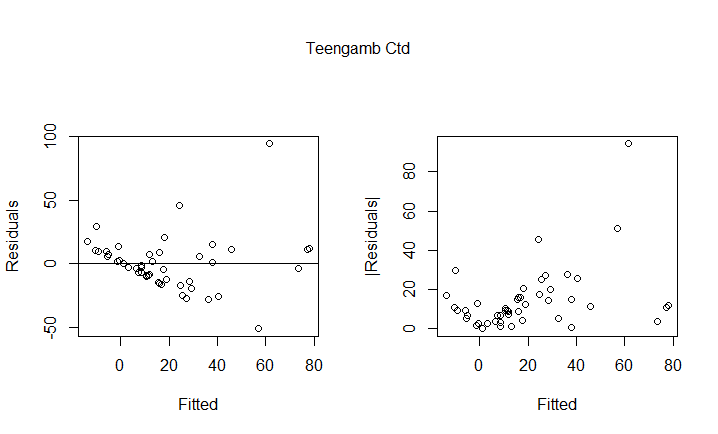
HW4 Solution

**• Check the constant variance assumption for the errors. Modify the model if necessary**

Using the teengamb dataset, first start with a linear regression of gamble on sex, status, income and verbal.

gamble = 22.56 – 22.12\*sex + 0.052\*status + 4.96\* income -2.96 \*verbal

To check the constant variance assumption for the errors. We plot the absolute values of the residuals against the fitted data as follows:



Also I had a look of the summary:

Coefficients:

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

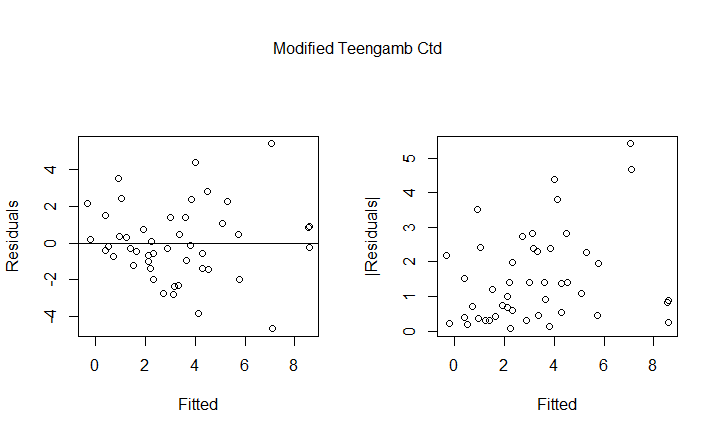
(Intercept) 9.3303 2.8789 3.241 0.00224 \*\*

tg.reg$fitted.values 0.2645 0.0968 2.732 0.00895 \*\*

The diagnostic plot revealed heteroscedasticity in residuals. The summary(with a significant positive coefficient) and the plot both show that residuals are increasing, indicating a non-constant variance.

Here I take the square root of the response, since the gamble variable has some zero values.

By fitting a new model as follows:



This time it passed the check of constant variance assumption with non-significant coefficient. Also we can see through the new diagnostic plot.

Coefficients:

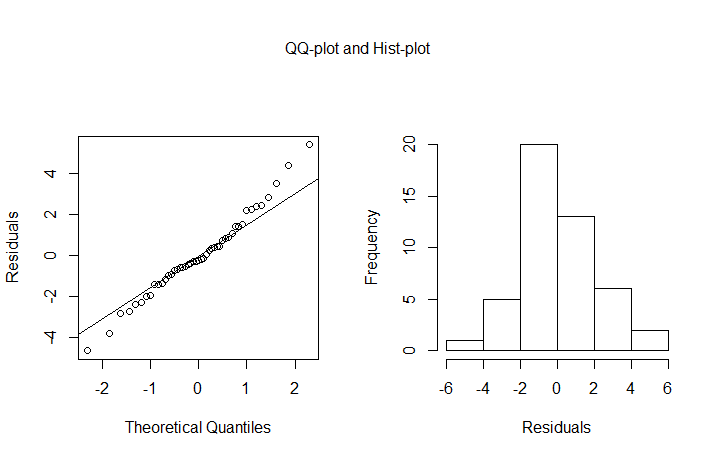
Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.01136 0.32365 3.125 0.00311 \*\*

tg.moreg$fitted.values 0.14957 0.08242 1.815 0.07623 .

• **Check the normality assumption.**

The qq-plot and hist-plot are as follows:



From the qq-plot and hist-plot and the Shapiro-Wilk test result, the new model passed the check of normality assumption.

>shapiro.test(tg.moreg$residuals)

Shapiro-Wilk normality test

data: tg.moreg$residuals

W = 0.98321, p-value = 0.7272

**• Check for large leverage points.**

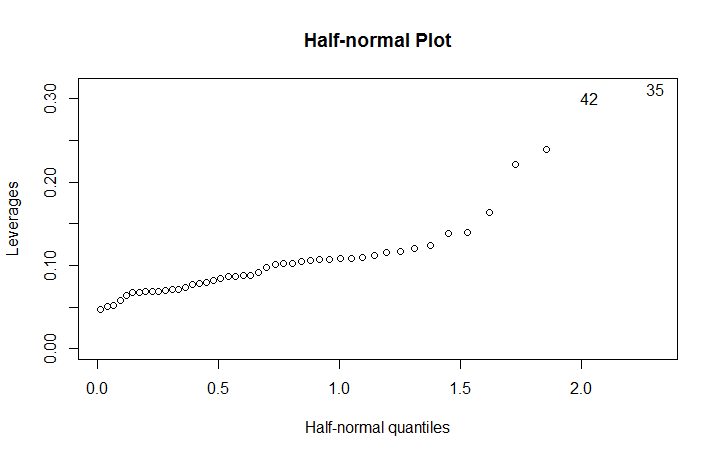
From the half-normal plot we can see that the points with large leverage is 35 and 42.

>teengamb[c(35, 42), ]

sex status income verbal gamble

35 0 28 1.5 1 14.1

42 0 61 15.0 9 69.7



• **Check for outliers.**

Through the externally studentized residuals and the Bonferroni Correction to mark the outliers.

24 has the largest residuals. However, after comparing the p-value with alpha/n, I found it is not an outlier. So other points who have smaller outlier-scores are not outliers, either.

> 2\*(1 - pt(max(abs(ti)), df = 47-5-1)) ## p-value

[1] 0.00414277

> 0.05/47 ## alpha/n

[1] 0.00106383

0.00414277 > 0.00106383, 24 is not an outlier.

>library(DMwR)

>outlier.scores <- lofactor(teengamb, k = 5)

> plot(density(outlier. scores))

> outliers <- order(outlier.scores, decreasing=T)[1:5]

> print(outliers)

[1] 24 35 42 36 19 none of the five is an outlier

There is no outlier in this model.

• **Check for influential points.**

I calculated the cook’s distance to identify the influential points.

The extract each of the influential points to see how the betas change. Use 24 as an example.

After extract 24, the model changes to:

The betas change significantly.

Some other influential points are identified in similar ways.

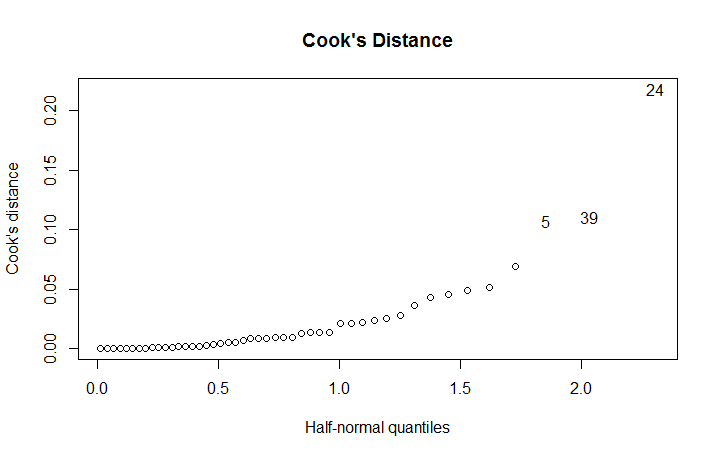
> identify(tg.moreg.inf$coef[, 1], tg.moreg.inf$coef[, 2])

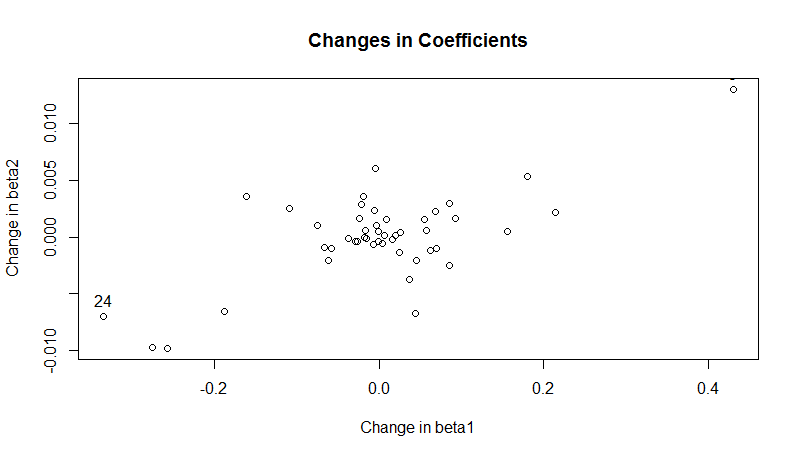
[1] 5 23 24

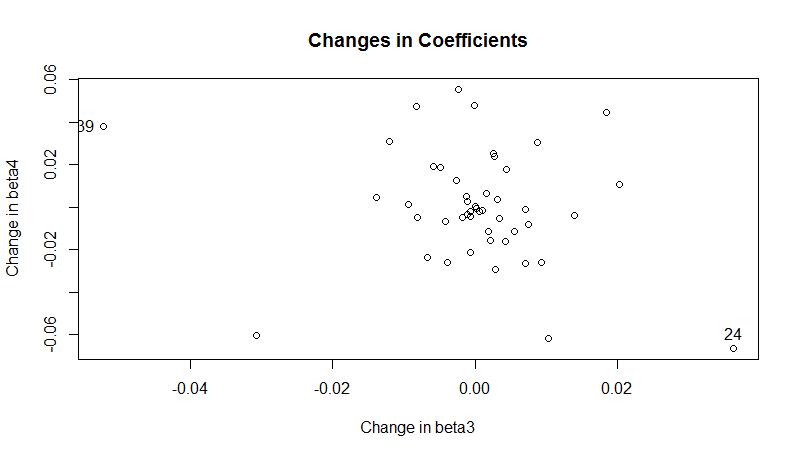
>plot(tg.moreg.inf$coef[, 3], tg.moreg.inf$coef[, 4], xlab = "Change in beta3",ylab="Change in beta4", main = "Changes in Coefficients")

> identify(tg.moreg.inf$coef[, 3], tg.moreg.inf$coef[, 4])

[1] 5 24 39







After analyzing the changes in coefficients, we found that:

The most influential point is 24. Some other influential points are 5, 39.

Codes Used

library(faraway)

data(teengamb)

attach(teengamb)

tg.reg=lm(gamble~sex+status+income+verbal,teengamb)

summary(lm(abs(tg.reg$residuals)~tg.reg$fitted.values))

par(mfrow=c(1,2),oma=c(0,0,3,0))

plot(tg.reg$fitted,tg.reg$residuals,xlab="Fitted",ylab="Residuals")

abline(h=0)

plot(tg.reg$fitted,abs(tg.reg$residuals),lty=2,xlab="Fitted",ylab="|Residuals|")

mtext("Teengamb Ctd",side=3,line=0,outer=T)

tg.moreg=lm(sqrt(gamble)~sex+status+income+verbal,teengamb)

par(mfrow=c(1,2),oma=c(0,0,3,0))

plot(tg.moreg$fitted,tg.moreg$residuals,xlab="Fitted",ylab="Residuals")

plot(tg.moreg$fitted,abs(tg.moreg$residuals),xlab="Fitted",ylab="|Residuals|")

mtext("Modified Teengamb Ctd",side=3,line=0,outer=T)

summary(lm(abs(tg.moreg$residuals)~tg.moreg$fitted.values))

qqnorm(tg.moreg$residuals,ylab="Residuals",main="")

qqline(tg.moreg$residuals)

hist(tg.moreg$residuals,xlab="Residuals",main="")

mtext("QQ-plot and Hist-plot",side=3,line=0,outer=T)

reg<-lm(gamble~.,data=teengamb)

halfnorm(lm.influence(reg)$hat,nlab=2,ylab="Leverages",main="Half-normal Plot")

teengamb[c(35,42),]

ti <- rstudent(tg.moreg)

max(abs(ti))

which(ti == max(abs(ti)))

2\*(1-pt(max(abs(ti)),df=47-5-1))

0.05/47

library(DMwR)

outlier.scores<-lofactor(teengamb,k=5)

plot(density(outlier.scores))

outliers <- order(outlier.scores, decreasing=T)[1:5]

print(outliers)

cook<-cooks.distance(tg.moreg)

halfnorm(cook,nlab=3,ylab="Cook's distance",main="Cook's Distance")

tg.moreg.24<-lm(gamble~sex+status+income+verbal,data=teengamb,subset=(cook<max(cook)))

plot(tg.moreg.inf$coef[,2],tg.moreg.inf$coef[,3],xlab="Change in beta2",ylab="Change in beta3",main="Changes in Coefficients")

identify(tg.moreg.inf$coef[,2],tg.moreg.inf$coef[,3])