STAT 423 Homework 3

Nan Tang 1662478

February 24, 2020

1

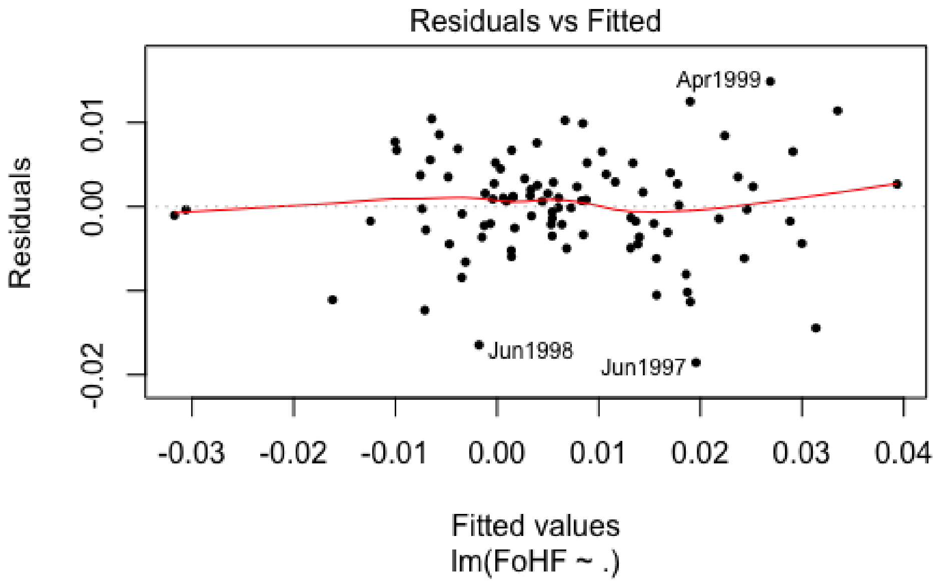
a

At the significance level of 5%, base on p-value of t-test, predictors RV, CA, FIA, CTA have significant effects on returns of FoHF. Among these significant factors, RV has negative impact, while others have positive impact, since estimated coefficient for RV is negative and are positive for others.

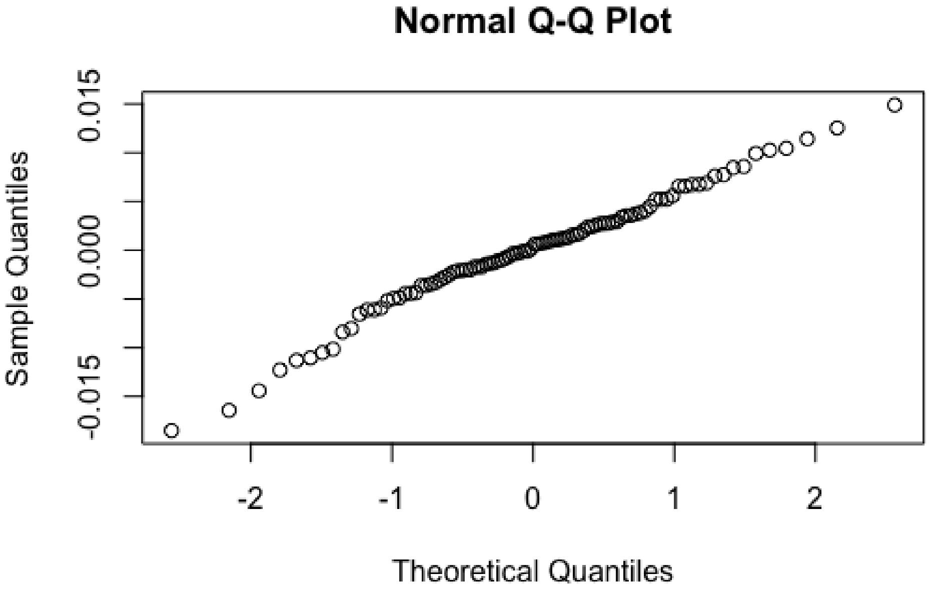
the small p-value of F test rejects the null hypothesis that all the predictors have no effect on FoHF, implying there is at least on significant factors among all these predictors that have impact.

R square represents goodness of fitting a linear regression on the data. The relatively large R-square in this case, 0.7798, implies linear regression may be a good fit for the FoHF data.

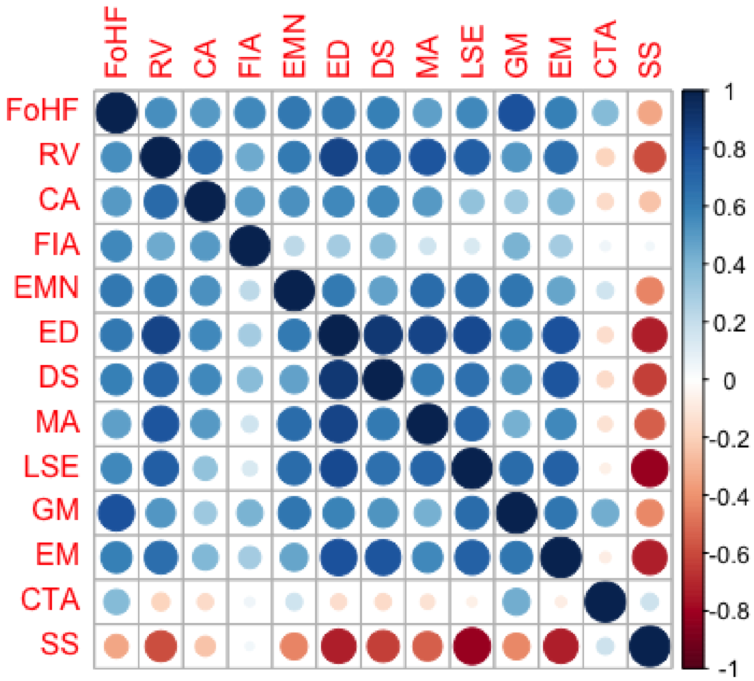
b



The TA plot shows an approximately horizontal line on zero, indicating the assumption of zero expected value for error is not violated. The width of residuals around zero line slightly increases as fitted value increases. This pattern implies the constant variance assumption might be violated. The distribution of residuals does not show particular pattern, therefore, assumption of independence is not violated.



QQplot shows approximately linear pattern, indicating the residuals follow normal distribution. Therefore the normality assumption of error is not violated.



```
> vif(fit1)
       RV        CA       FIA       EMN        ED        DS
 6.387024  2.982646  2.271113  3.672017 29.973694  9.404810
       MA       LSE        GM        EM       CTA        SS
 8.001994 10.046374  5.699120  4.255477  2.232320  4.972861
```

From the correlation matrix, we can perceive that predictor RV have strong correlation with other predictors such as ED, MA, LSE, EM and SS; ED is strongly correlated with DS, MA, LSE, EM, SS; DS is strongly correlated with EM.

The serious issue with multicollinearity can also be reflected on VIF values of predictors. ED, LSE, DS all have VIF values more than 5.

c

To solve the unequal variance problem, I will transform data of response variable via box-cox transformation. Multi-collinearity is the most serious problem. To fix this issue, I will keep ED and delete all other predictors that are strongly correlated with ED, such as RA, DS, MA, LSE, EM, SS.

d

Backward Selection

```
fit1 <- lm(formula=FoHF ~., data=FoHF)
```

```
step(object=fit1, direction=’backward’, k=log(nrow(FoHF)))


## result
Call:
lm(formula = FoHF ~ CA + FIA + ED + GM + CTA, data = FoHF)

Coefficients:
(Intercept)           CA          FIA           ED           GM          CTA
  -0.001857     0.175665     0.298492     0.265442     0.246942     0.153503
```

If start from full model and use BIC criterion, the correct model will contain predictors CA, FIA, ED, GM, CTA.

Forward Selection

```
fit2 <- lm(formula=FoHF ~ , data=FoHF)
step(object=fit2, direction=’forward’, k=log(nrow(FoHF)), scope=list(upper=fit1, lowwer=fit2))



## result
Call:
lm(formula = FoHF ~ GM + CA + FIA + CTA + ED, data = FoHF)

Coefficients:
(Intercept)           GM           CA          FIA          CTA           ED
  -0.001857     0.246942     0.175665     0.298492     0.153503     0.265442
```

If start from empty model and use BIC criterion, the correct model will contain GM, CA, FIA, CTA, ED.

All Subset Selection

```
regsub_out <- regsubsets(FoHF~., data=FoHF)
summary(regsub_out)$bic


Selection Algorithm: exhaustive
         RV  CA  FIA EMN ED  DS  MA  LSE GM  EM  CTA SS
1  ( 1 ) " " " " " " " " " " " " " " " " "*" " " " " " "
2  ( 1 ) " " "*" " " " " " " " " " " " " "*" " " " " " "
3  ( 1 ) " " " " "*" " " "*" " " " " " " " " " " "*" " "
4  ( 1 ) " " " " "*" " " "*" " " " " " " "*" " " "*" " "
5  ( 1 ) " " "*" "*" " " "*" " " " " " " "*" " " "*" " "
```

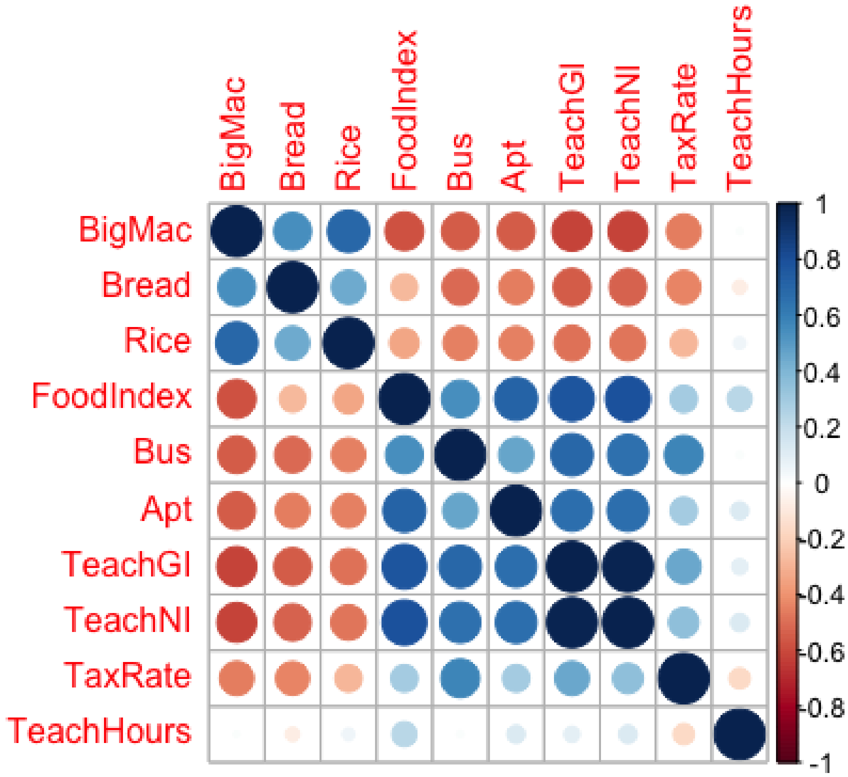
```
6  ( 1 ) "*" "*" "*" " " "*" " " " " " " "*" " " "*" " "
7  ( 1 ) "*" "*" "*" " " "*" " " " " "*" "*" " " "*" " "
8  ( 1 ) "*" "*" "*" " " "*" " " " " "*" "*" " " "*" "*"

> which(summary(regsub_out)$bic == min(summary(regsub_out)$bic))
[1] 5
```

The model that has minimum BIC is model 5, which contains CA, FIA, ED, GM, CTA.

conclusion: The procedure of forward, backward and all subset shows the best model based on BIC value is the model that contains predictors CA, FIA, ED, GM, CTA.

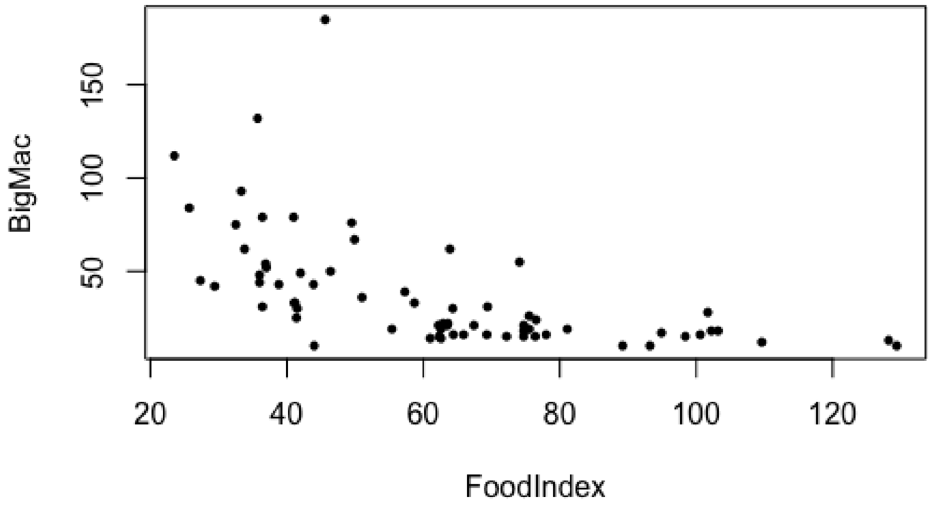
2

a



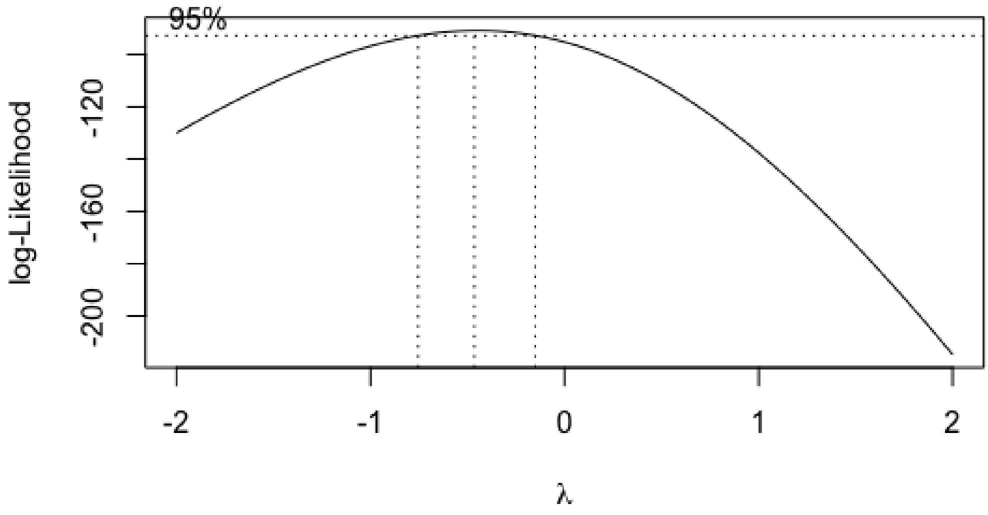
Base on the correlation matrix, predictor TeachNI has strong co-linearity with TeachGI (approximately 1). Strong multi-colinearity may exist between FoodIndex, Bus, Apt, TeachNI and TeachGI.

b

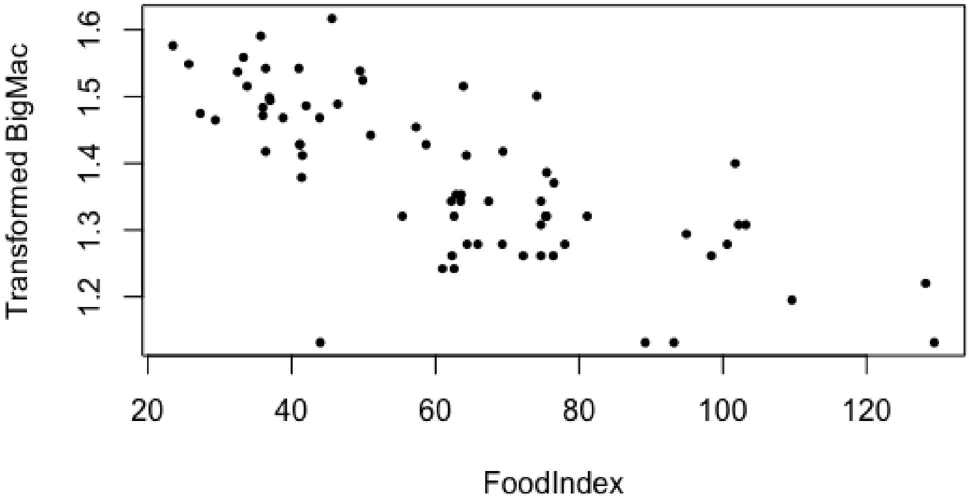


Without any data transformation, the scatterplot shows non-linear (more like exponential distribution) relationship between FoodIndex and BigMac.

c

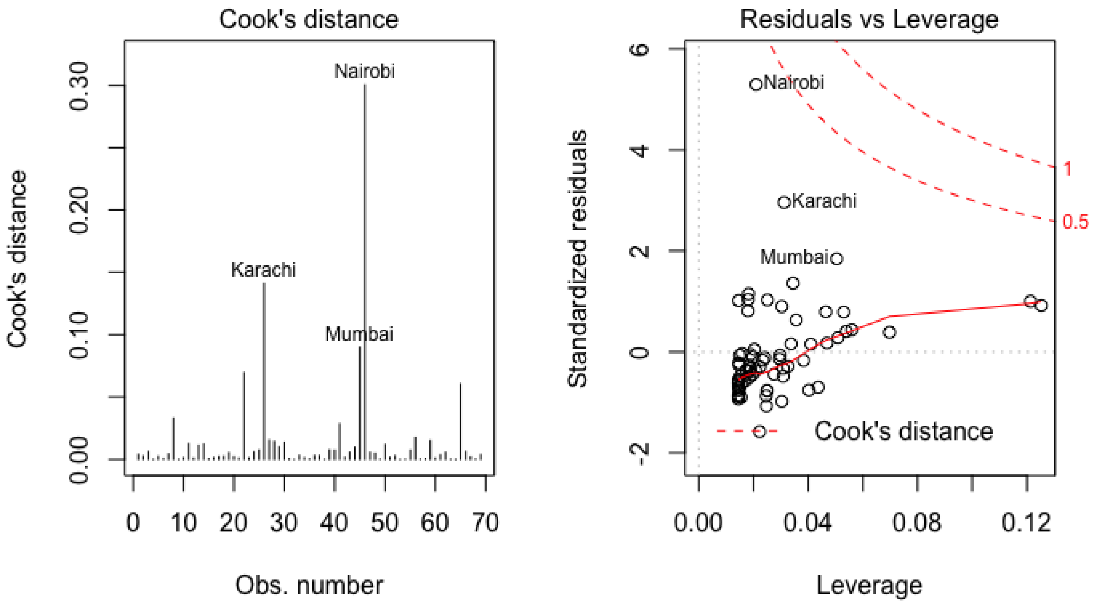


λ for maximum log-likelihood is equal to -0.4646, however, considering -0.5 also falls in 95% confidence interval, we can transform response variable $y \rightarrow \frac{1-log^{-1}(y)}{0.5}$



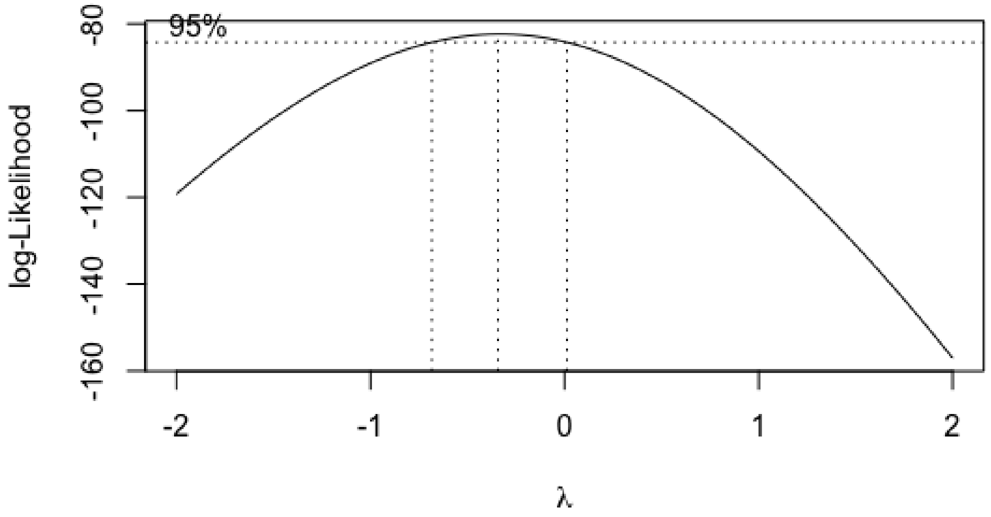
The scatter plot looks better as linear relation exists after transformation.

d

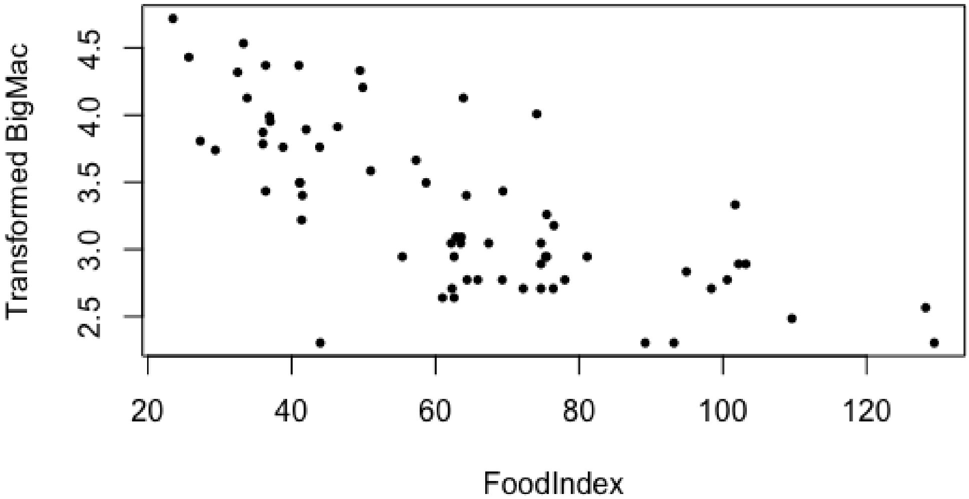


Base on Cook's distance, cities named as Nairobi (≈ 0.3) and Karachi (≈ 0.15) have largest influence on fitted line.

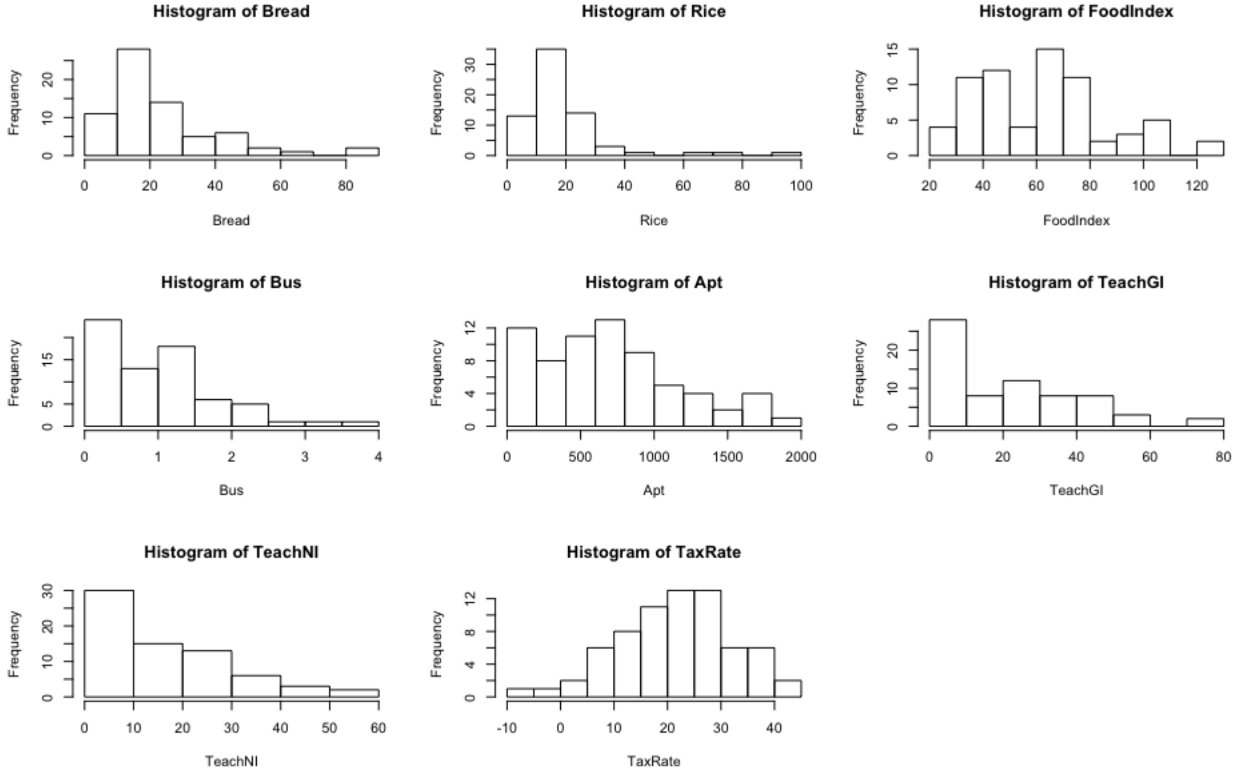
e



In the reduced dataset, λ for maximum log-likelihood changes to -0.3434. Since the 95% confidence for λ covers zero, we can transform response variable $y \to log(y)$.



f



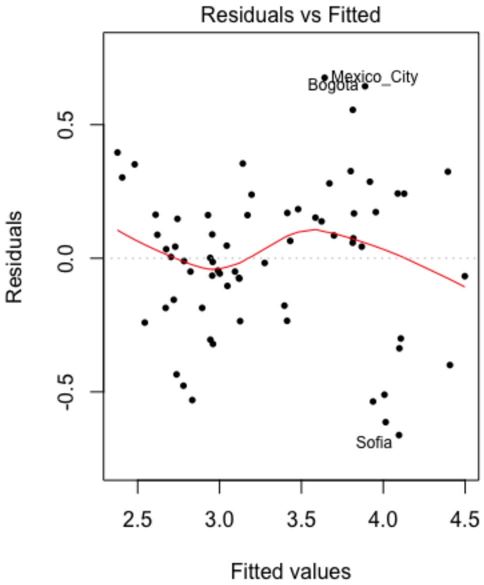
Distributions of Bread, Rice, Bus, Apt, TeachGI, TeachNI are right skewed.

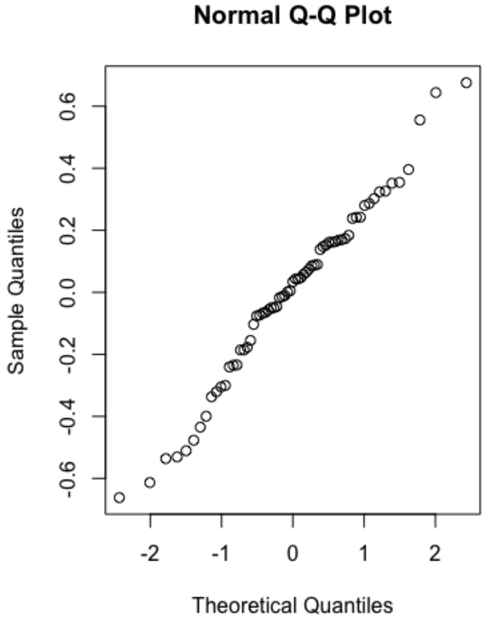
g

```
> c(loocv.lm(model1), loocv.lm(model2), loocv.lm(model3))
[1] 0.1984310 0.1353951 0.1008867
```

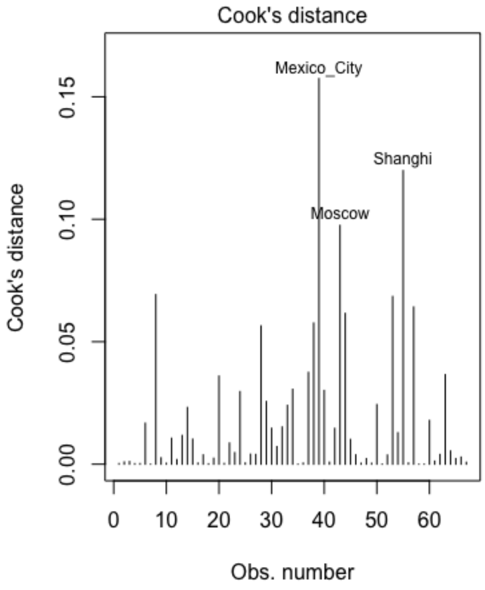
Base on LOOCV values of three models, model 3 which contains all predictors log(Bread), log(Rice), Apt, log(Bus), log(TeachNI) has the best score.

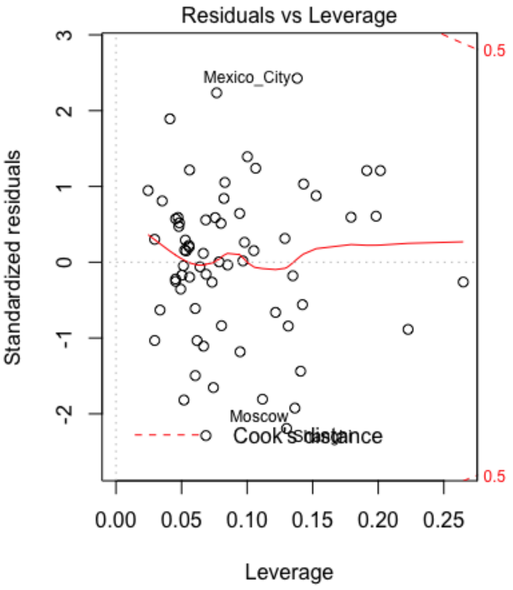
h





Normality, independence and zero mean assumptions of error are not violated, However, the variance of residuals is not constant along with fitted value. The assumption for constant error variance might be violated.





Cook's distance for all data points are below 0.15 after transformation of response variable. There is no effect from outliers.

3

(a) choose a)
since interaction of X_2 and X_3, X_4 and X_5 are significant, by hierarchy principle, theses main factors should be included in model.

(b) choose d)
if model exclude X_2, X_3, X_6, AIC will inflate, therefore these three factors should be included. Adding any other predictors will slightly inflate AIC, therefore, none of the others should be added.

(c) choose d)
information of sample size is missing, therefore cannot compare BIC scores with AIC scores.

(d) choose c)
M1 which contains none of the predictors has worst AIC score.

(e) choose d)
adding predictors will increase goodness of fit.

4

(a) choose a)
result of global F-test cannot be used to conclude all predictors are significant.

(b) choose b)
p-value and confidence interval are complementary.

(c) choose b)

(d) choose c)
qqplot can check normality assumption of error.

(e) choose b)
larger sample size leads to smaller standard error and larger test statistics, thereby having a smaller p-value.