

Take-home Graded Practice Opportunity

Due date: October 21, 2011, 5:00 p.m.

1. Prove that  $\mathbf{X}'\mathbf{X}$  is a positive definite matrix if  $\mathbf{X}$  is full rank. That is, show that  $\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} > 0$ , where  $\mathbf{c}$  is a scalar such that  $\mathbf{c} \neq 0$ . *Hint: if  $\mathbf{X}$  is full rank, then  $\mathbf{X}\mathbf{c} \neq 0$  for any  $c \neq 0$ .*

*Let  $\mathbf{X}\mathbf{c} = \mathbf{Z}$ . Then it follows that  $\mathbf{Z}'\mathbf{Z}$  is a positive definite matrix (intuitively, this is like squaring a term). Therefore, because  $\mathbf{c} \neq 0$ , then we know that  $\mathbf{X}$  must be a full rank matrix.*

2. Prove that the projection and residual maker matrices are (a) symmetric and (b) idempotent.

*To show symmetry, you must show that  $\mathbf{Z} = \mathbf{Z}'$ ; to show idempotence, you must show that  $\mathbf{Z} = \mathbf{Z} \cdot \mathbf{Z}$ .*

- (a) *The projection matrix is  $P_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . It is symmetric because:*

$$\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}' = \{(\mathbf{X}')'[(\mathbf{X}'\mathbf{X})^{-1}]'(\mathbf{X})'\} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

*Idempotence can be shown as follows:*

$$\begin{aligned} P_X \cdot P_X &= (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

- (b) *Residual maker matrix is  $M_X = I_n - P_X$ . It is symmetric because:*

$$M_X' = (I_n - P_X)' = I_n' - P_X' = I_n - P_X$$

*where  $P_X' = P_X$  is due to the result in part (a).*

*Idempotence can be shown as follows:*

$$\begin{aligned} M_X \cdot M_X &= (I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') (I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= (I_n \cdot I_n) - (I_n)(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') - (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') (I_n) + \\ &\quad (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= I_n - P_X - P_X + P_X \cdot P_X \\ &= I_n - P_X - P_X + P_X \\ &= I_n - P_X \end{aligned}$$

3. Consider a linear model:  $y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ . Suppose that  $E[\mathbf{X}'\varepsilon] = 0$  and  $\text{Var}[\varepsilon|\mathbf{X}] = \sigma^2$ . However,  $E[\varepsilon|\mathbf{X}] \neq 0$ .

(a) Does  $E[\varepsilon^2|\mathbf{X}] = \sigma^2$ ?

$$\sigma_\varepsilon^2 = E[\varepsilon^2|\mathbf{X}] - E[\varepsilon|\mathbf{X}]^2$$

*Because the second term,  $E[\varepsilon|\mathbf{X}]^2$ , does not equal zero, we cannot conclude that the equality holds.*

(b) What does  $E[\varepsilon|\mathbf{X}] \neq 0$  imply about the OLS estimators?

*This implies that the OLS estimator will not be unbiased.*

4. The following problem will help you understand the effects of scaling on linear OLS regression analysis.

(a) Let  $\hat{\boldsymbol{\beta}}_0$  and  $\hat{\boldsymbol{\beta}}_1$  be the estimated parameters from the regression of  $\mathbf{Y}$  on  $\mathbf{X}$ . Furthermore, assume that  $c_1$  and  $c_2$  are constants, with  $c_2 \neq 0$ . Now, assume that  $\tilde{\boldsymbol{\beta}}_0$  and  $\tilde{\boldsymbol{\beta}}_1$  are the estimated parameters from the regression of  $c_1\mathbf{Y}$  on  $c_2\mathbf{X}$ .

Show that  $\tilde{\boldsymbol{\beta}}_1 = (c_1/c_2)\hat{\boldsymbol{\beta}}_1$  and  $\tilde{\boldsymbol{\beta}}_0 = c_1\hat{\boldsymbol{\beta}}_0$ .

What does this imply about scaling the dependent and independent variables? How does scaling affect the estimated parameters?

*Recall that we can write the estimators using the following summation notations:*

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X} \\ \hat{\beta}_1 &= \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sum_i (x_i - \bar{X})^2}\end{aligned}$$

*Therefore, we can mimic this for the  $\tilde{\boldsymbol{\beta}}$  estimators. Starting with  $\tilde{\beta}_1$ :*

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_i (c_2 x_i - c_2 \bar{X})(c_1 y_i - c_1 \bar{Y})}{\sum_i (c_2 x_i - c_2 \bar{X})^2} \\
&= \frac{c_1 c_2 \sum_i (x_i - \bar{X})(y_i - \bar{Y})}{c_2^2 \sum_i (x_i - \bar{X})^2} \\
&= \frac{c_1}{c_2} \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sum_i (x_i - \bar{X})^2} \\
&= \frac{c_1}{c_2} \hat{\beta}_1
\end{aligned}$$

Now, we can plug in  $\tilde{\beta}_1$  to retrieve  $\tilde{\beta}_0$ :

$$\begin{aligned}
\tilde{\beta}_0 &= c_1 \bar{Y} - \tilde{\beta}_1 c_2 \bar{X} \\
&= c_1 \bar{Y} - \frac{c_1}{c_2} \hat{\beta}_1 c_2 \bar{X} \\
&= c_1 (\bar{Y} - \hat{\beta}_1 \bar{X}) \\
&= c_1 \hat{\beta}_0
\end{aligned}$$

Using matrix algebra, this result is even easier to show. Suppose that we write the regression equation with the scaling factors:

$$c_1 \mathbf{Y} = c_2 \mathbf{X} \boldsymbol{\beta} + \varepsilon$$

This implies that the estimated coefficient vector can be denoted as follows:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}' c_2' c_2 \mathbf{X})^{-1} \mathbf{X}' c_2' c_1 \mathbf{Y}$$

However,  $c_1$  and  $c_2$  are scalars, so they can simply be pulled outside:

$$\begin{aligned}
\tilde{\beta} &= (\mathbf{X}'c_2c_2\mathbf{X})^{-1}\mathbf{X}'c_2c_1\mathbf{Y} \\
&= \frac{1}{c_2^2}(\mathbf{X}'\mathbf{X})^{-1}(c_1c_2)\mathbf{X}'\mathbf{Y} \\
&= \frac{c_1}{c_2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \frac{c_1}{c_2}\hat{\beta}
\end{aligned}$$

The last step is simply from the fact that  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\beta}$ . The only intuitive step that has to be made is to understand that  $\tilde{\beta}_0$  describes the intercept of  $\mathbf{Y}$ , and so is not scaled by the factor  $c_2$ , which applies only to  $\mathbf{X}$ . Therefore,  $\tilde{\beta}_0 = c_1\hat{\beta}_0$ . This is the same result as above.

Thus, scaling of the variables simply implies that the intercept is scaled by the amount of the scaling of the dependent variable, and the independent parameter is scaled by the ratio of the scaling factors of  $Y$  and  $X$ . This is a useful result, because you can change the units in a particular regression model, but it will only change the fit of the regression line proportionately to the scaling. Therefore, you can estimate a regression with any units, and then scale the estimated coefficients without having to re-estimate.

- (b) Let  $\hat{\beta}_0$  and  $\hat{\beta}_1$  be OLS estimates from the regression of  $\log(y)$  on  $\mathbf{X}$ . For a constant  $c > 0$ , let  $\tilde{\beta}_0$  and  $\tilde{\beta}_1$  be estimates from a regression of  $\log(cy)$  on  $\mathbf{X}$ . Show how you can write  $\tilde{\beta}_0$  and  $\tilde{\beta}_1$  as a function of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . How does the constant scale the regression coefficients?

Consider that we can write:

$$\begin{aligned}
\ln(cy) &= \tilde{\beta}_0 + \tilde{\beta}_1X + u \\
\ln y &= (\tilde{\beta}_0 + \tilde{\beta}_1X + u) - \ln c
\end{aligned}$$

Therefore, only the  $\hat{\beta}_0$  coefficient is affected. We can re-write  $\tilde{\beta}_0 = \hat{\beta}_0 - (\ln c)\hat{\beta}_1$  and  $\tilde{\beta}_1 = \hat{\beta}_1$ .

5. Consider the following data set describing the number of days until (negative) and after (positive) harvest and the premium (in cents per bushel) for having a 1% increase in wheat protein levels.

Days	Premium
------	---------

-2	200
-1	180
0	20
1	22
2	20

- (a) Explain the economic intuition about what happened at harvest time to the quality level of wheat. How is that reflected in the premiums? What kind of quality was available before the harvest and what quality is available after the harvest?

*Prior to harvest time, the supply of higher protein wheat was low. Therefore, there was a very high premium for selling wheat that had a marginal increase in quality. At and after the harvest, the premium substantially declined. This indicates that there was an increase in wheat with high protein levels, and the excess supply drives down the premium.*

- (b) Using matrix algebra (by hand), perform the following:

- i. Compute estimated parameters for the model  $Premium = f(Days) + \varepsilon$ .

$$\mathbf{Y} = \begin{bmatrix} 200 \\ 180 \\ 20 \\ 22 \\ 20 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

To find the parameter vector, solve for:  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ .

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 88.4 \\ -51.8 \end{bmatrix}$$

- ii. Determine the variance-covariance matrix.

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.1 \end{bmatrix}$$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{5-2} = 2592.9$$

Therefore, the variance-covariance matrix are as follows:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 518.8 & 0 \\ 0 & 259.3 \end{bmatrix}$$

- iii. Construct a 95% confidence interval around the population parameter associated with the days.

*We need to determine the vector of standard associated with the estimated parameters. This is the square root of the main diagonal in the variance matrix.*

$$SE((\hat{\beta})) = \begin{bmatrix} 22.7 \\ 16.1 \end{bmatrix}$$

*The confidence interval is therefore:  $[-51.8 - 3.182 \cdot 16.1, -51.8 + 3.182 \cdot 16.1] = [-103, -0.561]$  where the critical value is from the  $t$ -distribution with 3 degrees of freedom.*

- iv. Test a 90% confidence level hypothesis that the average premium before the harvest was different than the average premium after the harvest. Interpret.

*Setting up the hypothesis is as follows:*

$$\begin{aligned} H_0 &: \text{Premium}_{t<0} = \text{Premium}_{t \geq 0} \\ H_a &: \text{Premium}_{t<0} \neq \text{Premium}_{t \geq 0} \end{aligned}$$

*The test statistic is:*

$$t_{stat} = \frac{\overline{\text{Premium}}_{t<0} - \overline{\text{Premium}}_{t \geq 0}}{\sqrt{\frac{s_{t<0}^2}{n_{t<0}} + \frac{s_{t \geq 0}^2}{n_{t \geq 0}}}} = \frac{190 - 20.7}{\sqrt{100 + 0.44}} = 16.89$$

*The associated critical value is from the  $t$ -distribution. The degrees of freedom is the total number of observations minus the total number of parameters that are in the test; that is, two. Therefore:  $t_{10\%,3} = 2.353$ .*

*The calculated  $t$ -statistic clearly exceeds the critical value. Therefore, we must reject the null in favor of the alternative that the premiums before and after harvest are not equal.*

## 6. Applied Practice

Use the data set “Heart” available under the “Homeworks” category on the class website. Complete the following:

- Explore the data using a software package. Provide some basic intuition about the relationships between individuals’ weight and their diastolic and systolic blood pressure measures. No regressions.
- Read only the columns “Weight” and “Diastolic” from this data set into IML (or another matrix language). Read each column into a separate vector.

- (c) Using matrix algebra, compute the parameter estimates of the model:  
 $Diastolic = \beta_0 + \beta_1 Weight + \varepsilon$ .
- (d) Compute the variance-covariance matrix and the standard errors vector. What are the standard errors for each estimated parameter?
- (e) For each estimated parameter, test the hypothesis (95% confidence level) that the true population value of the parameter is zero. At the 95% confidence level, what is the marginal effect of weight on the diastolic blood pressure?
- (f) Using the result in (e), what are the potential economic implications of a national health plan? Think about what is happening with the number of individuals in the United States who are overweight. Who might be better off? Who might be worse off?
- (g) Hypertension is a condition describing high blood pressure. What are some adverse effects of hypertension? At what level does hypertension become critical to one's health? How much weight, on average, would people in the sample need to get to a dangerous level of hypertension?

*SAS code is available on the course website.*