STATS 500 HOMEWORK 4 YUAN YIN

Problem 1

First, we fit the model with total SAT score as the response and *expand*, *salary*, *ratio* and *takers* as predictors. Summary the result and we found that the p-value for most predictors is too large to reject null hypothesis. To fit the model better, we perform regression diagnostics.

For the first step, we check whether the constant variance assumption holds for the errors, we plot the residuals variance with the variance of fitted values (i.e. estimated value $\hat{y}$):

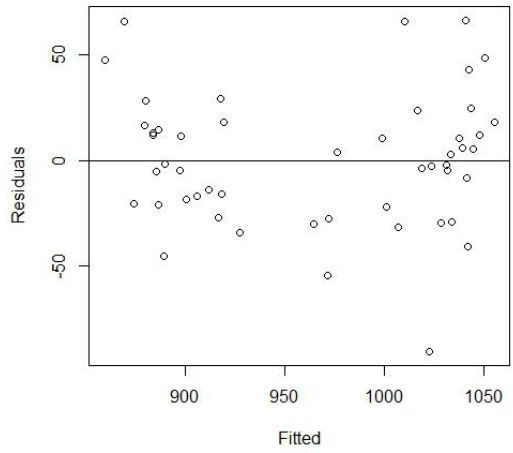

Fig 1 residuals vs fitted values

It looks like there is a non-linearity relationship between residuals and fitted values, which isn't supposed to be, as residuals should be independent with estimated values. Specifically, it looks like a quadratic relationship. Thus we want to find which predictors have a quadratic relationship with the fitted values.

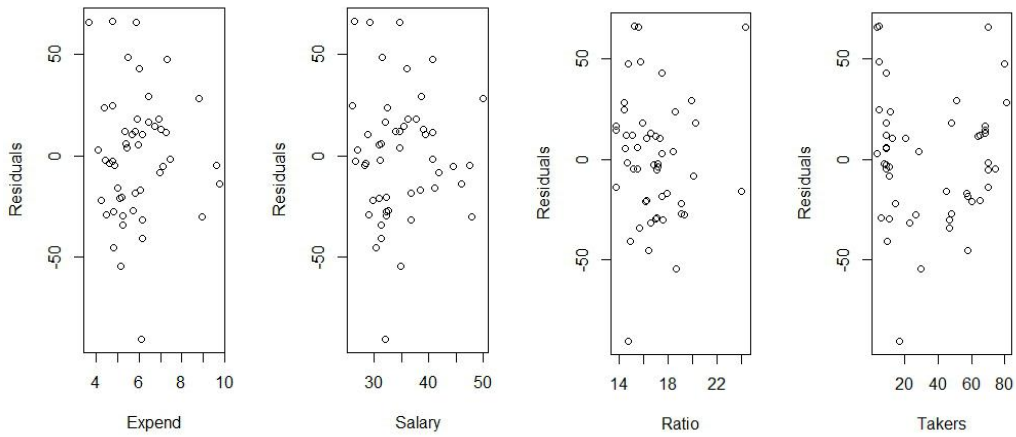

Fig 2 residuals vs each predictor variable

From the figure above, it's obvious that the data of *takers* has a quadratic relationship with residuals. In this way, we add a new variable which is quadratic term of takers into our fitted model. Summary the new model and we can see that p-value of *takers^2* = 8.58e-06 < 0.01, we can reject null hypothesis, which means *takers^2* can't be ignored in the model.

```
lm(formula = total ~ expend + salary + ratio + takers + I(takers^2),
    data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-68.181 -15.460  -1.722  18.063  52.163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1014.42328   43.04498  23.567  < 2e-16 ***
expend         9.11166    8.54525   1.066    0.292
salary        -0.07770    1.95288  -0.040    0.968
ratio          2.13991    2.83171   0.756    0.454
takers        -6.70915    0.77820  -8.621 5.26e-11 ***
I(takers^2)    0.05183    0.01029   5.035 8.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.34 on 44 degrees of freedom
Multiple R-squared:  0.8887,    Adjusted R-squared:  0.8761
F-statistic: 70.27 on 5 and 44 DF,  p-value: < 2.2e-16
```

Fig 3 summary of the new model

Plot the residuals vs new fitted values:

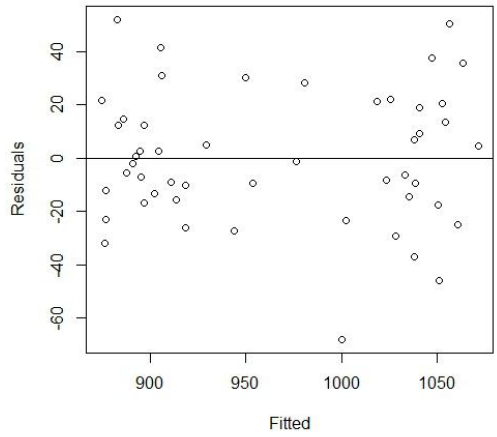

Fig 4 residuals vs updated fitted

Which looks much better than the original plot.

Next, we want to check whether the normality assumption holds. As studentized residuals are better than raw residuals, we draw QQ - plot with studentized residuals r_i .

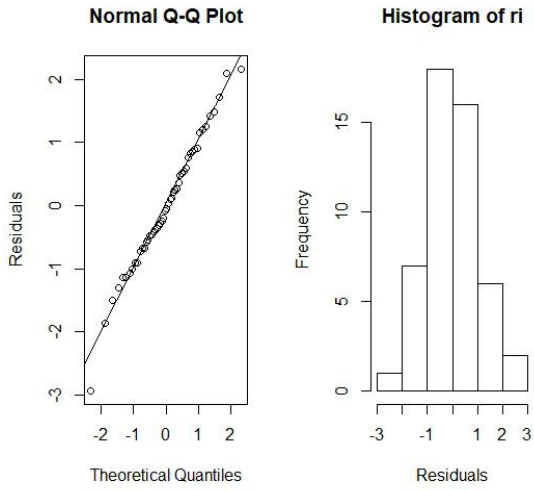

Fig 5 QQ - plot & histogram of

We found that almost all the points are on the line of normal distribution, also the right plot shows that ti satisfies normal distribution intuitively. Although the first point is a little far from the line of normal distribution, it doesn't affect the result because the data is finite (not vary large) and every point of residual is random, especially for the extreme residual (as their amount is very small so is more random). In conclusion, the normality assumption is reasonable.

The third step, we want to find points with large leverage. Using half-normal plot method, we want to find leverage of point $h_i = H_{ii}$ greater than $2(p+1)/n$, which will be considered with high leverage. The plot is as follows:

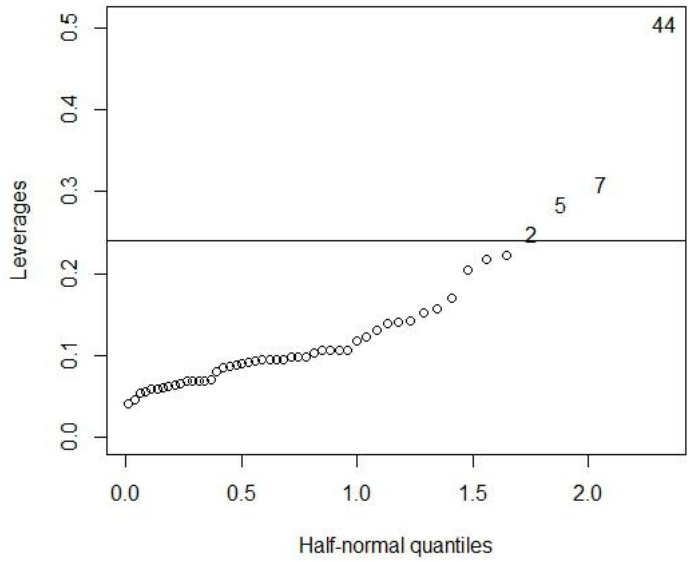

Fig 6 half - normal plot

The line in the figure above equals to $2(p+1)/n$, so all the points above the line have high leverage, which of them is point 2,5,7 and 44. the data of them are as follows (comparing with all the data):

```
    expend          ratio          salary          takers
Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
Median :5.768   Median :16.60   Median :33.29   Median :28.00
Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
    verbal           math           total
Min.   :401.0   Min.   :443.0   Min.   : 844.0
1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
Median :448.0   Median :497.5   Median : 945.5
Mean   :457.1   Mean   :508.8   Mean   : 965.9
3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
Max.   :516.0   Max.   :592.0   Max.   :1107.0

            expend ratio salary takers verbal math total
Alaska       8.963  17.6 47.951     47    445  489   934
California   4.992  24.0 41.078     45    417  485   902
Connecticut  8.817  14.4 50.045     81    431  477   908
Utah         3.656  24.3 29.082      4    513  563  1076
```

Fig 7 summary of data *sat* & data with high leverage

We can also tell from the data above with the reason they have high leverage. We see that "Alaska" has both high data in *expend* and *salary*, "California" has high data in *ratio*, "Connecticut" has high data in *expend*, and too much high data in *salary* and *takers*, and "Utah" has a very low data in *expend* and *takers*, too much high in *ratio*. What's more, they all have normal data in *total*, which makes them having a high leverage.

In the end, we want to check if there are outliers of the data. We do multiple hypothesis tests with Bonferroni Correction. First we found the maximum of studentized residual r_i, it's the 48th data "West Virginia", then we compute the p-value and compare it with adjusted $\alpha = \alpha / n$. We found that p-value = 0.005258305 is larger than the adjusted $\alpha = 0.001$. Hence, we cannot reject the null hypothesis, which means we aren't confident to say that the data with maximum r_i are outliers for the data, let alone the other points. In conclusion, there are no outliers for the data.

Appendix:

```
library(faraway)
data(sat)
temp = lm(total ~ expend + salary + ratio + takers, data = sat)
summary(temp)
## plot residuals vs fitted values
plot(temp$fitted, temp$residual, xlab = "Fitted", ylab = "Residuals")
abline(h = 0)
par(mfrow = c(1,2))
plot(sat$expend, temp$residual, xlab = "Expend", ylab = "Residuals")
plot(sat$salary, temp$residual, xlab = "Salary", ylab = "Residuals")
par(mfrow = c(1,2))
plot(sat$ratio, temp$residual, xlab = "Ratio", ylab = "Residuals")
plot(sat$takers, temp$residual, xlab = "Takers", ylab = "Residuals")
y = lm(total ~ expend + salary + ratio + takers + I(takers^2), data = sat)
summary(y)
par(mfrow = c(1,1))
plot(y$fitted, y$residual, xlab = "Fitted", ylab = "Residuals")
abline(h = 0)
## QQ-plot
ri = rstudent(y)
par(mfrow = c(1,2))
qqnorm(ri, ylab = "Residuals")
qqline(ri)
## Histogram
hist(ri, xlab = "Residuals")
## Half-normal plot for leverages
par(mfrow = c(1,1))
halfnorm(lm.influence(y)$hat, nlab = 4, ylab = "Leverages")
abline(h = 2*(5+1)/50)
sat[c(2,5,7,44),]
summary(sat)
max(abs(ri))
which(-ri == max(abs(ri)))
## compute p-value
2*(1 - pt(max(abs(ri)), df = 50-6-1))
## compare to alpha/n
0.05/50
```