5291 hw3

Yijin Wang

1.a) Fit a multiple linear regression model of predict ‘glu’, plasma glucose concentration in an oral glucose tolerance test, using the following set of predictors: ‘npreg’ number of pregnancies ‘bp’ diastolic blood pressure (mm Hg) ‘skin’ triceps skin fold thickness (mm) ‘bmi’ body mass index (weight in kg/(height in m)^2) ‘age’ age in years

```
library(MASS)
data(Pima.te)
head(Pima.te)
```

```
##   npreg glu bp skin  bmi   ped age type
## 1     6 148 72   35 33.6 0.627  50  Yes
## 2     1  85 66   29 26.6 0.351  31   No
## 3     1  89 66   23 28.1 0.167  21   No
## 4     3  78 50   32 31.0 0.248  26  Yes
## 5     2 197 70   45 30.5 0.158  53  Yes
## 6     5 166 72   19 25.8 0.587  51  Yes
```

```
mlr<-lm(glu ~ npreg + bp + skin + bmi + age, data=Pima.te)
mlr
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + age, data = Pima.te)
##
## Coefficients:
## (Intercept)        npreg           bp         skin          bmi          age
##     56.8314      -0.8753       0.1039       0.2626       0.7958       0.7638
```

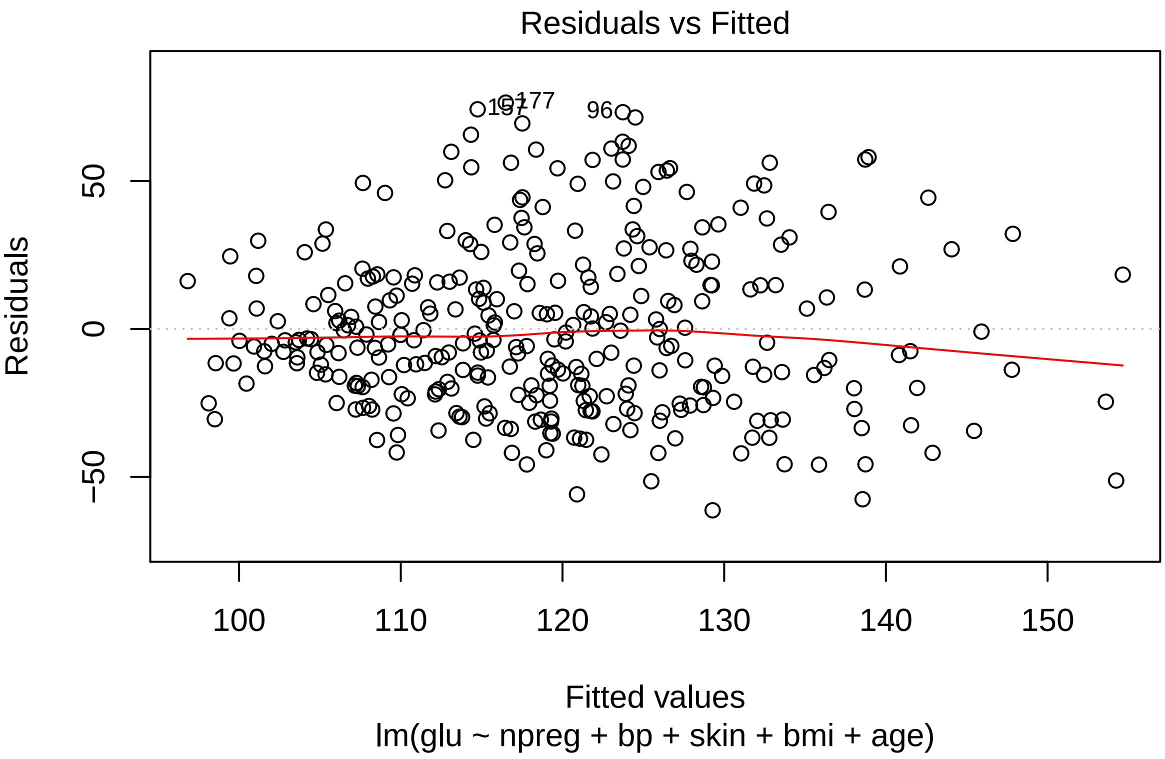
```
summary(mlr)
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + age, data = Pima.te)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.285 -20.556  -4.356  17.370  76.509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.8314    10.3090   5.513 7.19e-08 ***
## npreg        -0.8753     0.6475  -1.352  0.17735
## bp            0.1039     0.1385   0.750  0.45353
## skin          0.2626     0.2164   1.214  0.22575
## bmi           0.7958     0.3020   2.636  0.00880 **
## age           0.7638     0.2068   3.693  0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.6 on 326 degrees of freedom
## Multiple R-squared:  0.1338, Adjusted R-squared:  0.1205
## F-statistic: 10.07 on 5 and 326 DF,  p-value: 5.575e-09
```

Based on the summary, there are only two variables whose p value is smaller than 5% and the adjusted r-square is only 0.1205, which indicates the invalidity of the multiple linear regression model.

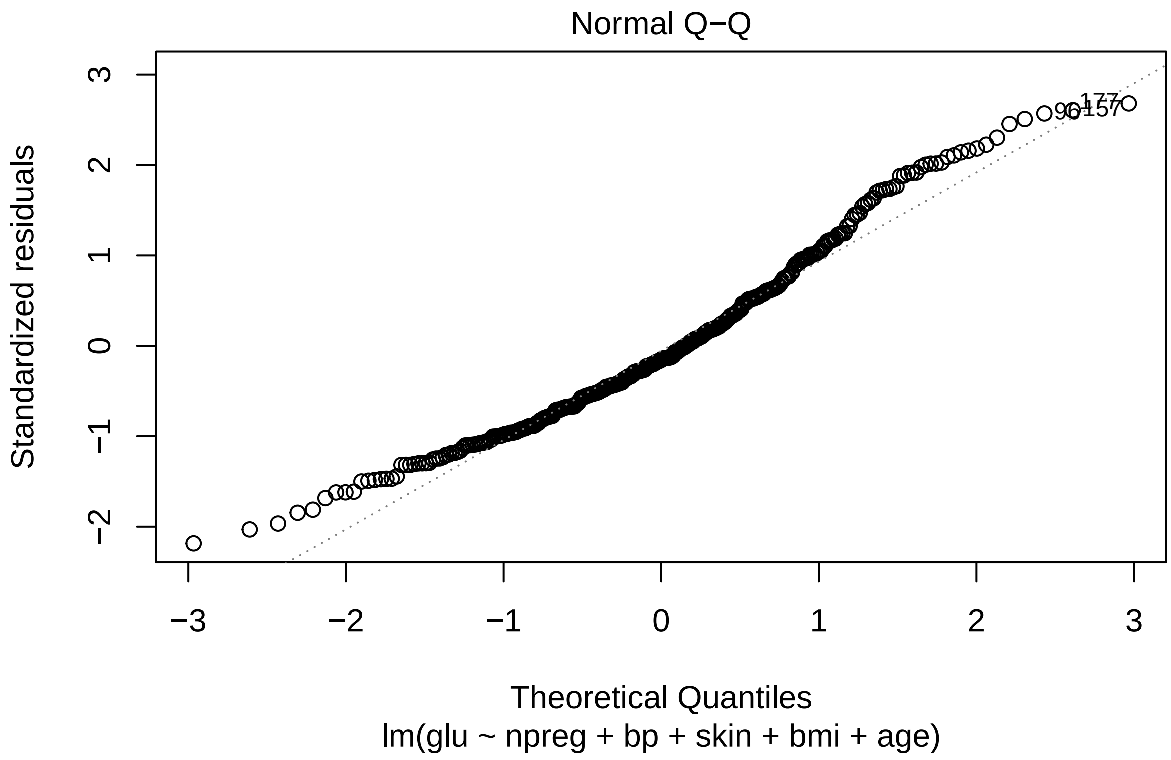
b) State and assess the validity of the underlying assumptions: Linearity/functional form, including the need for any interaction terms Normality Homoscedasticity Uncorrelated error, and Check for outliers and influential points.

```
#Linearity/functional form
plot(mlr,1)
```



```
#Data points are randomly scattered in the plot, so there is linearity.

#Normality
plot(mlr, 2)
```



```
#According to the qqplot, the residuals are not normally distributed.
shapiro.test(residuals(mlr))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mlr)
## W = 0.97032, p-value = 2.532e-06
```

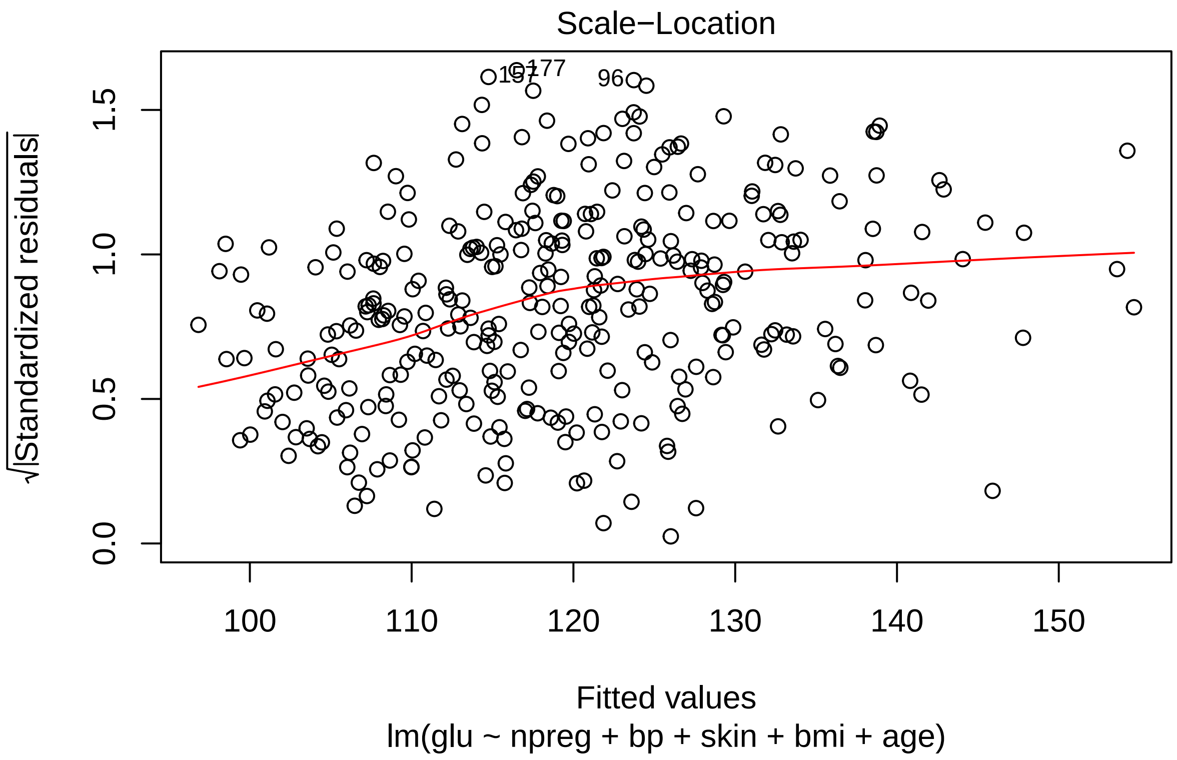
```
#According to shapiro test, the p-value is smaller than 0.05,
#so we should reject the null hypothesis.The residuals are not normally distributed.

#Homoscedasticity
plot(mlr,3)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```



```r
bptest(mlr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mlr
## BP = 19.578, df = 5, p-value = 0.0015
```

```r
#According to the plot and bptest, we conclude that
#there is no homoscedasticity and
#constant variance assumption is invalid

#Uncorrelated error
#install.packages("car")
library(car)
```

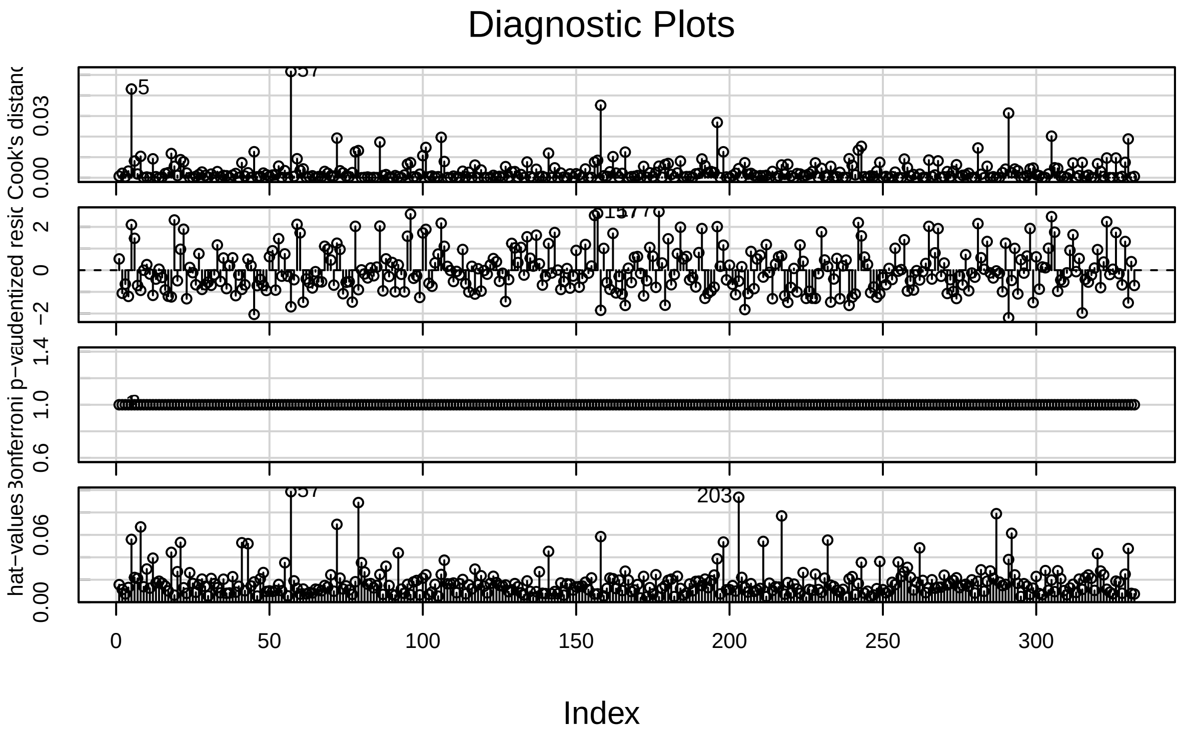
```
## Loading required package: carData
```

```r
durbinWatsonTest(mlr)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.029888      1.937881   0.596
##  Alternative hypothesis: rho != 0
```

```r
#According to the dubin-watson test, the p value is greater than 0.05,
#so we fail to reject null hypothesis.
#We conclude that there is no correlation among residuals.
#Thus, the errors are generally uncorrelated.

#Outliers and influential points
infIndexPlot(mlr)
```



```r
outlierTest(mlr)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 177 2.706896          0.0071504           NA
```

```r
cook<-cooks.distance(mlr)
#According to the influential plot, there are influential plots.
#However,the cook's distance suggest that they are not influential.
#According to the outlier test, the 177th data point might be an outlier.
#However, the p value suggests it is not an outlier.
```

c) Propose remedial measures in case of violations of any of the underlying assumptions 1.Linearity Non-linear model/Simple Transfomation

2.Normality Transfomation/Robust regression methods

3.Homoscedasticity Transformation

4.Uncorrelated error Transformation : Cochrane-Orcutt Procedure

5.Outliers and influential points remove outliers and influential points

2) Repeat (a) using Least Median of Squares Regression and compare the results with those obtained in (a).

```r
lmsr<-lmsreg(glu ~ npreg + bp + skin + bmi + age, data=Pima.te)
lmsr
```

```
## Call:
## lqs.formula(formula = glu ~ npreg + bp + skin + bmi + age, data = Pima.te,
##     method = "lms")
##
## Coefficients:
```

```
## (Intercept)       npreg          bp        skin         bmi         age
##    79.46184    -0.35625    -0.08628     0.88019    -0.18894     0.36836
##
## Scale estimates 24.99 24.79
```

```
#The coefficients from two methods are significantly different.
```