Stat 149 Final Project

Kenneth Chen, George Hu, Kay Lu

April 2019

Contents

1 Introduction

1.1 Project Task

Our goal is to model hospital length of stay (LOS) for heart attack patients in New York State based on data collected from $12,844$ diagnosis cases in 1993. The cases correspond to patients whose admitting diagnosis was an Acute Myocardial Infarction (AMI) and who did not have surgery, and we have been provided with information about personal details as well as the results of the stay for each of these cases. Our primary aim is to perform inferential analysis of length of stay to determine recommendations for New York hospital managers and insurance companies, but we will also incorporate predictions at various stages of this analysis to inform our recommendations.

1.2 Real World Implications

The American Heart Association estimates that there are over one million annual incidences of heart attacks in the United States, and that they are also consistently the most expensive condition treated in hospitals. For instance, in 2013, the treatment of heart attacks cost a total of $12.1 billion across the country [1]. While some costs cannot be avoided, a considerable amount is lost due to hospitals failing to properly allocate resources or assign treatments to patients because they lack insight on how long these patients will need to stay. This paper will attempt to address these issues, drawing inferential conclusions about a patient's length of stay to inform the following:

1. **Hospital Resource Allocation:** It is crucial to properly allocate resources to patients in advance, especially in the more densely populated areas of New York. Drawing inferences about `LOS` for newly admitted patients can allow hospitals to make informed decisions on the allocation of doctors, patient beds, medication, etc. Moreover, hospitals can use the inferences to decide how many new patients to admit.
2. **Measure Insurance Risks:** Finally, `LOS` is an important consideration for insurance companies trying calculate appropriate rates to charge their patients. Better understanding of what affects `LOS` will allow these companies to not only better balance their finances but also charge their clients more reasonable rates.

2 Data

2.1 Exploratory Analysis

We begin by visualizing the distributions of the quantitative variables in Figure 3 and note that `CHARGES` and `LOS` are heavily right-skewed. An initial check shows that these variables on a `log` scale still are quite non-normal distributions. As we will mention, we will not use `CHARGES` as a predictor and we will model use `AGE` to create an ordinal predictor, so rather than further examining `LOS` vs predictor relationships with transformations, we table this issue for later; instead of transforming any variables now, we will focus our modeling decisions on accounting for the skew through distribution assumptions. As for the categorical variables, we see that only about 11% of the cases result in a death and notice that `DRG` perfectly encapsulates all information that `DIED` provides, as the "123" group also holds 11% of all cases. Furthermore, our data is composed of 60.6% males and it is most common (40.6% of cases) for the AMI diagnosis to be at an unspecified location (ICD-9 code 410.91). Nothing from this preliminary analysis appears concerning.

We note that there is one observation in our data where `LOS` is 0. This person was still nonetheless present in the hospital since he or she was recorded in study, so we will treat this case as having an `LOS` value of 1.

2.2 Predictor Choices

First, we choose to ignore DIED, DRG, and CHARGES because it does not make sense to use post hoc predictors to model LOS. All three variables include information that is only available for certain after a patient leaves the hospital (whether by death or discharge); in other words, when we know LOS for certain as well. Since our goal is to make inferences and allocate resources when patients are first admitted, it doesn't make sense to use information that doesn't exist at that time. Please note that we will revisit each of these three variables at a later stage of analysis in this paper. Patient, which refers to patient number, is an interesting variable that warrants discussion. There may be some kind of serial time effect related to patient numbers, but a linear model fit on Patient, AGE, SEX, and DIAGNOSIS returns Patient as a insignificant predictor. VIF analysis does not flag Patient as collinear with other predictors and there is no convincing logic-backed reason to include Patient; thus, we do not use Patient as a predictor.

Finally, we do choose to use AGE, SEX, and DIAGNOSIS as predictors. We treat SEX and DIAGNOSIS as categorical predictors (SEX has 2 levels, DIAGNOSIS has 9) and AGE as a numerical predictor. However, we don't believe AGE should be purely numeric when modeling LOS because health is not necessarily something that deteriorates at clear-cut one year increments; it also declines with major life events such as diseases or injuries that are more characteristic of particular age groups than particular ages. Thus we introduce a new ordinal predictor, AGEGROUP, with 7 levels: "0-34","35-44","45-54","55-64","65-74","75-84","85+". This division is based on standard age groupings by the American Heart Association [2], which further backs our decision to create this variable. Note that we will wait until the model tuning step of our analysis to incorporate AGEGROUP, in order to avoid collinearity with AGE while we make modeling decisions.

3 Approach to Constructing the Models

3.1 Considering the Distribution of the Response

Our first key modeling decision is determining which distribution to assume for LOS. Since it is a discrete count, we begin by considering the Poisson and Negative Binomial (henceforth "NB") distributions. NB makes more sense based on first principles because it involves a sequence of trials, which are analogous to the days during a hospital stay (Poisson does not account for days happening in sequence). Based on the histogram of LOS in Figure 2, NB also seems more appropriate because the variance (26.2) is much greater than the mean (7.6), violating a key assumption of the Poisson.

However, when we fit our baseline models on a NB response, we obtain a problematic fitted vs deviance residuals plot with a large amount of residuals greater than 2 in Figure 8. See Table 2 for a full summary. These diagnostics, along with the original histogram, hint at positive skewness, so we fit a GLM assuming a Tweedie distribution for LOS. This model gives index parameter $p = 1.839$, so we know the mean-variance relationship corresponds most closely to a Gamma distributed response. This **Gamma GLM** fixes the issue of inflated deviance residuals and also produces a satisfactory binned plot and Cook's distances plot, seen in Figures 10, 11, and 12, so we choose to use it as our new baseline model for inferring on LOS.

3.2 Handling 1-Inflatedness of Response

Even with the steps above, there is still some overdispersion in the response we have not accounted for because of the disproportionate number of 1's in LOS (which hurts our Gamma assumption). We hypothesize the large number of 1-day stays, relative to 2-5 day stays, has two reasons. First, there are false alarms; people with heart attacks must necessarily stay overnight at the hospital for treatment [3]. Thus, people who are discharged after 1 day must likely have had an incorrect admitting diagnosis. Second, there are immediate deaths, or heart attacks that are so serious that the individual dies quickly after. Thus, there are three outcomes after day one: 1) false alarm, 2) death, or 3) stay at hospital. We create a factor variable `DAYONE` to keep track of these outcomes. We must account for these cases like we accounted for structural 0's in the hurdle models from lecture; the differences are that the response is one-inflated rather than zero-inflated, and predicting the structural 1's involves a multinomial (3 outcomes) rather than binomial distribution.

We attempt to predict the day one outcome with 1) multinomial logistic regression and 2) classification trees, but run into problems because the frequencies of outcomes are imbalanced. Specifically, 4.5% of the observations are day one deaths, 4.8% are day one discharges, and the rest stay after day one. Though this collective 9.3% of the observations is important in the grand scope of results a hospital must deal with, the two events are too rare to be predicted for. We see that the fitted regression predicts all individuals in the dataset will stay after day one, as does the fitted tree. To overcome this problem of imbalance, we apply the Synthetic Minority Over-sampling Technique (SMOTE) to increase the number of observations of minority outcomes (death and discharge) in the dataset to balance the two classes to roughly 50%/50% for the model fitting step. Even with this technique, however, overall accuracy ($\sim 81\%$) remains below the accuracy we would have obtained had we simply classified everyone as staying after day one ($\sim 91\%$). Moreover, only $\sim 24\%$ of the time that the new model predicts the person will die on day one does he or she actually die, a dangerous error because the hospital will under-allocate beds. The poor accuracy of the model and the high downside of type I error lead us to make a big decision: **we will wait until the end of the first day to model `LOS`**.

We choose this strategy because 1) on any given day, making resource allocation decisions altogether at the end of the day is more efficient than doing it every time a new patient enters, and 2) we can improve our ultimate model by waiting for extreme cases (immediate deaths and false alarms) to play out. After all, these cases are essentially noise obstructing our goal of drawing inferences on `LOS` for patients who actually *need* and would benefit from care. Note that to begin our model from day 2, we subtract one from all `LOS` values.

3.3 Deciding to Incorporate Post Hoc Variables

Thus far, we have modeled on `AGE`, `SEX`, and `DIAGNOSIS`. However, we should not ignore `CHARGES`, `DRG`, and `DIED` simply because they constitute information obtained after discharge. Instead of throwing them away, we can conduct likelihood ratio tests to determine which, if any, of the three provide significant additional information for inferring on `LOS`. We can do this by comparing, for example, Gamma GLM fits on `LOS`∼`AGE`+`SEX`+`DIAGNOSIS`+`DIED` vs `LOS`∼`AGE`+`SEX`+`DIAGNOSIS` with the `anova` function (using F-tests). We find that all three tests return significant p-values of $<$ 2.2e-16, meaning there is significant association between the response and `CHARGES`, `DRG`, and `DIED`. To capture this information in our model, we will attempt to predict these three variables from `AGE`, `SEX`, and `DIAGNOSIS` and use our predictions in turn as predictors in the GLM for `LOS`. We will construct these predictions using both linear and nonlinear methods to avoid introducing collinearity into the predictor set.

3.4 Incorporating CHARGES via Model Selection

As mentioned above, we begin our predictions from day 2 after subtracting 1 from all LOS values. To predict CHARGES, we first use the na-convert function from class to fill in missing values for CHARGES and introduce a CHARGES.na predictor with our other predictors, AGE, SEX, and DIAGNOSIS. We split our data into train and test sets, using 20% of the data as our test set. We predict using a Gamma GLM with a log link, a Gamma GAM with a log link, a pruned Regression Tree with optimal cp=.0011 from the 1-SE rule, and a Feedforward Neural Network (FFNN) with architecture defined in 10.1. For the FFNN, all variables were independently scaled between -1 and 1 to standardize for use with tanh layers. The dataset was then split for into train and test sets.

From Table 3, the GLM produced the lowest test RMSE and was henceforth used to produce CHARGES_PRED. The GLM predictions residuals on the entire data are centered somewhat normally around 0 in Figure 14 with the majority of differences residuals within 10,000 of the actual CHARGES, so we feel comfortable using them as predictors, considering that the range of CHARGES is 47,901. Comparison of models via histograms of test set residuals is in Figure 13.

3.5 Incorporating DIED & DRG via Model Selection

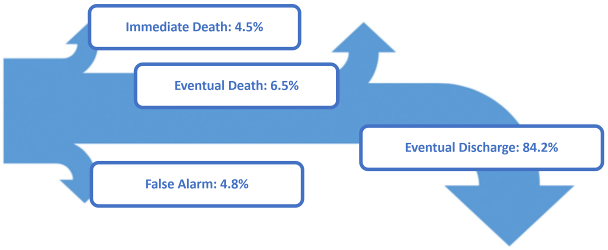

Figure 1: Distribution of DIED via Subsetting

After predicting CHARGES_PRED, we perform an analogous process for DIED using the same four models, with the difference being that the GLM and GAM are of the binomial family and that the pruned tree has optimized cp=0.00021 from the 1-SE rule. The models are run under two setups: once with the data as is, and once with SMOTE data to account for imbalance in the distribution of DIED (see 3.2). Without SMOTE data, all models predicted either only 0's or nearly only 0's on the test set. With SMOTE data, however, the pruned tree model and FFNN, with the architecture in 10.2, performed best in Table 4. Results on the original data set in Table 5 show that the models do not differ too much. Even though FFNN true positive rate shows better balance with true negative rate compared to the pruned tree model, we decided to use the DIED predictions from the tree model for two reasons. First, higher overall accuracy and higher F1 score to account for class imbalance is more preferable. Second, a lower true negative rate is not preferred because it means hospitals will be more likely to under-allocate resources as a result of estimating fewer deaths, when the number of actual survivors is higher than predicted. With a comparatively higher true negative rate, the pruned tree model is preferred over the FFNN. Lower true positive rate in the pruned tree model is less of a concern, as hospitals will over-allocate resources, which would not lead to deaths due to lack of resources.

Additionally, we can now utilize the DRG variable under the assumption that it should be clear that patients who are admitted and stay over a day should have been examined for clear signs of complications or no complications. We use only the DRG codes 121 and 122 for patients who do not die, since using the 123 code provides no additional information since maps only to those who died. In the case DIED_PRED is a false negative, code 123 is now a third level representing NA in DRG. Thus, for survivors, we use a DRG.na predictor that maps to 1 if DRG is 123 and to 0 otherwise.

4 Testing & Tuning the Final Models

After choosing a distribution assumption for the response, handling its one-inflatedness, and reincorporating the post-hoc variables with information we have on day one, we are ready to perform model selection on the final predictor set with the following three-steps:

4.1 Predictor selection with Gamma model

Based on our findings from 3.3, in which we found that DIED is a highly significant predictor of LOS, we believe that we can gain more granular insight from fitting separate models for patients who are predicted to survive and patients who are predicted to die. Thus, we split our data based on DIED_PRED. Starting from a null model for each of the two categories, we perform forward selection on main effects of predictors and the pair predictor interactions via likelihood ratio tests with $\alpha = .05$ to reach our final Gamma models. An abbreviated summary of model selection is as follows (see Tables 6 and 7 for full model selection summaries). Note that when performing forward selection on pair interaction terms, the full summary lists only the best models, for conciseness against the high number of predictor pairs. Below in the abbreviated summary, the best model under each setup is boxed.

Table 1: Abbrev. Model Selection for Survivals (Top) and Deaths (Bottom)

Model	Deviance	Coefficients	Residual df
1	3530.4	1	11227
1 + AGE	3354.7	2	11226
1 + AGE + DRG	3194.8	4	11224
1 + AGE + DRG + SEX	3182.2	5	11223
1 + AGE + DRG + SEX + CHARGES_PRED	3178.8	6	11222
1 + AGE + DRG + SEX + CHARGES_PRED + DIAGNOSIS	3172.9	14	11214
1 + AGE + DRG + SEX + CHARGES_PRED + DIAGNOSIS + AGEGROUP	3168.5	20	11208
1 + DRG * SEX + AGE + DIAGNOSIS + CHARGES_PRED + AGEGROUP	3162.1	22	11206
$\boxed{\text{1 + DRG ∗ SEX + AGE ∗ DIAGNOSIS + CHARGES\_PRED + AGEGROUP}}$	3157.1	30	11198
1 + DRG * SEX + AGE * DIAGNOSIS + CHARGES_PRED * AGEGROUP	3156.3	36	11192
1	161.55	1	420
$\boxed{\text{1 + CHARGES\_PRED}}$	158.38	2	419
1 + AGEGROUP	159.51	4	417
1 + CHARGES_PRED + AGE	157.73	3	418
1 + CHARGES_PRED + SEX	157.70	3	418
1 + CHARGES_PRED + DIAGNOSIS	155.76	10	411
1 + CHARGES_PRED + AGEGROUP	156.61	5	416

4.2 Comparison to Using Alternative Response Distributions

Summaries for the best models are found in Section 9. Though we determined earlier that using a Gamma GLM is most appropriate for performing inference on LOS, we now double check against Poisson and NB models since our models have substantially changed.

After fitting both alternative models for predicted survivors and predicted deaths, we examine the binned residual plots and Cook's distance plots across the three models. In all cases as seen in Section 8.4, the binned plots do not show concerning patterns and also have few extreme values, passing visually. As for the Cook's distance plots, none of them show any individual points which skew the results. The deviance residual plots, however, really set the three models apart. For predicted survivors, the magnitude of the residuals is upwards of 8 for the Poisson plot. The NB and Gamma models behave better; most of the residuals have magnitude less than 3. This is consistent with the fact that NB models address over-dispersion and Gamma models address positive skew. For predicted deaths, Poisson models perform poorly in the same manner. The NB plot shows some improvement, but the Gamma plot is clearly the best, with most residuals having magnitude less than 1.5.

4.3 Comparison to Using Generalized Additive Models

Lastly, we compare our two Gamma GLMs to the analogous GAMs with the same set of predictors. We smooth the numeric `CHARGES_PRED` predictor, with no cap for the degrees of freedom since there are many values of `CHARGES_PRED` in the data. We leave all other predictors alone because they are categorical/involved in interaction terms. We expect to see no significant difference between the GLMs and GAMs because of the categorical nature of the predictor set and the fact that a roughly linear relationship between the smoothed `CHARGES_PRED` and `LOS` makes sense. Our expectation is confirmed when we run likelihood ratio tests between the comparable GLM and GAM for predicted survivors, as the test returns an insignificant p-value, but not for predicted deaths. However, as `CHARGES_PRED` is the only predictor in this model, it makes sense that a smooth on the predictor would yield better performance.

Final Model: We thus arrive at our final three-part model. First, predict `CHARGES` using a Gamma GLM on `AGE`, `SEX`, `DIAGNOSIS`, and `CHARGES.na`. Second, predict whether a patient survives or dies using a classification tree on `AGE`, `SEX`, `DIAGNOSIS`, `CHARGES.na`, and `CHARGES_PRED`. Third, for those predicted to survive, fit a Gamma GLM fitted as in section 4.1; for those predicted to die, fit a Gamma GAM fitted as in section 4.1 with smoothing on `CHARGES_PRED`.

5 Conclusions

Our first major finding is that modeling on `LOS` after the end of the first day produces markedly better-fit models than if we did not implement this hurdle structure. We therefore recommend hospital managers to wait until the end of the first day after a person is admitted into the hospital to decide how many resources to allocate to this new patient (expected days of bed space needed, etc.). This modification makes sense because it removes the possibility that false alarms and immediate deaths will throw off hospital management; it also makes the hospital more efficient because management can make allocation decisions all at once at the end of each day. A second finding is that it is important to predict whether or not a patient dies based on what we know about him or her at the end of the first day. The importance is evident in how the chosen models for the predicted survivors and predicted deaths subsets of the data were different. Though the prediction is understandably not perfect, our explanation to hospital managers is that death is too crucial a piece of information to not at least attempt to gain insight about. The possibility of death not only affects length of stay, but also the amount of resources like medication and nurse attention that must be allocated to a patient.

As for particularly relevant predictors, for survivors, we see that the main effects of DRG are very significant, as well as the interaction of DRG123:SEXM. It is interesting to see that the only other significant predictors are interaction terms: AGE:DIAGNOSIS41021 and AGE:DIAGNOSIS41071. To interpret the significant interaction terms, as an example, AGE:DIAGNOSIS41071 suggests that the difference in effect on LOS between a subendocardial infarction and a infarction of the anterolateral wall is more pronounced in someone who is older. These interaction terms suggest that, while the diagnosis of a heart condition is important, the interaction with the patient's age is equally important, likely as the result of having other organs failing to work properly together at an older age. Seeing AGE but not AGEGROUP in the final model suggests that there is valuable information obtained from the raw age that would otherwise be lost in grouping ages. For those predicted to die, it is fascinating that AGE is not a significant predictor. This result could be because the predictors, including AGE, used to create CHARGES_PRED transfer significant information to CHARGES_PRED itself, suggesting that the prediction of charges holds a great deal of inferential importance. These results are all helpful to hospital managers because they indicate that, depending on whether a person is likely to die, old age, different diagnoses, and predicted charges can signal more serious cases.

6 Future Work

One of the major tasks we attempted was predicting CHARGES and subsequently DIED to use as a predictor variable and as a splitting criteria, respectively. Having strong prediction RMSE and accuracy/F1 score are crucial for our subsequent inferences, and we feel that this approach in subsetting individuals into survival and death before day 2 and on/after day 2 would help hospitals avoid drawing conclusions on aggregate data that may be misleading. While CHARGES_PRED is fairly normal looking, there is certainly room to improve, as some predictions are off in the order of 10K. Improvements to the prediction of DIED should be a major focus, as the true positive rate was low across all models. Further resampling approaches (bagging) and weighted approaches (weighted regressions, class weights for neural networks) should be explored. Additional models, such as SVMs and ensembles of trees (Random Forest) would be good ideas to explore as well, the latter in particular since the pruned tree model ended up being our best prediction model for DIED. With better predictions, different methods of selection (forward, backward, stepwise) can be explored in building the final inferential models.

7 Appendix for Tables

Table 2: Comparison of Models for 2.1

Model	Resid Dev/DF	Min Dev Resid	Max Dev Resid
Poisson	3.069	-3.697	8.468
NegBin	1.046	-2.331	4.287
Gamma	($\hat{\phi} = 0.421$)	-1.683	2.684

Table 3: Comparison of Models for Predicting `CHARGES`

Model	Test RMSE
Gamma GLM	6142.230
Gamma GAM	6142.390
Pruned Tree	6158.042
FFNN	15675.489

Table 4: Comparison of Models for Predicting `DIED` with SMOTE Data on Test Data

Model	Test True Positive Rate	Test True Negative Rate	Test Accuracy
Binomial GLM	.770	.702	.737
Binomial GAM	.789	.695	.744
Pruned Tree	.883	.963	.922
FFNN	.834	.955	.895

Table 5: Comparison of Models for Predicting `DIED` on Original Data

Model	True Positive Rate	True Negative Rate	Accuracy	F1 Score
Pruned Tree	.119	.970	.909	.952
FFNN	.167	.953	.897	.945

Table 6: Analysis of Deviance for Predicted Survivors

Model	Deviance	Coefficients	Residual df
1	3530.4	1	11227
1 + AGE	3354.7	2	11226
1 + SEX	3471.8	2	11226
1 + DIAGNOSIS	3518.7	9	11219
1 + DRG	3315.6	3	11225
1 + DRG.na	3516.9	2	11226
1 + CHARGES_PRED	3422.0	2	11226
1 + AGEGROUP	3357.9	7	11221
1 + AGE + SEX	3339.2	3	11225
1 + AGE + DIAGNOSIS	3344.2	10	11218
1 + AGE + DRG	3194.8	4	11224
1 + AGE + DRG.na	3322.0	3	11225
1 + AGE + CHARGES_PRED	3342.9	3	11225
1 + AGE + AGEGROUP	3348.9	8	11220
1 + AGE + DRG + SEX	3182.2	5	11223
1 + AGE + DRG + DIAGNOSIS	3185.6	12	11216
1 + AGE + DRG + CHARGES_PRED	3184.5	5	11223
1 + AGE + DRG + AGEGROUP	3190.0	10	11218
1 + AGE + DRG + SEX + DIAGNOSIS	3172.9	13	11215
1 + AGE + DRG + SEX + CHARGES_PRED	3178.8	6	11222
1 + AGE + DRG + SEX + AGEGROUP	3178.2	11	11217
1 + AGE + DRG + SEX + CHARGES_PRED + DIAGNOSIS	3172.9	14	11214
1 + AGE + DRG + SEX + CHARGES_PRED + AGEGROUP	3174.5	12	11216
1 + AGE + DRG + SEX + CHARGES_PRED + DIAGNOSIS + AGEGROUP	3168.5	20	11208
⋮	⋮	⋮	⋮
1 + DRG * SEX + AGE + DIAGNOSIS + CHARGES_PRED + AGEGROUP	3162.1	22	11206
$\boxed{1 + \texttt{DRG} * \texttt{SEX} + \texttt{AGE} * \texttt{DIAGNOSIS} + \texttt{CHARGES\_PRED} + \texttt{AGEGROUP}}$	3157.1	30	11198
1 + DRG * SEX + AGE * DIAGNOSIS + CHARGES_PRED * AGEGROUP	3156.3	36	11192

Table 7: Analysis of Deviance for Predicted Deaths

Model	Deviance	Coefficients	Residual df
1	161.55	1	420
1 + AGE	159.56	2	419
1 + SEX	159.63	2	419
1 + DIAGNOSIS	159.52	9	412
$\boxed{1 + \texttt{CHARGES\_PRED}}$	158.38	2	419
1 + AGEGROUP	159.51	4	417
1 + CHARGES_PRED + AGE	157.73	3	418
1 + CHARGES_PRED + SEX	157.70	3	418
1 + CHARGES_PRED + DIAGNOSIS	155.76	10	411
1 + CHARGES_PRED + AGEGROUP	156.61	5	416

8 Appendix for Figures

8.1 Distributions

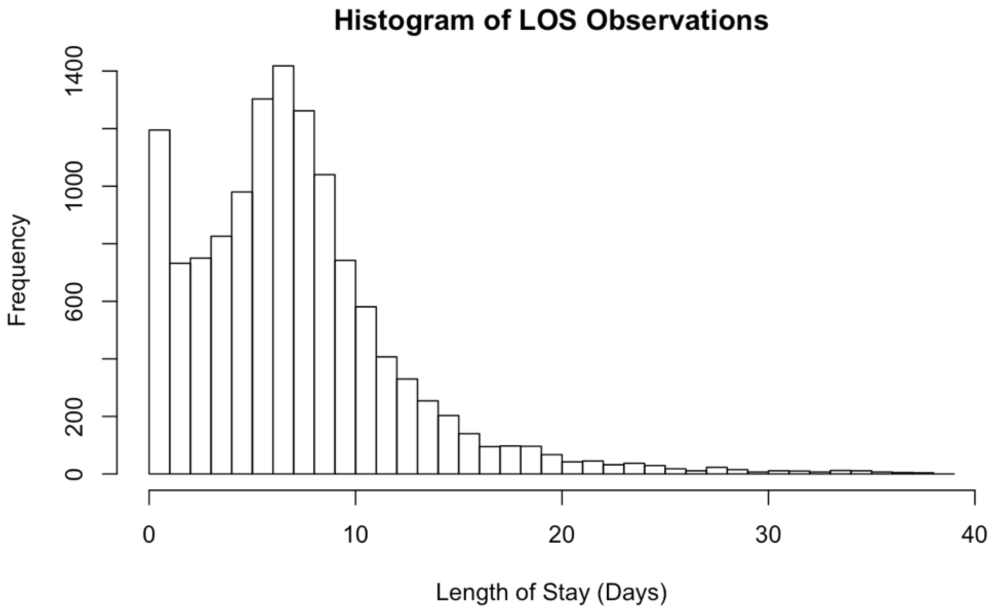

Figure 2: Histogram of Response Variable

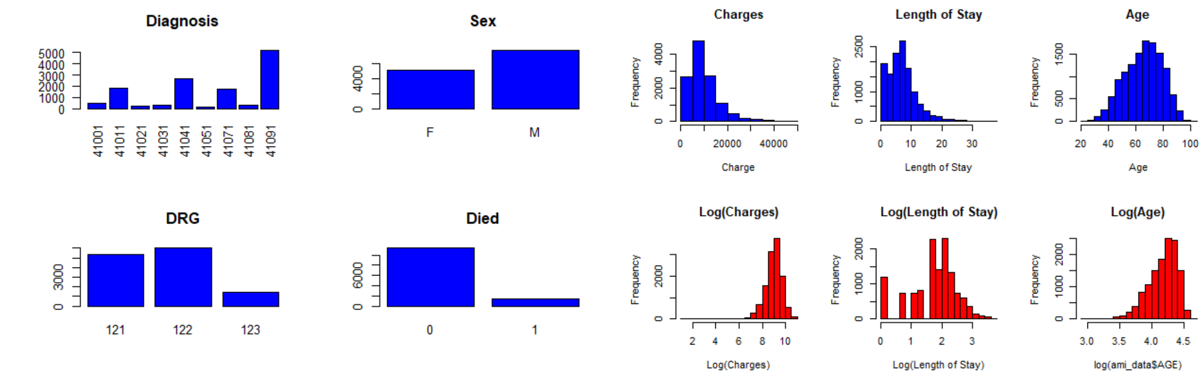

Figure 3: Distributions of Variables

8.2 Initial Models

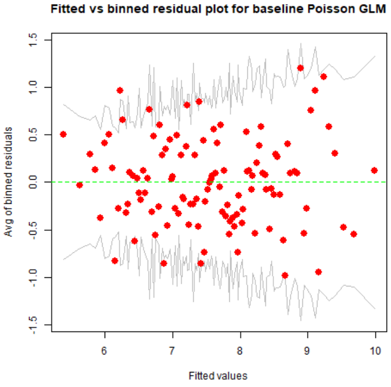

Figure 4: Fitted vs Binned Residual Plot for Poisson GLM

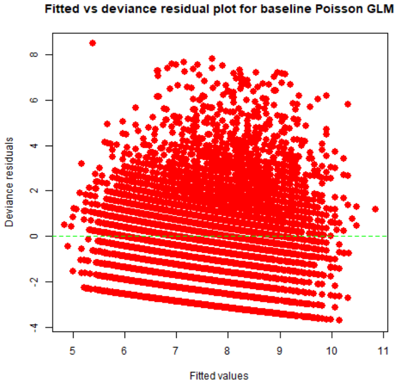

Figure 5: Fitted vs Deviance Residual Plot for Poisson GLM

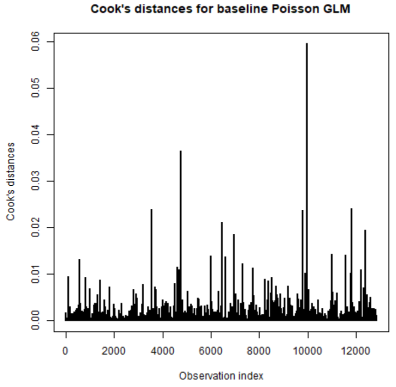

Figure 6: Cook's Distances Plot for Poisson GLM

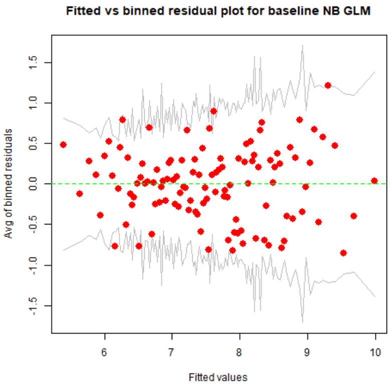

Figure 7: Fitted vs Binned Residual Plot for NegBin Model

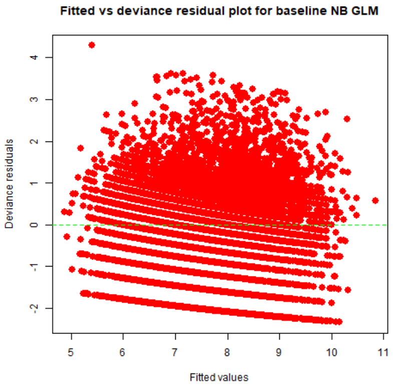

Figure 8: Fitted vs Deviance Residual Plot for NegBin Model

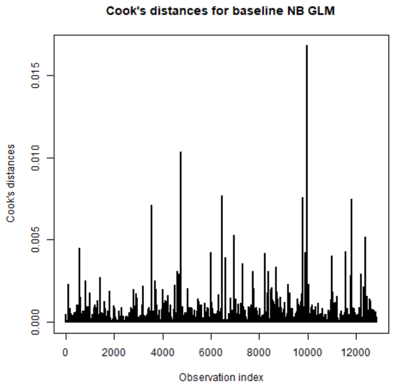

Figure 9: Cook's Distances Plot for NegBin Model

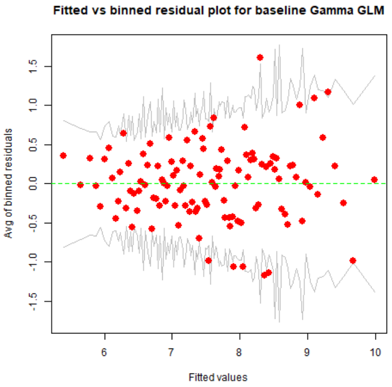

Figure 10: Fitted vs Binned Residual Plot for Gamma GLM

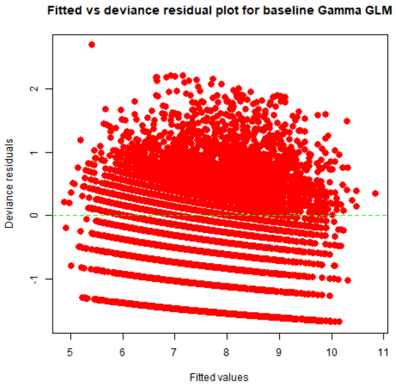

Figure 11: Fitted vs Deviance Residual Plot for Gamma GLM

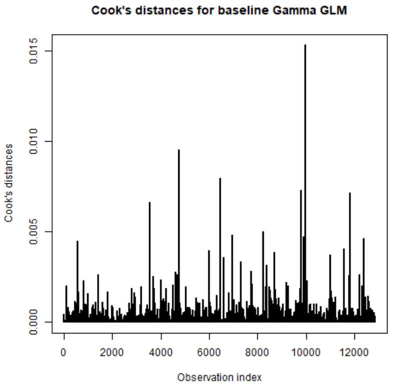

Figure 12: Cook's Distances Plot for Gamma GLM

8.3 CHARGES Predictions

8.3.1 Test Set Residuals

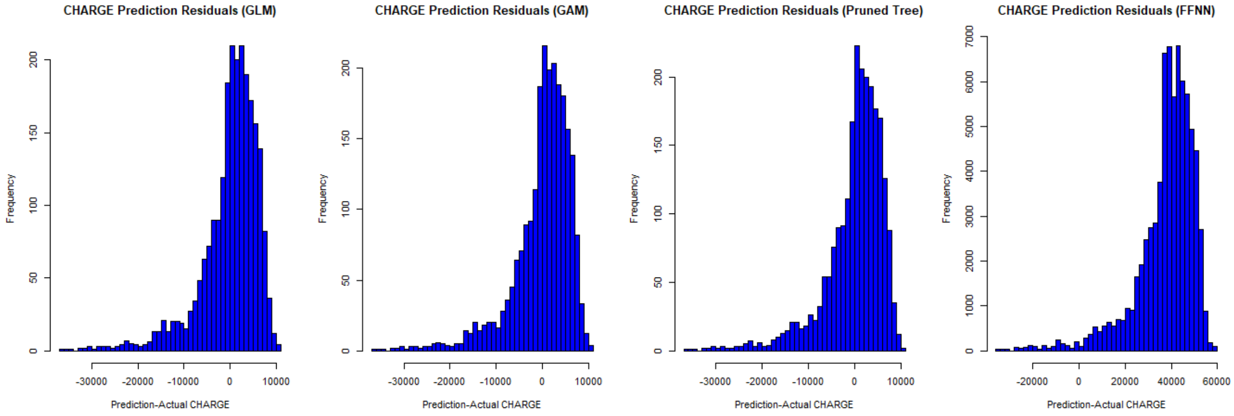

Figure 13: CHARGES Test Set Residuals

8.3.2 Final Model Full Set Residuals

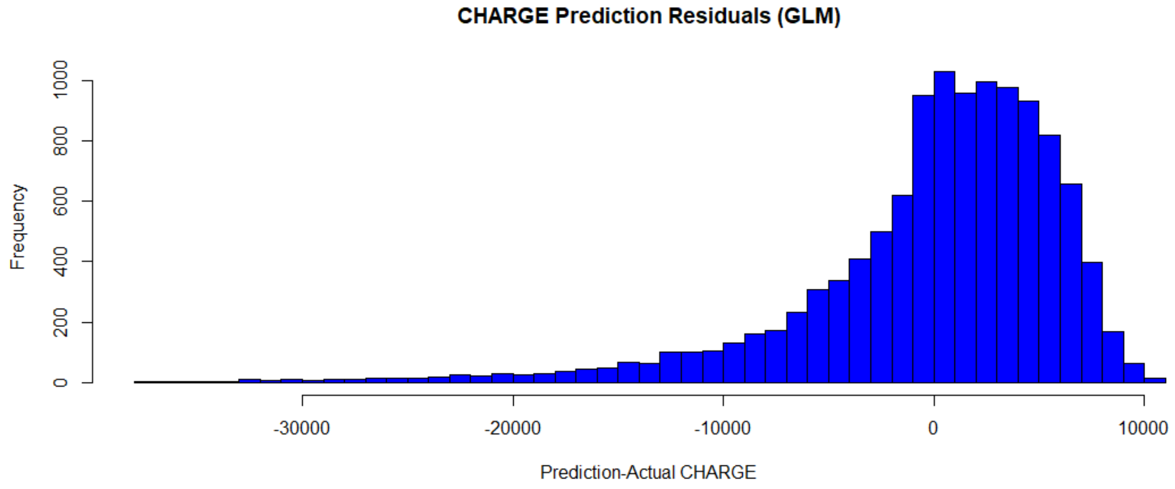

Figure 14: `CHARGES` Full Data Residuals for GLM

8.4 Testing and Tuning Models

8.4.1 Survivors

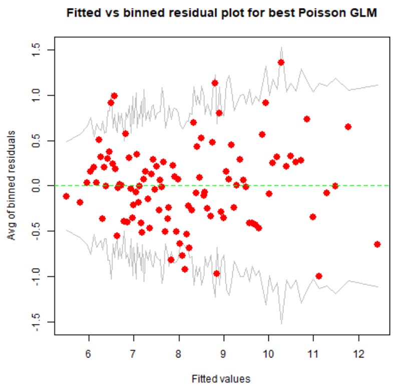

Figure 15: Fitted vs Binned Residual Plot for best Poisson GLM (predicted survivors)

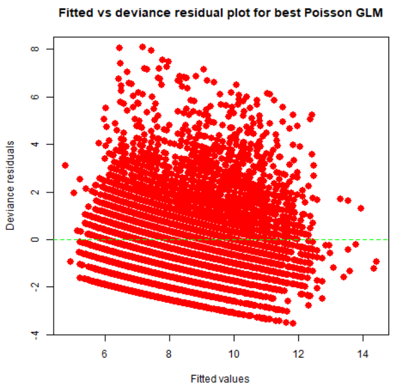

Figure 16: Fitted vs Deviance Residual Plot for best Poisson GLM (predicted survivors)

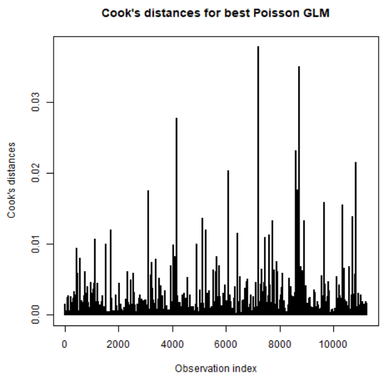

Figure 17: Cook's Distances Plot for best Poisson GLM (predicted survivors)

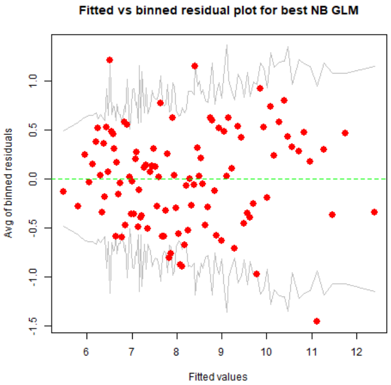

Figure 18: Fitted vs Binned Residual Plot for best NB GLM (predicted survivors)

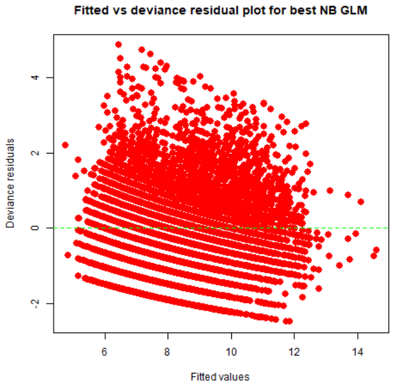

Figure 19: Fitted vs Deviance Residual Plot for best NB GLM (predicted survivors)

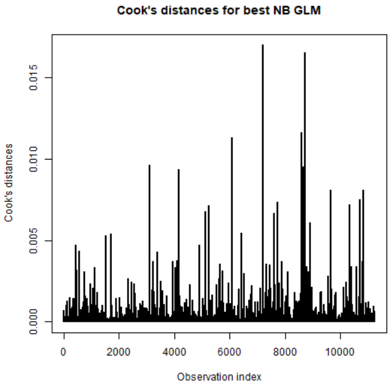

Figure 20: Cook's Distances Plot for best NB GLM (predicted survivors)

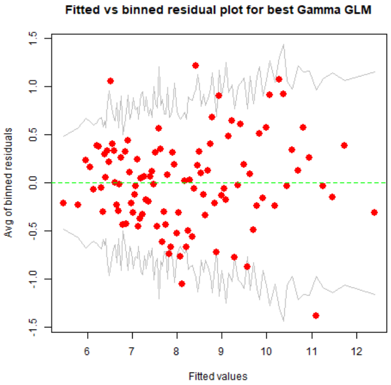

Figure 21: Fitted vs Binned Residual Plot for best Gamma GLM (predicted survivors)

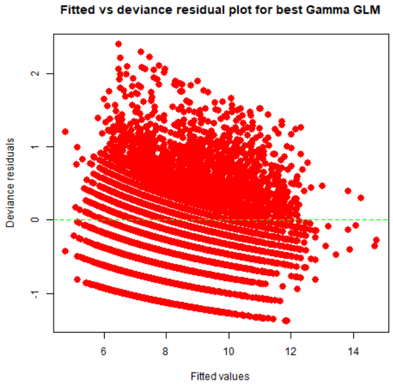

Figure 22: Fitted vs Deviance Residual Plot for best Gamma GLM (predicted survivors)

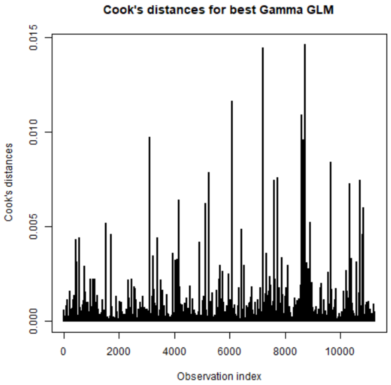

Figure 23: Cook's Distances Plot for best Gamma GLM (predicted survivors)

8.4.2 Predicted deaths

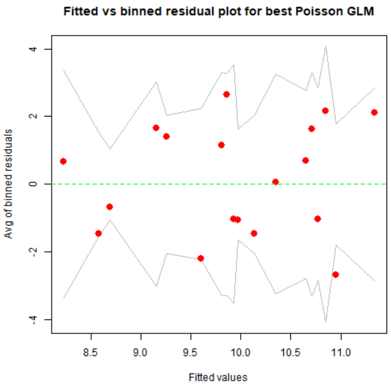

Figure 24: Fitted vs Binned Residual Plot for best Poisson GLM (predicted deaths)

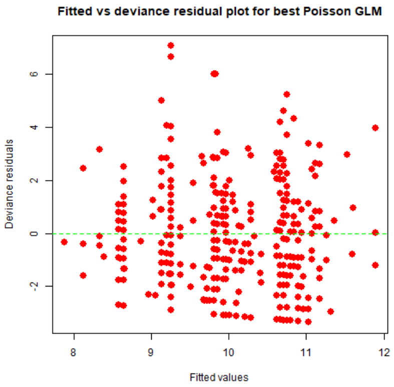

Figure 25: Fitted vs Deviance Residual Plot for best Poisson GLM (predicted deaths)

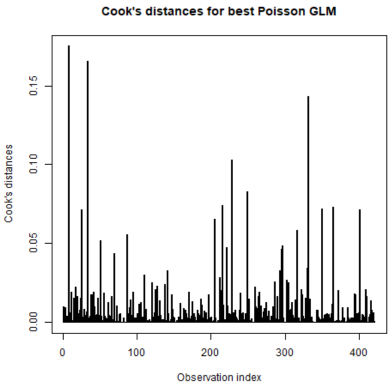

Figure 26: Cook's Distances Plot for best Poisson GLM (predicted deaths)

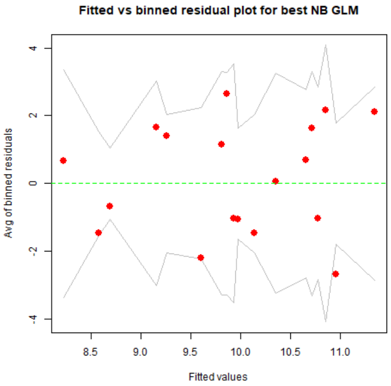

Figure 27: Fitted vs Binned Residual Plot for best NB GLM (predicted deaths)

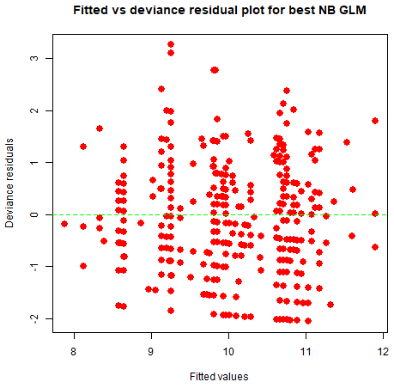

Figure 28: Fitted vs Deviance Residual Plot for best NB GLM (predicted deaths)

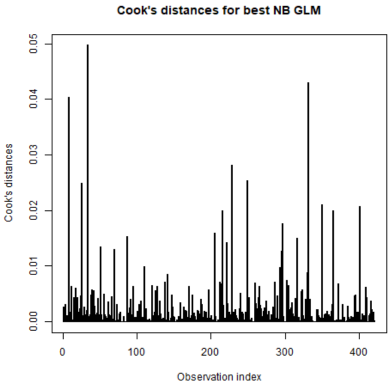

Figure 29: Cook's Distances Plot for best NB GLM (predicted deaths)

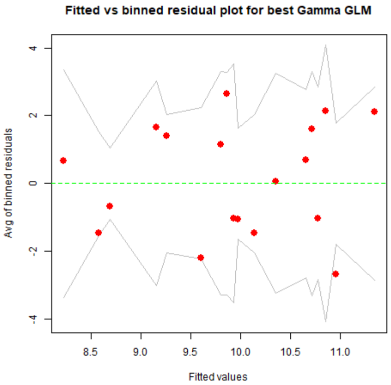

Figure 30: Fitted vs Binned Residual Plot for best Gamma GLM (predicted deaths)

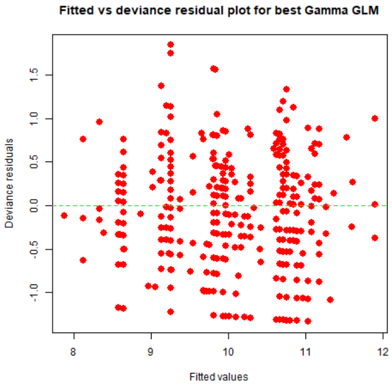

Figure 31: Fitted vs Deviance Residual Plot for best Gamma GLM (predicted deaths)

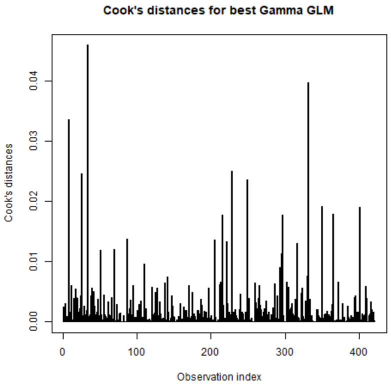

Figure 32: Cook's Distances Plot for best Gamma GLM (predicted deaths)

9 Appendix for Final Model Summaries

9.1 Predicted Survivors

```
Call:
glm(formula = LOS ~ 1 + DRG * SEX + AGE * DIAGNOSIS + CHARGES_PRED +
    AGEGROUP, family = Gamma(log), data = survivals)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-1.37851   -0.41143   -0.08689    0.21000    2.38961

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.3396574  3.1824937  -0.107   0.9150
DRG122             -0.2243883  0.0181280 -12.378  < 2e-16 ***
DRG123             -0.4318452  0.0304384 -14.188  < 2e-16 ***
SEXM                0.0425855  0.1825048   0.233   0.8155
AGE                -0.0026377  0.0115439  -0.228   0.8193
DIAGNOSIS41011     -0.0683466  0.2510721  -0.272   0.7855
DIAGNOSIS41021     -0.4329220  0.2880949  -1.503   0.1329
DIAGNOSIS41031     -0.1375705  0.3961162  -0.347   0.7284
DIAGNOSIS41041     -0.0058738  0.3988188  -0.015   0.9882
DIAGNOSIS41051     -0.0336972  0.7269001  -0.046   0.9630
DIAGNOSIS41071     -0.1490644  0.5288705  -0.282   0.7781
DIAGNOSIS41081     -0.0126084  0.5337232  -0.024   0.9812
DIAGNOSIS41091      0.1429056  0.5028688   0.284   0.7763
CHARGES_PRED        0.0002288  0.0003114   0.735   0.4625
AGEGROUP.L         -0.0999294  0.1004145  -0.995   0.3197
AGEGROUP.Q          0.0561958  0.0343046   1.638   0.1014
AGEGROUP.C         -0.0012939  0.0254812  -0.051   0.9595
AGEGROUP^4         -0.0125488  0.0202534  -0.620   0.5355
```

```
AGEGROUP^5          0.0072636  0.0154882   0.469   0.6391
AGEGROUP^6         -0.0060846  0.0122027  -0.499   0.6180
DRG122:SEXM         0.0114321  0.0227467   0.503   0.6153
DRG123:SEXM         0.2025672  0.0443507   4.567 4.99e-06 ***
AGE:DIAGNOSIS41011  0.0033215  0.0024454   1.358   0.1744
AGE:DIAGNOSIS41021  0.0091360  0.0036070   2.533   0.0113 *
AGE:DIAGNOSIS41031  0.0049336  0.0038784   1.272   0.2034
AGE:DIAGNOSIS41041  0.0042095  0.0026602   1.582   0.1136
AGE:DIAGNOSIS41051  0.0083788  0.0047312   1.771   0.0766 .
AGE:DIAGNOSIS41071  0.0076790  0.0030957   2.481   0.0131 *
AGE:DIAGNOSIS41081  0.0055119  0.0039307   1.402   0.1609
AGE:DIAGNOSIS41091  0.0039354  0.0029202   1.348   0.1778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.3072893)

    Null deviance: 3530.4  on 11227  degrees of freedom
Residual deviance: 3157.1  on 11198  degrees of freedom
AIC: 61873

Number of Fisher Scoring iterations: 5
```

9.2 Predicted Deaths

```
Call:
glm(formula = LOS ~ 1 + CHARGES_PRED, family = Gamma(log), data = deaths)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3334  -0.5024  -0.1419   0.2571   1.8410

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.411e-01  4.641e-01   2.028  0.04321 *
CHARGES_PRED 1.159e-04  3.988e-05   2.906  0.00385 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.379369)

    Null deviance: 161.55  on 420  degrees of freedom
Residual deviance: 158.38  on 419  degrees of freedom
AIC: 2582.8

Number of Fisher Scoring iterations: 5
```

10 Appendix for FFNN Architectures

10.1 CHARGES FFNN Architecture

```
model = keras_model_sequential()
model %>%
  layer_dense(units = 512, activation = "tanh", input_shape = c(ncol(X_train_nn))) %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 256, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 128, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 64, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 32, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 16, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 1, activation = "linear") %>%

summary(model)

model %>% compile(
  loss = "mse",
  optimizer = optimizer_rmsprop(),
  metrics = c("mse")
)

history = model %>% fit(
  as.matrix(X_train_nn), as.matrix(y_train_nn),
  epochs = 30, batch_size = 128*20,
  validation_split = 0.2,
  shuffle=TRUE
)
```

10.2 DIED FFNN Architecture

```
model = keras_model_sequential()
model %>%
  layer_dense(units = 512, activation = "tanh", input_shape = c(ncol(X_train_nn))) %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 256, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 128, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 64, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 32, activation = "tanh") %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 1, activation = "sigmoid")

summary(model)

model %>% compile(
  loss = "binary_crossentropy",
  optimizer = optimizer_rmsprop(),
  metrics = c("accuracy")
)

history = model %>% fit(
  as.matrix(X_train_nn), as.matrix(y_train_nn),
  epochs = 200, batch_size = 128*20,
  validation_split = 0.2,
  shuffle=TRUE
)
```

References

[1] Benjamin, EJ., et al. Heart Disease and Stroke Statistics 2018 At-a-Glance. (2018, January 31). Retrieved from https://www.heart.org/-/media/data-import/downloadables/heart-disease-and-stroke-statistics-2018—at-a-glance-ucm_498848

[2] Mozaffarian, D. et al. (2015, January 27). American Hearth Association [Powerpoint.] Retrieved from https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_449846.pdf

[3] Chest Heart & Stroke Scotland Head Office. (2019). Retrieved from https://www.chss.org.uk/heart-information-and-support/about-your-heart-condition/common-heart-conditions/heart-attack/what-happens/