Assignment 03

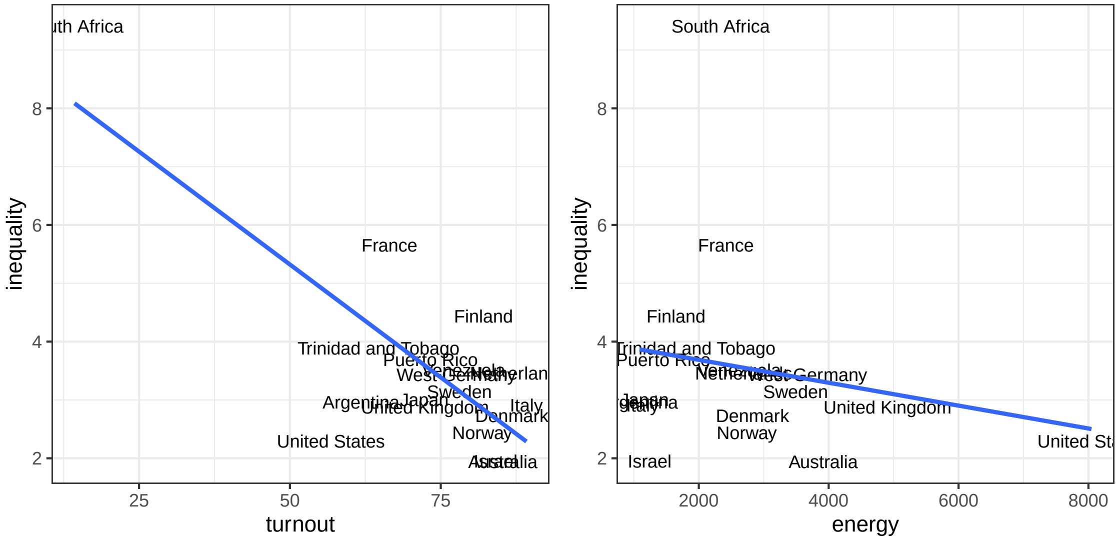
Regression Diagnostics

Answer Key

This assignment is worth 18pts.

Exploratory Analysis

1. Start by creating scatterplots to examine the relationship between each of the predictors and the outcome. Are there observations that look problematic in these plots? If so, identify the country(ies).

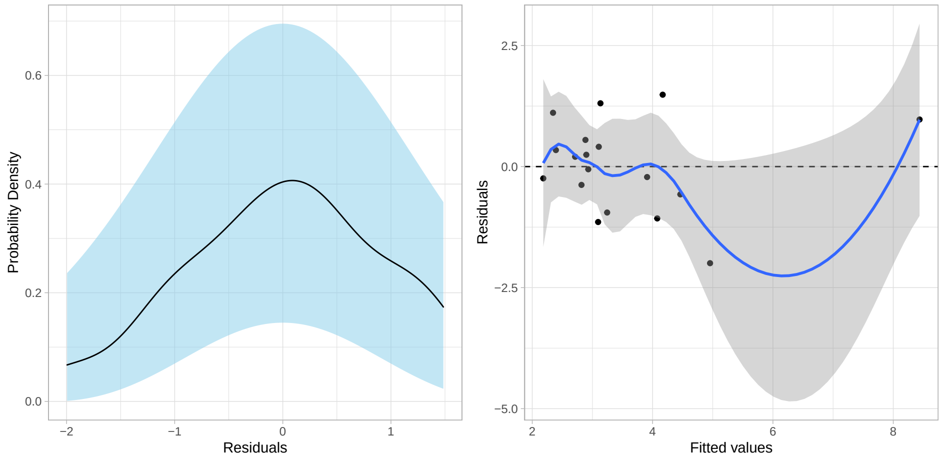


Based on these plots, South Africa seems to have an exceptionally large residual in several of the plots.

2. Fit the regression model specified earlier to the data. Report the fitted equation.

$$\widehat{\text{Inequality}}_i = 10.311 - 0.081(\text{Voter Turnout}_i) - 0.0003(\text{Energy}_i)$$

3. Create and include a set of plots that allow you to examine the assumptions for linear regression. Based on these plots, comment on the tenability of these assumptions.



Based on the QQ-plot, the assumption of normality seems tenable. The assumption that the conditional mean is zero (linearity) seems generally tenable; although the observation with a fitted value near 8 may be problematic. Similarly the homoskedasticity assumption also seems tenable apart from that same observation.

Outliers, Leverage, and Influence

4. Compute the studentized residuals for the observations based on the fitted regression. Based on these values, identify any countries that you would consider as regression outliers. Explain why you identified these countries as regression outliers.

Argentina ($t_i = -2.54$) and South Africa ($t_i = 2.22$) may be regression outliers. Both of these countries have an absolute studentized residual value that suggests their inequality is more than two standard errors from what they would be predicted to have given their predictor values.

5. Fit a mean-shift model that will allow you to test whether the observation with the largest absolute studentized residual is statistically different from zero. Report the coefficient-level output (B, SE, t, and p) for this model.

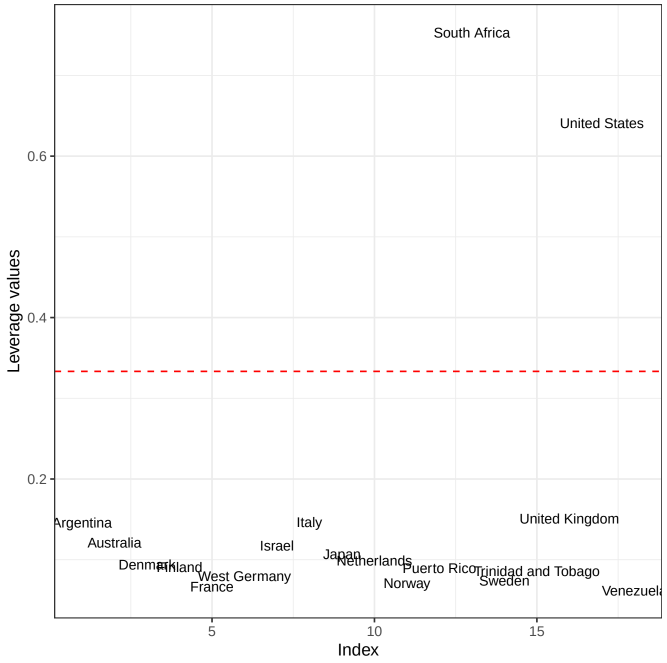
```
## # A tibble: 4 x 5
##   term         estimate std.error statistic      p.value
##   <chr>           <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept) 11.1       1.04         10.7  0.0000000411
## 2 turnout     -0.0875    0.0122       -7.20 0.00000460
## 3 energy      -0.000391  0.000126     -3.10 0.00786
## 4 d_argentina -2.34      0.917        -2.55 0.0231
```

6. Find (and report) the Bonferroni adjusted p-value for the observation with the largest absolute studentized residual. Based on this p-value, is there statistical evidence to call this observation a regression outlier? Explain.

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 1 -2.549854           0.023124      0.41623
```

Based on the Bonferroni-adjusted p-value of 0.416, it is likely that Argentina's studentized residual value is not significantly different than zero. The evidence suggests that Argentina is NOT a regression outlier.

7. Create and include an index plot of the leverage values. Include a line in this plot that displays the cutpoint for "high" leverage. Based on this plot, identify any countries with large leverage values.



There are two countries that have potentially high leverage values: South Africa, and the United States. These countries were identifed because they have leverage values that are more than twice as large as the mean leverage value.

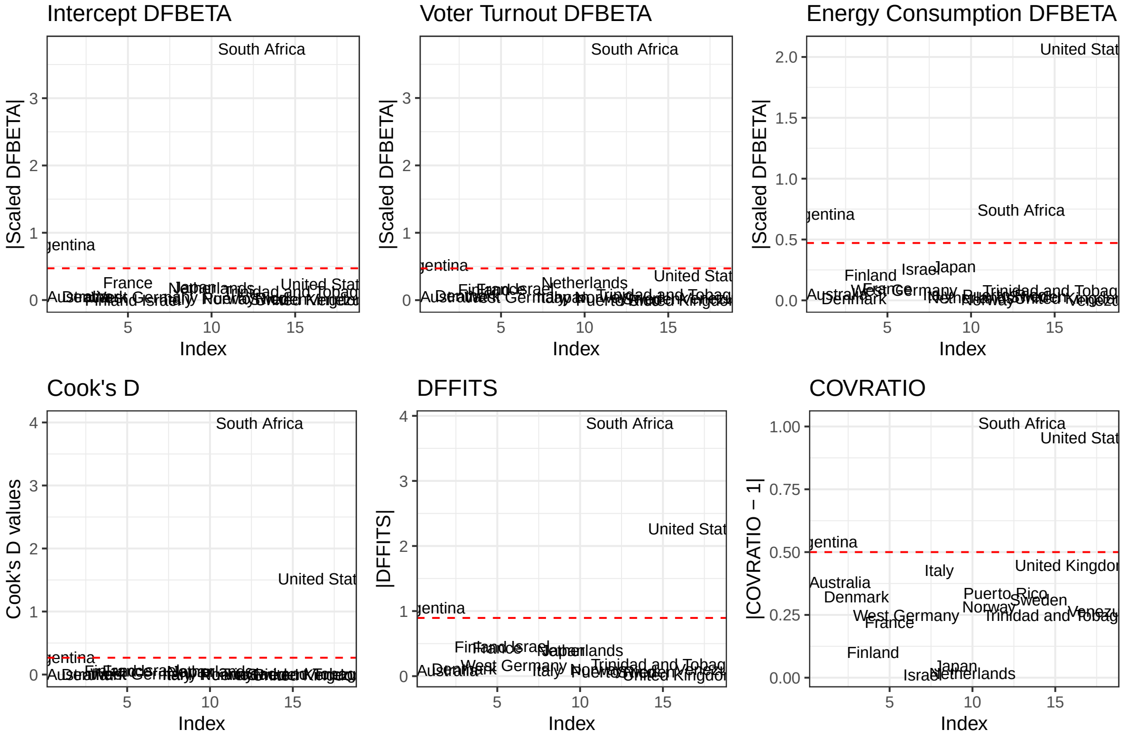
8. Based on the evidence you have looked at in Questions #4–7, do you suspect that any of the countries might influence the regression coefficients? Explain.

Because South Africa was identified as a potential regression outlier (even though the test was non-significant) and as having high leverage, it may be influential.

Influence Measures

9. For each of the influence measures listed below, create and include an index plot of the influence measure. For each plot, also include a line that displays the cutpoint for "high" influence. (3pts)

```
**a. Scaled (standardized) DFBETA values**
**b. Cook's $D$**
**c. DFFITS**
**d. COVRATIO**
```

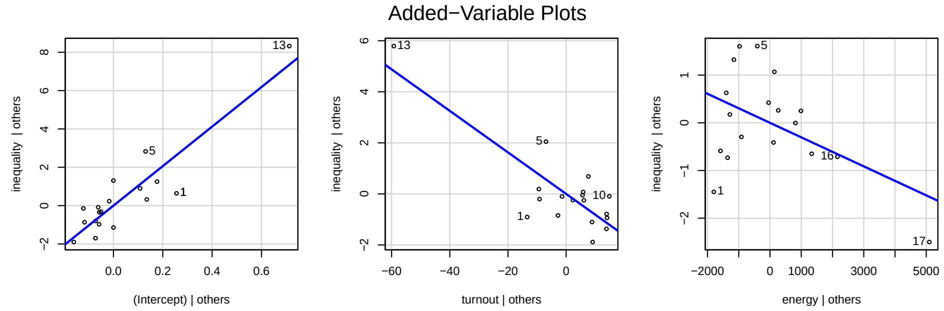


10. Show how the Cook's $D$ value for the country with the largest Cook's $D$ value is calculated using the country's leverage value and studentized residual.

South Africa has the largest Cook's $D$ value ($D_i = 3.99$). To compute this value:

$$
\begin{aligned}
D_i &= \frac{e_i'^2}{k+1} \times \frac{h_{ii}}{1-h_{ii}} \\
&= \frac{1.978^2}{3} \times \frac{0.754}{1-0.754} \\
&= 3.99
\end{aligned}
$$

11. Create and include the added-variable plots for each each of the coefficients. Based on these plots, identify any countries that you believe may be jointly influencing the regression coefficients.



Based on these plots, Observation 1 (Argentina) and 17 (United States) may jointly influence the regression coefficients. Observations 5 (France) and 13 (South Africa) were also identified, but given these plots do not look that problematic.

Remove and Refit

12. Based on all of the evidence from the different influence measures you examined, identify and report the country(ies) that are influential. Explain how you decided on this set of observations.

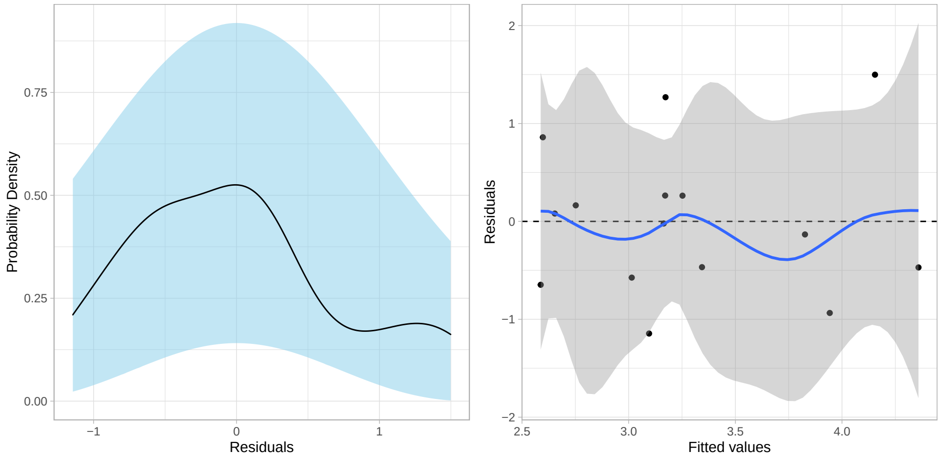
Answers will vary

Based on the evidence, there are three observations I identified as influential: Argentina, the United States, and South Africa. All three of these were identified as influential in the different index plots and also were identified in the added-variable plots.

13. Remove the observations you identified in Question #13 from the data and refit the regression model omitting these observations. Report the fitted equation.

$$\widehat{\text{Inequality}}_i = 9.24 - 0.071(\text{Voter Turnout}_i) - 0.0002(\text{Energy}_i)$$

14. Create and include a set of plots that allow you to examine the assumptions for linear regression. Based on these plots, comment on the tenability of these assumptions.



Based on the QQ-plot, the assumption of normality seems tenable. The assumption that the conditional mean is zero (linearity) and homoskedasticity assumptions also look tenable after removing the three influential observations.

15. Compare and contrast the coefficient-level inferences from the model fitted with the full data and that fitted with the omitted observations.

Table 1

Comparison of p-Values

Term	Full Data	Reduced Data
(Intercept)	0.00	0.00
turnout	0.00	0.03
energy	0.05	0.43

The p-values for the energy consumption effect has become non-significant once the three influential observations were removed.

16. Compare the model-level summaries, namely R^2 and the RMSE, from the model fitted with the full data and that fitted with the omitted observations.

Table 2

Comparison of p-Values

	R^2	RMSE
Full Data	0.71	0.99
Reduced Data	0.35	0.82

The R^2 value and RMSE both diminished once the three influential countries were removed.