ADA HW5

Minqian Guo(mg3418)

10/09/2014

Prob a. Fit a multiple linear regression model

```
require(MASS)
```

```
## Loading required package: MASS
```

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

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.29 -20.56  -4.36  17.37  76.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.831     10.309    5.51  7.2e-08 ***
## npreg         -0.875      0.647   -1.35  0.17735
## bp             0.104      0.138    0.75  0.45353
## skin           0.263      0.216    1.21  0.22575
## bmi            0.796      0.302    2.64  0.00880 **
## age            0.764      0.207    3.69  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:  0.134,  Adjusted R-squared:  0.121
## F-statistic: 10.1 on 5 and 326 DF,  p-value: 5.58e-09
```

Prob b. The underlying assumptions check

1. Nonlinearity: Using R-squared value in the result of linear regression.

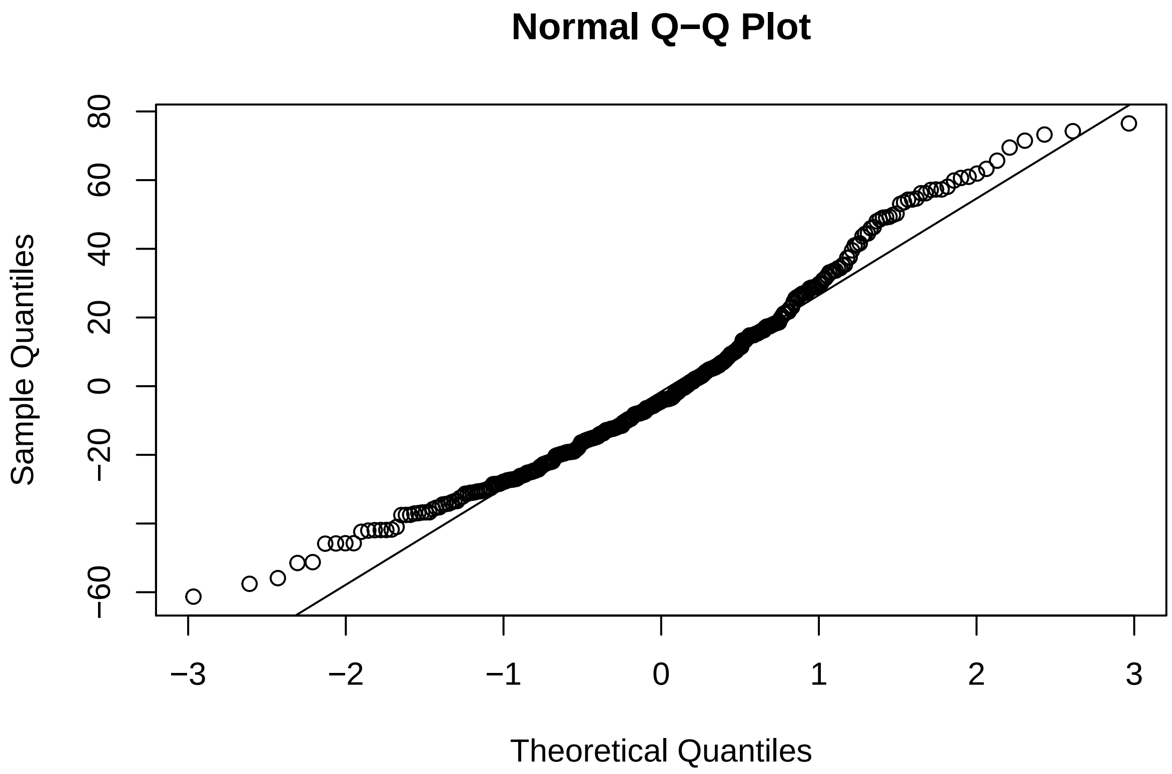
```
summary(fit)$r.squared
```

```
## [1] 0.1338
```

Since the R-squared is small, which means the lack of fit of fitted model.

2. Normality

```
qqnorm(fit$residuals)
qqline(fit$residuals)
```



The QQ plot shows the assumption of normality is invalid.

```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.9703, p-value = 2.532e-06
```

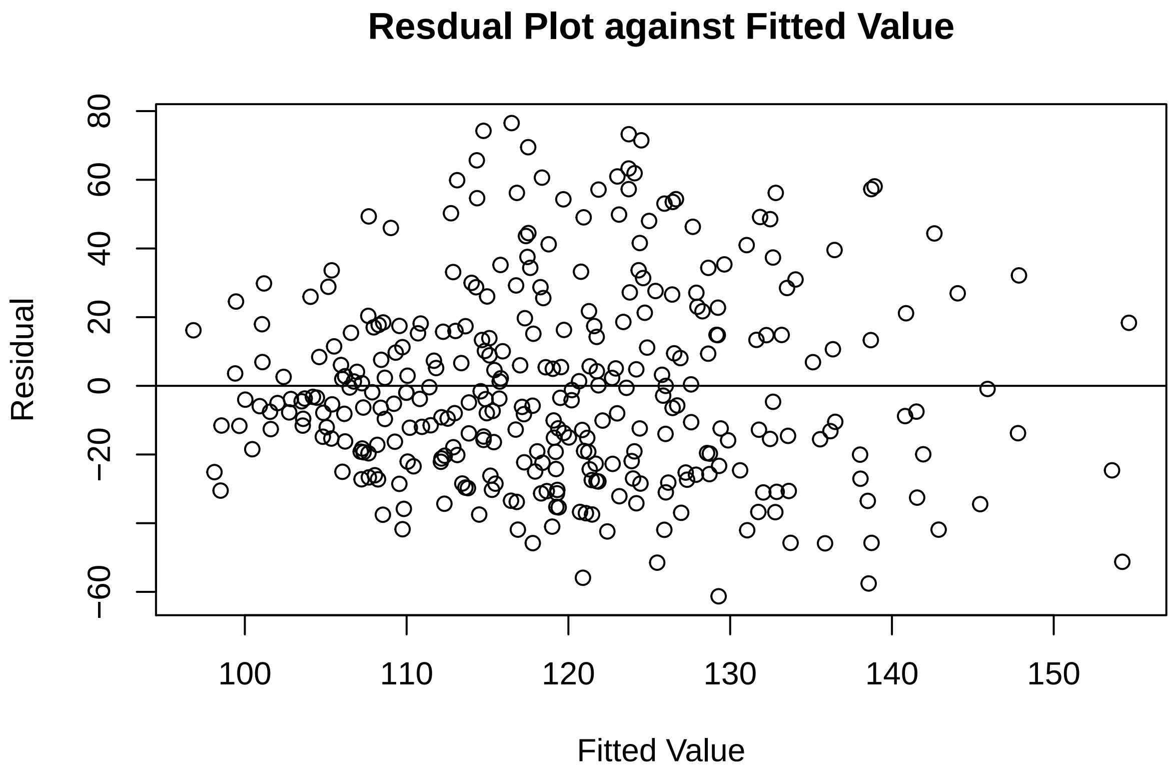
```
shapiro.test(fit$residuals)$p.value
```

```
## [1] 2.532e-06
```

By using Shapiro-Wilk test to check normality, Since the p-value is significantly small, we reject the null hypothesis that the distribution of the error is not normal.

3. Homoscedasticity

```
plot(fit$fitted.values,fit$residuals,xlab='Fitted Value',ylab='Residual',main="Resdual Plot against Fit
abline(h=0)
```



The plot of the residuals against the fitted values shows a scatter plot without certian pattern. Thus the assumption of constant error is valid.

4. Uncorrelation of Error

```
require(lmtest)
```

```
## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.0.3

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

```r
dwtest(fit)
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.938, p-value = 0.2847
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
dwtest(fit)$p.value
```

```
## [1] 0.2847
```

By using Durbin-watson test to do correlation check of error, Since p-value is greater than 0.05, we should accept the null hypothesis. So, the assumption of uncorrelated error is valid.

5. Outliers par(mfrow=c(2,2)) plot(fit) From Residuals VS Fitted and QQ plot, we further verify the conclusion that there

exist outliers and got the index of them, which are 8,101,179

6.. Influential Points

```r
lmi <- lm.influence(fit)
lms <- summary(fit)
e <- resid(fit)
s <- lms$sigma
si <- lmi$sigma
xxi <- diag(lms$cov.unscaled)
h <- lmi$hat
bi <- coef(fit)-t(coef(lmi))
dfbetas <- bi/t(si%o%xxi^0.5)
stand.resid <- e/(si*(1-h)^0.5)
DFFITS <- h^0.5*e/(si*(1-h))
which(abs(stand.resid)>2*sqrt(6/332))
```

```
##   1   2   3   4   5   6   7   8  12  13  15  16  17  18  19  20  21  22
##   1   2   3   4   5   6   7   8  12  13  15  16  17  18  19  20  21  22
##  23  26  27  28  29  30  31  33  34  35  36  38  39  41  42  43  45  46
##  23  26  27  28  29  30  31  33  34  35  36  38  39  41  42  43  45  46
##  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
##  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
##  66  67  68  69  70  71  72  73  74  75  76  77  78  79  82  86  87  88
##  66  67  68  69  70  71  72  73  74  75  76  77  78  79  82  86  87  88
##  89  90  91  94  95  96  97  99 100 101 102 103 104 105 106 107 110 113
##  89  90  91  94  95  96  97  99 100 101 102 103 104 105 106 107 110 113
## 114 115 117 119 123 124 125 127 128 129 130 131 132 134 135 137 138 139
## 114 115 117 119 123 124 125 127 128 129 130 131 132 134 135 137 138 139
## 140 141 143 145 146 148 149 150 151 152 153 156 157 158 159 160 161 162
## 140 141 143 145 146 148 149 150 151 152 153 156 157 158 159 160 161 162
## 163 164 165 166 168 169 170 172 173 174 175 176 177 178 179 180 181 183
## 163 164 165 166 168 169 170 172 173 174 175 176 177 178 179 180 181 183
## 184 185 186 187 188 189 190 191 192 193 194 195 196 198 199 201 202 203
## 184 185 186 187 188 189 190 191 192 193 194 195 196 198 199 201 202 203
## 205 206 207 208 209 210 211 212 214 215 216 217 218 219 220 222 223 224
## 205 206 207 208 209 210 211 212 214 215 216 217 218 219 220 222 223 224
## 225 226 227 228 230 231 233 234 235 236 238 239 240 241 242 243 244 246
## 225 226 227 228 230 231 233 234 235 236 238 239 240 241 242 243 244 246
## 247 248 249 250 251 253 254 257 258 259 260 262 264 265 266 267 268 269
## 247 248 249 250 251 253 254 257 258 259 260 262 264 265 266 267 268 269
## 270 271 272 273 274 276 277 278 280 281 282 284 286 289 290 291 292 293
## 270 271 272 273 274 276 277 278 280 281 282 284 286 289 290 291 292 293
## 294 295 297 298 299 300 301 304 305 306 307 308 309 311 312 314 315 316
## 294 295 297 298 299 300 301 304 305 306 307 308 309 311 312 314 315 316
## 317 320 321 322 323 326 328 329 330 331 332
## 317 320 321 322 323 326 328 329 330 331 332
```

Prob c. The remedial measures in case of violations of any of the underlying assumptions

1. Nonlinearity:

Simple trasformations, e.g, take log

Non-linear model

Other predictors

2. Non-Normality:

Transformation

Robust regression methods

3. Influential points:

Robust regression