ADA Homework 5 Wenxin Liang wl2455

Consider the Pima.te dataset, in R library MASS, on Diabetes in Pima Indian Women.

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

Run the R code followed, we obtain,

> mylm <- lm(glu ~ npreg+bp+skin+bmi+age,data = data )

> summary(mylm)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age, data = data)

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

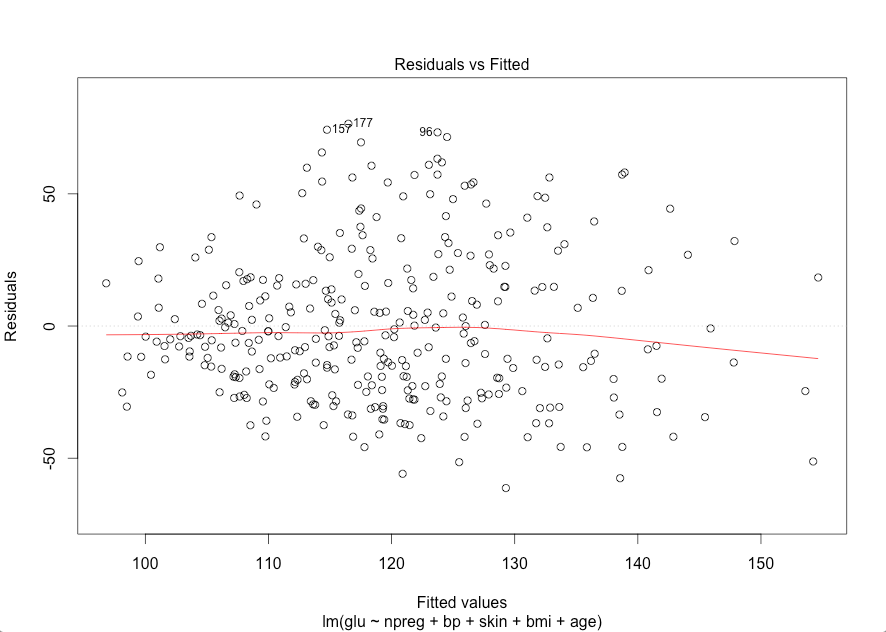
Multiple R-squared: 0.1338, Adjusted R-squared: 0.1205

F-statistic: 10.07 on 5 and 326 DF, p-value: 5.575e-09

The fitted linear model can be,

b) State and assess the validity of the underlying assumptions:

–Linearity/functional form, including the need for any interaction terms



Also based on the plot of the studentized residuals against the fitted values, there is although the points may gathered over the horizontal line at 0 since there is no obvious pattern then we can treat it as linearity.

For the interaction term, first we test whether or not we need the interaction term or based on the summary of mylm which do not have the interaction term with R-square equal to 0.1205 means if we add these predictor variables to our model, we only reduce the SSE(error sum of square) by 12.05% when compared to the model with no predictors, then we need interaction terms. Moreover, we can use the test followed to verify.

> library(phia)

> testInteractions(mylm)

F Test:

P-value adjustment method: holm

Value Df Sum of Sq F Pr(>F)

Mean 119.26 1 4721942 5771.1 < 2.2e-16 \*\*\*

Residuals 326 266733

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The null hypothesis here is H0:The coefficients of interaction terms are all zero. Since the p-value = 2.2e-16, thus there should be interaction terms existing.

Then we put all the interaction terms related two predict variables into the linear regression model. Since the interaction terms with more than two predict variables usually do not have so many influence to the linear model then we ignore them.

> mylm2 <- lm(glu ~ npreg+bp+skin+bmi+age+npreg\*bp+npreg\*skin+npreg\*bmi+npreg\*age+bp\*skin+bp\*bmi+bp\*age+skin\*bmi+skin\*age+bmi\*age,data = data )

> summary(mylm2)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + npreg \* bp +

npreg \* skin + npreg \* bmi + npreg \* age + bp \* skin + bp \*

bmi + bp \* age + skin \* bmi + skin \* age + bmi \* age, data = data)

Residuals:

Min 1Q Median 3Q Max

-63.424 -19.930 -4.356 19.575 75.418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.679e+01 4.855e+01 -0.346 0.7298

npreg 2.116e+00 5.318e+00 0.398 0.6909

bp 7.623e-01 6.614e-01 1.153 0.2500

skin 1.504e+00 1.384e+00 1.087 0.2780

bmi 5.571e-01 1.702e+00 0.327 0.7437

age 2.924e+00 1.631e+00 1.793 0.0740 .

npreg:bp -2.026e-02 5.869e-02 -0.345 0.7302

npreg:skin -1.323e-02 9.023e-02 -0.147 0.8835

npreg:bmi -2.978e-02 1.130e-01 -0.263 0.7924

npreg:age 7.744e-04 5.998e-02 0.013 0.9897

bp:skin 8.502e-03 1.829e-02 0.465 0.6424

bp:bmi -3.677e-03 1.813e-02 -0.203 0.8394

bp:age -2.441e-02 1.714e-02 -1.424 0.1555

skin:bmi -1.341e-02 2.034e-02 -0.659 0.5102

skin:age -4.546e-02 2.583e-02 -1.760 0.0793 .

bmi:age 3.097e-02 3.639e-02 0.851 0.3954

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.58 on 316 degrees of freedom

Multiple R-squared: 0.1615, Adjusted R-squared: 0.1217

F-statistic: 4.058 on 15 and 316 DF, p-value: 8.889e-07

Since the p-value of the interaction term skin\*age less than 0.1 then we conclude that the linear regression model should include the interaction term skin\*age. We verified by the following linear model.

> mylm3 <- lm(glu ~ npreg+bp+skin+bmi+age+skin\*age,data = data )

> summary(mylm3)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + skin \* age,

data = data)

Residuals:

Min 1Q Median 3Q Max

-64.849 -20.820 -4.357 17.453 75.701

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 28.66509 16.07608 1.783 0.07550 .

npreg -0.70926 0.64750 -1.095 0.27416

bp 0.08158 0.13795 0.591 0.55466

skin 1.38493 0.53847 2.572 0.01056 \*

bmi 0.73039 0.30143 2.423 0.01593 \*

age 1.78530 0.49410 3.613 0.00035 \*\*\*

skin:age -0.03614 0.01590 -2.273 0.02366 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.42 on 325 degrees of freedom

Multiple R-squared: 0.1474, Adjusted R-squared: 0.1316

F-statistic: 9.362 on 6 and 325 DF, p-value: 1.76e-09

We observe the value of R-square increase 0.0111 which shows it is useful to add the interaction term skin\*age. We check if we add the interaction terms have second and third small p-value to see the R-square.

> mylm4\_1 <- lm(glu ~.+skin\*age+bp\*age,data=data)

> summary(mylm4\_1)

Call:

lm(formula = glu ~ . + skin \* age + bp \* age, data = data)

Residuals:

Min 1Q Median 3Q Max

-61.52 -20.00 -4.30 18.24 75.77

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -24.73944 32.36098 -0.764 0.445136

npreg -0.79329 0.64643 -1.227 0.220643

bp 0.80056 0.40276 1.988 0.047690 \*

skin 1.33841 0.53688 2.493 0.013169 \*

bmi 0.74008 0.30027 2.465 0.014231 \*

age 3.61883 1.08370 3.339 0.000938 \*\*\*

skin:age -0.03636 0.01583 -2.296 0.022293 \*

bp:age -0.02360 0.01243 -1.899 0.058449 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.31 on 324 degrees of freedom

Multiple R-squared: 0.1567, Adjusted R-squared: 0.1385

F-statistic: 8.604 on 7 and 324 DF, p-value: 1.104e-09

We observe the value of R-square only increase 0.0069 which shows it is not so useful to add the interaction term bp\*age.

> mylm4\_2 <- lm(glu ~.+skin\*age+bp\*age+bmi\*age,data=data)

> summary(mylm4\_2)

Call:

lm(formula = glu ~ . + skin \* age + bp \* age + bmi \* age, data = data)

Residuals:

Min 1Q Median 3Q Max

-62.032 -19.660 -4.539 17.626 76.015

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -16.94875 33.42775 -0.507 0.61248

npreg -0.84556 0.64898 -1.303 0.19354

bp 0.91638 0.42155 2.174 0.03044 \*

skin 1.62748 0.62000 2.625 0.00908 \*\*

bmi -0.02748 0.87598 -0.031 0.97499

age 3.38583 1.11233 3.044 0.00253 \*\*

skin:age -0.04497 0.01833 -2.453 0.01469 \*

bp:age -0.02747 0.01310 -2.096 0.03684 \*

bmi:age 0.02419 0.02593 0.933 0.35164

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.32 on 323 degrees of freedom

Multiple R-squared: 0.159, Adjusted R-squared: 0.1382

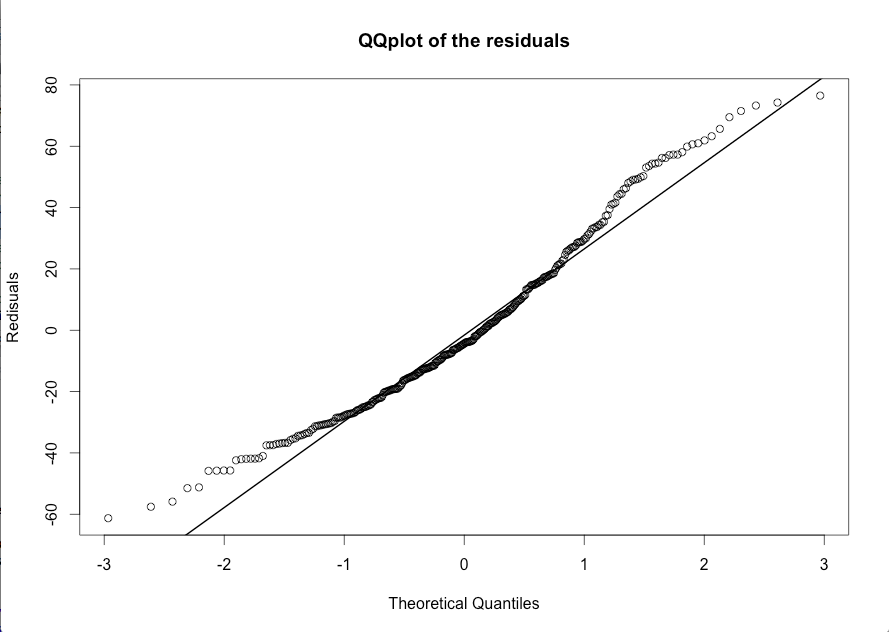
F-statistic: 7.634 on 8 and 323 DF, p-value: 2.308e-09

We observe the value of R-square only decrease 0.0003 which shows we should not add the interaction term bmi\*age to the linear regression model.

Then for the interaction term we can add the interaction term skin\*age and bp\*age to the linear regression model.

–Normality

Based on the qq plot we observed that it is not so close to a straight line so we conclude that the the data are not a sample from a normal distribution.



> # Shapiro-Wilk test for studentized residuals

> shapiro.test(r)

Shapiro-Wilk normality test

data: r

W = 0.9686, p-value = 1.321e-06

> #Ho:Normal data

> #Ha:Non-normal data

Since the p-value for the Shapiro-Wilk test is < 0.05 then we reject the null hypothesis so we conclude that the data are not a sample from a normal distribution.

–Homoscedasticity

For homoscedasticity, we want to assess the constancy of error variance.

> # Breush-Pagan Test for homoscedasticity

> # load lmtest package to gain access to bptest function

> library(lmtest)

> bptest(mylm, studentize = FALSE)

Breusch-Pagan test

data: mylm

BP = 17.1536, df = 5, p-value = 0.004218

> #H0:Variances of errors are constant

> #Ha:Variances of errors are not constant

Since the p-value <0.05 then we reject the null hypothesis, we conclude that the variances of errors are not constant.

Also we can use the Brown-Forsythe test to verify.

> ncvTest(mylm)

Non-constant Variance Score Test

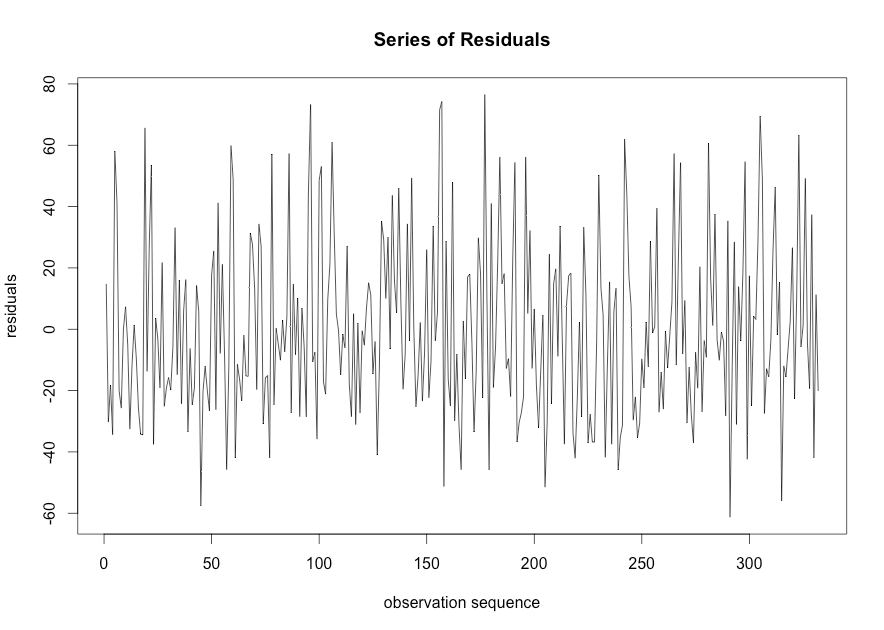
Variance formula: ~ fitted.values

Chisquare = 15.20531 Df = 1 p = 9.64318e-05

Therefore we reject the null hypothesis then we conclude that the variances of errors are not constant.

–Uncorrelated error

Based on the figure, we can see there is no obvious potential correlation. For further information, a Durin-Watson test can show the result.



> library("lmtest")

> library("zoo")

> dwtest(glu~npreg+bp+skin+bmi+age,data=data)

Durbin-Watson test

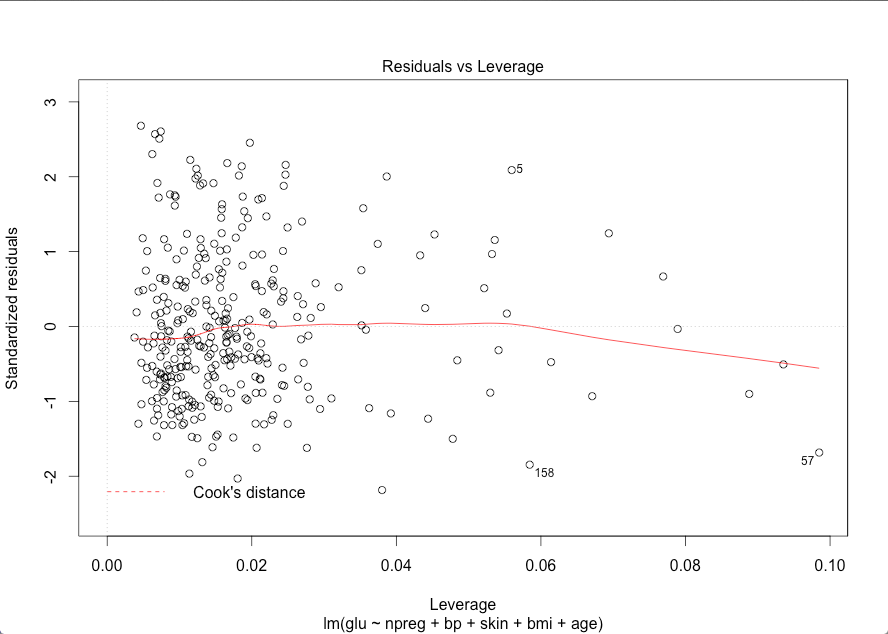
data: glu ~ npreg + bp + skin + bmi + age

DW = 1.9379, p-value = 0.2847

alternative hypothesis: true autocorrelation is greater than 0

Since p-value=0.2847>0.05, then we fail to reject the null hypothesis, we conclude that there is no correlation of errors.

–Check for outliers and influential points



Based on the figure above we can conclude the 5, 57 and 158 observations are the outliers.

For identifying outlying X Observations, we use hat matrix leverage values

> # Determine whether any leverages are large (i.e., larger than 2p/n)

> h <- hatvalues(mylm)

> which(h >= 2 \* 6/nrow(data))

5 8 12 18 21 41 43 57 72 79 92 107 141 158 196 198 203 211 217 232 249 262

5 8 12 18 21 41 43 57 72 79 92 107 141 158 196 198 203 211 217 232 249 262

287 291 292 320 330

287 291 292 320 330

Then we obtain observations 5,8,12,18,21,41,43,57,72,79,92,107,141,158,196,198,203,211,217,232,249,287,291,292,320,330 are outliers.

> nrow(data)

[1] 332

Based on R, for the influential, if we set n=332 as the small n, if the absolute value of its DFFITS exceeds one, if the maximum of the DFBETAS is greater than one for any observations, it is considered as an influential. For the Cook’s distance, it is considered as an outlier if its Cook’s distance exceeds the median of an distribution. According to this criterion, none of the observations are influential.

> D <- cooks.distance(mylm)

> which(D >= qf(.5, 5, 332-5))

named integer(0)

> DFFITS <- dffits(mylm)

> which(abs(DFFITS) > 1

+ )

named integer(0)

> DFBETAS <- dfbetas(mylm)

> max.DFBETAS <- apply(abs(DFBETAS), 1, max)

> which(max.DFBETAS > 1)

named integer(0)

If we consider n=332 as large n or large data set then if the absolute value of its DFFITS exceeds , if the maximum of the DFBETAS is greater than for any observations, it is considered as an influential.

> DFFITS <- dffits(mylm)

> which(abs(DFFITS) > 2\*sqrt(6/nrow(data)))

5 45 57 72 78 79 86 101 106 158 166 196 198 242 243 281 291 305 330

5 45 57 72 78 79 86 101 106 158 166 196 198 242 243 281 291 305 330

According to this criterion, by DFFITS observations 5, 45, 57, 72, 78, 79, 86, 101, 106, 158, 166, 196, 198, 242, 243, 281, 291, 305, 330 are influential.

> DFBETAS <- dfbetas(mylm)

> max.DFBETAS <- apply(abs(DFBETAS), 1, max)

> which(max.DFBETAS > 2/sqrt(332))

5 6 8 12 18 21 41 45 53 57 59 72 78 79 86 95 96 100 101 106 107 117

5 6 8 12 18 21 41 45 53 57 59 72 78 79 86 95 96 100 101 106 107 117

119 127 141 153 156 157 158 162 166 172 179 180 184 191 192 196 198 203 205 217 219 228

119 127 141 153 156 157 158 162 166 172 179 180 184 191 192 196 198 203 205 217 219 228

230 233 242 243 249 257 258 265 268 273 274 281 284 290 291 293 299 305 306 320 323 326

230 233 242 243 249 257 258 265 268 273 274 281 284 290 291 293 299 305 306 320 323 326

329 330

329 330

According to this criterion, by DFBETAS, observations 5, 6, 8, 12, 18, 21, 41, 45, 53, 57, 59, 72, 78, 79, 86, 95, 96, 100, 101, 106, 107, 117, 119, 127, 141, 153, 156, 157, 158, 162, 166, 172, 179, 180, 184, 191, 192, 196, 198, 203, 205, 217, 219, 228, 230, 233, 242, 243, 249, 257, 258, 265, 268, 273, 274, 281, 284, 290, 291, 293, 299, 305, 306, 320, 323, 326, 329, 330 are influential.

To make it more accurate we do the following process.

> IF <- influence.measures(mylm)

> DFFITS1 <- IF$is.inf[,7]

> DFBETAS1 <- IF$is.inf[,1:6]

> HAT1 <- IF$is.inf[,10]

> COOK1 <- IF$is.inf[,9]

> which(DFFITS1 == TRUE)

5 57 158 291

5 57 158 291

> which(DFBETAS1 == TRUE)

integer(0)

> which(HAT1 == TRUE)

5 8 57 72 79 158 203 217 232 287 292

5 8 57 72 79 158 203 217 232 287 292

> which(COOK1 == TRUE)

named integer(0)

Then we conclude the outliers are the 5, 8, 57, 72, 79, 158, 203, 217, 232, 287, 292 observations. The influential are the 5, 57, 158, 291 observations.

c) Propose remedial measures in case of violations of any of the underlying assumptions

–Linearity/functional form, including the need for any interaction terms

As the analyzing result shows, based on the summary of the linear model with interaction terms bp\*age and skin\*age we can use the remedial as followed. First do some transformations on predictors 'npreg' to make them significant. Second use other predictors such as add the interaction terms skin\*age and bp\*age into the regression, or remove predictor 'npreg'. Third we can use other non-linear models due to the very low R-squared of the linear regression shows there is not a strong linear relationship between 'glu' and the selected predictors.

The linear model we based on,

> mylm4\_1 <- lm(glu ~.+skin\*age+bp\*age,data=data)

> summary(mylm4\_1)

Call:

lm(formula = glu ~ . + skin \* age + bp \* age, data = data)

Residuals:

Min 1Q Median 3Q Max

-61.52 -20.00 -4.30 18.24 75.77

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -24.73944 32.36098 -0.764 0.445136

npreg -0.79329 0.64643 -1.227 0.220643

bp 0.80056 0.40276 1.988 0.047690 \*

skin 1.33841 0.53688 2.493 0.013169 \*

bmi 0.74008 0.30027 2.465 0.014231 \*

age 3.61883 1.08370 3.339 0.000938 \*\*\*

skin:age -0.03636 0.01583 -2.296 0.022293 \*

bp:age -0.02360 0.01243 -1.899 0.058449 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.31 on 324 degrees of freedom

Multiple R-squared: 0.1567, Adjusted R-squared: 0.1385

F-statistic: 8.604 on 7 and 324 DF, p-value: 1.104e-09

–Normality

Based on the conclusion we get from the tests we conclude that the data are not a sample from a normal distribution we could try the remedial followed to improve. First we still transform the data or we can use robust regression method such as we could change the square of the ordinary least square to the absolute value then take the minimum value.

–Homoscedasticity

Based on the conclusion we obtain from the test we conclude that variances of errors are not constant which means the heteroscedasticity exists in our project then we could use so we should use the following remedial measures to promote. First we could also do some transformation to the variables to make residuals variance stable. Second we could build the variance structure into other models such as weighted least square model.

–Uncorrelated error

We do not have this kind of problem in our project the population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria.

If we have we can use the following remedial. First we can use the Xt minus Xt-1, which make difference between the predictor before and predictor after. Then we use the difference, which is the delta we obtain to do the linear regression. Also we can transform data following Cochrane-Orcutt Procedure. Or we can use models that incorporate the correlation structure, such as Generalized Estimating Equations (GEE) method.

–Check for outliers and influential points

Based on our testing procedure, there are both ouliers and influential points existing in our project then we can use the following remedial methods. If the total number of the outliers and influential observations is small compared to the total number of samples, we could remove the outliers and report the influential points and then do the linear regression. If the total number of them is too many which cannot remove, we should try another methods to express the data.

The R code followed,

#ADA Homework 5

#Question 1 Part a

library(MASS)

data <- Pima.te[,c(1,2,3,4,5,7)]

mylm <- lm(glu ~ npreg+bp+skin+bmi+age,data = data )

summary(mylm)

mylm1 <- lm(data$glu ~data$npreg+data$bp+data$skin+data$bmi+data$age)

summary(mylm1)

#Question 1 Part b

e1 <- residuals(mylm) # OLS

r <- rstandard(mylm) # Studentized

t <- rstudent(mylm) # Deleted Studentized

# Assessing Linearity

# plot of studentized residuals vs fitted values

plot(mylm$fit, e1, xlab = "fitted values", ylab = "studentized residuals")

# Assessing interaction term

library(phia)

testInteractions(mylm)

mylm2 <- lm(glu ~ npreg+bp+skin+bmi+age+npreg\*bp+npreg\*skin+npreg\*bmi+npreg\*age+bp\*skin+bp\*bmi+bp\*age+skin\*bmi+skin\*age+bmi\*age,data = data )

summary(mylm2)

mylm3 <- lm(glu ~ npreg+bp+skin+bmi+age+skin\*age,data = data )

summary(mylm3)

mylm4\_1 <- lm(glu ~.+skin\*age+bp\*age,data=data)

summary(mylm4\_1)

mylm4\_2 <- lm(glu ~.+skin\*age+bp\*age+bmi\*age,data=data)

summary(mylm4\_2)

# Assessing Normality

# normal probability plot of studentized residuals

qqnorm(y=residuals(mylm),ylab="Redisuals",main="QQplot of the residuals")

qqline(residuals(mylm),lwd=2)

# See whether the residucal is from normal distribution

# quatity-quatity plot

# Shapiro-Wilk test for studentized residuals

shapiro.test(r)

#Ho:Normal data

#Ha:Non-normal data

# Assessing homoscedasticity

# Breush-Pagan Test for homoscedasticity

# load lmtest package to gain access to bptest function

library(lmtest)

bptest(mylm, studentize = FALSE)

#H0:Variances of errors are constant

#Ha:Variances of errors are not constant(if we change the predictor

#variance, the variances of errors are also changes)

e<- mylm$residuals

# Assessing Correlation of Error

plot(seq(1,332),e,type="l",xlab="observation sequence",ylab="residuals",main="Series of Residuals")

library("lmtest")

library("zoo")

dwtest(glu~npreg+bp+skin+bmi+age,data=data)

# Assessing Outlier

# Identify outlying responses

# Index plot of studentized residuals vs observation number

plot(rstandard(mylm), ylab = "studentized residuals", xlab = "observation")

# No unusually large studentized residuals

# Determine whether any deleted studentized residuals exceed

# what is expected (at a .95 confidence level) for an F distribution with p

# numerator degrees of freedom and n - p denominator degrees of freedom.

which(abs(rstudent(mylm)) >= qt(1 - .05/(2 \* nrow(data)), df = 332-6))

# named integer(0) means that non exceeded the treshhold

# Identify outlying X values

# Scatterplot of thigh circumference vs triceps skinfold thickness

# Label points by observation number. type = "n" means don't plot the points

plot(body$tri\_skin\_thick, body$thigh\_circ, xlab = "Triceps Skinfold Thickness",

ylab = "Thigh Circumference", type = "n")

text(body$tri\_skin\_thick, body$thigh\_circ, lab = 1:20)

# 3 and 15 may be a bit unusual

# Determine whether any leverages are large (i.e., larger than 2p/n)

h <- hatvalues(mylm)

which(h >= 2 \* 6/nrow(data))

# 3 and 15 are a bit large

# Identify influential observations

# Identify influential observations using DFFITS

DFFITS <- dffits(mylm)

which(abs(DFFITS) > 2\*sqrt(6/nrow(data)))

# For small data set, observation 3 influential

# Index plot of DFFITS

n <- nrow(data)

plot(DFFITS)

text(1:n,dffits(mylm),lab=1:n)

# Identify influential observations using Cook's Distances

D <- cooks.distance(mylm)

which(D >= qf(.5, 6, 332-6)) # none clearly identified as influential (though

# Index plot of Cook's Distances

plot(D, ylab = "Cook's Distance")

# Identify influential observations using DFBETAS

DFBETAS <- dfbetas(mylm)

max.DFBETAS <- apply(abs(DFBETAS), 1, max)

which(max.DFBETAS > 2/sqrt(332)) # For small data set, observation 3 influential

IF <- influence.measures(mylm)

DFFITS1 <- IF$is.inf[,7]

DFBETAS1 <- IF$is.inf[,1:6]

HAT1 <- IF$is.inf[,10]

COOK1 <- IF$is.inf[,9]

which(DFFITS1 == TRUE)

which(DFBETAS1 == TRUE)

which(HAT1 == TRUE)

which(COOK1 == TRUE)