PROJECT STA 138

Jared Yu 914640019

Burman
STA 138

(i) Introduction

The first dataset is from the file called baby.xls. The observations include descriptions of women and their babies. The goal is to try and determine if any of the available variables are suitable for predicting whether the infant will have a low birth weight due to these variables. Logistic regression will be used to analyze the model and determine an appropriate model. The reason is that there is the goal of binary classification.

The second dataset is from the file ischemic.xlsx. In this dataset, the observations are for health insurance consumers who had claims related to ischemic, also known as heart disease. The date of the dataset ranges from the beginning of 1998 to the end of 1999. Here the modeling will be related to a response variable which is the number of E.R. visits, and the other variables will serve as predictors. The goal is to model the mean of the variable as a function of 8 other variables within the dataset. Since we are dealing with count data, Poisson regression will be utilized for analysis and model selection.

(ii) Materials and Methods

In the first dataset for the file 'baby.xls,' there are a total of 189 observations and 7 columns. The columns include: age, weight, smoke, pre, hyp, visits, and birth. Some variables have a character datatype, and have binary values signifying 'yes' or 'no.' The other variables have datatype double and represent a numeric description for an observation. In Logistic regression it's important that the proportion of the binary variable is balanced. A table of the proportion of 1's and 0's indicating low birthweight for the dataset can be found in $A.1$ of the Appendix. It is apparent that the ratio is close to 70/30, which is not too extreme. Too extreme of a proportion would be closer to 90/10.

The methods that will be utilized in the first dataset will involve first using Logistic regression to examine the relationship of the independent variables which will serve as predictors for the response variable. Once the model is first fitted, plots will be used to examine the linearity of certain continuous variables with the estimated probabilities given by the fitted model. Afterwards, interaction terms will also be analyzed. Hypothesis tests can help to show whether certain interaction terms are useful for a model. The Likelihood Ratio statistic, G^2, which is a metric to understand if a new model has a greater Goodness-of-Fit, will also be implemented. Finally, Stepwise regression is a process of understanding what combination of predictors are best suited for predicting a response. The criterion used in this report is the AIC criterion, where the lower the AIC, the better.

The second dataset is 'ischemic.xlsx,' and in this dataset all the variables are of datatype double. There is a total of 788 observations and 9 different columns. The 9 columns include: cost, age, gender, inter, drugs, complications, comorbidities, duration, and visits. All the variables provide a numeric description of an observation. An important assumption with Poisson regression is that the mean and variance are roughly similar. In the Appendix ($A.2$) is a table of the sample mean and sample variance for the response variable 'visits.' The table shows that there is a rather large difference between the two, where the sample variance is roughly twice the sample mean. However, in an email conversation with the professor, it was stated that it was not necessary to calculate this since it was not covered in class. The actual method for checking the mean and variance involves using a calculation on the fixed $X_1, X_2, \cdots, X_n$, however this calculation was never explained in the notes or in class. Therefore, we will not worry about this assumption of the Poisson regression for the data during the modeling process, and rather it will be addressed in the end.

(iii) Results

First Dataset: First, we will begin by analyzing the first dataset through Logistic regression. The model logit of π, the probability of an infant having a low birthweight is: $\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$, where π' is the logit of π (full model). The variable X_1represents age of the mother, X_2 represents weight of the mother, X_3 represents smoking during pregnancy, X_4 represents previous low weight births, X_5 represents hypertension, and X_6 represents number of visits during the first trimester. A table which gives the parameter estimates, exponential parameter estimates, standard errors, and p-values for each of the $\beta's$ can be found in $B.1$ of the Code Appendix. The estimated logistic regression function is therefore:

$$\hat{\pi}(X_1, X_2, X_3, X_4, X_5, X_6) = \frac{\exp(-2.021488 + 0.059091X_1 + 0.016086X_2 - 0.513740X_3 - 1.798908X_4 - 1.772643X_5 + 0.032113X_6)}{1 + \exp(-2.021488 + 0.059091X_1 + 0.016086X_2 - 0.513740X_3 - 1.798908X_4 - 1.772643X_5 + 0.032113X_6)}.$$

The $\exp(\beta)$'s represent the change in odds of the infant having a low birthweight for each variable. For instance, the odds of the infant having a low birthweight at Age $(X_1 + 1)$ are estimated to be 1.060872 times the odds of having a low birthweight at Age X_1.

The continuous independent variables which are: 'age,' 'weight,' and 'visits,' have been plotted in a scatter plot against the estimated probability of having a low birthweight. The plots can be seen in $A.3$ of the appendix. It is apparent that the variables 'age' and 'weight' follow a much more linear trend. It is possible that certain transformations such as log would benefit the appearance of their linearity. The log of the variables are shown below, however it is not certain that they will be used later in place of the existing variables due to the possibility for interaction terms later. The 'visits' variable however lacks a strong linear appearance. This would help explain why the p-value for this variable is higher than the rest (β_6 in $B.1$ of the Code Appendix). In the same section are the categorical variables: 'smoke,' 'pre,' and 'hyp.' The boxplots indicate that 'pre' and 'hyp have the strongest difference in the distributions within the binary variables.

The following is a Goodness-of-Fit test included in the summary of the initial fit ($B.1$ in the Code Appendix). $G^2 = 202.15$ with $df = 182$. We wish to test H_0: $\pi_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}$ for all i, vs H_1: $\{\pi_i\}$ do not lie on a straight line. The χ^2 test comes back False, and so we fail to reject the null hypothesis with a p-value of 0.146. The conclusion is that a linear logistic fit is reasonable.

Interaction terms: Originally there were 6 predictors, and the response is Y: $0\ (no\ low\ birthweight), 1\ (low\ birthweight)$. Using some of the original predictors, we will now include some interactions: age*weight, weight*hyp, and weight*pre. Thus, the logistic regression has 9 predictor variables called: $X_1, \cdots, X_9$. Denoting $\pi_i = P(Y_i = 1)$ and π_i' as the logit transformation, the model is now:

$$\pi_i' = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{i9} X_9, i = 1, \cdots, n,$$

where $n = 189$, $X_{i1}, \cdots, X_{i6}$ have the same values as before, $X_{i7} = X_{i1}X_{i2}, X_{i8} = X_{i2}X_{i5}, X_{i9} = X_{i2}X_{i4}$

In the interaction terms: X_1X_2 represents age*weight, X_2X_5 represents weight*hyp, and X_2X_4 represents weight*pre. The hypothesis test here is: H_0: $\beta_7 = \beta_8 = \beta_9 = 0$ against H_1: at least one of them don't equal 0. The results of the test can be seen in $A.4$ of the Appendix.

These statistics were calculated under $\chi^2_{0.95,3}$ with $\alpha = 0.05$ (a table of the results can be found in $A.4$ of the Appendix). The Likelihood Ratio statistic, G^2, was not larger than the critical value under the χ^2distribution, and so we fail to reject the null hypothesis. Also, the p-value is significantly larger than 0.05. From this we can conclude that we do not need to include the interaction terms in the final model.

It is uncertain why the variable performed poorly. It seems that the interaction of age and weight could somehow be indicative of whether the baby is of low birthweight. For instance, mothers with low weights would possibly have low birthweight babies. It is possible that there are not enough observations to notice any sort of interaction effect due to the pattern being subtle. The interaction terms will be utilized next as part of a more general model to be placed into the Stepwise regression.

Stepwise Regression: Another strategy for trying to build a model is to use Stepwise regression. There are three possible directions: forward, backward, and both within R. However, in this case the only model utilized will be 'backwards,' in accordance with what was taught in class. Also, the metric used is AIC criterion. The entire trace of the stepwise function can be found in $B.2$ of the Code Appendix.

Using the same model with interaction terms (the final model in the stepwise is process is the same for both models with and without interaction terms, see $B.3$ in the Code Appendix), the final model chosen by AIC criterion is $\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$. The entire summary is included in $B.2$ of the Code Appendix. Included are the parameter estimates, standard errors, z-values, and p-values for the final model. With this model, the variables `visits` is dropped, along with all the interaction terms. This is logical due to the hypothesis test that were performed earlier. The variable 'visits' had the least linear relationship with the response variable, and so it is unlikely that it would be effective during the modeling process. The interaction terms performed poorly, so it makes sense that they are excluded. The AIC of the final model is 214.19, less than the original model's AIC of 216.15. Therefore, we choose the model chosen from the Stepwise regression over the original model. The estimated logistic regression is:

$$\hat{\pi} = \frac{\exp(-2.031969 + 0.060319\mathrm{X}_1 + 0.016154X_2 - 0.518366X_3 - 1.794038X_4 - 1.782710X_5)}{1 + \exp(-2.031969 + 0.060319\mathrm{X}_1 + 0.016154X_2 - 0.518366X_3 - 1.794038X_4 - 1.782710X_5)}$$

A Goodness-of-Fit test for the original model compared to the Stepwise regression model can be found at the end of $B.2$ in the Code Appendix. Based on the model for the Stepwise regression, the $H_0: \beta_6 = 0$ vs $H_1: \beta_6 \neq 0$. The test statistic is not significantly large, and so we fail to reject H_0. The p-value is quite large at 0.841. The conclusion is that we may drop variable X_6.

An additional model which was fitted was the same model that was just used in the Stepwise regression, except the following terms had their log taken first (including their second order versions): 'age' and 'weight.' The AIC for the model is 214.23, better than the original model, but worse than the model previously chosen by Stepwise regression. The summary of the model can be seen in $B.4$ of the Code Appendix.

Residuals: Further plots showing the Deviance and Pearson residuals including their standardized forms are included in $A.5$ of the Appendix. Also included is a table which shows the range of the data for each of the different plots ($A.6$ of the Appendix). Looking at the plots, it appears that most of the observations fall within ± 3 which is a positive result for the model that was chosen. There is

however one outlier in the Pearson residual plot, where the residual falls close to -3.5. This can be considered an outlier and does not put the model at risk. It is possible to drop these outliers and re-run the model, however that step will not be done in this report.

Additionally, a confusion matrix is created showing the comparison of the predicted values with the actual values from the dataset ($A.7$ in the Appendix). The Sensitivity of the model is $\frac{120}{130} = \frac{12}{13}$, the Specificity of the model is $\frac{21}{59}$. Although the model is powerful in predicting when an infant may be born with a low birthweight, it has issues in being able to predict when it is not the case. Therefore, the Specificity of the model is not highly effective. However, in practice Sensitivity is often what the emphasis is on, and so if the model can predict that a child may be of low birthweight action can be taken to help with the infant when it is born. The percentage of correct classification is taken by finding summing the correct predictions and dividing by the total number. This gives: $\frac{21+120}{189} = \frac{141}{189} = 0.7460317$. So roughly ~75% of the predictions are correct, however there is not a pre-existing metric to compare this to. So it is difficult to say if it is good or bad.

Interpretation of Features: Examining the estimated parameters and performing confidence intervals may help to understand what their meaning is within the model. Using the formula: $\hat{\beta}_i \pm 1.96s(\hat{\beta}_i)\ for\ i = 1, \cdots, 5$ gives a confidence interval for each of the β's associated with an independent variable at confidence level $\alpha = 0.05$ ($A.8$ in the Appendix). Two parameters which contain the value 0 at this confidence level are: 'age' and 'smoke.' Despite them being in the final model, their strength is not seen as very large. The variable 'weight' is positive, therefore it when controlling for other variables, as the weight of the mother is higher, there is a greater likelihood that the baby will be born with a low birthweight. The variable 'pre' and 'hyp' are both negative. This signifies that mothers who have a history of low birthweight babies and hypertension are less likely to have babies born in a low birthweight category.

Second Dataset: Next, we will examine the second dataset using Poisson regression. A Poisson regression model for this data is: $\mu = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8)$. The variable X_1 represents the total cost of all the claims per health insurance subscriber, X_2 represents the age of the health insurance subscriber in years, X_3 represents what gender the subscriber is, X_4 represents how many times interventions or procedures took place, X_5 represents number of drugs prescribed to a subscriber, X_6 represents number of complications that arose during the heart treatment, X_7 represents the number of other diseases present in the subscriber, X_8 represents the duration of the stay in days during the treatment. A table that gives the predicted parameters, exponential predicted parameters, standard errors, and p-values for each of the $\beta's$ can be found in $B.5$ of the Code Appendix. The $\exp(\beta)$'s represent the change in number of E.R. visits for each variable. For instance, the change in the number of E.R. visits at Cost $X + 1$ is about 1.000015 times the number of E.R. visits at Cost X.

The estimated Poisson regression function is therefore:

$$\hat{\mu} = \exp(4.994 \times 10^{-1} + 1.495 \times 10^{-5} X_1 + 6.724 \times 10^{-3} X_2 + 1.819 \times 10^{-1} X_3 + 1.007 \times 10^{-2} X_4 + 1.932 \times 10^{-1} X_5 + 6.125 \times 10^{-2} X_6 - 8.999 \times 10^{-4} X_7 + 3.529 \times 10^{-4} X_8)$$

A plot of each of the predictor variables have been plotted against the estimated counts ($A.9$ in the Appendix). It is apparent that the linearity of the plots is like the p-values which were seen in $B.5$.

Variables that have lower p-values, have plots which exhibit a more diagonal appearance. However, plots with higher p-values lack a significant expression of this diagonal effect. The 'gender' variable is plotted in a boxplot, where the difference between Male and Other show that the variable may be useful in predicting the response variable.

A Goodness-of-Fit test is also used on the current fitted model ($B.5$ in the Code Appendix). We wish to test $H_0: \log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8}$ for all i vs $H_1: \log(u_i)'s$ do not lie on a straight line. Ideally, all the counts would be 5 or larger, but this is not the case. This may be a sign that our model is not appropriate for the data. In the test, the test statistic is too large and so we reject H_0 in favor of the alternative. The p-value is very small in this case. So the conclusion is that the log-linear model $\mu = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8)$ is not reasonable for the data.

A transformation which is tested is taking the square root of all the predictor variables excluding gender. The resultant table of parameter estimates, standard errors, z-values, and p-values can be seen in $B.6$ of the Code Appendix. From the p-values that are given, in this model with the parameters square rooted, the strength is much weaker. Now an additional three variables, 'age,' 'gender,' and 'inter,' have p-values which are undesirable. Also, the AIC is shown to be ∞, and so the model will be disregarded.

Stepwise regression: Once again, Stepwise regression will be used for model selection. The Stepwise regression is used on the original model that was fitted first using all the first order variables. A table with the estimated parameters, standard errors, z-value, and p-values can be found in $B.7$ in the Code Appendix (the entire trace of the Stepwise regression can also be seen). The model chosen is:

$\hat{\mu} = \exp(5.208 \times 10^{-1} + 1.493 \times 10^{-5} X_1 + 6.334 \times 10^{-3} X_2 + 1.857 \times 10^{-1} X_3 + 1.025 \times 10^{-2} X_4 + 1.963 \times 10^{-1} X_5 + 3.453 \times 10^{-4} X_8)$

The p-values in this model are extremely low, and all the variables are quite effective. The AIC for the new model is 3268.1 which is lower than the original model's AIC of 3271. Therefore, this new model will be chosen as the final model.

A Goodness-of-Fit test for the original model compared to the Stepwise regression model can be found at the end of $B.7$ in the Code Appendix. Based on the model for the Stepwise regression, the $H_0: \beta_6 = \beta_7 = 0$ vs H_1: at least one is not 0. The test statistic is not significantly large at $\alpha = 0.05$, and the p-value is 0.577. So, we fail to reject H_0, and conclude that we may drop variables X_6 and X_7.

Residuals: The Deviance and Pearson residuals are plotted along with the standardized versions. The plots can be found in $A.10$ of the Appendix. Below is a table of the range of the residuals as well in $A.11$ of the Appendix. From the table in $A.11$, it is apparent that there are outliers which need to be considered. The plots in $A.10$ show that most of the data fall within the safe range of ± 3.5, however the tendency for some outliers to go too high means that the model is not perfect.

Looking at each of the different residual plots, first the positive residuals were subset from the data. Then, the residuals which were greater than 3.5 were subset, and then divided by the total number of positive residuals. This gives a percentage of outliers which are more severe than the rest of the residuals. In $A.12$ of the Appendix, a table is given which shows the proportion for each of the different residual types. It is apparent that only an insignificant amount of them are too large, and so it would be safe to conclude that the model fits the data well. In most cases it is less than 1% of the data,

and in one case it is under 5%. Again, it would be ideal to remove these outliers and re-run the model to see how the parameters fit, but this step will not be taken in the report.

Interpretation of Features: Looking at the confidence interval for each of the parameters makes it possible to understand the positive or negative effect on the final model. All the parameters have positive albeit extremely small values (they can be found in $A.13$ of the Appendix). Most all the parameters fall close to 0, but none have 0 within the upper or lower bound. Therefore, none of the parameters are ineffective in the model. Cost seems to have the weakest effect, whereas cost grows larger, the mean number of visits increases too. It would make sense that as a subscriber increases the cost of their claims, it follows that they are also increasing the number of visits. Age also has a similar result, as subscriber are older, the number of visits they make increase. This is also logical, since older patients have more health issues, and will require more visits to the doctor. Gender also has a similar strength. It is possible that men are more susceptible to health problems due to the lifestyles that they may lead. The intervention variable likewise has similar strength as the others. It is possible that as subscribers go through more procedures, they need more checkups with the doctor. Also, for the drugs variable, it is possible that when subscribers take more drugs, they must more regularly visit the doctor to update them on how drugs may be affecting their wellbeing. The duration variable makes sense if when subscribers stay at a hospital for long periods, they are possibly going through a procedure which requires checkups to find out the progress of their health.

(iv) Conclusion and Discussion

The first dataset utilized Logistic regression to deal with a response variable that was binary. The initial assumption regarding the balance of the proportion of 1's and 0's for the binary variable was somewhat further than what is desired. I believe that this would be one of the reasons for the ~70% accuracy on the classification test. Looking at the plots of the variables along with their p-values, it was apparent that some of the variables would likely not be used in the final model. During the steps of transformation, there was a goal of finding an efficient transformation. However, even the model with interaction terms was left behind. The variables themselves also became questionable later during the analysis. Variables which would logically seem to increase the likelihood of a lower birthweight had negative effects (e.g. smoking), while something such as higher weight in the mother lead to an increased likelihood of lower birthweight. It is possible that the binary variable for 1's and 0's in the 'birth' variable are mislabeled, causing this confusion. However, it may never be known what the case is.

The second dataset proved to be the more difficult dataset to fit. The initial assumption regarding a similar mean and variance of the term is possibly the reason why the fit was so poor in the final model. It became apparent during the plotting stages that it would be difficult to find a strong set of predictors. There was a lack of strong linear relationship between many variables. Also, the difference in the categorical variable was minor in appearance. The square root transformation also proved to not be useful, where the resultant AIC was ∞. The final chosen model using Stepwise regression also had a rather large number for the AIC, however it was nonetheless smaller than the original model. For this reason, the Stepwise regression model was chosen. The final interpretation of the variables for the second dataset was much more intuitive. The positive and negative directions of the variables seemed to follow what would be expected in a medical situation where biological factors are considered.

Appendix

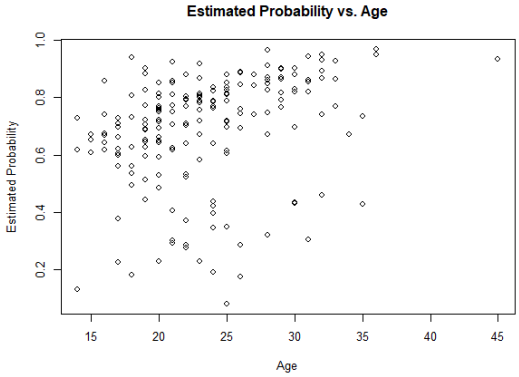
A. 1

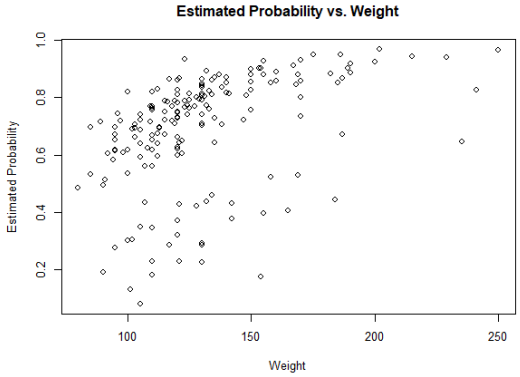
Value	Count	Sample Proportion
1	130	0.6878307
0	59	0.3121693

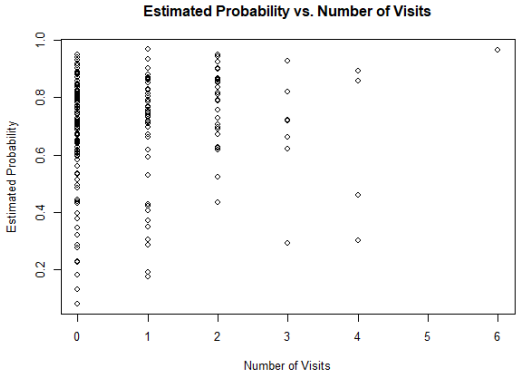
A. 2

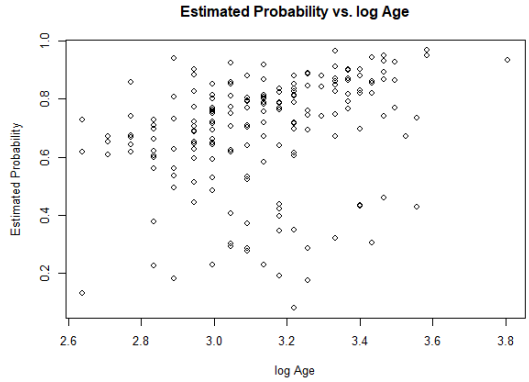
Mean	Variance
3.425127	6.956267

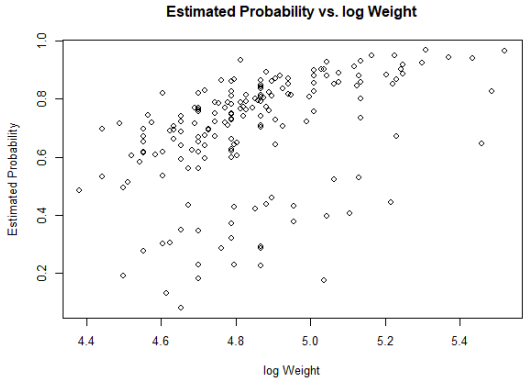
A. 3

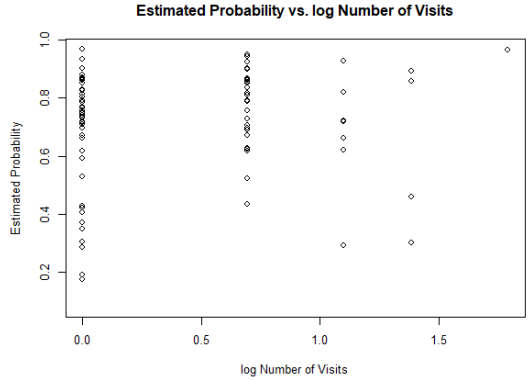


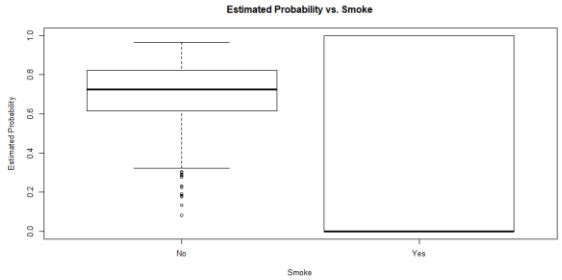


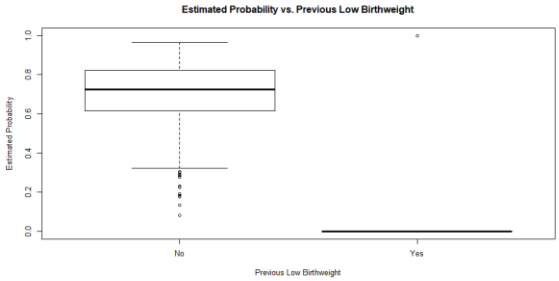


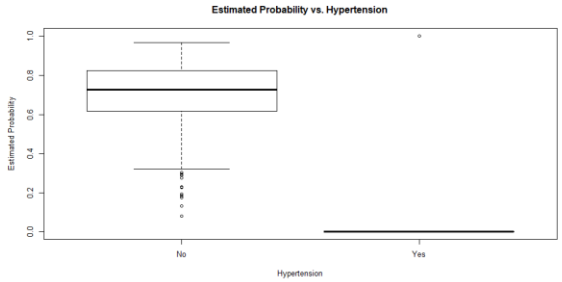








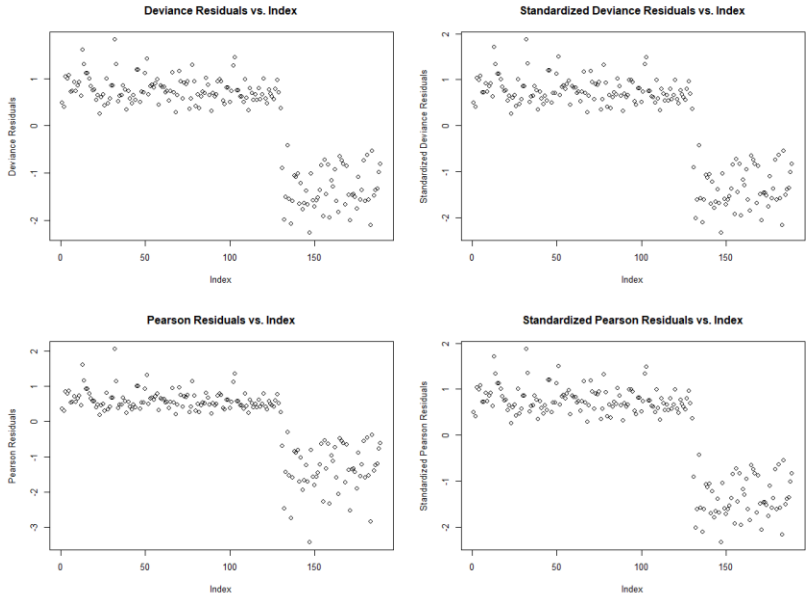




A. 4

Coefficients	Test statistic (G^2)	p-value
$\beta_7, \beta_8, \beta_9$	0.19	0.9791873

A. 5



A. 6

Range	Deviance Residuals	Standardized Deviance Residuals	Pearson Residuals	Standardized Pearson Residuals
Max	1.819297	1.872101	2.057345	1.860716
Min	−2.257052	−2.325042	−3.430822	−2.286726

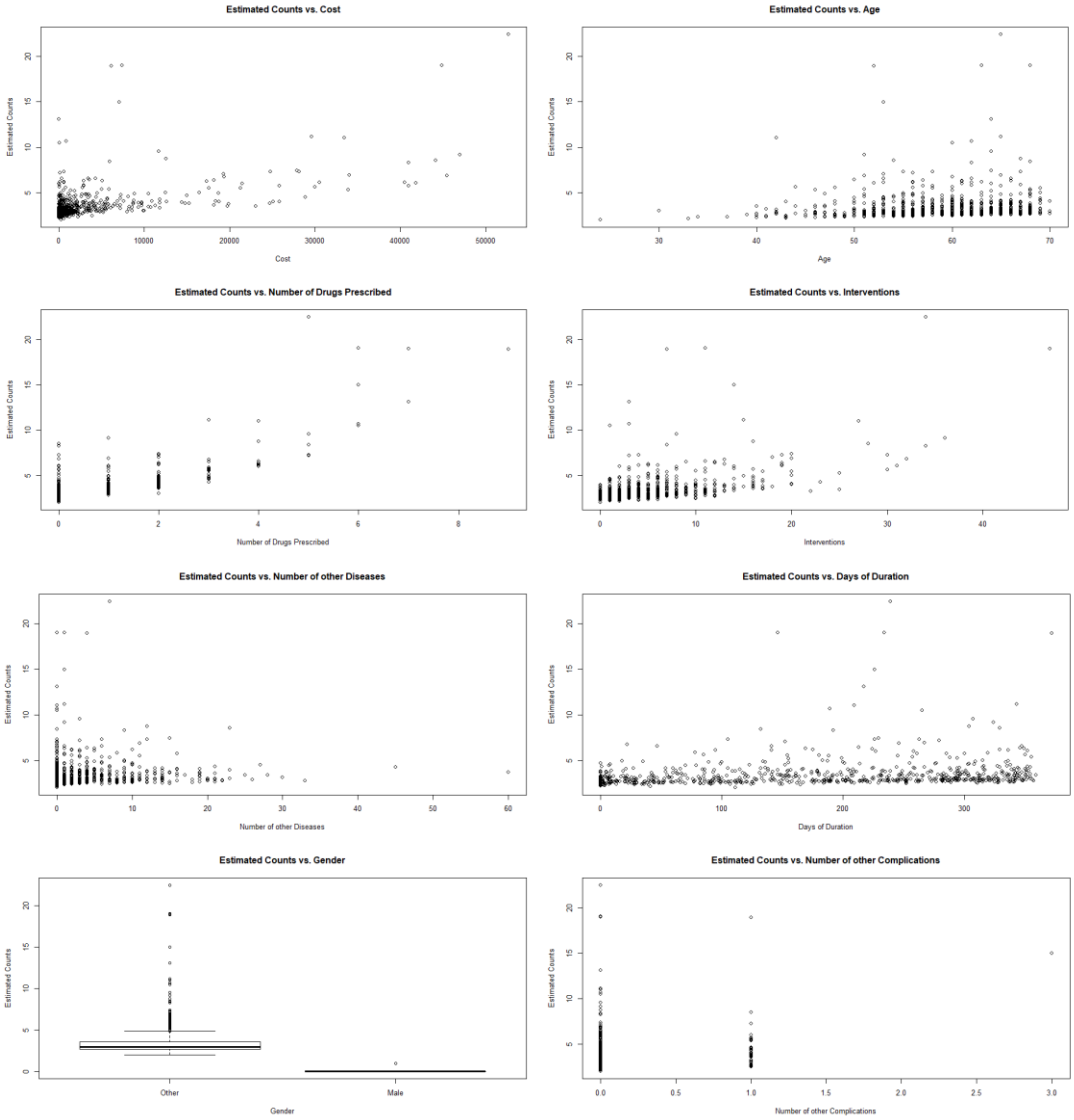
A. 7

$n = 189$	Predicted: NO	Predicted: YES	
Actual: NO	21	38	59
Actual: YES	10	120	130
	31	158	

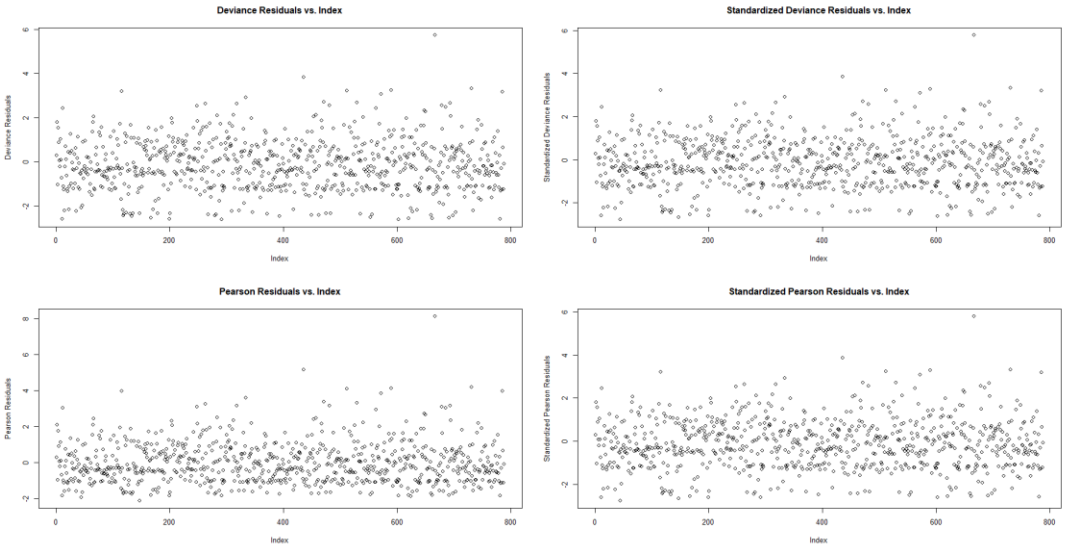
A. 8

Coefficient	Confidence Interval
age $\hat{\beta}_1$	$(-0.01086269, 0.13149973)$
weight $\hat{\beta}_2$	$(0.002585301, 0.029723494)$
smoke $\hat{\beta}_3$	$(-1.2010521, 0.1643193)$
pre $\hat{\beta}_4$	$(-2.791367, -0.796709)$
hyp $\hat{\beta}_5$	$(-3.1874385, -0.3779811)$

A. 9


Estimated Counts vs. Cost
Estimated Counts
Cost
Estimated Counts vs. Age
Estimated Counts
Age
Estimated Counts vs. Number of Drugs Prescribed
Estimated Counts
Number of Drugs Prescribed
Estimated Counts vs. Interventions
Estimated Counts
Interventions
Estimated Counts vs. Number of other Diseases
Estimated Counts
Number of other Diseases
Estimated Counts vs. Days of Duration
Estimated Counts
Days of Duration
Estimated Counts vs. Gender
Estimated Counts
Other
Male
Gender
Estimated Counts vs. Number of other Complications
Estimated Counts
Number of other Complications

A. 10



A. 11

Range	Deviance Residuals	Standardized Deviance Residuals	Pearson Residuals	Standardized Pearson Residuals
Max	5.745722	5.785878	8.119315	5.765866
Min	−2.605684	−2.769384	−2.117950	−2.879518

A. 12

	Deviance Residuals	Standardized Deviance Residuals	Pearson Residuals	Standardized Pearson Residuals
Proportion of Outliers > 3.5	0.005730659	0.005730659	0.02578797	0.005730659

A. 13

Coefficient	Confidence Interval
cost $\hat{\beta}_1$	$(9.357017 \times 10^{-6}, 2.050373 \times 10^{-5})$
age $\hat{\beta}_2$	$(0.0005747559, 0.0120932510)$
gender $\hat{\beta}_3$	$(0.09988022, 0.27154769)$
inter $\hat{\beta}_4$	$(0.002836101, 0.017658537)$
drugs $\hat{\beta}_5$	$(0.1723151, 0.2201964)$
duration $\hat{\beta}_6$	$(1.488865 \times 10^{-5}, 6.756963 \times 10^{-4})$

Resources:

https://stackoverflow.com/questions/40511202/mutating-multiple-columns-in-a-data-frame-using-dplyr

https://stackoverflow.com/questions/17319647/create-sequence-of-numbers-excluding-certain-numbers

https://stackoverflow.com/questions/43986118/replace-values-in-r-yes-to-1-and-no-to-0