Written Assignment

ECO 4000 Statistical Analysis for Economics and Finance

12 April, 2022

Maximum Points : 25

Due Date : 13 April, 2022 (Before Class)

1. $Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$**. where,** Y **are the observed values,** $\hat{\beta}_0 + \hat{\beta}_1 X = \hat{Y}$ **are the fitted values, and** e **are the residuals generated from the OLS regression. Prove the following algebraic properties of the OLS (5 points)(Answers at the end)**

a. $\overline{e} = 0$ (1 point)

b. $\overline{\hat{Y}} = \overline{Y}$ (1 point)

c. $cov(X, e) = 0$ (1 point)

d. $cov(\hat{Y}, e) = 0$ (2 points)

2. Determine whether following models are linear regression models or not.(2 points)

1. $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i} \right) + u_i$ - Linear, "because linear in parameters"
2. $Y_i = \beta_0 + \beta_1 lnX_i + u_i$ - Linear, "because linear in parameters"
3. $lnY_i = \beta_0 + \beta_1^2 lnX_i + u_i$ - Non-linear, "because not linear in parameters"
4. $Y_i = \beta_0 + (0.75 - \beta_0)e^{-\beta_2(X_i - 2)} + u_i$ - Non-linear, "because not linear in parameters"

3. Following are three data points on dependent (Y) and one explanatory variable(X). Fit a regression model by calculating $\hat{\beta}_0$ **and** $\hat{\beta}_1$**.(3 points)(Answers at the end)**

Y	X
8	10
15	7
24	3

4. The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep. where sleep and totwrk (total work) are measured in minutes per week and educ and age are measured in years.(6 points)

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u$$

a. If adults trade off sleep for work, what is the sign of β_1?(1 point)

   - If adults trade off sleep for work, more work implies less sleep (other things equal), so $\beta_1 < 0$.

b. What signs do you think β_2 and β_3 will have?(1 point)

   - The signs of β_2 and β_3 are not obvious, at least to me. One could argue that more educated people like to get more out of life, and so, other things equal, they sleep less ($\beta_2 < 0$). The relationship between sleeping and age is more complicated than this model suggests, and economists are not in the best position to judge such things.

c. Using the data, the estimated equation is

$$\widehat{sleep} = \underset{(112.28)}{3,638.25} - \underset{(0.017)}{0.148} \times totwrk - \underset{(5.88)}{11.13} \times educ + \underset{(1.45)}{2.20} \times age$$

$$n = 706, R^2 = 0.113$$

If someone works five more hours per week, by how many minutes is sleep predicted to fall?(1 point)

- Since *totwrk* is in minutes, we must convert five hours into minutes: Δ*totwrk* = 5(60) = 300. Then sleep is predicted to fall by .148(300) = 44.4 minutes. For a week, 45 minutes less sleep is not an overwhelming change.

d. Discuss the sign and magnitude of the estimated coefficient on *educ*.(1 point)

   - More education implies less predicted time sleeping, but the effect is quite small. If we assume the difference between college and high school is four years, the college graduate sleeps about 45 minutes less per week, other things equal.

e. Would you say *totwrk*, *educ*, and *age* explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are those likely to be correlated with *totwrk*? If we do not include those factors in our model, what kind of bias the coefficient on *totwrk* will have?(2 points)

   - Not surprisingly, the three explanatory variables explain only about 11.3% of the variation in sleep. One important factor in the error term is general health. Another is marital status, and whether the person has children. Health (however we measure that), marital status, and number and ages of children would generally be correlated with totwrk. (For example, less healthy people would tend to work less.)

5. Once again using the same study as above for our purposes, the estimated equation is as follows:(6 points)

$$\widehat{sleep} = \underset{(112.28)}{3,638.25} - \underset{(0.017)}{0.148} \times totwrk - \underset{(5.88)}{11.13} \times educ + \underset{(1.45)}{2.20} \times age$$

$$n = 706, R^2 = 0.113$$

a. Is either $educ$ or age **individually** significant at the 5% level against a two-sided alternative? Show your work.(1 point)

- With df = 706 – 4 = 702, we use the standard normal critical value which is 1.96 for a two-tailed test at the 5% level. Now t_{educ} = −11.13/5.88 ≈ −1.89, so $|t_{educ}|$ = 1.89 < 1.96, and we fail to reject H_0: β_{educ} = 0 at the 5% level. Also, t_{age} ≈ 1.52, so age is also statistically insignificant at the 5% level.

b. Construct a 95% confidence interval for a coefficient on *total work*. Show your work.(1 point)

$$[\hat{\beta}_1 - t_{(\frac{\alpha}{2}=0.025, df=702)} \times s.e(\hat{\beta}_1), \hat{\beta}_1 + t_{(\frac{\alpha}{2}=0.025, df=702)} \times s.e(\hat{\beta}_1)]$$

$$[-0.148 - 1.96 \times 0.017, -0.148 + 1.96 \times 0.017]$$

$$[-0.1812, -0.1148]$$

c. Dropping educ and age from the equation gives

$$\widehat{sleep} = \underset{(38.91)}{3,586.38} - \underset{(0.017)}{0.151} \times totwrk$$

$$n = 706, R^2 = 0.103$$

Are educ and age **jointly** significant in the original equation at the 5% level? Justify your answer.(2 points)

$H_0 : \beta_{educ} = 0 \ \& \ \beta_{age} = 0$
$H_1 : \beta_{educ} \neq 0 \ \& \ \beta_{age} \neq 0$

- We need to compute the R^2 form of the F - statistic for joint significance.

$$F = \frac{\frac{(0.113-0.103)}{2}}{\frac{(1-0.113)}{702}} \approx 3.96$$

- The 5% critical value in the $F_{2,702}$ distribution can be obtained from the F-table with denominator $df = \infty$, which is 3.00. Therefore, educ and age are jointly significant at the 5% level (3.96 > 3.00).

d. Does including educ and age in the model greatly affect the estimated trade off between sleeping and working?(1 point)

- Not really. These variables are jointly significant, but including them only changes the coefficient on *totwrk* from –.151 to –.148.

e. Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts **a** and **c**?(1 point)

- The standard t and F statistics that we used assume homoskedasticity, in addition to the other CLM assumptions. If there is heteroskedasticity in the equation, the tests are no longer valid.

6. Continuing with the same study as above, now consider the estimated equation as follows:(3 points)

$$\widehat{sleep} = \underset{(235.11)}{3,840.83} - \underset{(0.018)}{0.163} \times totwrk - \underset{(5.86)}{11.71} \times educ - \underset{(11.21)}{8.70} \times age + \underset{(0.134)}{0.128} \times age^2 + \underset{(34.33)}{87.75} \times male$$

$$n = 706, R^2 = 0.123, \overline{R^2} = 0.117$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

a. All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?(1 point)

   - The coefficient on male is 87.75, so a man is estimated to sleep almost one and one-half hours more per week than a comparable woman. Further, tmale = 87.75/34.33 ≈ 2.56, which is close to the 1% critical value against a two-sided alternative (about 2.58). Thus, the evidence for a gender differential is fairly strong.

b. Is there a statistically significant tradeoff between working and sleeping? What is the estimated trade off?(1 point)

   - The t statistic on *totwrk* is −.163/.018 ≈ −9.06, which is very statistically significant. The coefficient implies that one more hour of work (60 minutes) is associated with .163(60) ≈ 9.8 minutes less sleep.

c. What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?(1 point)

   - To obtain R_r^2 , the R-squared from the restricted regression, we need to estimate the model without *age* and age^2. When *age* and age^2 are both in the model, age has no effect only if the parameters on both terms are zero.

1. (a) $\bar{e} = 0$

$\Rightarrow e = Y - \hat{Y}$

$\Rightarrow e = Y - \hat{\beta}_0 - \hat{\beta}_1 X$ $\qquad \because \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Taking expectations on both sides,

$\Rightarrow E(e) = E[Y - \hat{\beta}_0 - \hat{\beta}_1 X]$ $\qquad E(X) = \bar{X}$

$\Rightarrow \bar{e} = E(Y) - E(\hat{\beta}_0) - E(\hat{\beta}_1 X)$ $\qquad \because E(X+Y) = E(X) + E(Y)$

$\Rightarrow \bar{e} = \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}$ $\qquad \because E(AX) = AE(X)$

$\Rightarrow \bar{e} = \bar{Y} - [\bar{Y} - \hat{\beta}_1 \bar{X}] - \hat{\beta}_1 \bar{X}$ $\qquad \because \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$\Rightarrow \bar{e} = \cancel{\bar{Y}} - \cancel{\bar{Y}} + \cancel{\hat{\beta}_1 \bar{X}} - \cancel{\hat{\beta}_1 \bar{X}}$

$\Rightarrow \bar{e} = 0$

1. (b) $\bar{\hat{Y}} = \bar{Y}$

$e = Y - \hat{Y}$

Taking expectations on both sides,

$E(e) = E(Y) - E(\hat{Y})$

$\Rightarrow \bar{e} = \bar{Y} - \bar{\hat{Y}}$ $\qquad \because E(X) = \bar{X}$

$\Rightarrow 0 = \bar{Y} - \bar{\hat{Y}}$ $\qquad \because \bar{e} = 0$

$\Rightarrow \bar{\hat{Y}} = \bar{Y}$

1 (c) $\operatorname{cov}(X, e) = 0$

$= \operatorname{cov}(X, Y - \hat{Y})$ ∵ $\operatorname{cov}(X, Y+Z) = \operatorname{cov}(X+Y, X+Z)$

$= \operatorname{cov}(X, Y - \hat{\beta}_0 - \hat{\beta}_1 X)$

$= \operatorname{cov}(X, Y) + \operatorname{cov}(X, -\hat{\beta}_0) + \operatorname{cov}(X, -\hat{\beta}_1 X)$

$\operatorname{cov}(X, AX) = A\operatorname{cov}(X,X)$

$= \operatorname{cov}(X, Y) + 0 - \hat{\beta}_1 \operatorname{cov}(X, X)$ ∵ $\operatorname{cov}(A, X) = 0$.

$= \operatorname{cov}(X, Y) - \hat{\beta}_1 \cdot \operatorname{var}(X)$ ∵ $\operatorname{cov}(X, X) = \operatorname{var}(X)$

$= \operatorname{cov}(X, Y) - \dfrac{\operatorname{cov}(X, Y)}{\cancel{\operatorname{var}(X)}} \cdot \cancel{\operatorname{var}(X)}$ ∵ $\hat{\beta}_1 = \dfrac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$

$= 0$

1 (d) $\operatorname{cov}(\hat{Y}, e) = 0$

$= \operatorname{cov}(\hat{\beta}_0 + \hat{\beta}_1 X, e)$ ∵ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$= \operatorname{cov}(\hat{\beta}_0, e) + \operatorname{cov}(\hat{\beta}_1 X, e)$ ∵ $\operatorname{cov}(A, X) = 0$

$= \underbrace{\operatorname{cov}(\hat{\beta}_0, e)}_{0} + \hat{\beta}_1 \underbrace{\operatorname{cov}(X, e)}_{0}$ ∵ $\operatorname{cov}(AX, Y) = A\operatorname{cov}(X, Y)$

$\operatorname{cov}(X, e) = 0$ from 1(c).

$= 0$.

3.

y	x	$\hat{y} = b_0 + b_1 x$	$e = y - \hat{y}$
8	10	$b_0 + 10b_1$	$8 - b_0 - 10b_1$
15	7	$b_0 + 7b_1$	$15 - b_0 - 7b_1$
24	3	$b_0 + 3b_1$	$24 - b_0 - 3b_1$

(I am using b_0 and b_1 as estimates for the sake of simplicity)

The Goal: We want to find estimates b_0 and b_1 such that $SSR = e_1^2 + e_2^2 + e_3^2$ is minimized.

$$SSR = e_1^2 + e_2^2 + e_3^2$$

$$SSR = (8 - b_0 - 10b_1)^2 + (15 - b_0 - 7b_1)^2 + (24 - b_0 - 3b_1)^2$$

Now, to minimize this function with respect to b_0,

$$\frac{\partial SSR}{\partial b_0} = 2(8 - b_0 - 10b_1)(-1) + 2(15 - b_0 - 7b_1)(-1) + 2(24 - b_0 - 3b_1)(-1) = 0.$$

$$\Rightarrow 8 - b_0 - 10b_1 + 15 - b_0 - 7b_1 + 24 - b_0 - 3b_1 = 0$$

$$\Rightarrow 47 - 3b_0 - 20b_1 = 0$$

$$\Rightarrow 3b_0 + 20b_1 = 47 \quad \text{— (1)}$$

To minimize the function with respect to b_1,

$$\frac{\partial SSR}{\partial b_1} = 2(8-b_0-10b_1)(-10) + 2(15-b_0-7b_1)(-7) + 2(24-b_0-3b_1)(-3) = 0.$$

$$\Rightarrow 10(8-b_0-10b_1) + 7(15-b_0-7b_1) + 3(24-b_0-3b_1) = 0$$

$$\Rightarrow 80 - 10b_0 - 100b_1 + 105 - 7b_0 - 49b_1 + 72 - 3b_0 - 9b_1 = 0$$

$$\Rightarrow 257 - 20b_0 - 158b_1 = 0$$

$$\Rightarrow 20b_0 + 158b_1 = 257 \quad \text{— } ②$$

from equations ① & ②,

$$3b_0 + 20b_1 = 47$$
$$20b_0 + 158b_1 = 257$$

{ 2 unknowns, 2 equations, I used calculator }

$$\hat{\beta}_0 = b_0 = 30.89$$

$$\hat{\beta}_1 = b_1 = -2.28$$

Regression Model:

$$\hat{Y} = 30.89 - 2.28X$$

must answer in this way