015

J. Angrist (angrist@mit.edu)
Luu Nguyen (hqn@mit.edu)

Problem Set 2

Due: Thursday, March 5th

A. From Stock and Watson: 4.5, 4.10, 5.5, 5.10

B. Additional problems

1. Consider the simple regression model $y = \alpha + \beta x + \epsilon$.
   (a) Derive the formulas for the OLS estimates of the slope and intercept in two ways
      i. Solve the sample least squares problem
      ii. Substitute sample moments for population moments in the formula for α and β
   (b) Show that these estimates are unbiased
2. Suppose X is a dummy variable that equals one with probability p and is zero otherwise. Prove that the CEF and the regression of Y on X are the same in this case. Why is the regression of Y on X the same as $E[Y|X]$ when the regressor is a dummy variable?
3. The 14.32 web page contains a Stata data set with data from the March 2013 CPS. The data set contains observations on annual earnings, weeks worked last year, usual hours worked/week, age, race, and sex for men and women aged 25-59.
   (a) List the variables in the data set and use Stata's *sum* command to produce descriptive stats. Construct a measure of average weekly earnings and average hourly earnings and provide descriptive stats for these as well.
   (b) Construct a t-test and 95 percent confidence intervals for the difference in the log of average hourly earnings (AHE) by sex for white men and women aged 30-39. What is the approximate percentage difference in AHE between men and women?
   (c) Use Stata's reg command to prove the result in question B.2 above "by computer."
   (d) Stata's testing procedure calculates the standard error of differences in averages under two alternative assumptions about variances. State these assumptions in words. Which calculation corresponds to the default for regression? Explain and check your answer.
   (e) Calculate the sex differential in AHE for a sample of men and women in their 40s. How does this compare with the sex differential in the younger sample (estimated in part b). What might explain the change in sex differentials?
   (f) Regress wage on age, separately for men and women in their 30s. Whose age-earnings profile is more steeply sloped?
4. The RAND Health Insurance Experiment (HIE)
   (a) What causal questions was the RAND HIE designed to answer?
   (b) Download the Stata data associated with Tables 1.3 and 1.4 in MM from the MM Resources page. The "person_years.dta" dataset contains information on the RAND HIE sample, including demographic characteristics and treatment assigned. The "annual_spend.dta" dataset contains information on annual hospital expenditures. To link these together, merge "person_years.dta"

with "annual_spend.dta" using the variables *person* and *year*. Keep only those person/year observations that appear in both datasets.

Generate a variable for total hospital spending, equal to the sum of dollars spent on inpatient care (*inpdol*) and outpatient care (*outsum*). Calculate the difference in average hospital spending between people who report being in excellent health (*exc_health*) versus those who report being in bad health (*bad_health*). Is this difference statistically significant at the 5% level?

(c) As described in MM Chapter 1, the RAND HIE had many small treatment groups - in fact, the variable *plan* in your dataset shows that there were 24 different groups. Define a new variable "plantype" that divides these into 4 larger categories as follows. Plan Type 1 ("Free") is plan 24; Plan Type 2 ("Individual Deductible") is plans 1 and 5; Plan Type 3 ("Cost-Sharing") is plans 9-23, inclusive; and Plan Type 4 ("Catastrophic") is plans 2-4 and plans 6-8, both inclusive. What is the average hospital spending in each group? Is the difference in hospital spending between Plan Types 1 and 4 significant at the 5% level?

(d) Clear your Stata session and read in "rand_initial_sample_2.dta". The four plan types have already been defined in this dataset, which also contains the variable *ghindx*, a general health index. Is the difference in the average health between Plan Type 1 and 4 significant at the 5% level? How do your results from parts (c) and (d) relate to the HIE findings discussed in MM chapter ?

5. Regression application: the effects of class size

The Angrist data archive (http://econ-www.mit.edu/faculty/angrist/data1) contains data from the following article (posted on Stellar): J. Angrist and V. Lavy, "Using Maimonides Rule to estimate the Effect of Class Size on Student Achievement," The Quarterly Journal of Economics, May 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, we'll use the data to explore regression basics.

(a) 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 what's 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 average math (*avgmath*) score and an average verbal (*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=. if mathsize==0.

(b) Economists and educators have long debated whether it's worth paying the extra labor costs (i.e., teachers' wages) 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?

(to be continued)

MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2015 **Luu Nguyen (hqn@mit.edu)**

Problem Set 3

Due: Tuesday, March 17th

A. From Stock and Watson: 5.3, 6.6

B. Additional problems

1. Let

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

denote a bivariate regression of y_i on x_i,with regressor residual ε_i. Derive a formula for the sampling variance of the estimated regression *intercept* (assuming the regressors are fixed in repeated samples).

2. Biddle and Hamermesh (1990) study the determinants of time spent sleeping (outside of the classroom!). Using a sample from the 1975-1976 Time Use Study, they run regressions like this:

$$sleep = \beta_0 + \beta_1 totwork + \beta_2 educ + \beta_3 age + u$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* (education) and *age* are measured in years. Means and standard deviations of these variables are as follows:

Variable	Mean	Standard Deviation
sleep	3266	444
totwork	2122	947
educ	12.8	2.8
age	38.8	11.3

   (a) What would you expect such a regression to show and why?

   (b) Suppose sleep and work are measured in hours. How would this affect the work and education coefficients?

   (c) A friend who took 14.32 last year proposes to add work experience to the model. His experience variable is defined as years since graduation. Does the addition of such a variable to this model make sense?

   (d) One set of estimates looks like this:

$$sleep = 3638.25 - .148\, towork - 11.13\, educ + 2.20\, age + \varepsilon;\ n = 706,\ R^2 = .113$$

   Interpret the coefficients on work and education.

   (e) Sleep is negatively correlated with both hours of work and schooling. Are these likely to be causal relationships? Explain.

3. Regression in practice

   (a) Reload the NHIS data you used in Part D of Problem Set 1. Again, execute the Stata code in NHIS2009_hicompare.do through line 35 to make sure that you use the same selection criteria that were used to produce MM Table 1.1.

      i. Use the *sum* command to calculate average health separately for husbands with and without health insurance. What is the difference in average health by insurance status? Is this difference statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

ii. Use the NHIS data to construct a variable such that a regression of health on this variable reproduces the difference calculated in quesiton (i), above. Compare the difference, t-statistic, and confidence interval for your regression estime of differences in health with those you computed in (i).

(b) In Part D of Problem Set 1, we showed that some of the difference in average health between those with and without health insurance in the NHIS can be attributed to the fact that the insured differ from the uninsured along many relevant dimensions. We can also show this using regressions. Starting with your regression from part a.ii above, sequentially add controls for age (*age*), years of education (*yedu*), and income (*inc*). Does any set of controls eliminate the difference in health between insured and uninsured? Explain how the results change as you add controls and what changes in the estimates as you add more controls might mean.

(c) Reload the RAND HIE dataset "rand_initial_sample_2.dta" used in Question B4 of Problem Set 2. Define a dummy variable called *anydum*, which is equal to 1 for individuals with Plan Types 1-3 ("any insurance") and equal to 0 for individuals with Plan Type 4 (only "catastrophic" insurance). Regress the general health index *ghindx* on a dummy for any insurance (as a reminder, *ghindx*, is a general health index similar to that in the NHIS, but scaled differently).

i. Interpret your estimates of this model.

ii. Sequentially add controls for age (*age*), education (*educper*), and income (*income1*). Do these controls have much of an effect on your estimates? Why is the effect of adding these demographic controls so different from what you saw in question 4b? (Hint: Think about Section 1.3 of LN7 - Regression Anatomy - and the differences between the NHIS and the RAND HIE data.)

4. More on class size

(a) Go back to the class size data in problem set 2. To establish a benchmark, start by replicating your earlier results, that is, run a bivariate regression of test scores on class size.

(b) A possible concern with the bivariate regression of test scores on class size is that bigger schools have bigger classes and also better students. Check this by adding enrollment (in the grade, measured at the school level) to your regressions.

(c) Construct the correlation matrix of test scores (math or verbal, your choice), class size, and enrollment. Use this matrix and your regression results to explain why and how the coefficient on class size changes when you add enrollment controls. (Hint: This requires an application of the omitted variables bias formula)

(d) Add the percent of students who came from disadvantaged backgrounds (PD) instead of enrollment. How does this affect the class size coefficient? Use the correlation matrix for test scores, class size, and PD, along with the correlation matrix in Part (c) above and your regression results, to explain why the coefficient changes more here than when you added enrollment in response to Part (b).

(e) Estimate the effects of class size in a model that includes both PD and enrollment controls. Is the class size coefficient here closer to the one in Part (c) or Part (d)? Why?

(f) All told, does the analysis in this question suggest that smaller classes are good, bad, or neutral?

(to be continued...)

MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2015 **Luu Nguyen (hqn@mit.edu)**

Problem Set 4

Due: Thursday, April 9

A. From Stock and Watson: 6.8, 6.9, 6.10 (a-b); 7.9

B. Additional problems

1. Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$ for $i = 1, 2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$ for $i = 1, 2, \ldots, n$ are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. (Hint: Use the fact that the $\tilde{\beta}_j$ satisfy OLS first order conditions for the rescaled dependent and independent variables.)

2. Consider the regression model $\ln Y_i = \alpha + \beta C_i + \epsilon_i$ where Y_i is annual income and C_i is a dummy for being a college graduate.

   (a) Solve for Y_i in the above expression, differentiate with respect to C_i, and derive an expression for β that shows why the slope coeffiicent in the "semilog model" can be interpreted as the percent increase in annual income associated with being a college graduate.

   (b) Consider instead $\ln Y_i = \gamma + \rho \ln X_i + \nu_i$ where Y_i is annual income and X_i is parental income. Solve for Y_i, differentiate with respect to X_i, and derive an expression for ρ. What do economists call parameters like ρ?

   (c) Take the original model and modify as follows:

   $$\ln Y_i = \alpha + \beta C_i + \delta\left(C_i \times F_i\right) + \epsilon_i$$

   where F_i is a dummy for being female. Interpret the parameters in this case in terms of differences in means.

3. Here are some estimates of the relationship between a firm's research and development expenditures and its sales (measured in millions of dollars). These are from a sample of 32 firms:

   $$\hat{rdintensity} = 2.613 + .00030\, sales - .0000000070\, sales^2;\ n = 32,\ R^2 = .1484$$

   $$(0.429) \quad (.00014) \quad (.0000000037)$$

   (a) Define *salesbil* as sales measured in billions of dollars so that $salesbil = sales/1000$. Rewrite the estimated equation and coefficients with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the R-squared.

   (b) Is the quadratic term in this model important? Explain.

   (c) At what point does the marginal effect of sales on $\hat{rdintensity}$ become negative?

4. The 14.32 Stellar page contains a CSV data set (PS4.csv) with observations on log weekly wages, log hourly wages, age, sex (1=male), race (1=White, 2=Black, 3=Native American, 4= Asian or Pacific Islander, 5=Other), and years of schooling for men and women aged 25-50 in the March 1992 CPS.

   Besides schooling, a variable that figures importantly in the analysis of individual wage rates is labor market experience. Wages generally increase as we get more experience working, though not necessarily at a constant rate.

   (a) Construct a measure of potential work experience by defining:

   $$experience = age - education - 6$$

   What is the rationale for this? Why does this measure potential experience and not actual experience?

(b) Check the distribution of your constructed potential experience variable to see whether the values make sense. Set any implausible values to missing, so that they will be excluded from the statistical analysis for the remainder of the question.

(c) Compute the multivariate regression of log hourly wages on: sex, a full set of race dummies, a quadratic function of potential work experience, and years of schooling.

(d) As we discussed in class, check your regression by computing the multivariate schooling coefficient manually in two steps: (i) regress schooling on all the other covariates and save the residuals (ii) regress the dependent variable on the residuals.

(e) Plot the estimated experience profile, and use calculus to compute the level of experience at which earnings reach their peak, according to this model. How old is a high school graduate who reaches this level of experience?

(f) Re-estimate the model allowing the relationship between potential experience and wages to differ by sex. Construct an F-test for the null hypothesis that the relationship between potential experience and wages is the same for men and women. What happens to the coefficient on the sex dummy when the effects of potential experience differ by sex? What do you think is going on here?

(g) Re-estimate the model in part (c), allowing the relationship between wages and schooling to differ for Blacks in a sample limited to Blacks and Whites. Construct an F-test to test the null hypothesis that the returns to an additional year of schooling are the same for both racial groups.

(h) Re-estimate the model in part (c) with a full set of interactions between race (Black/non-Black) and sex. How many new variables do you need to add to your regression? Using your coefficients, calculate the expected log hourly wage in each race-sex category for a 25-year-old worker with 12 years of schooling.

5. Twins and the returns to schooling.

(a) Consider the regression model $\ln Y_i = \alpha + \beta S_i + \epsilon_i$ where Y_i are wages and S_i is years of schooling. What does "ability bias" refer to in this context? Isβ likely to be an under- or over-estimate of the economic returns to schooling?

(b) Go to http://dataspace.princeton.edu/jspui/handle/88435/dsp01rv042t084 and download the pub-twins.dta file. These are data on individual-level wages and schooling for 340 twin pairs, interviewed at an annual Twins Days festival in Twinsburg, Ohio (there are 680 total observations). As a baseline, regress log wages (*lwage*) on years of schooling (*educ*), age (*age*), age-squared (*age2*), and dummies for female (*female*) and white (*white*). Interpret the estimated schooling and age coefficients.

(c) Consider the long regression $\ln Y_{if} = \alpha' X_{if} + \beta S_{if} + \gamma A_f + \epsilon_{if}$ where f stands for family and subscript $i = 1, 2$ indexes twin siblings. The vector X_{if} includes the covariates from part (b), while A_f is an unobserved ability measure assumed to be fixed within families. Show (i.e., derive on paper, not in the data) that a model where the dependent variable is the difference in log wages between twin 1 and twin 2 eliminates family-specific ability bias.

(d) Regress the difference in wages between twins (*dlwage*) on their difference in education (*deduc*). be careful with your data handling - you should have only one obs per twin pair. Interpret the coefficient on *deduc*. What do your results suggest about the direction of ability bias in the undifferenced model? What should the constant be in this model?