Statistics 138 Project

PART II

Troy Lui
Due: December 12, 2018

PART II: Ischemic Heart Disease

I. Introduction:

To preface, the time period of the data collected was between January 1, 1998 December 31, 1999. In addition, the data set has variables cost, age, gender, total number of procedures carried out, number of tracked drugs prescribed, number of other complications that arose during the heart disease treatment, comorbidities, and duration of treatment condition. The goal is to use these variables to model the mean of the number of emergency room visits from an insurance company's subscribers.

II. Materials and Methods:

To be a little more specific about the data, the ischemic data set has gender listed as 0 if otherwise and 1 if male. In addition, complications, drugs, inter, comorbidities, and duration are all numbers that don't technically have bounds like gender.

The type of statistical method that will be used is a Poisson regression; however, do note that the mean visits (3.425) do not equal the variance of visits (6.956). Therefore, we should be considering a different model from the start. To start, the strategy will be to fit a Poisson regression model where the mean visits is modeled as mean(visits) = $\exp(B_0+B_1X_1+\ldots+B_8X_8)$ [not including interaction variables]. The saturated model that was first put into use included 25 variables, which were (cost + age + gender + inter + drugs + complications + comorbidities + duration + cost*age + cost*gender + cost*inter + cost*complications + cost*comorbidities + cost*duration + age*inter + age*drugs + age*complications + age*comorbidities + age*duration + inter*drugs + inter*complications + inter*comorbidities + inter*duration + complications*comorbidities + comorbidities*duration).

Transformations of the variables were considered. For example, taking the square root of all the predictor variables (except gender) was considered; however, when run through the diagnostics and run test, the untransformed data seemed to be a better fit than the transformed data. Therefore, the untransformed predictor variables were used.

The summary statistics can be found in figure 9 in the appendix. Diagnostics were then taken from the untransformed data and figures 11, 12, and 13 were created. From these diagnostics, the deviance and Pearson residuals did not look to have different distributions, which does not indicate lack-of-fit. In addition, figures 12 and 13 of the residuals against the fitted values show that the smoothed line does run along 0. Lastly, when using runs test, the p-value came out to 0.354, which suggests we can't reject a good fit. All in all, the saturated model seemed to fit the data well.

III. Results:

Although the saturated model fit the data well, stepAIC function was implemented in a backwards direction to receive a final model to drop any unnecessary variables since the initial model was already so concentrated.

From this, the final model that was chosen was:

$$\text{Mean(visits)} = \exp(0.0842 + 0.000032X_1 + 0.0146X_2 + 0.18178X_3 - 0.00955X_4 + 0.2355X_5 - 0.8075X_6 + 0.002296X_8 - 0.0000004X_1X_4 - 0.00000007X_1X_8 + 0.01996X_2X_6 - 0.00005X_2X_8 - 0.004X_4X_5 - 0.02799X_4X_6 + 0.0001898X_4X_8)$$

[final ischemic model summary statistics: Figure 10 appendix]

For the interpretation of the model, variables cost, age, gender, drugs, duration, age*complications, and inter*duration, if given values made the mean(visits) variable increase while the rest of the betas made the mean(visits) variable decrease.

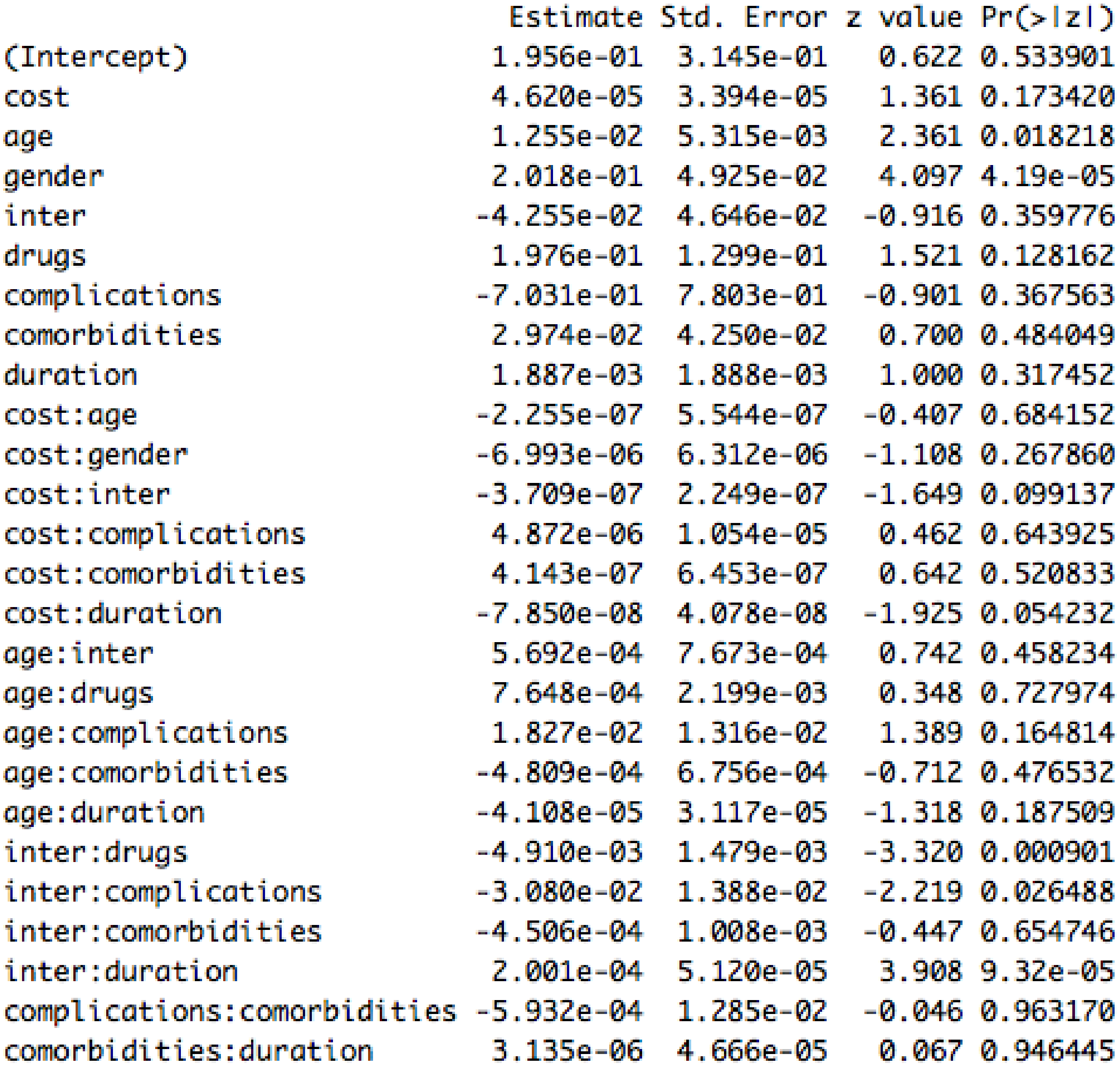
Taking a look at figure 15 and 16, the plots actually look very similar to the initial model's plots. This could mean that there wasn't much change in goodness of fit when switching from the initial model to the final model. When a run test was applied to the final model, the p-value was equal to 0.7215, which is convincing evidence to fail to reject the null hypothesis that there are no systematic patterns.

IV. Conclusion and Discussion:

In conclusion, the interesting points to points out are that the initial saturated model being chosen might not be too far off in goodness of fit when compared to the final model. Another interesting points to note again like in Part I is that having an over saturated model doesn't seem to squander any part of the process; therefore, for further analysis, I would add more interaction variables and run it through stepAIC to see if there are any other models with smaller AIC.

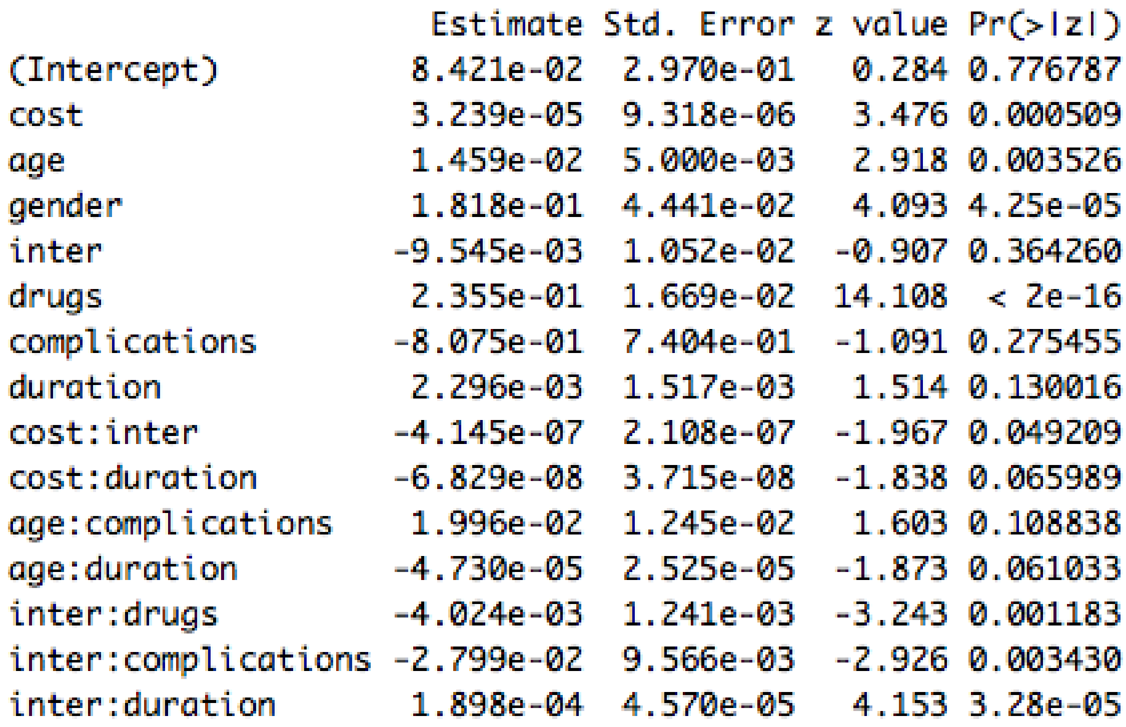
Appendix

Figure 9: (Initial ischemic model summary statistics)



```
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.956e-01  3.145e-01   0.622 0.533901
cost                        4.620e-05  3.394e-05   1.361 0.173420
age                         1.255e-02  5.315e-03   2.361 0.018218
gender                      2.018e-01  4.925e-02   4.097 4.19e-05
inter                      -4.255e-02  4.646e-02  -0.916 0.359776
drugs                       1.976e-01  1.299e-01   1.521 0.128162
complications              -7.031e-01  7.803e-01  -0.901 0.367563
comorbidities               2.974e-02  4.250e-02   0.700 0.484049
duration                    1.887e-03  1.888e-03   1.000 0.317452
cost:age                   -2.255e-07  5.544e-07  -0.407 0.684152
cost:gender                -6.993e-06  6.312e-06  -1.108 0.267860
cost:inter                 -3.709e-07  2.249e-07  -1.649 0.099137
cost:complications          4.872e-06  1.054e-05   0.462 0.643925
cost:comorbidities          4.143e-07  6.453e-07   0.642 0.520833
cost:duration              -7.850e-08  4.078e-08  -1.925 0.054232
age:inter                   5.692e-04  7.673e-04   0.742 0.458234
age:drugs                   7.648e-04  2.199e-03   0.348 0.727974
age:complications           1.827e-02  1.316e-02   1.389 0.164814
age:comorbidities          -4.809e-04  6.756e-04  -0.712 0.476532
age:duration               -4.108e-05  3.117e-05  -1.318 0.187509
inter:drugs                -4.910e-03  1.479e-03  -3.320 0.000901
inter:complications        -3.080e-02  1.388e-02  -2.219 0.026488
inter:comorbidities        -4.506e-04  1.008e-03  -0.447 0.654746
inter:duration              2.001e-04  5.120e-05   3.908 9.32e-05
complications:comorbidities -5.932e-04  1.285e-02  -0.046 0.963170
comorbidities:duration      3.135e-06  4.666e-05   0.067 0.946445
```

Figure 10: (Final ischemic model summary statistics)



```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              8.421e-02  2.970e-01   0.284 0.776787
cost                     3.239e-05  9.318e-06   3.476 0.000509
age                      1.459e-02  5.000e-03   2.918 0.003526
gender                   1.818e-01  4.441e-02   4.093 4.25e-05
inter                   -9.545e-03  1.052e-02  -0.907 0.364260
drugs                    2.355e-01  1.669e-02  14.108  < 2e-16
complications           -8.075e-01  7.404e-01  -1.091 0.275455
duration                 2.296e-03  1.517e-03   1.514 0.130016
cost:inter              -4.145e-07  2.108e-07  -1.967 0.049209
cost:duration           -6.829e-08  3.715e-08  -1.838 0.065989
age:complications        1.996e-02  1.245e-02   1.603 0.108838
age:duration            -4.730e-05  2.525e-05  -1.873 0.061033
inter:drugs             -4.024e-03  1.241e-03  -3.243 0.001183
inter:complications     -2.799e-02  9.566e-03  -2.926 0.003430
inter:duration           1.898e-04  4.570e-05   4.153 3.28e-05
```

Figure 11: (initial ischemic)

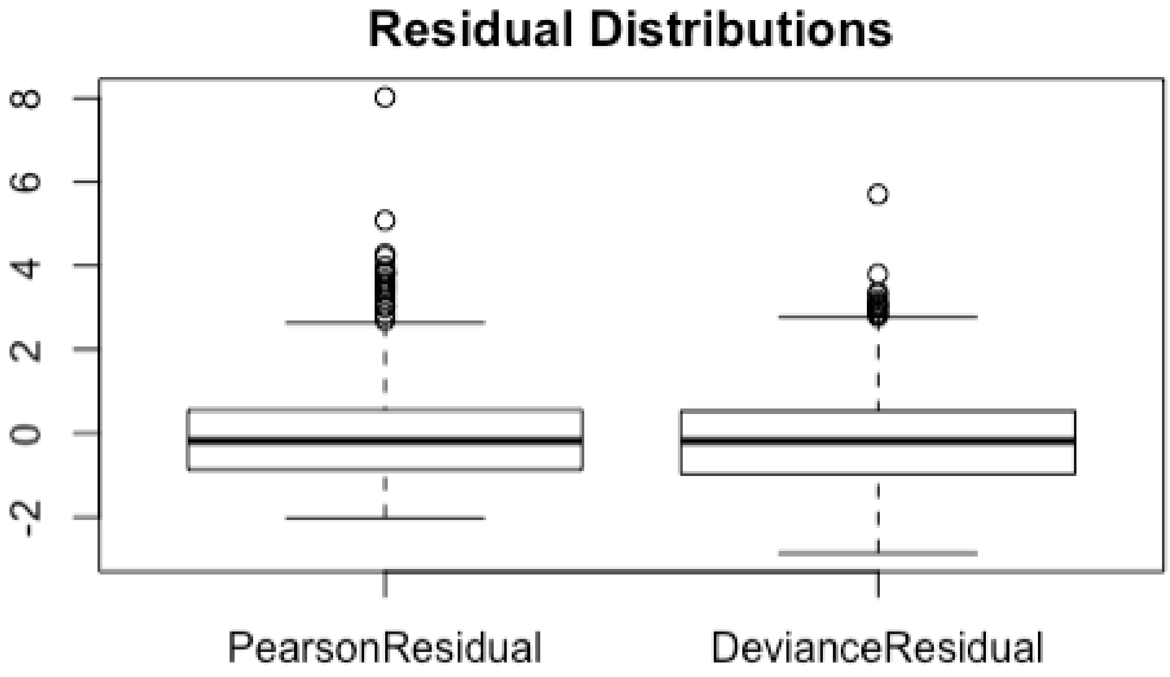


Figure 12: (initial ischemic)

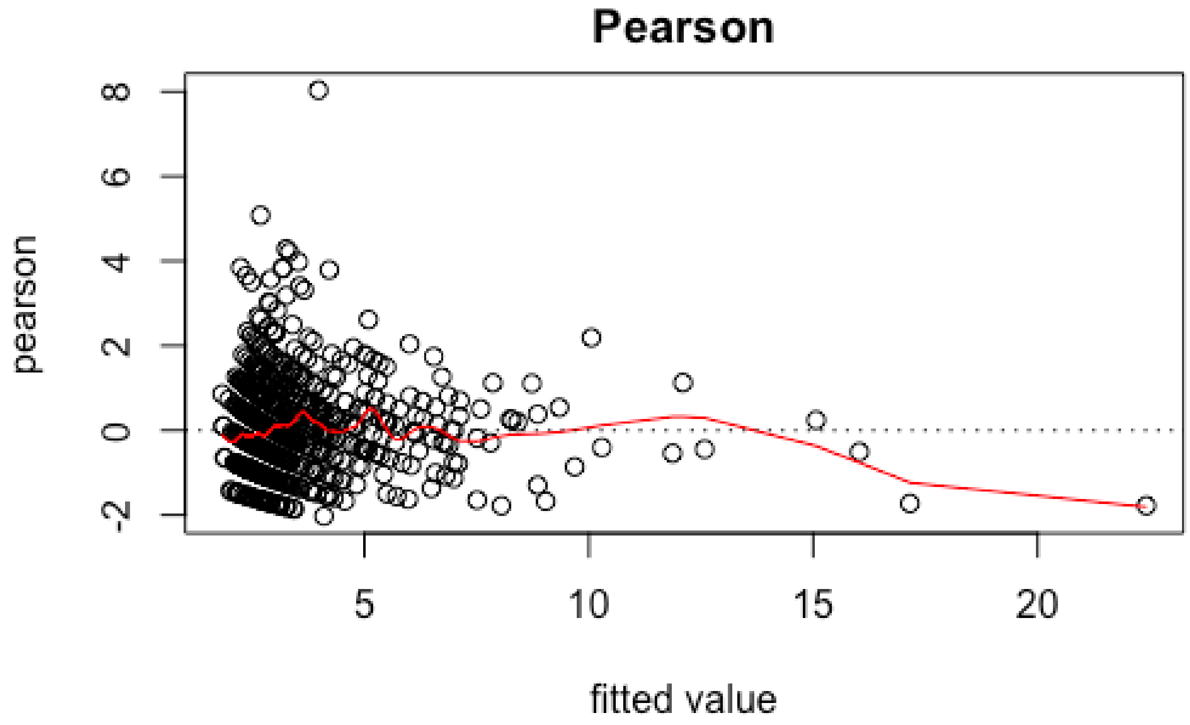


Figure 13: (initial ischemic)

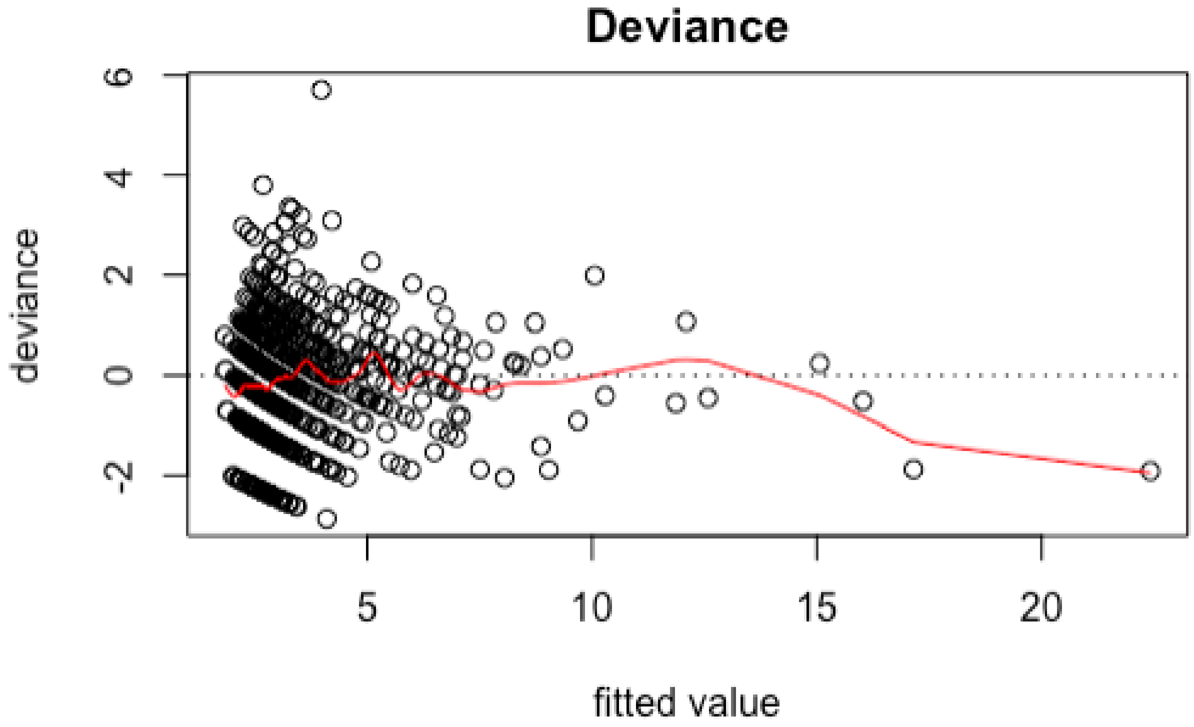


Figure 14: (final ischemic)

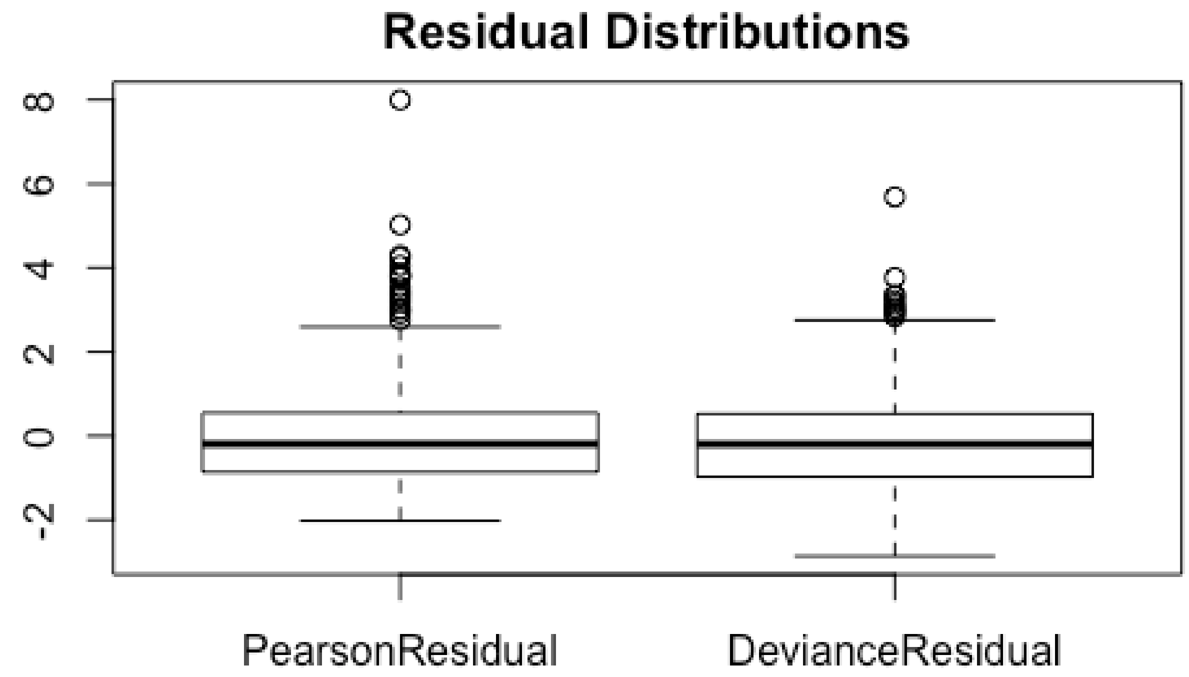


Figure 15: (final ischemic)

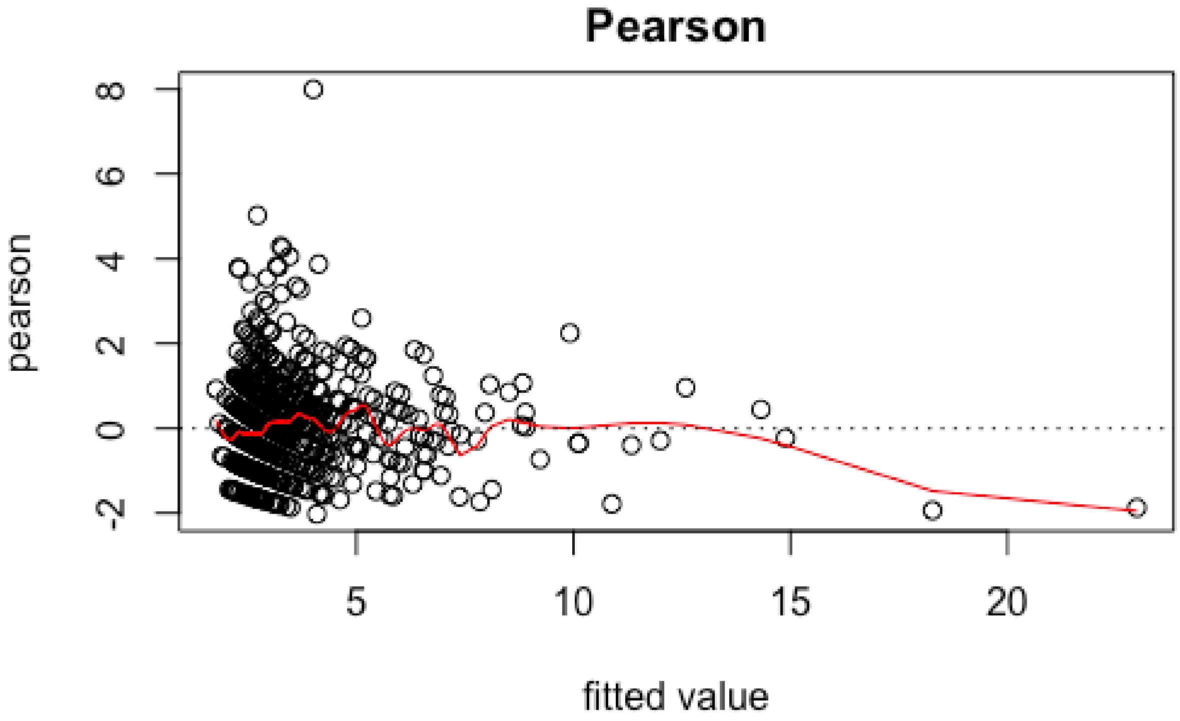
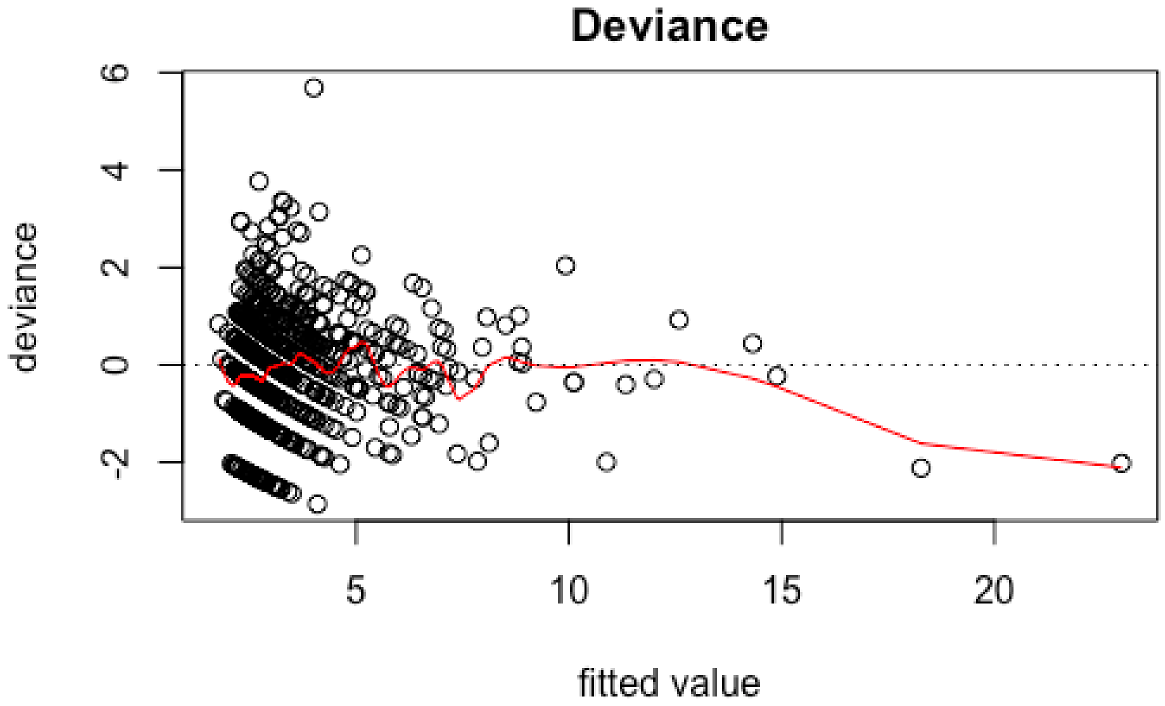


Figure 16: (final ischemic)



```
#PART II


library("readxl")
ischemic = read_excel("/Users/Troy/Downloads/ischemic.xlsx")
fit.poisson = glm(visits ~ cost + age + gender + inter + drugs + complication
s + comorbidities + duration + cost*age + cost*gender + cost*inter + cost*com
plications + cost*comorbidities + cost*duration + age*inter + age*drugs + age
*complications + age*comorbidities + age*duration + inter*drugs + inter*compl
ications + inter*comorbidities + inter*duration + complications*comorbidities
 + comorbidities*duration, data = ischemic, family = poisson())
summary(fit.poisson)
#Diagnostics
PearsonResidual = residuals(fit.poisson, type = "pearson")
DevianceResidual = residuals(fit.poisson, type = "deviance")
Dsquare = sum(PearsonResidual)
Gsquare = sum(DevianceResidual)
# of betas minus 1
Dsquare > qchisq(0.95, length(fit.poisson)-24)
Gsquare > qchisq(0.95, length(fit.poisson)-24)
boxplot(cbind(PearsonResidual, DevianceResidual), labels = c("Pearson", "Devi
ance"), main = "Residual Distributions")
plot(fit.poisson$fitted.values,PearsonResidual,xlab="fitted value",ylab="pear
son", main = "Pearson")
fit.cv = smooth.spline(fit.poisson$fitted.values,PearsonResidual, cv = FALSE,
spar=0.9) # CV fit
lines(fit.cv, col = "red")
```

```
abline(h=0,lty=3)
plot(fit.poisson$fitted.values,DevianceResidual,xlab="fitted value",ylab="dev
iance", main = "Deviance")
fit.cv = smooth.spline(fit.poisson$fitted.values,DevianceResidual, cv = FALSE
,spar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
library(lawstat)
runs.test(y = PearsonResidual, plot.it = FALSE)
ischemic.step = stepAIC(fit.poisson, direction = "backward", k=2)
ischemic.step$anova
fin.ischemic = glm(visits ~ cost + age + gender + inter + drugs + complicatio
ns + duration + cost*inter + cost*duration + age*complications + age*duration
 + inter*drugs + inter*complications + inter*duration, data = ischemic, famil
y = poisson())
summary(fin.ischemic)
#Diagnostics
PearsonResidual = residuals(fin.ischemic, type = "pearson")
DevianceResidual = residuals(fin.ischemic, type = "deviance")
Dsquare = sum(PearsonResidual)
Gsquare = sum(DevianceResidual)
# of betas minus 1
Dsquare > qchisq(0.95, length(fin.ischemic)-24)
Gsquare > qchisq(0.95, length(fin.ischemic)-24)
boxplot(cbind(PearsonResidual, DevianceResidual), labels = c("Pearson", "Devi
ance"), main = "Residual Distributions")
plot(fin.ischemic$fitted.values,PearsonResidual,xlab="fitted value",ylab="pea
rson", main = "Pearson")
fit.cv = smooth.spline(fin.ischemic$fitted.values,PearsonResidual, cv = FALSE
,spar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
plot(fin.ischemic$fitted.values,DevianceResidual,xlab="fitted value",ylab="de
viance", main = "Deviance")
fit.cv = smooth.spline(fin.ischemic$fitted.values,DevianceResidual, cv = FALS
E,spar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
library(lawstat)
runs.test(y = PearsonResidual, plot.it = FALSE)
#transformed
tischemic = ischemic
tischemic$cost = sqrt(tischemic$cost)
tischemic$age = sqrt(tischemic$age)
tischemic$inter = sqrt(tischemic$inter)
tischemic$drugs = sqrt(tischemic$drugs)
tischemic$complications = sqrt(tischemic$complications)
tischemic$comorbidities = sqrt(tischemic$comorbidities)
tischemic$duration = sqrt(tischemic$duration)
trans.poisson = glm(visits ~ cost + age + gender + inter + drugs + complicati
ons + comorbidities + duration, data = tischemic, family = poisson())
#+ cost*age + cost*gender + cost*inter + cost*complications + cost*comorbidit
ies + cost*duration + age*inter + age*drugs + age*complications + age*comorbi
dities + age*duration + inter*drugs + inter*complications + inter*comorbiditi
es + inter*duration + complications*comorbidities + comorbidities*duration
summary(trans.poisson)
tischemic.step = stepAIC(trans.poisson, trace = FALSE)
tischemic.step$anova
tischemic.step
```