Replication: Two Linear Regression Models

Simu Huang, Xijia Luo, Minqi Li, Jiachen Feng

1 Checking the Assumptions

Our client plans to use two simple linear regression to see the relationship between Mobility and GNI&FSI. For the simple linear regression model works appropriately, the datasets must satisfy the following conditions.

First of all, all the response variables and the explanatory variable should be continuous.

Varible	Range
Mobility	[25,189]
FSI	[17.9,113.4]
GNI	[6.348e+07,1.286e+14]
ln(GNI)	[17.966,32.488]

From the table above, we can conclude that all these variables we are going to use in the model are continuous.

Next, we need to ensure that the outcome variable Y has a roughly linear relationship with the explanatory variable X. More simply, make a plot based on the dataset we used, then check with eyes if they seem to be linear. We have put the codes in the Appendix. For the Mobility~Log(GNI) plot, we found a weak positive relation. And for the Mobility~FSI plot, we found a strong negative linear relation. Thus, both of them satisfied the SLM conditions.

Lastly, check the independence assumption.

Mobility Score: It can be understood as the ability for citizens holding this passport access to pre-authorized cross-border movement. For each country, we count the number of countries that it has direct access with visa free. That is saying, we do not take dependence pattern across countries and diplomatic reciprocity into account. Therefore, the countries are all independence in this case.

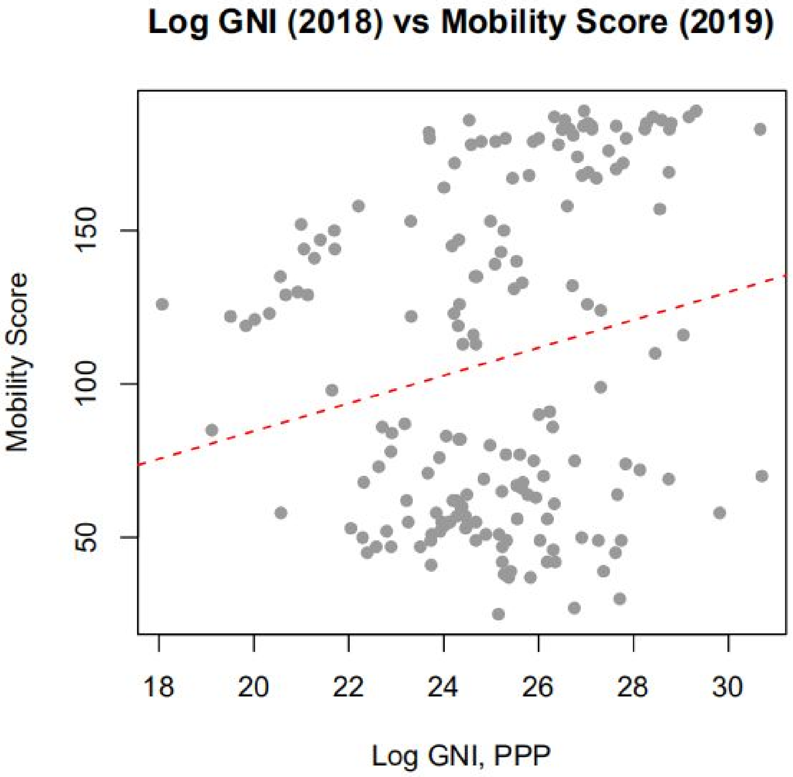
FSI: Fragile States Index is calculated from the situation within the country, External countries will not be able to affect most indicators. For the sake of preciseness, our group considers that the dependent indicators of FSI index have “External Intervention” and

"Human Flight and Brain Drain". Since there are 12 indicators in all and each worth 10 points. As a result, we can guarantee 100/120= 83% independence of the dataset which we think is enough to do the simple linear regression model.

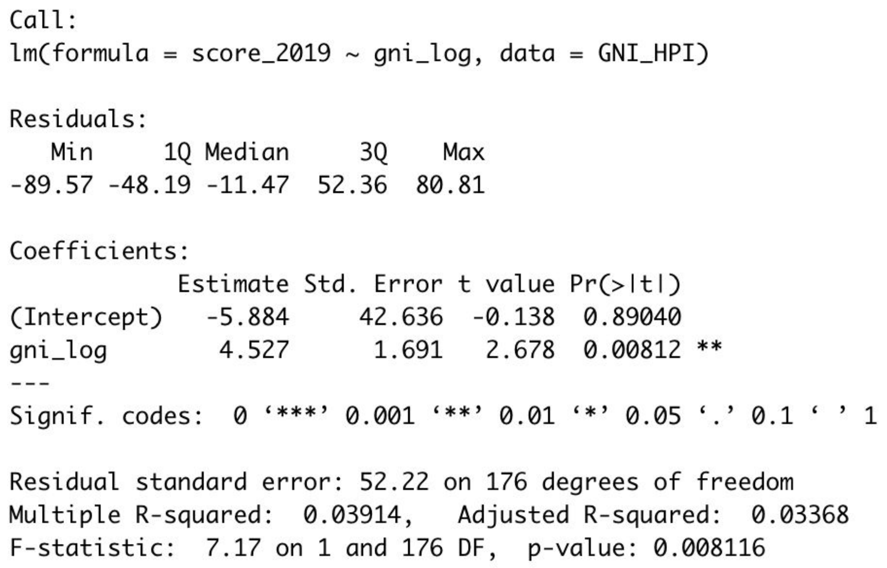
GNI: This index means the per capita gross national income. The per capita income of each country is independent because the per capita income of foreigners does not affect domestic data.

2 Model

2.1 Log GNI (2018) vs Mobility Score (2019)

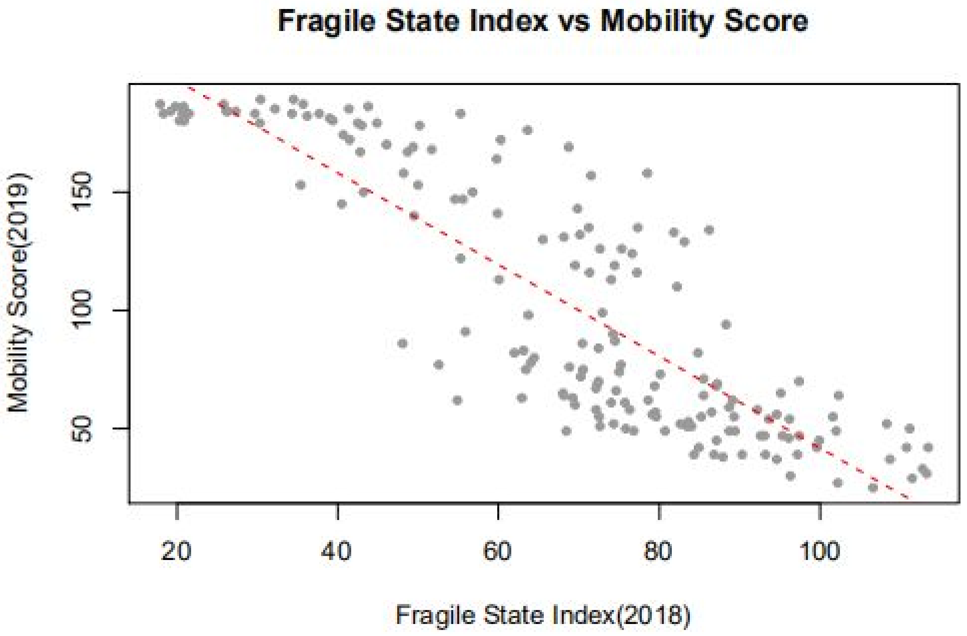


Result:

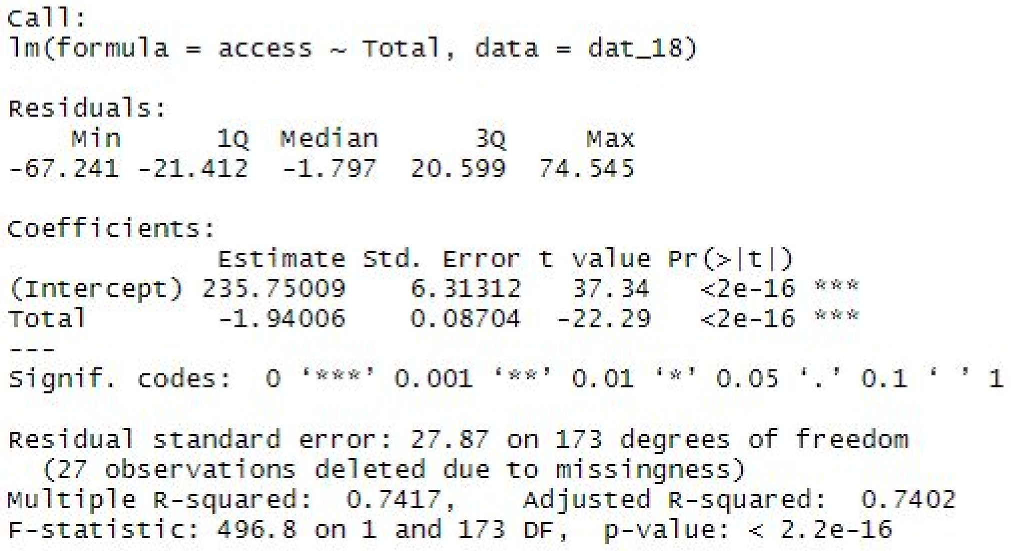


```
Call:
lm(formula = score_2019 ~ gni_log, data = GNI_HPI)

Residuals:
   Min     1Q Median     3Q    Max 
-89.57 -48.19 -11.47  52.36  80.81 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -5.884     42.636  -0.138  0.89040   
gni_log        4.527      1.691   2.678  0.00812 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.22 on 176 degrees of freedom
Multiple R-squared:  0.03914,	Adjusted R-squared:  0.03368 
F-statistic:  7.17 on 1 and 176 DF,  p-value: 0.008116
```

2.2 Fragile State Index vs Mobility Score



Result:



```
Call:
lm(formula = access ~ Total, data = dat_18)

Residuals:
    Min      1Q  Median      3Q     Max 
-67.241 -21.412  -1.797  20.599  74.545 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 235.75009    6.31312   37.34   <2e-16 ***
Total        -1.94006    0.08704  -22.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.87 on 173 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.7417,	Adjusted R-squared:  0.7402 
F-statistic: 496.8 on 1 and 173 DF,  p-value: < 2.2e-16
```

3 Interpretation

3.1 Log GNI (2018) vs Mobility Score (2019)

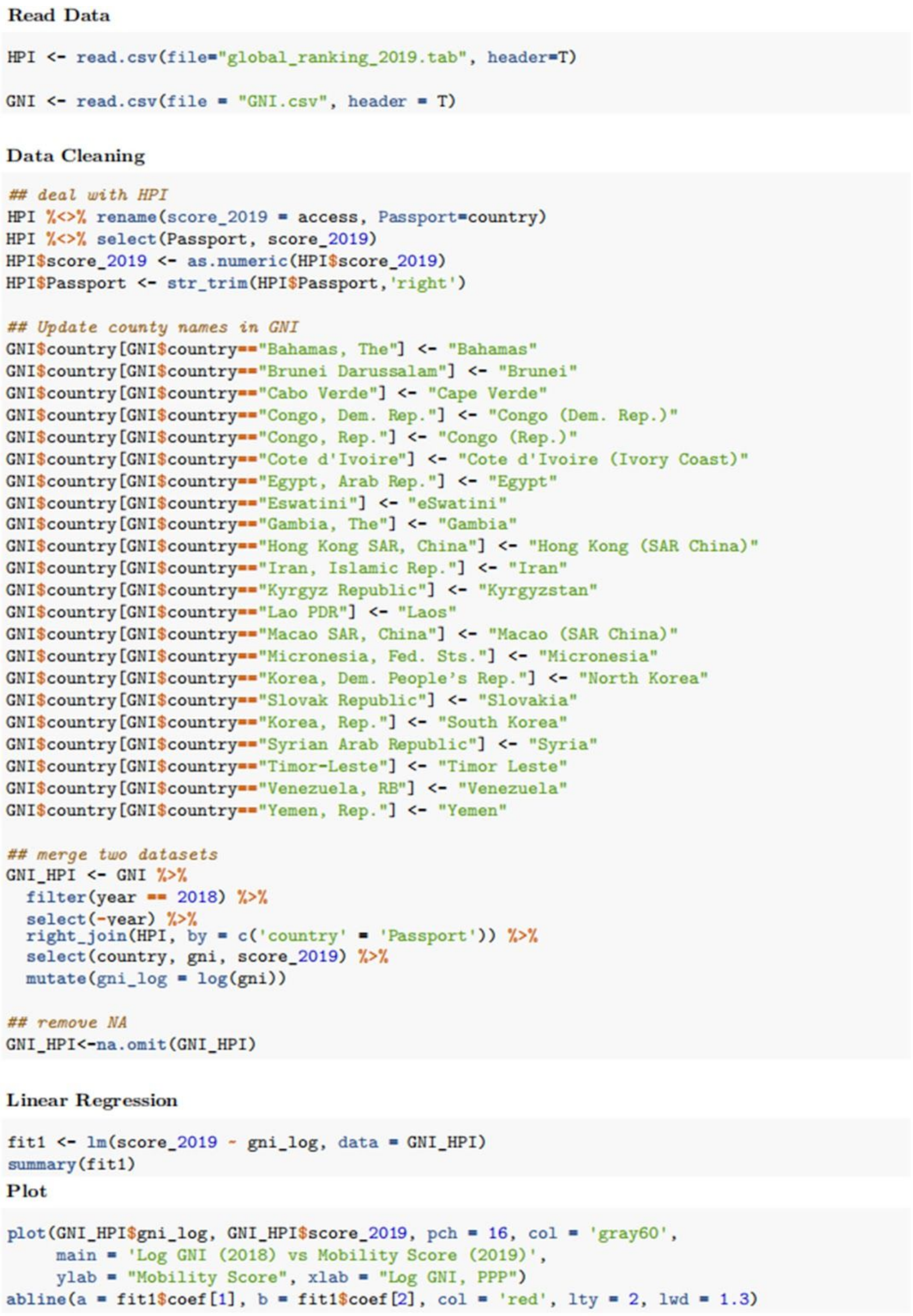
- We create gni_log to take the log of the gni coefficient because the original data are too large for model fitting; the score_2019 represents the mobility score of this country in 2019.
- The Intercept -5.884 implies that when the gni_log equals to zero, which means GNI is 1, then the mobility score in the 2019 will be -5.884
- The coefficient 4.527 implies that with every 1 unit increase in the log of the gni coefficients, there will be a 4.527 unit increase in the mobility score in 2019.

3.2 Fragile State Index vs Mobility Score

- The Total variable represents the fsi score of this country.
- The intercept 235.75 implies that when the Total variable equals 0, the access number would be 236.
- The coefficient -1.94 implies that with every 1 unit increase in the Total variable, the mobility score would decrease by two units in the number of access, in other words, the number of access will decrease by 2.

4 Appendix

4.1 Mobility Score ~ Log (GNI)



Read Data

```
HPI <- read.csv(file="global_ranking_2019.tab", header=T)

GNI <- read.csv(file = "GNI.csv", header = T)
```

Data Cleaning

```
## deal with HPI
HPI %<>% rename(score_2019 = access, Passport=country)
HPI %<>% select(Passport, score_2019)
HPI$score_2019 <- as.numeric(HPI$score_2019)
HPI$Passport <- str_trim(HPI$Passport,'right')

## Update county names in GNI
GNI$country[GNI$country=="Bahamas, The"] <- "Bahamas"
GNI$country[GNI$country=="Brunei Darussalam"] <- "Brunei"
GNI$country[GNI$country=="Cabo Verde"] <- "Cape Verde"
GNI$country[GNI$country=="Congo, Dem. Rep."] <- "Congo (Dem. Rep.)"
GNI$country[GNI$country=="Congo, Rep."] <- "Congo (Rep.)"
GNI$country[GNI$country=="Cote d'Ivoire"] <- "Cote d'Ivoire (Ivory Coast)"
GNI$country[GNI$country=="Egypt, Arab Rep."] <- "Egypt"
GNI$country[GNI$country=="Eswatini"] <- "eSwatini"
GNI$country[GNI$country=="Gambia, The"] <- "Gambia"
GNI$country[GNI$country=="Hong Kong SAR, China"] <- "Hong Kong (SAR China)"
GNI$country[GNI$country=="Iran, Islamic Rep."] <- "Iran"
GNI$country[GNI$country=="Kyrgyz Republic"] <- "Kyrgyzstan"
GNI$country[GNI$country=="Lao PDR"] <- "Laos"
GNI$country[GNI$country=="Macao SAR, China"] <- "Macao (SAR China)"
GNI$country[GNI$country=="Micronesia, Fed. Sts."] <- "Micronesia"
GNI$country[GNI$country=="Korea, Dem. People’s Rep."] <- "North Korea"
GNI$country[GNI$country=="Slovak Republic"] <- "Slovakia"
GNI$country[GNI$country=="Korea, Rep."] <- "South Korea"
GNI$country[GNI$country=="Syrian Arab Republic"] <- "Syria"
GNI$country[GNI$country=="Timor-Leste"] <- "Timor Leste"
GNI$country[GNI$country=="Venezuela, RB"] <- "Venezuela"
GNI$country[GNI$country=="Yemen, Rep."] <- "Yemen"

## merge two datasets
GNI_HPI <- GNI %>%
  filter(year == 2018) %>%
  select(-year) %>%
  right_join(HPI, by = c('country' = 'Passport')) %>%
  select(country, gni, score_2019) %>%
  mutate(gni_log = log(gni))

## remove NA
GNI_HPI<-na.omit(GNI_HPI)
```

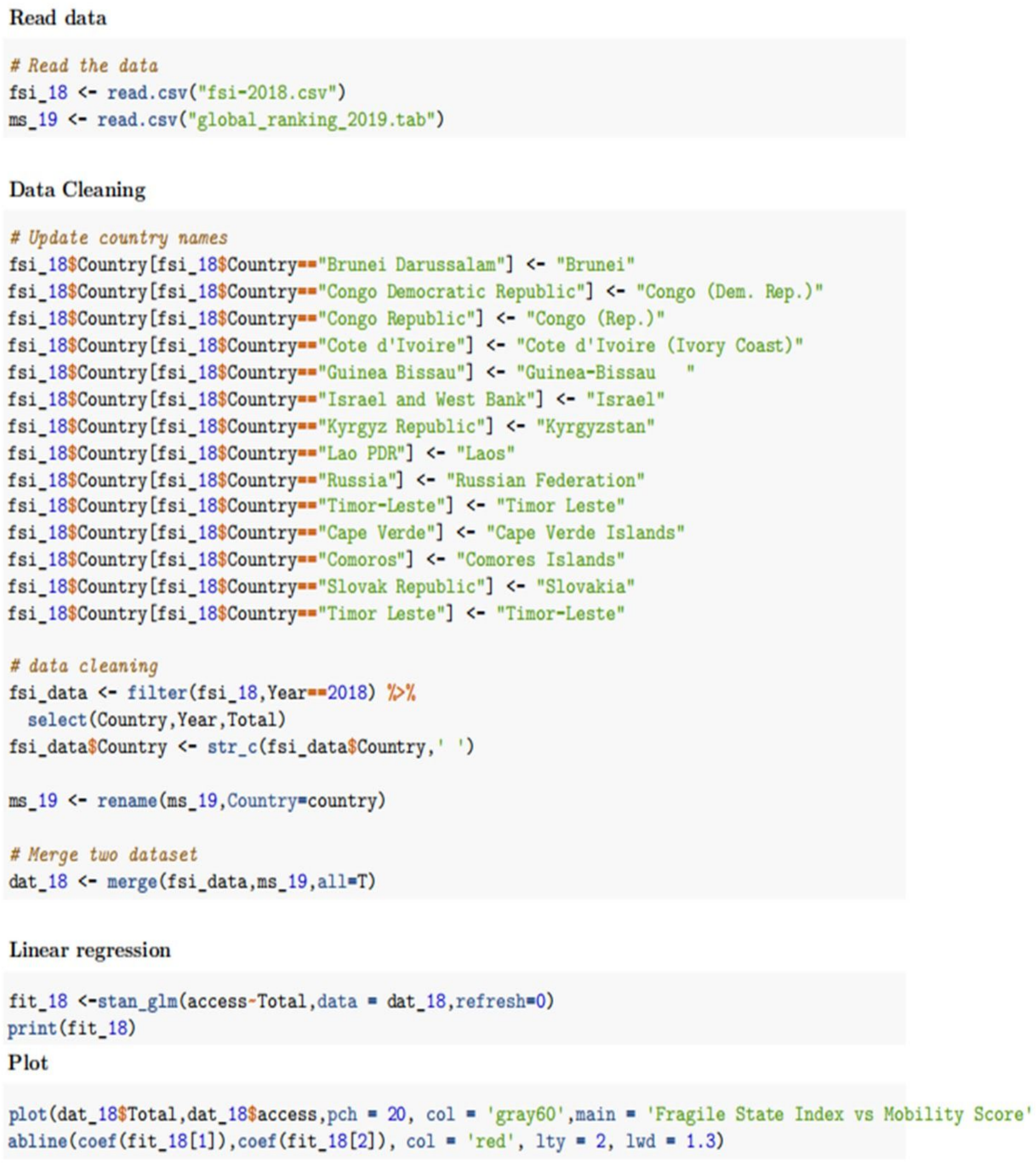
Linear Regression

```
fit1 <- lm(score_2019 ~ gni_log, data = GNI_HPI)
summary(fit1)
```

Plot

```
plot(GNI_HPI$gni_log, GNI_HPI$score_2019, pch = 16, col = 'gray60',
     main = 'Log GNI (2018) vs Mobility Score (2019)',
     ylab = "Mobility Score", xlab = "Log GNI, PPP")
abline(a = fit1$coef[1], b = fit1$coef[2], col = 'red', lty = 2, lwd = 1.3)
```

4.2 Mobility Score ~ Fragile State Index



Read data

```
# Read the data
fsi_18 <- read.csv("fsi-2018.csv")
ms_19 <- read.csv("global_ranking_2019.tab")
```

Data Cleaning

```
# Update country names
fsi_18$Country[fsi_18$Country=="Brunei Darussalam"] <- "Brunei"
fsi_18$Country[fsi_18$Country=="Congo Democratic Republic"] <- "Congo (Dem. Rep.)"
fsi_18$Country[fsi_18$Country=="Congo Republic"] <- "Congo (Rep.)"
fsi_18$Country[fsi_18$Country=="Cote d'Ivoire"] <- "Cote d'Ivoire (Ivory Coast)"
fsi_18$Country[fsi_18$Country=="Guinea Bissau"] <- "Guinea-Bissau   "
fsi_18$Country[fsi_18$Country=="Israel and West Bank"] <- "Israel"
fsi_18$Country[fsi_18$Country=="Kyrgyz Republic"] <- "Kyrgyzstan"
fsi_18$Country[fsi_18$Country=="Lao PDR"] <- "Laos"
fsi_18$Country[fsi_18$Country=="Russia"] <- "Russian Federation"
fsi_18$Country[fsi_18$Country=="Timor-Leste"] <- "Timor Leste"
fsi_18$Country[fsi_18$Country=="Cape Verde"] <- "Cape Verde Islands"
fsi_18$Country[fsi_18$Country=="Comoros"] <- "Comores Islands"
fsi_18$Country[fsi_18$Country=="Slovak Republic"] <- "Slovakia"
fsi_18$Country[fsi_18$Country=="Timor Leste"] <- "Timor-Leste"

# data cleaning
fsi_data <- filter(fsi_18,Year==2018) %>%
  select(Country,Year,Total)
fsi_data$Country <- str_c(fsi_data$Country,' ')

ms_19 <- rename(ms_19,Country=country)

# Merge two dataset
dat_18 <- merge(fsi_data,ms_19,all=T)
```

Linear regression

```
fit_18 <-stan_glm(access~Total,data = dat_18,refresh=0)
print(fit_18)
```

Plot

```
plot(dat_18$Total,dat_18$access,pch = 20, col = 'gray60',main = 'Fragile State Index vs Mobility Score'
abline(coef(fit_18[1]),coef(fit_18[2]), col = 'red', lty = 2, lwd = 1.3)
```