Problem Set 7

Nicholas R. Gonzalez

Problem 1

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(haven)
tabellini <- read_dta("pset_7/tabellini.dta")
# View(tabellini)
library(car)
```

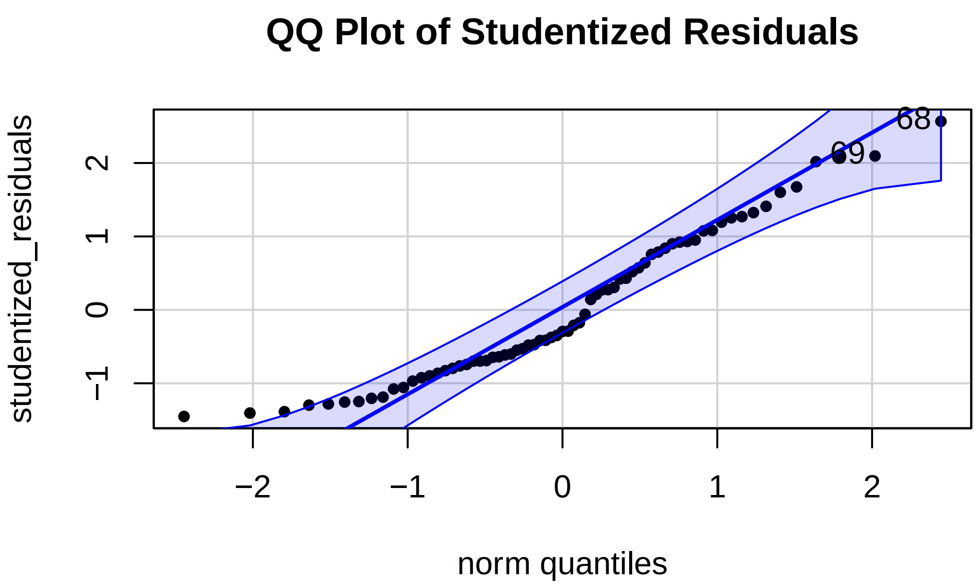
```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':
```

some

a

i

```r
library(car)
# install.packages('car')
library(ggplot2)

model <- lm(rgdph ~ polityIV, data = tabellini)

studentized_residuals <- rstudent(model)

qqPlot(studentized_residuals, main = "QQ Plot of Studentized Residuals", pch = 20)
```




```
[1] 68 69
```

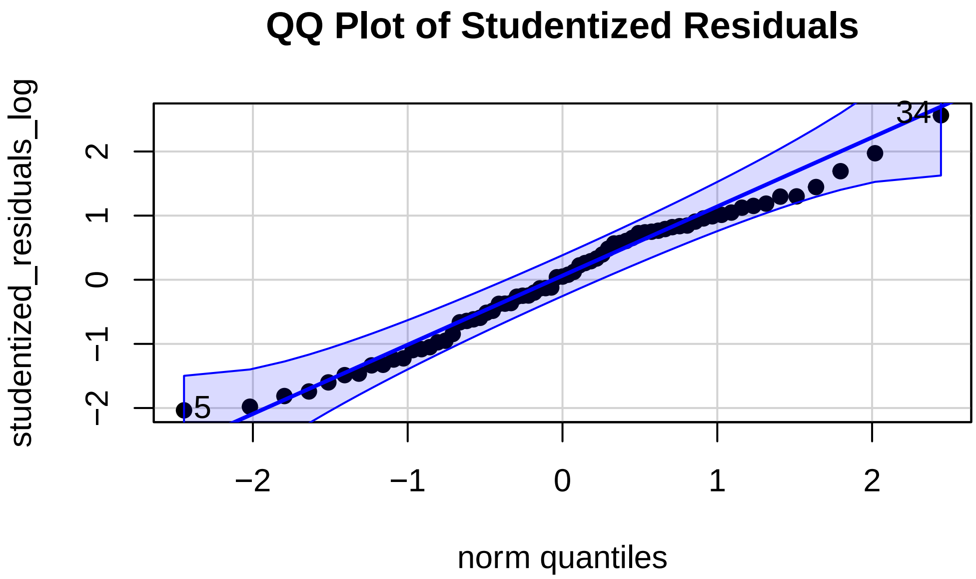
```r
summary(model)
```

```
Call:
lm(formula = rgdph ~ polityIV, data = tabellini)

Residuals:
   Min     1Q Median     3Q    Max 
 -6637  -3534  -1348   3867  11337 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -698.1     1346.4  -0.518    0.606    
polityIV      1014.4      166.5   6.094 6.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4651 on 67 degrees of freedom
Multiple R-squared:  0.3566,    Adjusted R-squared:  0.347 
F-statistic: 37.14 on 1 and 67 DF,  p-value: 6.086e-08
```

ii

In our log of GDP per capita, the QQ plot is better. The data is generally more fitted on our regression line. The following of the line better in plot two, suggest as well that the data has more of a normal shape as opposed to the plot in (a) (i).

```r
tabellini$log_gdp <- log(tabellini$rgdph)

log_model <- lm(log_gdp ~ polityIV, data = tabellini)

studentized_residuals_log <- rstudent(log_model)

qqPlot(studentized_residuals_log, main = "QQ Plot of Studentized Residuals", pch = 19)
```




```
[1] 34  5
```

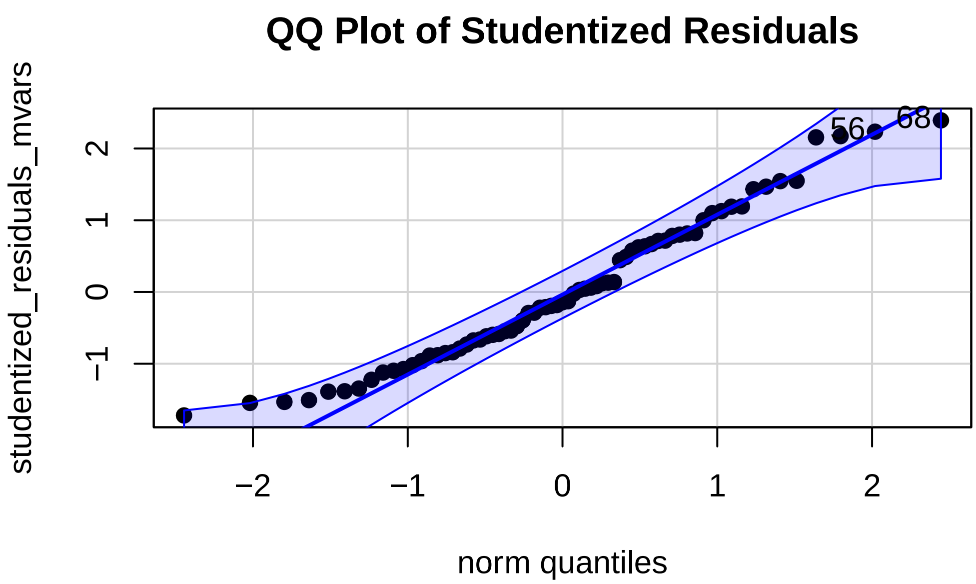
```
summary(log_model)
```

```
Call:
lm(formula = log_gdp ~ polityIV, data = tabellini)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38861 -0.46293  0.03467  0.55127  1.50815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.84420    0.20359  33.618  < 2e-16 ***
polityIV     0.21010    0.02517   8.347  5.7e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7033 on 67 degrees of freedom
Multiple R-squared:  0.5098,    Adjusted R-squared:  0.5025
F-statistic: 69.67 on 1 and 67 DF,  p-value: 5.701e-12
```

iii

When non-normality occurs, I think for this dataset, and in general the topics I am interested in it would be better to add relevant predictors. As I usually am interested in race, class and political economy, adding more predictors such as demographic data, or economic data, in theory would help my models. Which is the case for this dataset, and model, as we add the polity, gini and trade variables, we get better results.

```r
model_mvars <- lm(rgdph ~ polityIV + gini_8090 + trade, data = tabellini)

studentized_residuals_mvars <- rstudent(model_mvars)

qqPlot(studentized_residuals_mvars, main = "QQ Plot of Studentized Residuals", pch = 19)
```




```
[1] 68 56
```

```r
summary(model_mvars)
```

```
Call:
lm(formula = rgdph ~ polityIV + gini_8090 + trade, data = tabellini)

Residuals:
    Min      1Q  Median      3Q     Max
```

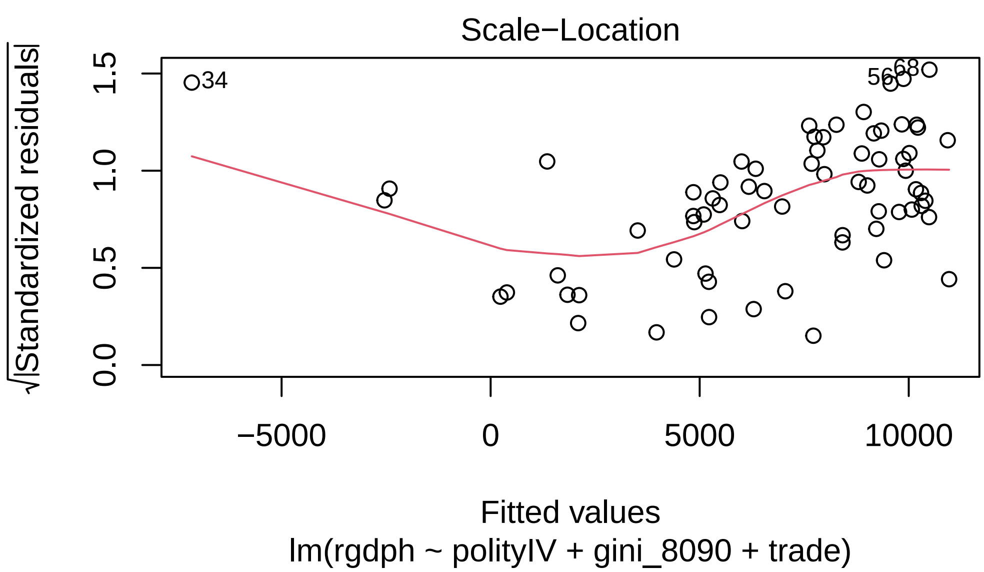
```
-7491.1 -3527.4  -588.8  3188.8 10287.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5802.284   3311.947   1.752   0.0845 .  
polityIV     871.682    174.719   4.989 4.8e-06 ***
gini_8090   -128.241     56.734  -2.260   0.0272 *  
trade         -6.252     15.620  -0.400   0.6903    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4547 on 65 degrees of freedom
Multiple R-squared:  0.4035,	Adjusted R-squared:  0.376 
F-statistic: 14.66 on 3 and 65 DF,  p-value: 2.147e-07
```

b

i

The errors of our data appear to be homoskedastic. This is because the red line increases at higher values, which tells our residual variance grows as predicted values increases. We also have some outliers, specifically 34, 56 and 58.

```r
plot(model_mvars, which = 3)
```



ii

We can use the Breusch-Pagan test to check for heteroskedasticity. It does so by testing if the variance of the residuals relies on the independent variable(s).

The results of our BP text tell us that there is heteroskedasticity, but not a whole lot. This is because we only marginally reject the null (homoskedasticty), based on our p-value.

```r
# install.packages('lmtest')
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
bptest(model_mvars)
```

```

	studentized Breusch-Pagan test

data:  model_mvars
BP = 8.0464, df = 3, p-value = 0.04506
```

iii

The coefficients are the same, but the standard errors are different. This might happen because the data might sensitive to influence points, or their might be strong outliers within our data. If this were true, latter would have higher leverage and disproportionately effect our model. The standard error then would be different for our robust standard errors because robust standard errors are designed to help deal with that issue.

```r
library(modelsummary)
```

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```r
library(sandwich)
library(lmtest)

hc_se <- vcovHC(model_mvars, type = "HC3")

robust_test <- coeftest(model_mvars, vcov. = hc_se)

modelsummary(list("OLS" = model_mvars, "HC3 Robust SE" = robust_test),
             gof_omit = "R2|AIC|BIC|Log.Lik",
             coef_map = c("polityIV" = "Polity Score",
                          "gini_8090" = "Gini Index",
                          "trade" = "Trade"),
             title = "Regression Results: GDP per Capita")
```

c

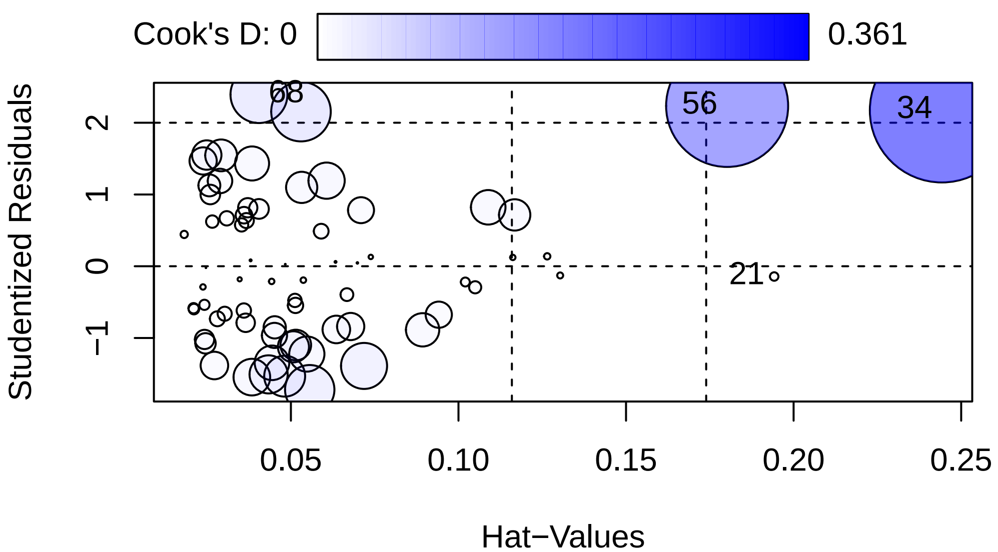
i

The observations that stand out are " "Zimbabwe" & "Luxembourg". When looking at the data we can notice that Luxembourg and Zimbabwe are both on the far ends of the polity score.

```r
influencePlot(model_mvars)
```

Table 1: Regression Results: GDP per Capita

	OLS	HC3 Robust SE
Polity Score	871.682	871.682
	(174.719)	(219.733)
Gini Index	−128.241	−128.241
	(56.734)	(48.986)
Trade	−6.252	−6.252
	(15.620)	(19.604)
Num.Obs.	69	69
F	14.659	
RMSE	4412.86	




```
      StudRes         Hat       CookD
21 -0.1431775 0.19420717 0.001254082
34  2.1722042 0.24425180 0.360613373
56  2.2339894 0.18016541 0.258327215
68  2.3923617 0.04046796 0.056257496
```

```
model_mvars$model # confirming no observations have dropped
```

```
        rgdph   polityIV gini_8090     trade
```

1	3582.3389	10.0000000	28.2600	107.99532
2	4986.0034	10.0000000	24.3800	73.65478
3	2426.2827	9.3333330	41.7900	122.17721
4	1431.4984	8.4444447	32.0000	21.88374
5	893.3627	6.3750000	25.7100	70.71071
6	1451.7408	7.7777781	31.7350	38.71352
7	4952.1670	7.1666670	19.4900	119.40155
8	2543.5754	6.0000000	35.8200	153.66084
9	5217.1069	8.0000000	24.7650	94.51885
10	1794.2019	9.0000000	42.0000	48.88424
11	4330.1152	8.1111107	25.5550	48.27744
12	1704.8525	6.0000000	22.9000	55.68492
13	1759.4392	7.8888888	45.5000	80.63206
14	2858.5952	8.8888893	43.0000	57.26163
15	1322.4855	6.8888888	50.0000	87.20344
16	3697.9712	10.0000000	46.7800	85.90728
17	530.2225	0.0000000	62.0000	62.48260
18	6666.7651	10.0000000	41.3500	128.11646
19	1395.1619	6.0000000	54.5000	84.09262
20	2202.6362	5.7777781	42.5000	86.43324
21	6459.1099	10.0000000	49.6750	176.05992
22	2480.1443	9.0000000	54.0000	87.46591
23	1147.1674	5.0000000	30.0000	49.82251
24	2019.5516	6.8888888	48.0000	54.02488
25	7001.2891	10.0000000	35.0000	41.74717
26	2396.8171	6.4444451	46.0000	67.85826
27	10102.7666	10.0000000	25.5000	45.26283
28	659.1872	2.6666670	47.2550	73.76629
29	2441.6929	5.0000000	37.5500	76.01492
30	8104.7588	10.0000000	36.5000	68.45359
31	3995.7478	8.1111107	44.0000	41.00135
32	4093.8086	5.0000000	42.5000	118.66697
33	1611.7850	4.7777781	33.6350	25.35251
34	1201.9840	-6.0000000	56.8300	69.16644
35	4400.0298	7.2222219	49.5000	84.86580
36	1116.2789	-1.0000000	54.1200	65.39255
37	13729.2109	10.0000000	26.5000	132.14191
38	11390.4893	10.0000000	35.0000	128.91702
39	2298.2178	4.6666670	58.5300	43.20914
40	3587.2847	8.0000000	51.0000	34.79187
41	7616.8940	9.0000000	42.0000	86.71415
42	763.5707	0.2222222	39.0000	121.57738
43	3117.3628	7.8571429	62.3000	43.85324

```
44 13820.4297 10.0000000   29.0000 100.33990
45 13313.1348 10.0000000   29.0000  78.26205
46 13072.6719 10.0000000   28.4859  60.83865
47 10428.7412  9.0000000   37.0000  77.46720
48  5491.7866  8.0000000   56.0000  59.19077
49  6660.6182  8.2222223   49.5000  52.77753
50  2410.2461  2.1111109   44.0000  25.88982
51  4185.5527  8.0000000   58.6900  17.56221
52  1569.1210  0.8750000   29.0000 116.52354
53 13586.1787 10.0000000   29.5000  54.16156
54 12929.4160 10.0000000   32.5000  44.32747
55 12088.6270 10.0000000   38.0000  58.07035
56 18801.7090 10.0000000   27.0000 188.98882
57 14414.4385 10.0000000   32.4800  68.05694
58 10115.8965  5.7777781   29.0350  90.74010
59  6207.8105  2.6666670   50.5000  47.54999
60   980.6061 -0.6250000   35.3600  53.48115
61 14917.8770 10.0000000   33.0000  66.56440
62 17040.5840 10.0000000   30.5000  66.52283
63 14185.7725  9.0000000   35.0000  44.88788
64 16720.4785 10.0000000   33.0000  72.23154
65 15878.5566 10.0000000   37.0000  68.89883
66 15255.6289 10.0000000   34.2000  18.75606
67 15499.7227 10.0000000   39.8600  38.79273
68 20782.8145 10.0000000   29.0000  48.75680
69 18844.2773 10.0000000   37.5000  23.03455
```

```r
tabellini$country[c(34, 56)]
```

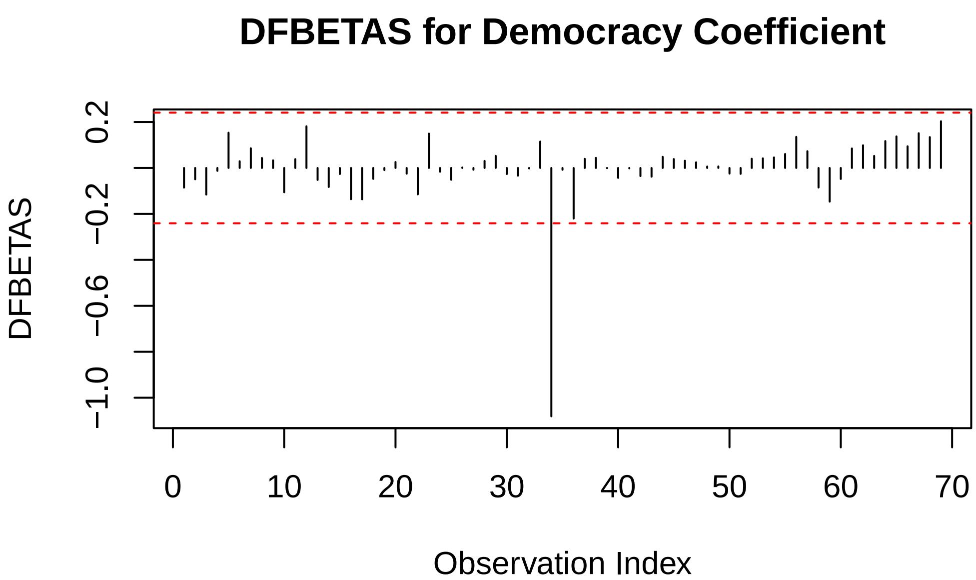
```
[1] "Zimbabwe"   "Luxembourg"
```

ii

Once again, Zimbabwe stands out. This tells us that it has a significant impact on the democracy coefficient (polityIV) and may very large outlier. This is confirmed by the data as well as it is 6x lower than the next lowest polity score.

I also made a model to examine the coefficients without Zimbabwe. We can see removing Zimbabwe drastically changes the results of the polity coefficient.

```
dfbetas_values <- dfbetas(model_mvars)

dfbetas_democracy <- dfbetas_values[, "polityIV"]

plot(dfbetas_democracy, type = "h", main = "DFBETAS for Democracy Coefficient",
     ylab = "DFBETAS", xlab = "Observation Index")
abline(h = c(-2/sqrt(length(dfbetas_democracy)), 2/sqrt(length(dfbetas_democracy))), col = "
```




```
influential <- which(abs(dfbetas_democracy) > 2/sqrt(length(dfbetas_democracy)))
influential
```

```
34
34
```

```
model_no_zim <- lm(rgdph ~ polityIV + gini_8090 + trade, data = tabellini[-34, ])

modelsummary(list("OLS" = model_mvars, "No Zimbabwe" = model_no_zim),
             gof_omit = "R2|AIC|BIC|Log.Lik",
             coef_map = c("polityIV" = "Polity Score",
                          "gini_8090" = "Gini Index",
                          "trade" = "Trade"),
             title = "Removing Zimbabwe")
```

Table 2: Removing Zimbabwe

	OLS	No Zimbabwe
Polity Score	871.682	1055.379
	(174.719)	(189.807)
Gini Index	−128.241	−133.093
	(56.734)	(55.223)
Trade	−6.252	−6.684
	(15.620)	(15.193)
Num.Obs.	69	68
F	14.659	16.535
RMSE	4412.86	4289.87

iii

As stated in the previous question, this issue occurs because how extreme Zimbabwe is. We could remove the observation from our dataset, or model, as I did in the previous question. We could also use robustness checks. For example robust standard errors. This is why our standard errors are different in the HC3 model in one of the previous questions.