Lab 6 - Homework, Part C1

Nikita Grabher-Meyer

11/1/2020

PART C (coding)

Part 1

Exercise 1

Setup: set working directory, load packages and data set

```
setwd("/Users/nikitagrabher-meyer/Desktop/PHD/Econometrics/Labs/Lab 6, Homework")

library(data.table)
library(ggplot2)
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

load("hprice1.RData")
dt.hprice <- data.table(data)
rm(data)
```

Analysis

Use the data in hprice1.RData to estimate the model price = BO + B1 sqrft+ B2 bdrms + u where price is the house price measured in thousands of dollars

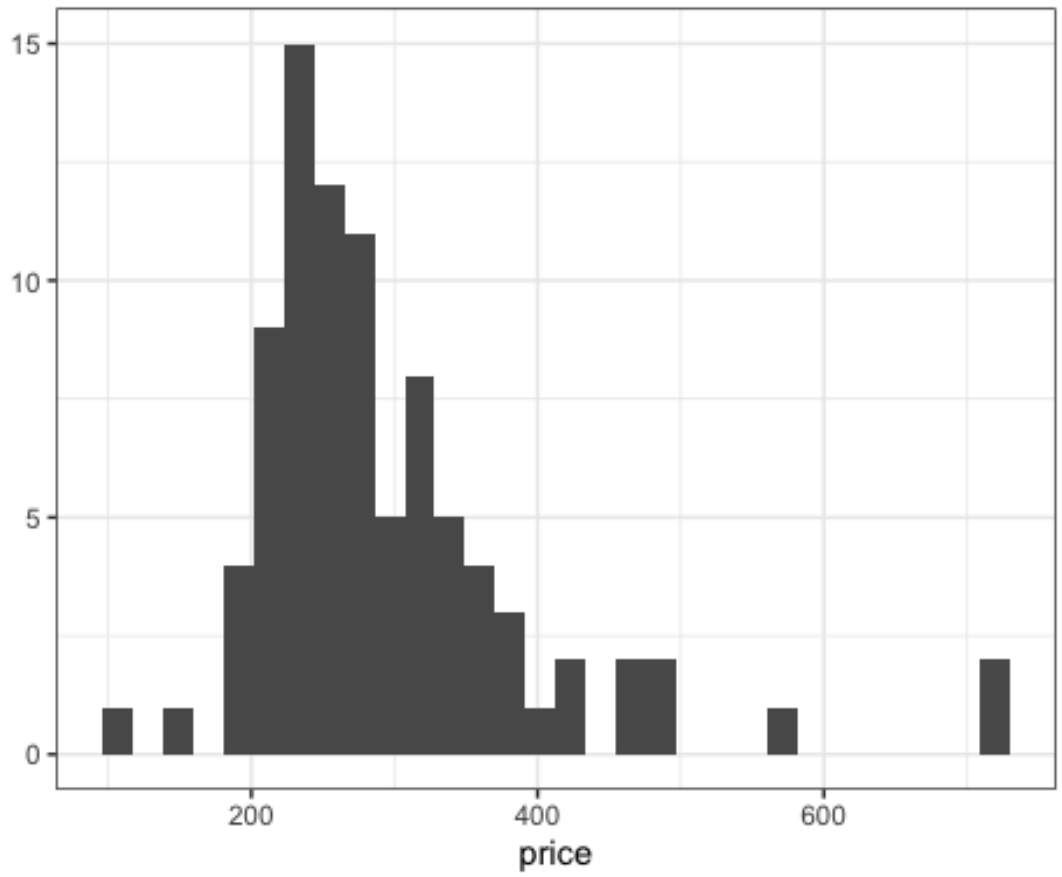
```
stargazer(dt.hprice, type = "text")

##
## ===============================================================
## Statistic N    Mean     St. Dev.    Min   Pctl(25) Pctl(75)   Max
## ---------------------------------------------------------------
## price     88  293.546   102.713     111     230     326.2     725
## assess    88  315.736    95.314   198.700 253.900  352.125  708.600
## bdrms     88   3.568     0.841       2       3        4        7
## lotsize   88 9,019.864 10,174.150  1,000  5,732.8  8,583.2  92,681
```

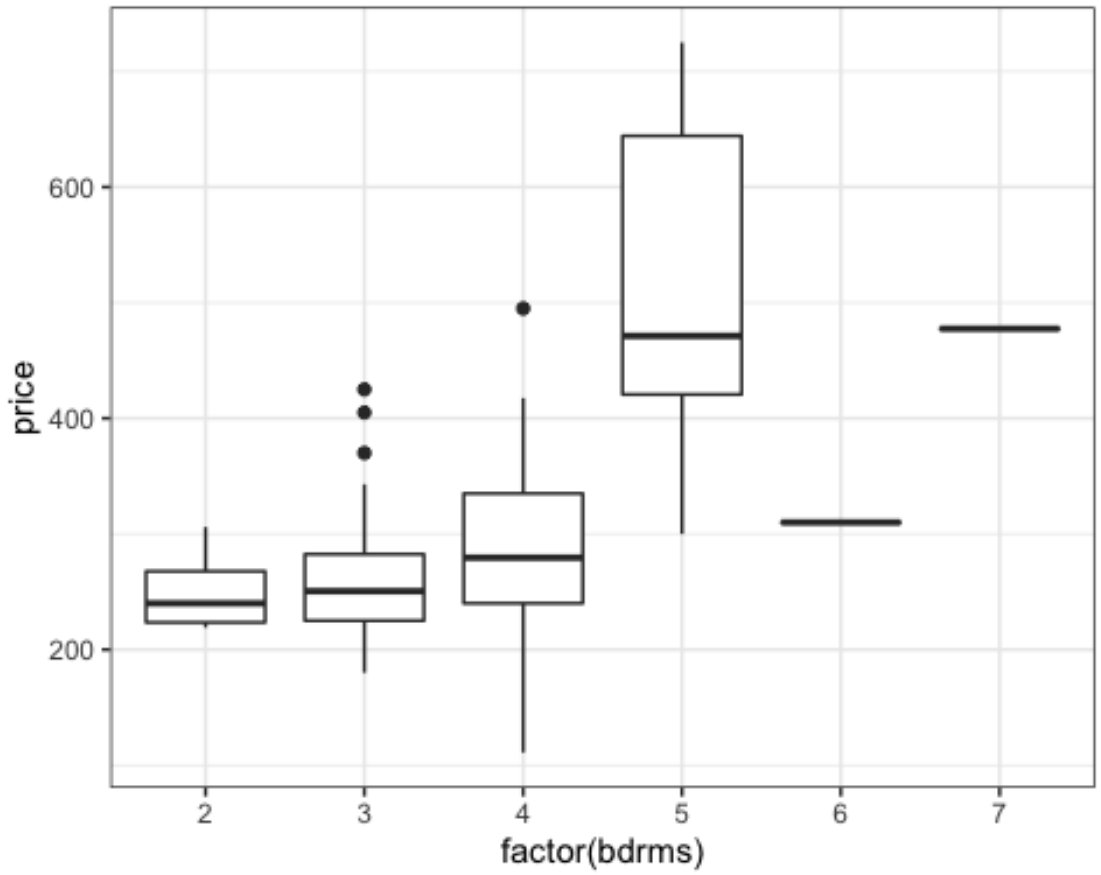
```
## sqrft     88 2,013.693  577.192    1,171  1,660.5   2,227    3,880
## colonial  88   0.693     0.464       0       0        1        1
## lprice    88   5.633     0.304     4.710   5.438    5.788    6.586
## lassess   88   5.718     0.262     5.292   5.537    5.864    6.563
## llotsize  88   8.905     0.544     6.908   8.654    9.058   11.437
## lsqrft    88   7.573     0.259     7.066   7.415    7.708    8.264
## ----------------------------------------------------------------------
```

```
qplot( data = dt.hprice
, x = price
, geom = "histogram") + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot( data = dt.hprice
, x = factor(bdrms)
, y = price
, geom ="boxplot") + theme_bw()
```



i) Write out the results in equation form

```
lm.price1 <- lm( price ~ bdrms + sqrft
, data = dt.hprice)
stargazer(lm.price1 , type = "text")

##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                price
## ------------------------------------------------
## bdrms                          15.198
##                                (9.484)
##
## sqrft                         0.128***
##                                (0.014)
##
## Constant                       -19.315
##                               (31.047)
##
## ------------------------------------------------
## Observations                     88
```

```
## R2                             0.632
## Adjusted R2                    0.623
## Residual Std. Error      63.045 (df = 85)
## F Statistic           72.964*** (df = 2; 85)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

price = -19.315 + 0.128 sqrft + 15.198 bdrms

ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant? 15.198 gives us the increase in price that results from a 1 unit increase in the number of bedrooms. However the coefficient is not statistically significant.

iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii)

```
new.bdrms1 <- data.table( bdrms = 1
, sqrft = 140)
new.bdrms1

##    bdrms sqrft
## 1:     1   140

my.pred <- predict(lm.price1, newdata = new.bdrms1)
my.pred

##        1
## 13.86426
```

Including an additional bedroom of 140 square feet gives us a coefficient of 13.864.

iv) What percentage of the variation in price is explained by square footage and number of bedrooms? R2=63.2% gives us the percentage of the variation in price that is explained by the current model.

v) The first house in the sample has sqrft = 2,438 and bdrms = 4. Find the predicted selling price for this house from the OLS regression line

```
new.bdrms2 <- data.table( bdrms = 4
, sqrft = 2438)
new.bdrms2

##    bdrms sqrft
## 1:     4  2438

my.pred <- predict(lm.price1, newdata = new.bdrms2, interval="prediction",
level = .95)
my.pred

##        fit      lwr      upr
## 1 354.6052 228.1436 481.0669
```

The predicted selling price for this house is 354.6.

vi) The actual selling price of the first house in the sample was $300,000 (so price=300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house? The residual e=y-y' is 354-300=54. The buyer underpaid for the house.

vii) Now add the variable colonial to your model. Interpret its coefficient. Is it significant?

```
lm.price2 <- lm( price ~ bdrms + sqrft + colonial
, data = dt.hprice)
stargazer(lm.price1 , lm.price2, type = "text")

##
## ===============================================================
##                                  Dependent variable:
##                     ---------------------------------------------
##                                         price
##                              (1)                    (2)
## ---------------------------------------------------------------
## bdrms                        15.198                 12.487
##                             (9.484)                (10.024)
##
## sqrft                       0.128***               0.130***
##                             (0.014)                (0.014)
##
## colonial                                           13.078
##                                                   (15.436)
##
## Constant                    -19.315                -21.552
##                             (31.047)               (31.210)
##
## ---------------------------------------------------------------
## Observations                  88                     88
## R2                           0.632                  0.635
## Adjusted R2                  0.623                  0.622
## Residual Std. Error    63.045 (df = 85)       63.150 (df = 84)
## F Statistic         72.964*** (df = 2; 85) 48.720*** (df = 3; 84)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

The coefficient 13.078 of the dummy variable colonial tells us that, on average, colonial style houses report higher prices. However the coefficient is not statistically significant.

Exercise 4

Setup: set working directory, load packages and data set

```
setwd("/Users/nikitagrabher-meyer/Desktop/PHD/Econometrics/Labs/Lab 6,
Homework")

library(data.table)
```

```
library(ggplot2)
library(stargazer)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
load("meap93.RData")
dt.mathpass <- data.table(data)
rm(data)
```

Analysis

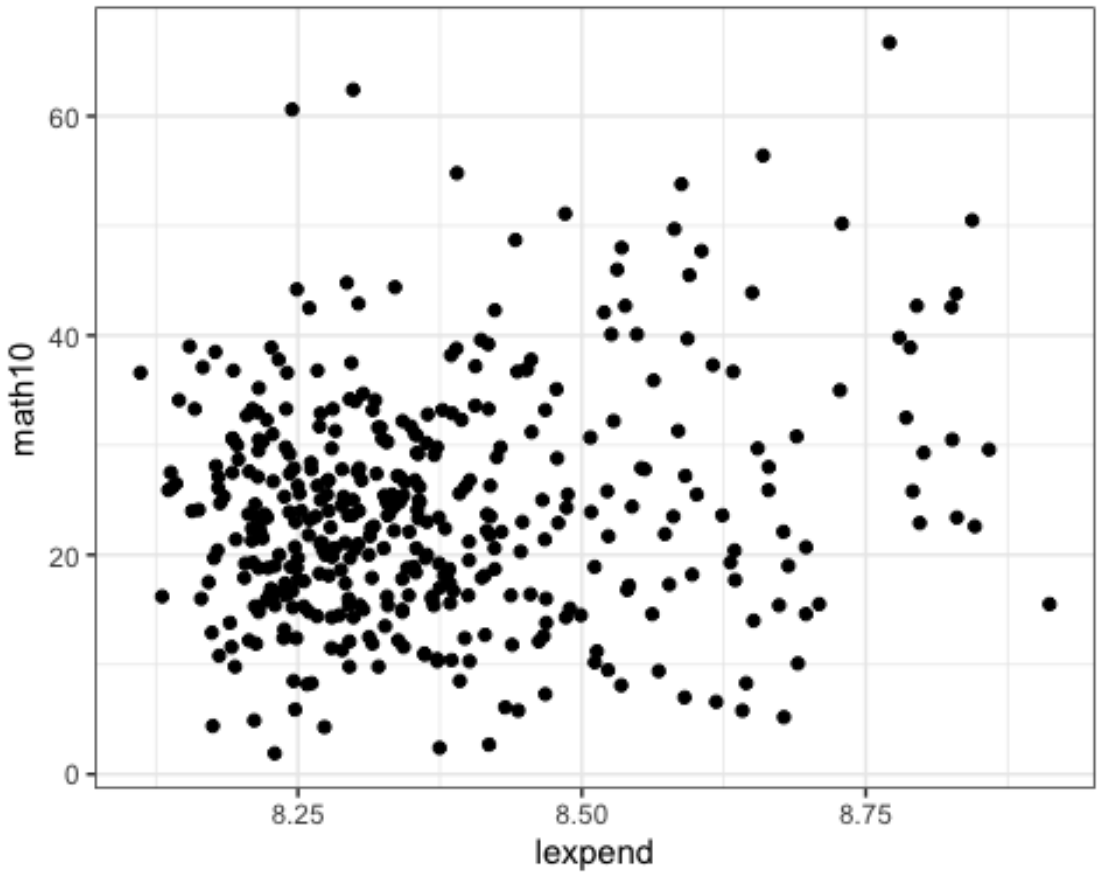
i) Load the dataset MEAP93.RData and obtain the summary statistics

```
stargazer(dt.mathpass, type = "text")
```

```
##
## ===============================================================
## Statistic  N     Mean    St. Dev.   Min   Pctl(25) Pctl(75)   Max
## ---------------------------------------------------------------
## lnchprg   408   25.201    13.610   1.400   14.625   33.825  79.500
## enroll    408 2,663.806  2,696.821  212   1,037.5  3,084.8  16,793
## staff     408  100.642    13.300   65.900  91.450  108.025  166.600
## expend    408 4,376.578   775.790  3,332  3,821.2  4,658.8   7,419
## salary    408 31,774.510 5,038.304 19,764 28,185.5 34,499.8 52,812
## benefits  408 6,463.429  1,456.338   0    5,536.5   7,228   11,618
## droprate  408   5.066      5.485   0.000   1.900    6.500   61.900
## gradrate  408   83.652    13.368   23.500  77.000   93.225  127.100
## math10    408   24.107    10.494   1.900   16.625   30.050  66.700
## sci11     408   49.183    12.525   7.200   41.300   57.150  85.700
## totcomp   408 38,237.940 5,985.086 24,498  34,032   41,637  63,518
## ltotcomp  408   10.540     0.151   10.106  10.435   10.637  11.059
## lexpend   408   8.370      0.162   8.111   8.248    8.447    8.912
## lenroll   408   7.510      0.867   5.357   6.945    8.034    9.729
## lstaff    408   4.603      0.127   4.188   4.516    4.682    5.116
## bensal    408   0.205      0.038   0.000   0.188    0.220    0.450
## lsalary   408   10.354     0.154   9.892   10.247   10.449  10.874
## ---------------------------------------------------------------
```

ii) We want to explore the relationship between the math pass rate (math10) and spending per student (expend). Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate?

```
qplot( data = dt.mathpass
, x = lexpend
, y = math10
, geom = "point") +
theme_bw()
```




```
dt.mathpass[, cor(math10, lexpend)]

## [1] 0.1722303

dt.mathpass[, rcorr(math10, lexpend)]

##      x    y
## x 1.00 0.17
## y 0.17 1.00
##
## n= 408
##
##
## P
```

```
##   x     y
## x       5e-04
## y 5e-04
```

The plot suggests a soft positive correlation between the two variables, also confirmed by the correlation coefficient.

iii) In the population model math10 = B0 + B1log(expend) + μ argue that B1/10 is the percentage point change in math10 given a 10% increase in expend As the independent variable is log transformed, we should divide the coefficient by 100. This tells us that a 1% increase in the independent variable increases (or decreases) the dependent variable by (coefficient/100) units. For an x percent increase, we need to multiply the coefficient by log(1.x). Example: For a 10% increase in the expend, math10 increases by about B1*log(1.10).

iv) Use the data in MEAP93.RAW to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and R-squared

```
lm.math10 <- lm( math10 ~ lexpend
, data = dt.mathpass)
stargazer(lm.math10 , type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                math10
## ------------------------------------------------
## lexpend                       11.164***
##                                (3.169)
##
## Constant                     -69.341***
##                                (26.530)
##
## ------------------------------------------------
## Observations                    408
## R2                             0.030
## Adjusted R2                    0.027
## Residual Std. Error      10.350 (df = 406)
## F Statistic           12.411*** (df = 1; 406)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

math10 = -69.341 + 11.164 lexpend

v) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in math10? A 1% increase in the independent variable increases the dependent variable by (11.164/100) units, that is 0.1116. Therefore, if expend increases by 10%, the estimated percentage point increase in math10 is: 1.064 = 11.164 * log(1.10)= 11.164 * 0.09531018.

vi) One might worry that regression analysis can produce fitted values for math10 that are greater than 100. Why is this not much of a worry in this data set? Math10 is a pass rate expressed in percentages.