# Statistics 138 Project
## PART I

Troy Lui
Due: December 12, 2018

## PART I: Birth Rates

### I. Introduction:

For the baby.xls file, data was given that showed the low or not low birth weights in infants. Accompanied with this, data about mother's age, weight, smoking status, history of pre-mature labor, history of hypertension, and number of visits during the first trimester were also taken. A little more specifically, weight is measured in pounds and visits can be any number greater than 0. Smoking, pre-mature labor, and hypertension variables were listed as either TRUE or FALSE. Lastly, the birth variable had 0 representing no low birth weight and 1 representing low birth weight. The goal of the data is to investigate whether birth weights of infants is related to any of the mother's statuses provided. In short, we are to fit a model to the birth variable.

### II. Materials and Methods:

To start, the type of statistical method that was used within the analysis was logistic regression. Through logistic regression, a linear model was obtained; however, do note that the data for the variable birth is highly imbalanced since there are 59 0's and 130 1's. To begin with, the general strategy that was to fit a linear model that had similar form of $B_0+B_1X_1+\ldots+B_nX_n$ with interactions of the variables age and weight, weight and hypertension, and weight and pre-mature labor. Aside from these given interactions that were to be considered, additional interactions that were put into a saturated initial model included hypertension and smoke, pre-mature labor and age, hypertension and age, weight and smoke, and smoke and age. Therefore, the saturated initial model included birth as a function of (age + weight + smoke + pre-mature labor + hypertension + visits + age*weight + weight*hypertension + weight*pre-mature labor + hypertension*smoke + pre-mature labor*age + hypertension*age + weight*smoke + smoke*age). When taking the summary statistics of the initial frame, figure 1 was produced, which includes the estimate, standard error, z value, and p value of the intercept and each beta variable within it [reference figure 1 appendix].

When checking the diagnostics of the initial analysis, the Pearson residual and Deviance residuals were found. From this, both distributions were plotted in side-by-side boxplots to determine if there is any lack of fit in the model. From figure 2, we see that the Pearson residual distribution and deviance residual distribution match each other quite well, which suggests no indication of a lack-of-fit. Other diagnostics used include plotting residuals (deviance or pearson) against fitted values of the saturated model [reference figure 3 & 4 appendix]. As shown, the red smoothing line does not quite follow 0, which indicates the possibility of a lack-of-fit. Lastly, using runs test with a null hypothesis that there are no systematic patterns, we received a significantly small p-value of 3.476e-09, which indicates a lack-of-fit in the model. All in all, running through all diagnostics is important since, as shown, the first diagnostic of boxplots showed no indication of lack-of-fit; however, the last two diagnostics said otherwise.

## III. Results:

When discovering the lack-of-fit from the diagnostics of the saturated model, the stepAIC( ) function was applied to the saturated model to receive a final model. When applying the step function, 9 variables were removed from the saturated model to create a model with the lowest AIC. The initial model ended up keeping six variables; therefore, the final birth model was a function of the betas age + weight + smoke + pre-mature labor + hypertension + age*hypertension.

In the end, the final model ended up being:

$$\text{Birth} = -2.342043 + 0.073073X_1 + 0.016360X_2 - 0.514877X_3 - 1.809716X_4 + 3.214987X_5 - 0.222579X_1X_5$$

[final model summary statistics: figure 8 appendix]

For the interpretation of the model, a higher age, weight, and history of hypertension are all variables that increase birth weight while smoking, a history of pre-mature labor, and the interaction variable between age and history of hypertension are variables that decrease birth weight.

Taking a look at the figure 6 and 7, the smoothed line looks a little more controlled along 0 when compared to the saturated model. Although, the plots do seem to oscillate, which could be a sign of potential lack-of-fit. Also, the boxplots, like the saturated model, are have no indication of lack-of-fit.

When using the final model to estimate the percentage of correct classification, we used the original data, baby.xls, and fitted it to the final model. Once fitted, we saw how any of the birth variables matched the newly predicted variable. 134 results ended up to be TRUE and 55 results ended up to be FALSE. With this in mind, the estimated percentage of correct classification using the baby.xls data is approximately 70.9%.

## IV. Conclusion and Discussion:

In conclusion, the final model had a fairly decent prediction percentage of the actual baby.xls data. With this in mind, running a completely saturated model within the stepAIC function could output a model with even lower AIC then the final model produced here. Also, some points to note is that the stepAIC function actually got rid of an initial variable (visits) while producing AIC; however, the variable was ultimately dropped since the variable visits also did not have anything to do with any significant interaction variables. Another interesting point is that when figuring out an initial model, it doesn't hurt to make the initial model more and more saturated since the stepAIC function will eventually filter out the variables that aren't significant.

# Appendix

Figure 1: (Initial baby model summary statistics)

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.7632455 | 4.8390914 | -0.364 | 0.716 |
| age | 0.0071591 | 0.2124432 | 0.034 | 0.973 |
| weight | 0.0092870 | 0.0391938 | 0.237 | 0.813 |
| smoke1 | 3.1862959 | 2.4836124 | 1.283 | 0.200 |
| pre1 | -6.6859538 | 4.7300498 | -1.414 | 0.158 |
| hyp1 | 3.3339549 | 6.3511530 | 0.525 | 0.600 |
| visits | 0.0459113 | 0.1867255 | 0.246 | 0.806 |
| age:weight | 0.0006206 | 0.0016961 | 0.366 | 0.714 |
| weight:hyp1 | 0.0057843 | 0.0229329 | 0.252 | 0.801 |
| weight:pre1 | 0.0050373 | 0.0250341 | 0.201 | 0.841 |
| smoke1:hyp1 | -0.5481738 | 1.8627042 | -0.294 | 0.769 |
| age:pre1 | 0.1715404 | 0.1223920 | 1.402 | 0.161 |
| age:hyp1 | -0.2522931 | 0.2523773 | -1.000 | 0.317 |
| weight:smoke1 | -0.0144994 | 0.0152507 | -0.951 | 0.342 |
| age:smoke1 | -0.0850124 | 0.0790718 | -1.075 | 0.282 |

Figure 2:

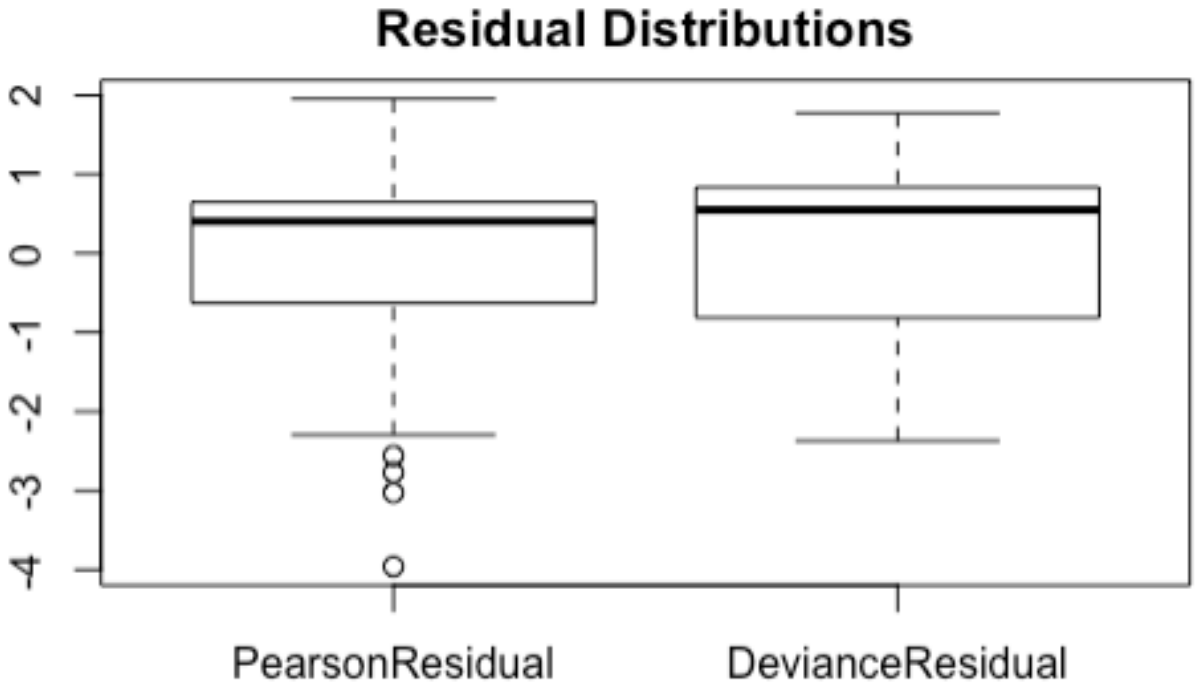


Figure 3:

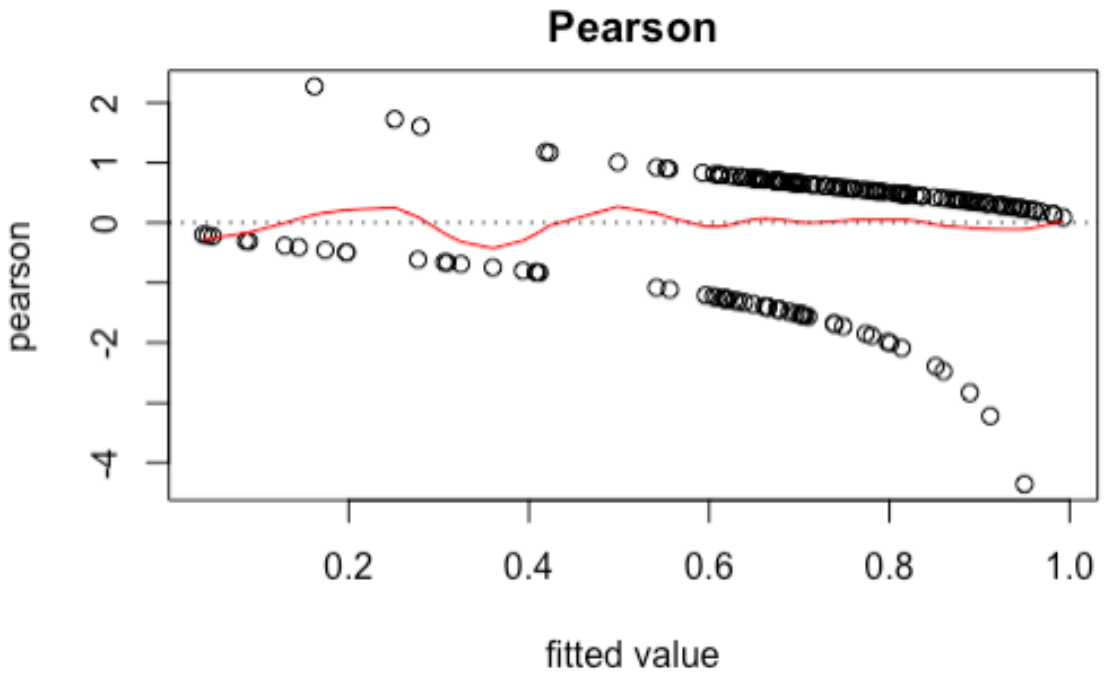



Figure 4:

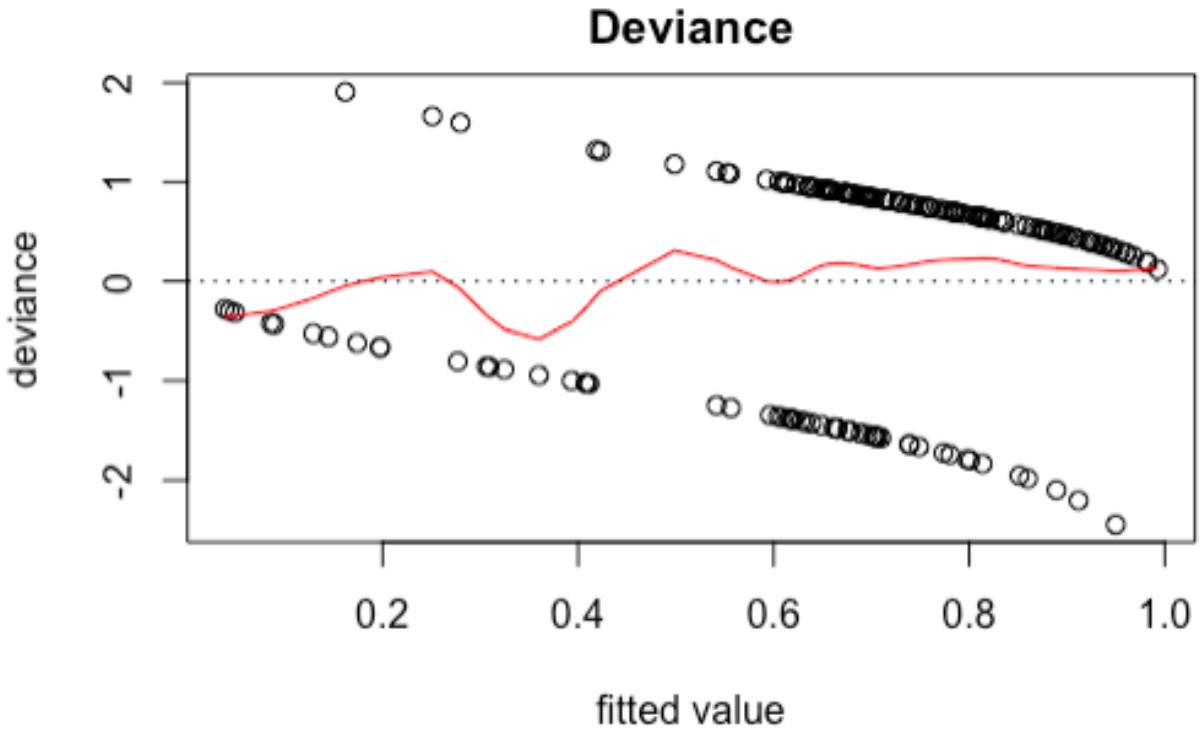



Figure 5:

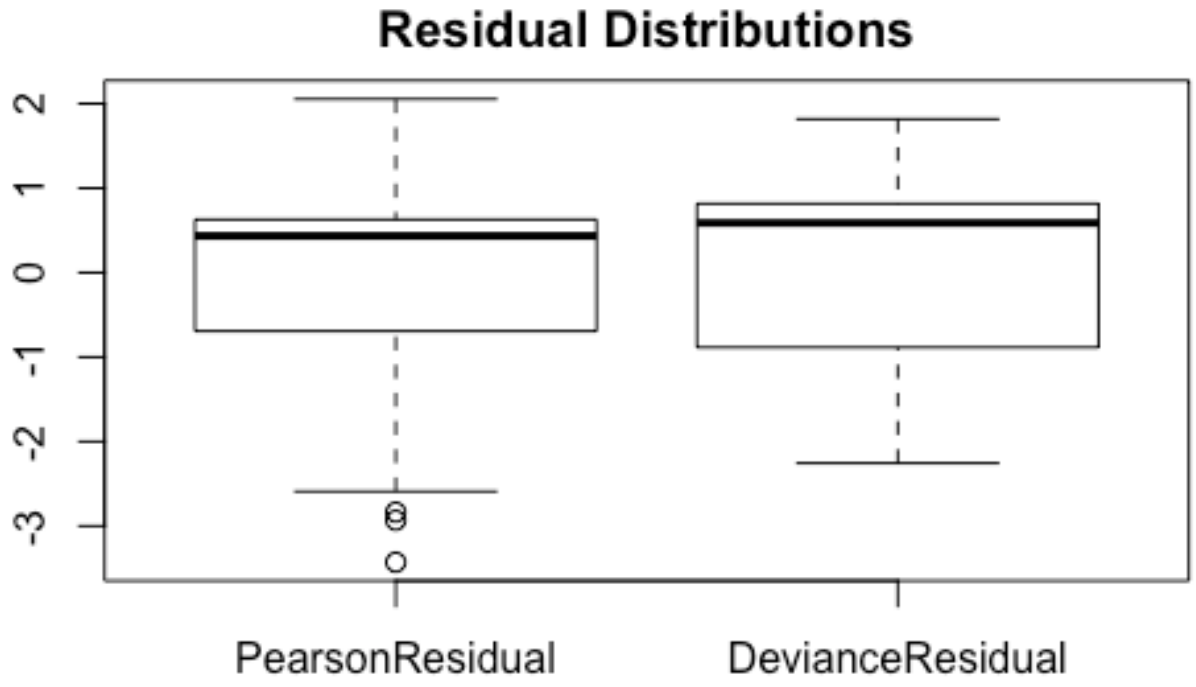

Residual Distributions
2
1
0
-1
-2
-3
PearsonResidual
DevianceResidual

Figure 6:

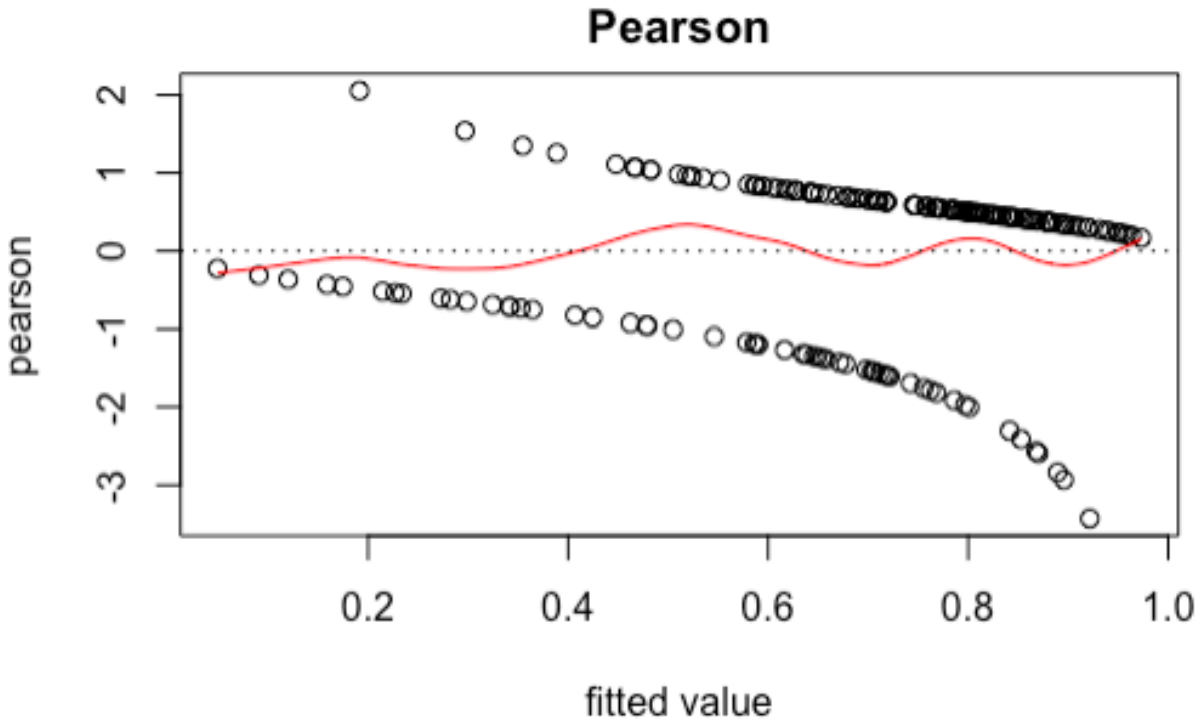

Pearson
pearson
2
1
0
-1
-2
-3
0.2
0.4
0.6
0.8
1.0
fitted value

Figure 7:

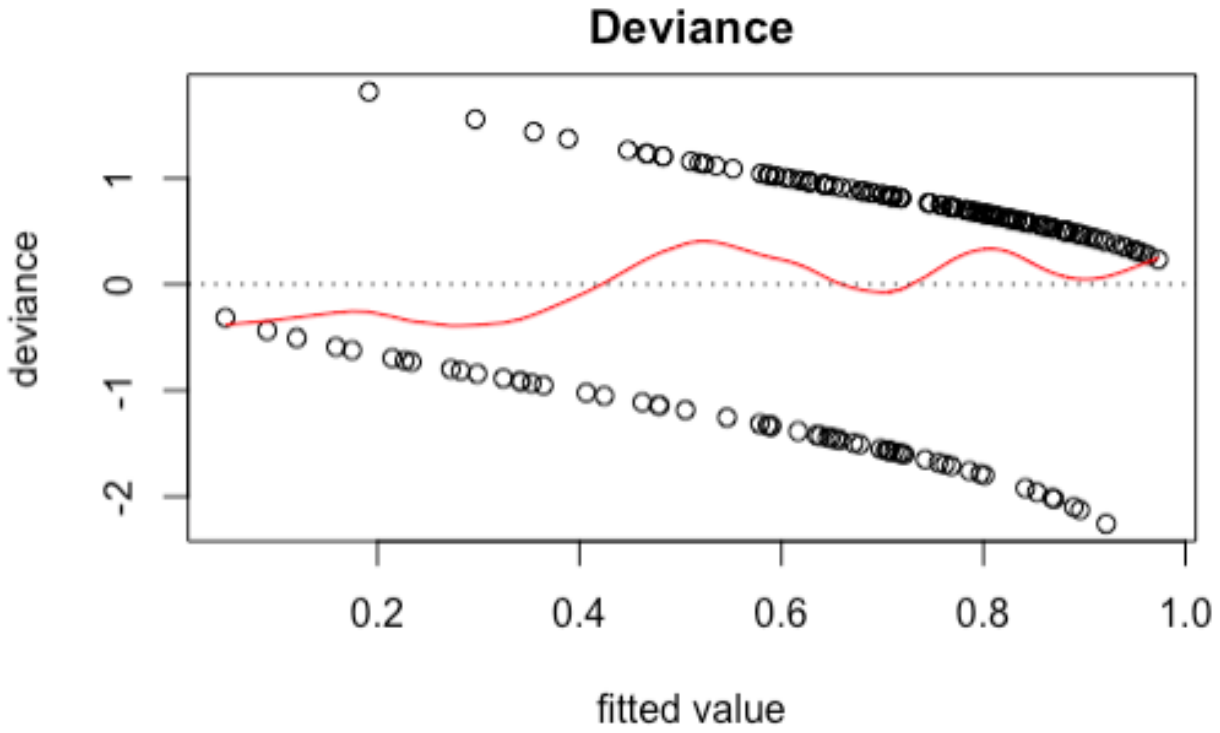



Figure 8: (Final baby model summary statistics)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.34204256 | 1.147235118 | -2.0414669 | 0.0412044368 |
| age | 0.07307318 | 0.038034357 | 1.9212413 | 0.0547012876 |
| weight | 0.01636045 | 0.007002709 | 2.3363027 | 0.0194754685 |
| smoke | -0.51487744 | 0.351001016 | -1.4668830 | 0.1424078750 |
| pre | -1.80971628 | 0.517780667 | -3.4951407 | 0.0004738119 |
| hyp | 3.21498663 | 3.991727490 | 0.8054124 | 0.4205817657 |
| age:hyp | -0.22257931 | 0.177334785 | -1.2551362 | 0.2094293117 |

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)

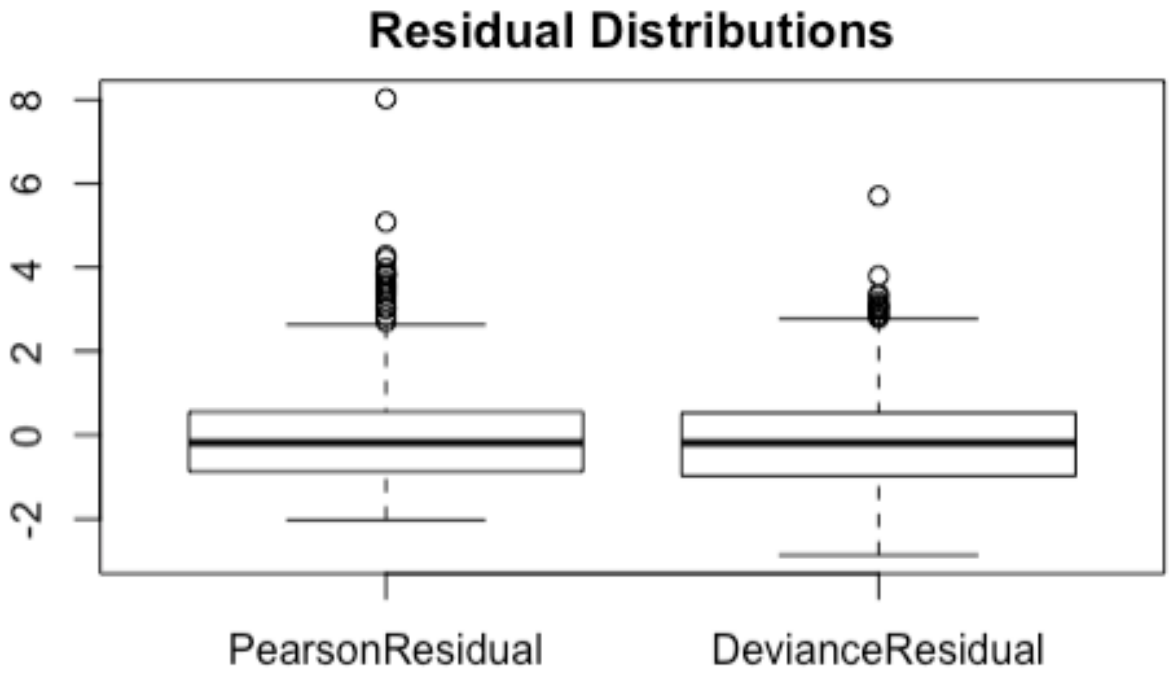



Figure 12: (initial ischemic)

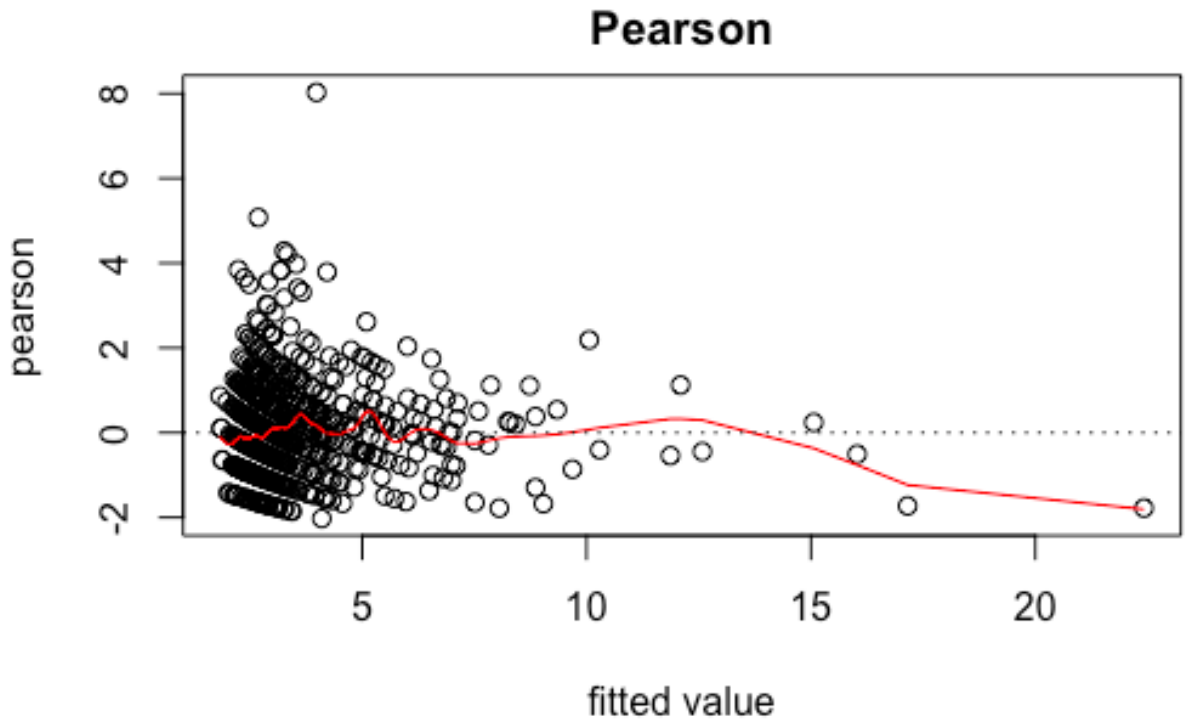



Figure 13: (initial ischemic)

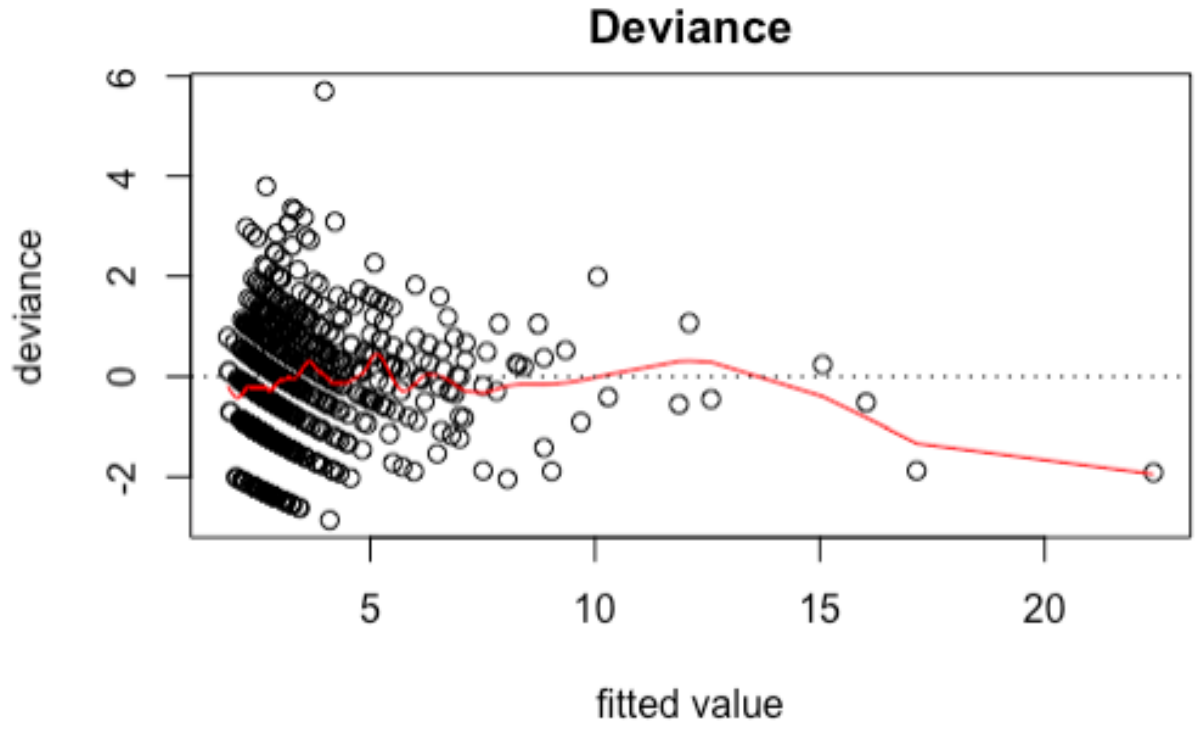



Figure 14: (final ischemic)

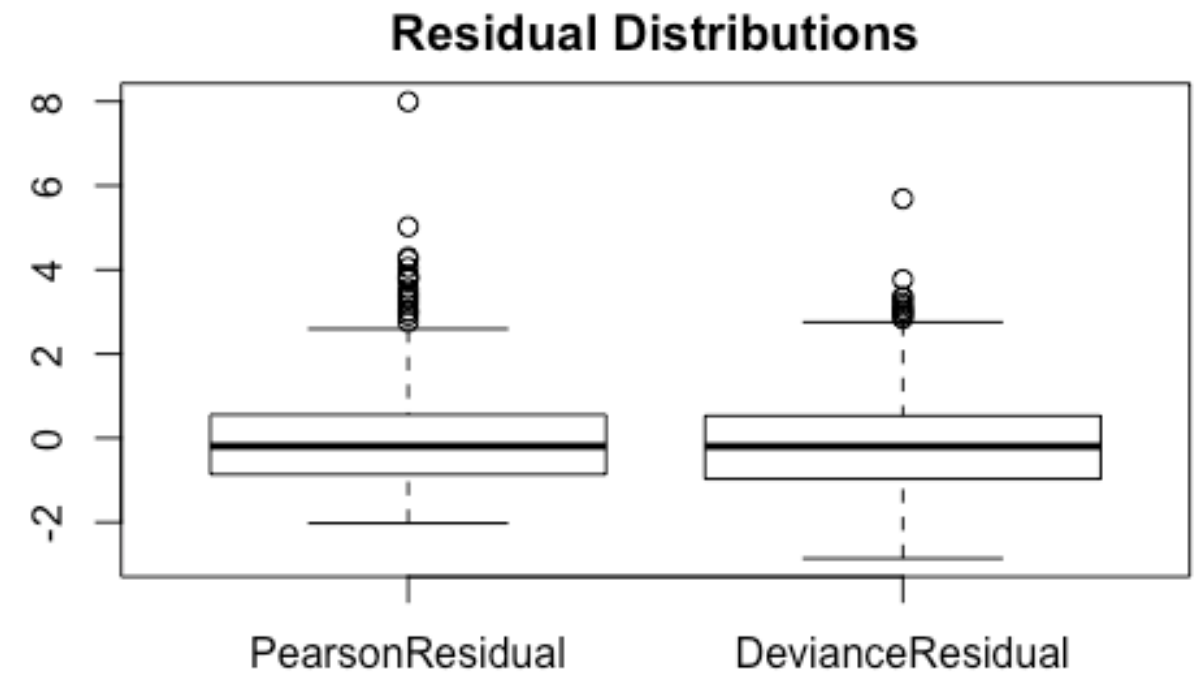



Figure 15: (final ischemic)

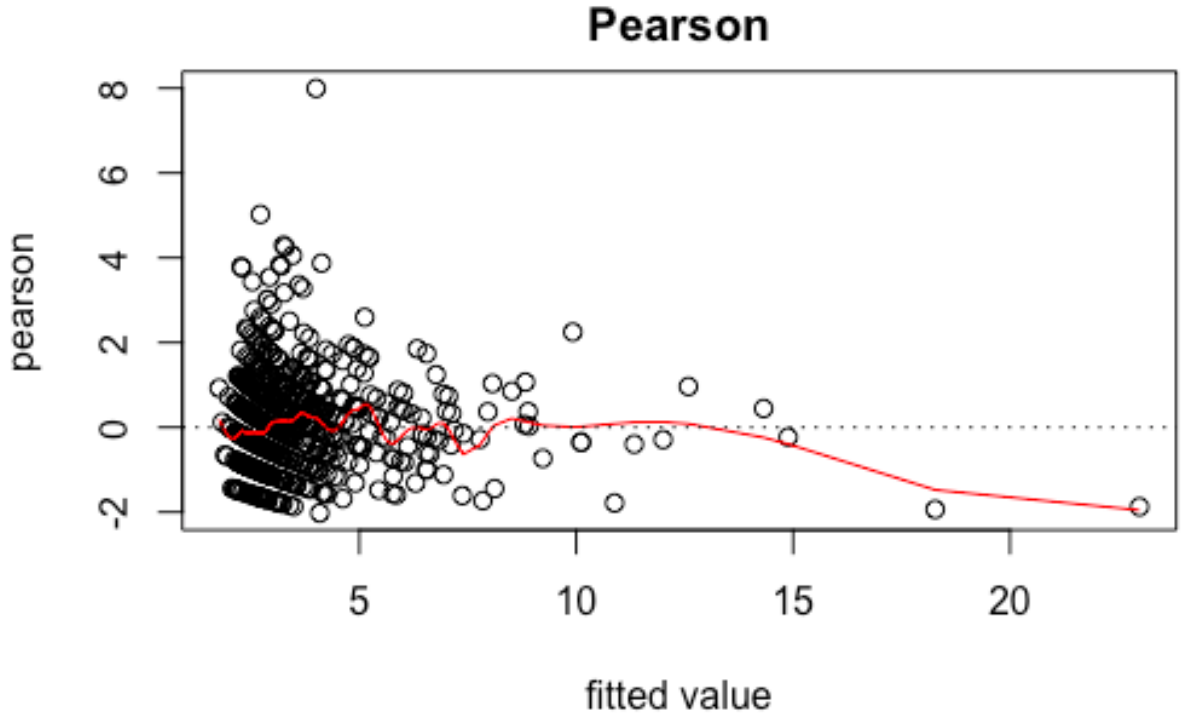
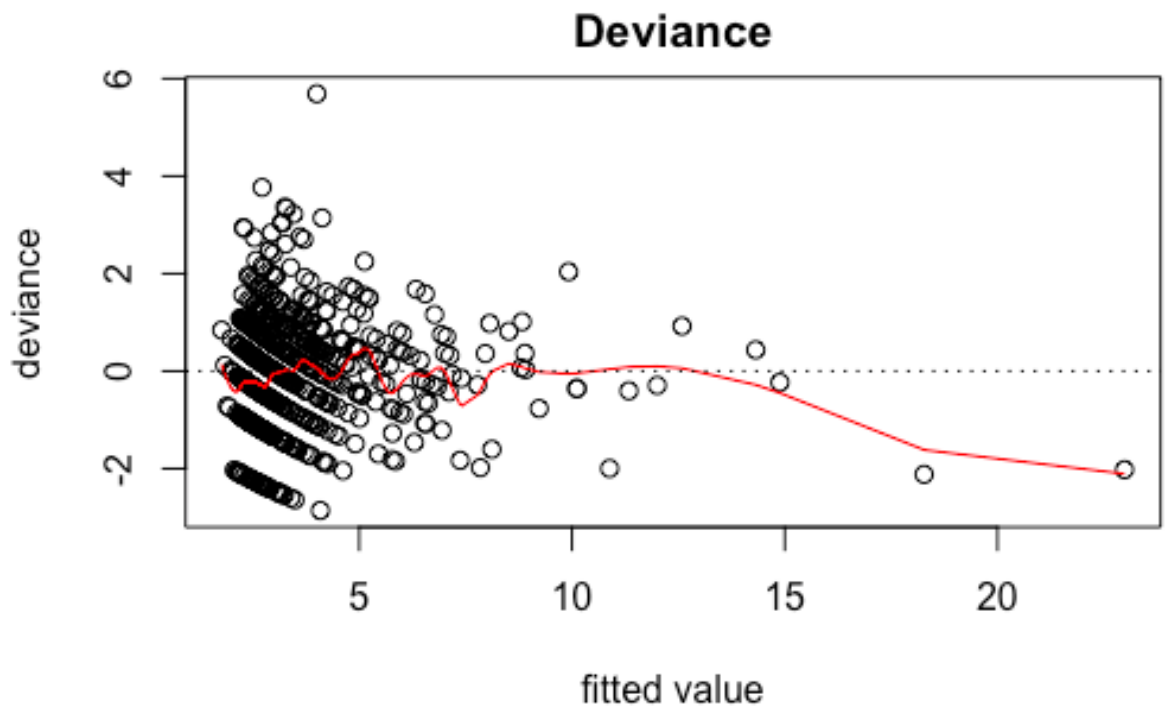



Figure 16: (final ischemic)




```
#PART I

library(MASS)
library("readxl")
baby = read_excel("/Users/Troy/Downloads/baby.xls")
babyf = baby
babyf$smoke[babyf$smoke == "yes"] = 1
babyf$smoke[babyf$smoke == "no"] = 0
babyf$pre[babyf$pre == "yes"] = 1
babyf$pre[babyf$pre == "no"] = 0
babyf$hyp[babyf$hyp == "yes"] = 1
babyf$hyp[babyf$hyp == "no"] = 0
#babyf$visits[babyf$visits >= 2] = "2+"
inbabymod = glm(birth~age + weight + smoke + pre + hyp + visits + age*weight
+ weight*hyp + weight*pre + hyp*smoke + pre*age + hyp*age + weight*smoke + sm
oke*age, family = binomial(), data = babyf)
summary(inbabymod)
#initial saturated model
#diagnostics for initial model
PearsonResidual = residuals(inbabymod, type = "pearson")
DevianceResidual = residuals(inbabymod, type = "deviance")
Dsquare = sum(PearsonResidual)
Gsquare = sum(DevianceResidual)
# of betas minus 1
Dsquare > qchisq(0.95, length(inbabymod)-14)
Gsquare > qchisq(0.95, length(inbabymod)-14)
```

```r
boxplot(cbind(PearsonResidual, DevianceResidual), labels = c("Pearson", "Devi
ance"), main = "Residual Distributions")
#more diagnostics
plot(inbabymod$fitted.values,PearsonResidual,xlab="fitted value",ylab="pearso
n", main = "Pearson")
fit.cv = smooth.spline(inbabymod$fitted.values,PearsonResidual, cv = FALSE,sp
ar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
#more diagnostics
plot(inbabymod$fitted.values,DevianceResidual,xlab="fitted value",ylab="devia
nce", main = "Deviance")
fit.cv = smooth.spline(inbabymod$fitted.values,DevianceResidual, cv = FALSE,s
par=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
#last diagnostic
library(lawstat)
runs.test(y = DevianceResidual, plot.it = FALSE)
#choosing the model from the saturated model
baby.step = stepAIC(inbabymod, trace = FALSE)
baby.step$anova
#final model
baby.fit = glm(formula = birth ~ age + weight + smoke + pre + hyp + age*hyp,
family = binomial(), data = babyf)
summary(baby.fit)
#estimate percentage of correct classification
babyf$age = as.numeric(babyf$age)
babyf$weight = as.numeric(babyf$weight)
babyf$smoke = as.numeric(babyf$smoke)
babyf$pre = as.numeric(babyf$pre)
babyf$hyp = as.numeric(babyf$hyp)
babyf$predict = (-2.342043 + 0.073073*babyf$age + 0.016360*babyf$weight - 0.5
14877*babyf$smoke - 1.809716*babyf$pre + 3.214987*babyf$hyp - 0.222579*babyf$
age*babyf$hyp)
babyf$predict[babyf$predict>.5] = 1
babyf$predict[babyf$predict<=.5] = 0
babyf$result = (babyf$birth == babyf$predict)
#table(babyf$result)
#55 false 134 true
#134/(134+55) = 0.7089947
#Diagnostics
PearsonResidual = residuals(baby.fit, type = "pearson")
DevianceResidual = residuals(baby.fit, type = "deviance")
Dsquare = sum(PearsonResidual)
Gsquare = sum(DevianceResidual)
# of betas minus 1
Dsquare > qchisq(0.95, length(baby.fit)-6)
Gsquare > qchisq(0.95, length(baby.fit)-6)
boxplot(cbind(PearsonResidual, DevianceResidual), labels = c("Pearson", "Devi
```

```
ance"), main = "Residual Distributions")
plot(baby.fit$fitted.values,PearsonResidual,xlab="fitted value",ylab="pearson
", main = "Pearson")
fit.cv = smooth.spline(baby.fit$fitted.values,PearsonResidual, cv = FALSE,spa
r=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
plot(baby.fit$fitted.values,DevianceResidual,xlab="fitted value",ylab="devian
ce", main = "Deviance")
fit.cv = smooth.spline(baby.fit$fitted.values,DevianceResidual, cv = FALSE,sp
ar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
library(lawstat)
runs.test(y = PearsonResidual, plot.it = FALSE)

#Diagnostics
PearsonResidual = residuals(tischemic.step, type = "pearson")
DevianceResidual = residuals(tischemic.step, type = "deviance")
Dsquare = sum(PearsonResidual)
Gsquare = sum(DevianceResidual)
# of betas minus 1
Dsquare > qchisq(0.95, length(tischemic.step)-24)
Gsquare > qchisq(0.95, length(tischemic.step)-24)
boxplot(cbind(PearsonResidual, DevianceResidual), labels = c("Pearson", "Devi
ance"), main = "Residual Distributions")
plot(tischemic.step$fitted.values,PearsonResidual,xlab="fitted value",ylab="p
earson", main = "Pearson")
fit.cv = smooth.spline(tischemic.step$fitted.values,PearsonResidual, cv = FAL
SE,spar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
plot(tischemic.step$fitted.values,DevianceResidual,xlab="fitted value",ylab="
deviance", main = "Deviance")
fit.cv = smooth.spline(tischemic.step$fitted.values,DevianceResidual, cv = FA
LSE,spar=0.9) # CV fit
lines(fit.cv, col = "red")
abline(h=0,lty=3)
library(lawstat)
runs.test(y = PearsonResidual, plot.it = FALSE)
```