Assignment 07

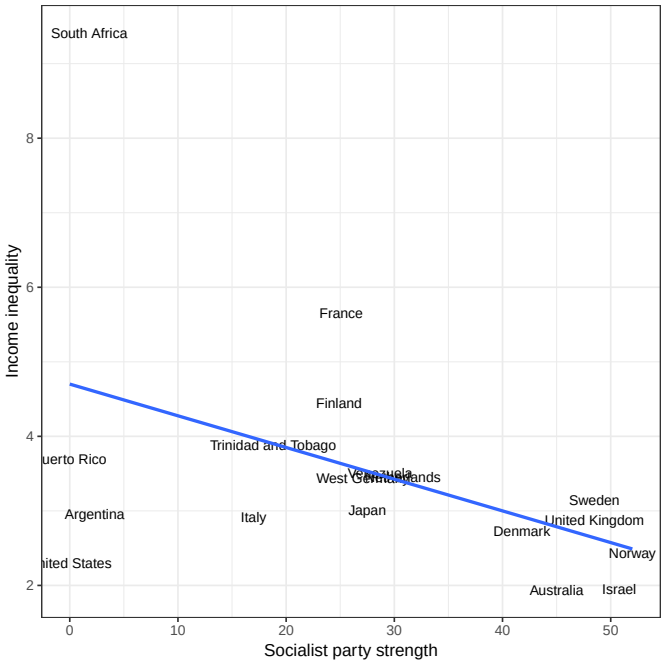
WLS for Addressing Outliers

Answer Key

This assignment is worth 18pts.

Exploratory Analysis

1. Start by creating a scatterplot to examine the relationship between socialist party strength and income inequality (outcome).



2. Are there observations that look problematic in this plot? If so, identify the country(ies).

Based on the plot, South Africa seems to have an exceptionally large residual. France and the United States also seem to have large residuals, but not as big as South Africa's.

3. Fit a linear model regressing income inequality on socialist party strength. Examine and report a set of regression diagnostics that allow you to identify any observations that are regression outliers.

```
## # A tibble: 1 x 3
##   country      studentized .std.resid
##   <chr>              <dbl>      <dbl>
## 1 South Africa        5.81       3.33
```

South Africa is the only country whose absolute studentized residual is greater than 2, suggesting that it is a regression outlier. The Bonferroni-adjusted outlier test also indicates that South Africa is a regression outlier ($p = .0006$).

Weighted Least Squares Estimation

4. Compute the empirical weights that you will use in the WLS estimation. Report the weight for the United States. (Hint: We do not know the true variances in the population.)

```
## # A tibble: 1 x 2
##   country          w_i
##   <chr>          <dbl>
## 1 United States 0.0253
```

5. Fit the WLS model. Report the fitted equation.

$$\hat{\text{Inequality}}_i = 5.22 - 0.060(\text{Socialist}_i)$$

6. Based on the model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?

The coefficient associated with Socialist party strength is negative and statistically significant ($p = .008$). This indicates support for the hypothesis that countries with more socialist tendencies tend to have less income inequality.

7. Create a scatterplot that shows the relationship between socialist party strength and income inequality. Include the country names as labels (or instead of the points). Include both the OLS and WLS regression lines on this plot.

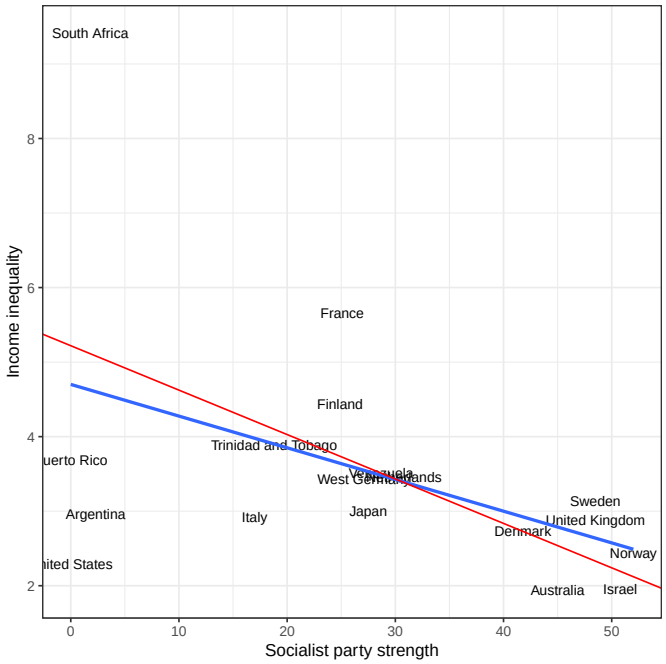


Figure 1: Scatterplot showing income inequality versus Socialist party strength for 18 countries. The OLS (blue) and WLS (red) lines are also presented.

8. Based on the plot, comment on how the residuals from the WLS model compare to the residuals from the OLS model.

The residuals based on the WLS model are very similar to the residuals from the OLS model; the WLS line is not that different from the OLS line. The biggest difference is that the residual for the regression outlier (South Africa) is smaller in the WLS model.

9. Based on your response to Question #8, how will the model-level R^2 value from the WLS model compare to the model-level R^2 from the OLS model. Explain.

Since the residuals are comparable, the R^2 values should also be quite similar.

10. The mathematical formulaa for computing the studentized residuals for both the OLS and WLS models is given below. Compute and report the studentized residuals, using this formula, from both the OLS and WLS models for any regression outliers you identified in Question #2. (Hint: Remember that in an OLS regression the weight is 1 for each observation.)

$$e_i' = \frac{e_i}{s_{e(-i)}\sqrt{1-h_{ii}}} \times \sqrt{w_i}$$

```
# OLS studentized residual for South Africa (Observation 13)
e = augment(lm.1)$.resid[13]
s = augment(lm.1)$.sigma[13]
h = augment(lm.1)$.hat[13]
w = 1
e / (s * sqrt(1 - h)) * sqrt(w)
```

```
## [1] 5.811155
```

```
# WLS studentized residual for South Africa (Observation 13)
e = augment(wls_1)$.resid[13]
s = augment(wls_1)$.sigma[13]
h = augment(wls_1)$.hat[13]
w = w_i[13]
e / (s * sqrt(1 - h)) * sqrt(w)
```

```
##       13
## 1.371407
```

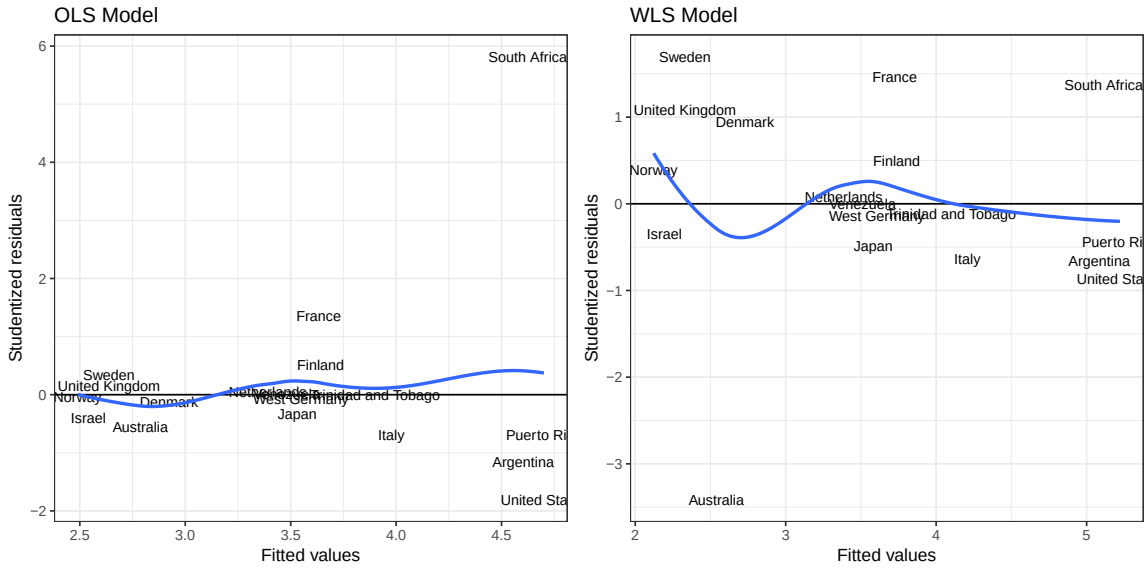
11. Based on the values of the studentized residuals in the WLS model, are the observations you identified as regression outliers from the OLS model still regression outliers in the WLS model? Why or why not?

No. In the WLS model, South Africa's studentized residual is only 0.160. This is smaller than the cutoff of two.

12. Explain why the is the case by referring to the formula.

The residual values, deleted residual error term, and the leverage values for South Africa are slightly different across the two models, but the differences aren't that great. The big difference is the weight computation. This reduces the OLS studentized residual value by approximately 83%.

13. Create and report residual plots of the studentized residuals versus the fitted values for the OLS and WLS models. Comment on which model better fits the assumptions.



The WLS model seems to better meet the assumptions. Linearity seems tenable; while in the OLS model the loess line trends up for higher fitted values potentially indicating some violation of the linearity assumption. Homoskedasticity also seems more tenable in the WLS model than in the OLS model. While Australia looks to have a large residual, it isn't beyond extreme (like South Africa was in the OLS model). Furthermore, the OLS model looks to exhibit a higher variance for the highest fitted values, in general.

Including Covariates

14. Use matrix algebra to compute the empirical weights based on the two-predictor model and report the weight for the United States.

```
# Obtain weights
X = stack %>% mutate(b0 = 1) %>% select(b0, energy, socialist) %>% data.matrix()
Y = stack$inequality

# Obtain weights
B = solve(t(X) %*% X) %*% t(X) %*% Y
e_sq = (Y - X %*% B) ^ 2
b = solve(t(X) %*% X) %*% t(X) %*% e_sq
w_i = 1 / ( (X %*% b) ^ 2)

# Get weight for United States
w_i[17]
```

```
## [1] 0.04408061
```

** Fit the two-predictor WLS model using matrix algebra. Report the fitted equation.**

```
# Matrix algebra
W = as.numeric(w_i) * diag(18)

B = solve(t(X) %*% W %*% X) %*% t(X) %*% W %*% Y
B
```

```
##                    [,1]
## b0         5.269921e+00
## energy    -6.972714e-05
## socialist -5.662168e-02
```

$$\hat{\text{Inequality}}_i = 5.27 - 0.00007(\text{Energy}_i) - 0.057(\text{Socialist}_i)$$

16. Compute and report the standard errors of the two-predictor WLS model using matrix algebra.

```
e_i = Y - X %*% B
mse = sum( w_i * (e_i ^ 2) ) / 15
SE = sqrt(diag(mse * solve(t(X) %*% W %*% X)))
SE
```

```
##           b0       energy    socialist
## 0.8764045692 0.0001415291 0.0188582464
```

17. Using your results from Questions #14 and #15, compute and report the t-values and p-values. Show your work or syntax. While you can use the output of the `tidy()`, `summary()`, or other functions that automatically compute p-values to check your work, you can not use them to answer this question. (Hint: Use the `pt()` function.)

```
data.frame(
  Term = c("Intercept", "Energy", "Socialist"),
  B = B,
  SE = SE,
  t = B / SE
) %>%
  mutate( p = 2 * pt(-abs(t), 15) )
```

```
##        Term     B     SE      t       p
## 1 Intercept  5.27 0.8764  6.013 0.00002
## 2    Energy  0.00 0.0001 -0.493 0.62938
## 3 Socialist -0.06 0.0189 -3.002 0.00893
```

18. Based on the two-predictor WLS model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?

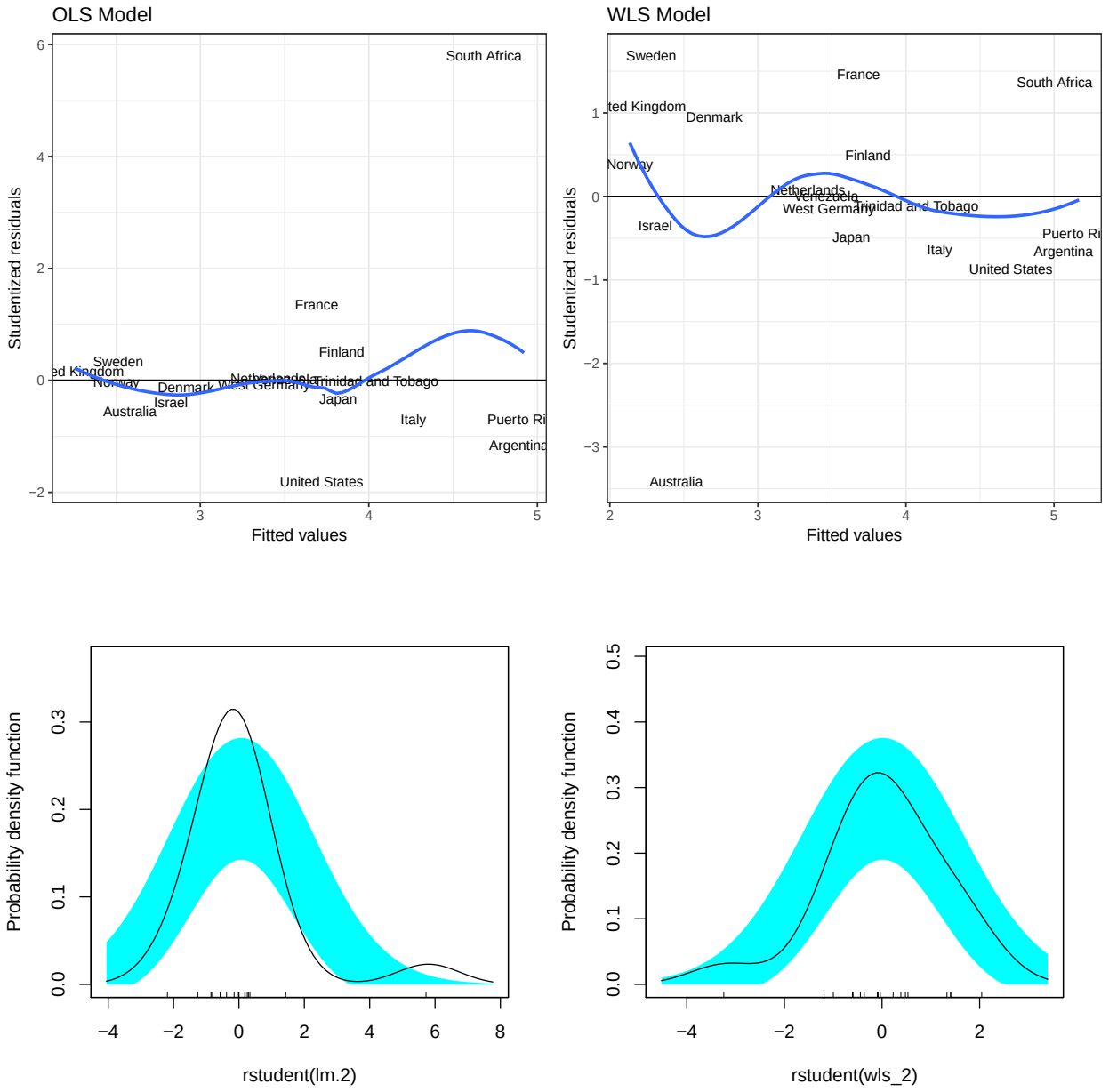
The coefficient associated with Socialist party strength is negative and statistically significant ($p = .009$), even after controlling for differences in economic development. This indicates support for the hypothesis that countries with more socialist tendencies tend to have less income inequality.

19. Based on the two-predictor OLS model results, what is suggested about the research hypothesis that countries with more socialist tendencies have less income inequality?

```
## # A tibble: 3 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  5.19      0.922        5.63  0.0000482
## 2 energy      -0.000182  0.000227    -0.801 0.436
## 3 socialist   -0.0420    0.0215      -1.96  0.0693
```

The coefficient associated with Socialist party strength is negative, however it is not statistically significant ($p = .069$), after controlling for differences in economic development. This is evidence against the hypothesis that countries with more socialist tendencies tend to have less income inequality.

20. Which set of the model results should we trust. Explain by referring to the tenability of the assumptions.



Since the WLS model better meets the assumptions, we should base our interpretations around those results.