University of California, Santa Barbara **Olivier Deschenes**
Department of Economics **Winter 2018**

ESM 296
Individual Assignment 1: Answer Key

Some of these exercises are taken from Stock and Watson textbook.

Question 1:

Sir Francis Galton, a cousin of James Darwin, examined the relationship between the height of children and their parents towards the end of the 19th century. It is from this study that the name "regression" originated. You decide to update his findings by collecting data from 110 college students, and estimate the following relationship:

$$\hat{Studenth} = \underset{(7.2)}{19.6} + \underset{(0.10)}{0.73} \times Midparh, \quad R^2 = 0.45$$

where *Studenth* is the height of students in inches, and *Midparh* is the average of the parental heights. (Following Galton's methodology, both variables were adjusted so that the average female height was equal to the average male height.)

(a) Interpret the estimated coefficients.

<u>Answer</u>: *For every one inch increase in the average height of their parents, the student's height increases by 0.73 of an inch. There is no reasonable interpretation for the intercept (no one has parents with height 0 inches).*

(b) What is the meaning of the regression R^2?

<u>Answer</u>: *The model (average parent height) explains 45 percent of the variation in the height of students. Explains here refers to correlation and not causation.*

(c) What is the prediction for the height of a child whose parents have an average height of 70.06 inches?

<u>Answer</u>: *$19.6 + 0.73 \times 70.06 = 70.74$.*

(d) Given the positive intercept and the fact that the slope lies between zero and one, what can you say about the height of students who have quite tall parents? Who have quite short parents?

<u>Answer</u>: *Tall parents will have, on average, tall students, but they will not be as tall as their parents. Short parents will have short students, although on average, they will be somewhat taller than their parents.*

(e) Test for the statistical significance of the slope coefficient.

Answer: *We would like to test whether or not slope coefficient is statistically different than 0. Form the null hypothesis as H0: Slope = 0 (the alternative is that Slope does not equal 0). To test this, construct the test statistic:*

$$\hat{t} = \frac{0.73 - 0}{0.10} = 7.3$$

Since $|\hat{t}| > 1.98$ *(for a two-sided hypothesis with 108 degrees of freedom (110 - 2)), we reject the null hypothesis and conclude that at least 95% of the time a reasonable range of estimates for the Slope coefficient excludes 0. Using a critical value of 1.96 (from the N(0,1)) would also be fine.*

(f) If children, on average, were expected to be of the same height as their parents, then this would imply two hypotheses, one for the slope and one for the intercept.

(i) What should the null hypothesis be for the intercept? Calculate the relevant *t*-statistic and carry out the hypothesis test at the 1% level.
(ii) What should the null hypothesis be for the slope? Calculate the relevant *t*-statistic and carry out the hypothesis test at the 5% level.

Answer: *(i)* $H_0 : \beta_0 = 0$*, t=2.72, for* $H_1 : \beta_0 \neq 0$*, the critical value for a two-sided alternative at the 1% level is 2.62 (df = 108). Hence we reject the null hypothesis in (i).*

(ii) $H_0 : \beta_1 = 1$*, t=-2.70, for* $H_1 : \beta_1 \neq 1$*, the critical value for a two-sided alternative is 1.98 (df = 108). Hence we reject the null hypothesis in (ii), but note that this just barely passes the 95% level.*

Question 2

The data for this question contain information on the reported value and characteristics of houses in the Boston area we used in class. The STATA data file “HPRICE2.dta” is available on the class website. The same file also available in spreadsheet format “HPRICE2.csv”.

Consider the following linear regression model for the price and characteristics of houses:

$$\text{Price}_\text{i} = \beta_0 + \beta_1 NOx_i + \beta_2 Rooms + \beta_3 STratio_i + u_i$$

Where *price* is the value of the house, NOx is a measure of NOx concentration in the Census track (in parts per 100 million), *Rooms* is the number of rooms in the house, and STratio is the student-teacher ratio in the nearest school.

(a) What is the effect of adding an additional room on the house price, holding NOx concentrations and student-teacher ratio constant?

```
. regress price nox rooms stratio, robust;

Linear regression                                      Number of obs =     506
                                                       F(  3,   502) =  134.45
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6005
                                                       Root MSE      =  5837.8

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nox |  -1693.326   217.4923    -7.79   0.000    -2120.633   -1266.018
       rooms |   7000.633   657.8602    10.64   0.000     5708.135    8293.131
     stratio |   -1166.96   115.7563   -10.08   0.000    -1394.386   -939.5333
       _cons |   9458.008   5843.485     1.62   0.106    -2022.693    20938.71
------------------------------------------------------------------------------
```

<u>**ANSWER:**</u> β_2 *is the effect of an addition room on the house price, holding square footage constant. The estimate implies that holding NOx concentrations and student-teacher ratio constant, an additional room will add $7,000.63 to the selling price.*

(b) What is the estimated effect on house values of reducing NOx concentrations by 2.5 parts per 100 million?

<u>**ANSWER:**</u> *The regression predicts that reducing NOx concentrations by 2.5 parts per 100 million would increase house values by $4,233.31b. (see code below).*

```
. lincom -2.5*nox;

 ( 1)  - 2.5*nox = 0

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   4233.314   543.7308     7.79   0.000     3165.046    5301.583
```

```
------------------------------------------------------------------------------
```

(c) What percent of the variation in house values is explained by NOx concentrations, number of rooms, and student-teacher ratio? What percent of the variation in house values is explained by NOx concentrations alone?

> **<u>ANSWER:</u>** *In the regression in part (a) above, R^2 = 0.60, so 60% of the variation in price is explained by Nox, Rooms, and STratio. To explain the amount of variation in house values explained by NOx concentrations alone, run the auxiliary regression below. In this regression, R^2 = 0.1815, so 18.2% of the variation is explained by NOx concentrations.*

```
. reg price nox, robust

Linear regression                                      Number of obs =     506
                                                       F(  1,   504) =  141.30
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1815
                                                       Root MSE      =  8339.6

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nox |  -3386.853   284.9188   -11.89   0.000    -3946.628   -2827.078
       _cons |   41307.81   1588.999    26.00   0.000     38185.93    44429.68
------------------------------------------------------------------------------
```

(d) What is the predicted value of a house with NOx concentrations of 6, with 7 rooms, and a student-teacher ratio of 20? The actual price for that house was $20,000. Did the buyer overpay for this house?

> **<u>ANSWER:</u>** *To predict housing values, just multiply the coefficients by the desired characteristics. This is easy with STATA's lincom command, which also gives a standard errors estimate. The predicted price is $24,963, roughly $5,000 more than the actual price, so the buyer did not overpay for the house. Eyeballing the standard errors, this $5,000 is greater than 2*688, so it this is significant difference both monetarily and statistically.*

```
. lincom  _cons + nox*6 + rooms*7 + stratio*20;

 ( 1)  6*nox + 7*rooms + 20*stratio + _cons = 0

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   24963.29   688.5342    36.26   0.000     23610.53    26316.05
------------------------------------------------------------------------------
```

(e) Test the null hypothesis that β_1 and β_2 jointly equal zero.

ANSWER: *Based on the value of the test statistic and p-value, we reject the null hypothesis. There are a couple ways to set up (input) this test:*

Here's one way:

```
. test nox=rooms=0

 ( 1)  nox - rooms = 0
 ( 2)  nox = 0

       F(  2,   502) =  136.86
            Prob > F =    0.0000
```

And here's another:

```
. test nox rooms

 ( 1)  nox = 0
 ( 2)  rooms = 0

       F(  2,   502) =  136.86
            Prob > F =    0.0000
```

(f) Test the null hypothesis that $\beta_1 = \beta_3$ against the two-sided alternative.

ANSWER: *Based on the value of the test statistic and p-value, we reject the null hypothesis. The p-value doesn't ensure the same level of significance as before, but this is still robust.*

```
. test nox=stratio;

 ( 1)  nox - stratio = 0

       F(  1,   502) =    6.54
            Prob > F =    0.0109
```

R SCRIPT

Load useful libraries

```
library(ggplot2)
library(dplyr)
library(lmtest)
library(sandwich)
library(tidyr)
library(broom)
library(knitr)
library(car)

poss_packages = installed.packages()[,1]

has_robust = any('RobustRegression' %in% poss_packages)

if (has_robust == F){ devtools::install_github('DanOvando/RobustRegression',build_vignettes = T)}
library(RobustRegression)
```

Question 1

I'm only providing R scripts for more complex operations

1.e

You need the critical t values for questions *1.e* and *1.f*. R allows you to look those up using the `qt` function (quantiles of the *t* distribution).

We have 110 data points and 2 parameters (slope and intercept), so we have 108 degrees of freedom. We want a two-tailed test, meaning we don't really care if the alternative hypothesis is above or below the null, just that it's different. We get this from `qt` per

```
crit_t_95 = qt(c(0.025,0.975),df = 108)
```

The `c(0.025,0.975)` tells us that we want the values from the Student *t* distribution associated with the .025th and .975th quantiles. Meaning, that values inside that range should make up 95% of the distribution. For a one-tailed test at the 5% level, you could do. In this case, we don't really care which one we pick, they're the same, so let's report the absolute value of the top end `abs(crit_t_95)[2]` = 1.98

```
crit_t_95_onetailed = qt(0.95,df = 108)
```

Since for the one tailed (and in this case greater than), 95% of the distribution has to fall **below* a value

1.f.i

Same drill but now we want the 1% level (or 99% if you're counting that way).

```
crit_t_99 = qt(c(0.005,0.995),df = 108)
```

Question 2

Read in data. Every time you read in a .csv without `stringsAsFactors = F` a dolphin gets punched in the snout.

```
HPrice<- read.csv('Homework 1/HPRICE2.csv', stringsAsFactors = F ) #Change to
your own directory as needed
```

I'm going to use the `RobustRegression` package, but to see the nuts and bolts for calculating HSC robust SE's, here we go

The `sandwhich` package takes apart the variance covariance matric using the `vcovHC` function

```
  FullModel =   lm(price ~ nox+rooms+stratio,data=HPrice)

  hsc_vcov <- vcovHC(FullModel,type='HC1')

  summary(FullModel)

##
## Call:
## lm(formula = price ~ nox + rooms + stratio, data = HPrice)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14352  -3386   -297   2222  40007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9458.0     4371.4   2.164    0.031 *
## nox          -1693.3      236.3  -7.168 2.75e-12 ***
## rooms         7000.6      409.1  17.110  < 2e-16 ***
## stratio      -1167.0      128.8  -9.063  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5838 on 502 degrees of freedom
## Multiple R-squared:  0.6005, Adjusted R-squared:  0.5981
## F-statistic: 251.5 on 3 and 502 DF,  p-value: < 2.2e-16

    SEs<- data.frame(t(sqrt(diag(hsc_vcov))),stringsAsFactors=F) %>% gather('
variable','HSC Robust SEs')

kable(SEs)
```

variable	HSC Robust SEs
X.Intercept.	5843.4856
nox	217.4923
rooms	657.8602
stratio	115.7563

```
FullReg<- RobustRegression(lm(price ~ nox+rooms+stratio,data=HPrice), dat = H
price)
```

Full Regression - Parentheses are HSC robust standard errors

```
                    Dependent variable:
                ---------------------------
                           price
```

nox -1,693.326*** (217.492)

rooms 7,000.633*** (657.860)

stratio -1,166.960*** (115.756)

Constant 9,458.009 (5,843.486)

Observations 506
R2 0.601
Adjusted R2 0.598
Residual Std. Error 5,837.767 (df = 502)
F Statistic 251.545*** (df = 3; 502)
== Note: *p<0.1;* ***p<0.05;*** p<0.01

2.a

`FullReg$TidyModel` has the coefficients with robust standard errors. We can extract the marginal effect of adding an additional room on the house from the `rooms` coefficient using the `coef` command per

`coef(FullReg$model)['rooms']` = 7000.6331831

2.b

Extract the `nox` coefficient and multiply by -2.5. You can get the HSC robust CI by using the data in TidyModel, pulling out the HSC robust 95% CI, and under the assumption of normality, multiply them by -2.5 as well

```
coef(FullReg$model)['nox'] * -2.5

##      nox
## 4233.314
```

```
hsc_nox_cis = FullReg$TidyModel %>%
  filter(variable == 'nox') %>%
  select(LCI95, UCI95)

hsc_nox_cis * -2.5

##      LCI95    UCI95
## 1 5301.583 3165.046
```

2.c

use the `glance` command to extract summary statistics like R^2

```
FullReg = RobustRegression(lm(price ~ nox+rooms+stratio,data=HPrice), dat = H
Price)

PartialReg = RobustRegression(lm(price ~ nox ,data=HPrice), dat = HPrice)

glance(FullReg$model)['r.squared']

##   r.squared
## 1 0.6005205

glance(PartialReg$model)['r.squared']

##   r.squared
## 1 0.1815076
```

2.d

We'll use the `predict` function for this. We can use the res.var option to pass the HSC robust variance-covariance matric to `predict`, set `interval` to 'confidence' to get 95% CIs back, and se.fit = T to get fit diagnostics

```
new_dat <- data.frame(nox = 6, rooms = 7, stratio = 20)

predicted_house = predict(FullReg$model,new_dat, res.var = FullReg$model$VCOV
, interval = 'confidence', se.fit = T)

predicted_house

## $fit
##        fit   lwr      upr
## 1 24963.29 23975 25951.58
##
## $se.fit
## [1] 503.024
##
## $df
## [1] 502
##
```

```
## $residual.scale
## [1] 5837.767
```

This is one of those where I can't get the standard errors to match up perfectly to Stata, will work on this.

2.e

We're going to use the `linearHypothesis` function from the `car` package here.

```
linearHypothesis(model = FullReg$model,hypothesis.matrix = c('nox = 0',
'rooms = 0'),
                 white.adjust = 'hc0')

## Linear hypothesis test
## 
## Hypothesis:
## nox = 0
## rooms = 0
## 
## Model 1: restricted model
## Model 2: price ~ nox + rooms + stratio
## 
## Note: Coefficient covariance matrix supplied.
## 
##   Res.Df Df      F    Pr(>F)    
## 1    504                        
## 2    502  2 137.95 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.f

```
linearHypothesis(model = FullReg$model,hypothesis.matrix = c('nox = stratio')
,
                 white.adjust = 'hc0')

## Linear hypothesis test
## 
## Hypothesis:
## nox - stratio = 0
## 
## Model 1: restricted model
## Model 2: price ~ nox + rooms + stratio
## 
## Note: Coefficient covariance matrix supplied.
## 
##   Res.Df Df      F  Pr(>F)  
## 1    503                    
## 2    502  1 6.5881 0.01055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```