Practice Final
Econ471
Fall 2020
Solutions

1. In a study of early retirements, the following equations were estimated using 1980 census data for 44 States (values in parentheses are t ratios):

$$\text{RETRD} = \underset{(-0.4)}{-3.930} + \underset{(5.4)}{1.627}\ \text{HLTH} - \underset{(-0.26)}{0.0005}\ \text{MSSEC} + \underset{(0.28)}{0.00005}\ \text{MPUBAS}$$

(A) $$+ \underset{(2.2)}{0.549}\ \text{UNEMP} + \underset{(1.6)}{0.153}\ \text{DEP} + \underset{(2.3)}{0.077}\ \text{RACE}$$

$R^2 = 0.702$ s = 2.175 SSR = 175.088

$$\text{RETRD} = \underset{(-1.6)}{-5.093} + \underset{(6.5)}{1.596}\ \text{HLTH} + \underset{(2.3)}{0.549}\ \text{UNEMP} + \underset{(1.8)}{0.153}\ \text{DEP}$$

(B) $$+ \underset{(3.5)}{0.083}\ \text{RACE}$$

$R^2 = 0.701$ s = 2.121 SSR = 175.524

Where

RETRD = retired men who are between the ages of 16 and 65
HLTH = percent of people between 16 and 64 years who are prevented from working due to disability
MSSEC = mean social security income
MPUBAS = mean public assistance income
UNEMP = unemployment rate (%)
DEP = percent of households that represent married couples with children under 18
RACE = percent of men who are nonwhite.

a. Calculate the adjusted R^2 for models A and B.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

For A: $n = 44,\ k = 6,\ R^2 = 0.702 \rightarrow \bar{R}^2 = 1 - \frac{44-1}{44-6-1}(1 - 0.702) = 0.654$

For B: $n = 44,\ k = 4,\ R^2 = 0.701 \rightarrow \bar{R}^2 = 1 - \frac{44-1}{44-4-1}(1 - 0.701) = 0.670.$

b. Test both models for overall significance. (Note: For this part and parts c and d below, state your null and alternative hypotheses, test statistics, their distribution and degrees of freedom).

Model A:

$H_o = \beta_1 = \beta_2 = \cdots = \beta_6 = 0$

H_1: H_o not true

$$F = \frac{R^2/k}{(1- \ R^2)/(n-k-1)} = \frac{0.702/6}{(1-0.702)/(44-6-1)} = 14.53$$

Under H_o, *F* has *F* distribution with 6 and 37 degrees of freedom. The 1% critical value with 6 and 40 degrees of freedom is 3.29. We reject H_o. The model is overall relevant.

Model B:

H_o: $\beta_1 = \beta_2 = \cdots = \beta_4 = 0$

H_1: H_o not true

$$F = \frac{R^2/k}{(1- \ R^2)/(n-k-1)} = \frac{0.701/4}{(1-0.701)/(44-4-1)} = 22.86$$

Under H_o, *F* has *F* distribution with 4 and 39 degrees of freedom. The 1% critical value with 4 and 40 degrees of freedom is 3.83. We reject H_o. The model is overall relevant.

c. Test for each regression coefficient (except the constant term) for significance at α = .05.

Model A:

H_o: $\beta_i = 0 \qquad i = 1,2, \ldots\ldots,6$

H_1: $\beta_i \neq 0$

$$t_i = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)}$$

Under H_o, t_i has *t* distribution with $n-k-1 = 37$ degrees of freedom. The 5% critical value with 40 degrees of freedom is 2.01. The calculated t values are given in parentheses below the estimated coefficients. The absolute value of the calculated t is greater than 2.01 for HLTH, UNEMP, and RACE., so we reject H_o for these variables. For the rest we do not reject H_o.

Model B:

$H_o: \beta_i = 0 \qquad i = 1,2, \dots\dots, 4$

$H_1: \beta_i \neq 0$

$$t_i = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)}$$

Under H_o, t_i has *t* distribution with $n - k - 1 = 39$ degrees of freedom. The 5% critical value with 40 degrees of freedom is 2.01. The absolute value of the calculated t is greater than 2.01 for HLTH, UNEMP, and RACE., so we reject H_o for these variables. For DEP the calculated test statistic is less than 2.01 so we do not reject H_o.

d. Test the hypothesis that the coefficients for both MSSEC and MPUBAS in model A are jointly insignificant. Use α = .05.

$H_o: \beta_{MSSEC} = \beta_{MPUBAS} = 0$

H_1: H_o not true

$$F = \frac{(R_A^2 - R_B^2)/q}{(1 - R_A^2)/(n - k - 1)} = \frac{(0.702 - 0.701)/2}{(1 - 0.702)/(44 - 6 - 1)} = 0.062$$

This is smaller than any reasonable critical value so we do not reject H_o.

2. (True or false) In multiple regression, correlation in the sample among the regressors is called multicollinearity and does not imply bias in the least squares coefficients. Justify your answer.

True. As long as there is no perfect multicollinearity among the regressors, the OLS estimators are still unbiased. In fact, they will still be BLUE as long as the assumptions about the linear regression model (MLR1 toMLR5) hold. When there is a high degree of collinearity, it becomes difficult to isolate the effect of a given regressor on the dependent variable.

3.

a. .For the simple linear regression model $Y_i = \beta_1 + \beta_2 X_i + e_i$ what happens to the estimates of the intercept and slope if you add a constant *c* to each observation on Y? Each observation on X?

$$Let\, Y_i^* = Y_i + c \rightarrow \bar{Y}^* = \bar{Y} + c.$$

$$\hat{\beta}_1^* = \frac{\sum (X_t - \bar{X})(Y_t^* - \bar{Y}^*)}{\sum (X_t - \bar{X})^2} = \frac{\sum (X_t - \bar{X})(Y_t + c - \bar{Y} - c)}{\sum (X_t - \bar{X})^2}$$

$$= \frac{\sum (X_t - \bar{X})(Y_t - \bar{Y})}{\sum (X_t - \bar{X})^2} = \hat{\beta}_1.$$

$$\hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X} = \bar{Y} + c - \hat{\beta}_1 \bar{X} = \bar{Y} - \hat{\beta}_1 \bar{X} + c = \hat{\beta}_0 + c.$$

Similarly,

$$Let\, X_i^* = X_i + c \rightarrow \bar{X}^* = \bar{X} + c.$$

$$\hat{\beta}_1^* = \frac{\sum (X_i^* - \bar{X}^*)(Y_i - \bar{Y})}{\sum (X_i^* - \bar{X}^*)^2} = \frac{\sum (X_i + c - \bar{X} - c)(Y_i - \bar{Y})}{\sum (X_i + c - \bar{X} - c)^2}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \hat{\beta}_1.$$

$$\hat{\beta}_0^* = \bar{Y} - \hat{\beta}_1^* \bar{X}^* = \bar{Y} - \hat{\beta}_1(\bar{X} + c) = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 c = \hat{\beta}_0 - \hat{\beta}_1 c.$$

b. What happens to the estimates of the intercept and slope if you subtract the sample mean of X from each observation on X and the sample mean of Y from each observation on Y?

$$Let\, Y_i^* = Y_i - \bar{Y} \rightarrow \bar{Y}^* = \bar{Y} - \bar{Y} = 0$$

Similarly, let

$$X_i^* = X_i - \bar{X} \rightarrow \bar{X}^* = \bar{X} - \bar{X} = 0$$

$$\hat{\beta}_1^* = \frac{\sum (X_i^* - \bar{X}^*)(Y_i^* - \bar{Y}^*)}{\sum (X_i^* - \bar{X}^*)^2} = \frac{\sum (X_i - \bar{X} - 0)(Y_i - \bar{Y} - 0)}{\sum (X_i - \bar{X} - 0)^2}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \hat{\beta}_1.$$

$$\hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X}^* = 0 - \hat{\beta}_1(0) = 0.$$

c. What happens to the estimates of the intercept and slope if you multiply each observation on X by a constant *c*?

$$Let\ X_i^* = cX_i \to \bar{X}^* = c\bar{X}.$$

$$\hat{\beta}_1^* = \frac{\sum(X_i^* - \bar{X}^*)(Y_i - \bar{Y})}{\sum(X_i^* - \bar{X}^*)^2} = \frac{\sum(cX_i - c\bar{X})(Y_i - \bar{Y})}{\sum(cX_i - c\bar{X})^2} = \frac{\sum c(X_i - \bar{X})(Y_i - \bar{Y})}{\sum c^2(X_i - \bar{X})^2}$$

$$= \frac{c\sum(X_i - \bar{X})(Y_i - \bar{Y})}{c^2\sum(X_i - \bar{X})^2} = \frac{1}{c}\hat{\beta}_1.$$

$$\hat{\beta}_0^* = \bar{Y} - \hat{\beta}_1^*\bar{X}^* = \bar{Y} - \frac{1}{c}\hat{\beta}_1(c\bar{X}) = \bar{Y} - \hat{\beta}\bar{X} = \hat{\beta}_0.$$

d. What happens to the estimates of the intercept and slope if you multiply each observation on X and Y by the same constant *c*?

Let

$$X_i^* = cX_i$$

and

$$Y_i^* = cY_i.$$

$$Then\ \bar{X}^* = c\bar{X}\ and\ \bar{Y}^* = c\bar{Y}.$$

$$\hat{\beta}_1^* = \frac{\sum(X_i^* - \bar{X}^*)(Y_i^* - \bar{Y}^*)}{\sum(X_i^* - \bar{X}^*)^2} = \frac{\sum(cX_i - c\bar{X})(cY_i - c\bar{Y})}{\sum(cX_i - c\bar{X})^2} = \frac{\sum c^2(X_i - \bar{X})(Y_i - \bar{Y})}{\sum c^2(X_i - \bar{X})^2}$$

$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \hat{\beta}_1.$$

$$\hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^*\bar{X}^* = c\bar{Y} - \hat{\beta}_1(c\bar{X}) = c(\bar{Y} - \hat{\beta}_1\bar{X}) = c\hat{\beta}_0.$$

4. You are estimating the relationship between a firm's sales and advertising expenditures in an industry. It becomes apparent to you that half the firms in the industry are large relative to the other half, and you are concerned about the proper estimation technique in such a situation. Assume that the error variances associated with the large firms are twice the error variances associated with the small firms.

a. If you used ordinary least squares to estimate the regression of sales on advertising (assuming that advertising is uncorrelated with the error term), would your estimated parameter be unbiased? Consistent? Efficient?

The OLS estimators will still be unbiased since to prove unbiasedness we don’t need homoskedasticity. OLS estimators will also be consistent but they will not be BLUE.

b. If the method used in part a is not efficient, how might you revise the estimation procedure to eliminate or resolve inefficiency?

Since $Var\left(u_{large\ firms}\right) = 2Var(u_{small\ firms})$, we should assign less weight to observations for large firms. To obtain the weighted least squares (WLS) we first multiply the observations for the large firms by $\frac{1}{\sqrt{2}}$ and then apply OLS to the transformed observations.

5. Suppose a sample of adults is classified into groups 1, 2, and 3 on the basis of whether their education stopped in (or at the end of) elementary school, high school, or university, respectively. The relationship $y = \beta_1 + \beta_2 D_2 + \beta_3 D_3 + u$ is specified, where y is income, $D_i = 1$ for those in group *i* and zero for all others.

a. In terms of the parameters of the model, what is the expected income of those whose education stopped in university?

$$E(y|D_2 = 0, D_3 = 1) = \beta_1 + \beta_2(0) + \beta_3(1) = \beta_1 + \beta_3$$

b. How would you test the hypothesis that going on to university after high school makes no contribution to adult income?

For those who didn't go to college, expected income is

$$E(y|D_2 = 1, D_3 = 0) = \beta_1 + \beta_2(1) + \beta_3(0) = \beta_1 + \beta_2.$$

So the difference in the expected income of those who go to college and those who only go to high school is

$$E(y|D_2 = 0, D_3 = 1) - E(y|D_2 = 1, D_3 = 0) = \beta_3 - \beta_2.$$

To test the hypothesis that going on to university after high school makes no contribution to adult income, the null and alternative hypotheses are:

$$H_o: \beta_3 - \beta_2 = 0$$

$$H_1: \beta_3 - \beta_2 > 0$$

The test statistic is

$$t = \frac{\hat{\beta}_3 - \hat{\beta}_2}{[\widehat{Var}(\hat{\beta}_3) + \widehat{Var}(\hat{\beta}_2) - 2\widehat{Cov}(\hat{\beta}_3, \hat{\beta}_2)]^{1/2}}$$

Reject H_o if the calculate value of the test statistic is greater than the critical value (df = n-3). If we fail to reject H_o we conclude that going on to university after high school makes no contribution to adult income.

6. A study tried to find the determinants of the increase in the number of households headed by a female. Using 1940 and 1960 historical census data, a logit model was estimated to predict whether a woman is the head of a household (living on her own) or whether she is living within another's household. The limited dependent variable takes on a value of one if the female lives on her own and is zero if she shares housing. The results for 1960 using 6,051 observations on prime-age whites and 1,294 on nonwhites were as shown in the table:

Regression	**(1) White**	**(2) Nonwhite**
Regression model	Logit	Logit
Constant	1.459 (0.685)	–2.874 (1.423)
Age	–0.275 (0.037)	0.084 (0.068)
age squared	0.00463 (0.00044)	0.00021 (0.00081)
Education	–0.171 (0.026)	–0.127 (0.038)
farm status	–0.687 (0.173)	–0.498 (0.346)
South	0.376 (0.098)	–0.520 (0.180)
expected family earnings	0.0018 (0.00019)	0.0011 (0.00024)
family composition	4.123 (0.294)	2.751 (0.345)
Pseudo-R^2	0.266	0.189
Percent Correctly Predicted	82.0	83.4

where *age* is measured in years, *education* is years of schooling of the family head, *farm status* is a binary variable taking the value of one if the family head lived on a farm, *south* is a binary variable for living in a certain region of the country, *expected family earnings* was generated from a separate OLS regression to predict earnings from a set of regressors, and *family composition* refers to the number of family members under the age of 18 divided by the total number in the family.

The mean values for the variables were as shown in the table.

Variable	**(1) White mean**	**(2) Nonwhite mean**
age	46.1	42.9
age squared	2,263.5	1,965.6
education	12.6	10.4
farm status	0.03	0.02
south	0.3	0.5
expected family earnings	2,336.4	1,507.3
family composition	0.2	0.3

(a) Interpret the results. Do the coefficients have the expected signs? Why do you think age was entered both in levels and in squares?

Answer: Since these are logit estimates, the value of the coefficients cannot be interpreted easily. However, statements can be made about the direction of the relationship between the dependent variable and the regressors. There is a decrease in the probability of females living on their own with an increase in years of education. Living on a farm also lowers the probability. These results hold both for whites and nonwhites. In addition, for whites the probability of living on her own increases beyond a certain point with age. This is the result of age entering as a level and the square of age. This relationship with regard to age is not statistically significant for nonwhites. In the south, white females are more likely to live on their own, but nonwhites are not. An increase in expected family earnings and family composition increase the probability of females living on their own.

(b) Calculate the difference in the predicted probability between whites and nonwhites at the sample mean values of the explanatory variables. Why do you think the study did not combine the observations and allowed for a nonwhite binary variable to enter?

Answer:

$$\begin{aligned} z_{white} &= 1.459 - .275(46.1) + .00463(2263.5) - .171(12.6) \\ &\quad - .687(.03) + .376(.3) + .0018(2336.4) + 4.123(.2) \\ &= 2.23 \end{aligned}$$

$$\begin{aligned} z_{nonwhite} &= -2.874 + .084(42.9) + .00021(1965.6) - .127(10.4) \\ &\quad - .498(.02) - .520(.5) + .0011(1507.3) + 2.751(.3) \\ &= 2.03 \end{aligned}$$

For whites, the probability is $\frac{\exp(2.23)}{1+\exp(2.23)} = 0.90$, while for nonwhites, it is $\frac{\exp(2.03)}{1+\exp(2.03)} = 0.88$. In the above approach, all coefficients are allowed to vary, whereas in a combined sample, the coefficients on the variables other than the binary race variable would have to be identical.

(c) What would be the effect on the probability of a nonwhite woman living on her own, if *education* and *family composition* were changed from their current mean to the mean of whites, while all other variables were left unchanged at the nonwhite mean values?

Answer:

$$\begin{aligned} z_{nonwhite} &= -2.874 + .084(42.9) + .00021(1965.6) - .127(12.6) \\ &\quad - .498(.02) - .520(.5) + .0011(1507.3) + 2.751(.2) \\ &= 1.48 \end{aligned}$$

The probability would decrease to $\frac{\exp(1.48)}{1+\exp(1.48)} = 0.81$.