

Problem Set 7: Linear Regression

Joe Ornstein

Due November 11, 2020

Create an R project, an R script, load tidyverse, and import the modified Freedom In The World dataset you created for the midterm.

```
library(tidyverse)
library(readxl)

# load Freedom in the World dataset
FIW <- read_xlsx(path = 'data/2020_All_Data_FIW_2013-2020.xlsx', skip = 1, sheet = 'FIW13-20') %>%
  filter(`C/T` == 'c',
         Edition == 2020) %>%
  mutate(democracy_score = A1 + A2 + A3 + B1 + B2 + B3 + B4 + C1 + C2 + C3,
         liberty_score = D1 + D2 + D3 + D4 + E1 + E2 + E3 + F1 + F2 + F3 + F4 + G1 + G2 + G3 + G4)
```

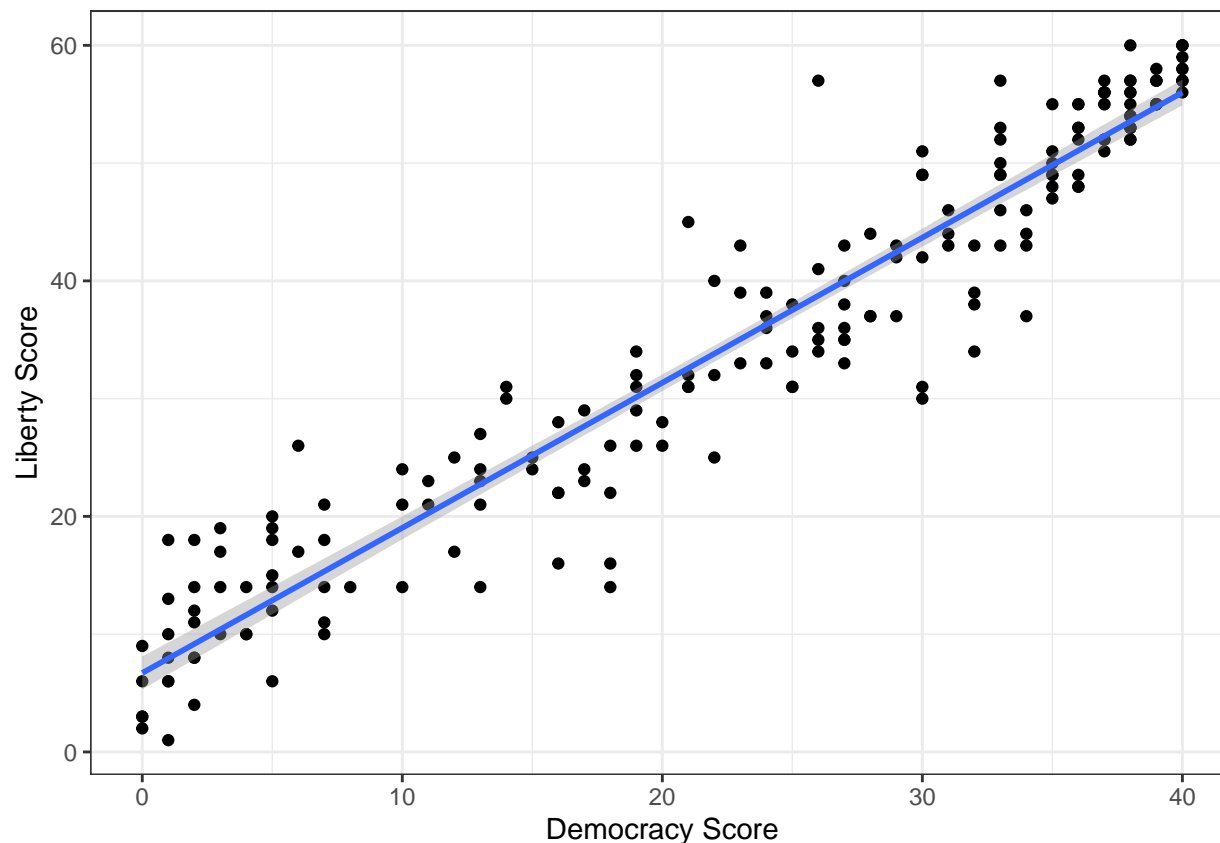
Line of Best Fit

1. Estimate a line of best fit with `liberty_score` as the outcome and `democracy_score` as the explanatory variable.

```
lm1 <- lm(liberty_score ~ democracy_score, data = FIW)
```

2. Visualize the line of best fit with `geom_smooth(method = 'lm')`.

```
ggplot(data = FIW) +
  geom_point(mapping = aes(x=democracy_score, y=liberty_score)) +
  geom_smooth(mapping = aes(x=democracy_score, y=liberty_score),
             method = 'lm') +
  labs(x='Democracy Score', y='Liberty Score') +
  theme_bw()
```



3. What is the slope of the relationship? What is the 95% confidence interval around that slope?

```
coef(lm1)['democracy_score']
```

```
## democracy_score
##      1.232325
```

```
confint(lm1)['democracy_score',]
```

```
##      2.5 %   97.5 %
## 1.180489 1.284161
```

4. Are there any outliers with higher or lower liberty scores than you would expect given their democracy score? Which countries are they? (Hint: there is a vector called `residuals` in the `lm` object you just created. Add it to your FIW dataframe).

```
FIW <- FIW %>%
  mutate(lm1_residual = lm1$residuals)
```

```
# sort by residuals and display the top 5
```

```
FIW %>%
  arrange(-lm1_residual) %>%
  select(`Country/Territory`, democracy_score, liberty_score, lm1_residual) %>%
  head(5)
```

```
## # A tibble: 5 x 4
##   `Country/Territory` democracy_score liberty_score lm1_residual
##   <chr>                <dbl>         <dbl>         <dbl>
## 1 Monaco                26             57          18.3
```

## 2 Benin	21	45	12.4
## 3 Thailand	6	26	11.9
## 4 Eswatini	1	18	10.1
## 5 Liechtenstein	33	57	9.63

Monaco, Benin, Thailand, Eswatini, and Liechtenstein are all more free than we would expect given their level of democracy. (BTW: I had no idea Swaziland was officially renamed in 2018.)

```
FIW %>%
  arrange(lm1_residual) %>%
  select(`Country/Territory`, democracy_score, liberty_score, lm1_residual) %>%
  head(5)
```

```
## # A tibble: 5 x 4
##   `Country/Territory` democracy_score liberty_score lm1_residual
##   <chr>                <dbl>          <dbl>          <dbl>
## 1 Iraq                18            14          -14.9
## 2 Bhutan              30            30          -13.7
## 3 Myanmar             18            16          -12.9
## 4 Indonesia           30            31          -12.7
## 5 El Salvador         32            34          -12.1
```

Iraq, Bhutan, Myanmar, Indonesia, and El Salvador are all *less* free than we would expect given their level of democracy.

Data Wrangling and OLS

I'm including another dataset that I pulled from the World Bank with GDP per capita figures for each country since 1960. Man, it's a mess. We're going to have to tidy it up.

5. Read the dataset into R and pivot the data so each row represents a country-year.
6. Keep the most recent year of gdp per capita
7. Create a new variable, `log_gdp_per_capita`, equal to the logarithm of GDP per capita.

```
WB <- read_csv('data/API_NY.GDP.PCAP.PP.CD_DS2_en_csv_v2_1495136.csv',
  skip = 4) %>%
  select(-c(`Indicator Name`, `Indicator Code`, X66)) %>%
  pivot_longer(cols = `1960`:`2020`,
    names_to = 'year',
    values_to = 'gdp_per_capita') %>%
  filter(year == 2019) %>%
  mutate(log_gdp_per_capita = log(gdp_per_capita))
```

8. Save your cleaned up World Bank dataset to the `data/` folder.

```
write_csv(WB, 'data/cleaned_WB.csv')
```

9. Merge your cleaned up World Bank dataset with the FIW dataset.

```
# make sure the key variables have the same name
WB <- WB %>%
  mutate(country_name = `Country Name`)

# merge with left_join
FIW <- FIW %>%
  mutate(country_name = `Country/Territory`) %>%
  left_join(WB, by = 'country_name')
```

10. Estimate a multivariable linear model with `liberty_score` as the outcome variable and `democracy_score` and `log_gdp_per_capita` as the explanatory variables. What are the slope coefficients and 95% confidence intervals?

```
lm2 <- lm(liberty_score ~ democracy_score + log_gdp_per_capita,
          data = FIW)

summary(lm2)

##
## Call:
## lm(formula = liberty_score ~ democracy_score + log_gdp_per_capita,
##     data = FIW)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9465  -2.2936  -0.1101   2.6223  13.9915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.59154    2.90273  -1.237  0.217779
## democracy_score    1.17446    0.03106  37.812 < 2e-16 ***
## log_gdp_per_capita 1.22092    0.33068   3.692 0.000304 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.415 on 161 degrees of freedom
## (31 observations deleted due to missingness)
## Multiple R-squared:  0.9213, Adjusted R-squared:  0.9203
## F-statistic: 942.4 on 2 and 161 DF,  p-value: < 2.2e-16
```

The estimated slope coefficient for `democracy_score` is a bit smaller than it was before, but not by much. Democracies tend to be wealthier on average, and wealthier places tend to have higher Freedom House scores, but it is not enough to fully explain the relationship we found between democracy and liberty in the midterm.

```
ggplot(FIW) +
  geom_text(aes(x=log_gdp_per_capita, y=liberty_score,
               label=`Country Code`)) +
  scale_x_log10() +
  labs(x='Log GDP Per Capita', y='Liberty Score') +
  theme_bw()
```

