MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2021 **Juliette Fournier (jfournie@mit.edu)**

Problem Set 4

Due: Thursday, April 15

I. Regression Theory

1. Consider the regression model:
$$\ln Y_i = \alpha + \rho C_i + \gamma X_i + \varepsilon_i, \tag{1}$$
where Y_i is worker i's weekly earnings at age 40, C_i is a dummy for being a college graduate and X_i is $i's$ family income when he or she was aged 16.

   (a) Show that ρ can be interpreted as measuring the percent change in Y_i as a function of C_i, conditional on X_i.

   (b) Consider a version of (1) that replaces X_i with $\ln X_i$. How should γ be interpreted in this case?

   (c) Consider
$$\ln Y_i = \alpha + \beta_1 C_i + \beta_2 W_i + \beta_{12}\left(C_i \times W_i\right) + \gamma_1 X_i + \gamma_{12}(X_i \times W_i) + \varepsilon_i$$
   where W_i is a dummy for women. Assuming this regression model describes $E[\ln Y_i | C_i, W_i, X_i]$, explain how the model parameterizes the effects of college graduation and family income on average wages for men and women.

2. For bivariate regression model, $Y_i = \alpha + \beta X_i + \varepsilon_i$, define the regression *fitted values*, $\hat{Y}_i = \alpha + \beta X_i$, and note that $Y_i = \hat{Y}_i + \varepsilon_i$.

   (a) Prove that $E[\hat{Y}_i \varepsilon_i] = 0$, that is, fitted values and residuals are uncorrelated.

   (b) Use this to show that the variance of Y_i can be written as the sum:
$$V(Y_i) = V(\hat{Y}_i) + V(\varepsilon_i). \tag{2}$$
   It's customary to say that $V(\hat{Y}_i)$ is the "explained" variance associated with a particular regression model, while $V(\varepsilon_i)$ is the "unexplained" or *residual variance* for this model. Equation (2) is the regression version of the ANOVA formula reviewed at the start of the course (LN8 covers this point in detail; see also the appendix to MM Chpt 2).

   (c) Define the *estimated fitted values* $\hat{Y}_i^* = \hat{\alpha} + \hat{\beta} X_i$ where $\hat{\alpha}$ and $\hat{\beta}$ are OLS estimates of the bivariate slope and intercept. Likewise, define *estimated residuals* $e_i = Y_i - \hat{Y}_i^*$. Prove that $\sum \hat{Y}_i^* e_i = 0$ in the sample used to compute OLS estimates. Use this fact to show that ANOVA holds in samples, i.e., that:
$$s_Y^2 = s_{\hat{Y}^*}^2 + s_e^2,$$
   in your data.

3. Suppose the CEF of Y_i given X_i is linear:
$$E[Y_i | X_i] = a + bX_i. \tag{3}$$
We know from the regression-CEF theorem that $b = \beta = \frac{C(Y_i, X_i)}{V(X_i)}$ and $a = \alpha = E[Y_i] - \beta E[X_i]$. Typically, we estimate β with the OLS estimator, $\hat{\beta}_{OLS} = \frac{s_{XY}}{s_X^2}$. But there are many ways to fit a line. Here's one: split the data in half by dividing the sample into observations with values above and below median X_i. Compute above-median and below-median average Y_i and X_i; call these $\bar{y}_1, \bar{x}_1$ for means above and $\bar{y}_0, \bar{x}_0$ for means below. Finally, define an alternative slope estimator,
$$\hat{\beta}_w = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}.$$
($\hat{\beta}_w$ is called a Wald estimator, after Abe Wald, who first thought of it in 1940)

(a) Treating the X_i as fixed in repeated random samples, show that $\hat{\beta}_w$ is an unbiased estimator of the regression slope.

(b) Treating the X_i as fixed in repeated random samples and assuming homoskedastic residuals, derive a formula for the sampling variance of $\hat{\beta}_w$ as a function of the variance of residuals (analogous to the classical formula for the sampling variance of $\hat{\beta}_{OLS}$).

(c) Continue to assume the X_i are fixed in repeated random samples and residuals are homoskedastic. Is $\hat{\beta}_{OLS}$ or $\hat{\beta}_w$ a more precise estimator of β? (hint: this is the consequence of a well-known theoretical result. No actual math needed.)

(d) More challenging: Using the same assumptions as in q3c, compare the sampling variance of $\hat{\beta}_w$ and $\hat{\beta}_{OLS}$ directly, rather than relying on the theorem you used for question (c) (hint: this requires math! Start by showing that the sampling variance of $\hat{\beta}_w$ is inversely proportional to the square of the denominator of $\hat{\beta}_w$. Next, note that the term $(\bar{x}_1 - \bar{x}_0)^2$ is proportional to the variance of fitted values from a regression of X_i on a dummy that indicates values of X_i above the median. Finally, complete the argument using the variance decomposition for regression established in q2c, above.)

II. Regression Practice

1. The Canvas tab for Pset4 contains a CSV file (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.

   (a) Labor economists often model wages as a function of *potential experience*, which is a crude calculation of work experience that adjusts for time out of the labor force while in school. Potential experience is usually defined as $potex = age - years\ of\ education - 6$. Use PS4.csv to compute *potex* and check its distribution. Set implausible values to missing, or to a plausible value that seems consistent with the underlying data.

   (b) Limit the sample to men and regress log weekly wages (ln_uwe) on potential experience and its square, along with years of schooling.

      i. Is the quadratic term significantly different from zero? What do we learn from it?

      ii. Re-estimate this model with robust standard errors - does this change your conclusions?

   (c) Labor economists call the relationship between log wages and potential experience the *experience profile*. Use Stata's `graph twoway function` command to plot the experience profile. At what experience level do wages peak? Explain how to use your regression estimates to compute this value.

   (d) In nonlinear models, econometricians refer to the average of the derivative of the conditional mean function with the respect to an independent variable as this variable's *marginal effect*. The model linking potential experience with wages in q1b is nonlinear. Compute the marginal effect of potential experience. Use Stata to compute a standard error for this, treating the value of average experience in the marginal effects formula as non-random.

   (e) (More challenging) Use Stata to obtain a standard error for the estimated peak-earnings age computed in q1c, above. How precisely is this estimated? (Hint: you'll need to find a Stata routine for the calculation of standard errors of nonlinear functions of regression estimates)

2. Add women back to your extract from PS4.csv. Focus here on hourly wages (ln_ahe).

   (a) Estimate a version of the model explored in q1b above that allows the relationship between schooling and (log) hourly wages to differ for men and women, including a female main effect (i.e., allow the intercept to differ for men and women). Use this model to test the hypothesis that the returns to schooling are the same for men and women.

(b) Estimate a version of the model explored in q1b above that allows the relationship between potential experience and (log) wages to differ for men and women.

   i. What happens to the female main effect when experience returns differ by sex? Many labor economists have commented on this pattern. Why is this of economic interest?

   ii. Use this model to construct an F-test for the joint null hypothesis that the relationship between potential experience and wages is the same for men and women. How many restrictions are you testing?

(c) Estimate a version of this model allowing schooling coefficients to vary freely by race and freely by sex, but not by race within sex. Maintain a common experience profile for all. Use an F-test to evaluate the joint null hypothesis that schooling coefficients are the same for all racial groups, allowing for differences by sex. How many restrictions are you testing here?

(d) Test the joint null hypothesis that the regression of wages on potential experience and schooling is the same for men and women, after allowing for a female main effect (that is, retain a dummy for women under both null and alternative).