

Econ 203B Econometrics

Raphael Lam

1 Problem Set 2 Solutions¹

Greene Chapter 4 Question 8 Consider the multiple regression of y on k variables X and an additional variable z . Prove that, under the assumptions 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 non-stochastic and that the coefficient on z is not zero.

Solution:

Consider the linear regression $y = X\beta + \varepsilon$. Assume X is of dimension of $(N \times K)$

The variance of the coefficient estimator $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1} \quad (1)$$

The addition of a new variable z turns the linear regression into $y = X\tilde{\beta} + z\gamma + \varepsilon$. We have then the estimator of the variance will become $\text{Var}(\phi|W) = \sigma^2 (W'W)^{-1}$ where $W = [X \ z]$ and $\phi = [\tilde{\beta} \ \gamma]'$. Rewrite the matrix into

$$W'W = \begin{bmatrix} x_1'x_1 & \cdots & x_1'x_n & x_1'z \\ \vdots & \ddots & \vdots & \vdots \\ x_n'x_1 & \cdots & x_n'x_n & x_n'z \\ z'x_1 & \cdots & z'x_n & z'z \end{bmatrix} = \begin{bmatrix} X'X & X'z \\ z'X & z'z \end{bmatrix}$$

where $z'z$, a scalar (1x1).

Using the handout given in the Partition Regression, it can be seen that

$$\tilde{\beta} = (X'M_zX)^{-1}X'M_zy \quad \text{where } M_z = (I_n - z(z'z)^{-1}z') \quad (2)$$

The variance of the coefficient estimator $\tilde{\beta}$ is

¹I have included in part the solutions from M. Heusch and A. Mezza.

$$\begin{aligned}
\text{Var}(\tilde{\beta}|W) &= \sigma^2 [(X'M_zX)' X'M_zX]^{-1} \\
&= \sigma^2 (X'M_zX)^{-1} = \sigma^2 [X'(I_n - z(z'z)^{-1}z')X]^{-1} \\
&= \sigma^2 [(X'I_nX) - X'z(z'z)^{-1}z'X]^{-1}
\end{aligned}$$

Compare $\text{Var}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$ and $\text{Var}(\tilde{\beta}|W) = \sigma^2 [X'M_zX]^{-1}$.
With σ^2 known, we can write it as

$$\begin{aligned}
\text{Var}(\tilde{\beta}|W) - \text{Var}(\hat{\beta}|X) &= \sigma^2 \left\{ (X'X)^{-1} - [(X'I_nX) - X'z(z'z)^{-1}z'X]^{-1} \right\} \\
&= \sigma^2 (X'z(z'z)^{-1}z'X)^{-1}
\end{aligned}$$

Notice that if $\text{Var}(\tilde{\beta}|W) - \text{Var}(\hat{\beta}|X) \geq 0$ if the inverse, i.e.
 $\text{Var}(\hat{\beta}|X)^{-1} - \text{Var}(\tilde{\beta}|W)^{-1} \geq 0$. Applying the inverse on the left
hand side,

$$\text{Var}(\hat{\beta}|X)^{-1} - \text{Var}(\tilde{\beta}|W)^{-1} = \frac{1}{\sigma^2} X'z(z'z)^{-1}z'X \quad (3)$$

The matrix from above equation is non-negative since $\sigma^2 > 0$ and $z'z$
is nonnegative. The entire matrix will be positive semi-definite. There-
fore, under the assumptions 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.

Let's look at the case where σ^2 is unknown. From the lecture notes,
we can estimate the variance of residual errors with $\hat{\sigma}^2 = \text{Var}(e|X)$,
where $e = \hat{\varepsilon}$. Under the regression of y on X without z , we get $\hat{\sigma}^2 = \frac{e'_XeX}{n-k}$,
while under the regression of y on X and z , we get $\tilde{\sigma}^2 = \text{Var}(e|W) =$
 $\frac{e'_we'w}{n-k-1}$. So, the sample variance of the two estimators will be equal to

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\beta}|X) &= \hat{\sigma}^2 (X'X)^{-1} = \frac{e'_XeX}{n-k} (X'X)^{-1} \\
\widehat{\text{Var}}(\tilde{\beta}|W) &= \tilde{\sigma}^2 (X'M_zX)^{-1} = \frac{e'_we'w}{n-k-1} (X'M_zX)^{-1} \quad (4)
\end{aligned}$$

Using the same procedure as above by comparing the inverse of the
two, we may verify that $(X'X) > (X'M_zX)$. On the sample variance

term, we can see from the numerator that $e'_X e_X \geq e'_w e_w$ as the sum of squared residuals is nondecreasing with an additional variable, while the denominator $(n - k) > (n - k - 1)$. We cannot conclude which sample variance is greater. It's not possible to conclude which estimated variance is larger.

Greene Chapter 4 Question 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 $F[K, n - K] = (R^2/K) [(1 - R^2) / (n - K)]$?

Solution

Remark 1 *It is not clear which R^2 is referring to.*

Using different representations of F - statistics, we can write down:

$$\begin{aligned} F[K, n - K] &= \frac{R^2}{1 - R^2} \frac{(n - K)}{K} = \frac{1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}}{1 - \left[1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}\right]} \frac{(n - K)}{K} \\ &= \frac{\frac{y'y - \hat{\varepsilon}'\hat{\varepsilon}}{y'y} (n - K)}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} K} = \frac{y'y - \hat{\varepsilon}'\hat{\varepsilon} (n - K)}{\hat{\varepsilon}'\hat{\varepsilon} K} \\ &= \frac{\frac{y'y}{\sigma^2} - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} (n - K)}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} K} \end{aligned}$$

Note that

$$\begin{aligned} \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} &\sim \chi^2_{n-k} \\ \frac{y'y}{\sigma^2} &= \sum_{i=1}^n \left(\frac{y_i}{\sigma}\right)^2 \sim \chi^2_n \end{aligned}$$

The numerator follows a χ^2 distribution with degree of freedom k .

$$\frac{y'y}{\sigma^2} - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} \sim \chi^2_{n-(n-k)} = \chi^2_k$$

The denominator follows a χ^2 distribution with degree of freedom $n - k$.

Remark 2 It can be shown that the two χ^2 distributions in the numerator and denominator of F is indeed independent.

Intuition: $y = X\hat{\beta} + \hat{\varepsilon}$ and the numerator $y'y - \hat{\varepsilon}'\hat{\varepsilon} = \hat{\beta}'X'X\hat{\beta} = (\beta + (X'X)^{-1}X'\varepsilon)'(X'X)((\beta + (X'X)^{-1}X'\varepsilon))$

Given the above assumption that the true value of β is equal to 0, we can arrange the above as $y'y - \hat{\varepsilon}'\hat{\varepsilon} = \varepsilon'X(X'X)^{-1}X'\varepsilon = \varepsilon'P\varepsilon$. On the other hand, the denominator $\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon'M\varepsilon$ from derivations given in last problem set. It can be seen that the two terms are independent as $PM = X(X'X)^{-1}X' - (I - X(X'X)^{-1}X') = 0$

Therefore,

$$F[K, n - K] = \frac{\chi_k^2/k}{\chi^2/(n - k)}$$

and we can verify

$$\begin{aligned} \frac{R^2}{1 - R^2} \frac{(n - K)}{K} &\sim F[K, n - K] \\ E \left[\frac{R^2}{1 - R^2} \frac{(n - K)}{K} \right] &= \frac{n - K}{n - K - 2} \end{aligned}$$

As the two terms are independent we can rewrite

$$E \left[\frac{R^2}{1 - R^2} \frac{(n - K)}{K} \right] = \frac{(n - K)}{K} E[R^2/\sigma^2] E \left[\frac{1}{(1 - R^2)/\sigma^2} \right] \quad (5)$$

Using the fact that $E(\chi_k^2) = k$, so the second term $E[R^2/\sigma^2] = k$. The expectation of last term follows

$$\text{Define } X = (1 - R^2)/\sigma^2 \sim \chi_{n-k}^2$$

Using the density function of a χ^2 distribution (a specific case of Gamma distribution with $\alpha = \frac{n-k}{2}; \beta = 2$

$$\begin{aligned} E\left[\frac{1}{X}\right] &= \int \left[\frac{1}{\Gamma(\frac{n-k}{2})2^{\frac{n-k}{2}}} x^{\frac{n-k}{2}-1} e^{-\frac{x}{2}} \frac{1}{x} \right] dx \\ &= \frac{1}{\Gamma(\frac{n-k}{2})2^{\frac{n-k}{2}}} \int \left[x^{\frac{n-k-2}{2}-1} e^{-\frac{x}{2}} \right] dx \\ &= \frac{\Gamma(\frac{n-k-2}{2})2^{\frac{n-k-2}{2}}}{\Gamma(\frac{n-k}{2})2^{\frac{n-k}{2}}} \underbrace{\int \frac{1}{\Gamma(\frac{n-k-2}{2})2^{\frac{n-k-2}{2}}} \left[x^{\frac{n-k-2}{2}-1} e^{-\frac{x}{2}} \right] dx}_{1} \end{aligned}$$

One property in the Gamma distribution: $\Gamma(a) = (a - 1)!$, and the last term is an integral from 0 to ∞ (CDF=1). Rewrite the above,

$$E\left[\frac{1}{X}\right] = \frac{\Gamma\left(\frac{n-k-2}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)} \frac{1}{2} = \frac{1}{n-k-2}$$

Going back to above

$$\begin{aligned} E\left[\frac{R^2}{1-R^2} \frac{(n-K)}{K}\right] &= \frac{(n-K)}{K} \frac{k}{n-k-2} \\ &= \frac{n-k}{n-k-2} \end{aligned}$$

Additional Problems:

Problem 1

The CNLR model applies to

$$E(Y|X_1, X_2) = X_1\beta_1 + X_2\beta_2$$

A sample of size $n = 102$ gives

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}; (X'X) = \begin{pmatrix} 21 \\ 13 \end{pmatrix}; \hat{\varepsilon}'\hat{\varepsilon} = 80$$

Let $\theta = \beta_1 - \beta_2$. Test at the 5% significance level the null hypothesis that $\theta = 1$.

Solution:

First we rewrite the hypothesis as the following:

$$H_0 : \theta = \beta_1 - \beta_2 = \Gamma\beta = 1$$

$$\text{where } \Gamma = [1 \quad -1] \quad (1 \times 2)$$

$$\begin{aligned} \widehat{V}(\hat{\beta}|X) &= \frac{e'e}{n-k} (X'X)^{-1} = \frac{80}{102-2} \cdot \begin{pmatrix} 21 \\ 13 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.48 & -0.16 \\ -0.16 & 0.32 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \widehat{V}(\Gamma\hat{\beta}|X) &= \Gamma\widehat{V}(\hat{\beta})\Gamma' = \Gamma \frac{e'e}{n-k} (X'X)^{-1} \Gamma' \\ &= (1 \quad -1 \quad 0.48 \quad -0.16 \quad 1) \\ &= 1.12 \end{aligned}$$

Remark 3 You may easily compute it with Scientific Workplace in a second or Matlab.

To test the statistical significance of the null hypothesis H_0 under the CLNR model, we construct a test statistic

$$T = \frac{\Gamma\hat{\beta} - \Gamma\beta}{\hat{\sigma}_t} \sim t_{102-2}.$$

The sample calculation for the test statistics is equal to $\frac{3-1}{\sqrt{1.12}} = 1.8898$. The critical level for 5% significant level for a t-distribution with d.f. 100 is around 1.98. We cannot reject the null hypothesis H_0 , i.e. θ is not statistically significantly different from 1

Problem 2. It's the same as in Problem 1 with the CNLR model except the variance is known.

$$\sigma^2 = Var(Y|X_1, X_2) = 2$$

$$X'X = \begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix}$$

The sample produces the following estimate for β : $\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

1. Construct a 95% confidence interval for the null hypothesis $\theta = \beta_1 + \beta_2$.

Similar to problem 1, $\Gamma = [1 \ 1]$ (only 1 restriction: dimension 1×2)

Given σ^2 , we may construct the confidence interval as the following

$$CI = \{\theta : \theta = \beta_1 + \beta_2; \theta \in \Gamma\hat{\beta} \pm c_{N(0,1), \alpha/2=0.025}^* \sigma(\Gamma\hat{\beta})\}$$

Notice that $\Gamma\hat{\beta} = [1 \ 1] \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 5$

$$\begin{aligned} \sigma(\Gamma\hat{\beta}) &= \sqrt{Var(\Gamma\hat{\beta})} \\ &= \sqrt{\sigma^2 \Gamma (X'X)^{-1} \Gamma'} \\ &= \sqrt{2 [1 \ 1] \begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}} \\ &= 0.79057 \end{aligned}$$

The critical level $C_{N(0,1),\alpha/2=0.025}^* \approx 1.96$

The confidence interval becomes

$$\begin{aligned} CI &= 5 \pm 1.96(0.79057) \\ &= [3.45, 6.55] \end{aligned}$$

If θ falls between the CI , we cannot reject the null hypothesis.

2. Construct a 90% confidence region for the pair (β_1, β_2) .

Similar to the above except this time we have more than one restriction s.t. t-statistics may not be appropriate. We then use the F-statistics or χ -statistics (σ^2 known).

$$F = (\Gamma\hat{\beta} - \theta)' (\sigma^2\Gamma(X'X)^{-1}\Gamma')^{-1} (\Gamma\hat{\beta} - \theta)$$

$$\Gamma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ (dimension } 2 \times 2\text{)}$$

The test statistic that we consider $W = (\Gamma\hat{\beta} - \theta)' (\sigma^2\Gamma(X'X)^{-1}\Gamma')^{-1} (\Gamma\hat{\beta} - \theta)$

The proof can be found in the lecture notes or derived from the restricted least square method (see handouts given)

$$\begin{aligned} \Gamma\hat{\beta} - \theta &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} - \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\ &= \begin{bmatrix} 3 - \beta_1 \\ 2 - \beta_2 \end{bmatrix} \end{aligned}$$

$$(\sigma^2\Gamma(X'X)^{-1}\Gamma')^{-1} = \left(2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} = \frac{1}{2} \begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix}$$

$$W = \frac{1}{2} [3 - \beta_1 \ 2 - \beta_2] \begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 3 - \beta_1 \\ 2 - \beta_2 \end{bmatrix}$$

$$W = \frac{1}{2} (5\beta_1^2 + 4\beta_1\beta_2 + 4\beta_2^2 - 38\beta_1 - 28\beta_2 + 85)$$

The critical value for Chi-squared distribution for d.f. of 2 at 10% is ≈ 4.61 . Therefore, we cannot reject the null hypothesis if $W \leq c_{\chi^2(2),\alpha=0.1}^* \approx 4.61$. The confidence region (eclipse) [see notes] is therefore

$$CR = \{(\beta_1, \beta_2) : W \leq 4.61\}$$

Problem 3. A multiple regression of y on a constant, x_1 and x_2 produces the following results:

$$\hat{y} = 4 + 0.4x_1 + 0.9x_2, \quad R^2 = 8/60, \quad \hat{\varepsilon}'\hat{\varepsilon} = 520, \quad n = 29, \quad X'X = \begin{bmatrix} 290 & 0 \\ 0 & 5010 \\ 0 & 1080 \end{bmatrix}$$

(a) Test the hypothesis that the two slopes sum to 1 at the 5% significance level under normality of y .

(b) Test the hypothesis at the 5% significance level that the slope on x_1 is 0 by running the restricted regression and comparing the two sums of squared deviations. As in part (a), carry the test assuming normality of y .

Solution

$$H_0 : \theta = \beta_1 + \beta_2 = \Gamma\beta = 1$$

$$\text{where } \Gamma = [0 \quad 1 \quad 1] \quad [\text{dimension } 1 \times 3]$$

$$\beta = [1 \quad \beta_1 \quad \beta_2]'$$

$$\hat{\beta} = [4 \quad 0.4 \quad 0.9]'$$

$$\begin{aligned} \widehat{V}(\hat{\beta}) &= \frac{e'e}{n-k} (X'X)^{-1} = \frac{520}{29-3} \cdot \begin{bmatrix} 290 & 0 \\ 0 & 5010 \\ 0 & 1080 \end{bmatrix}^{-1} \\ &= 20 \cdot \begin{pmatrix} \frac{1}{29} & 0 & 0 \\ 0 & \frac{8}{390} & -\frac{1}{390} \\ 0 & -\frac{1}{390} & \frac{5}{390} \end{pmatrix} \\ \implies \widehat{V}(\Gamma\hat{\beta}) &= \Gamma\widehat{V}(\hat{\beta})\Gamma' = 20 \left[\frac{8}{390} + \frac{5}{390} - \frac{2}{390} \right] = \frac{22}{39}. \end{aligned}$$

Under the null and the CNR model, the test statistic

$$T = \frac{\Gamma\hat{\beta} - \theta}{\widehat{\sigma}(\Gamma\hat{\beta})} \sim t_{26}.$$

$$|T| = \frac{1.3 - 1}{\sqrt{22/39}} = 0.3994 < 1.71 \simeq C_{5\%}(t_{26})$$

We cannot reject H_0 (θ is not statistically significantly different from 1).

b)

$$H_0 : \beta_1 = 0$$

Again using the partitioned regression handout, we can obtain the residuals from the restricted regression (regress y on a constant and x_2), i.e. $X_2 = [\tilde{1} \ x_2]$

$$\begin{aligned} e_r &= M_2 y = M_2 [x_1 b_1 + X_2 b_2 + e] \\ &= M_2 x_1 b_1 + M_2 e = M_2 x_1 b_1 + e \\ \text{where } M_2 &= I_n - X_2 (X_2' X_2)^{-1} X_2'. \end{aligned}$$

where e is the residuals from the regression of y on a constant, x_1 , and x_2 .

$$\begin{aligned} e_r' e_r &= (M_2 x_1 b_1 + e)' (M_2 x_1 b_1 + e) \\ &= (b_1' x_1' M_2' + e') (M_2 x_1 b_1 + e) \\ &= b_1 x_1' M_2 x_1 b_1 + e' e \end{aligned} \tag{6}$$

[Notice that $M_2 e = e$ and e is orthogonal to x_1]

We want to find the value of $b_1 x_1' M_2 x_1 b_1$. Note that

$$\begin{aligned} X'X &= \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} 1' & x_1' & x_2' \end{bmatrix} = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i} x_{1i}' & \sum x_{1i} x_{2i}' \\ \sum x_{2i} & \sum x_{2i} x_{1i}' & \sum x_{2i} x_{2i}' \end{bmatrix} \\ \text{Var}(\hat{\beta}) &= \text{Var} \begin{bmatrix} e \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \sigma^2 (X'X)^{-1} = \sigma^2 \begin{bmatrix} (X'X)_{1,1}^{-1} & (X'X)_{1,2}^{-1} & (X'X)_{1,3}^{-1} \\ (X'X)_{2,1}^{-1} & (X'X)_{2,2}^{-1} & (X'X)_{2,3}^{-1} \\ (X'X)_{3,1}^{-1} & (X'X)_{3,2}^{-1} & (X'X)_{3,3}^{-1} \end{bmatrix} \end{aligned}$$

From above and remember on page 2 of partitioned regression handout, $\text{Var}(\hat{\beta}_1) = \sigma^2 (X'X)_{2,2}^{-1} = \sigma^2 (x_1' M_2 x_1)^{-1}$. (Other off-diagonal terms are covariance between estimators (e.g. $(X'X)_{2,3}^{-1} = \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$). We can see that $(X'X)_{2,2}^{-1} = (x_1' M_2 x_1)^{-1} = \frac{390}{8}$. So, the restricted regression provides an estimates of $b_1 = 0.4$. We can calculate $b_1 x_1' M_2 x_1 b_1 = 0.4^2 \left(\frac{390}{8}\right)$

Hence

$$\begin{aligned} e_r' e_r &= b_1 x_1' M_2 x_1 b_1 + e' e \\ e_r' e_r - e' e &= b_1 x_1' M_2 x_1 b_1 = 7.8 \end{aligned}$$

Under the null hypothesis and CLNR model,

$$F = \frac{(e'_r e_r - e' e) / p}{(e' e) / (n - k)} \sim F_{(1,26)}$$

with $p = 1; n - k = 29 - 3 = 26$

Given the information above,

$$F = 26 \times \frac{7.8}{520} = 0.39 < 4.23 = C_{5\%}(F_{(1,26)})$$

We cannot reject null hypothesis $H_0: \beta_1 = 0$

Problem 4. Production data for 22 firms in a certain industry produce the following, where $y = \ln(\text{output})$ and $x = \ln(\text{labor hours input})$:

$$\bar{y} = 20, \bar{x} = 10, \sum_{i=1}^n (y_i - \bar{y})^2 = 100, \sum_{i=1}^n (x_i - \bar{x})^2 = 60, \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = 30$$

(a) Write down the model in matrix notation and state the assumptions that justify running OLS to compute the unknown coefficients. Form the $(X'X)$, $(X'X)^{-1}$ and $(X'Y)$ matrices and compute the least squares estimator of $\beta = (\beta_1, \beta_2)'$.

Solution

The regression looks like $y = \beta_1 + \beta_2 x + \varepsilon$; with $y = \ln(\text{output})$ and $x = \ln(\text{labor})$

We can write down different matrices: $(X'X)$, $(X'X)^{-1}$, $(X'Y)$. Construct i to be a vector of 1 and use the same notation as in lecture note that i and x denote a row vector, i.e. $X = [1' x']$

$$\begin{aligned} X'X &= \begin{bmatrix} ii' & ix' \\ ix' & xx' \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \\ &= \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n (x_i - \bar{x})^2 + \bar{x} \sum_{i=1}^n x_i \end{bmatrix} \\ &= \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2 \end{bmatrix} = \begin{bmatrix} n & 10n \\ 10n & 60 + 100n \end{bmatrix} \\ &= \begin{bmatrix} 22 & 220 \\ 220 & 2260 \end{bmatrix} \end{aligned}$$

Similarly solving for $(X'X)^{-1}$

$$\begin{aligned}
 (X'X)^{-1} &= \frac{1}{\det(X'X)} \left[(X'X)_{j,k} \right]' \\
 &= \frac{1}{60n + 100n^2 - 100n^2} \begin{bmatrix} 60 + 100n - 10n & \\ -10n & n \end{bmatrix} \\
 &= \frac{1}{60n} \begin{bmatrix} 60 + 100n - 10n & \\ -10n & n \end{bmatrix} = \frac{1}{1320} \begin{bmatrix} 2260 & -220 \\ -220 & 22 \end{bmatrix} \\
 &= \begin{bmatrix} 1.7121 & -0.16667 \\ -0.16667 & 1.6667 \times 10^{-2} \end{bmatrix}
 \end{aligned}$$

Finally for $(X'Y)$: i and x denotes row vector of dimension $1 \times n$ while y denotes a column vector $n \times 1$. It's somehow confusing to match with the notation in lecture notes.

$$\begin{aligned}
 X'Y &= \begin{bmatrix} iy \\ xy \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\
 &= \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + \bar{x} \sum_{i=1}^n y_i + \bar{y} \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i \end{bmatrix} \\
 &= \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + n\bar{x}\bar{y} + n\bar{x}\bar{y} - n\bar{x}\bar{y} \end{bmatrix} \\
 &= \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + n\bar{x}\bar{y} \end{bmatrix} \\
 &= \begin{bmatrix} 20n \\ 30 + 200n \end{bmatrix} = \begin{bmatrix} 440 \\ 4430 \end{bmatrix}
 \end{aligned}$$

Finally, computing the least squares estimator $\hat{\beta}$

$$\begin{aligned}
 \hat{\beta} &= (X'X)^{-1} X'Y \\
 &= \frac{1}{60n} \begin{bmatrix} 60 + 100n - 10n & \\ -10n & n \end{bmatrix} \begin{bmatrix} 20n \\ 30 + 200n \end{bmatrix} \\
 &= \begin{bmatrix} 15 \\ 0.5 \end{bmatrix}
 \end{aligned}$$

Assumptions in CLR model are taken to justify the OLS estimation method. We know that the OLS is BLUE under CLR model. Although the asymptotic properties are important, here we have only a few observations $n = 22$. So, the asymptotic properties may not be applicable.

(b) Test the statistical significance of your estimates at the 5% significance level assuming that the y_i 's are jointly normally distributed with variance-covariance matrix the identity matrix.

You may proceed of testing the coefficient separately ($\beta_1 = 0$) and ($\beta_2 = 0$) or testing them jointly as ($\beta_1 = 0$ and $\beta_2 = 0$), here shows a joint significance test of both restrictions.

$$\text{Construct } \Gamma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \gamma_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ such that } \Gamma \hat{\beta} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 15 \\ 0.5 \end{bmatrix}$$

Given the variance covariance matrix is an identity matrix, i.e. $Var(y|X) = Var(X\beta + \varepsilon|X) = Var(\varepsilon|X) = \sigma^2 = I_2$. We can construct a test statistics and compare it with the critical value from a Chi-square distribution

$$W \leq c_{\chi^2(2), \alpha=0.05}^* \approx 5.99$$

$$\begin{aligned} W &= (\Gamma \hat{\beta} - \gamma_0)' (\sigma^2 \Gamma (X'X)^{-1} \Gamma')^{-1} (\Gamma \hat{\beta} - \gamma_0) \\ &= \left(\begin{bmatrix} 15 \\ 0.5 \end{bmatrix} \right)' \left(I_2 \cdot \frac{1}{1320} \begin{bmatrix} 2260 & -220 \\ -220 & 22 \end{bmatrix} \cdot I_2 \right)^{-1} \left(\begin{bmatrix} 15 \\ 0.5 \end{bmatrix} \right)' \\ &= [15 \ 0.5] \begin{bmatrix} 22 & 220 \\ 220 & 2260 \end{bmatrix} \begin{bmatrix} 15 \\ 0.5 \end{bmatrix} = 8815 \end{aligned}$$

We reject the null hypothesis as $8815 > 5.99$.

(c) Test the hypothesis that there exist constant returns to labor at the 5% significance level under the same assumption as in part (b).

Here we want to test if the slope is significantly different from zero, so $\Gamma = [0 \ 1]$ and $\gamma_0 = 1$. Given the $\sigma^2 = I_2$ is known, we may use the standard normal distribution and calculate the Z statistics. We cannot reject the null hypothesis if

$$|Z| \leq c_{N(0,1), \alpha/2=0.025}^* \approx 1.96$$

where Z is calculated as

$$\begin{aligned} Z &= \frac{\Gamma \hat{\beta} - \gamma_0}{\sqrt{\sigma^2 \Gamma (X'X)^{-1} \Gamma'}} \\ &= \frac{[0 \ 1] \begin{bmatrix} 15 \\ 0.5 \end{bmatrix} - 1}{\sqrt{I_2 [0 \ 1] \frac{1}{1320} \begin{bmatrix} 2260 & -220 \\ -220 & 22 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}}} \\ &\approx -3.87 \end{aligned}$$

We reject the null hypothesis and there is not constant returns to scale.

A previous remark from T. Xiang: One could make an argument (Goldberger 1991) on economic versus statistical significance of the coefficient. That is, even when we are statistically rejecting the null of constant returns to labor, our point estimate and the associated confidence interval are more consistent with constant rather than with increasing returns to scale.

But in that case, we may take a one-sided test that $\beta_2 \geq 1$. I would prefer rejecting the null, but keep in mind about what we are doing is on statistical significance.

Problem 5.

I posted two versions of program codes: my Matlab .m file and Alvaro's STATA .do file for your reference.

a) The two measures are not identical: the 1st measure of mean wage is 7.8396 and the 2nd is 9.0239.

b) The sample means of *SOUTH*, *FE*, *UNIO*, *NONWH*, *HISP* are 0.2921, 0.4588, 0.1798, 0.1255, 0.0506, respectively.

There are 289 males, 245 females, 145 south residents, 378 non-south residents, 96 union workers, 438 non-union workers,

440 whites, 27 hispanics, 67 non-whites and non-hispanics in the sample.

c)

	<i>LNWAGE</i>	<i>EDU</i>	<i>EX</i>
Sample Mean	2.0592	13.0187	17.8221
Sample Std. Deviation	0.5277	2.6154	12.3797

LNWAGE :

	Male	Female	Whites	Hisp.	Non-Whites & Non-Hisp	Union	Non-union workers
Sample Mean	2.1653	1.9340	2.0881	1.8185	1.9663	2.2934	2.0078
Std. Deviation	0.5344	0.4921	0.5279	0.5295	0.4968	0.4237	0.53475

Union workers have the highest average log-wage.

Non-union workers have the highest log-wage dispersion.

EDU :

	Male	Female	Whites	Hisp.	Non-Whites & Non-Hisp	Union	Non-union workers
Sample Mean	13.0138	13.0245	13.1682	11.5185	12.6418	12.8854	13.0479
Std. Deviation	2.7676	2.4292	2.4761	4.0513	2.6036	2.6353	2.6131

Whites have the highest average education.

Hispanic workers have the highest dispersion in schooling.

d)

$$LNWAGE = 0.5941 + 0.0964 * EDU + 0.0118 * EX$$

$$(0.1244) \quad (0.0083) \quad (0.0018)$$

Since

$$\frac{\Delta wage}{wage} \approx \Delta LNWAGE = \beta_2 \Delta EDU + \beta_3 \Delta EX,$$

β_2 and β_3 measures the *percentage change* in wage "associated" with a unit change in schooling and experience, holding everything else constant, respectively.

Since EX is defined to be age minus schooling minus 6, age is a linear combination of the constant regressor, EDU and EX . So adding age would violate the full-column-rank condition for the X matrix: there would be no unique solution to the least-squares regression problem.

e)

$$R^2 = 0.2115,$$

which means EDU and EX (or, equivalently, EDU and age) could "explain" about 1/5 of the sample variation in $LNWAGE$: "The linear model fits the data better than nothing".

f)

$$H_0 : \beta_3 = 0$$

$$H_A : \beta_3 > 0$$

Note this is a one-sided test. Under the CNR model, the t-stat is

$$\frac{0.0118}{0.0018} = 6.7069 > 1.64 \approx t_{5\%}(531)$$

Therefore we can accept the alternative hypothesis that the coefficient on EX is statistically significantly greater than 0.

g)

$$LNWAGE = \begin{matrix} 0.5203 & + & 0.0898 * EDU & + & 0.0349 * EX & - & 0.0005 * EXSQ \\ (0.1236) & & (0.0083) & & (0.0056) & & (0.0001) \end{matrix}$$

The coefficient on EX^2 is statistically significant ($t = -4.3067$) and its sign is consistent with the hypothesis.

To find the level of experience where wage is maximized, take FOC with respect to EX :

$$\begin{aligned} \frac{\partial LNWAGE}{\partial EX} &= \beta_3 + 2\beta_4 EX^* = 0 \\ \implies EX^* &= -\frac{\beta_3}{2\beta_4} = \frac{0.0349}{2 \times 0.0005} = 32.5803. \end{aligned}$$

As a result of adding $EXSQ$, the coefficient on EDU drops from 0.0964 to 0.0898, but its std. error stays roughly the same. The coefficient on EX increases from 0.0118 to 0.0349 and its std. error more than triples. Adding $EXSQ$ does not appreciably increase the std. error of the coefficient on EDU but increases that of the coefficient on EDU , because it is highly correlated with EX .

h)

$$\begin{aligned} Wage &= \alpha_M \cdot e^{\beta EDU + \dots + \varepsilon} & \rightarrow & \ln wage = \ln \alpha_M + \beta EDU + \dots + \varepsilon \\ Wage &= \alpha_F \cdot e^{\beta EDU + \dots + \varepsilon} & \rightarrow & \ln wage = \ln \alpha_F + \beta EDU + \dots + \varepsilon \end{aligned}$$

which is equivalent to running the following regression:

$$LNWAGE = \begin{matrix} 0.6007 & - & 0.2570 * FE & + & 0.0913 * EDU & + & 0.0360 * EX & - & 0.0005 * EXSQ \\ (0.1195) & & (0.0387) & & (0.0080) & & (0.0054) & & (0.0001) \end{matrix}$$

where the coefficients $\alpha_1 + \alpha_2 = \ln \alpha_F$ and $\alpha_1 = \ln \alpha_M$.
To test for gender discrimination, the null hypothesis is

$$H_0 : \alpha_2 = 0$$

$$|t| = \left| \frac{-0.2570}{0.0387} \right| = |6.6408| > t_{cv}. \text{ So we reject the null.}$$

i) γ_1 and γ_2 allow for different intercepts for males and females.
 $\gamma_1 = \ln \alpha_M$ and $\gamma_2 = \ln \alpha_F$. The regression result is

$$LNWAGE = \begin{array}{r} 0.6007 * MA + 0.3437 * FE + 0.0913 * EDU + 0.0360 * EX - 0.0005 * EXSQ \\ (0.1195) \qquad (0.1218) \qquad (0.0080) \qquad (0.0054) \qquad (0.0001) \end{array}$$

So $\gamma_1 = \alpha_1$ and $\gamma_2 = \alpha_1 + \alpha_2$.

To test for gender discrimination, formulate the null hypothesis $H_0 : \gamma_1 = \gamma_2$. The t-test statistic

$$v = \frac{hb}{\sqrt{h\widehat{V}(b)h'}} = 6.6409 > t_{cv},$$

where $h = [1 \ -1 \ 0 \ 0 \ 0]$. So we reject the null. Including an intercept term would cause the regressor matrix to be short-ranked, because $MA + FE = 1$.

j) To test whether the returns to education is different for male and females we can run the following regression:

$$LNWAGE = \begin{array}{r} 0.7896 * MA + 0.0358 * FE + 0.0762 * EDU \\ (0.1402) \qquad (0.1713) \qquad (0.0099) \\ + 0.0381 * (EDU \cdot FE) + 0.0369 * EX - 0.0006 * EXSQ \\ (0.0150) \qquad (0.0054) \qquad (0.0001) \end{array}$$

Adding the regressor $(EDU \cdot FE)$ allows for different coefficients on education across genders.

Test: $H_0 : \beta_4 = 0$. We get $|t| = 2.54 > t_{cv} \Rightarrow$ reject H_0 .

k) Add the interaction term $(EDU \cdot EX)$ to detect the different returns to experience across education level (see notes)

$$LNWAGE = \begin{array}{r} 0.3100 - 0.2544 * FE + 0.1114 * EDU + 0.0529 * EX \\ (0.2119) \ (0.0387) \qquad (0.0145) \qquad (0.0115) \\ - 0.0006 * EXSQ - 0.0010 * (EDU \cdot EX) \\ (0.0001) \qquad (0.0006) \end{array}$$

$H_0 : \beta_6 = 0$. $|t| = 1.6667 < t_{cv} \Rightarrow$ accept H_0 . Be careful with the economic implication of this result. Economic significance is quite another matter.

1) Run the following regression:

$$\begin{aligned}
 LNWAGE = & \quad 0.5803 - 0.2309 * FE + 0.0907 * EDU + 0.0344 * EX \\
 & \quad (0.1179) \quad (0.0387) \quad \quad (0.0079) \quad \quad (0.0054) \\
 & \quad \quad \quad - 0.0005 * EXSQ + 0.2022 * UNIO \\
 & \quad \quad \quad (0.0001) \quad \quad (0.0505)
 \end{aligned}$$

To test for union wage premium, specify the null hypothesis: $H_0 : \beta_6 = 0$. $|t| = 4.0040 > t_{cv} \Rightarrow$ reject H_0 . Note: keep FE in the regressor in view of the fact that female workers are less likely to be union workers, because

$$\rho_{FE,UNIO} = -0.1570.$$

Thus failing to do so would lead to an exaggerated union wage premium, part of which should be attributed to gender difference rather than union status.