

Econ 203B: Single Equation Models

Solutions for Problem Set 2

Michael Powell

Department of Economics, UCLA

January 24th, 2006

1 Greene Chapter 4

8. Consider the multiple regression of y on K variables X and an additional variable z . Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on X is larger when z is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that X and z are nonstochastic and that the coefficient on z is nonzero.

Solution This question asks us to estimate the following models:

$$y = X\beta + \varepsilon \quad (1)$$

$$\begin{aligned} y &= \underset{n \times k}{X} \underset{k \times 1}{\beta} + \underset{n \times 1}{z} \underset{1 \times 1}{\gamma} + \underset{n \times 1}{\varepsilon} \\ &= \begin{bmatrix} X & z \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \varepsilon \end{aligned} \quad (2)$$

And asks us to compare $Var(\hat{\beta})$ with $Var(\tilde{\beta})$ where $\hat{\beta}$ is the OLS estimator of (1) and $\tilde{\beta}$ is the first k OLS estimators of (2). Recall that the respective OLS estimates are:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'y \\ \tilde{\beta} &= (X'M_zX)^{-1} X'M_zy \end{aligned} \quad (3)$$

Where (3) was derived using the partitioned regression formula. This gives us:

$$\begin{aligned} Var(\hat{\beta}) &= Var\left((X'X)^{-1} X'y\right) \\ &= (X'X)^{-1} X' \cdot Var(y) \cdot X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

And

$$\begin{aligned} Var(\tilde{\beta}) &= Var\left((X'M_zX)^{-1} X'M_zy\right) \\ &= (X'M_zX)^{-1} X'M_z \cdot Var(y) \cdot M'_zX(X'M_zX)^{-1} \\ &= \sigma^2 (X'M_zX)^{-1} X'M_zX(X'M_zX)^{-1} \\ &= \sigma^2 (X'M_zX)^{-1} \end{aligned}$$

Where I used the fact that $M_z = M_zM_z = M'_z$. (annihilator matrices are symmetric and idempotent). The question asks us to show that

$$Var(\tilde{\beta}) \geq Var(\hat{\beta})$$

In the matrix sense. (That is, $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semi-definite). It can be shown that for two nonsingular, symmetric, positive semi-definite matrices A and B , $A - B$ is positive semi-definite if and only if $A^{-1} - B^{-1}$ is negative semi-definite. I will use this fact.

$$\begin{aligned} Var(\tilde{\beta})^{-1} - Var(\hat{\beta})^{-1} &= \frac{1}{\sigma^2}(X'M_zX) - \frac{1}{\sigma^2}X'X \\ &= -\frac{1}{\sigma^2}(X'X - X'M_zX) \\ &= -\frac{1}{\sigma^2}X'(I - M_z)X \\ &= -\frac{1}{\sigma^2}X'P_zX \end{aligned}$$

Clearly $X'P_zX$ is positive semi-definite and therefore $-\frac{1}{\sigma^2}X'P_zX$ is negative semi-definite. This gives us the desired result that $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semi-definite.

For the second part of the question, we are asked to determine whether the matrix $\widehat{Var}(\tilde{\beta}) - \widehat{Var}(\hat{\beta})$ is positive or negative semi-definite. Letting $\hat{\varepsilon}_1$ denote the residuals from the estimate of (1) and $\hat{\varepsilon}_2$ denote the residuals from the estimate of (2), we have:

$$\begin{aligned} \widehat{Var}(\tilde{\beta}) &= \hat{\sigma}^2(X'M_zX)^{-1} = \frac{\hat{\varepsilon}'_2\hat{\varepsilon}_2}{n-k-1}(X'M_zX)^{-1} \\ \widehat{Var}(\hat{\beta}) &= \hat{\sigma}^2(X'X)^{-1} = \frac{\hat{\varepsilon}'_1\hat{\varepsilon}_1}{n-k}(X'X)^{-1} \end{aligned}$$

What can we say about $\widehat{Var}(\tilde{\beta}) - \widehat{Var}(\hat{\beta})$?

$$\widehat{Var}(\tilde{\beta}) - \widehat{Var}(\hat{\beta}) = \frac{\hat{\varepsilon}'_2\hat{\varepsilon}_2}{n-k-1}(X'M_zX)^{-1} - \frac{\hat{\varepsilon}'_1\hat{\varepsilon}_1}{n-k}(X'X)^{-1}$$

From the previous part of the question, we saw that $(X'M_zX)^{-1} - (X'X)^{-1}$ was positive semi-definite. Also, we know that since (2) involves more regressors than (1), $\hat{\varepsilon}'_2\hat{\varepsilon}_2 \leq \hat{\varepsilon}'_1\hat{\varepsilon}_1$. Finally, $n-k-1 < n-k$ and thus $\frac{1}{n-k-1} > \frac{1}{n-k}$. Putting this all together, we cannot conclude anything about the semi-definiteness of this matrix without knowing the relevant magnitudes.

9. For the classical normal regression model $y = X\beta + \varepsilon$ with no constant term and K regressors, assuming that the true value of β is zero, what is the exact expected value of $F[K, n-K] = \frac{R^2/K}{(1-R^2)/(n-K)}$?

Solution First, note that if $\beta = 0$, then we have:

$$\begin{aligned} \hat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'Y = X(X'X)^{-1}X' \left(\underbrace{X\beta}_{=0} + \varepsilon \right) \\ &= X(X'X)^{-1}X'\varepsilon \equiv P\varepsilon \end{aligned}$$

And, of course,

$$\begin{aligned} \hat{\varepsilon} &= Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = \underbrace{(I - X(X'X)^{-1}X')}_{\equiv M} Y \\ &= MY = M(X\beta + \varepsilon) = M\varepsilon \end{aligned}$$

Where I used the fact that

$$MX = (I - X(X'X)^{-1}X')X = X - X(X'X)^{-1}X'X = X - X = 0$$

Further, since $\varepsilon \sim N(0, \sigma^2 I_n)$, we have that

$$\begin{aligned} \begin{bmatrix} \hat{Y} \\ \hat{\varepsilon} \end{bmatrix} &= \begin{bmatrix} P \\ M \end{bmatrix} \varepsilon \sim N\left(0, \begin{bmatrix} P \\ M \end{bmatrix} \sigma^2 I_n \begin{bmatrix} P' & M' \end{bmatrix}\right) \\ &= N\left(0, \sigma^2 \begin{bmatrix} PP' & PM' \\ MP' & MM' \end{bmatrix}\right) \end{aligned}$$

Thus, \hat{Y} and $\hat{\varepsilon}$ are independent iff $PM' = 0$. This indeed does hold since:

$$PM' = PM = P(I - P) = P - PP = P - P = 0$$

Where I used symmetry of M and idempotence of P . (Facts we proved several times in Hahn's class)

Since \hat{Y} and $\hat{\varepsilon}$ are independent, for any function g of \hat{Y} alone and any function h of $\hat{\varepsilon}$ alone, $g(\hat{Y})$ is independent of $h(\hat{\varepsilon})$.

Since our regression does not involve a constant, it only makes sense to use the uncentered R^2 (recall that with no constant, R_C^2 is not restricted to be between 0 and 1):

$$R_{UC}^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{Y'Y} = \frac{\hat{Y}'\hat{Y}}{Y'Y}$$

Giving us:

$$\begin{aligned} F &= \frac{R^2/K}{(1 - R^2)/(n - K)} = \left(\frac{n - K}{K}\right) \frac{R^2}{1 - R^2} \\ &= \left(\frac{n - K}{K}\right) \frac{\hat{Y}'\hat{Y}}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{Y'Y}} = \left(\frac{n - K}{K}\right) \frac{\hat{Y}'\hat{Y}}{\hat{\varepsilon}'\hat{\varepsilon}} \end{aligned}$$

Substituting in the above expressions, we have:

$$\begin{aligned} F &= \left(\frac{n - K}{K}\right) \frac{\varepsilon'P'\varepsilon}{\varepsilon'M'\varepsilon} = \left(\frac{n - K}{K}\right) \frac{\varepsilon'P\varepsilon}{\varepsilon'M\varepsilon} \\ &= \left(\frac{n - K}{K}\right) \frac{\left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right)}{\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)} \end{aligned}$$

Recall that

$$\left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right) \sim \chi^2(\text{tr}(P)) = \chi^2(K)$$

and

$$\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right) \sim \chi^2(\text{tr}(M)) = \chi^2(n - K)$$

By the above argument, these two random variables are independent and therefore:

$$\begin{aligned} E[F] &= E\left[\left(\frac{n - K}{K}\right) \frac{\left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right)}{\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)}\right] = \left(\frac{n - K}{K}\right) E\left[\frac{\left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right)}{\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)}\right] \\ &= \left(\frac{n - K}{K}\right) E\left[\left(\frac{\varepsilon}{\sigma}\right)' P \left(\frac{\varepsilon}{\sigma}\right)\right] E\left[\frac{1}{\left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)}\right] \end{aligned}$$

Using the well-known result that the expected value of a random variable which is $\chi^2(K)$ distributed is K , it remains to determine what the expected value of the inverse of a chi-square random variable with degrees of freedom equal to $n - K$ is.

Let $Y \sim \chi^2(p)$. Then (since $Y \sim \chi^2(p) \iff Y \sim \Gamma(\frac{p}{2}, 2)$):

$$\begin{aligned} E\left[\frac{1}{Y}\right] &= \int_0^\infty \frac{1}{y} \frac{1}{\Gamma(\frac{p}{2}) 2^{\frac{p}{2}}} y^{\frac{p}{2}-1} e^{-\frac{y}{2}} dy \\ &= \int_0^\infty \frac{1}{\frac{p-2}{2} \Gamma(\frac{p-2}{2}) 2^{\frac{p-2}{2}} \cdot 2} y^{\frac{p-2}{2}-1} e^{-\frac{y}{2}} dy \text{ since } \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \\ &= \frac{1}{p-2} \int_0^\infty \underbrace{\frac{1}{\Gamma(\frac{p-2}{2}) 2^{\frac{p-2}{2}}} y^{\frac{p-2}{2}-1} e^{-\frac{y}{2}}}_{\text{pdf of } \Gamma(\frac{p-2}{2}, 2) \text{ r.v.}} dy = \frac{1}{p-2} \end{aligned}$$

This gives us the final result that:

$$E[F] = \left(\frac{n-K}{K}\right) (K) \left(\frac{1}{n-K-2}\right) = \frac{n-K}{n-K-2}$$

2 Other Questions

1. This problem uses the data set CPS85 that may be downloaded from the class web site, along with the file ReadCPS that describes the data. The data set is a random sample from the May 1985 Current Population Survey conducted by the U.S. Census Bureau. It contains observations on 12 variables for 534 individuals. The first variable in the data set is years of schooling (EDU), and the next six entries are 0-1 dummy variables taking on the value 1 if the individual resides in the south (SOUTH), is non-white and non-hispanic (NONWH), is Hispanic (HISP), is female (FE), is married (MAR) and is female and married (MARFE). The next two variables measure potential years of experience (EX), computed as age minus years of schooling minus 6, and this potential experience measure squared (EXSQ). The next entry is a dummy variable taking on the value 1 if the individual works at a union job (UNIO). The next column is the natural logarithm of the individual's average hourly in dollars earnings (LNWAGE). The next variable is the individual's age in years (AGE). The rest of the variables in the data set will not be used in this problem. Whenever necessary in the questions below, assume that the assumptions of the Classical Normal Regression model hold.
 - a. Compute the average hourly wage for the entire sample. There are two ways of doing this. Either compute the arithmetic mean of $LNWAGE$ and exponentiate it (which gives the geometric mean of average hourly wage, $WAGE = e^{LNWAGE}$), or exponentiate $LNWAGE$ for each individual and then compute the arithmetic mean. Are these two measures identical?

Solution For the first part of the question, MATLAB gives the following computed values:

$$\begin{aligned} avgwage1 &= \exp\left\{\frac{1}{534} \sum_{i=1}^{534} LNWAGE_i\right\} \approx 7.8396 \\ avgwage2 &= \frac{1}{534} \sum_{i=1}^{534} \exp\{LNWAGE_i\} \approx 9.0239 \end{aligned}$$

Clearly, these two measures are not identical. Intuitively, since $f(x) = e^x$ is a convex function, we would expect, by Jensen's inequality that $E[f(X)] > f(E[X])$. Here, this is confirmed, since $avgwage2 > avgwage1$.

- b. Compute the sample means of the following dummy variables: *SOUTH*, *FE*, *UNIO*, *NONWH*, *HISP*. How many males, females, south residents, non-south residents, union workers, non-union workers, whites, Hispanics, non-whites and non-Hispanics are in the sample?

Solution MATLAB gives the following output:

$$\begin{aligned} \frac{1}{534} \sum_{i=1}^{534} SOUTH_i &\approx 0.2921; & \frac{1}{534} \sum_{i=1}^{534} FE_i &\approx 0.4588 \\ \frac{1}{534} \sum_{i=1}^{534} UNIO_i &\approx 0.1798; & \frac{1}{534} \sum_{i=1}^{534} NONWH_i &\approx 0.1255 \\ \frac{1}{534} \sum_{i=1}^{534} HISP_i &\approx 0.0506 \end{aligned}$$

Since, for a dummy variable D_i , we have that:

$$\begin{aligned} [\# \text{ of } i \text{ s.t. } D_i = 1] &= \sum_{i=1}^n D_i = n\bar{D} \\ [\# \text{ of } i \text{ s.t. } D_i = 0] &= n - \sum_{i=1}^n D_i = n(1 - \bar{D}) \end{aligned}$$

This gives us:

$$\begin{aligned} [\# \text{Males}] &= 534(1 - 0.4588) = 289 \\ [\# \text{Females}] &= 534(0.4588) = 245 \\ [\# \text{South residents}] &= 534(0.2921) = 156 \\ [\# \text{Non-South residents}] &= 534(1 - 0.2921) = 378 \\ [\# \text{Union workers}] &= 534(0.1798) = 96 \\ [\# \text{Non-union workers}] &= 534(1 - 0.1798) = 438 \\ [\# \text{Whites}] &= 534(1 - 0.1255) = 467 \\ [\# \text{Non-Whites}] &= 534(0.1255) = 67 \\ [\# \text{Hispanics}] &= 534(0.0506) = 27 \\ [\# \text{Non-Hispanics}] &= 534(1 - 0.0506) = 507 \end{aligned}$$

Since this is an econometrics course, and since statistics look nicer in a table, here is a table of the preceding results:

Variable	Number
Males	289
Females	245
South Residents	156
Non-South Residents	378
Union Workers	96
Non-Union Workers	438
Whites	467
Non-Whites	67
Hispanics	27
Non-Hispanics	507

- c. Compute the means and standard deviations of $LNWAGE$, EDU , and EX for the entire sample, and then by gender (male/female), by race (white/nonwhite/Hispanic), and by union status (union/non-union). Within each of the three groups sorted by gender, race, and union status, find which subgroup has the highest average $LNWAGE$ and the highest dispersion as measured by the standard deviation. Do the same for EDU .

Solution This question is way easier to do in STATA, so here are my results:

	MeanLNWAGE	StdLNWAGE	MeanEDU	StdEDU	MeanEX	StdEX
Entire Sample	2.06	0.53	13.02	2.62	17.82	12.38
Females	1.93	0.49	13.02	2.43	18.83	12.61
Males	2.17	0.53	13.01	2.77	16.97	12.13
Non-Whites	1.97	0.50	12.64	2.60	18.79	12.22
Whites	2.09	0.53	13.17	2.47	17.73	12.27
Hispanics	1.82	0.53	11.52	4.05	17	14.70
Union Workers	2.29	0.42	12.89	2.64	20.94	12.59
Non-Union Workers	2.01	0.53	13.05	2.61	17.14	12.24

- d. Using Least Squares, estimate the parameters in a sample model where $LNWAGE$ is regressed on a constant, years of schooling (EDU), and experience (EX), and report these along with their estimated standard errors. What do the slope coefficients on EDU and EX measure? Intuitively, age may affect wages as well. What would be the problem of including AGE as an additional regressor?

Solution Estimating the model:

$$LNWAGE_i = \beta_1 + \beta_2 EDU_i + \beta_3 EX_i + \varepsilon_i$$

MATLAB gives the following results (standard errors in parentheses):

$$\widehat{LNWAGE}_i = \underbrace{0.5941}_{(0.1244)} + \underbrace{0.0964}_{(0.0083)} EDU_i + \underbrace{0.0118}_{(0.0018)} EX_i$$

Now, proceeding to interpret the slope coefficients:

$$\hat{\beta}_2 = \frac{\partial \widehat{LNWAGE}_i}{\partial EDU_i} = \frac{\frac{\partial \widehat{WAGE}_i}{\widehat{WAGE}_i}}{\frac{\partial EDU_i}{\widehat{WAGE}_i}}$$

$$\hat{\beta}_3 = \frac{\partial \widehat{LNWAGE}_i}{\partial EX_i} = \frac{\frac{\partial \widehat{WAGE}_i}{\widehat{WAGE}_i}}{\frac{\partial EX_i}{\widehat{WAGE}_i}}$$

That is, $\hat{\beta}_2$ measures the proportional increase in wages associated with a one unit increase in education. Similarly, $\hat{\beta}_3$ measures the proportional increase in wages associated with a one unit increase in experience.

Recall that the definition of EX_i is

$$EX_i = AGE_i - EDU_i - 6$$

Here, if we had included the variable for AGE , we would have a problem with exact collinearity. (i.e. there is a perfect linear relationship between the variables) This would lead to problems inverting the matrix $X'X$ as it would no longer be of full column rank.

- e. Compute and interpret the R^2 coefficient for this model.

Solution Here, I will calculate the centered R^2 for this regression: (using MATLAB)

$$R_C^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{LNWAGE'M^0LNWAGE}$$

$$= 0.2115$$

A possible interpretation for this value is that this regression explains approximately 21% of the variation in $LNWAGE$.

f. According to the human capital theory, experience affects wages. Test this hypothesis at the 5% significance level.

Solution This question asks us to perform a 5% t-test of $H_0 : \beta_3 = 0$. We have $n - k = 534 - 3 = 531$ degrees of freedom, so the critical value is

$$c_{0.05,t(531)}^* \approx c_{0.05,N(0,1)}^* = 1.95$$

We have (from part d):

$$t_0 = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} = \frac{0.0118}{0.0018} = 6.5556 > 1.95$$

Therefore, we reject the null hypothesis that experience does not affect wages, lending support to the human capital theory.

g. Because human capital depreciates with age (and hence with the measure of experience we are using), we expect decreasing returns to experience. To see whether this assumption is correct, estimate a linear model with an additional quadratic in experience variable:

$$LNWAGE = \beta_0 + \beta_1 EDU + \beta_2 EX + \beta_3 EXSQ + \varepsilon$$

Is the sign of β_3 consistent with what you would expect? Is β_3 statistically significant? At what level of experience is $LNWAGE$ maximized? What happens to the estimated coefficients of EDU and EX and their standard errors?

Solution Estimating the model:

$$LNWAGE_i = \beta_0 + \beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i$$

MATLAB gives the following results (standard errors in parentheses):

$$\widehat{LNWAGE}_i = \underbrace{0.5203}_{(0.1236)} + \underbrace{0.0898}_{(0.0083)} EDU_i + \underbrace{0.0349}_{(0.0056)} EX_i - \underbrace{0.0005}_{(0.0001)} EXSQ_i$$

The sign of $\hat{\beta}_3$ is consistent with the theory that human capital depreciates with age. $\hat{\beta}_3$ is statistically significant since

$$|t_0| = \left| \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} \right| = \left| \frac{-0.0005}{0.0001} \right| = 5 > 1.96 = c_{0.05,N(0,1)}^*$$

Holding education constant, if we take first order conditions of \widehat{LNWAGE}_i with respect to EX_i , we get:

$$0 = \frac{\partial \widehat{LNWAGE}_i}{\partial EX_i} = 0.0349 - 0.0010 EX_i$$

Which occurs when

$$\begin{aligned} 0.0010 EX_i &= 0.0349 \\ EX_i &= \frac{0.0349}{0.0010} = 34.9 \end{aligned}$$

That is, \widehat{LNWAGE}_i is maximized when $EX_i = 34.9$.

Finally, we see that the coefficient for EDU_i decreases from 0.0964 to 0.0898, while its standard error stays constant at 0.0083 and the coefficient for EX_i increases from 0.0118 to 0.0349, and its standard error increases from 0.0018 to 0.0056.

- h. The specification in part (g) assumes that the intercept and slope coefficient on EDU are the same for all individuals. We may think that the effect of schooling on wages differs by a *constant factor of proportionality* for males and females, i.e. that for males,

$$WAGE = \alpha_M e^{\beta_1 EDU + \beta_2 EX + \beta_3 EXSQ} e^\varepsilon$$

while for females,

$$WAGE = \alpha_F e^{\beta_1 EDU + \beta_2 EX + \beta_3 EXSQ} e^\varepsilon$$

where α_M and α_F are differing factors of proportionality, β is the common return to schooling and ε is a random error term. Show that this implies that when the dependent variable is $LNWAGE$ rather than $WAGE$, males and females have different intercept terms but common slope coefficients $\beta_1, \beta_2, \beta_3$. To estimate these different intercepts, run the following regression

$$LNWAGE = \alpha_1 + \alpha_2 FE + \beta_1 EDU + \beta_2 EX + \beta_3 EXSQ + \varepsilon$$

Interpret the estimates of α_1 and α_2 relating them to α_M and α_F above. Formulate and test the hypothesis (at a 5% significance level) that there is no gender discrimination using your estimates of α_1 and α_2 .

Solution Since

$$\begin{aligned} WAGE_i &= \alpha_M e^{\beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i} e^{\varepsilon_i} \text{ for males} \\ WAGE_i &= \alpha_F e^{\beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i} e^{\varepsilon_i} \text{ for females} \end{aligned}$$

Is equivalent to

$$\begin{aligned} LNWAGE_i &= \ln \alpha_M + \beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i \text{ for males} \\ LNWAGE_i &= \ln \alpha_F + \beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i \text{ for females} \end{aligned}$$

The assertion that $\alpha_M = \alpha_F$ is equivalent to the assertion that $\ln \alpha_M = \ln \alpha_F$. Define

$$\begin{aligned} \alpha_1 &\equiv \ln \alpha_M \\ \alpha_1 + \alpha_2 &\equiv \ln \alpha_F \end{aligned}$$

Then we see that $\alpha_M = \alpha_F$ if our results from the regression

$$LNWAGE_i = \alpha_1 + \alpha_2 FE_i + \beta_1 EDU_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i$$

Are such that we can conclude that $\alpha_2 = 0$. Estimating this model, MATLAB gives:

$$\widehat{LNWAGE}_i = \underbrace{0.6007}_{(0.1195)} - \underbrace{0.2570}_{(0.0387)} FE_i + \underbrace{0.0913}_{(0.0080)} EDU_i + \underbrace{0.0360}_{(0.0054)} EX_i - \underbrace{0.0005}_{(0.0001)} EXSQ_i$$

There is no gender discrimination if $H_0 : \alpha_2 = 0$. Testing this at the 5% significance level, we see that:

$$|t_0| = \left| \frac{-0.2570}{0.0386} \right| = 6.658 > 1.96 = c_{0.05, N(0,1)}^*$$

Therefore, we can reject the null.

- i An alternative procedure for formulating the regression relationship in the previous question is as follows: First create a dummy variable called MA , defined as $MA = 1 - FE$. (What does this variable denote?) Then estimate the following model

$$LNWAGE = \gamma_1 MA + \gamma_2 FE + \beta_1 EDU + \beta_2 EX + \beta_3 EXSQ + \varepsilon$$

Interpret γ_1 and γ_2 and relate them to α_M and α_F above. According to econometric theory, what should be the relationship among the estimates α_1 and γ_1 and α_2 and γ_2 ? Are your estimates consistent with this relationship? Formulate a test for the null hypothesis that there is no gender discrimination using the estimates of γ_1 and γ_2 . What would be the problem of including an intercept term in the model?

Solution This new variable, MA takes on the value 1 if an individual is not female (i.e. is Male, except for very weird cases) and 0 if the individual is female. Estimating this model using MATLAB:

$$\widehat{LNWAGE}_i = \underbrace{0.6007}_{(0.1195)} MA_i + \underbrace{0.3437}_{(0.1218)} FE_i + \underbrace{0.0913}_{(0.0080)} EDU_i + \underbrace{0.0360}_{(0.0054)} EX_i - \underbrace{0.0005}_{(0.0001)} EXSQ_i$$

Here, γ_1 is equivalent to $\ln \alpha_M$ from above and γ_2 is equivalent to α_F . Econometric theory suggests that

$$\begin{aligned}\gamma_1 &= \alpha_1 \\ \gamma_2 &= \alpha_1 + \alpha_2\end{aligned}$$

Which is confirmed in this estimation since:

$$\begin{aligned}\hat{\gamma}_1 &= 0.6007 = \hat{\alpha}_1 \\ \hat{\gamma}_2 &= 0.3437 = 0.6007 - 0.2570 = \hat{\alpha}_1 + \hat{\alpha}_2\end{aligned}$$

To say that there are no gender effects is to say that

$$H_0 : \gamma_1 = \gamma_2 \text{ or } H_0 : \gamma_1 - \gamma_2 = 0$$

Testing this at the 5% significance level, we see that:

$$t_0 = \left| \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{se(\hat{\gamma}_1 - \hat{\gamma}_2)} \right| = \left| \frac{0.2570}{0.0387} \right| = 6.6408 > 1.96 = c_{0.05, N(0,1)}^*$$

And we can once again reject the null that there is no gender discrimination.

In this regression, we cannot include a constant, since there would then be a direct linear relationship among the variables FE , MA , and the constant:

$$FE_i + MA_i = 1 \quad \forall i$$

This would lead to noninvertibility of $X'X$.

j The specifications in parts (h) and (i) above take the returns to education to be constant for males and females. How would you test whether this assumption is correct? Perform the test.

Solution Using MATLAB to estimate the model:

$$LNWAGE_i = \alpha_1 + \alpha_2 FE_i + \beta_1 EDU_i + \gamma_1 EDU_i * FE_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i$$

We have:

$$\begin{aligned}\widehat{LNWAGE}_i &= \underbrace{0.7896}_{(0.1402)} - \underbrace{0.7539}_{(0.1992)} FE_i + \underbrace{0.0762}_{(0.0099)} EDU_i + \\ &\quad \underbrace{0.0381}_{(0.0150)} FE_i * EDU_i + \underbrace{0.0369}_{(0.0054)} EX_i - \underbrace{0.0006}_{(0.0001)} EXSQ_i\end{aligned}$$

Here, we want to test $H_0 : \gamma_1 = 0$

$$|t_0| = \left| \frac{0.0381}{0.0150} \right| = 2.54 > 1.96 = c_{0.05, N(0,1)}^*$$

Therefore, we can reject the null that returns to education is the same for males and females.

k. Consider the specification

$$LNWAGE = \beta_0 + \beta_1 EDU + \beta_2 EX + \beta_3 EXSQ + \varepsilon$$

Does the return to experience differ by years of schooling? How would you allow that the return to experience differs by education level? Test whether the return to experience differs by education level.

Solution In this part, we are asked to estimate the model:

$$LNWAGE_i = \beta_0 + \beta_1 EDU_i + \gamma_1 EDU_i * EX_i + \beta_2 EX_i + \beta_3 EXSQ_i + \varepsilon_i$$

MATLAB gives the following output:

$$\begin{aligned} \widehat{LNWAGE}_i &= \underbrace{0.1840}_{(0.2193)} - \underbrace{0.1131}_{(0.0151)} EDU_i - \underbrace{0.0012}_{(0.00063)} EDU_i * EX_i \\ &\quad + \underbrace{0.0545}_{(0.0120)} EX_i - \underbrace{0.0007}_{(0.0001)} EXSQ_i \end{aligned}$$

Here, we are asked to test $H_0 : \gamma_1 = 0$:

$$|t_0| = \left| \frac{-0.0012}{0.00063} \right| = 1.9048 \leq 1.96 = c_{0.05, N(0,1)}^*$$

Therefore, we fail to reject the null that the return to experience differs by education level.

l Describe how you would test whether there is a wage premium for union jobs.

Solution I would estimate the model:

$$LNWAGE_i = \beta_1 + \beta_2 UNIO_i + \beta_3 EDU_i + \beta_4 EX_i + \beta_5 EXSQ_i + \varepsilon_i$$

And I would do a hypothesis test to see whether or not $\beta_2 = 0$. If $\beta_2 \neq 0$, then there is a wage premium for union jobs.