Huiwen Chen 02156341 huiwenc@umich.edu

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

Code:

```
>library(faraway)

>data(infmort)

>attach(infmort)

>sum(is.na(infmort))

>4

##remove samples with missing data

>df = na.omit(infmort)

>df = df[, -4]
```

There are 4 missing data with income = NA in the data set. After removing them, there are 101 observations left.

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

Code:

```
library(MASS)

g1 <- lm(mortality ~ income+ region+ income:
region,df)

##find the outlier

cook <- cooks.distance(g1)

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

##remove the ourliers and then use boxcox to find
lambda

df <- df[-c(25, 72,27),]

g1 <- lm(mortality ~ income+ region+ income:
region,df)

boxcox(g1, data = df, plotit =T)

##transform the response

g2 <- lm(log(mortality)~ income + region+ income:
region,df)

boxcox(g2, data = df,plotit =T,lambda = seq(-2, 4, by
=0.05))

summary(g1)

Call:

lm(formula = mortality ~ income + region +
income:region, data = df)

Coefficients:

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

(Intercept) 152.79825     10.09018    15.143   < 2e-16

income      -0.08033      0.03901    -2.059     0.0424

regionEurope -118.14375     24.32762    -4.856
5.01e-06

regionAsia          -78.19063     13.74903    -5.687
1.59e-07
```

regionAmericas -88.18746 14.44664 -6.104 2.56e-08

income:regionEurope 0.07526 0.03957 1.902 0.0604

income:regionAsia 0.06056 0.03979 1.522 0.1315

income:regionAmericas 0.07062 0.03950 1.788 0.0772

Residual standard error: 39.25 on 90 degrees of freedom Multiple R-squared: 0.6112, Adjusted R-squared: 0.581 F-statistic: 20.21 on 7 and 90 DF, p-value: 4.692e-16

Summary(g2)

Call:

lm(formula = log(mortality) ~ income + region + income:region, data = df)

Coefficients:

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

(Intercept) 5.0014097 0.1302556 38.397 < 2e-16

income -0.0007607 0.0005036 -1.511 0.134

regionEurope -1.4753373 0.3140487 -4.698 9.40e-06

regionAsia -0.9683024 0.1774881 -5.456 4.26e-07

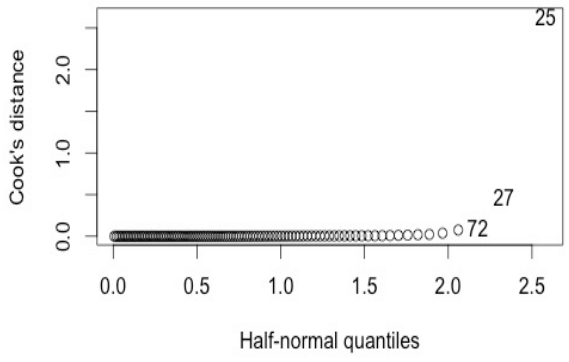
regionAmericas -0.8756323 0.1864938 -4.695 9.49e-06

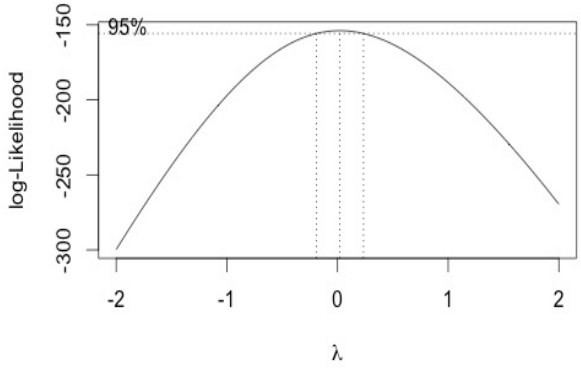
income:regionEurope 0.0005371 0.0005108 1.051 0.296

income:regionAsia 0.0003865 0.0005137 0.752 0.454

income:regionAmericas 0.0005083 0.0005100 0.997 0.322

Residual standard error: 0.5067 on 90 degrees of freedom Multiple R-squared: 0.715, Adjusted R-squared: 0.6929 F-statistic: 32.26 on 7 and 90 DF, p-value: < 2.2e-16





By checking cook's distance, we find 3 influential points. After removing the three points, and check boxcox plot we select lambda = 0 for transforming response. Then comparing the summary of the full model before and after response transformation, we find that the interaction term is not significant indicating that we can fit the same slope within each group. So we can remove the interaction term. And after response transformation, the adjusted R-squared increased a lot.

(3) Now investigate transform on predictor variable income

Code:

```
> g4 <- lm(log(mortality)~ log(income) + region+
log(income): region,df)
> summary(g4)
```

Call:

lm(formula = log(mortality) ~ log(income) + region + log(income):region, data = df)

Residuals:

Min 1Q Median 3Q Max

-1.2371 -0.2592 -0.0069 0.2958 1.4260

Coefficients:

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

(Intercept) 5.7995 0.6030 9.618 1.78e-15

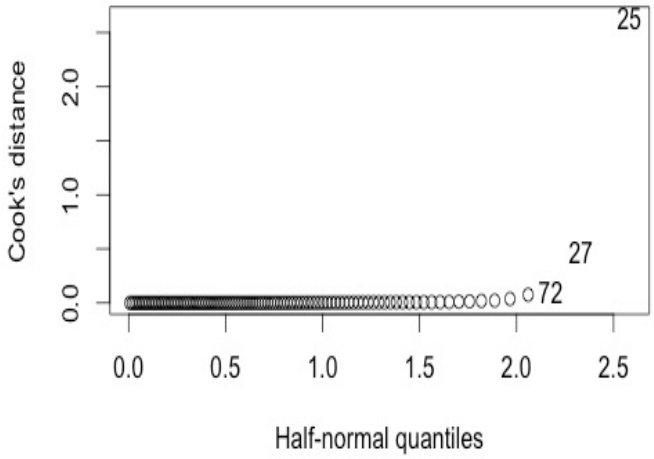
```
log(income)                    -0.1888    0.1197
-1.577    0.1182
regionEurope                    1.2271    1.4519
0.845     0.4003
regionAsia                      0.7702    0.7623
1.010     0.3151
regionAmericas                  0.7050    0.9735
0.724     0.4708
log(income):regionEurope       -0.3428    0.2058
-1.666    0.0992
log(income):regionAsia         -0.2992    0.1442
-2.075    0.0408
log(income):regionAmericas     -0.2202    0.1682
-1.309    0.1937

(Intercept)                    ***
log(income)
regionEurope
regionAsia
regionAmericas
log(income):regionEurope       .
log(income):regionAsia         *
log(income):regionAmericas
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 0.4561 on 90 degrees of
freedom
Multiple R-squared:  0.7691,    Adjusted R-squared:
0.7511
F-statistic: 42.82 on 7 and 90 DF,  p-value: < 2.2e-16
```

By adding log term of predictor income, we find that all of the terms are significant. So the model after predictor transformation is log(mortality) = log(income) + region+ log(income): region

(4) Investigate outliers simply, first simply looking at plots of infant mortality vs. income.

Code:

```
##simply check
plot(df$income, df$mortality)
 identify(df$income, df$mortality)
## [1] 25 27 72
##compute cook's distance to check
influential points
plot(mortality ~income, df)
identify(income,mortality)
cook<-cooks.distance(g4)
halfnorm(cook,nlab=3,ylab="Cook's
Distance")
##identify the influential points
df[c(72,25,27),]
## Afganistan libya Saudi_Arabia
ti<-rstudent(g4)
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 0.00573199    not outlier
pt(ti[25],df=101-5-1)
2*(1-pt(ti[25],df=101-5-1))-0.05/101
##Libya 0.001954266 is not an outlier
pt(ti[27],df=101-5-1)
2*(1-pt(ti[27],df=101-5-1))-0.05/101
##Saudi_Arabia -0.0004949912 not outlier
##remove 27th point and refit
df <- df[-27,]
 g4 <- lm(log(mortality)~ log(income) + region,df)
```



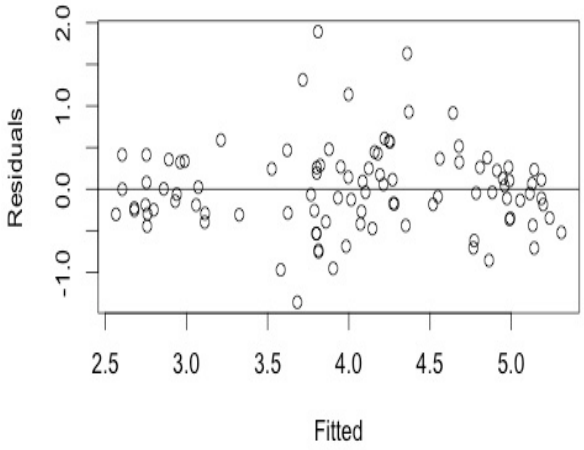
Through simply plot of the predictor against response, we detected that 25th, 27th,72th observations are outliers possibly.
Then by checking cook's distance, we find 72th, 25th, 27th observations are influential points.
After t-test, we find the 27th point is an outlier. So we remove it and refit the model.

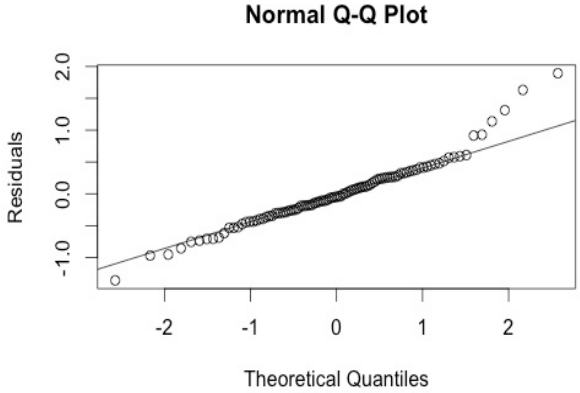
(5) Continue with the standard analysis of covariance, to determine a final model, and include the usual diagnostics for linear models.

Code:

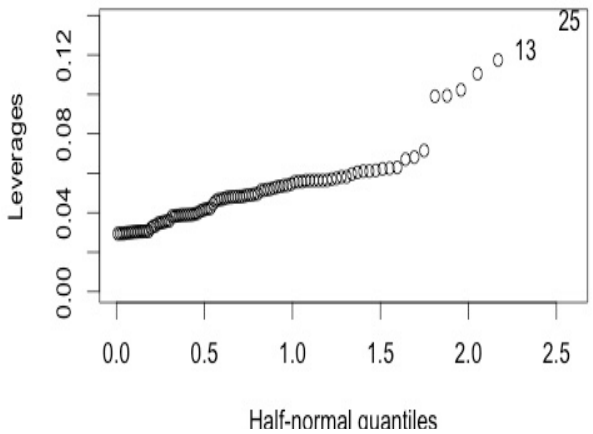
```
##standard analysis of covariance
g_cov <- lm(log(mortality)~ log(income)*region,df)
anova(g_cov)
Analysis of Variance Table

Response: log(mortality)
                    Df Sum Sq Mean Sq    F value    Pr(>F)
log(income)          1 51.117  51.117 212.1189 < 2.2e-16 ***
region               3 10.484   3.495  14.5015 8.477e-08 ***
log(income):region   3  4.318   1.439   5.9721  0.00092 ***
Residuals           91 21.930   0.241
plot(g4$fitted.values,g4$residuals,xlab="Fitted",ylab="Residuals",main="")
abline(h=0)
qqnorm(g4$residual,ylab="Residuals")
qqline(g4$residual)
halfnorm(lm.influence(g4)$hat,nlab=2,ylab="Leverages")
plot(g4$residuals/((summary(g4)$sig)*sqrt(1-lm.influence(g4)$hat)), g4$residuals,xlab="Studentized Residuals",ylab="Raw Residuals")
```

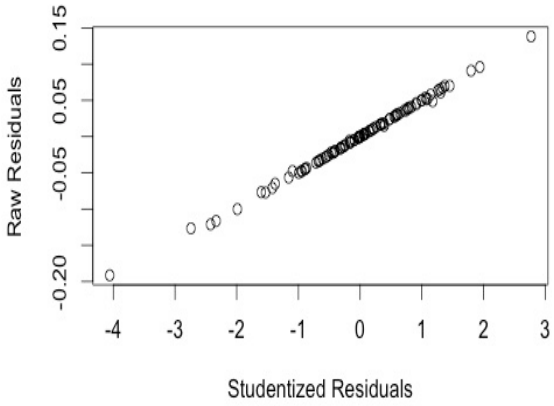
By checking anova, we find the interaction term is not significant , so we remove it.
Based on the analysis above, the final model we selected is log(mortality)= log(income)+ region.





In the plots above, we can see residuals scattered symmetrically around 0 against fitted values, so the model shows constant variance. From qq-plot, we find the residuals are normal.





Also the final model shows a good fit. This can also be confirmed by the two plots above.