Econ 103 Chapter 6 Discussion Questions

Ryan Martin

February 13, 2018

Note, all questions are 6.3, 6.4 and 6.14. 6.3 is a hand problem

6.3

Consider the model

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

and suppose that application of least squares to 20 observations on these variables yields the following results

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} .96587 \\ .69914 \\ 1.7769 \end{bmatrix}$$

and

$$\widehat{cov(b)} = \begin{bmatrix} .21812 & .019195 & -.050301 \\ .019195 & .048526 & -.031223 \\ -.050301 & -.031223 & .037120 \end{bmatrix}$$

with

$$\hat{\sigma}^2 = 2.5193, \qquad R^2 = .9466$$

a

Find the total variation, unexplained variation and explained variation for this model

Solution Recall that the unexplainted variation is the SSE (sum of square errors). Also recall the definition of our variance estimate, $\hat{\sigma}^2 = \frac{1}{N-K}\sum_i (y_i - \hat{y}_i)^2 = \frac{1}{N-K}SSE$ Plugging in and multiplying each side by N-K yields

```
SSE = (20 - 3)*2.5193
SSE
```

```
## [1] 42.8281
```

To get the other variations, we must recall

$$R^2 = 1 - \frac{SSE}{SST}$$

and that, as *as long as there is an intercept in the model* (which there is), $SSR + SSE = SST$. Algebra yields

$$SST = \frac{SSE}{1 - R^2}$$

which we can now plug in to get

```
SST = 42.8281/(1 - .9466)
SSR = SST - SSE
SST
```

```
## [1] 802.0243
```

```
SSR
```

```
## [1] 759.1962
```

b

Find 95% interval estimates for β_2 and β_3.

Solution Our intervals are $b_j \pm t_{.975,20-3} se(b_j)$ for each j = 1,2. This gives

```
#b2:
tc = qt(.975,20-3)
.69914 + tc *sqrt(.04852)
```

```
## [1] 1.163874
```

```
.69914 - tc *sqrt(.04852)
```

```
## [1] 0.2344055
```

```
#b3
1.7769 + tc *sqrt(.03712)
```

```
## [1] 2.183389
```

```
1.7769 - tc *sqrt(.03712)
```

```
## [1] 1.370411
```

c

Use a t-test to test the hypothesis $H_0 : \beta_2 \geq 1$ against the alternative $H_1 : \beta_2 < 1$

Solution

```
b2 = .69914
test_stat = (b2 - 1)/sqrt(.048526)
test_stat
```

```
## [1] -1.365769
```

```
tc #same as before
```

```
## [1] 2.109816
```

```
abs(test_stat) <tc
```

```
## [1] TRUE
```

```
#true so fail to reject
```

d

Use your answers in part (a) to test the joint hypothesis $H_0 : \beta_2 = 0, \beta_3 = 0$.

Solution a joint test is an F-test. We have a test of two variables so our number of restristictions is J = 2. Recall that the F-statistic is given by

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

where K is the number of coefficients in hte *unrestricted* model and N is the number of observations. Also, SSE_R is the sum of square errors in the restricted model and SSE_U is the sum of squared errors in the full (or unrestricted) model. Our SSE_U is the sum of squared errors we calculated in part a. K = 3 and N = 20. Now, all that is needed is SSE_R.

To get, SSE_R, notice that if $\beta_2 = 0$ and $\beta_3 = 0$, then our model is just an intercept. Recall that if our estimate is just an intercept, then the OLS estimate of the estimate is just the average y. That is, in the special case where $y = \beta_1$, $b_{R,1} = \bar{y}$ where the R stands for restricted. But this means that

$$\begin{aligned} SSE_R &= \sum_i (y_i - b_{R,1})^2 \\ &= \sum_i (y_i - \bar{y})^2 && (\bar{y} \text{ solves OLS when only constant in model}) \\ &= SST \end{aligned}$$

So, *in the special case of the constant estimate in the restricted model* $SSE_R = SST$. And we have SST from A. Plugging in, we get

```
J = 2
N = 20
K = 3
F.stat = ((SST- SSE)/J )/ (SSE/(N - K))
F.stat
```

```
## [1] 150.676
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 3.591531
```

```
F.stat>F.c #True so reject the null.
```

```
## [1] TRUE
```

```
#At least one of the coeffients truly different from 0
```

e

Test the hypothesis $H_0 : 2\beta_2 = \beta_3$

Solution Note this is equivalent to the hypothesis $2\beta_2 - \beta_3 = 0$. We can use a t-test. Our test-statistic is just

$$t = \frac{2b_2 - b_3 - 0}{se(2b_2 - b_3)}.$$

Recall that

$$se(2b_2 - b_3) = \sqrt{4\widehat{var}(b_2) + (-1)^2\widehat{var}(b_3) - 2 \cdot 2 \cdot 1\widehat{cov}(b_2, b_3)}$$

Plugging in from the covariance matrix, we get

```
b2 = .69914
b3 = 1.7769
my_se = sqrt(4*.04853 + .037120 - 4*(-.031223))
test_stat = (2*b2 - b3)/my_se
test_stat
```

```
## [1] -0.6344509
```

```
t_c = qt(.975, 20-3)
t_c
```

```
## [1] 2.109816
```

```
abs(test_stat) < t_c #TRUE so fail to reject H0
```

```
## [1] TRUE
```

6.4

Consider the wage equation

$$\log(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6 (EDUC \times EXPER) + \beta_7 HRSWK + e$$

where the explanatory variables are years of education, years of experience and hours worked per week. Estimation results for this equation, and for modified versions of it obtained by dropping some of the variables, are displayed in Table 6.4. These results are from the 1000 observations in the file `cps4c_small.dat`

a

Using an approximate 5% critical value of $t_c = 2$, what coefficient estimates are not significantly different form zero? *Solution*

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "cps4c_small.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)

reg_out <- lm(data = dat, I(log(wage)) ~ educ + I(educ^2) + exper + I(exper^2) + educ:exper + hrswk)
summary(reg_out)
```

```
##
## Call:
## lm(formula = I(log(wage)) ~ educ + I(educ^2) + exper + I(exper^2) +
##     educ:exper + hrswk, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30371 -0.29260 -0.00782  0.31469  1.82924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.055e+00  2.659e-01   3.969 7.74e-05 ***
## educ         4.983e-02  3.969e-02   1.255   0.2096
## I(educ^2)    3.193e-03  1.693e-03   1.886   0.0595 .
## exper        3.727e-02  8.144e-03   4.577 5.32e-06 ***
## I(exper^2)  -4.849e-04  9.013e-05  -5.380 9.29e-08 ***
## hrswk        1.145e-02  1.374e-03   8.336 2.53e-16 ***
```

```
## educ:exper  -5.104e-04  4.824e-04  -1.058   0.2903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4733 on 993 degrees of freedom
## Multiple R-squared:  0.3173, Adjusted R-squared:  0.3132
## F-statistic: 76.93 on 6 and 993 DF,  p-value: < 2.2e-16
```

Looking at the t-values, we see educ, and the interaction term $educ \times exper$ are not significant at the .95 level

b

What restriction on the coefficients of Eqn (A) gives Eqn(B)? Use an F-test to test this restriction. Show how the same results can be obtained using a t-test.

Solution Looking at the table, we see just the interaction term is missing from B. One term can be done with either an F-test or a t-test.

```
N = 1000
J = 1
K = 7
SSE_U = 222.4166
SSE_R = 222.6675
F.stat = ((SSE_R- SSE_U)/J )/ (SSE_U/(N - K))
F.stat
```

```
## [1] 1.120167
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 3.85084
```

```
F.stat>F.c #False so fail to reject the null.
```

```
## [1] FALSE
```

```
  #The coeefficient is likely 0 (or doesn't improve our model)
1 - pf(F.stat,df1 = J, df2 = N - K) #p-value of F-test
```

```
## [1] 0.2901399
```

```
#t-test
test_stat = -.000510/.000482
2*pt(test_stat, df = N-K) #exact same pvalue!
```

```
## [1] 0.2902711
```

c

What restrictions on the coefficients of Eqn (A) gives Eqn(C)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

Solution Restrict $\beta_4 = 0 = \beta_5, \beta_6$. Now the test is easy, just copy, paste and edit. We conclude that, given the data, at least one of these coefficients is likely nonzero. That is, at least one of Experience or Experience squared or Experience interacted with education, has predictive power for wages.

```
N = 1000
J = 3
K = 7
SSE_U = 222.4166
SSE_R = 233.8317
F.stat = ((SSE_R- SSE_U)/J )/ (SSE_U/(N - K))
F.stat
```

```
## [1] 16.98793
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 2.613866
```

```
F.stat>F.c #Reject the null.
```

```
## [1] TRUE
```

```
  #At least one coefficient is likely nonzero
1 - pf(F.stat,df1 = J, df2 = N - K)
```

```
## [1] 9.147039e-11
```

d

What restrictions on the coefficients of Eqn (B) give Eqn (D)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

Solution Restrict $\beta_2 = 0 = \beta_3$. We conclude that, given the data, at least one of these coefficients is likely nonzero. That is, at least Education or Education squared has an influence on wages.

```
N = 1000
J = 2
K = 6 #reduced by 1 because B is our full model
SSE_U = 222.6674
SSE_R = 280.5061
F.stat = ((SSE_R- SSE_U)/J )/ (SSE_U/(N - K))
F.stat
```

```
## [1] 129.0976
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 3.004779
```

```
F.stat>F.c #Reject the null.
```

```
## [1] TRUE
```

```
  #At least one coefficient is likely nonzero
1 - pf(F.stat,df1 = J, df2 = N - K)
```

```
## [1] 0
```

e

What restrictions on the coefficients of Eqn (A) give Eqn (E)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

Solution Restrict $\beta_2 = 0 = \beta_6$. Fail to reject the null. We conclude that, given the data, β_2 and β_6's exclusion from the model does not significantly change the model's explained variance.

```
N = 1000
J = 2
K = 7 #reduced by 1 because B is our full model
SSE_U = 222.4166
SSE_R = 223.6716
F.stat = ((SSE_R- SSE_U)/J )/ (SSE_U/(N - K))
F.stat
```

```
## [1] 2.801533
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 3.004788
```

```
F.stat>F.c #False, Fail to Reject the null.
```

```
## [1] FALSE
```

```
  #beta_2 and beta_6 likely 0 in comparison. That is,
  #beta_2 and beta_6's exclusion from the model does not significantly
  #reduce the variation explained by the model
1 - pf(F.stat,df1 = J, df2 = N - K)
```

```
## [1] 0.06119689
```

f

Based on your answers to parts (a) to (e), which model would you prefer? Why?

Solution Probably B or E. They acheiveve relatively good prediction without unneeded regressors.

g

Compute the missing AIC value for Eqn (D) and the missing SC value for Eqn (A). Which model is favored by the AIC? Which model is favored by the SC?

Solution Model B is favored by the AIC. Model E Is favored by the SC. To calculate AIC, recall

$$AIC = \log\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

while calculating SC utilizes the formula

$$SC = \log\left(\frac{SSE}{N}\right) + \frac{K\log(N)}{N}$$

Note that SC is almost always referred to (outside of this text) as BIC (Bayesian Information Criterion).

Plugging into these formulas leaves us with

Table 6.4 **Wage Equation Estimates for Exercises 6.4 and 6.5**

Variable	Coefficient Estimates and (Standard Errors)				
	Eqn (A)	Eqn (B)	Eqn (C)	Eqn (D)	Eqn (E)
C	1.055 (0.266)	1.252 (0.190)	1.573 (0.188)	1.917 (0.080)	0.904 (0.096)
EDUC	0.0498 (0.0397)	0.0289 (0.0344)	0.0366 (0.0350)		0.1006 (0.0063)
$EDUC^2$	0.00319 (0.00169)	0.00352 (0.00166)	0.00293 (0.00170)		
EXPER	0.0373 (0.0081)	0.0303 (0.0048)		0.0279 (0.0054)	0.0295 (0.0048)
$EXPER^2$	−0.000485 (0.000090)	−0.000456 (0.000086)		−0.000470 (0.000096)	−0.000440 (0.000086)
EXPER × *EDUC*	−0.000510 (0.000482)				
HRSWK	0.01145 (0.00137)	0.01156 (0.00137)	0.01345 (0.00136)	0.01524 (0.00151)	0.01188 (0.00136)
SSE	222.4166	222.6674	233.8317	280.5061	223.6716
AIC	−1.489	−1.490	−1.445		−1.488
SC		−1.461	−1.426	−1.244	−1.463

Figure 1:

```
modelE_SC = log(222.4166/1000) + 7*log(1000)/1000
modelA_AIC = log(280.5061/1000) + 2*4/1000
modelA_AIC
```

```
## [1] -1.26316
```

```
modelE_SC
```

```
## [1] -1.454849
```

6.14

Following on from the example in section 6.3, the file `hwage.dat` contains another subset of the data used by labor economist Tom Mroz. The variables with which we are concerned are

- HW - Husband's wage in 2006 dollars
- HE - Husband's education attainment in years
- HA - Husband's age
- CIT - a variable equal to one if living in a large city, otherwise zero.

a

Estimate the model

$$HW = \beta_1 + \beta_2 HE + \beta_3 HA + e$$

What effects do changes in the level of education and age have on wages?

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "hwage.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
```

```r
reg_out <- lm(data = dat, hw ~ he + ha)
summary(reg_out)
```

```
##
## Call:
## lm(formula = hw ~ he + ha, data = dat)
##
## 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
```

The current estimates have increases in HE and HA positively increasing wages. All coefficients are significant at the .05 level.

b

Does RESET suggest that the model in part (a) is adequate?

Solution To test RESET, we had $\hat{y}$ to the regression and see if it is significant.

All in all, RESET tells us, (through the significance on the residuals squared), that the original model is inadequate. Both methods of testing RESET reject the null of no misspecification.

c

Add the variables HE^2 and HA^2 to the original equation and re-estimate it. Describe the effect that education and age have on wages in this newly estimated model.

```r
reg_outv2 <- lm(data = dat, hw ~ he + ha + I(he^2) + I(ha^2))
summary(reg_outv2)
```

```
##
## Call:
## lm(formula = hw ~ he + ha + I(he^2) + I(ha^2), data = dat)
##
## 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
## ha            2.889541   0.732887   3.943 8.81e-05 ***
## I(he^2)       0.151143   0.045828   3.298 0.001020 **
## I(ha^2)      -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
```

Note the more complicated story these regressions paint. the linear terms and the quadratic terms are opposite in sign, implying a decreasing then increasing relationship for HE and an increasing then decreasing relationship for HA (because the squared term will dominate as the terms grow)

d

Does RESET suggest that the model in part (c) is adequate?

Solution Repeating the RESET test on our larger model.

```
dat2 = cbind(dat, mypredsq = reg_out$fitted.values^2,
             mypredcu = reg_out$fitted.values^3)
dat2 <- as.data.frame(dat2)
reg_out2v2 <- lm(data = dat2, hw ~ he + ha + mypredsq +
                   I(he^2) + I(ha^2))
summary(reg_out2v2)
```

```
##
## Call:
## lm(formula = hw ~ he + ha + mypredsq + I(he^2) + I(ha^2), data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.886  -8.387  -1.954   6.345 125.374
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -47.579236  24.541826  -1.939 0.052914 .
## he           -1.411313   1.191858  -1.184 0.236739
## ha            2.933483   0.823508   3.562 0.000391 ***
## mypredsq     -0.003134   0.026719  -0.117 0.906653
## I(he^2)       0.164921   0.126095   1.308 0.191307
## I(ha^2)      -0.030212   0.008176  -3.695 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 747 degrees of freedom
## Multiple R-squared:  0.1918, Adjusted R-squared:  0.1864
## F-statistic: 35.45 on 5 and 747 DF,  p-value: < 2.2e-16
```

```
reg_out3v2 <- lm(data = dat2, hw ~ he + ha + mypredsq + mypredcu +
                   I(he^2) + I(ha^2))
summary(reg_out3v2)
```

```
##
```

```
## Call:
## lm(formula = hw ~ he + ha + mypredsq + mypredcu + I(he^2) + I(ha^2),
##     data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.248  -8.413  -1.919   6.276 125.400
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.890e+01  4.706e+01  -1.252 0.211133
## he           1.538e-02  5.199e+00   0.003 0.997640
## ha           3.097e+00  1.008e+00   3.071 0.002208 **
## mypredsq    -3.118e-02  1.030e-01  -0.303 0.762183
## mypredcu     3.376e-04  1.197e-03   0.282 0.778053
## I(he^2)      1.729e-01  1.293e-01   1.337 0.181589
## I(ha^2)     -3.042e-02  8.213e-03  -3.704 0.000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.43 on 746 degrees of freedom
## Multiple R-squared:  0.1919, Adjusted R-squared:  0.1854
## F-statistic: 29.52 on 6 and 746 DF,  p-value: < 2.2e-16
```

```
#F-test
N = nrow(dat)
J = 2
K = 6
SSE_U = sum((reg_out3v2$residuals)^2)
SSE_R = sum((reg_outv2$residuals)^2)
F.stat = ((SSE_R- SSE_U)/J )/ (SSE_U/(N - K))
F.stat
```

```
## [1] 0.04668423
```

```
F.c = qf(.95, df1 = J, df2 = N - K)
F.c
```

```
## [1] 3.007778
```

```
F.stat>F.c #False, fail to reject the null. None of the terms
```

```
## [1] FALSE
```

```
    #appear necessary
1 - pf(F.stat,df1 = J, df2 = N - K)
```

```
## [1] 0.9543915
```

Our data passes both RESET tests of misspecification.

e

Reestimate the model in part(c) with the variable CIT included. What can you say about the level of wages in large cities relative to outside those cities?

Solution Note that cit is a binary random variable, taking 0 if the person is not inside a city and 1 if the person is in a big city. From the coefficient, larger cities seem to have much higher expected wages than small

cities.

```
reg_outv3 <- lm(data = dat, hw ~ he + ha + I(he^2) + I(ha^2) +
                  cit)
summary(reg_outv3)
```

```
##
## Call:
## lm(formula = hw ~ he + ha + I(he^2) + I(ha^2) + cit, data = dat)
##
## 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 *
## ha            2.621256   0.710107   3.691 0.000239 ***
## I(he^2)       0.168760   0.044410   3.800 0.000156 ***
## I(ha^2)      -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
```

```
cor( cbind(dat$he, dat$ha, dat$cit)) #not too strong
```

```
##            [,1]        [,2]       [,3]
## [1,]  1.0000000 -0.19532255 0.23334852
## [2,] -0.1953226  1.00000000 0.06762408
## [3,]  0.2333485  0.06762408 1.00000000
```

f

Do you think CIT should be included in the equation?

Solution It's t-statistic says it is a statistically significant variable. It makes sense that there is a bump in wages from cities, too. Thus, leaving it in seems fine and perhaps most appropriate.

g

For both the model estimated in part (c) and the model estimated in part (e) evaluate the following four derivatives:

- $\frac{\partial HW}{\partial HE}$ for $HE = 6$ and $HE = 15$
- $\frac{\partial HW}{\partial HA}$ for $HA = 35$ and $HA = 50$

Does the omission of CIT lead to omitted-variable bias? Can you suggest why?

Solution Our model in part c has

$$HW = \beta_1 + \beta_2 HE + \beta_3 HA + \beta_4 HE^2 + \beta_5 HA^2$$

Thus, for our model from (c):

$$\frac{\partial HW}{\partial HE} = \beta_2 + 2\beta_4 HE^2$$

and

$$\frac{\partial HW}{\partial HA} = \beta_3 + 2\beta_5 HA^2$$

Note that the model from part (e) has the same derivatives! It's just the coefficients are changing in value with the inclusion of CIT. So, plugging in the numbers we get the following marginal effects

```
#Model from (c)

##marginal with respect to he
coef(reg_outv2)[2] + 2* coef(reg_outv2)[4]*6 #at 6
```

```
##      he
## 0.35574
```

```
coef(reg_outv2)[2] + 2* coef(reg_outv2)[4]*15 #at 15
```

```
##       he
## 3.076306
```

```
##marginal with respect to ha
coef(reg_outv2)[3] + 2* coef(reg_outv2)[5]*6 #at 35
```

```
##       ha
## 2.528086
```

```
coef(reg_outv2)[3] + 2* coef(reg_outv2)[5]*15 #at 50
```

```
##       ha
## 1.985904
```

```
#Model from e,

##marginal with respect to he
coef(reg_outv3)[2] + 2* coef(reg_outv3)[4]*6 #at 6
```

```
##         he
## -0.1824573
```

```
coef(reg_outv3)[2] + 2* coef(reg_outv3)[4]*15 #at 15
```

```
##       he
## 2.855218
```

```
##marginal with respect to ha
coef(reg_outv3)[3] + 2* coef(reg_outv3)[5]*6 #at 35
```

```
##       ha
## 2.288041
```

```
coef(reg_outv3)[3] + 2* coef(reg_outv3)[5]*15 #at 50
```

```
##       ha
## 1.788219
```

Note that the estimated marginal effects are lower in model e, uniformly (that is, for all points, the models in c have larger marginal effects than the corresponding estimates from part e). This makes sense, since people

in the city tend to marry later and have higher education. So, the positive correlation between HE and CIT as well as between HA and CIT was leading to positive bias of the estimates

Indeed, recall from equation 6.23 of the text that the bias should be $bias(b_x) = \beta_{CIT} \hat{cov}(CIT, x)/\hat{var}(x)$ where β_{CIT} is the true coefficient on CIT, b_x is the estimated coefficient of x and x is either HA or HE. Thus, since the coefficient on CIT is positive, positive correlation between CIT and x corresponds to positive bias. The positive correlation is confirmed by the correlation matrix I calculated in part (e)

6.22

In Chapter 5.7 we used the data in file pizza4.dat to estimate the model

$$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4 (AGE \times INCOME) + e$$

a

Test the hypothesis that age does not affect pizza expenditure - that is, test the joint hypothesis $H_0 \beta_2 = 0, \beta_4 = 0$. What do you conclude?

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "pizza4.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
attach(dat)
reg_out_unr <- lm(data = dat, pizza ~ age*income)
summary(reg_out_unr)
```

```
##
## Call:
## lm(formula = pizza ~ age * income, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -200.86  -83.82   20.70   85.04  254.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 161.46543  120.66341   1.338   0.1892
## age          -2.97742    3.35210  -0.888   0.3803
## income        6.97991    2.82277   2.473   0.0183 *
## age:income   -0.12324    0.06672  -1.847   0.0730 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127 on 36 degrees of freedom
## Multiple R-squared:  0.3873, Adjusted R-squared:  0.3363
## F-statistic: 7.586 on 3 and 36 DF,  p-value: 0.0004681
```

```
reg_out_res <- lm(data = dat, pizza ~ income)
summary(reg_out_res)
```

```
##
## Call:
## lm(formula = pizza ~ income, data = dat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -260.17 -103.81  -49.86  122.59  337.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 128.9803    34.5913   3.729 0.000626 ***
## income        1.1213     0.4595   2.440 0.019461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.8 on 38 degrees of freedom
## Multiple R-squared:  0.1355, Adjusted R-squared:  0.1127
## F-statistic: 5.954 on 1 and 38 DF,  p-value: 0.01946
```

```
#R's built in f-test method
anova(reg_out_unr, reg_out_res)
```

```
## Analysis of Variance Table
##
## Model 1: pizza ~ age * income
## Model 2: pizza ~ income
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     36 580609
## 2     38 819286 -2   -238677 7.3995 0.002033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#or, do it by hand
F.test = ( (sum(reg_out_res$residuals^2) -
            sum(reg_out_unr$residuals^2) )/2)/
        (sum(reg_out_unr$residuals^2)/(nrow(dat) - 3))
F.test
```

```
## [1] 7.604999
```

```
1 - pf(F.test, 2, nrow(dat) - 3)
```

```
## [1] 0.001711153
```

```
#Note, the same!
```

b

Construct point estimates and .95 CI for the marginal propensity to spend on pizza for individuals of ages 20, 30, 40, 50 and 55. Comment on these estimates.

solution marginal propensity to spend is just $\beta_3 + \beta_4 \times AGE$.

```
my_cov_mat <- vcov(reg_out_unr)
my_cov_mat
```

```
##              (Intercept)          age       income   age:income
## (Intercept) 14559.658424 -386.7975523 -269.1518796  6.552844006
## age          -386.797552   11.2365799    6.4637905 -0.166083924
## income       -269.151880    6.4637905    7.9680170 -0.185924335
```

```
## age:income      6.552844   -0.1660839   -0.1859243  0.004451389
```

```
#as a function
age = 20
my_CI_func <- function(my_age) {
se.mp = as.numeric(sqrt( my_cov_mat[3,3] + my_age^2*my_cov_mat[4,4] +
               2*my_age*my_cov_mat[3,4]))
my_est = as.numeric(reg_out_unr$coefficients[3] +
      reg_out_unr$coefficients[4]*my_age)
t.c = qt(.975, nrow(dat) - 3)
cup = my_est + t.c*se.mp
clower = my_est - t.c*se.mp
out = list(my_est, se.mp, cup, clower)
names(out) <- c("estimate", "se", "Upper", "Lower")
return(out)
}

my_CI_func(20)
```

```
## $estimate
## [1] 4.515118
##
## $se
## [1] 1.520394
##
## $Upper
## [1] 7.59573
##
## $Lower
## [1] 1.434506
```

```
my_CI_func(30)
```

```
## $estimate
## [1] 3.282725
##
## $se
## [1] 0.9048794
##
## $Upper
## [1] 5.116184
##
## $Lower
## [1] 1.449265
```

```
my_CI_func(40)
```

```
## $estimate
## [1] 2.050331
##
## $se
## [1] 0.4650721
##
## $Upper
## [1] 2.992657
##
```

```
## $Lower
## [1] 1.108005
```

```
my_CI_func(50)
```

```
## $estimate
## [1] 0.8179375
##
## $se
## [1] 0.7099684
##
## $Upper
## [1] 2.25647
##
## $Lower
## [1] -0.6205952
```

```
my_CI_func(55)
```

```
## $estimate
## [1] 0.2017408
##
## $se
## [1] 0.9908536
##
## $Upper
## [1] 2.209401
##
## $Lower
## [1] -1.805919
```

c

Modify the equation to permit a "life-cycle" effect in which the marginal effect of income on pizza expenditure increases with age, up to a point, then falls. Do so by adding the term $(AGE^2 \times INC)$ to the model. What sign do you anticipate on this term? Estimate the model and test the significance of the coefficient for this variable. Did the estimate have the expected sign?

Solution We expect increasing into early adulthood, then decreasing when health concerns and more money make you not want pizza as much. Thus, a negative sign on $AGE^2 \times INC$ is expected, since this will dominate for large ages. This predicted sign is not what we see in the regression. It seems, over the range of ages, $age \times income$ and $age^2 \times income$ move together (we'll see the evidence of this in part f), which may be why we don't get the sign here

```
reg_out_unr_2 <- lm(data = dat, pizza ~ age*income + I(age^2*income))
summary(reg_out_unr_2)
```

```
##
## Call:
## lm(formula = pizza ~ age * income + I(age^2 * income), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -212.080  -79.979    7.395   81.429  260.074
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       109.720767 135.572473   0.809    0.424
## age                -2.038273   3.541904  -0.575    0.569
## income             14.096163   8.839862   1.595    0.120
## I(age^2 * income)   0.004205   0.004948   0.850    0.401
## age:income         -0.470371   0.413908  -1.136    0.264
##
## Residual standard error: 127.5 on 35 degrees of freedom
## Multiple R-squared:  0.3997, Adjusted R-squared:  0.3311
## F-statistic: 5.826 on 4 and 35 DF,  p-value: 0.001057
```

d

Using the model in (c) e marginal propensity to spend on pizza for individuals of ages 20, 30, 40, 50 and 55. Comment on these estimates. In light of these values and of the range of the age in the sample data, what can you say about the quadratic function of age that describes the marginal propensity to spend on pizza?

Solution Now marginal propensity to spend on pizza is

$$\beta_3 + \beta_4 \times AGE + \beta_5 \times AGE^2$$

Note that the standard error would be a little painful to calculate. That's why they only have us do the estimate

The quadratic function of age is always increasing over the interval. We see that because the marginal effect is always positive, but decreasingly positive.

```
summary(dat$age) #age is between 18 and 55. Median 32
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   24.75   32.00   33.48   40.00   55.00
```

```
my_cov_mat <- vcov(reg_out_unr_2)
my_cov_mat
```

```
##                     (Intercept)           age        income
## (Intercept)       18379.8955619 -457.08644098 -781.06066096
## age                -457.0864410   12.54508213   15.76705680
## income             -781.0606610   15.76705680   78.14315850
## I(age^2 * income)    -0.3012341    0.00546731    0.04142764
## age:income           31.4726406   -0.61873740   -3.60749340
##                   I(age^2 * income)   age:income
## (Intercept)           -3.012341e-01 31.472640629
## age                    5.467310e-03 -0.618737404
## income                 4.142764e-02 -3.607493402
## I(age^2 * income)      2.447827e-05 -0.002020842
## age:income            -2.020842e-03  0.171319722
```

```
#as a function
age = 20
my_CI_func_2 <- function(my_age) {
my_est = as.numeric(reg_out_unr_2$coefficients[3] +
      reg_out_unr_2$coefficients[5]*my_age +
      reg_out_unr_2$coefficients[4]*my_age^2)
return(my_est)
}
```

```
my_CI_func_2(20)
```

```
## [1] 6.370658
```

```
my_CI_func_2(30)
```

```
## [1] 3.769337
```

```
my_CI_func_2(40)
```

```
## [1] 2.008969
```

```
my_CI_func_2(50)
```

```
## [1] 1.089555
```

```
my_CI_func_2(55)
```

```
## [1] 0.9452059
```

e

For the model in part (c), are each of the coefficient estimates for AGE ($AGE \times INC$) and ($AGE^2 \times INC$) significantly different from 0 at a 5% level? Carry out a joint test for the signifiance of these variables. Comment on your results.

Solution Just use the R built in for this one, to do the F-test. Note, they are not jointly significantly different from 0 at the .05 level. That is, the test fails to reject the null that both of the coefficients are 0.

```
model_res <- lm(data = dat, pizza ~ age + income)
anova(reg_out_unr_2, model_res )
```

```
## Analysis of Variance Table
##
## Model 1: pizza ~ age * income + I(age^2 * income)
## Model 2: pizza ~ age + income
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     35 568869
## 2     37 635637 -2    -66768 2.054 0.1434
```

f

Check the model used in part (c) for collinearity. Add the term ($AGE^3 \times INC$) to the model in (c) and check the resulting model for collinearity.

Solution Remember, a large vif would be anything over 5. These vifs are very large! We indeed have collinearity,

```
reg_alt <- lm(data = dat, pizza ~
                age*income + I(age^2*income) +
                I(age^3*income))
#install.packages("olsrr")
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_vif_tol(reg_out_unr_2)
```

```
## # A tibble: 4 x 3
##           Variables    Tolerance         VIF
##               <chr>        <dbl>       <dbl>
## 1               age 0.3160018717    3.164538
## 2            income 0.0020371525  490.881263
## 3 I(age^2 * income) 0.0014135530  707.437216
## 4        age:income 0.0004365421 2290.729711
```

```
ols_vif_tol(reg_alt)
```

```
## # A tibble: 5 x 3
##           Variables    Tolerance          VIF
##               <chr>        <dbl>        <dbl>
## 1               age 1.740195e-01 5.746483e+00
## 2            income 1.708152e-04 5.854281e+03
## 3 I(age^2 * income) 9.798099e-06 1.020606e+05
## 4 I(age^3 * income) 6.236831e-05 1.603378e+04
## 5        age:income 1.371601e-05 7.290751e+04
```

As a side note, this package "olsrr" is very nice! Check out more here https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html or at their site www.rsquaredacademy.com