STATS 500 - Homework 9

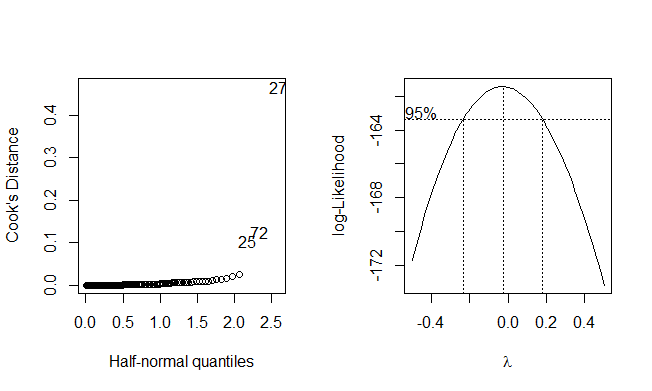
Using the infmort data, find a simple model for the infant mortality in terms of the variables income and regions (do not include oil). Be alert to transformations and unusual points. Interpret your model by explaining how to interpret the estimates of the regression parameters.

1. **Handling missing values**

There are 4 missing values in mortality. After simply deleting them, there remain 101 observations.

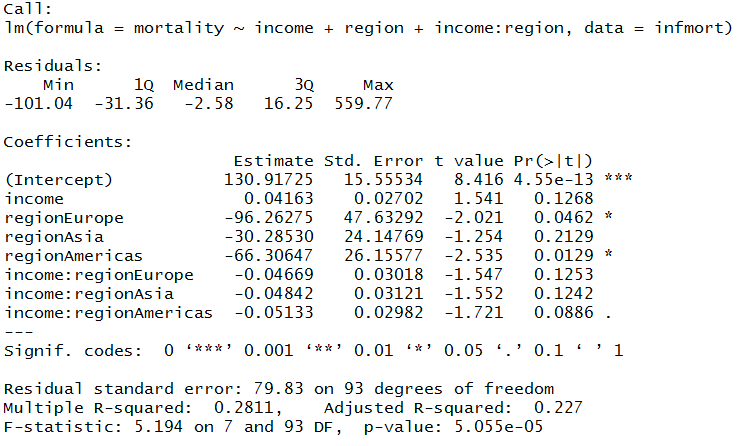
1. **Looking at Box-Cox plot of response vs. the full model of income and region(with interactions)**

First do the diagnostics, using the Cook’s distance, from the halfnorm plot we found 3 influential points, which are Afganistan, Saudi\_Arabia, and Libya.

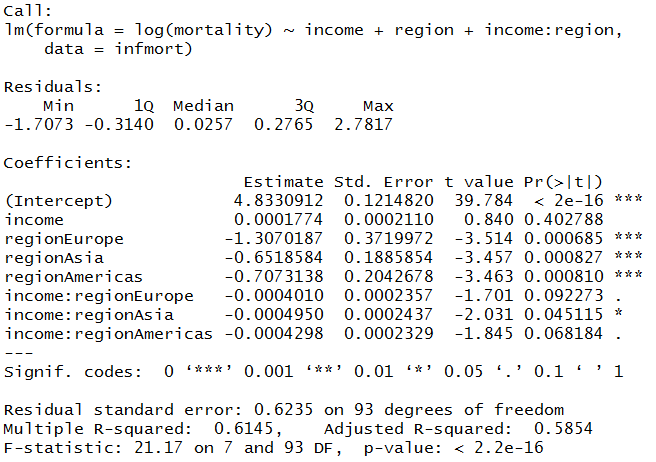


We look at the Box-Cox plot(excluding the outliers). Select λ= 0, using the transformation that .

The full model of income and region(with interactions)



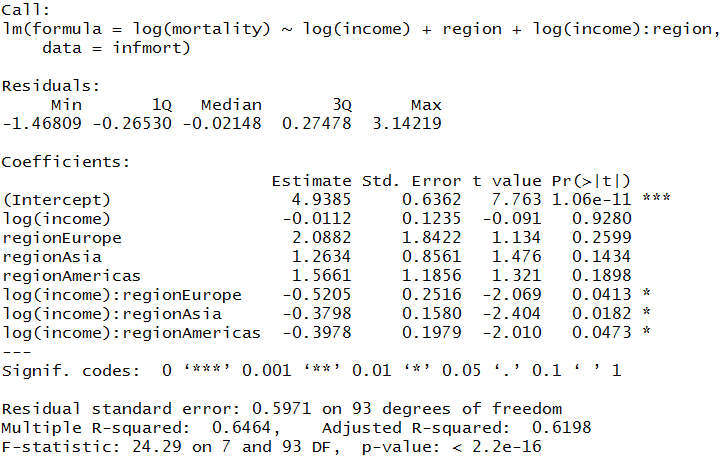
After the transformation of response:



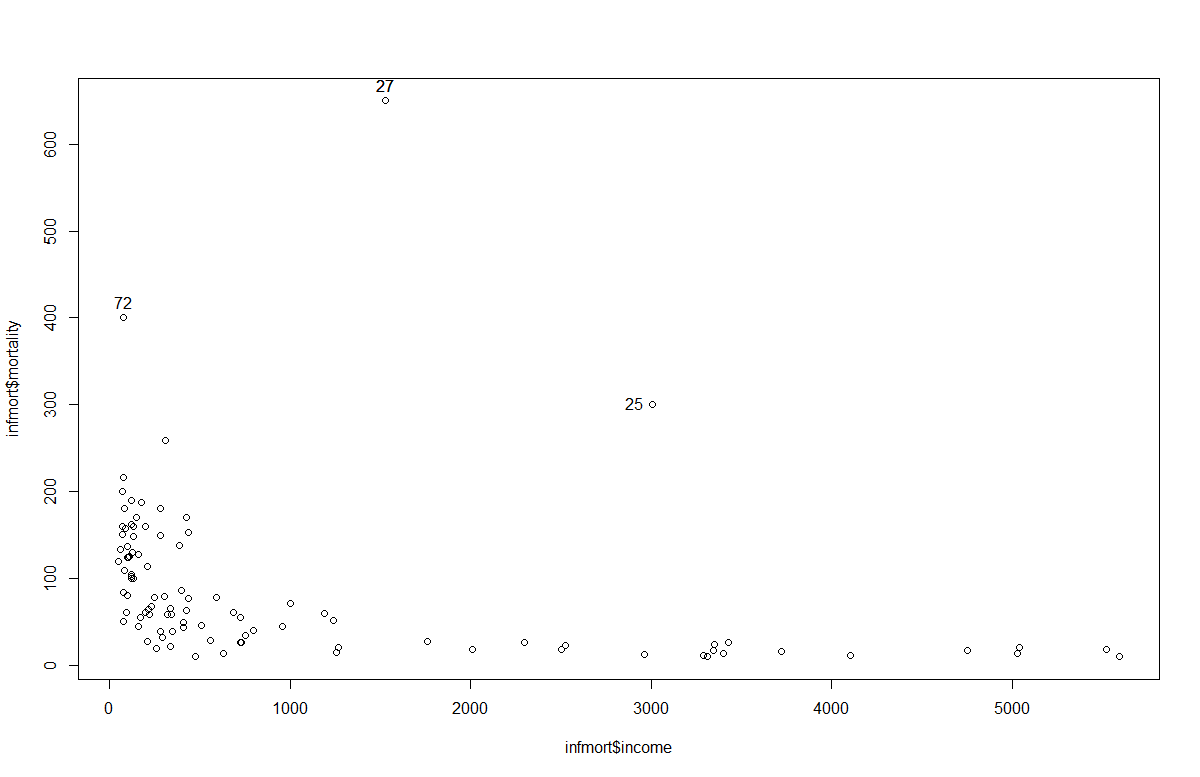
1. **Using transform of response, now investigate transform on predictor variable income:**

Introduce log terms of variable income:

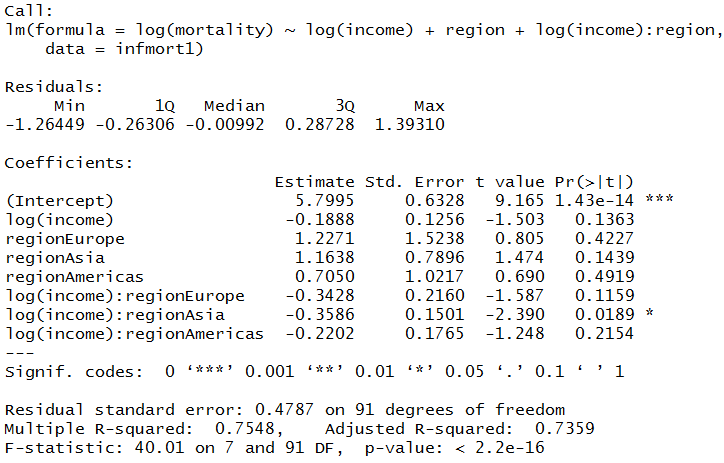
Replace the income with log(income), we get satisfying significance.



1. **Once settled on possibly transforms for both response and income predictor, investigate outliers.** There are a clear few outliers. Simply looking at plots of infant mortality vs. income, we identify three outliers. They are Saudi\_Arabia, and Libya.

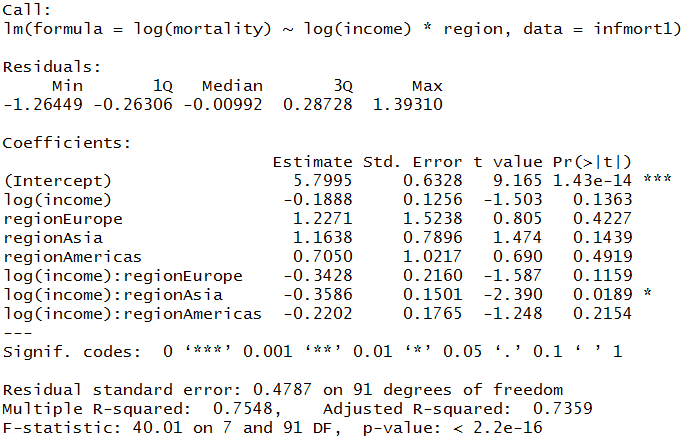


After remove the outliers(27 and 25). We got a better fit( Adjusted R-squared has increased)



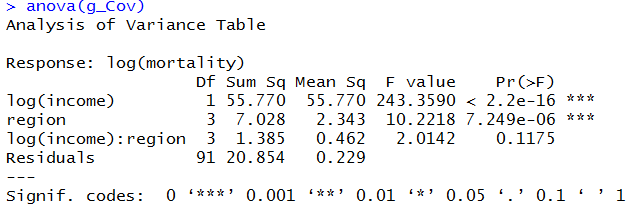
1. **Continue with the standard analysis of covariance.**

Fit the full model with all terms



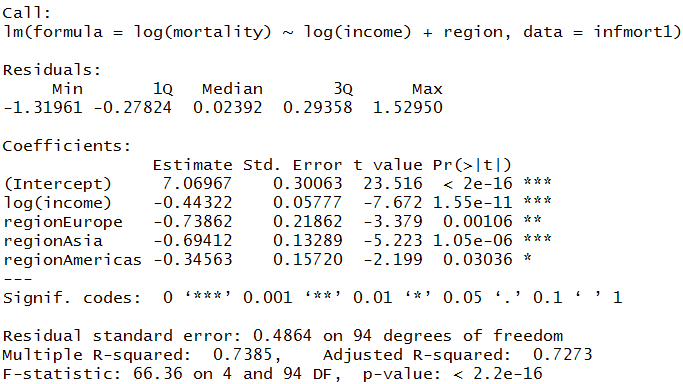
Among the interactions, the interaction between log(income) and regionAsia is significant

Let’s do the sequential analysis of variance table

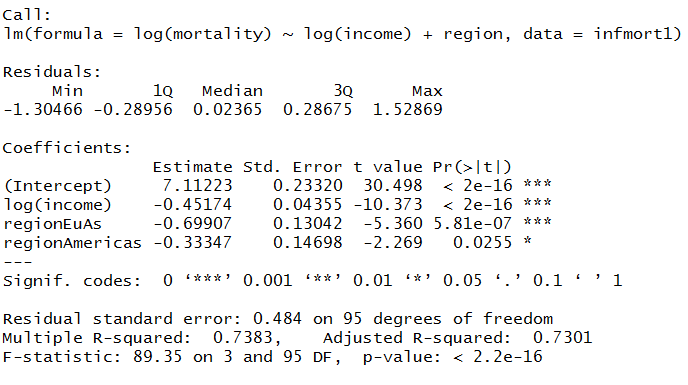


We found that the interactions are not significant using the ANOVA.

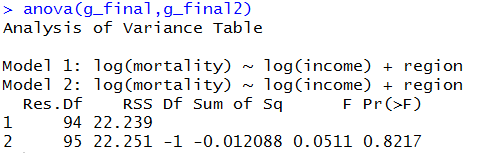
So that we fit the final model without interaction.



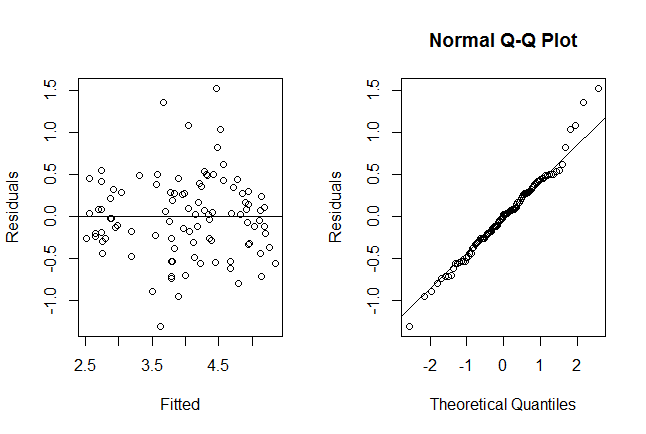
The regionEurope and regionAsia are not significantly different.Re-group into three levels, Africa, EuAs, and Americas. We got the final model.



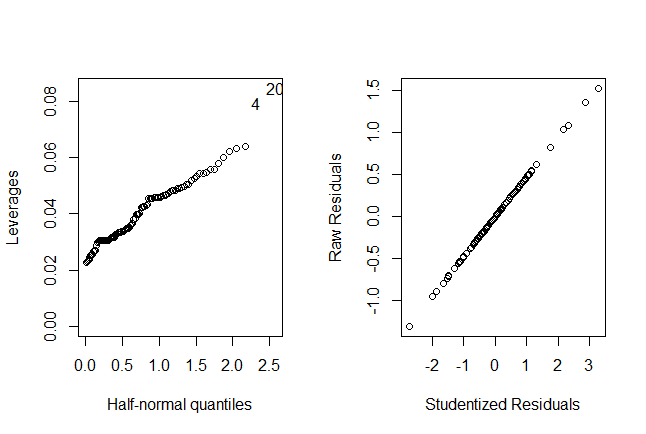
Do the anova to check with 3 levels for region.



After determine the final model. We do diagnostics to check the validation of the model.



The residuals showed homoscedasticity and the QQ-plot showed the residuals are normal.



The halfnorm plot and the studentized residuals plot also showed that the model has a good fit.

**R-code used:**

library(faraway)

data(infmort)

attach(infmort)

sum(is.na(infmort))

summary(infmort)

## Remove NAs

infmort=na.omit(infmort[,-4])

library(MASS)

g=lm(mortality~income+region+income:region, data = infmort)

## Compute Cook's Distance

cook<-cooks.distance(g)

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

## Find the outliers

infmort[c(25,72,27),]

ti<-rstudent(g)

pt(ti[25],df=101-5-1)

##Compute the p-value and compare with alpha/n

2\*(1-pt(ti[25],df=101-5-1))-0.05/101

# Libya is not an outlier

# 0.02631397

pt(ti[72],df=101-5-1)

##Compute the p-value and compare with alpha/n

2\*(1-pt(ti[72],df=101-5-1))-0.05/101

# Afganistan is an outlier

# -0.0004287065

pt(ti[27],df=101-5-1)

##Compute the p-value and compare with alpha/n

2\*(1-pt(ti[27],df=101-5-1))-0.05/101

##Saudi\_Arabia is an outlier

# -0.0004950495

g1=lm(mortality~income+region+income:region, data = infmort)

par(mfrow=c(1,2))

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

boxcox(g1, plotit =TRUE,lambda=seq(-0.5,0.5,by=0.05))

##Transform the response to ln

g2=lm(log(mortality)~income+region+income:region, data = infmort)

g3=lm(log(mortality)~log(income)+region+log(income):region,data=infmort)

#Identify outliers

identify(infmort$income,infmort$mortality)

infmort1=infmort[-c(27,25),]

#Remove the outliers

g4=lm(log(mortality)~log(income)+region+log(income):region,data=infmort1)

g\_Cov=lm(log(mortality)~log(income)\*region,infmort1)

anova(g\_Cov)

g\_final=lm(log(mortality)~log(income)+region,data=infmort1)

levels(infmort1$region)=c("Africa","EuAs","EuAs","Americas")

g\_final2=lm(log(mortality)~log(income)+region,data=infmort1)

anova(g\_final,g\_final2)

par(mfrow=c(1,2))

plot(g\_final2$fitted.values,g\_final2$residuals,xlab="Fitted",ylab="Residuals",main="")

abline(h=0)

qqnorm(g\_final2$residual,ylab="Residuals")

qqline(g\_final2$residual)

halfnorm(lm.influence(g\_final2)$hat,nlab=2,ylab="Leverages")

plot(g\_final2$residuals/((summary(g\_final2)$sig)\*sqrt(1-lm.influence(g\_final2)$hat)), g\_final2$residuals,xlab="Studentized Residuals",ylab="Raw Residuals")ls",ylab="Raw Residuals")