HW5_ahcooper

Andrew Cooper

11/2/2020

```
library(tidyverse)
library(readr)
library(janitor)
library(MASS)
library(ggfortify)
library(gridExtra)
```

Problem 3

My r-session crashed everytime I attempted to load the file "EdStatsData.csv" in its entirety. TO avoid this issue I only read in the first 1000 rows, which makes my compuations of the number of rows in the data before and after munging it innacurate.

```
EdStatsCountry <- read_csv("~/STAT5014/EdStatsCountry.csv")
EdStatsCountry_Series <- read.csv("~/STAT5014/EdStatsCountry-Series.csv", col.names = c("Country Code",
EdStatsData <- read_csv("~/STAT5014/EdStatsData.csv", n_max = 1000)
EdStatsFootNote <- read_csv("~/STAT5014/EdStatsFootNote.csv")
EdStatsSeries <- read_csv("~/STAT5014/EdStatsSeries.csv")
```

```
df1 <- left_join(EdStatsCountry, EdStatsCountry_Series, by = c("Country Code" = "Country.Code")) %>%
  left_join(EdStatsData, by = "Country Code") %>%
  left_join(EdStatsFootNote, by = c("Country Code" = "CountryCode")) %>%
  left_join(EdStatsSeries, by = c("SeriesCode" = "Series Code"))
```

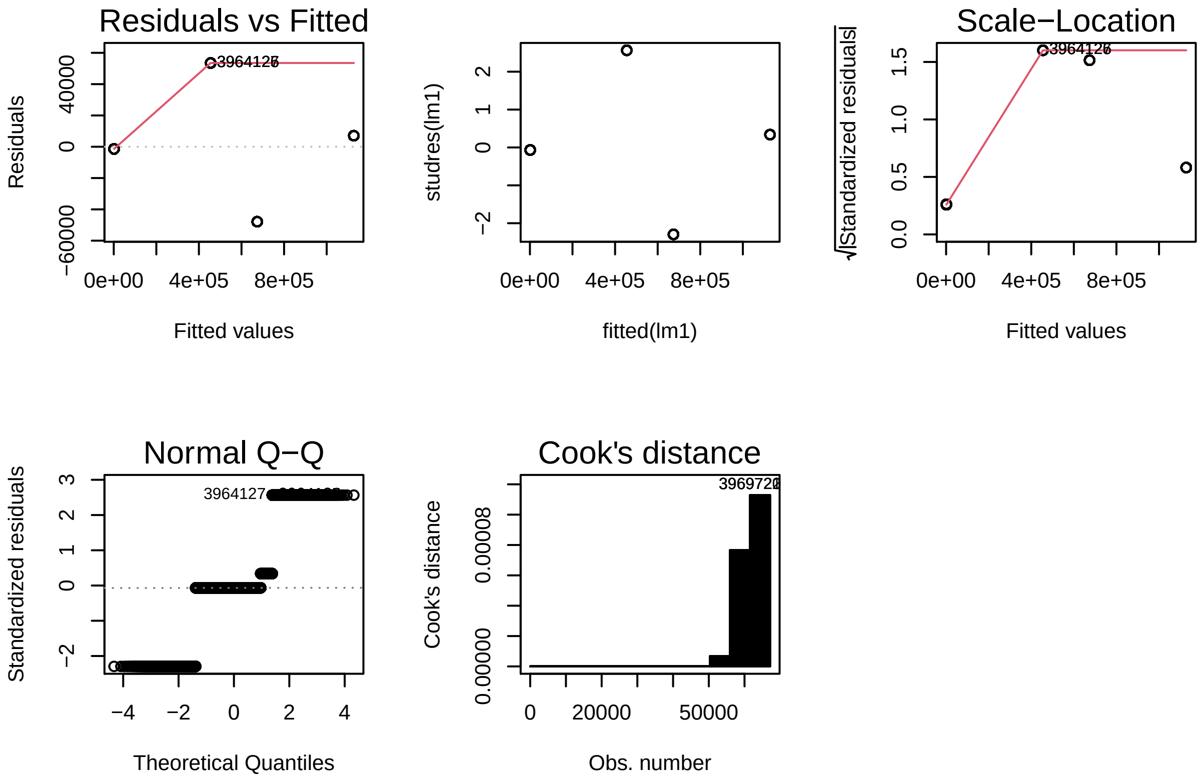
```
df1_clean <- df1 %>%
  janitor::clean_names()
```

There were 6.49157×10^5 rows in the original data. There are now 7.317784×10^6 rows in the cleaned data.

Problem 4

```
lm1 <- lm(x1971 ~ x1970, df1_clean)
```

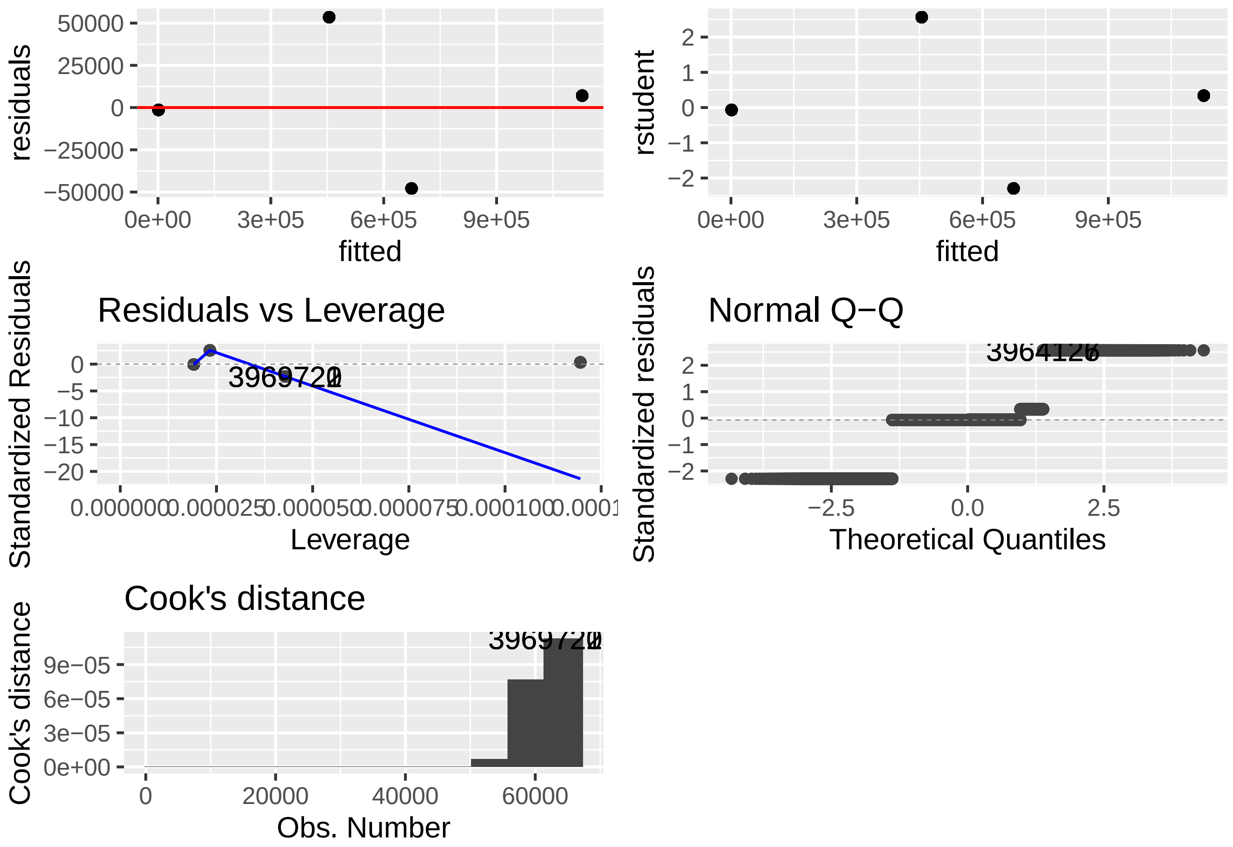
```r
par(mfrow = c(2, 3))
plot(lm1, which = 1)
plot(fitted(lm1), studres(lm1))
plot(lm1, which = 3)
plot(lm1, which = 2)
plot(lm1, which = 4)
```



Problem 5

```r
lm_df <- data_frame("fitted" = fitted(lm1), "residuals" = residuals(lm1), "rstudent" = studres(lm1))
```

```r
grid.arrange(
lm_df %>% ggplot(aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_abline(slope = 0, intercept = 0, col = "red"),
lm_df %>% ggplot(aes(x = fitted, y = rstudent)) +
  geom_point(),
autoplot(lm1, which = 5)[[1]],
autoplot(lm1, which = 2)[[1]],
autoplot(lm1, which = 4)[[1]]
)
```


residuals
50000
25000
0
−25000
−50000
0e+00
3e+05
6e+05
9e+05
fitted
rstudent
2
1
0
−1
−2
0e+00
3e+05
6e+05
9e+05
fitted
Residuals vs Leverage
Standardized Residuals
0
−5
−10
−15
−20
3969720
0.000000
0.000025
0.000050
0.000075
0.000100
0.0001
Leverage
Normal Q−Q
Standardized residuals
3964126
2
1
0
−1
−2
−2.5
0.0
2.5
Theoretical Quantiles
Cook's distance
Cook's distance
3969720
9e−05
6e−05
3e−05
0e+00
0
20000
40000
60000
Obs. Number