**Department of Sociology and Anthropology**

**Carleton College**

Statistical Tools for Quantitative Reasoning Peter D. Brandon

Soan 280 Leighton Hall 229

Winter Term 2010 Phone: 222-7199

Email: pbrandon@carleton.edu

**ASSIGNMENTS**

The goals of assignments one through four are to help you gain firsthand experience in constructing and interpreting coefficients on dummy variables and on interaction terms, to enable you to discover by yourself what factors affect individuals' wages and earnings, to estimate rates of return to schooling, to calculate union-nonunion wage differentials, and to compute measures of discrimination by race and by gender.

**ASSIGNMENT 1: UNDERSTANDING YOUR DATA**

The purpose of this assignment is to help you become familiar with features of the data in CPS78 and CPS85. In examining the data you will compute and compare arithmetic means, geometric means, and standard deviations for the entire sample and for the data sorted into various subgroups.

1. Using the data in CPS78, compute the arithmetic mean and standard deviation of LNWAGE. The geometric mean is calculated by first computing the arithmetic mean of LNWAGE, and then exponentiating it. Compute this geometric mean wage rate, in dollars. Assuming that a work year consists of 2000 hours, what is the implied geometric mean of annual wages? Exponentiate LNWAGE, and then compute its mean. Compare this arithmetic with the geometric mean wage rate. What are the arithmetic means and standard deviations of years of schooling (ED) and years potential experience (EX)?
2. The dummy variable NONWH in CPS78 equals 1 if the individual is nonwhite and non-Hispanic (the vast proportion of NONWH individuals are black). What is the sample mean of NONWH? Since there are 550 observations in CPS78, how many individuals in your sample are nonwhite and non-Hispanic? Another dummy variable is HISP, and it equals 1 if the individual is nonwhite and is of Hispanic origin. What is its sample mean, and how many individuals in your sample are Hispanic? Finally, the dummy variable FE equals 1 if the individual is female. What proportion of the sample is female? How many individuals are female?
3. A common procedure used to inspect a data set is to sort it into subgroups by certain variables and then compute and compare means and standard deviations for the various subgroups. Sort your data in CPS78 by gender. What are the mean and standard deviation of years of schooling (ED) for males and for females? The geometric mean wage rates? Next sort your data by race into whites, nonwhites and non-Hispanics (NONWH), and Hispanics (HISP). For each of these groups, compare the arithmetic mean and standard deviation of ED and the geometric mean and standard deviations of the wage rate. Whose LNWAGE has the largest dispersion, as measured by the standard deviation?
4. Now repeat parts (a) through (c), using the 1985 data in CPS85. (Note: There are 534 observations in CPS85 and 550 in CPS78.) Are there substantial changes in the sample proportions of individuals by gender or race in 1985 as compared to 1978? Using a consumer price index deflator that equals 1.000 in 1978 and 1.649 in 1985, compute and compare the geometric mean of real wage rates in 1978 and 1985, for (i) the entire sample, (ii) by gender, and (iii) by race. Which subgroups appear to have attained higher real wages? Have any subgroups suffered real wage losses? Finally, compare years of schooling by gender and by race for 1978 and 1985. Are changes in years of schooling and mean geometric wages broadly consistent with the human capital model for the various subgroups? Why or why not?

**ASSIGNMENT 2: Confirming Relationships among Alternative Dummy Variable Specifications: Earnings and Returns to Schooling**

The purpose of this assignment is to have you confirm empirically the relationships that theoretically should occur among alternative simple dummy variable specifications in the linear regression model. This exercise involves examining least squares coefficients in a stylized human capital model where LNWAGE is affected by years of schooling (ED) with a slope coefficient that is identical for everyone but where the intercept terms differ by union status (UNION = 1 if the individual works on a union job).

Choose either the 1978 (CPS78) or the 1985 (CPS85) data set. Employ this data set for each section of this exercise.

1. Using least squares, estimate parameters in a simple model where LNWAGE is regressed on a constant term and years of schooling (ED). What is the value of the estimated slope coefficient? When can this coefficient be interpreted as the rate of return to schooling? Construct and interpret a 95% confidence interval for this coefficient. What is the *R2* in this model? Interpret this *R2,* and compare it (as well as the estimated rate of return to schooling) to that reported by Mincer in his classic book (1974) based on 1959 data. (See Mincer, J. 1974. Schooling, Experience, and Earnings, New York: Columbia University Press for the National Bureau of Economic Research.)

(b) The specification in part (a) assumes that the intercept and slope coefficients are identical for all individuals. Now assume instead that the effect of schooling on wages differs between union and nonunion members by a constant factor of proportionality; that is, for nonunion members WAGE = αneβEDei whereas for union members, WAGE = αUeβEDei where WAGE is the wage rate, αnand αU are differing factors of proportionality, β is the common rate of return to schooling, and є is a random disturbance term. Show that this implies that when LNWAGE rather than WAGE is the dependent variable, union and nonunion workers have different intercept terms αnand αU but a common slope coefficient β. Now estimate by least squares parameters in a regression model

LNWAGE = α+ α*'* • UNION + β ED + є

where UNION is the dummy variable equal to 1 if the individual works at a union job and є is a random disturbance term. Interpret your estimates of αnand αUrelating α and α'*'* • to αnand αU above. Formulate and test the null hypothesis that intercept terms for union and nonunion individuals are identical, using a 5% level of significance. What happens to the estimated rate of return to schooling when union and non-union individuals have different intercepts? Can you interpret this?

(c) An alternative procedure for formulating the regression relationship in (b) is as follows. First, create a new nonunion dummy variable called NONU, where NONU = 1.0 - UNION. (Note that for each individual, NONU + UNION = 1.0. Why is this important?) Then estimated parameters in an equation without a constant term:

LNWAGE = αo NONU + α1 UNION + β ED + є

Interpret αo and α1  and relate them to αn and αU in part (b). What should the relationship be among estimates of αo and α1 and of α and α' (see part (b))? Why? Are your estimates numerically consistent with this relationship? Why or why not? Formulate and test the null hypothesis that intercept terms for union and nonunion workers are identical. Your test result should be precisely equivalent to that in part (b). Is this the case? Compare the estimated of return to schooling and R2 here with that obtained in part (b). Why does this occur?

(d) A seemingly plausible specification is the equation

LNWAGE = α\* + α\*o NONU + α\*1 UNION + β ED + є

in which LNWAGE is related to a common intercept term α*\*,* to separate terms α\*o and α\*1 for nonunion and union members, and to ED by a common coefficient β*.* What is wrong with this specification? Why is least squares unable to estimate parameters of this model uniquely? What happens when you try to estimate this model using STATA? Interpret this phenomenon.

(e) One other logical possibility is to estimate parameters in an equation in which the NONU dummy variable in part (b) replaces the UNION dummy variable. Such an equation would be of the form

LNWAGE = α0 + α1 NONU + α\*1 UNION + β ED + є

Interpret the regression estimates of α0, α1, and β in this equation. What do you expect to be the relationship among estimates of α0, α1, and β here with those of α, α', and β in part (b)? Why? Verify your conjectures empirically by estimating the above equation.

**ASSIGNMENT 3: Dummy Interaction Variables: The Earnings of Single and Married Males and Females**

The purpose of this exercise is to help you construct and interpret dummy involving interaction terms. The empirical implementation focuses earnings of single and married males and females. To the extent that married females differ from single females and from married or single males in their frequency of leaving the labor force temporarily to rear children, other things being equal, the earnings functions of married females might differ from those of married or single males or from single females. In this exercise you take a preliminary look at this empirical issue, using dummy interaction variables. Choose either the 1978 (CPS78) or the 1985 (CPS85) data set. Employ this data set for each section of this assignment.

1. Begin by estimating a model allowing for no differences by gender or marital status. In particular, using least squares procedures, estimate parameters in the simple, stylized “human capital model”:

LNWAGE = α + β1•ED + β2•EX + β3•EXSQ + є

where EX and EXSQ are potential experience (age minus years schooling minus 6) and potential experience squared. Are your estimates plausible? Why?

1. Now allow for intercepts to vary by gender. The dummy variable FE is 1 if the individual is female and 0 if the individual is male. Using least squares, estimate parameters in the expanded equation

LNWAGE = α + αF FE + β1•ED + β2•EX + β3•EXSQ + є

Interpret these parameter estimates. Using a reasonable level of significance, formulate and test the null hypothesis that there are no differences in intercept terms for males and females.

1. Next examine whether marital status affects the earnings function in a linear manner. The dummy variable MARR equals 1 if the individual is married with spouse present or in the military and otherwise is O. Using least squares, estimate parameters in the equation

LNWAGE = α + αF FE + αM MARR + β1•ED + β2•EX + β3•EXSQ + є

Interpret estimates of α, αF,and αM*.* Show that the effect of marital status on LNWAGE in this specification, given ED, EX, and EXSQ, is the same for males and females. Within this specification, formulate and then test the null hypothesis that marital status has no effect on LNWAGE.

1. As was noted above, the effect of marital status on earnings might be expected to depend on gender; in particular, being married may have a larger negative effect on earnings for females than for males, *ceteris paribus.* The restrictive specification in part (c) does not allow you to test this hypothesis. Why? Construct an interaction dummy variable for each of the *i* individuals in your sample that takes on the value of unity if the individual is married with spouse present or in the military and is female. For each spouse present or in the military *and* is female. For each of the *i* individuals this variable is constructed as INFMARi =; FEi \* MARRi. Generate the dummy interaction variable INFMARi. INFMAR should be identical to the dummy variable in your data set entitled MARRFE. To verify this, construct a temporary variable called CHECK, defined as CHECKi =; INFMARi - MARRFEi, for *i* = 1, ... , *n.* If you have properly constructed INFMARi, CHECK should equal 0 for all observations. Print out CHECK for the entire sample, and verify that you have properly constructed this dummy interaction variable. Comment on why INFMARi is an interaction variable rather than a simple additive dummy variable.
2. Finally, using least squares procedures, estimate parameters in the equation with both additive and interaction dummy variables:

LNWAGE = α + αF FE + αM MARR + αMF INFMAR + β1•ED + β2•EX + β3•EXSQ + є

Interpret the parameter estimates of α, αF,αM*,* and αMF*.* Are they consistent with the argument noted above? Why or why not? Formulate and test the null hypothesis that the effect of marital status does not depend on gender. How do you interpret this finding?

**ASSIGNMENT 4: Measuring and Interpreting Wage Discrimination by Race & GENDER**

The purpose of this exercise is to have you implement empirically and then interpret three alternative procedures for measuring wage discrimination by race and gender. You need to conduct a Chow test for parameter equality. Choose either the 1978 or the 1985 data set in CPS78 and CPS85, respectively, and use this data set for all portions of this exercise.

1. Begin by estimating the effects of wage discrimination by race using simple dummy variable procedures. Specifically, using OLS, estimate parameters in the equation

LNWAGE = α + αF•FE + αu•UNION + αN•NONWH + αH•HISP + β1•ED + β2•EX + β3•EXSQ + є

Interpret the estimated coefficients αN and αH. Test the null hypothesis that, gender, union status, education, and experience being held fixed, racial status has no effect on log wages. Then test and interpret the null hypothesis that the coefficients on NONWH and HISP are equal to each other but are not necessarily equal to zero.

1. Next sort your data set into three subsamples, according to whether the individual is nonwhite and not Hispanic (NONWH = 1), Hispanic (HISP = 1), or other (primarily whites, hereafter denoted OTHER). How many observations are there in each of the three subsamples? Calculate sample means for the variables LNW AGE, FE, UNION, ED, EX, and EXSQ separately for each of the three subsamples. Then for each of these variables, calculate differences in the sample means between the OTHER and NONWH samples and between OTHER and HISP. What is the mean difference in LNWAGE in the OTHER and NONWH samples? What is the mean difference in LNWAGE between OTHER and HISP? On average, do NONWH and HISP workers have less or more schooling than OTHER workers? What are the potential experience differences among these three subsamples? Are there substantial gender (FE) differences by race?
2. Using the sorted data by race from part (b), estimate parameters in separate NONWH, HISP, and OTHER equations, each of the form

LNWAGE = α + αF•FE + αu•UNION + β1•ED + β2•EX + β3•EXSQ + є

Now compare magnitudes of the estimated coefficients in the three equations, calculating the standard errors of differences in estimated coefficients between subsamples using the formula discussed in class related to this question. In particular, is the effect of schooling (ED) on LNWAGE different between OTHER and NONWH? Between OTHER and HISP? Interpret these results. Does the gender-related (FE) wage differential vary significantly by race? In what ways does the age-earnings profile differ by race? How do you interpret this?

1. Using a reasonable level of significance, perform a Chow test of the null hypothesis that all intercept and slope coefficients are simultaneously equal in the NONWH, HISP, and OTHER equations. To do this, first reestimate the equation in part (a), omitting the NONWH and HISP variables, and retrieve the sum of squared residuals from this equation. Using the Chow test procedure (be careful in computing degrees of freedom), compare this sum of squared residuals with the sum from the three separate regressions in part (c). Interpret your results. Outline what additional information would be required in order to test the null hypothesis that parameters from the HISP sample are equal to those from the OTHER sample. Calculate and interpret results from such a test.
2. On the basis of these analyses of wage discrimination by race, what do you conclude concerning the relative importance of racial discrimination in the United States? Defend your conclusions.

Next, you will compute and interpret the effects of wage discrimination by gender, using three alternative procedures: a traditional dummy variable approach and the two weighted differences in mean endowments procedures introduced by Blinder and by Oaxaca. You will also do a Chow test and then examine whether wage discrimination by gender has declined in the United States over the 1978-1985 time period.

1. Using data in the data file CPS78, begin by estimating the effects of wage discrimination by race using simple dummy variable procedures. Specifically, using OLS, estimate parameters in the equation

LNWAGE = α + αF•FE + αu•UNION + αN•NONWH + αH•HISP + β1•ED + β2•EX + β3•EXSQ + є

Interpret the estimated coefficient αFand its value exponentiated. Test the null hypothesis that, holding race, union status, education, and experience fixed, gender has no effect on log wages.

1. Next sort your CPS78 data set into two subsamples, according to whether the individual is female (FE = 1) or male (FE = 0). How many observations are there in each of the two subsamples? Calculate sample means for the variables LNWAGE, NONWH, HISP, UNION, ED, EX, and EXSQ separately for males and for females. Then for each of these variables, calculate differences in the sample means between the FE = 1 and FE = 0 samples. What is the mean difference in LNWAGE in the male and female samples? On average, do male workers have less or more schooling than females? What are the potential experience differences between males and females? For which gender is the potential experience variable EX more likely to measure on-the-job training? Why? Are there substantial differences in the proportions of HISP, NONWH, and all other individuals by gender? Does the proportion of unionized workers vary considerably between males and females?
2. Using the data sorted by gender from part (g), estimate parameters in separate male and female equations, each of the form

LNWAGE = α + αN•NONWH + αH•HISP + αu•UNION + β1•ED + β2•EX + β3•EXSQ + є

Now compare magnitudes of the estimated coefficients in the male and female equations, calculating the standard errors of differences in estimated coefficients between subsamples using the formula discussed in class related to this question. In particular, does the effect of schooling (ED) on LNWAGE differ by gender? What about the effect of race on LNWAGE-does it differ by gender? Interpret these results. In what ways does the age-earnings profile differ between males and females? How do you interpret this?

1. Using a reasonable level of significance, do a Chow test of the null hypothesis that all intercept and slope coefficients are simultaneously equal in the male and female equations. To do this, first re-estimate the equation in part (f), omitting the FE dummy variable. Retrieve the sum of squared residuals from this regression. Then sum the sum of squared residuals in the two gender-specific regressions of part (h). Using the Chow test procedure based on the comparison of sum of squared residuals (be careful in computing degrees of freedom), calculate the Chow test statistic and compare it to the critical value. Interpret your results.
2. Next implement the Blinder-Oaxaca procedures discussed in class to obtain the first of two alternative measures that bracket the estimated effects of wage discrimination by gender. Specifically, using sample mean data from part (g) the equation noted in class for this assignment, where males are the advantaged group and females are the disadvantaged group, decompose the difference in mean LNWAGE by gender into that portion due to differences in endowments (weight the differences in sample means of the regressors by the estimated coefficients from the male equation) and the residual part due to discrimination by gender. (Note that according to class discussion the discrimination portion can be calculated indirectly as the above residual or directly as the difference in the estimated coefficients ∆βweighted by mean values of the regressors in the female equation.) What proportion of mean differences in LNWAGE by gender is due to differences in endowments? Which endowment differences appear to be particularly important? What proportion of the mean differences in LNWAGE can be attributed to effects of discrimination by gender? How do these proportions compare with the dummy variable procedure in part (f) of this assignment?
3. The alternative Blinder-Oaxaca measure of wage discrimination by gender weights mean differences in endowments by the estimated female (rather than male) coefficients. Using the estimated parameters from the female equation from part (c) and male/female mean differences in regressors, substitute the other equation discussed in class and compute this alternative measure of wage discrimination. What proportion of the mean difference in LNWAGE by gender is now estimated as being due to differences in mean endowments? What portion represents wage discrimination by gender? Compare these estimates to results obtained in part (j). What range brackets the estimated proportion of differences in LNWAGE by gender due to discrimination?
4. Now examine empirically whether discrimination by gender has been declining in the United States over the 1978-1985 time period. In particular, repeat steps (f) - (k) using data from the CPS85 data file, rather than from CPS78. What has happened to the size of the LNWAGE differential from 1978 to 1985? Have differences in mean endowments by gender declined from 1978 to 1985? Interpret your results. Has the proportion of gender differences in LNWAGE due to discrimination changed from 1978 to 1985? Interpret these results, and comment on the reliability of your estimates. Are there any surprises here? Why or why not?