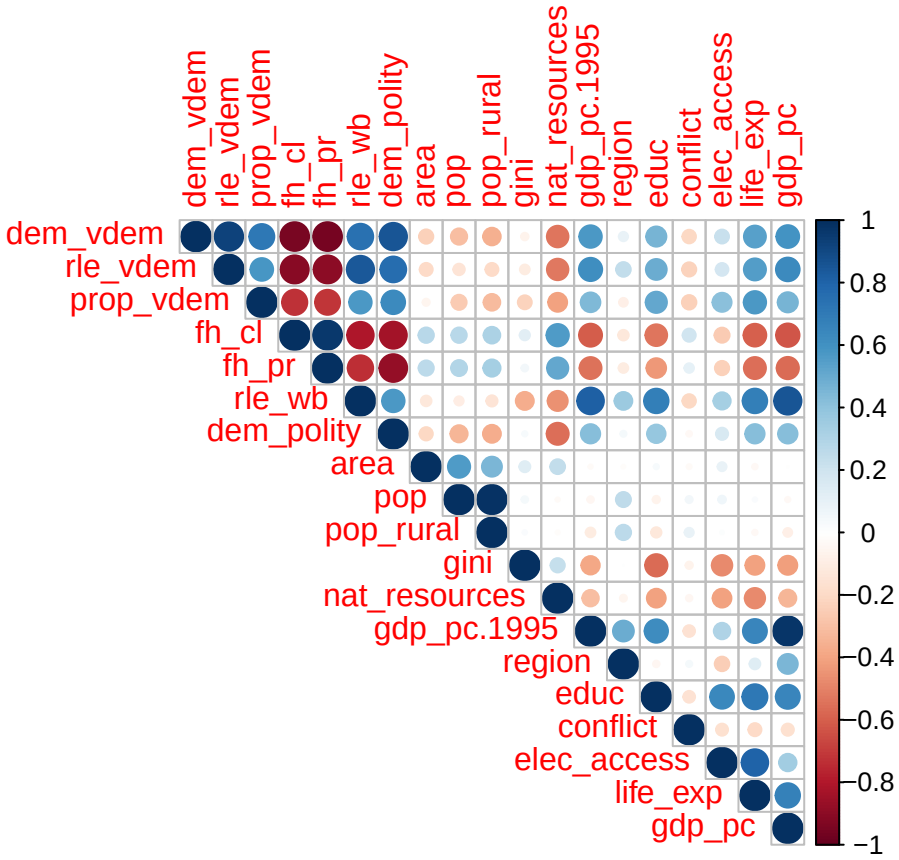
Midterm

Colden Johnson

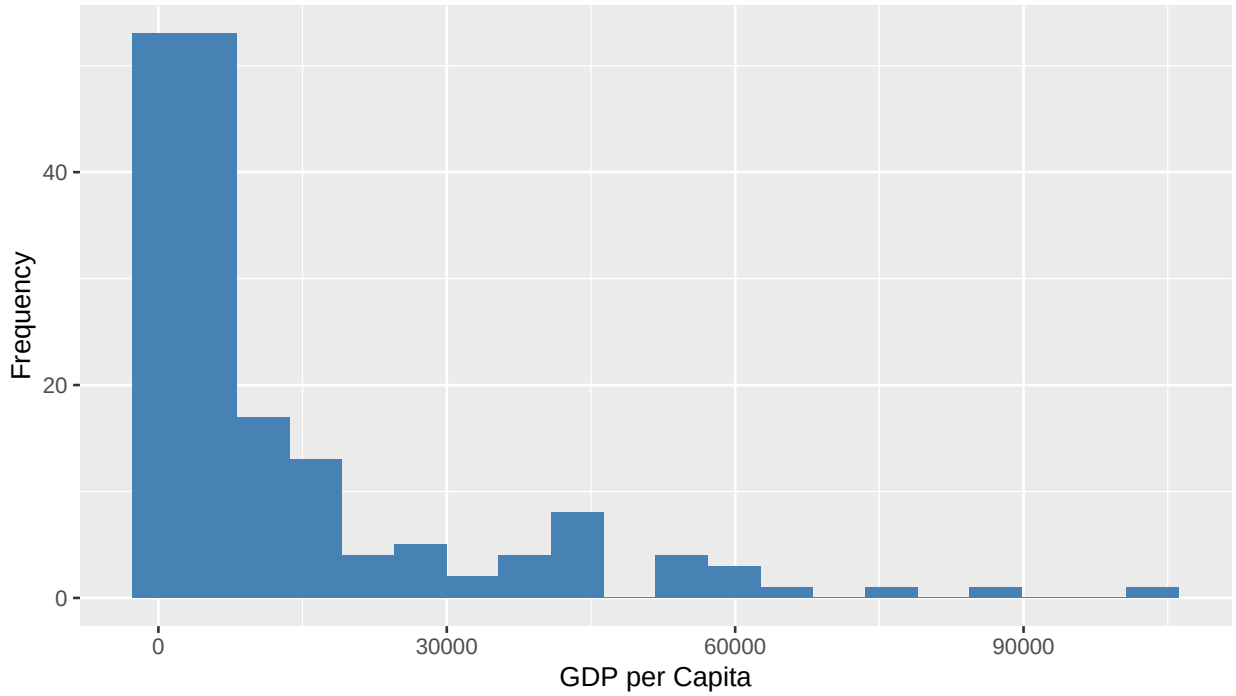
2022-11-21

```
# I'm starting out by creating an exploratory correlation plot to look at how the various indices in th
corrplotdata <- demindex[-c(1,2)]
corrplotdata <- drop_na(corrplotdata)
corrplotdata.cor = cor(corrplotdata)
corrplot(corrplotdata.cor, type = "upper")
```

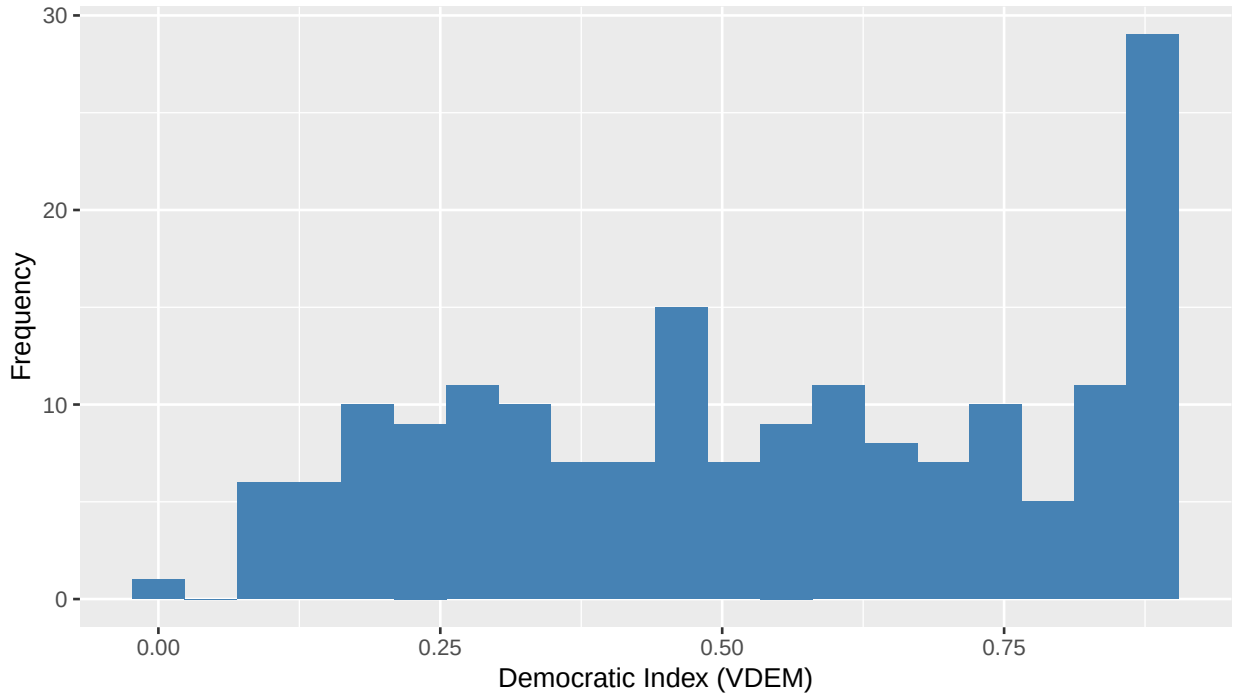



```
# This data clearly needs to be log transformed
ggplot(demindex, aes(x=gdp_pc)) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('GDP per Capita') +
  ylab('Frequency')
```

```
## Warning: Removed 9 rows containing non-finite values (‘stat_bin()‘).
```

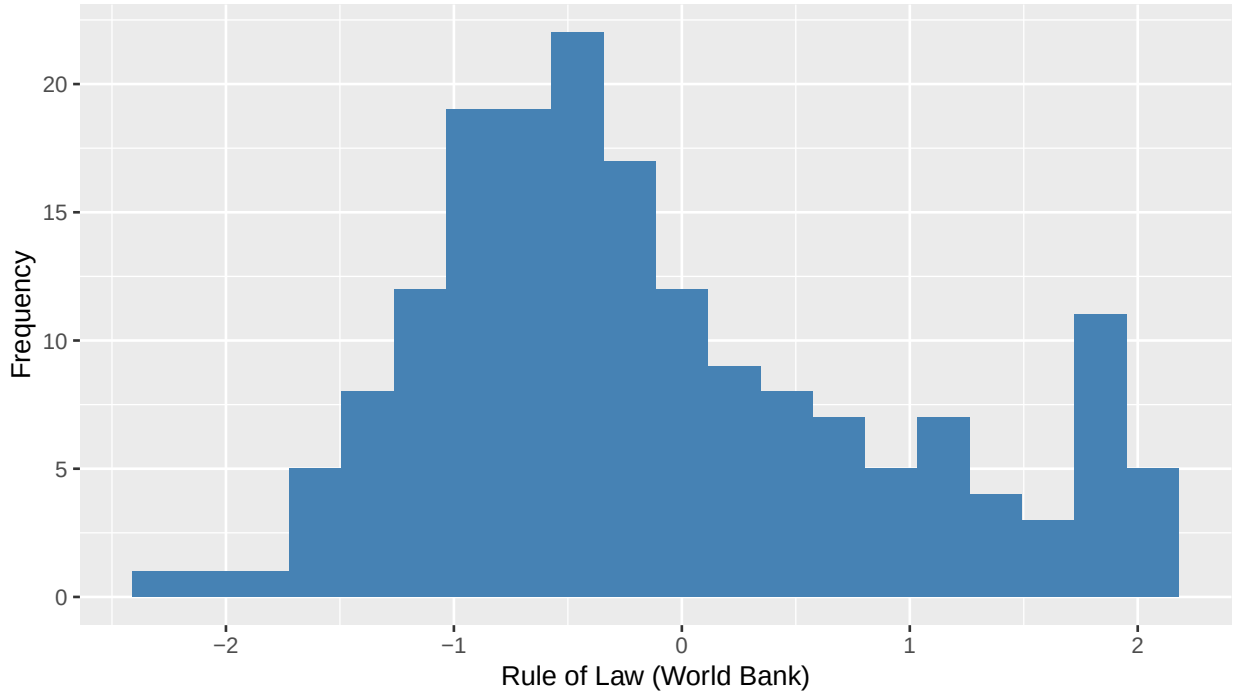



```
ggplot(demindex, aes(x=dem_vdem)) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Democratic Index (VDEM)') +
  ylab('Frequency')
```



```r
ggplot(demindex, aes(x = rle_wb)) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Rule of Law (World Bank)') +
  ylab('Frequency')
```

```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```

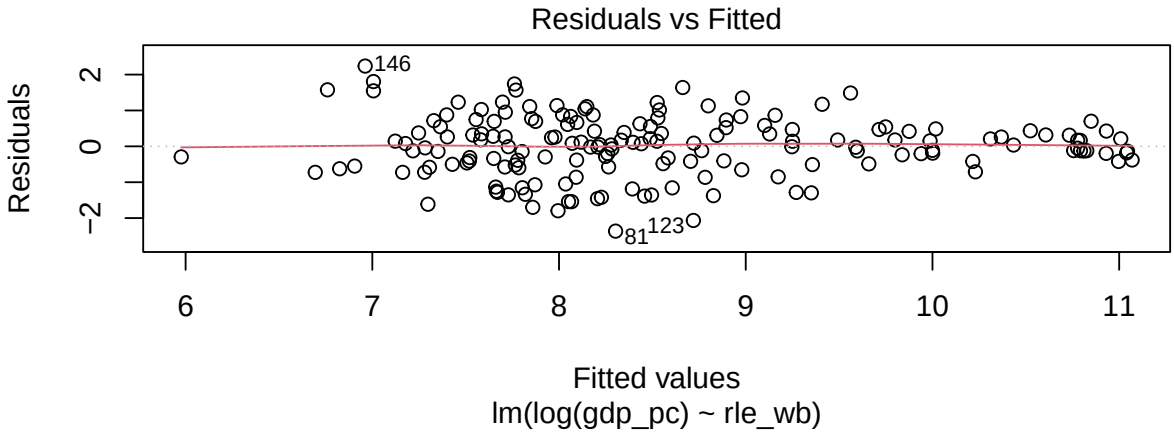


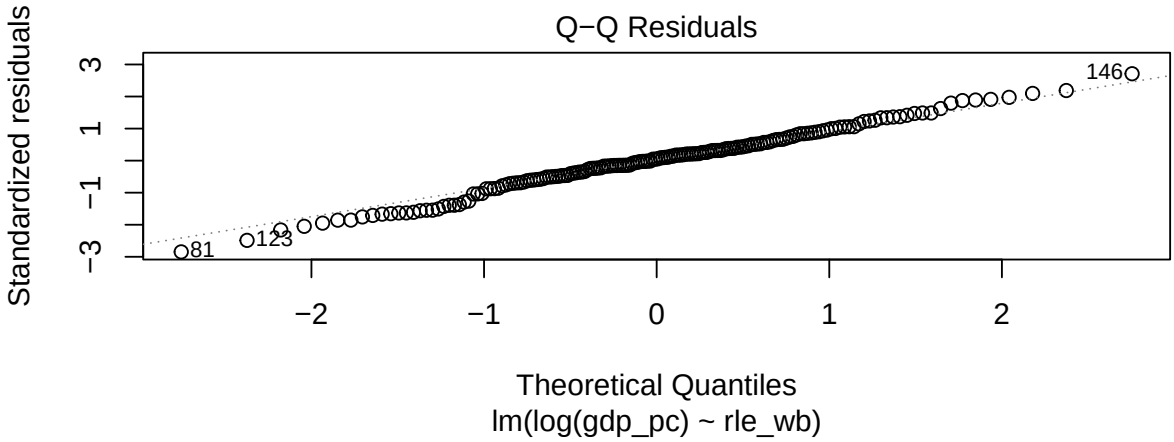
How to measure democracy and rule of law?

To measure democracy and rule of law we should select one of the indices provided in the database. While fh_cl (civil liberties) and fh_pr (political rights) are likely both strong measures of democracy, the dem_vdem democracy index is a better choice here, as it directly measures what is being discussed. However, it is nice to notice that in the corrplot fh_cl and fh_pr are extremely strongly correlated with dem_vdem, indicating internal validity in our data. I also considered using dem_polity, but I notice that it is almost entirely not correlated with gdp_pc and is heavily right skewed, so decide against it.

Second, rle_vdem and rle_wb are the two possible rule of law indices we can choose. Once again, it is good to see that they are moderately strongly correlated with each other. I graph histograms of the two distributions, and notice that rle_wb is much more normally distributed than rle_vdem, and so choose rle_wb as the indicator to be used.

```r
# Residuals Graph (Identifying Outliers)
outliersmodel <- lm(log(gdp_pc) ~ rle_wb, demindex)
plot(outliersmodel, c(1,2))
```






```
# dataframe of outliers
outliertibble <- data.frame(demindex[146,1:4])
outliertibble <- outliertibble %>%
  add_row(demindex[81,1:4]) %>%
  add_row(demindex[123,1:4]) %>%
  add_row(demindex[124,1:4])
outliertibble
```

```
##        country_name country_code dem_vdem rle_vdem
## 1 Equatorial Guinea          GNQ    0.183    0.035
## 2            Malawi          MWI    0.592    0.545
## 3            Rwanda          RWA    0.275    0.747
## 4           Somalia          SOM    0.171    0.125
```

```
# Remove these 4 countries
demindex <- demindex %>%
filter(!(country_code %in% c('SOM', 'GNQ', 'MWI', 'RWA')))
```

Chosen variables explained

For all 3 variables, there are very few missing values, excluding Zanzibar. GDP per capita is strongly skewed and so clearly needs to be log transformed. rle_wb is largely normal, although there is some degree of bimodality and/or outliers on the upper end of the scale. This will be addressed below.

Dem_vdem is honestly horrifying. This is entirely not a normal distribution, and cannot even be approximated using a bell curve. I attempted to square the data to normalize it, but it has a very small impact on the resulting RVF plot, and adds a lot of uneccessary complication and clutter to the model (our regression would be comparing a log scale to a squared scale, drastically reducing the actual utility). Unfortunately, this would not be solved by using a different democracy index as they all suffer from this distribution type. So, despite sticking with the indicator as is, it is important to note its limitations of use.

I also chose to remove 4 outliers, as identified using RVF plots, as well as an initial scatter plot of the indices graphed against gdp per capita. Somalia does not show up on the RVF plot, but is a very large outlier that can be seen on scatterplots, so I have chosen to remove it as well.

```
dem_econ_model1 <- lm(log(gdp_pc) ~ dem_vdem, data = demindex)
summary(dem_econ_model1)
```

```
##
## Call:
## lm(formula = log(gdp_pc) ~ dem_vdem, data = demindex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5998 -0.8619  0.0602  0.8810  3.6513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.1691     0.2353  30.463  < 2e-16 ***
## dem_vdem      2.6124     0.3883   6.728 2.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.235 on 164 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.2163, Adjusted R-squared:  0.2115
## F-statistic: 45.26 on 1 and 164 DF,  p-value: 2.742e-10
```

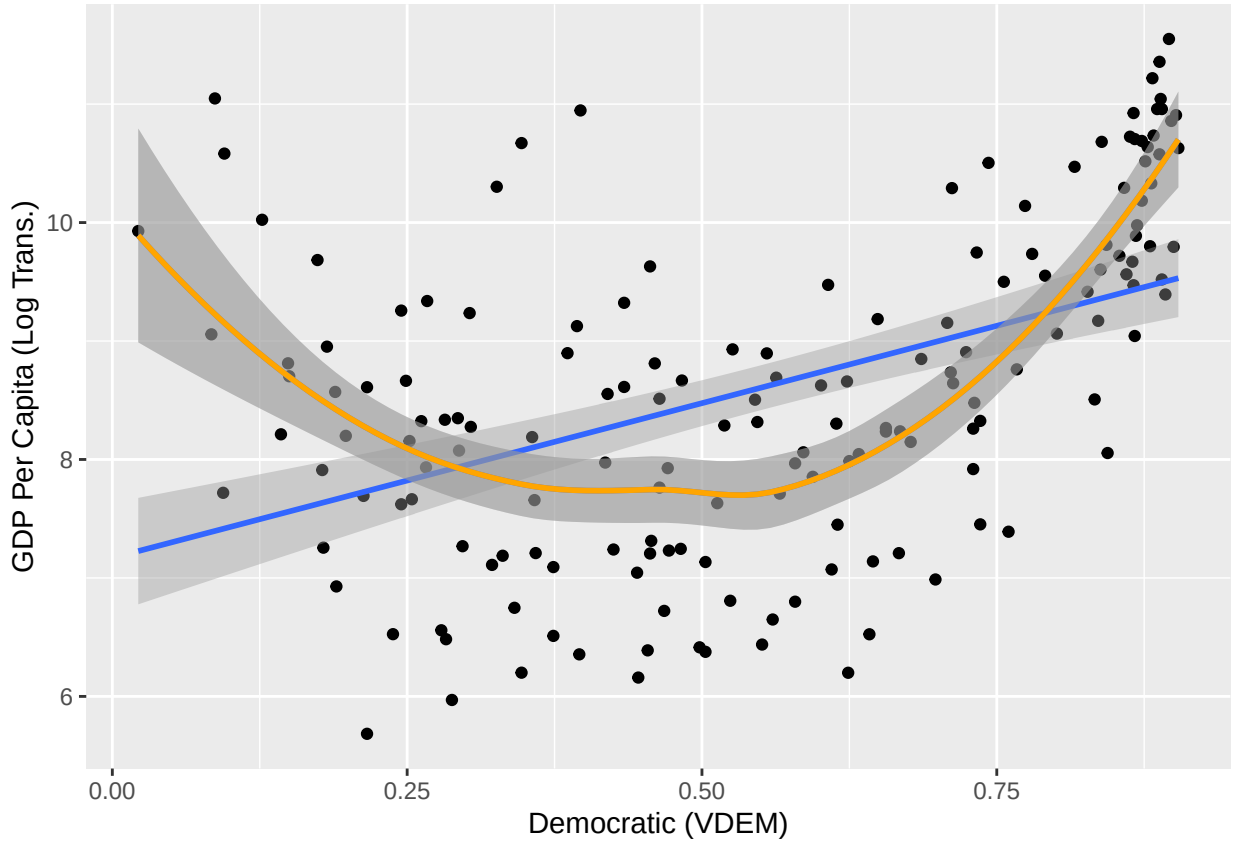
```
ggplot(demindex, aes(dem_vdem, log(gdp_pc))) +
  geom_point() +
  geom_smooth(method = 'lm', formula = 'y~x') +
  geom_smooth(method = ) +
  geom_smooth(method = 'loess', formula = 'y~x', color = 'orange') +
  xlab('Democratic (VDEM)') + ylab('GDP Per Capita (Log Trans.)')
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
## Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



Democracy linear model

In terms of models, this is simply not very good. I have included the lowess line to show just how poorly a linear model fits to this data. The reason the fit is so bad is that we have not met our pre-regression assumption of normality. Overall, the R squared of 0.2 is quite low. Interpreting coefficients, the intercept is 7.17 and for every increase of 1 in dem_vdem log(gdp_pc) is predicted to increase by 2.6. Next, we will apply some techniques to address these problems with the model.

```
rule_econ_model1 <- lm(log(gdp_pc) ~ rle_wb, data = demindex)
summary(rule_econ_model1)
```

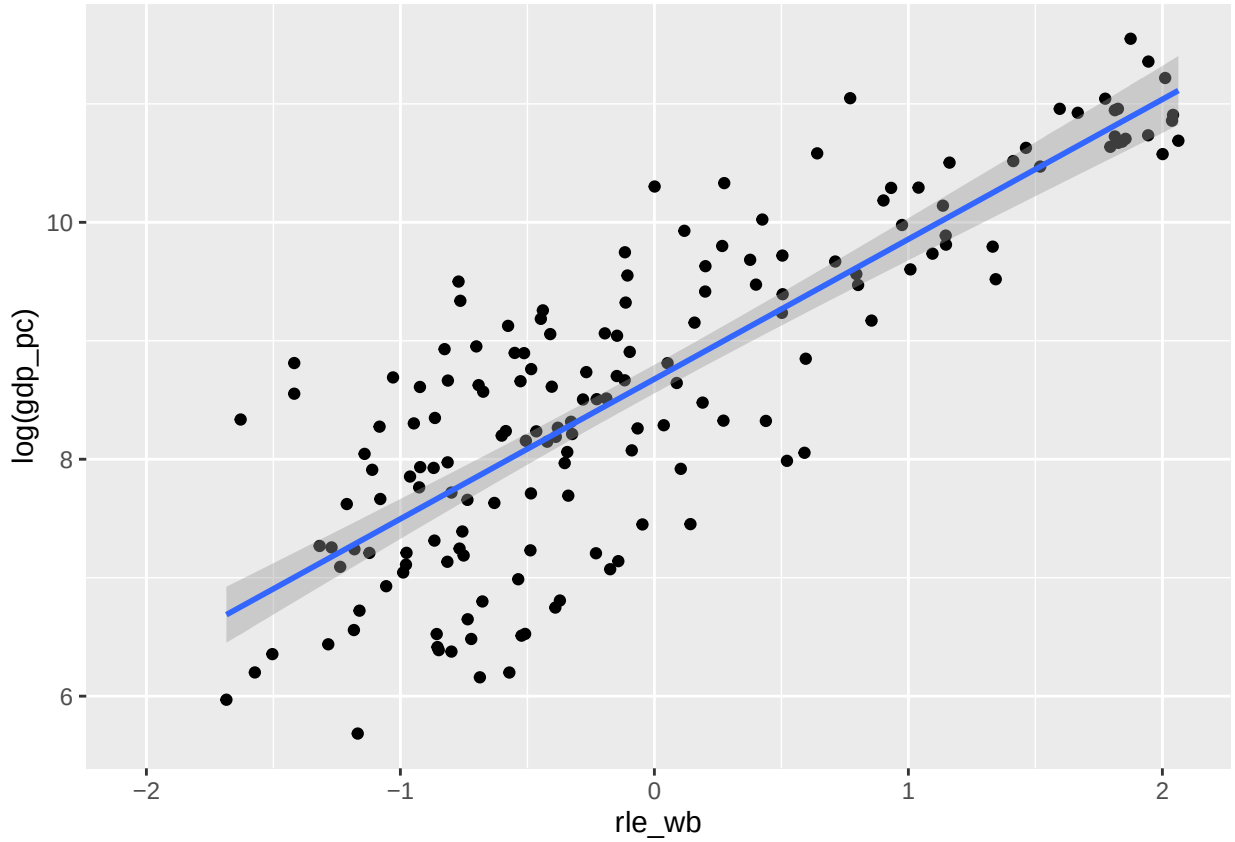
```
##
## Call:
## lm(formula = log(gdp_pc) ~ rle_wb, data = demindex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80324 -0.46467  0.04839  0.49081  1.80926
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.67688    0.06116  141.88   <2e-16 ***
```

```
## rle_wb       1.18063    0.06293   18.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7868 on 164 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6802
## F-statistic:   352 on 1 and 164 DF,  p-value: < 2.2e-16
```

```
ggplot(demindex, aes(rle_wb, log(gdp_pc))) +
  geom_point() +
  geom_smooth(method = 'lm', formula = 'y~x')
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



This model is much more acceptable. The R squared is 0.68, which is higher. Here, the intercept is 8.68, meaning at the lowest value of rle_wb for rule of law log(gdp_pc) is predicted to be at 8.68. Notice that this intercept does not actually appear as a point on the graph. The coefficient of rle_wb is 1.18, meaning for every 1 unit in rle_wb log(gdp_pc) is predicted to increase by 1.18.

Defining Democracy Boolean

While I could not find a hard number for when countries were considered democratic/not in the dem_vdem documentation (they stated that it is more of a spectrum than anything), according to Pew Research (https://

www.pewresearch.org/fact-tank/2019/05/14/more-than-half-of-countries-are-democratic/), as of 2017 57% of countries with pop > 500,000 are democratic. First, I filter for countries with a pop > 500,000.

```
demindex_pop_over500k <- demindex %>%
  filter(pop > 500000)
```

This yields a dataframe of 167 countries (the same number as classified by Pew). Pew places the number of democratic countries at 96. Pulling the 96th country from our modified dataset gives us Guinea with a value of 0.468. Using this as an approximate value, we will define a 'democracy' as any country with a score greater than or equal to 0.46. I like the broader definition of democracy as it allows for edge cases to be included, making the '0' boolean closer to an indicator of authoritarianism than simply 'not western-style democracy'.

```
# Use 0.46 as the cutoff for a democratic country
demindex <- demindex %>%
  mutate(democratic_country = case_when(dem_vdem >= 0.46  ~ 1, TRUE ~ 0))

ROL_econ_wbmodel1 <- lm(log(gdp_pc) ~ democratic_country * rle_wb, data = demindex)
summary(ROL_econ_wbmodel1)
```

```
##
## Call:
## lm(formula = log(gdp_pc) ~ democratic_country * rle_wb, data = demindex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8147 -0.4658  0.0350  0.5002  1.8383
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8.8589     0.1270  69.768   <2e-16 ***
## democratic_country        -0.2635     0.1502  -1.754   0.0813 .
## rle_wb                     1.2882     0.1382   9.324   <2e-16 ***
## democratic_country:rle_wb -0.0772     0.1596  -0.484   0.6293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7841 on 162 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.6882, Adjusted R-squared:  0.6824
## F-statistic: 119.2 on 3 and 162 DF,  p-value: < 2.2e-16
```

Multivar Regression Using democratic_country and rle_wb

Looking at p-values, the interaction between democratic_country and rle_wb is clearly not statistically significant, and should be excluded from the model. The democratic_country boolean also has a p-value greater than 0.05, although it is less large than the interaction democratic_country:rle_wb.

Because democratic_country is a boolean value between 0 or 1, it will modify either the intercept or the slope by a constant. By this I mean: the intercept is predicted to decrease by -.026 when democracy is present, and the slope of the line rle_wb is flattened by -.08 when democracy is present in a country. According to this model, when democracy is present, the impact of a higher rule of law is actually predicted to be

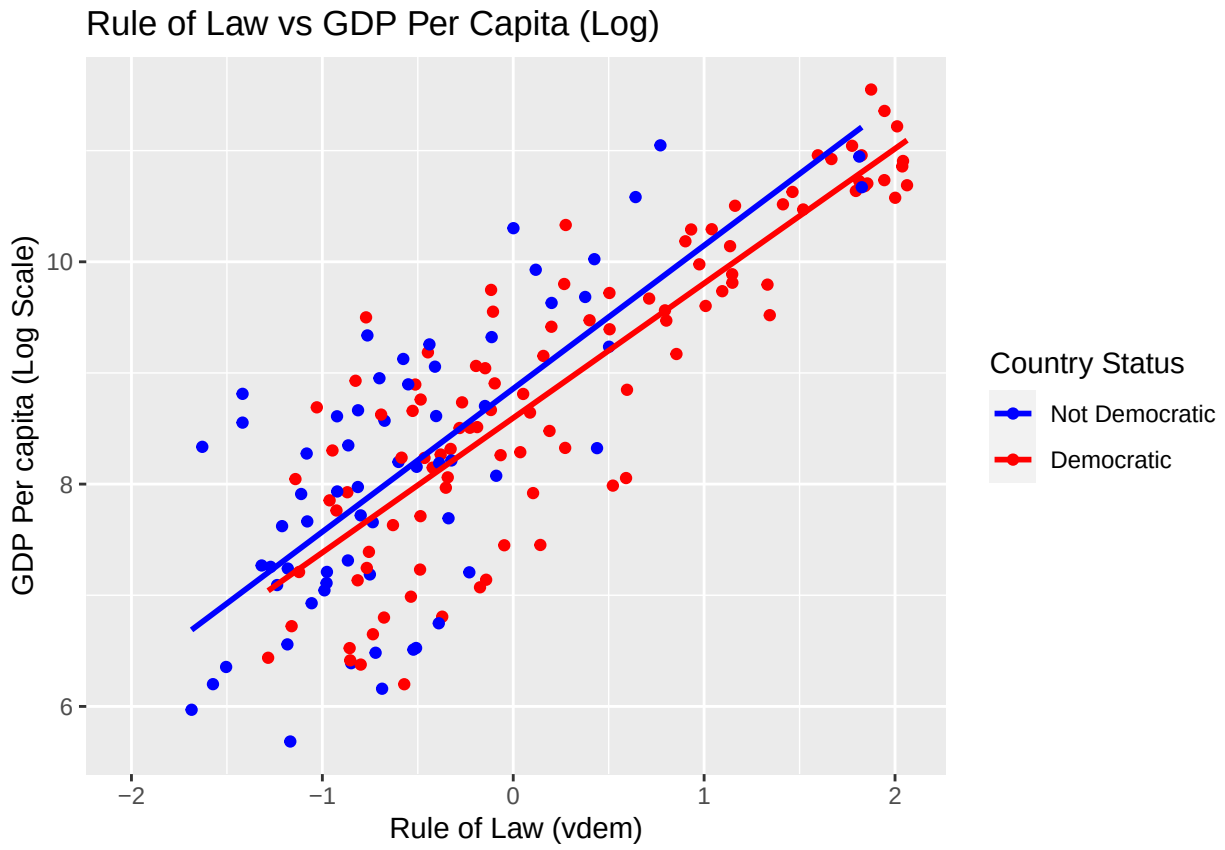
less effective. This may appear counterintuitive, but it is important to remember that this model is largely meaningless given the non-normality of dem_vdem distribution and low p-value.

We can interpret the remaining two coefficients the same basic way as we did previously. The intercept is 8.86 and rle_wb gives the regression line a base slope of 1.29 (when democracy is not present).

```
# This is a second graph, with the interaction between variables included. This means, the slopes will
ggplot(demindex, aes(rle_wb,log(gdp_pc), color = factor(democratic_country))) +
  geom_point() +
  geom_smooth(method = 'lm', formula = 'y ~ x', se = FALSE) +
  labs(color = 'Country Status',   title = "Rule of Law vs GDP Per Capita (Log)",
  x = "Rule of Law (vdem)",
  y = "GDP Per capita (Log Scale)") +
  scale_color_manual(labels = c('Not Democratic', 'Democratic'), values = c('blue', 'red'))
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



Note that, if I were to include confidence intervals on this graph they would be overlapping. This is because the similarity of slopes and intercepts predicted by these regression lines are close enough to not be statistically significant.

What else affects development?

To find what variables might influence gdp p/c, I looked back at my initial exploratory corrplot. Unsurprisingly, 2015 GDP is extremely highly correlated with 1995 GDP. Also unsurprisingly, education and life

expectancy are relatively strongly correlated. To build the most powerful possible model from the data (if that is our goal) while at the same time attempting to not overfit, I would try to include many of these variables in the final model. At the same time, it is also possible that there may be collinearity between some of these variables (I am specifically thinking that education and life expectancy are outcomes of the same type of affluence, and so adding both of them to the model may not provide any new unique information). For these two models, I will choose to use life_exp and educ as additional predictive factors. I will not be interacting these variables, as I don't know of any theoretical framework that specifies a need to do so, and the p-values are not significant.

```
# To be clear, I am not excited about adding dem_vdem to this model. It should not be included :)
finalmodel1 <- lm(log(gdp_pc) ~ rle_wb + dem_vdem + life_exp, data = demindex)

finalmodel2 <- lm(log(gdp_pc) ~ rle_wb + dem_vdem + educ, data = demindex)

summary(finalmodel1)
```

```
##
## Call:
## lm(formula = log(gdp_pc) ~ rle_wb + dem_vdem + life_exp, data = demindex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47351 -0.39276 -0.01946  0.34826  1.62614
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.149270   0.625281   3.437 0.000747 ***
## rle_wb       0.706265   0.078338   9.016 5.31e-16 ***
## dem_vdem    -0.384453   0.238641  -1.611 0.109124
## life_exp     0.093128   0.008332  11.177  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5904 on 162 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.8232, Adjusted R-squared:   0.82
## F-statistic: 251.5 on 3 and 162 DF,  p-value: < 2.2e-16
```

```
summary(finalmodel2)
```

```
##
## Call:
## lm(formula = log(gdp_pc) ~ rle_wb + dem_vdem + educ, data = demindex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45697 -0.37766 -0.01821  0.39984  1.59835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.57463    0.27439  23.961  < 2e-16 ***
## rle_wb       0.62233    0.09216   6.753 4.66e-10 ***
## dem_vdem     0.20513    0.30354   0.676      0.5
```

```
## educ         0.23928    0.02397   9.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5952 on 127 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.8296, Adjusted R-squared:  0.8256
## F-statistic: 206.1 on 3 and 127 DF,  p-value: < 2.2e-16
```

Overall, it appears that both of these variables do help explain development, and have significant p-values when included in the two models. The rle_wb slope for both is positive, indicating that stronger rule of law absolutely does correlate to greater gdp p/c. dem_vdem flips sign between negative and positive, but because of the low p-values this is not really so important to interpret. life_exp is positively correlated with greater log(gdp p/c) (.093), and similarly education is positively correlated (0.24). R squared in both models is actually quite high, sitting at around 0.82, which is a step up from single variable regression including only rle_wb. Looking at the coefficients in front of education and life_exp, it would be possible to say that education has greater predictive power on GDP per capita. However, because the R squared are almost identical, I think it is more fair to say that education and life expectancy have similar predictive power on gdp_pc, and while not entirely colinear, are also somewhat internally correlated.

Finally, both finalmodel1 and finalmodel2 give solid looking QQ and RVF plots, indicating that the models are drawing from normally distributed data, there are no homoscedasticity issues, and our pre-regression assumptions are met.

ColdenJohnson/Development_Index_Predictors