HW5_xinyis

Xinyi Song

10/27/2020

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Exploratory Data Analysis and plotting. To begin the homework, we will as usual, start by loading, munging and creating tidy data sets. In this homework, our goal is to create informative (and perhaps pretty) plots showing features or perhaps deficiencies in the data.

Problem 2

Create a new R Markdown file within your local GitHub repo folder (file–>new–>R Markdown–>save as).

The filename should be: HW5_lastname, i.e. for me it would be HW5_Settlage

You will use this new R Markdown file to solve the following problems.

Problem 3

`Solution`

Using tidy concepts, get and clean the following data on education from the World Bank.

http://databank.worldbank.org/data/download/Edstats_csv.zip

How many data points were there in the complete dataset? In your cleaned dataset?

Choosing 2 countries, create a summary table of indicators for comparison.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts --------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
dat <- read.csv("~/Desktop/VTCourse/STAT 5014/HW5/Edstats_csv/EdStatsData.csv")
num_obs_complete = dim(dat)[1]
print(num_obs_complete)
```

```
## [1] 886930
```

Based on the results above, we can see that there are 886930 observations in the complete dataset.

```
# Column 70 is the empty column
# Delete the empty column
k = matrix(0,1, dim(dat)[2])
for (i in 1:dim(dat)[2]){
  k[i] = FALSE %in% (is.na(dat[,i]))
}
full_null = which(k==0)
data = dat[,-full_null]
clean_dat = data%>%gather(key = 'Year', value = 'Value', 5:69, na.rm = TRUE)
print(dim(clean_dat)[1])
```

```
## [1] 5082201
```

In my cleaned dataset, there are 5082201 observations.

```
summary_table = clean_dat %>% select(Indicator.Code, Country.Code, Value) %>% filter(Country.Code == c(
```

```
## `summarise()` regrouping output by 'Indicator.Code' (override with `.groups` argument)
```

```
knitr::kable(summary_table[1:20,])
```

Indicator.Code	Country.Code	Mean	Median
BAR.NOED.1519.FE.ZS	MUS	3.2580	2.700
BAR.NOED.1519.FE.ZS	ZWE	4.2220	1.200
BAR.NOED.1519.ZS	MUS	2.1500	0.750
BAR.NOED.1519.ZS	ZWE	7.5825	4.300
BAR.NOED.15UP.FE.ZS	MUS	18.2550	14.180
BAR.NOED.15UP.FE.ZS	ZWE	19.2525	18.425
BAR.NOED.15UP.ZS	MUS	15.9660	17.220
BAR.NOED.15UP.ZS	ZWE	18.4680	14.580
BAR.NOED.2024.FE.ZS	MUS	5.3375	1.405
BAR.NOED.2024.FE.ZS	ZWE	11.6200	10.920
BAR.NOED.2024.ZS	MUS	4.8660	3.500
BAR.NOED.2024.ZS	ZWE	7.0840	3.330
BAR.NOED.2529.FE.ZS	MUS	9.3375	2.165
BAR.NOED.2529.FE.ZS	ZWE	12.0175	10.920
BAR.NOED.2529.ZS	MUS	7.4420	8.500
BAR.NOED.2529.ZS	ZWE	12.2500	11.000
BAR.NOED.25UP.FE.ZS	MUS	29.1360	32.400
BAR.NOED.25UP.FE.ZS	ZWE	32.3040	29.300
BAR.NOED.25UP.ZS	MUS	17.5350	12.970
BAR.NOED.25UP.ZS	ZWE	19.9275	17.755

For this part, I choose the country 'ZWE' and 'MUS'. And the first 20 obs of summary_table is above.

Problem 4

Using *base* plotting functions, create a single figure that is composed of the first two rows of plots from SAS's simple linear regression diagnostics as shown here: https://support.sas.com/rnd/app/ODSGraphics/examples/reg.html. Demonstrate the plot using suitable data from problem 3.

```
library(tidyverse)
library(dplyr)
library(MASS)
```

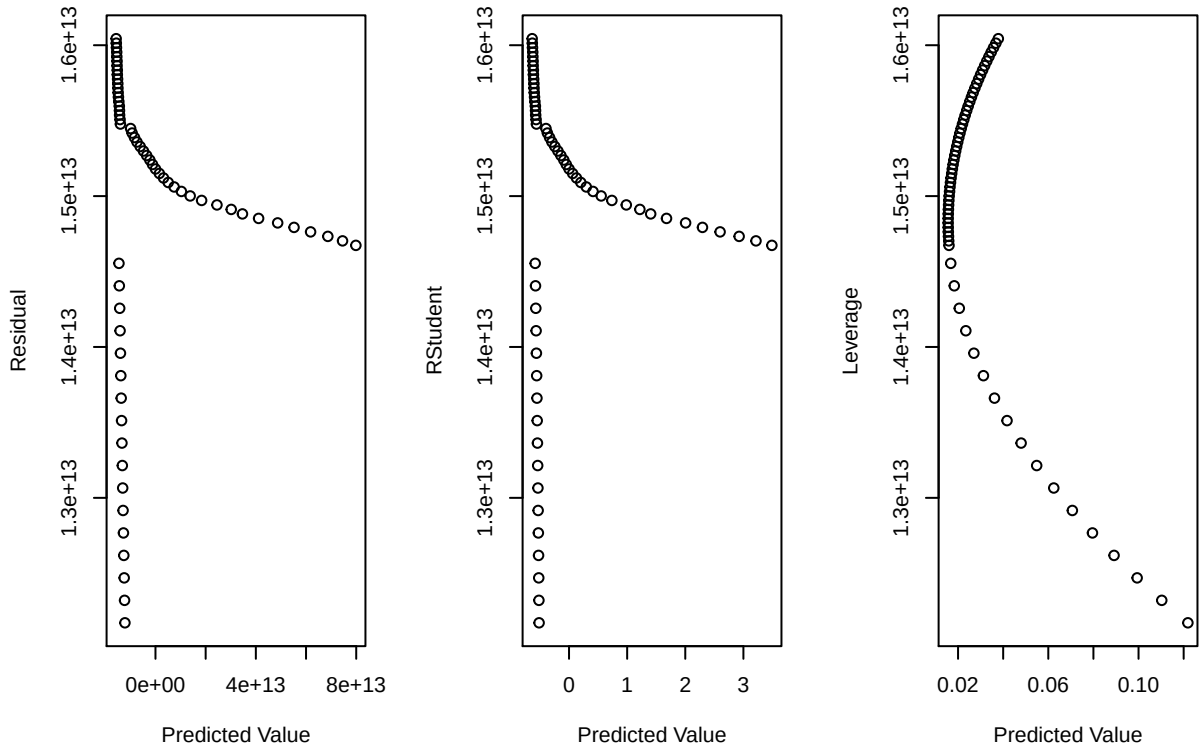
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

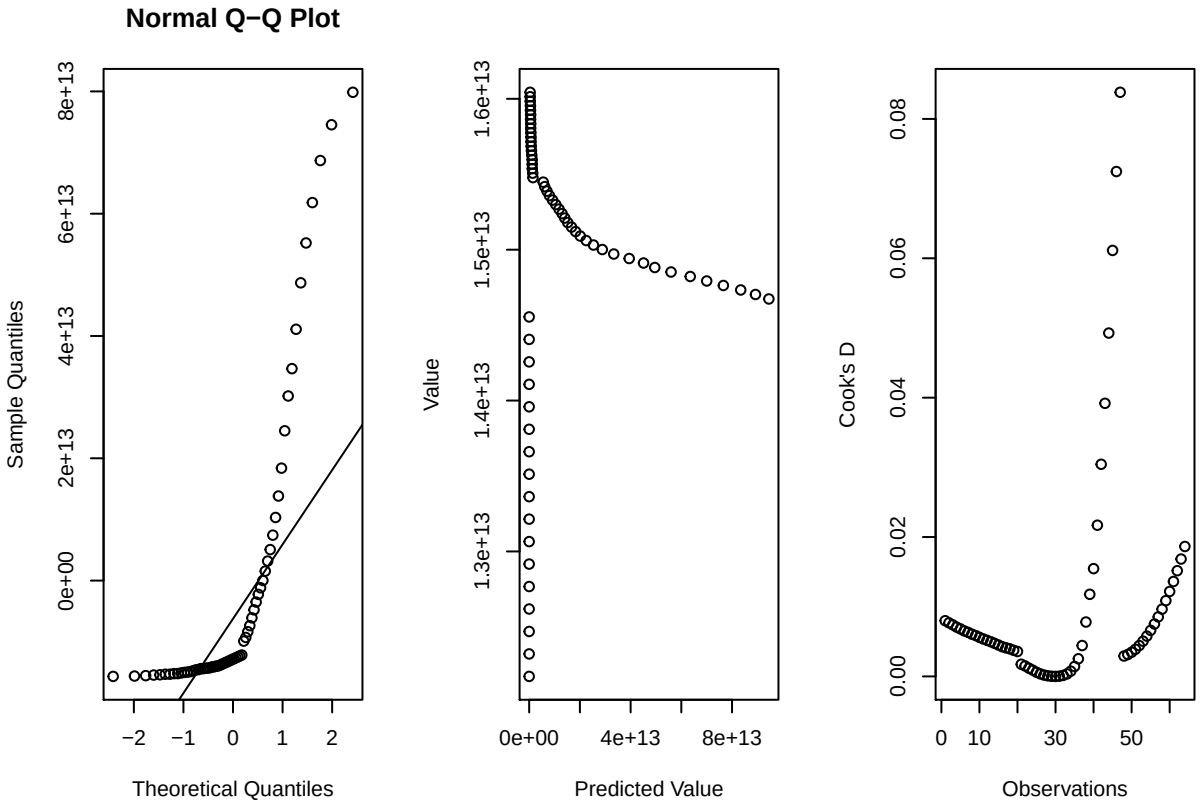
```
CHN_dat <- clean_dat %>% filter(Country.Code == 'CHN')%>% group_by(Year)%>% summarise(sum(Value))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
dat = as.data.frame(CHN_dat)
k = as.numeric(gsub('X', '', dat$Year))
data_CHN = as.data.frame(cbind(k, dat$`sum(Value)`))
colnames(data_CHN) = c('Year', 'Value')
fit = lm(Value~ as.numeric(Year), data = as.data.frame(data_CHN))
par(mfrow=c(1,3))
plot(fit$residuals, fit$fitted.values, xlab = 'Predicted Value', ylab = 'Residual')
plot(studres(fit) ,  fit$fitted.values, xlab = 'Predicted Value', ylab = 'RStudent')
plot(hatvalues(fit),fit$fitted.values, xlab = 'Predicted Value', ylab = 'Leverage')
```




```
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(data_CHN$Value, fit$fitted.values, xlab = 'Predicted Value',ylab = 'Value')
plot(cooks.distance(fit), xlab = 'Observations', ylab = "Cook's D")
```



Here for the linear regression, I mainly focus analyze country of China, the relationship between value and year.

Problem 5

Recreate the plot in problem 3 using ggplot2 functions. Note: there are many extension libraries for ggplot, you will probably find an extension to the ggplot2 functionality will do exactly what you want.

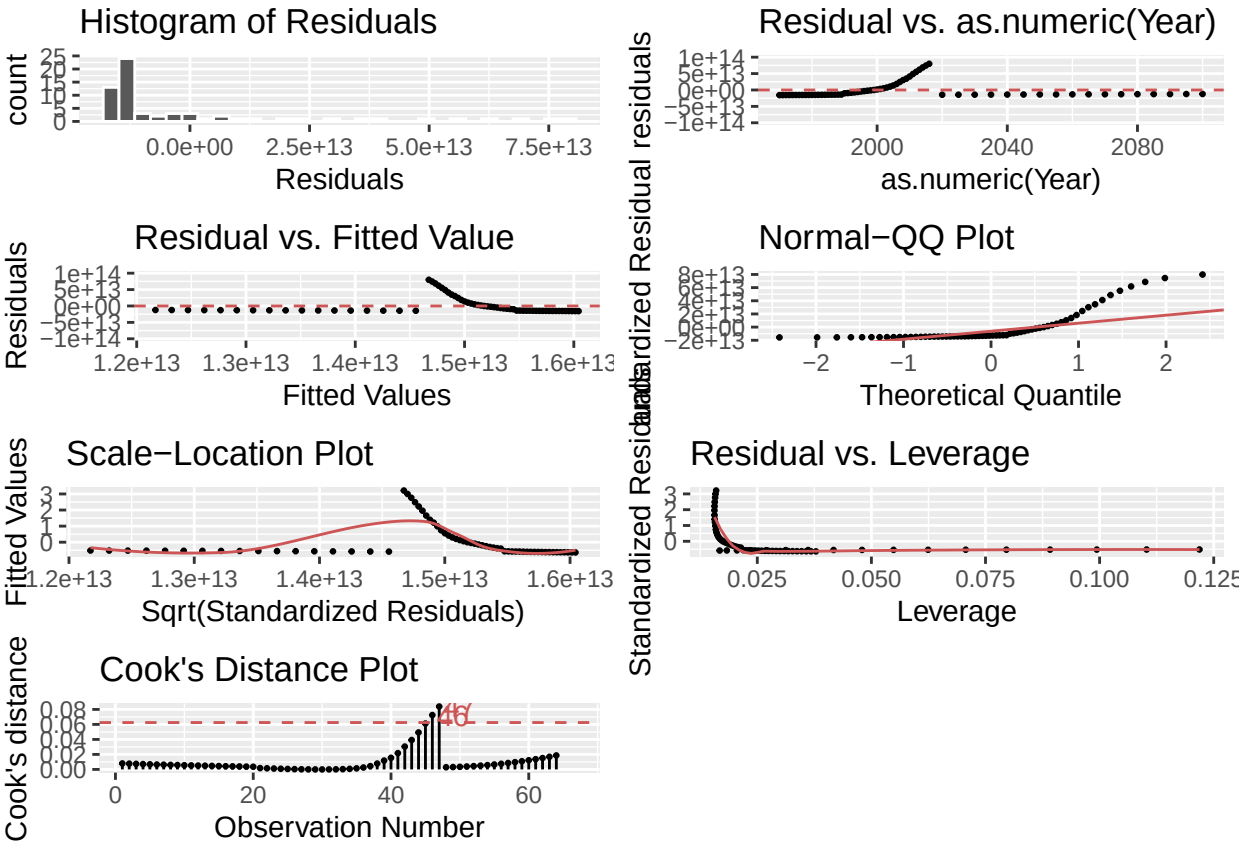
```
devtools::install_github("yeukyul/lindia")
```

```
## Skipping install of 'lindia' from a github remote, the SHA1 (9853c34f) has not changed since last ins
##   Use `force = TRUE` to force installation
```

```
library(lindia)
gg_diagnose(fit)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Problem 6

Finish this homework by pushing your changes to your repo.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW5_lastname_firstname.Rmd and HW5_lastname_firstname.pdf