# 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 datset
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)
```

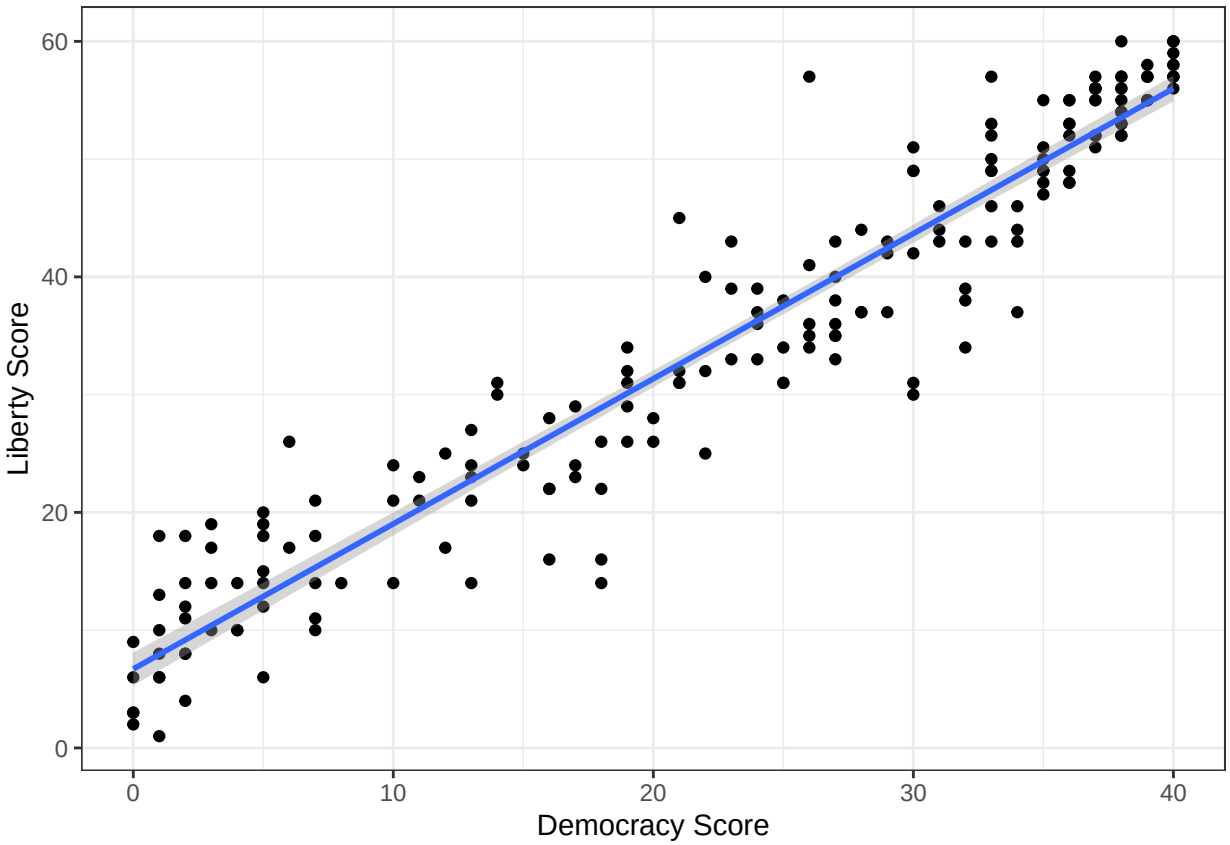
## Line of Best Fit

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 %>%
  rename(country_name = `Country Name`)

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

```r
# Which countries failed to merge?
data %>%
  filter(is.na(`Country Code`)) %>%
  pull(country_name)
```

```
##  [1] "Bahamas"             "Brunei"              "Congo (Brazzaville)"
##  [4] "Congo (Kinshasa)"    "Egypt"               "Iran"
##  [7] "Kyrgyzstan"          "Laos"                "Micronesia"
## [10] "North Korea"         "Russia"              "Slovakia"
## [13] "South Korea"         "Syria"               "Taiwan"
## [16] "The Gambia"          "Venezuela"           "Yemen"
```

```r
# Rename those countries in the WB dataset
WB <- WB %>%
  mutate(country_name = case_when(country_name == 'Bahamas, The' ~ 'Bahamas',
                                  country_name == 'Brunei Darussalam' ~ 'Brunei',
                                  country_name == 'Congo, Rep.' ~ 'Congo (Brazzaville)',
                                  country_name == 'Congo, Dem. Rep.' ~ 'Congo (Kinshasa)',
                                  country_name == 'Egypt, Arab Rep.' ~ 'Egypt',
                                  country_name == 'Iran, Islamic Rep.' ~ 'Iran',
                                  country_name == 'Kyrgyz Republic' ~ 'Kyrgyzstan',
                                  country_name == 'Lao PDR' ~ 'Laos',
                                  country_name == 'Micronesia, Fed. Sts.' ~ 'Micronesia',
                                  country_name == 'Korea, Dem. People's Rep.' ~ 'North Korea',
                                  country_name == 'Russian Federation' ~ 'Russia',
                                  country_name == 'Slovak Republic' ~ 'Slovakia',
                                  country_name == 'Korea, Rep.' ~ 'South Korea',
                                  country_name == 'Syrian Arab Republic' ~ 'Syria',
                                  country_name == 'Gambia, The' ~ 'The Gambia',
                                  country_name == 'Venezuela, RB' ~ 'Venezuela',
                                  country_name == 'Yemen, Rep.' ~ 'Yemen',
                                  TRUE ~ country_name))

# Try the merge again
data <- FIW %>%
  rename(country_name = `Country/Territory`) %>%
  left_join(WB, by = 'country_name')

# Now only Taiwan fails to merge
# (there's no separate entry in the World Bank Data)
data %>%
  filter(is.na(`Country Code`)) %>%
  pull(country_name)
```

```
## [1] "Taiwan"
```

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?

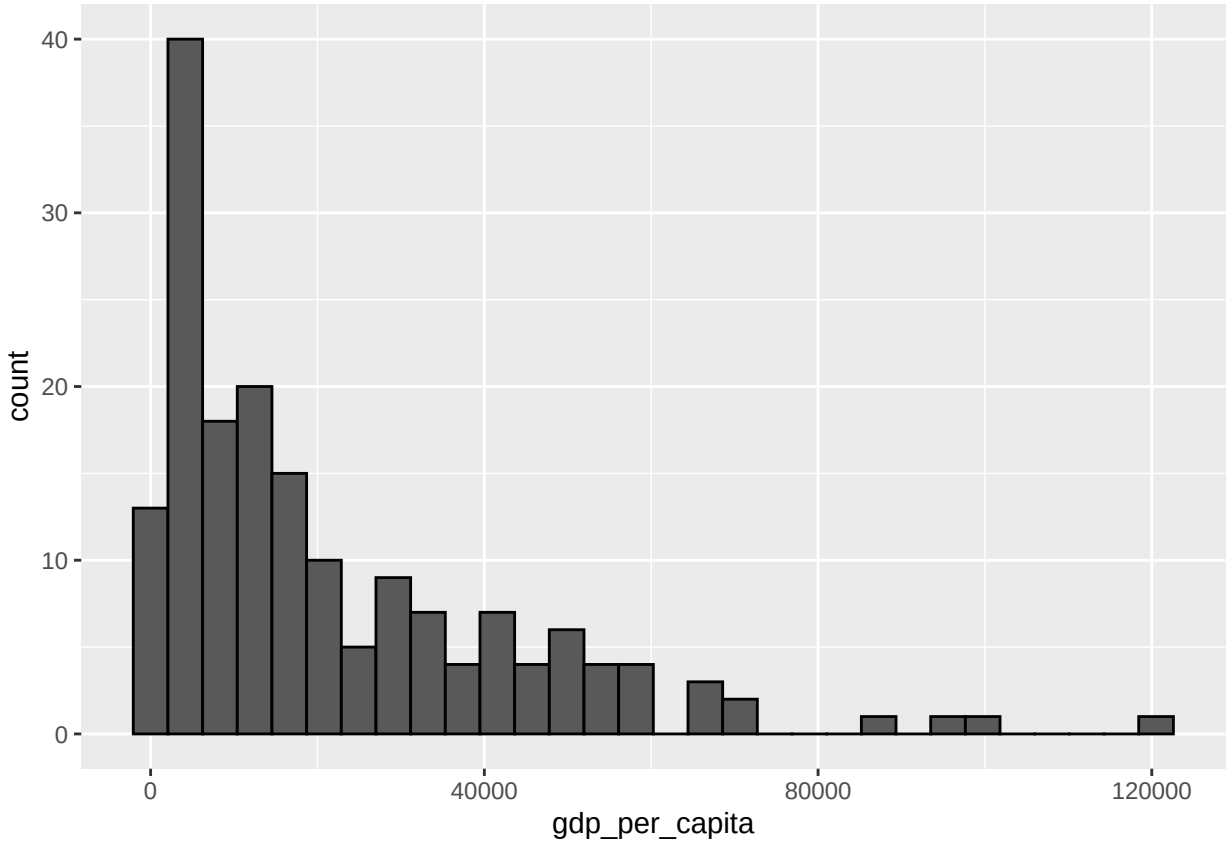
```r
lm2 <- lm(liberty_score ~ democracy_score + log_gdp_per_capita,
          data = data)

summary(lm2)
```

```
##
## Call:
## lm(formula = liberty_score ~ democracy_score + log_gdp_per_capita,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1459  -2.2123  -0.0718   2.5399  13.8364
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.14990    2.76569  -1.139 0.256319
## democracy_score     1.16026    0.02906  39.930  < 2e-16 ***
## log_gdp_per_capita  1.22235    0.31446   3.887 0.000145 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.375 on 172 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.9243, Adjusted R-squared:  0.9234
## F-statistic:  1049 on 2 and 172 DF,  p-value: < 2.2e-16
```

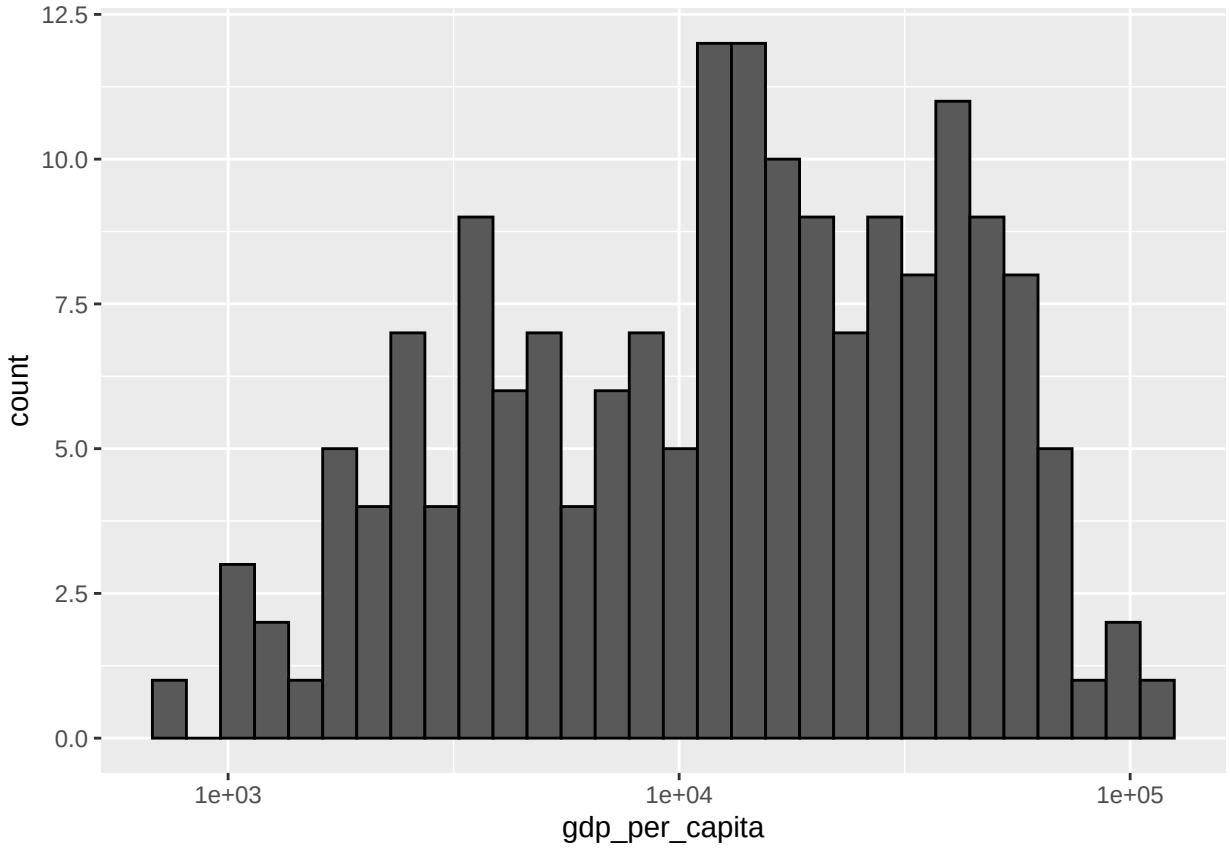
The estimated slope coefficient for `democracy_score` is a bit smaller than it was before, but not by much. Democracies tend to be wealthier on average, and wealthier places tend to have higher Freedom House scores, but it is not enough to fully explain the relationship we found between democracy and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

### Why Log GDP Per Capita?

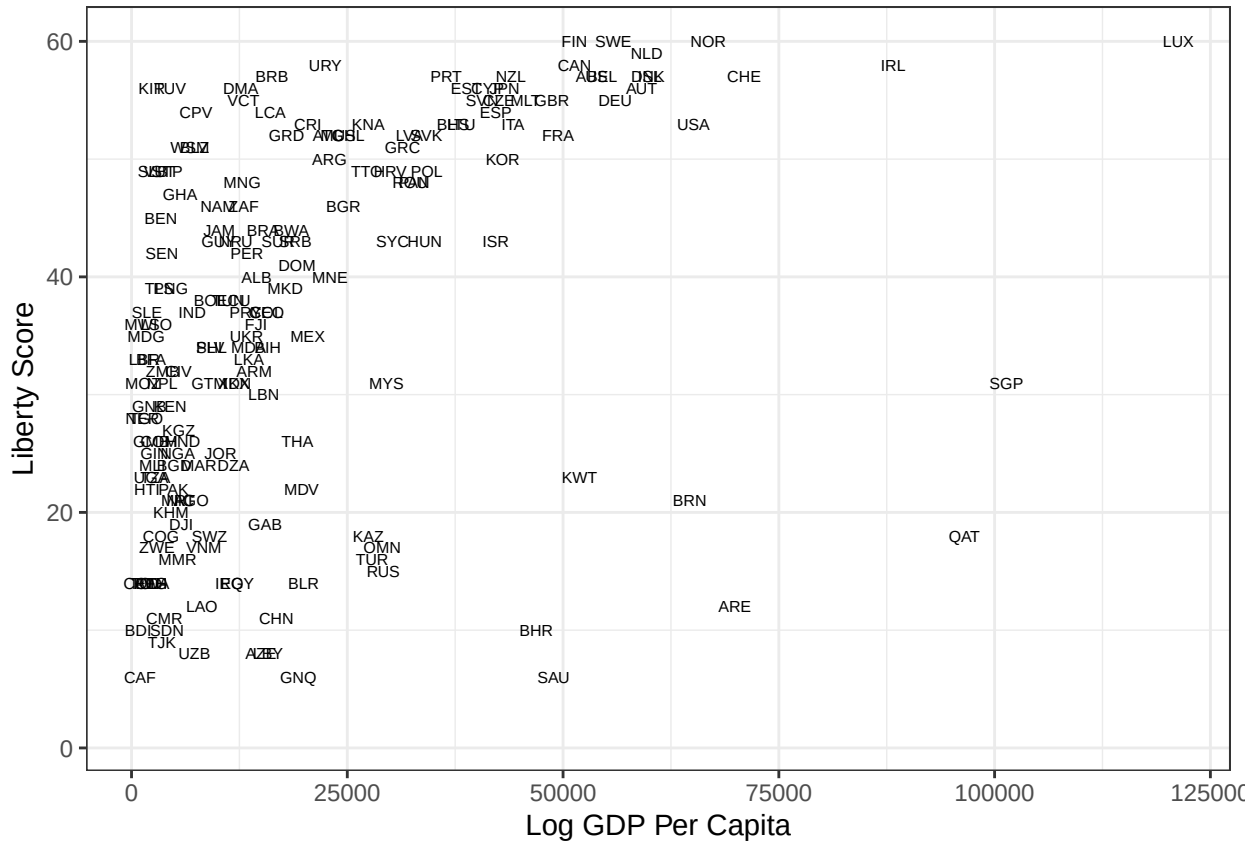
```
# GDP per capita is right-skewed
ggplot(data) +
  geom_histogram(aes(x=gdp_per_capita), color = 'black')
```



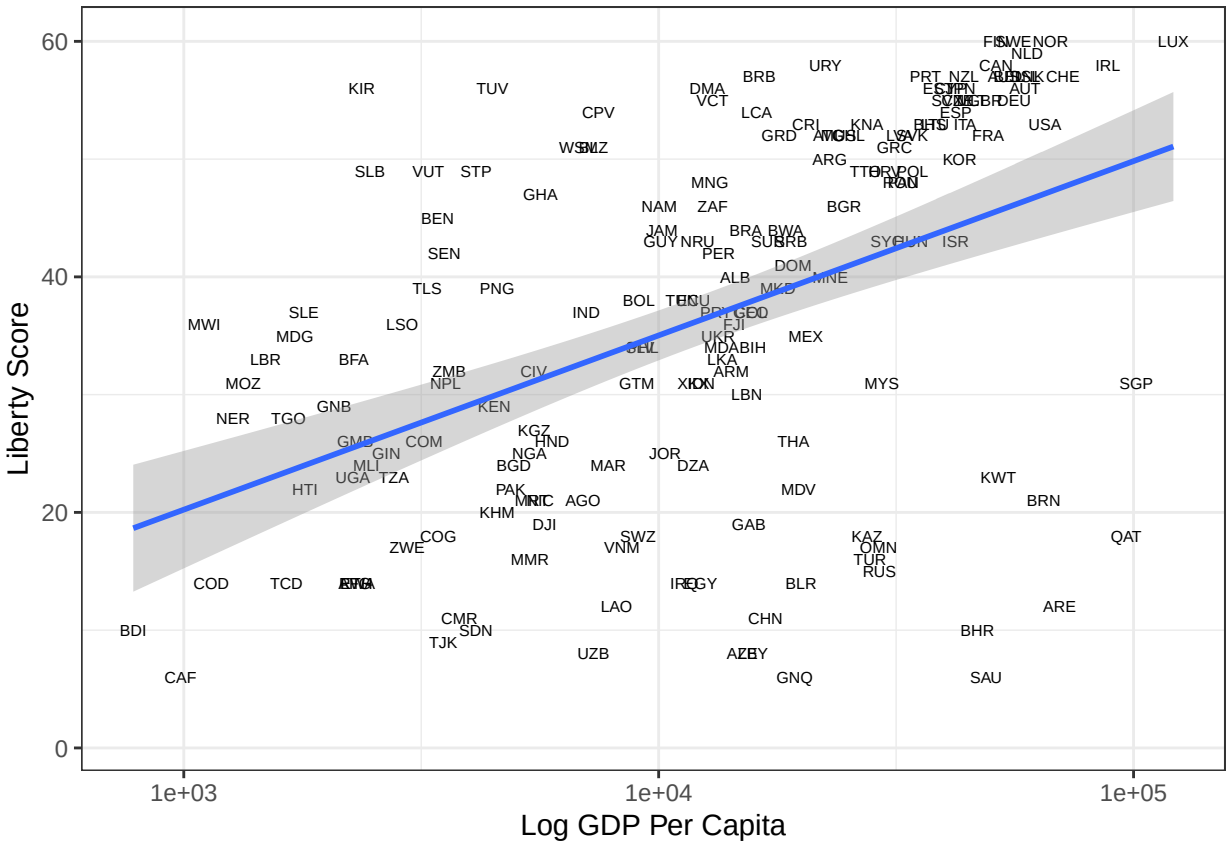

```
# On a log scale it's a bit more normal
ggplot(data) +
  geom_histogram(aes(x=gdp_per_capita), color = 'black') +
  scale_x_log10()
```




```
# Not a linear relationship!
ggplot(data) +
  geom_text(aes(x=gdp_per_capita, y=liberty_score,
                label=`Country Code`), size = 2) +
  labs(x='Log GDP Per Capita',y='Liberty Score') +
  theme_bw()
```




```
# This is more plausibly linear
ggplot(data) +
  geom_text(aes(x=gdp_per_capita, y=liberty_score,
                label=`Country Code`), size = 2) +
  scale_x_log10() +
  geom_smooth(aes(x=gdp_per_capita, y = liberty_score),
              method = 'lm') +
  labs(x='Log GDP Per Capita',y='Liberty Score') +
  theme_bw()
```


Liberty Score
Log GDP Per Capita
60
40
20
0
1e+03
1e+04
1e+05
LUX
IRL
KIR
TUV
DMA
BRB
URY
CHE
CPV
LCA
USA
FRA
SLB
VUT
STP
KOR
GHA
MNG
BEN
NAM
ZAF
BGR
SEN
PER
DOM
ALB
TLS
PNG
MWI
SLE
IND
MDG
LSO
FJI
MEX
LBR
BFA
ZMB
CIV
MOZ
NPL
GTM
MYS
SGP
NER
TGO
GNB
KEN
LBN
GMB
COM
KGZ
HND
THA
GIN
NGA
JOR
MLI
BGD
MAR
DZA
KWT
UGA
TZA
HTI
PAK
MDV
AGO
BRN
KHM
DJI
GAB
ZWE
COG
SWZ
VNM
KAZ
QAT
MMR
TUR
RUS
COD
TCD
BLR
LAO
ARE
CMR
CHN
BDI
SDN
BHR
TJK
UZB
CAF
GNQ
SAU