Ec 122 - HW5

Ben Juarez

10/27/2021

HGL 6.1

We have $N = 40$ to estimate model $y = \beta_1 + \beta_2 x + \beta_3 z + e$ giving us $SSE = 979.830$ and $s_y = 13.45222$.

Part a

Let us find R^2 as follows:

$$SST = s_y^2 \cdot (N-1) = (13.45222)^2 * (40-1) = 7057.527$$

$$\implies R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{979.830}{7057.527} = 0.861$$

Part b

Let us find the value of the F-statistic for testing $H_0 : \beta_2 = \beta_3 = 0$ as follows:

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} = \frac{(7057.27 - 979.83)/(3-1)}{979.83/(40-3)} = 114.75$$

To determine whether we reject or accept H_0 (at 5% level of significance), we compare F to $F_c = F_{(2,37)} = 3.25$. Since $F > F_c$, we reject H_0, implying that at least $\beta 2$ or β_3 is not equal to zero.

HGL 6.2

Let us consider the model $y = \beta_1 + \beta_2 x + \beta_3 z + e$ such that after augmentation with squares and cubes of predictions $\hat{y}^2$ and $\hat{y}^3$, we obtain $SSE = 696.5357$. Let us use RESET to test for misspecification.

So, for $y = \beta_1 + \beta_2 x + \beta_3 z + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$, let us test $H_0 : \gamma_1 = \gamma_2 = 0$:

$$F = \frac{(979.83 - 696.5357)/2}{696.5357/(40-5)} = 7.12$$

With $F_c = F_{(2,35)} = 3.27$, we have $F > F_c$, so we reject H_0, implying that the model is misspecified (inadequate and can be improved).

HGL 6.4

Let us consider the equation $ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6 (EDUC \times EXPER) + \beta_7 HRSWK + e$.

Part a

Using $t_c = 2$, let us determine which coefficient estimates are not significantly different from zero. For each b_i, let us calculate $t = \frac{b_i}{se(b_i)}$.

```
A = c()
B = c()
C = c()
D = c()
E = c()

Eqn_A = data.frame("EST" = c(1.055,0.0498,0.00319,0.0373,-0.000485,-0.000510,0.01145), "SE" = c(0.266,0
for (i in 1:length(Eqn_A[[1]])) {
  A[i] = Eqn_A[[1]][i] / Eqn_A[[2]][i]
}

Eqn_B = data.frame("EST" = c(1.252,.0289,.00352,.0303,-.000456,0,.01156), "SE" = c(.19,.0344,.00166,.004
for (i in 1:length(Eqn_B[[1]])) {
  B[i] = Eqn_B[[1]][i] / Eqn_B[[2]][i]
}

Eqn_C = data.frame("EST" = c(1.573,0.0366,.00293,0,0,0,.01345), "SE" = c(.188,.0350,.00170,0,0,0,.00136
for (i in 1:length(Eqn_C[[1]])) {
  C[i] = Eqn_C[[1]][i] / Eqn_C[[2]][i]
}

Eqn_D = data.frame("EST" = c(1.917,0,0,.0279,-.000470,0,.01524), "SE" = c(.08,0,0,.0054,.000096,0,.0015
for (i in 1:length(Eqn_D[[1]])) {
  D[i] = Eqn_D[[1]][i] / Eqn_D[[2]][i]
}

Eqn_E = data.frame("EST" = c(.904,.1006,0,.0295,-.0044,0,.01188), "SE" = c(.096,.0063,0,.0048,.000086,0
for (i in 1:length(Eqn_E[[1]])) {
  E[i] = Eqn_E[[1]][i] / Eqn_E[[2]][i]
}

tvalues = data.frame(cbind(A,B,C,D,E))
rownames(tvalues) = c("B_1", "B_2", "B_3", "B_4", "B_5", "B_6", "B_7")
tvalues
```

```
##             A          B        C         D          E
## B_1  3.966165  6.5894737 8.367021 23.962500   9.416667
## B_2  1.254408  0.8401163 1.045714       NaN  15.968254
## B_3  1.887574  2.1204819 1.723529       NaN        NaN
## B_4  4.604938  6.3125000      NaN  5.166667   6.145833
## B_5 -5.388889 -5.3023256      NaN -4.895833 -51.162791
## B_6 -1.058091        NaN      NaN       NaN        NaN
## B_7  8.357664  8.4379562 9.889706 10.092715   8.735294
```

Using the calculated t-values from the table, we reject $H_0 : \beta_i = 0$ if $|t| > t_c = 2$, implying that their corresponding estimates are significantly different from zero. So, the estimates that are not significantly different than zero are as follows:

$$Eqn(A) \implies \beta_2, \beta_3, \beta_6 \quad Eqn(B) \implies \beta_2 \quad Eqn(C) \implies \beta_2, \beta_3$$

Part b

Let us use an F-test to test the restriction on the coefficients of Eqn(A) that gives Eqn(B). Let us test $H_0 : \beta_6 = 0$ vs $H_1 : \beta_6 \neq 0$:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(222.6674 - 222.4166)/1}{222.4166/(1000-7)} = 1.12$$

Since $F_c = F_{(1,993)} = 3.85$, we have $F < F_c$, so we fail to reject H_0 which implies that $EDUC \times EXPER$ is not very meaningful.

Let us show that this same result can be obtained using a t-test. Testing $H_0 : \beta_6 = 0$ vs the same H_1 gives us $t = -1.058$ (from table in Part A). With $t_c = 1.96$, we achieve the same result since since $-t_c \leq t \leq t_c$.

Part c

Let us use an F-test to test the restriction on the coefficients of Eqn(A) that gives Eqn(C). Let us test $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ vs H_1 such that at least one of them is not equal to zero.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(222.8317 - 222.4166)/3}{222.4166/(1000-7)} = 16.99$$

Since $F_c = F_{(3,993)} = 2.614$, we have $F > F_c$, so we reject H_0 which implies that at least one of $\beta_4 \neq 0$, $\beta_5 \neq 0$, $\beta_6 \neq 0$ is true. By performing this test, we are essentially asking whether or not EXPER (years of experience) is a meaningful factor for determining wage. The answer to this question seems to be that experience is meaningful in this sense.

Part d

Let us use an F-test to test the restriction on the coefficients of Eqn(B) that gives Eqn(D). Let us test $H_0 : \beta_2 = \beta_3 = 0$ vs H_1 such that at least one of them is not equal to zero.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(280.5061 - 222.6674)/2}{222.6674/(1000-6)} = 129.1$$

Since $F_c = F_{(2,994)} = 3.01$, we have $F > F_c$, so we reject H_0 which implies that at least one of $\beta_2 \neq 0$, $\beta_3 \neq 0$ is true. By performing this test, we are essentially asking whether or not EDUC (years of education) is a meaningful factor for determining wage. The answer to this question seems to be that education is meaningful in this sense.

Part e

Let us use an F-test to test the restriction on the coefficients of Eqn(A) that gives Eqn(E). Let us test $H_0 : \beta_3 = \beta_6 = 0$ vs H_1 such that at least one of them is not equal to zero.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(223.6716 - 222.4166)/2}{222.4166/(1000-7)} = 2.80$$

Since $F_c = F_{(2,993)} = 3.01$, we have $F < F_c$, so we fail to reject H_0 which implies that $\beta_2 = 0$, $\beta_3 = 0$ is a valid assumption. By performing this test, we are essentially asking whether or not EDUC (years of education) needs to be incorporated beyond its linear term. The answer to this question seems to be that EDUC as linear term is sufficient for determining wage.

Part f

Considering Parts A-E, it seems that Eqn(E) should be the most preferred model. Each of its coefficients are significantly different than zero and it utilizes EDUC efficiently while also including EXPER.

Part g

Let us compute the missing AIC value for Eqn(D):

$$AIC = ln(\frac{SSE}{N}) + \frac{2K}{N} = ln(280.5061/1000) + (8/1000) = -1.26$$

Let us compute the missing SC value for Eqn(A):

$$SC = ln(\frac{SSE}{N}) + \frac{K \cdot ln(N)}{N} = ln(222.4166/1000) + (7 \cdot ln(1000)/1000) = -1.46$$

Considering the AIC values, we conclude Eqn(B) is favored by AIC. Considering SC values, we conclude Eqn(E) is favored by SC.

HGL 6.6

We know that RESET suggests augmenting an existing model with the squares of the predictions $\hat{y}^2$, or with their squares and cubes $(\hat{y}^2, \hat{y}^3)$. Let us consider the following model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

With (b_1, b_2, b_3) as the least squares estimates, we also have $\hat{y} = b_1 + b_2 x_x + b_3 x_3$. Let us consider what would happen if we augmented this model with the predictions themselves. This augmentation would give us

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma \hat{y} + e$$

The issue arises from the exact collinearity between $\hat{y}$ with x_2 and x_3 due to the linear relationships. Thus, the least squares estimator is not defined.

HGL 6.14

```
setwd("~/Desktop/R")
hwage = read.table("dat/hwage.dat")
colnames(hwage) = c("hw", "he", "ha", "cit")
```

Part a

Let us estimate the model $HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HA + e$

```
HW = hwage$hw
HE = hwage$he
HA = hwage$ha

hwage_lm = lm(HW ~ HE + HA)
summary(hwage_lm)
```

```
##
## Call:
## lm(formula = HW ~ HE + HA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.083  -8.636  -1.762   5.883 125.939
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.12358    4.15832  -1.954  0.05112 .
## HE           2.19329    0.18005  12.182  < 2e-16 ***
## HA           0.19966    0.06749   2.958  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.63 on 750 degrees of freedom
## Multiple R-squared:  0.1655, Adjusted R-squared:  0.1633
## F-statistic: 74.37 on 2 and 750 DF,  p-value: < 2.2e-16
```

We can see from the model estimation that an increase in 1 year of HE (husband's education) implies a 2.19 dollar increase in wages. We can also see that this same increase corresponds with a 0.20 dollar increase in wages.

Part b

Let us determine if RESET suggests that the model is adequate. Let us first consider the one variable augmented model:

$$HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HA + \gamma_1 \cdot \hat{HW}^2 + e$$

```
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(hwage_lm, power=2)
```

```
##
##  RESET test
##
## data:  hwage_lm
## RESET = 9.5279, df1 = 1, df2 = 749, p-value = 0.002098
```

So, we have $F = 9.5$ and p-value $= 0.0021 < 0.05$, meaning that the null hypothesis is rejected, implying that the model in part (a) is inadequate.

Let us now consider the two variable augmented model:

$$HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HA + \gamma_1 \cdot \hat{HW}^2 + \gamma_2 \cdot \hat{HW}^2 + e$$

```
resettest(hwage_lm, power=2:3)
```

```
##
##  RESET test
##
## data:  hwage_lm
## RESET = 4.7884, df1 = 2, df2 = 748, p-value = 0.008583
```

So, we have $F = 4.8$ and p-value $= 0.0085 < 0.05$, meaning that the null hypothesis is rejected, implying that the model in part (a) is inadequate.

Part c

Let us estimate the model $HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HE^2 + \beta_4 \cdot HA + \beta_5 \cdot HA^2 + e$.

```
HA2 = HA * HA
HE2 = HE * HE

hwage_qm = lm(HW ~ HE + HE2 + HA + HA2)
summary(hwage_qm)
```

```
##
## Call:
## lm(formula = HW ~ HE + HE2 + HA + HA2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.098  -8.364  -1.940   6.343 125.385
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.567540  17.543637  -2.597 0.009578 **
## HE           -1.457971   1.122786  -1.299 0.194506
## HE2           0.151143   0.045828   3.298 0.001020 **
## HA            2.889541   0.732887   3.943 8.81e-05 ***
## HA2          -0.030121   0.008134  -3.703 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 748 degrees of freedom
## Multiple R-squared:  0.1918, Adjusted R-squared:  0.1874
## F-statistic: 44.37 on 4 and 748 DF,  p-value: < 2.2e-16
```

We can examine the partial derivatives to see the effect that education and age have on wages with this new model:

$$\frac{\partial \hat{HW}}{\partial HE} = -1.457971 + 0.302286 \cdot HE$$

This tells us that there is a negative effect on wage as education increases up until $HE \approx 4.8$ years. After this point, the effect becomes positive such that with each additional year of education, wage increases by approx. 0.3.

$$\frac{\partial \hat{HW}}{\partial HA} = 2.889541 - 0.060242 \cdot HA$$

This tells us that there is a positive effect on wage as age increases up until $HA \approx 48$ years. After this point, the effect becomes negative such that with each additional year of age, wage decreases by approx. 0.06.

Part d

Let us determine if RESET suggests that the model is adequate. Let us first consider the one variable augmented model:

$$HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HE^2 + \beta_4 \cdot HA + \beta_5 \cdot HA^2 + \gamma_1 \cdot \hat{HW}^2 + e$$

```
library("lmtest")
resettest(hwage_qm, power=2)
```

```
##
##  RESET test
##
## data:  hwage_qm
## RESET = 0.32637, df1 = 1, df2 = 747, p-value = 0.568
```

So, we have $F = 0.33$ and p-value $= 0.568 > 0.05$, meaning that the null hypothesis is not rejected, implying that there is no detection of inadequacy with the part (c) model.

Let us now consider the two variable augmented model:

$$HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HE^2 + \beta_4 \cdot HA + \beta_5 \cdot HA^2 + \gamma_1 \cdot \hat{HW}^2 + \gamma_2 \cdot \hat{HW}^2 + e$$

```
resettest(hwage_qm, power=2:3)
```

```
##
##  RESET test
##
## data:  hwage_qm
## RESET = 0.88222, df1 = 2, df2 = 746, p-value = 0.4143
```

So, we have $F = 0.88222$ and p-value $= 0.4143 > 0.05$, meaning that the null hypothesis not rejected, implying that there is no detection of inadequacy with the part (c) model.

Part e

Let us reestimate the model in part (c) with CIT included as follows:

$$HW = \beta_1 + \beta_2 \cdot HE + \beta_3 \cdot HE^2 + \beta_4 \cdot HA + \beta_5 \cdot HA^2 + +\beta_6 \cdot CIT + e$$

```
CIT = hwage$cit

hwage_qm_2 = lm(HW ~ HE + HE2 + HA + HA2 + CIT)
summary(hwage_qm_2)
```

```
##
## Call:
## lm(formula = HW ~ HE + HE2 + HA + HA2 + CIT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.261  -7.562  -1.385   5.442 122.699
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -37.054033  17.016012  -2.178 0.029748 *
## HE           -2.207574   1.091357  -2.023 0.043452 *
## HE2           0.168760   0.044410   3.800 0.000156 ***
## HA            2.621256   0.710107   3.691 0.000239 ***
## HA2          -0.027768   0.007877  -3.525 0.000449 ***
## CIT           7.937853   1.101249   7.208 1.39e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.95 on 747 degrees of freedom
## Multiple R-squared:  0.2443, Adjusted R-squared:  0.2393
## F-statistic:  48.3 on 5 and 747 DF,  p-value: < 2.2e-16
```

We can say that the average wage rate is $7.94 greater relative to outside those cities.

Part f

Logically, I would think that factoring in whether or not the setting is a large city would be meaningful in determining wage, and this logic seems to be confirmed by summary() which gives CIT and significance code of ***. Further testing could confirm this thought.

Part g

Let us estimate the derivatives for part (c):

$$\frac{\partial \hat{HW}}{\partial HE} = -1.457971 + 0.302286 \cdot HE \qquad \frac{\partial \hat{HW}}{\partial HA} = 2.889541 - 0.060242 \cdot HA$$

Evaluating these derivatives respectively gives us

$$HE = 6 \implies 0.36 \quad HE = 15 \implies 3.08 \quad HA = 35 \implies 0.78 \quad HA = 50 \implies -0.12$$

Let us estimate the derivatives for part (e):

$$\frac{\partial \hat{HW}}{\partial HE} = -2.207574 + 0.3376 \cdot HE \qquad \frac{\partial \hat{HW}}{\partial HA} = 2.621256 - 0.0556 \cdot HA$$

Evaluating these derivatives respectively gives us

$$HE = 6 \implies -0.18 \quad HE = 15 \implies 2.86 \quad HA = 35 \implies 0.68 \quad HA = 50 \implies -0.16$$

It does not appear that the omission of CIT leads to omitted-variable bias considering the similar model estimates of part(c)/(e) and the results of the derivative evaluations. Although the derivative evaluations were not very similar for HE=6, they were quite similar for HE=16,HA=35,HA=50. It is also important to consider that HE=6 may not have been an appropriate value to evaluate since it corresponds to 6 years of education, and very few of the observations have an education value this small.