

# Chapter 3 Lab

Allen Church

September 25, 2019

Prepare packages and load data

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

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

#Add working directory, assign data to lab variable
lab <- read_dta("Ch2_lab_survey_data.dta")
```

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.

```
#Set y variable to therm_trump column, and set x variable to education column
y=lab$therm_trump
x=lab$education

#Run linear model with therm_trump as dependent variable (y), education as independent variable (x) and
results1 <- lm(y~x, data = lab)
summary(results1)
```

```
##
## Call:
## lm(formula = y ~ x, data = lab)
##
## 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 ***
## x             -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 coefficient above is -3.038, meaning that education is negatively correlated with the Trump thermometer. Specifically, for every increase in unit of education, the Trump thermometer decreases by -3.038.

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

```
#Perform tests of estimated coefficients and specify heteroskedasticity-consistent estimation of the co
coeftest(results1, vcov. = vcovHC(results1,type="HC1"))
```

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

In the model above, estimating using robust standard errors produces the same intercept and Beta hat 1. However, there are higher values for standard errors and t values. The model above produces a higher standard error, meaning that the model fits less precisely with the previous model, since the standard error of the regression corresponds to average distance of observations from fitted values.

2) For the fourth observation: What is

(a) the value of education

```
#Isolate education column in lab and select fourth observation
lab$education[4]
```

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

(b) the fitted value

```
#Use predict function which predicts values based on linear model stored in results1
#Select fourth observation from the model
predict(results1)[4]
```

```
## [1] 14
```

(c) the actual therm\_trump value

```
#Select actual value from the therm_trump column, and select fourth value
lab$therm_trump[4]
```

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

(d) the residual

```
#Calculate the difference between the fitted value and the actual value
#Select fourth observation from results1 linear model
residuals(results1)[4]
```

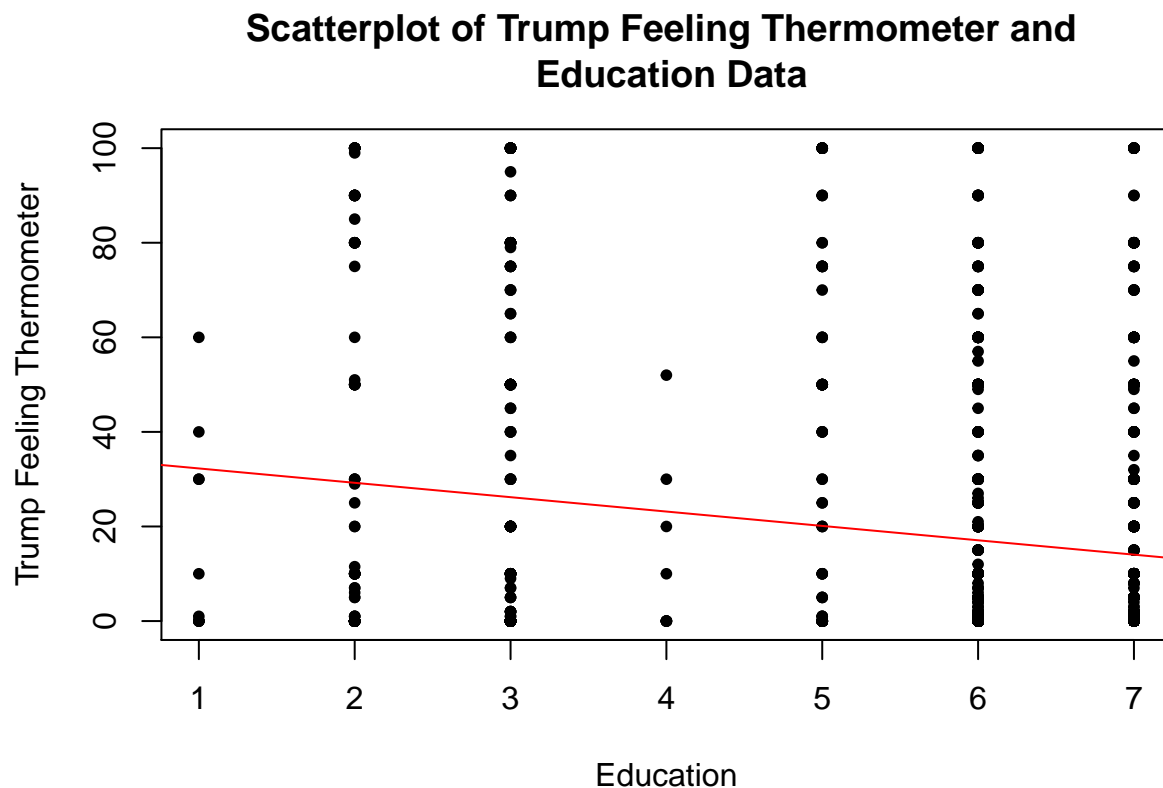
```
## 4
## -14
```

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.)

Trump scatterplot before jitter

```
#Create a plot specifying education on x-axis, Trump thermometer on y-axis, labels of axes and graph title
plot(lab$education, lab$therm_trump,
     type = "p",
     pch = 20,
     xlab = 'Education',
     ylab = 'Trump Feeling Thermometer',
     main = 'Scatterplot of Trump Feeling Thermometer and \n Education Data')

#Add fitted line using regression results (results1)
abline(a=coef(results1)[1], b=coef(results1)[2], col=2)
```



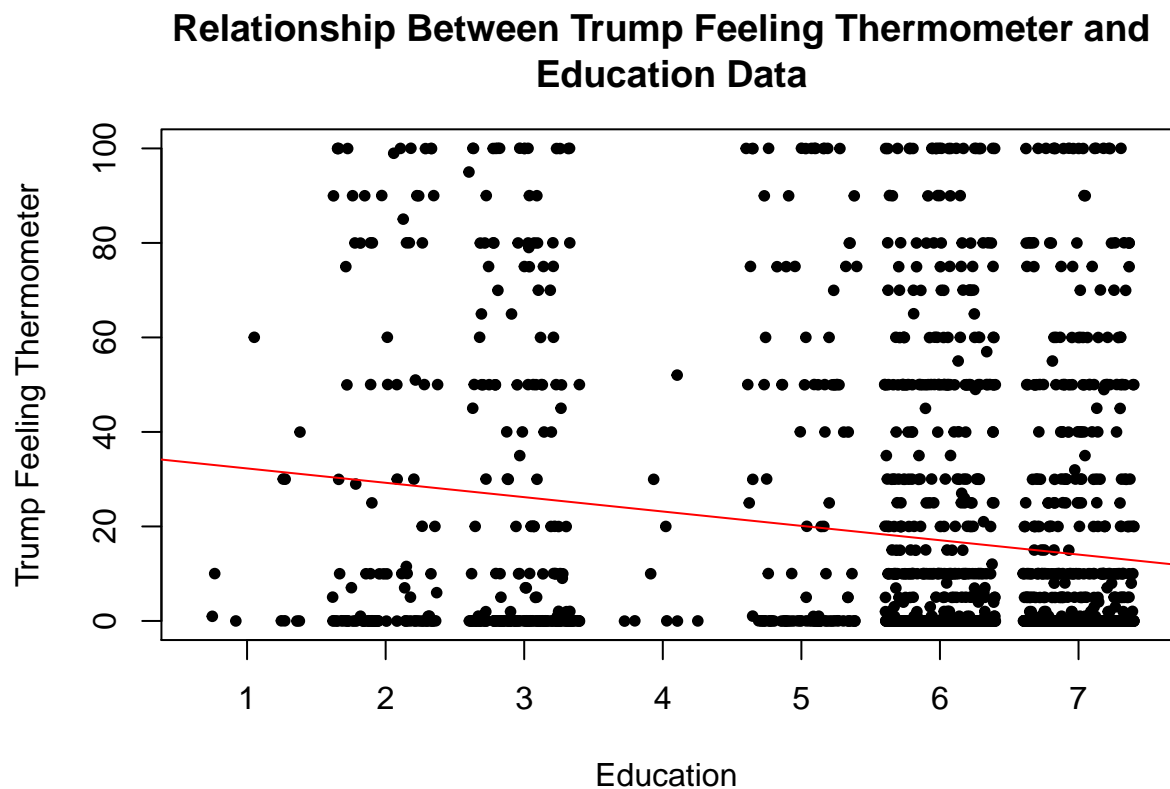
Trump scatterplot after jitter

```

#Create a plot specifying education on x-axis, Trump thermometer on y-axis, labels of axes and graph title
#Add jitter subcommand that adds random noise to plot and aids visualization
plot(jitter(lab$education, 2), jitter(lab$therm_trump, 2),
     type = "p",
     pch = 20,
     xlab = 'Education',
     ylab = 'Trump Feeling Thermometer',
     main = 'Relationship Between Trump Feeling Thermometer and \n Education Data')

#Add fitted line using regression results (results1)
abline(a=coef(results1)[1], b=coef(results1)[2], col=2)

```



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

```

#Set y2 variable to therm_clinton column, and set x2 variable to education column
y2 = lab$therm_clinton
x2 = lab$education

#Run linear model with therm_clinton as dependent variable (y2), education as independent variable (x2)
results2 <- lm(y2~x2, data = lab)
summary(results2)

```

```

##
## Call:
## lm(formula = y2 ~ x2, data = lab)

```

```
##
## 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 ***
## x2             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
```

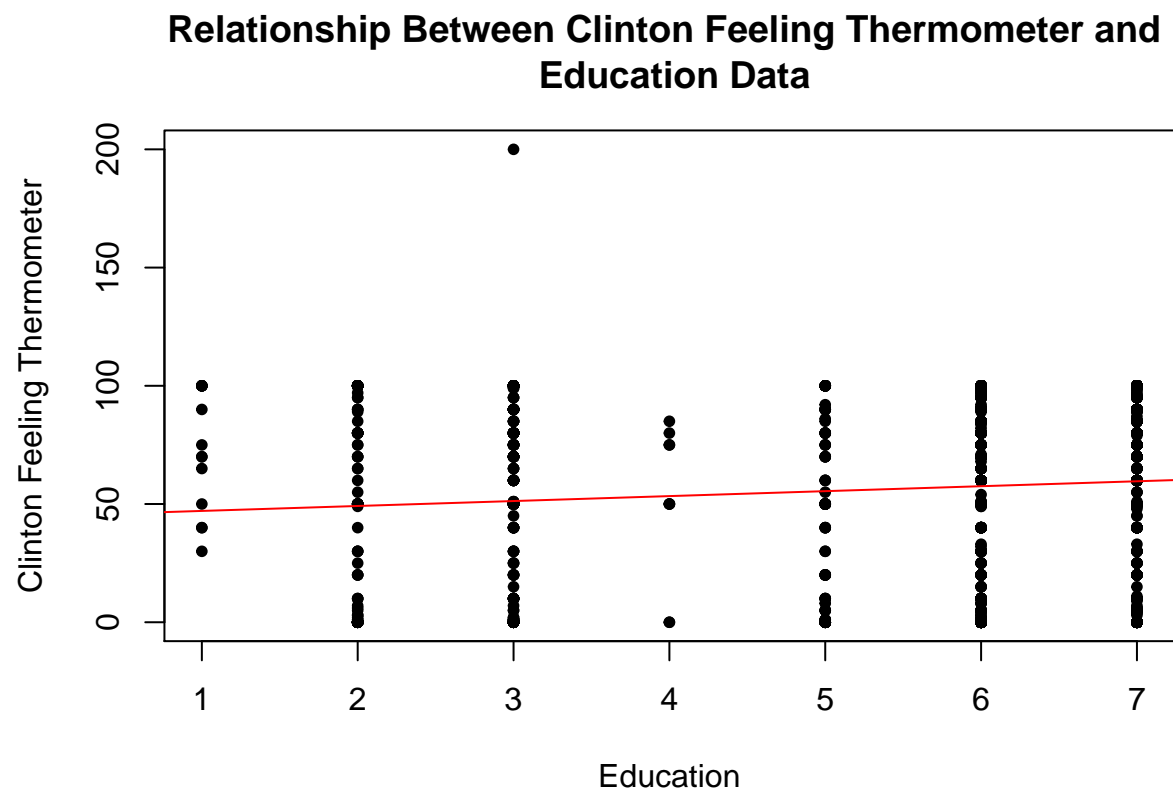
The slope coefficient from the model above is 2.090, meaning that education is positively correlated with the Clinton thermometer. Specifically, for every increase in unit of education, the Clinton thermometer increases by 2.090. One can infer from the observations that a higher level of education would indicate a higher favoring of 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.)**

Clinton scatterplot before jitter

```
#Create a plot specifying education on x-axis, Clinton thermometer on y-axis, labels of axes and graph
plot(lab$education, lab$therm_clinton,
     type = "p",
     pch=20,
     xlab= 'Education',
     ylab= 'Clinton Feeling Thermometer',
     main='Relationship Between Clinton Feeling Thermometer and \n Education Data')

#Add fitted line using regression results (results2)
abline(a=coef(results2)[1], b=coef(results2)[2], col=2)
```

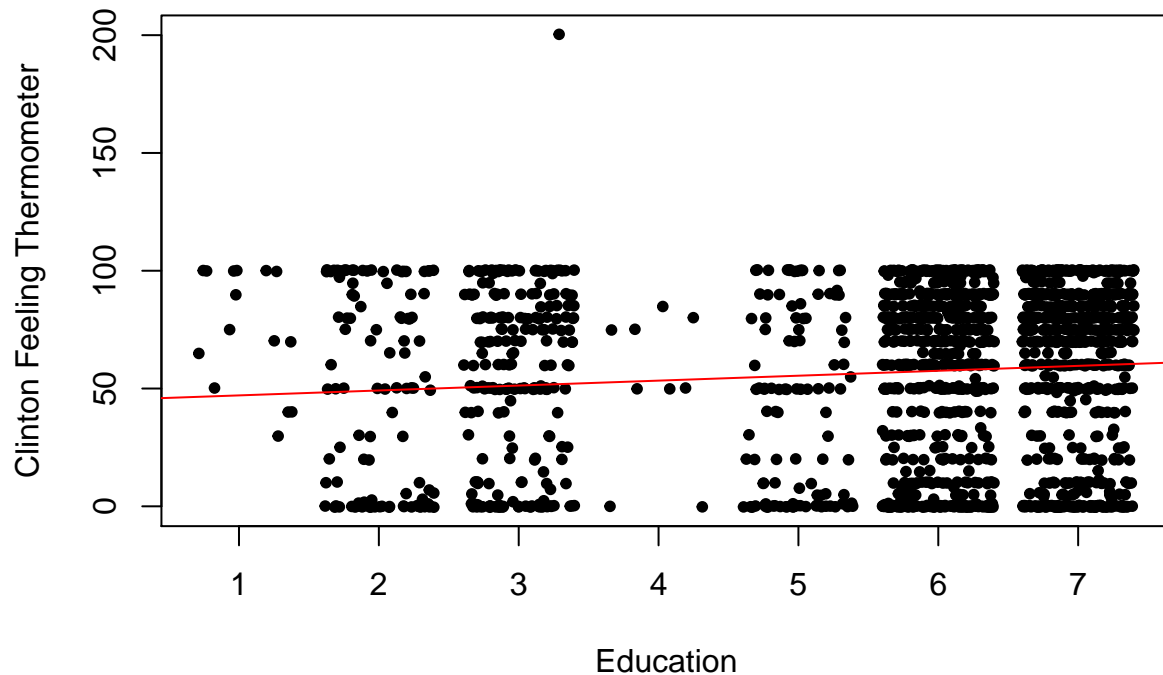


Clinton scatterplot after jitter

```
plot(jitter(lab$education, 2), jitter(lab$therm_clinton, 2),
     type = "p",
     pch = 20,
     xlab = 'Education',
     ylab = 'Clinton Feeling Thermometer',
     main = 'Relationship Between Clinton Feeling Thermometer and \n Education Data')

#Add fitted line using regression results (results1)
abline(a=coef(results2)[1], b=coef(results2)[2], col=2)
```

## Relationship Between Clinton Feeling Thermometer and Education Data



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.

```
#Create linear model as above, but only select rows 1-400 from lab dataset, and summarize results of li
results3 <- lm(formula = therm_clinton ~ education, data=lab[1:400,])
summary(results3)
```

```
##
## Call:
## lm(formula = therm_clinton ~ education, data = lab[1:400, ])
##
## 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 ***
## education       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
```

The Beta hat 1 for the regression model above is now 2.42, which is a higher OLS estimator than the original value of 2.090 with the full dataset. The standard error, however, increases to 1.53, which is higher than the original value of 0.517 with the full dataset. Since the standard error increases, this implies that our second regression model with only 400 observations does not fit the data as well as our original regression model which contained 2359 observations.

```
#Counting number of rows in original dataset - used for answer above
nrow(lab)
```

```
## [1] 2359
```

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?

```
#Create Republican dummy variable that specifies political party can equal 5, 6, or 7 (taken from survey)
lab$Rep <- (lab$pol_party == 5 | lab$pol_party == 6 | lab$pol_party == 7)

#Run linear model with therm_trump as dependent variable, education as independent variable, and specifying data=lab
results4 <- lm(therm_trump ~ education, data=lab[lab$Rep==1,])

#Use robust standard errors and display OLS results (results4) from linear model above
coeftest(results4, vcov. = vcovHC(results4, 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
```

As seen above, the slope coefficient is -6.010 and t-statistic is -6.11.

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.

```
#Create new female variable, with gender equalling 2
lab$female <- (lab$gender==2)

#Create linear model with Trump thermometer as a function of gender for women
results5 <- lm(therm_trump ~ female, data = lab)

#Use robust standard errors and display OLS results (results5) from linear model above
coeftest(results5, vcov. = vcovHC(results5, 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 from the model above is -4.42, which suggests that female education is negatively correlated with the Trump thermometer. Specifically, for every increase in unit of education, the Trump thermometer drops by -4.42. One can infer from the observations that a higher level of education among women would indicate a more negative feeling towards Trump. The intercept is 20.32, which indicates the Trump thermometer value when female education is zero.

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.

```
#Create variable that stores the mean value for Trump thermometer for women
therm_trump_female <- mean(lab$therm_trump[lab$female==1], na.rm = TRUE)

#Create variable that stores the mean value for Trump thermometer for others
therm_trump_others <- mean(lab$therm_trump[lab$female==0], na.rm = TRUE)

#Calculate the difference between the mean value for Trump thermometer for women and mean value for Trump thermometer for others
therm_trump_female - therm_trump_others
```

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

```
#View result for the mean value for Trump thermometer for women
therm_trump_female
```

```
## [1] 15.9
```

```
#View result for the mean value for Trump thermometer for others
therm_trump_others
```

```
## [1] 20.3
```

The mean value for Trump thermometer for women is 15.9, while the mean value Trump thermometer for others is 20.3. We can relate the low value of 15.9 to the regression model above. As seen in the model above, the Beta hat 1 value was -4.42, implying that every increase in unit of education for females caused the Trump thermometer to drop by -4.42.