STA Homework Template

Amy Kim

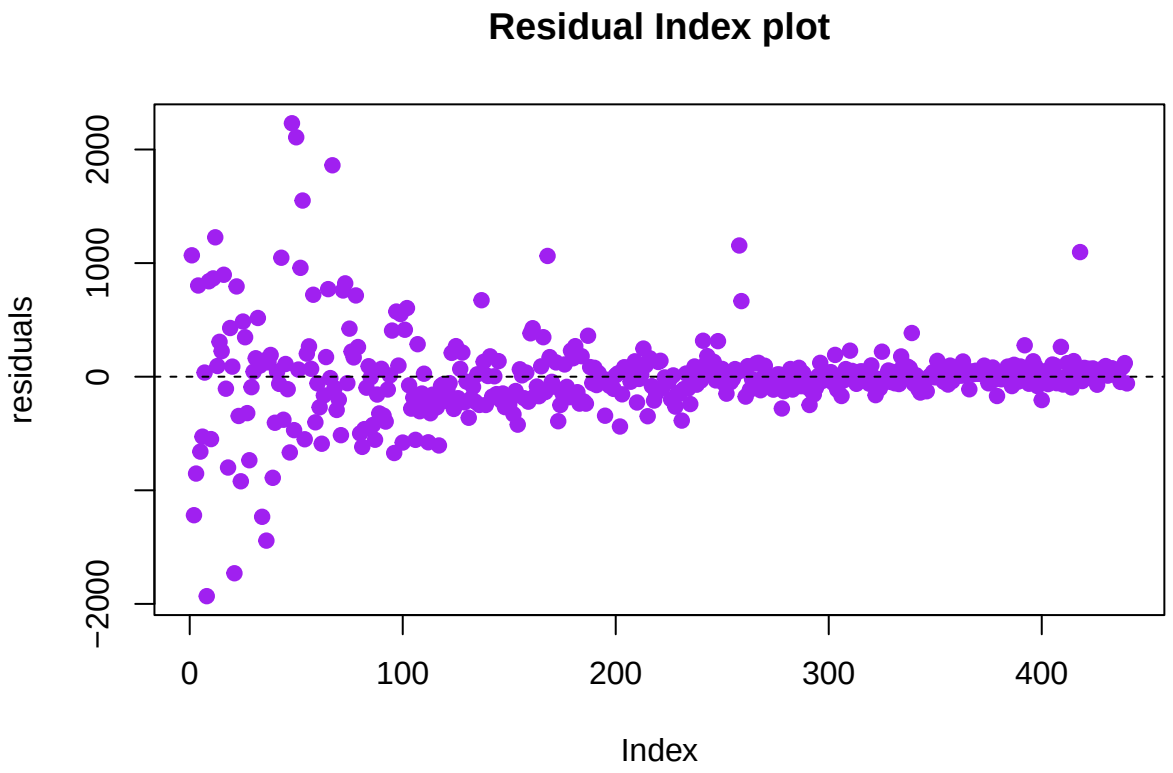
```
corrplot 0.92 loaded
```

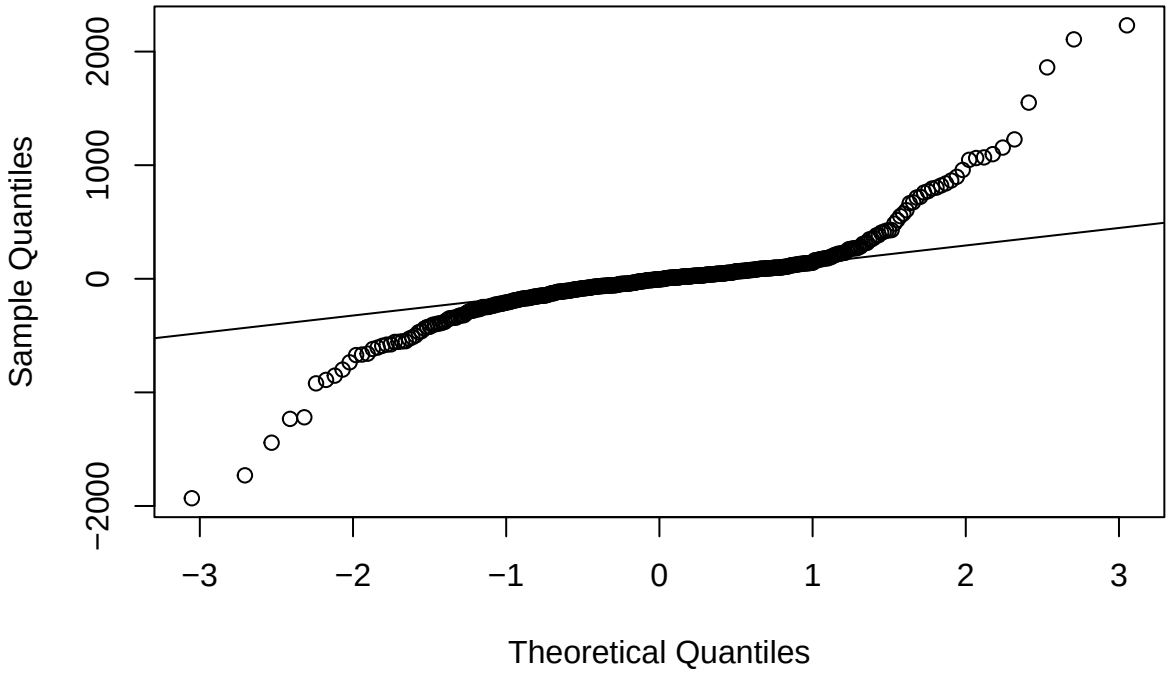
```
            physician  area population senior   bed  crime income
physician       1.000 0.078      0.940 -0.003 0.950  0.820  0.948
area            0.078 1.000      0.173  0.006 0.073  0.129  0.127
population      0.940 0.173      1.000 -0.029 0.924  0.886  0.987
senior         -0.003 0.006     -0.029  1.000 0.053 -0.035 -0.023
bed             0.950 0.073      0.924  0.053 1.000  0.857  0.902
crime           0.820 0.129      0.886 -0.035 0.857  1.000  0.843
income          0.948 0.127      0.987 -0.023 0.902  0.843  1.000
```


```
   physician           area          population        senior     
 Min.   :   39.0   Min.   :   15.0   Min.   : 100043   Min.   : 3.000  
 1st Qu.:  182.8   1st Qu.:  451.2   1st Qu.: 139027   1st Qu.: 9.875  
 Median :  401.0   Median :  656.5   Median : 217280   Median :11.750  
 Mean   :  988.0   Mean   : 1041.4   Mean   : 393011   Mean   :12.170  
 3rd Qu.: 1036.0   3rd Qu.:  946.8   3rd Qu.: 436064   3rd Qu.:13.625  
 Max.   :23677.0   Max.   :20062.0   Max.   :8863164   Max.   :33.800  
      bed              crime            income      
 Min.   :   92.0   Min.   :   563   Min.   :  1141  
 1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:  2311  
 Median :  755.0   Median : 11820   Median :  3857  
```

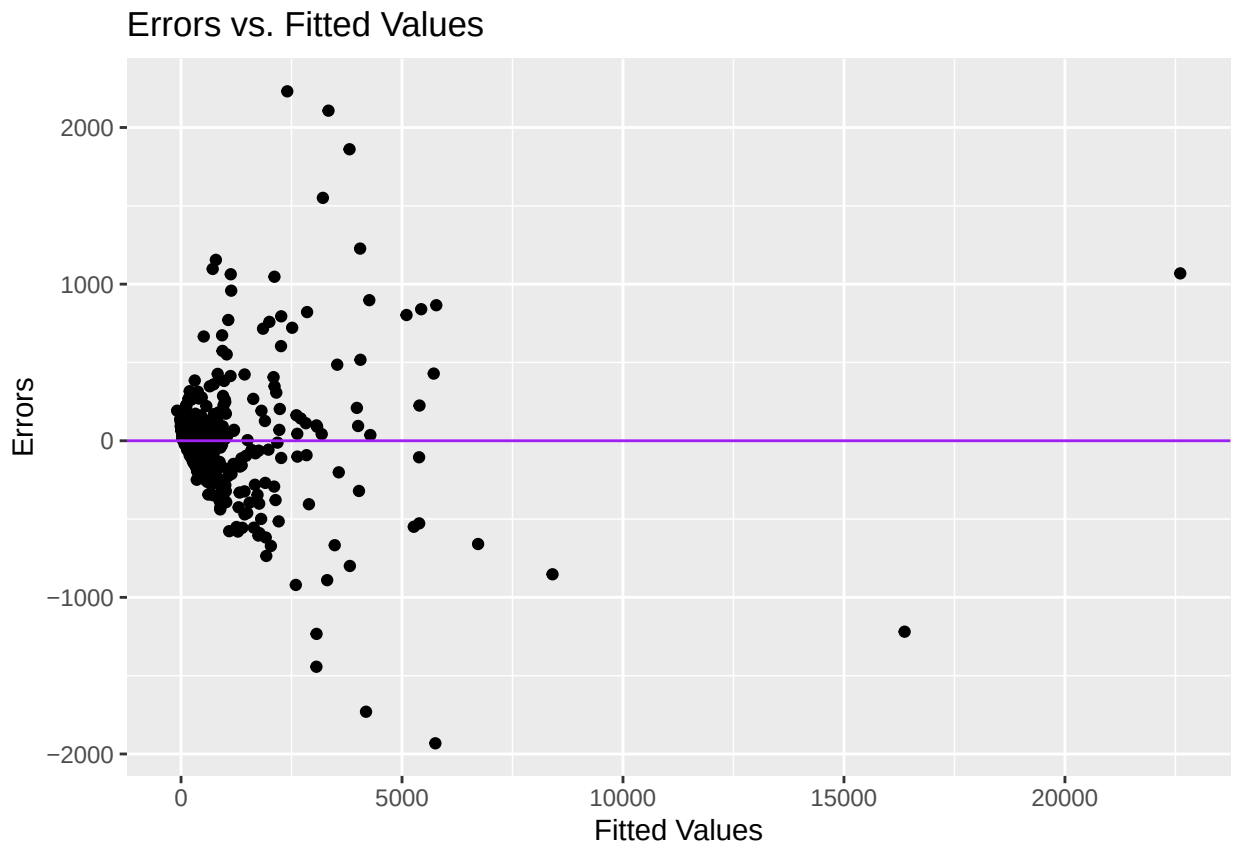
```
 Mean   : 1458.6   Mean   : 27112   Mean   :  7869
 3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:  8654
 Max.   :27700.0   Max.   :688936   Max.   :184230

$physician
[1] 1789.75

$area
[1] 1549.922

$population
[1] 601987

$senior
[1] 3.992666

$bed
[1] 2289.134

$crime
[1] 58237.51

$income
[1] 12884.32



Call:
lm(formula = physician ~ population + income + bed, data = CDI2)

Residuals:
     Min       1Q   Median       3Q      Max
-1931.75  -118.96    -4.76    88.95  2230.98

Coefficients:
              Estimate  Std. Error t value             Pr(>|t|)
(Intercept) -89.1047384  21.9773223  -4.054            0.0000595 ***
population   -0.0018325   0.0002116  -8.661 < 0.0000000000000002 ***
income        0.1381999   0.0087725  15.754 < 0.0000000000000002 ***
bed           0.4866043   0.0209173  23.263 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 379.8 on 436 degrees of freedom
Multiple R-squared:  0.9553,    Adjusted R-squared:  0.955
F-statistic:  3104 on 3 and 436 DF,  p-value: < 0.00000000000000022



Call:
lm(formula = physician ~ pdensity + senior + income, data = CDI2)

Residuals:
     Min       1Q   Median       3Q      Max
-3055.75  -175.30   -38.05    72.88  3045.81

Coefficients:
              Estimate  Std. Error t value             Pr(>|t|)
(Intercept) -170.574223   83.532885  -2.042               0.0418 *
pdensity       0.096159    0.012238   7.857    0.000000000000031 ***
senior         6.339841    6.383772   0.993               0.3212
```

```
income         0.126566    0.002084  60.723 < 0.0000000000000002 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 533.5 on 436 degrees of freedom
Multiple R-squared:  0.9117,    Adjusted R-squared:  0.9111
F-statistic:  1501 on 3 and 436 DF,  p-value: < 0.00000000000000022
```

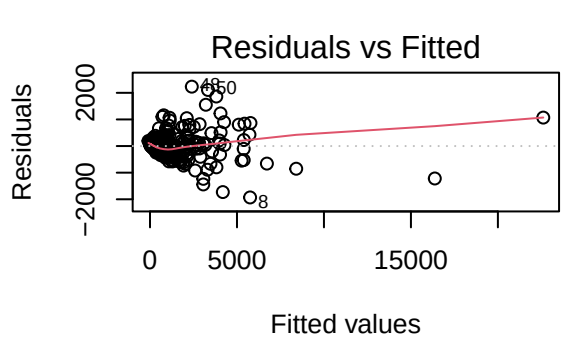


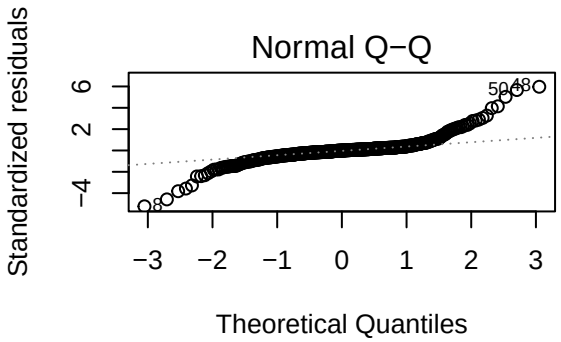


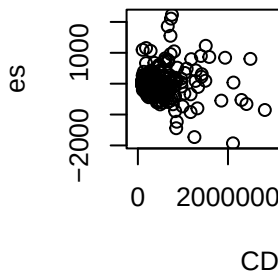

```
    Shapiro-Wilk normality test

data:  model_1$residuals
W = 0.8028, p-value < 0.00000000000000022
```

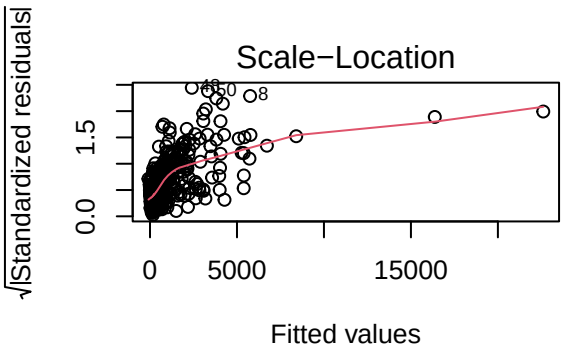


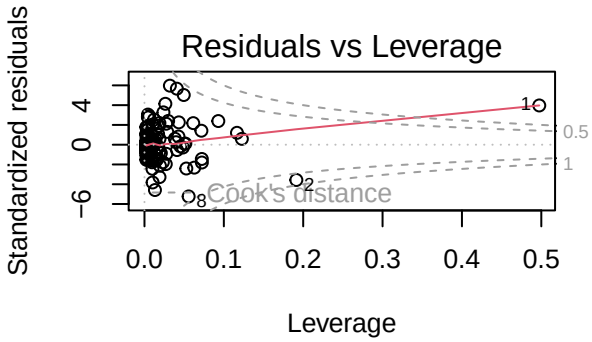

```
    Fligner-Killeen test of homogeneity of variances

data:  CDI2$ei and CDI2$Group
Fligner-Killeen:med chi-squared = 133.01, df = 1, p-value <
0.00000000000000022
```

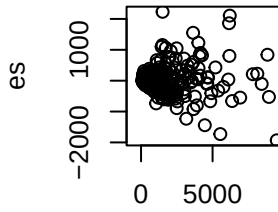













```
  1   2   3   4   5   6   8   9  11  12  16  18  19  21  22  24  25  28  32  34
  1   2   3   4   5   6   8   9  11  12  16  18  19  21  22  24  25  28  32  34
 36  39  47  48  50  52  53  58  65  67  72  73 102 117 123 258 418
 36  39  47  48  50  52  53  58  65  67  72  73 102 117 123 258 418
```

```
Rows: 440 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (7): physician, area, population, senior, bed, crime, income

i Use ‘spec()‘ to retrieve the full column specification for this data.
i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

physician	area	population	senior	bed	crime	income	row
23677	4060	8863164	9.7	27700	688936	184230	1
15153	946	5105067	12.4	21550	436936	110928	2
3823	614	2111687	12.5	9490	193978	36872	8
5280	2126	1507319	11.1	4009	124959	35843	12
2456	1209	1255488	20.7	5543	107386	28066	21
1833	1974	863518	24.4	3164	76142	23141	34
1620	280	851659	26.0	4458	62344	18404	36
4635	495	757027	10.2	1507	34754	22772	48
5444	81	736014	13.7	6203	87355	12706	50
4761	47	723959	14.5	3640	71234	20656	53
5674	59	663906	12.1	6154	68808	15369	67
1944	291	181835	10.7	1496	15477	3498	258

physician	area	population	senior	bed	crime	income	row
23677	4060	8863164	9.7	27700	688936	184230	1
15153	946	5105067	12.4	21550	436936	110928	2
7553	1729	2818199	7.1	12449	253526	55003	3
5905	4205	2498016	10.9	6179	173821	48931	4
6062	790	2410556	9.2	6369	144524	58818	5
4861	71	2300664	12.4	8942	680966	38658	6
4320	9204	2122101	12.5	6104	177593	38287	7
3823	614	2111687	12.5	9490	193978	36872	8
6274	1945	1937094	13.9	8840	244725	34525	9
4718	880	1852810	8.2	6934	214258	38911	10
6641	135	1585577	15.2	10494	109148	26512	11
5280	2126	1507319	11.1	4009	124959	35843	12
4101	1291	1497577	8.7	3342	77009	37728	13
2463	20062	1418380	8.8	3349	83110	23260	14
5620	458	1412140	15.6	8132	73150	29776	15
5158	824	1398468	12.5	4152	35825	35398	16
5281	730	1336449	17.4	8436	50186	27639	17
3021	911	1321864	10.8	3904	66723	32071	18
6147	287	1287348	14.2	5200	43203	40782	19
3169	738	1279182	10.6	3284	107338	28331	20
2456	1209	1255488	20.7	5543	107386	28066	21
3062	1247	1185394	9.9	4086	133098	18383	22
1385	7208	1170413	13.2	2435	95494	20114	23
4020	873	1083592	10.9	3254	50964	29131	25
3706	557	1032431	11.3	5395	71753	24474	27
1194	508	993529	13.1	1056	42595	24062	28
4577	433	874866	14.4	3540	37118	29159	32
1833	1974	863518	24.4	3164	76142	23141	34
2417	626	827645	13.3	2494	44374	26768	39
2489	755	826330	10.4	4918	67032	15229	40
3226	234	825380	15.3	2279	28521	26602	41
1694	396	818584	6.5	135	30202	23738	42
1761	720	803732	10.9	1781	51243	20514	44
2936	396	797159	11.7	4654	61004	15264	45
2157	334	781666	8.7	1842	29708	20927	46
2811	126	778206	12.7	4841	75595	19084	47
4635	495	757027	10.2	1507	34754	22772	48
5444	81	736014	13.7	6203	87355	12706	50
2094	737	725956	8.5	2076	58610	11179	52
4761	47	723959	14.5	3640	71234	20656	53
1269	599	692134	14.0	641	46789	16244	57
3237	483	678111	15.0	2425	20335	19300	58
5674	59	663906	12.1	6154	68808	15369	67
2532	1113	651525	14.0	4602	55604	12134	68
1814	449	649623	12.3	1642	30473	18721	69
3368	529	648951	10.0	5757	93025	14808	70
3674	61	606900	12.8	4262	64393	14325	73
795	1013	591610	8.1	1650	54002	6830	76
2293	502	510784	11.6	3847	45237	9963	90
2500	181	496938	13.0	4018	54238	8238	95
2867	153	467610	13.8	3652	37466	10360	102
1147	469	421353	10.6	1599	12147	13281	117
4189	62	396685	16.6	7814	64103	7185	123
311	1569	383545	10.1	860	26712	3413	128
1001	520	230096	12.3	488	9460	8638	206

R Appendix

```
knitr::opts_chunk$set(echo = FALSE, comment = NA)
options(scipen = 999) #Remove the scientific notation
#### LOADING IN DATASET ####
library(readr)
CDI2 <- read_csv("CDI2.csv")
#### SUMMARY ####
  # Correlation matrix
library(corrplot)
round(cor(CDI2),3)
corrplot(cor(CDI2))
  # Numerical Summaries
summary(CDI2)
lapply(CDI2, sd)
#### MODEL COMPARISON & FIT ####
model_1 = lm(physician ~ population + income + bed, data = CDI2)
CDI2$pdensity = CDI2$population/CDI2$area
model_2 = lm(physician ~ pdensity + senior + income, data = CDI2)
summary(model_1)
summary(model_2)
#### DIAGNOSTICS ####
  # Assessing Independence
plot(model_1$residuals,main = "Residual Index plot",xlab = "Index",ylab = "residuals",pch = 19, col = "purple"
abline(h = 0, lty = 2)
  # Assessing Normality
    # Normal Q-Q Plot
qqnorm(model_1$residuals)
qqline(model_1$residuals)
    # Shapiro-Wilks Test
the.SWtest = shapiro.test(model_1$residuals)
the.SWtest
  # Assessing Constant Variance
    # Plotting Errors vs. Fitted Values
library(ggplot2)
CDI2$ei = model_1$residuals
CDI2$yhat = model_1$fitted.values
qplot(yhat, ei, data = CDI2) +  ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
  ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
    # Formal Testing
Group = rep("Lower",nrow(CDI2))
Group[CDI2$physician < median(CDI2$physician)] = "Upper"
Group = as.factor(Group)
CDI2$Group = Group
the.FKtest= fligner.test(CDI2$ei, CDI2$Group)
the.FKtest
es = residuals(model_1)
par(mfrow = c(2,2))
plot(model_1)
plot(CDI2$population, es)
plot(CDI2$income, es)
plot(CDI2$bed, es)
#### OUTLIERS ####
  # Cook's Distance
n = nrow(CDI2)
CD = cooks.distance(model_1)
which(CD > 4/n)
  # Leverage
p = 4
```

```r
h = hatvalues(model_1)
leverage = which(h > (p+1)/n)
  # Studentized Residuals
sei = rstudent(model_1)
outliers = which(abs(sei) > 3)
  # Table of outliers and leverage points
CDI2 <- read_csv("CDI2.csv")
outlier_table = CDI2[outliers,]
outlier_table$row = outliers
leverage_table = CDI2[leverage,]
leverage_table$row = leverage
knitr::kable(outlier_table)
knitr::kable(leverage_table)
```