Chapter 3 Lab

Tianwei Liu

Sep 25 2019

Preparation

```
require(knitr)
require(haven)
require(AER)

opts_chunk$set(echo = TRUE)
options(digits = 3)

#add your working directory here
opts_knit$set(root.dir ="~/Desktop/GU/Stats/Lab2(Chapter3)")
```

1) Estimate a regression model explaining Trump feeling thermometer as a function of education.

(a) What is the slope coefficient? Briefly explain what this coefficient means.

```
my_reg1 <- lm(dta$therm_trump ~ dta$education, data=dta)
summary(my_reg1)
```

```
##
## Call:
## lm(formula = dta$therm_trump ~ dta$education, data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.27 -17.08 -14.04   7.92  85.96
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.312      2.514   14.05  < 2e-16 ***
## dta$education   -3.038      0.425   -7.15  1.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.6 on 1779 degrees of freedom
##   (578 observations deleted due to missingness)
## Multiple R-squared:  0.028,  Adjusted R-squared:  0.0274
## F-statistic: 51.2 on 1 and 1779 DF,  p-value: 1.23e-12
```

The slope of this coeffieicnt is -3.038. This means that Trump feeling thermometer is negatively correlated with the level of education. In other words, the more education one receives, the less likely the person supports Trump.

(b) Estimate the model with robust standard errors and explain similarities and differences from results in part (a).

```
coeftest(my_reg1, vcov.=vcovHC(my_reg1,type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.312      3.010   11.73  < 2e-16 ***
## dta$education   -3.038      0.484   -6.28  4.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Robust standard errors are bigger than std. errors result from part (a), and this makes t-values smaller.

2) For the fourth observation: What is

(a) the value of education

```
dta$education[4]
```

```
## [1] 7
```

(b) the fitted value

```
predict(my_reg1)[4]
```

```
##  4
## 14
```

(c) the actual therm_trump value

```
dta$therm_trump[4]
```

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

(d) the residual

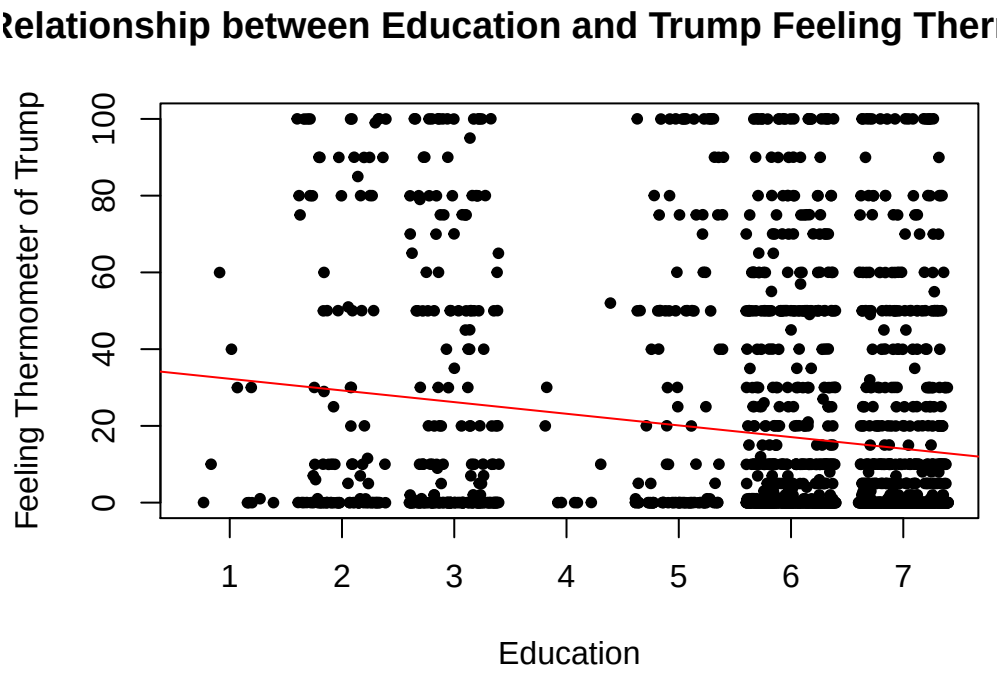
```
residuals(my_reg1)[4]
```

```
##   4
## -14
```

3) Scatterplot the Trump feeling thermometer and education data. Add a fitted line from your regression. Use the jitter(2) subcommand when making the scatterplot. (If you are filling this lab sheet in by hand, you may simply produce a quick sketch what your output looks like.)

```r
plot(jitter(dta$therm_trump,2) ~ jitter(dta$education,2) , pch = 20, xlab="Education", ylab="Feeling Th

abline (a = coef(my_reg1)[1], b=coef(my_reg1)[2], col="red")
```



4) Estimate a regression model explaining Clinton feeling thermometer as a function of education. What is the coefficient? Briefly explain what this model means.

```r
my_reg2 <- lm(dta$therm_clinton ~ dta$education)
summary(my_reg2)
```

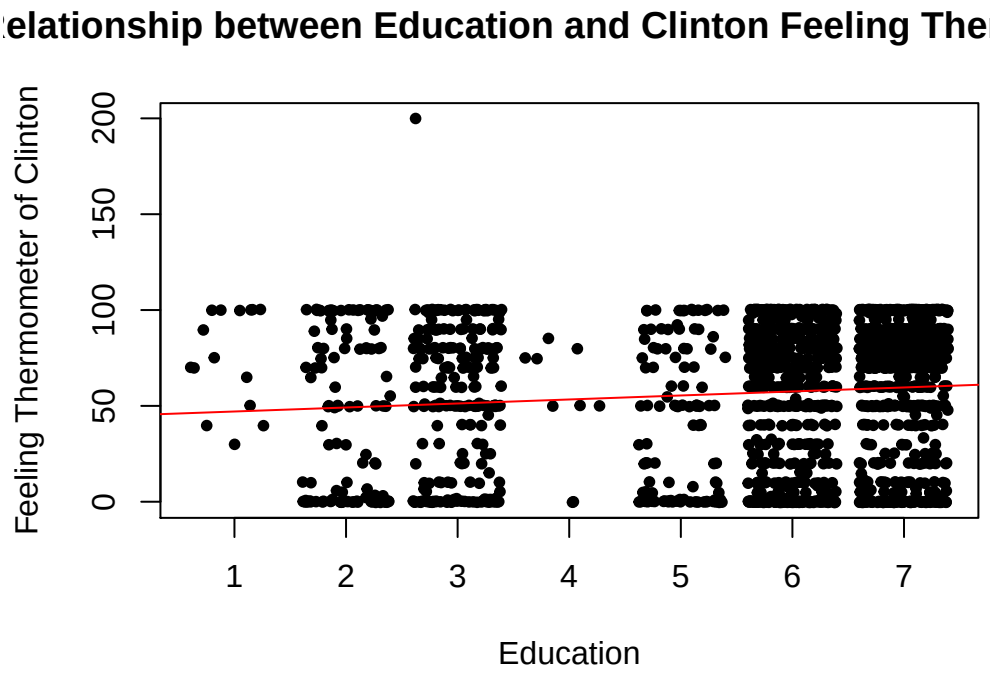
```
##
## Call:
## lm(formula = dta$therm_clinton ~ dta$education)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -59.6  -37.5   10.4   30.4  148.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     44.989      3.048   14.76  < 2e-16 ***
## dta$education    2.090      0.517    4.05  5.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.7 on 1820 degrees of freedom
##   (537 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.00892,    Adjusted R-squared:  0.00837
## F-statistic: 16.4 on 1 and 1820 DF,  p-value: 5.42e-05
```

The slope coefficient is a positive value at 2.090. This means that feeling thermometer of Clinton is positively correlated with education level; the more education one receives, the more the person supports Clinton.

5) Scatterplot the Clinton feeling thermometer and education data. Add a fitted line from your regression. (If you are filling this lab sheet in by hand, you may simply produce a quick sketch what your output looks like.)

```
plot(jitter(dta$education,2), jitter(dta$therm_clinton,2), pch = 20, xlab="Education", ylab="Feeling Th
abline (a = coef(my_reg2)[1], b=coef(my_reg2)[2], col="red")
```



6) Estimate a regression model explaining Clinton feeling thermometer as a function of education for the first 400 observations only. What is $\hat{\beta}_1$? What is the standard error of $\hat{\beta}_1$^? Compare to results from the entire sample.

```
my_reg3 <- lm(dta$therm_clinton[1:400] ~ dta$education[1:400], data=dta)
summary(my_reg3)
```

```
##
## Call:
## lm(formula = dta$therm_clinton[1:400] ~ dta$education[1:400],
##     data = dta)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##  -64.5  -20.3   10.5   25.5   50.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              47.57       9.57    4.97  1.1e-06 ***
## dta$education[1:400]      2.42       1.53    1.58     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.6 on 306 degrees of freedom
##   (92 observations deleted due to missingness)
## Multiple R-squared:  0.00813,    Adjusted R-squared:  0.00489
## F-statistic: 2.51 on 1 and 306 DF,  p-value: 0.114
```

If we run regression on only the first 400 observations, the new beta1 hat is 2.42 with a standard error of 1.53, compared with regression on the entire dataset, where beta1 hat is 2.090 with a standard error 0.517. It is notable that the standard error is three times than the previous result, and it is not statistically significant. This indicates that having a larger sample size can reduce the variance of beta1 hat (recall that var(beta1_hat) = sigmaˆsquared / N * var(X)).

7) Estimate a regression model explaining Trump feeling thermometer as a function of education for Republicans only. Use robust standard errors. What are the slope coefficient and t-statistic?

```
#First, we need to make a dummy variable for republican
dta$rep <- (dta$pol_party == 5 | dta$pol_party == 6 | dta$pol_party == 7)

my_reg4 <- lm (therm_trump ~ education, data = dta[dta$rep==1,])
coeftest(my_reg4, vcov. = vcovHC(my_reg4, type = "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.629      5.666   14.23  < 2e-16 ***
## education     -6.010      0.984   -6.11  2.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The slope is -6.010. The t-statistic is -6.11 and it is significant. This means that Trump feeling thermometer is negatively correlated with education, and the correlation effect is stronger than running the regression on the entire sample.

8) Estimate a regression model explaining Trump feeling thermometer as a function of gender (e.g., a dummy variable for women). What is the slope coefficient? What is the intercept? Explain what they mean.

```
dta$female <- (dta$gender == 2)
my_reg5 <- lm(therm_trump ~ female, data = dta)
coeftest(my_reg5, vcov. = vcovHC(my_reg5, type = "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20.32       1.03    19.7   <2e-16 ***
## femaleTRUE     -4.42       1.38    -3.2   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The slope coefficient has a value of -4.42, meaning that female tends to dislike trump. The intercept coefficient has a value of 20.32, meaning that when female=0 (male), the Trump feeling thermometer is predicted to be at 20.32.

9) Use the summarize command to calculate the mean value of *therm_trump* for men and women. Try to connect these values back to the regression model above.

```
therm_trump_female <- mean (dta$therm_trump[dta$female==1], na.rm=TRUE)
therm_trump_other <- mean (dta$therm_trump[dta$female==0], na.rm=TRUE)
therm_trump_female - therm_trump_other
```

```
## [1] -4.42
```

The difference between the mean values is the slope coefficient. It's not hard to understand that when running a regression on a dummy variable, the slope is the difference between the values of the two variables considered.

(10)

10) [Advanced with new material: looping] Use a loop to estimates models on all feeling thermometer variables; use education and female as the only independent variables. Save coefficients in a matrix. For which feeling thermometer is the magnitude of the education coefficient the largest? For which feeling thermometer is the magnitude of the female coefficient the largest? (Don't worry about statistical significance at this point.)

Note: to create a matrix of the feeling thermometer dependent variables, use *grep("^therm", names(dta))*.

```
therm_vars <- grep("^therm",names(dta)) ## Grep the index of names in dta starting with "therm"
therm_vars_no <- length (therm_vars) ## Number of variables containing "therm"
coeffmat <- matrix (NA, nrow = therm_vars_no, ncol = 3) ## Create a coefficient matrix
row.names (coeffmat) <- colnames (dta[ ,therm_vars]) ## Name the coefficient matrix
colnames(coeffmat) <- c("Intercept", "Education", "Female") ## Name the columns
therm_dta <- data.frame(dta[,therm_vars]) ## create a df
for (i in 1 : therm_vars_no) {
  my_reg <- lm (therm_dta[,i] ~ dta$education + dta$female)
  coeffmat[i,] = my_reg$coefficients
}
print (coeffmat)
```

```
##               Intercept Education Female
## therm_clinton      38.1    2.3260 10.114
```

```
## therm_obama        52.9    2.4277  6.794
## therm_blm          46.8    0.7821 14.008
## therm_repub        40.6   -1.3514 -0.379
## therm_trump        38.1   -3.1248 -4.377
## therm_putin        19.8   -1.5630 -3.612
## therm_ryan         34.1   -0.0419 -3.314
## therm_dem          43.9    1.0338  9.311
```

11) [Advanced with new material: creating a function and using list apply] Use lapply to estimate and save models using feeling thermometers as dependent variables and education and female as independent variables. Display the coefficients.

```
therm_vars_list <- dta[,therm_vars]
counter = 0
## list apply function
OLS.temp = lapply(therm_vars_list, function(x) {
  counter = counter + 1
  print(noquote(c("Dependent variable: ", names(dta [,therm_vars[counter]]))))
  temp = lm(x ~ dta$education  + dta$female)
  print(summary(temp))
temp }   )
```

```
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -64.5  -34.4   10.5   27.8  144.8
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.142      3.200   11.92  < 2e-16 ***
## dta$education     2.326      0.516    4.51  6.9e-06 ***
## dta$femaleTRUE   10.114      1.670    6.05  1.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.3 on 1797 degrees of freedom
##   (559 observations deleted due to missingness)
## Multiple R-squared:  0.0299, Adjusted R-squared:  0.0288
## F-statistic: 27.7 on 2 and 1797 DF,  p-value: 1.41e-12
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -76.7  -19.9   15.1   25.8   44.7
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      52.875      3.239   16.33  < 2e-16 ***
## dta$education     2.428      0.522    4.65  3.6e-06 ***
## dta$femaleTRUE    6.794      1.674    4.06  5.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.2 on 1779 degrees of freedom
##   (577 observations deleted due to missingness)
## Multiple R-squared:  0.0204, Adjusted R-squared:  0.0193
## F-statistic: 18.6 on 2 and 1779 DF,  p-value: 1.05e-08
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -66.24 -27.23   8.76  30.83  52.46
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      46.760      3.301   14.17   <2e-16 ***
## dta$education     0.782      0.531    1.47     0.14
## dta$femaleTRUE   14.008      1.688    8.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35 on 1724 degrees of freedom
##   (632 observations deleted due to missingness)
## Multiple R-squared:  0.0391, Adjusted R-squared:  0.038
## F-statistic: 35.1 on 2 and 1724 DF,  p-value: 1.12e-15
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.22 -26.11  -3.81  18.06  69.27
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      40.570      2.658   15.27   <2e-16 ***
## dta$education    -1.351      0.427   -3.16   0.0016 **
## dta$femaleTRUE   -0.379      1.353   -0.28   0.7791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28 on 1721 degrees of freedom
```

```
##   (635 observations deleted due to missingness)
## Multiple R-squared:  0.0058, Adjusted R-squared:  0.00464
## F-statistic: 5.02 on 2 and 1721 DF,  p-value: 0.00672
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.02 -16.27 -11.89   8.11  88.11
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.144      2.652   14.39  < 2e-16 ***
## dta$education    -3.125      0.427   -7.32  3.8e-13 ***
## dta$femaleTRUE   -4.377      1.362   -3.21   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.5 on 1760 degrees of freedom
##   (596 observations deleted due to missingness)
## Multiple R-squared:  0.0343, Adjusted R-squared:  0.0332
## F-statistic: 31.3 on 2 and 1760 DF,  p-value: 4.56e-14
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.23  -8.85  -5.24   1.15  91.63
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      19.795      1.638   12.09  < 2e-16 ***
## dta$education    -1.563      0.264   -5.91  4.0e-09 ***
## dta$femaleTRUE   -3.612      0.826   -4.37  1.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.9 on 1667 degrees of freedom
##   (689 observations deleted due to missingness)
## Multiple R-squared:  0.0309, Adjusted R-squared:  0.0298
## F-statistic: 26.6 on 2 and 1667 DF,  p-value: 4.22e-12
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -34.03 -28.78  -5.61  19.50  69.54
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.0685     2.8690   11.87   <2e-16 ***
## dta$education   -0.0419     0.4622   -0.09    0.928
## dta$femaleTRUE  -3.3143     1.4476   -2.29    0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.4 on 1647 degrees of freedom
##   (709 observations deleted due to missingness)
## Multiple R-squared:  0.00317,    Adjusted R-squared:  0.00196
## F-statistic: 2.62 on 2 and 1647 DF,  p-value: 0.073
##
## [1] Dependent variable:  therm_clinton
##
## Call:
## lm(formula = x ~ dta$education + dta$female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -60.4  -29.4    8.9   24.9   55.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      43.859      3.036   14.45  < 2e-16 ***
## dta$education     1.034      0.489    2.12    0.035 *
## dta$femaleTRUE    9.311      1.564    5.95  3.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.6 on 1741 degrees of freedom
##   (615 observations deleted due to missingness)
## Multiple R-squared:  0.022,  Adjusted R-squared:  0.0209
## F-statistic: 19.6 on 2 and 1741 DF,  p-value: 3.96e-09
```