5291 - HW3

Hongyu Ji - hj2475

9/26/2018

Problem a

```
# loading data
library(MASS)
attach(Pima.te)
```

```
lm1 <- lm(glu ~ npreg + bp + skin + bmi + age )
summary(lm1)
```

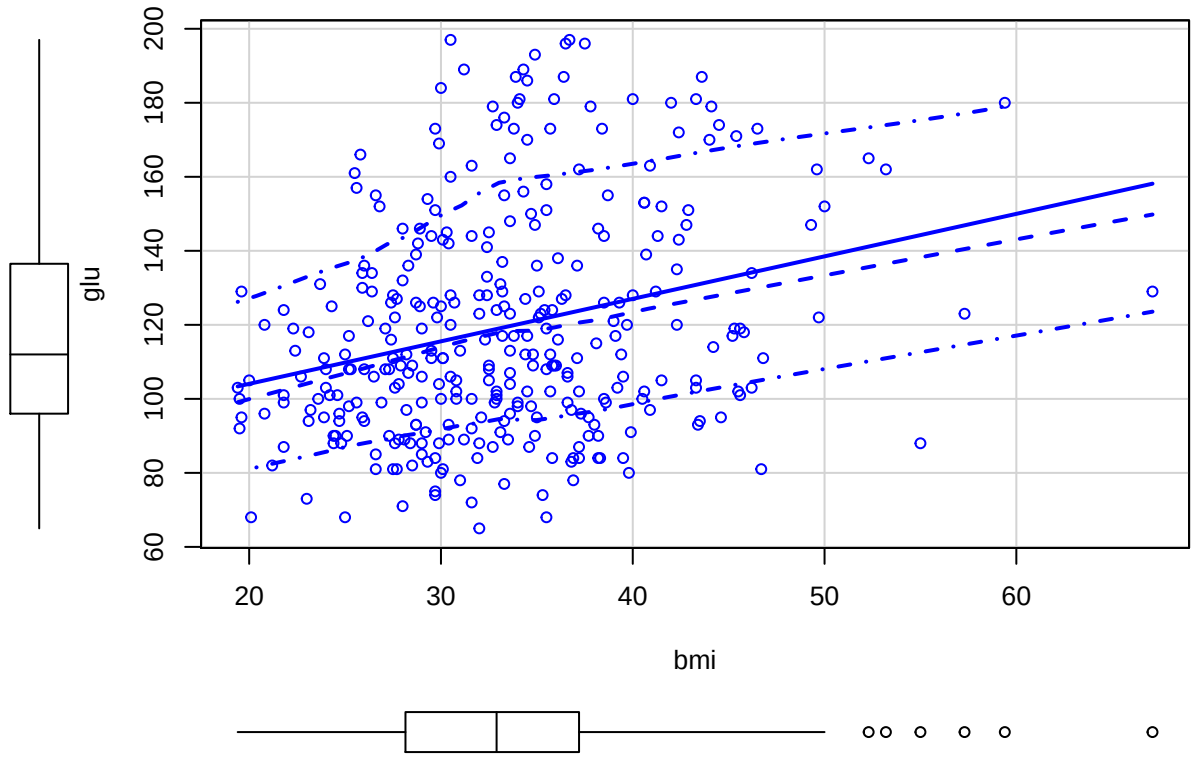
```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.285 -20.556  -4.356  17.370  76.509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.8314    10.3090   5.513 7.19e-08 ***
## 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
```

From the linear regression model we can see that bmi and age are the significant variables since their p-values are smaller than 0.05.

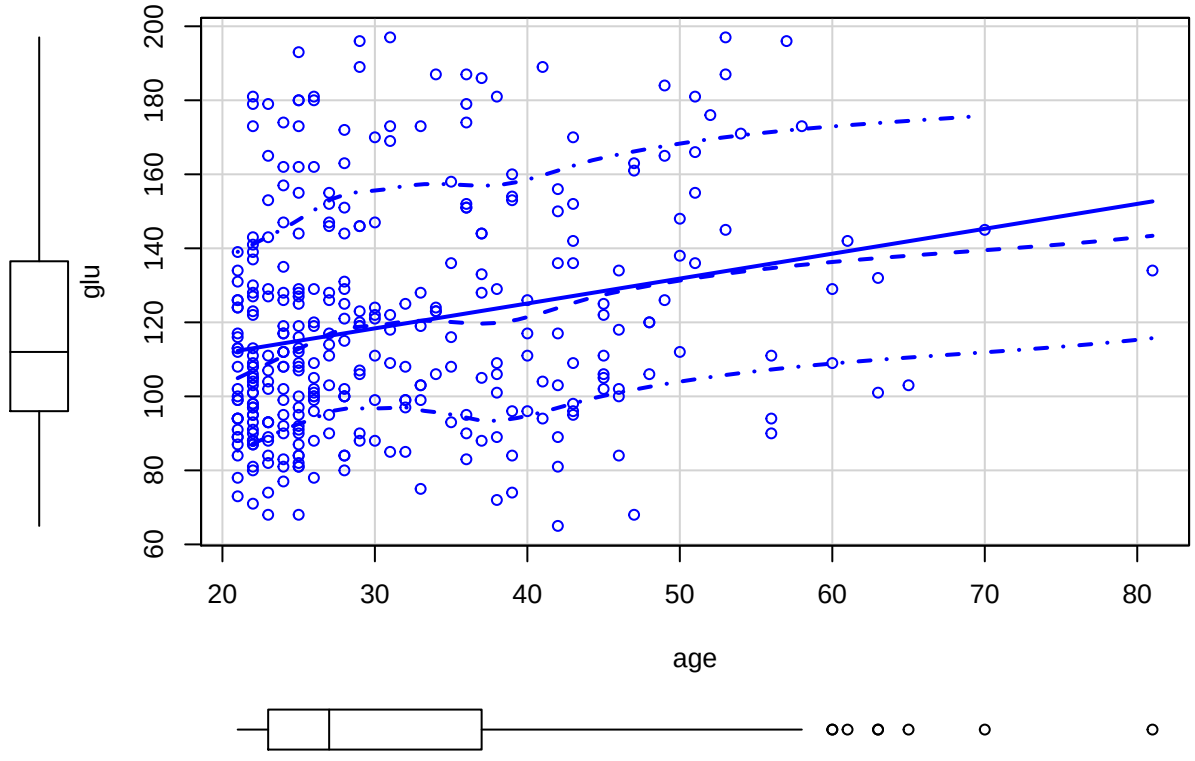
Problem b

Check linearity/funtional form

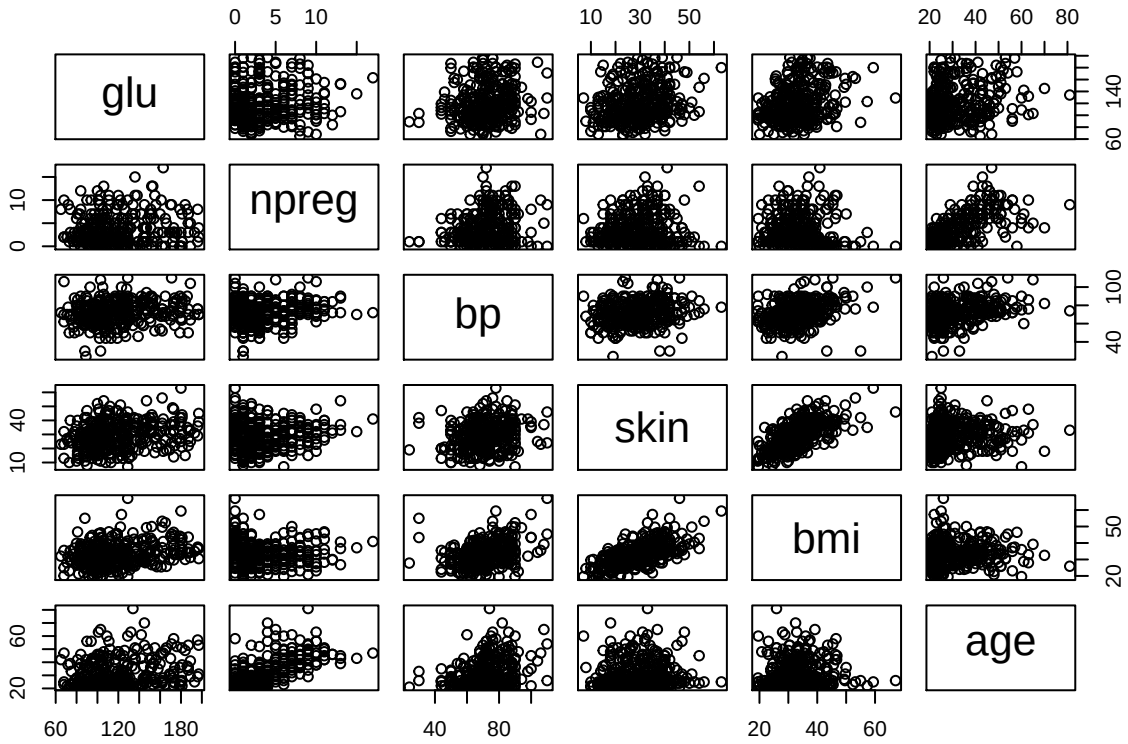
```
library(car)
scatterplot(glu ~ bmi)
```




```
scatterplot(glu ~ age)
```




```
pairs(glu ~ npreg + bp + skin + bmi + age)
```



From the scatterplot we can see that there are no strong linear relationship between dependent variable glu and other independent variables. Also the R^2 is 13.38% which shows weak association.

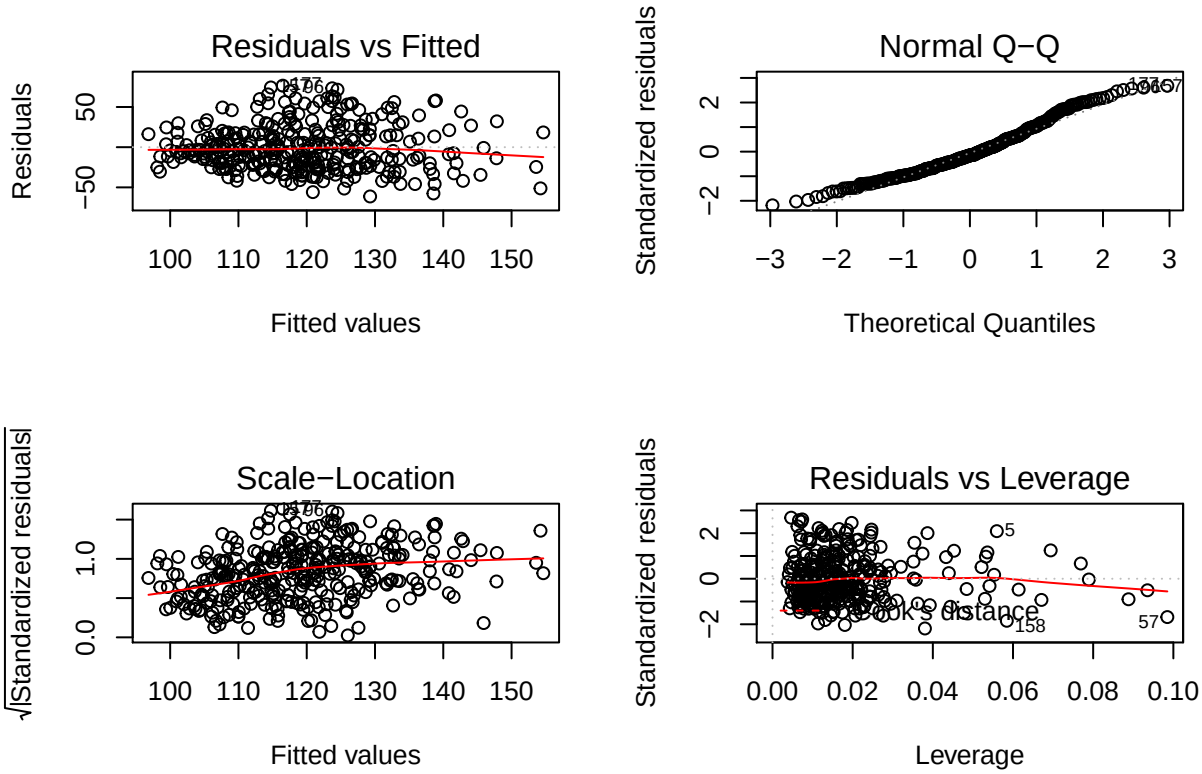
Check Normality

```
#pi <- Pima.te
#c1 <- c(1:4, 7); c2 <- c(1:5,7)
#pi[c1] <- apply(pi[c1], 2, as.numeric)
#apply(pi[c2], 2, shapiro.test)
shapiro.test(lm1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm1$residuals
## W = 0.97032, p-value = 2.532e-06
```

From the shapiro test, the p-value is smaller than 0.05, thus we have enough evidence to reject the null hypothesis that it is normally distributed.

```
par(mfrow=c(2,2))
plot(lm1)
```



Also from the Q-Q plot, points are deviated from the line in the beginning and show a curve at the tail, thus it looks like not normally distributed.

Check for homoscedasticity

```
ncvTest(lm1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 15.20531, Df = 1, p = 9.6432e-05
```

From the residual plots and the test (p-value is smaller than 0.05), we have enough evidence to reject the null hypothesis. Thus it is Non-constant (not homoscedasticity)

Check for uncorrelated errors

```
library(lmtest)
dwtest(lm1)
```

```
##
##  Durbin-Watson test
##
## data:  lm1
## DW = 1.9379, p-value = 0.2847
## alternative hypothesis: true autocorrelation is greater than 0
```

p-value is larger than 0.05, thus we do not have enough evidence to reject the null hypothesis. we can have the conclusion with errors are uncorrelated.

Check for outliers and influential points

```
# check outleirs
lmi <- lm.influence(lm1)
lms <- summary(lm1)
e <- resid(lm1)
s <- lms$sigma
si <- lmi$sigma
xxi <- diag(lms$cov.unscaled)
h <- lmi$hat
```

```
bi <- coef(lm1)-t(coef(lmi))
dfbetas <- bi/t(si%o%xxi^0.5)
stand.resid <- e/(s*(1-h)^0.5)
student.resid <- e/(si*(1-h)^0.5)
DFFITS <- h^0.5*e/(si*(1-h))
```

```
outlierTest(lm1)
```

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

```
# influencePlot(fit1)

all(dffits(lm1) < 1)
```

```
## [1] TRUE
```

```
all(abs(dfbetas(lm1)) < 1)
```

```
## [1] TRUE
```

```
head(sort(cooks.distance(lm1), decreasing = T))
```

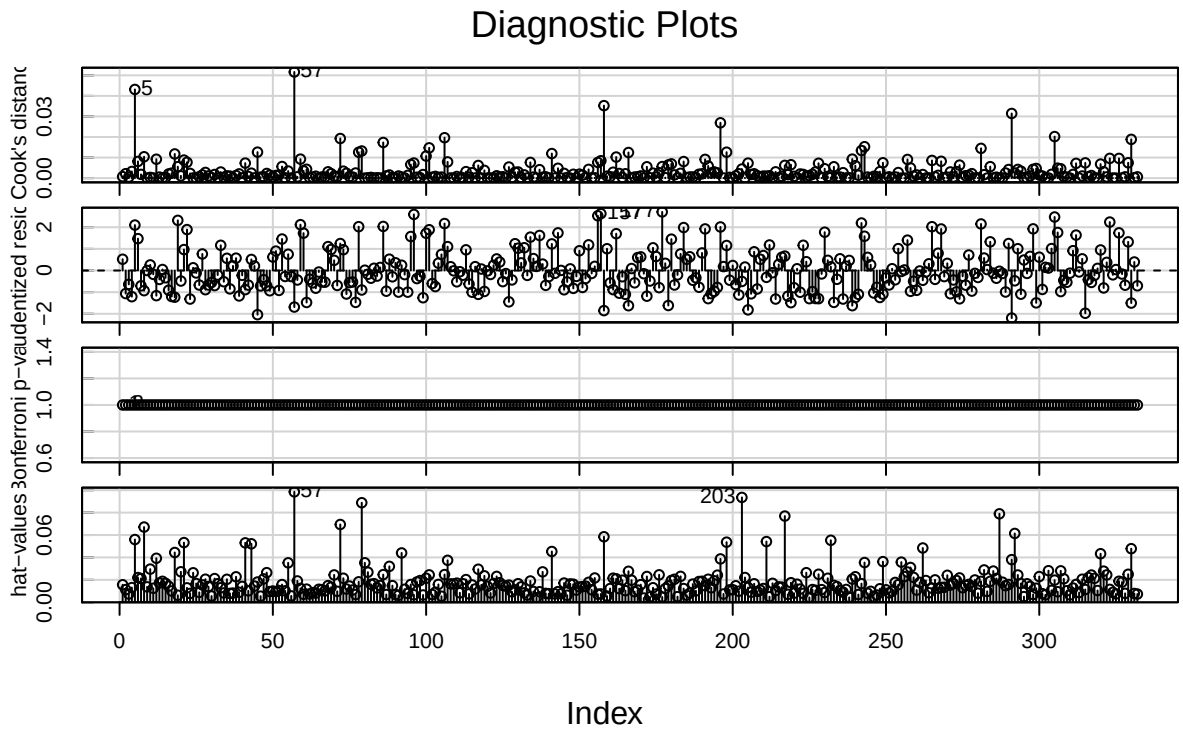
```
##         57          5        158        291        196        305
## 0.05163723 0.04313546 0.03526009 0.03143735 0.02690254 0.02020202
```

```
qf(0.2, 6, 326)
```

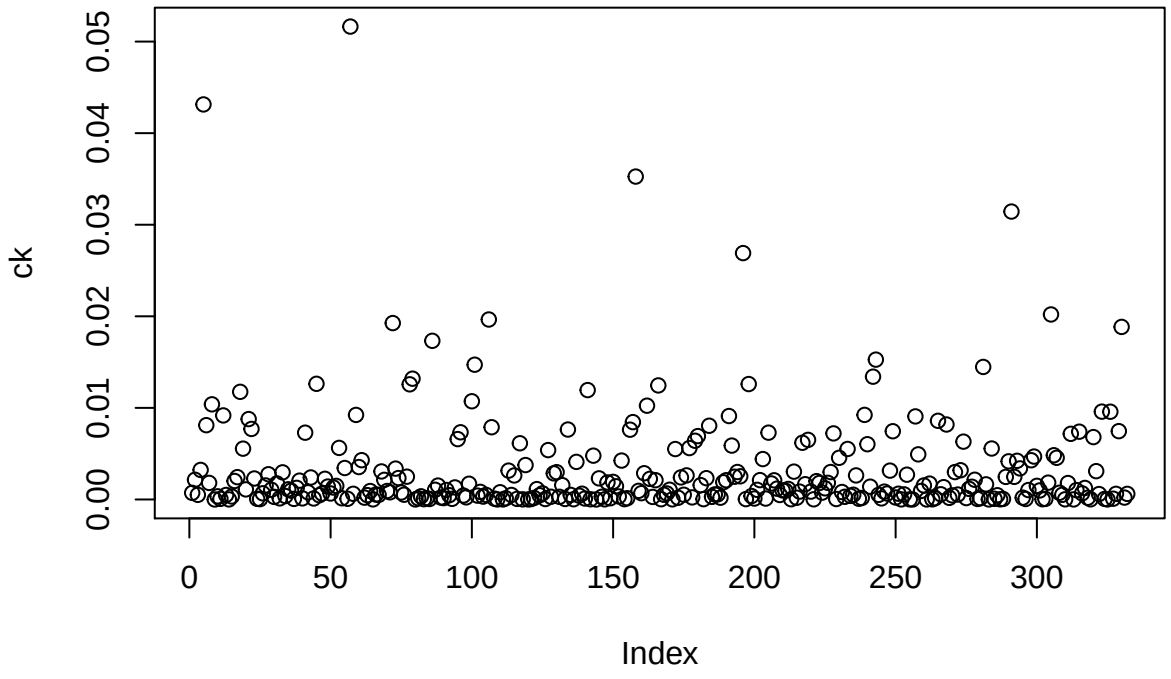
```
## [1] 0.5109577
```

No outliers detected from the test since p-value is not samller than 0.05; we don't have enough evidence to reject the null hypothesis.

```
influenceIndexPlot(lm1)
```




```
ck <-cooks.distance(lm1)
plot(ck)
```



Also from the plots we detect no influential points.

Problem c

Remedy for Linearity/Functional form: start from simple transformation such as log, square-root, or use box-cox. Otherwise, we can use non-linear model

Remedy for Normality: we can use transformation or use other robust regression methods

Remedy for Homoscedasticity: we can use transformation or use weighted least squares to give each data point different weight to maximize the efficiency of parameter estimation

Remedy for uncorrelated error: we can do transformation using Cochrane-Orcutt estimation. We can also use some models that incorporate correlation structures, such as generalized estimating equations.

Remedy for outliers and influential points: we can use robust regression, or delete the outliers.

Problem d

```
set.seed(123)
lmsreg(glu ~ npreg + bp + skin + bmi + age )
```

```
## Call:
## lqs.formula(formula = glu ~ npreg + bp + skin + bmi + age, method = "lms")
##
## Coefficients:
## (Intercept)        npreg           bp         skin          bmi
##    43.58028      2.06788      0.28193      0.09056      0.73620
##         age
##     0.40583
##
## Scale estimates 24.56 25.99
```

Each coefficients obtained here is different from the ones in a. Depends on set.seed will have different results.