STAT 149: Final Project Report

Karina Huang, Nicholas Stern, Phoebe Wong

May 8, 2019

1. BACKGROUND

Modeling patients' length of stay is an important task for hospital administration. By understanding what factors into predicting a patient's length of stay, hospitals can coordinate and distribute medical resources more efficiently. In the current study, we modeled the length of stay of 12,844 patients with Acute Myocardial Infarction (AMI) across hospitals in New York State, in 1993. Our objectives were to investigate how the provided attributes relate to the subjects' length of stay, and identify the best generalized linear, or additive model to parameterize these relationships.

2. EXPLORATORY DATA ANALYSIS

We first examined pairwise correlations of all variables in the data. Figure 1 shows that the response variable, length of stay (*los*), and *charges* are strongly, positively correlated ($r^2 = 0.714$). This is intuitive because a longer hospital stay is going to naturally be more expensive.

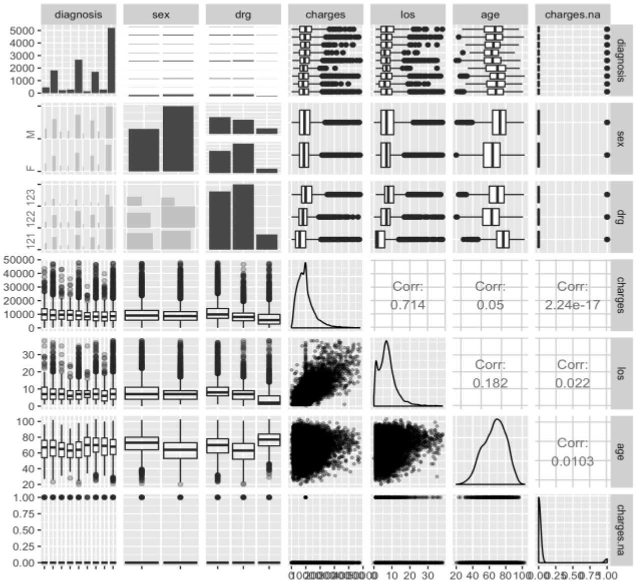


Figure 1. Pairwise correlation and comparison of all variables in the dataset.

Figure 2 shows that *los* is right-skewed, ranging from 0 to 38 days, with a median of 7 days. Further inquiry into the distribution of the response variable revealed that there was only a single zero-valued point. This prompted a discussion over where this originates, detailed further in the discussion section.

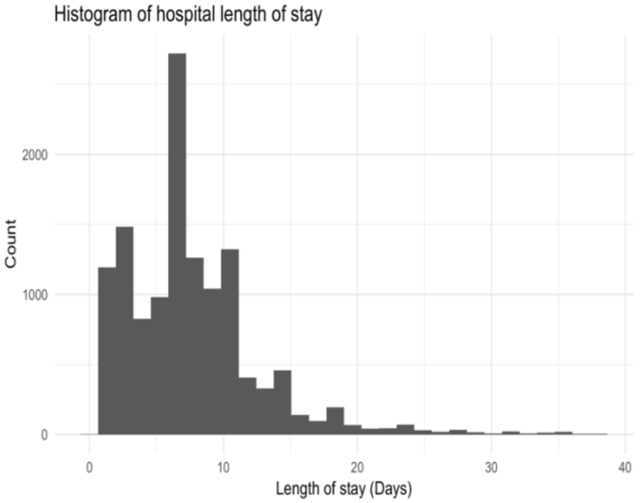


Figure 2. Histogram of response variable - hospital length of stay in days.

The distribution of *los* by survival status (*died*) of patients indicated that a substantial fraction of admitted patients who died, perished on the first day of their hospital stay from the heart attack (see Figure 3 for visualization).

3. METHODOLOGY

3.1 Data Preprocessing

Before building our models, we preprocessed the data in the following ways:

- Cast the categorical variables as factors.
- Removed *patient id* as it should have no connection to the response.

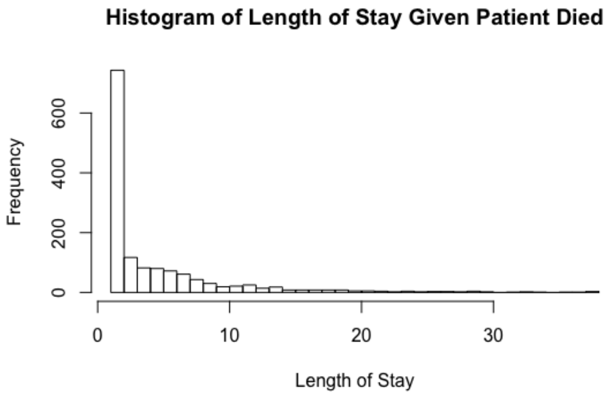


Figure 3. Distribution of length of stay given the patient died.

- Removed *died*, as it was perfectly collinear with *drg* of 123.
- Imputed missing values in the *charges* column with the mean, and added a binary missing indicator, *charges.na*, as an additional predictor.
- For gamma models only, we imputed the single, zero-valued datapoint in the dataset with 1. We chose to impute in the absence of conclusive evidence that this was an erroneous datapoint. Our decision to round to 1 was to emulate how we would cast a gamma model prediction that is less than 0.5 to a positive integer value.
- Split the data into training and test sets for inter-model comparisons later on.

3.2 Model Selection

We chose to model the response variable with three different distributions:

- *Poisson*: Because our response variable, *los*, is a discrete count, our first idea was to model the variable as a Poisson process.
- *Negative Binomial*: We can treat *los* as a negative binomial random variable with r = 1, if we consider each day to be a trial, and the patient's departure to be a failure.
- *Gamma distribution*: Time, inherently, is a continuous variable. Hence, *los* can be thought of as aggregated chunks of a latent variable that spans the real number line. Additionally, our response variable exhibits right-skewness, therefore it would be appropriate to fit our data with a gamma model.

For each type of model, we selected the best subset of predictors using stepwise, bidirectional feature selection with the AIC criterion. We chose to perform feature selection on the base set of predictors and set with interactions separately, then determine whether the inclusion of interaction terms significantly improved the residual deviance using likelihood ratio tests.

3.3 Model Diagnostics

The following diagnostics were performed on the selected models:

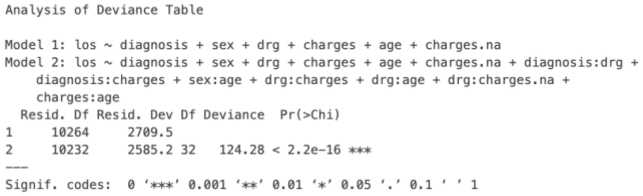
- Plotted the deviance residuals vs. fitted values to look for nonlinearity and excessive spread.
- Visualized the Cook's distances to check for influential points.
- For GLMs, we compared the ratio of residual deviance over degrees of freedom to 1.
- Examined the VIF for a model trained without interactions. This is because the inclusion of interactions inflates the multicollinearity, while we are interested in the collinearity between the original set of predictors.

As an additional step, for each type of model we smoothed the quantitative predictors to create a Generalized Additive Model (GAM). We then performed a likelihood ratio test comparing the GAM with previously selected model of the same type. Note that for the negative binomial model, in order to make this comparison with the *anova* function in R, we had to cast the GAM as a negative binomial model (using glm.nb).

Therefore, we trained a proxy model on predictors transformed according to the exponential power approximated by the smooths. For each type of model, we picked the best fit and summarized the Root Mean Squared Error (RMSE) on a held out test set.

4. RESULTS

As outlined in the Methodology section, for each of the three probability distributions, we pre-processed the data and performed feature selection to choose the best fit with and without interaction terms. In each scenario, a likelihood ratio test (LRT) between the two suggested that the model with interaction terms performed better. Figure 4 shows example output from the LRT for the gamma GLM's.



```
Analysis of Deviance Table

Model 1: los ~ diagnosis + sex + drg + charges + age + charges.na
Model 2: los ~ diagnosis + sex + drg + charges + age + charges.na + diagnosis:drg +
    diagnosis:charges + sex:age + drg:charges + drg:age + drg:charges.na +
    charges:age
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     10264     2709.5
2     10232     2585.2 32   124.28 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4. Analysis of deviance for gamma GLM's.

The diagnostic plots of deviance residuals versus fitted values for the Poisson and negative binomial models fell well outside the expected range of -2 to 2. Meanwhile, for all three probability distributions, the residuals were not symmetric about 0, veering negative for large fitted values as demonstrated in Figure 5. This was because the fitted estimations consistently overshot the length of stay at larger values. For the Poisson GLM, we noted that the ratio of the residual deviance over the degrees of freedom was 1.5, which is larger than the ideal value of 1. For all three distributions, visuals of Cook's distances did not reveal any influential points, and the VIF indicated there was no multicollinearity of note.

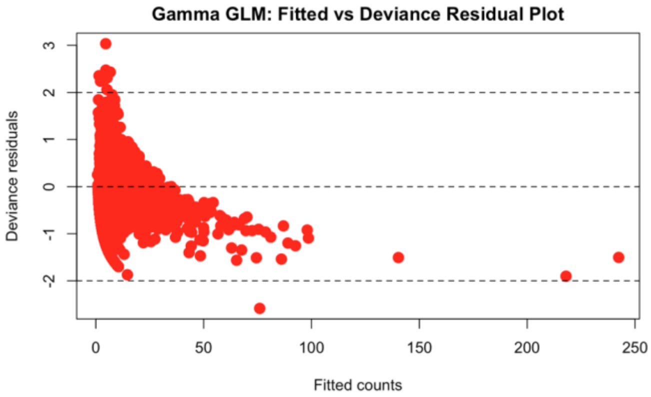


Figure 5. Deviance residuals vs. fitted values plot for the gamma GLM with interaction effects.

Next, we examined nonlinear contributions of the quantitative predictors using GAM's. We chose to smooth *age* and *charge* as they were the only quantitative predictors in our dataset. For each distribution we compared our selected regression with its corresponding GAM through an LRT. In all cases, the LRT's indicated that the GAM yielded the better fit. After visualizing the deviance residuals vs. fitted values once more, we saw that the deviance residuals were more symmetric about 0 at higher values, and that the fitted values did not overestimate as much as shown in Figure 6. These three GAM models, smooth exponent estimates, and train/test set performances are summarized in Table 1.

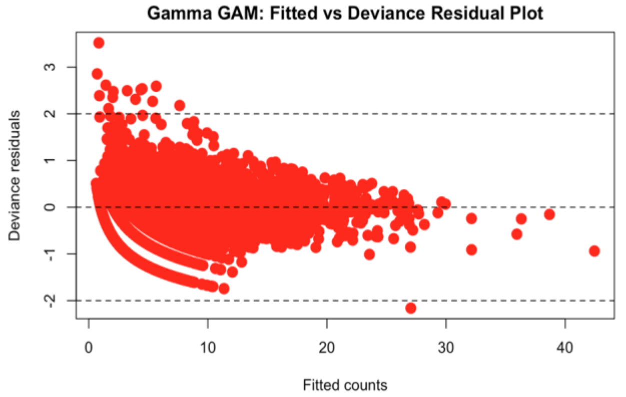
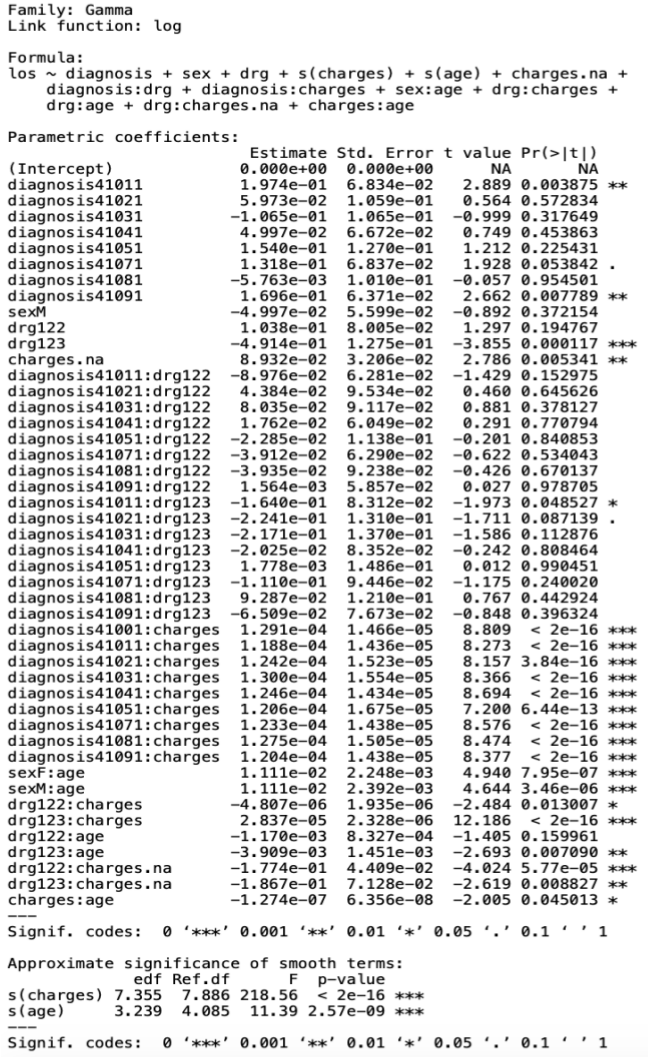


Figure 6. Deviance residuals vs. fitted values plot for gamma GAM.

Source Code: https://github.com/phoebewong/Stat149-Final-Project

5. CONCLUSION

Our results indicated that the GAM's outperformed their respective regression counterparts for all types of probability distributions. This makes sense because the GAM's capture non-linear contributions of the predictors to our response variable. For conciseness, we provide below detailed interpretations of the coefficients of the gamma GAM, which yielded the best-behaving deviance residuals. For summaries of other models, please refer to our GitHub page linked in the upper right corner.



```
Family: Gamma
Link function: log

Formula:
los ~ diagnosis + sex + drg + s(charges) + s(age) + charges.na +
    diagnosis:drg + diagnosis:charges + sex:age + drg:charges +
    drg:age + drg:charges.na + charges:age

Parametric coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             0.000e+00  0.000e+00      NA       NA    
diagnosis41011          1.974e-01  6.834e-02   2.889 0.003875 ** 
diagnosis41021          5.973e-02  1.059e-01   0.564 0.572834    
diagnosis41031         -1.065e-01  1.065e-01  -0.999 0.317649    
diagnosis41041          4.997e-02  6.672e-02   0.749 0.453863    
diagnosis41051          1.540e-01  1.270e-01   1.212 0.225431    
diagnosis41071          1.318e-01  6.837e-02   1.928 0.053842 .  
diagnosis41081         -5.763e-03  1.010e-01  -0.057 0.954501    
diagnosis41091          1.696e-01  6.371e-02   2.662 0.007789 ** 
sexM                   -4.997e-02  5.599e-02  -0.892 0.372154    
drg122                  1.038e-01  8.005e-02   1.297 0.194767    
drg123                 -4.914e-01  1.275e-01  -3.855 0.000117 ***
charges.na              8.932e-02  3.206e-02   2.786 0.005341 ** 
diagnosis41011:drg122  -8.976e-02  6.281e-02  -1.429 0.152975    
diagnosis41021:drg122   4.384e-02  9.534e-02   0.460 0.645626    
diagnosis41031:drg122   8.035e-02  9.117e-02   0.881 0.378127    
diagnosis41041:drg122   1.762e-02  6.049e-02   0.291 0.770794    
diagnosis41051:drg122  -2.285e-02  1.138e-01  -0.201 0.840853    
diagnosis41071:drg122  -3.912e-02  6.290e-02  -0.622 0.534043    
diagnosis41081:drg122  -3.935e-02  9.238e-02  -0.426 0.670137    
diagnosis41091:drg122   1.564e-03  5.857e-02   0.027 0.978705    
diagnosis41011:drg123  -1.640e-01  8.312e-02  -1.973 0.048527 *  
diagnosis41021:drg123  -2.241e-01  1.310e-01  -1.711 0.087139 .  
diagnosis41031:drg123  -2.171e-01  1.370e-01  -1.586 0.112876    
diagnosis41041:drg123  -2.025e-02  8.352e-02  -0.242 0.808464    
diagnosis41051:drg123   1.778e-03  1.486e-01   0.012 0.990451    
diagnosis41071:drg123  -1.110e-01  9.446e-02  -1.175 0.240020    
diagnosis41081:drg123   9.287e-02  1.210e-01   0.767 0.442924    
diagnosis41091:drg123  -6.509e-02  7.673e-02  -0.848 0.396324    
diagnosis41001:charges  1.291e-04  1.466e-05   8.809  < 2e-16 ***
diagnosis41011:charges  1.188e-04  1.436e-05   8.273  < 2e-16 ***
diagnosis41021:charges  1.242e-04  1.523e-05   8.157 3.84e-16 ***
diagnosis41031:charges  1.300e-04  1.554e-05   8.366  < 2e-16 ***
diagnosis41041:charges  1.246e-04  1.434e-05   8.694  < 2e-16 ***
diagnosis41051:charges  1.206e-04  1.675e-05   7.200 6.44e-13 ***
diagnosis41071:charges  1.233e-04  1.438e-05   8.576  < 2e-16 ***
diagnosis41081:charges  1.275e-04  1.505e-05   8.474  < 2e-16 ***
diagnosis41091:charges  1.204e-04  1.438e-05   8.377  < 2e-16 ***
sexF:age                1.111e-02  2.248e-03   4.940 7.95e-07 ***
sexM:age                1.111e-02  2.392e-03   4.644 3.46e-06 ***
drg122:charges         -4.807e-06  1.935e-06  -2.484 0.013007 *  
drg123:charges          2.837e-05  2.328e-06  12.186  < 2e-16 ***
drg122:age             -1.170e-03  8.327e-04  -1.405 0.159961    
drg123:age             -3.909e-03  1.451e-03  -2.693 0.007090 ** 
drg122:charges.na      -1.774e-01  4.409e-02  -4.024 5.77e-05 ***
drg123:charges.na      -1.867e-01  7.128e-02  -2.619 0.008827 ** 
charges:age            -1.274e-07  6.356e-08  -2.005 0.045013 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
             edf Ref.df      F  p-value    
s(charges) 7.355  7.886 218.56  < 2e-16 ***
s(age)     3.239  4.085  11.39 2.57e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7. gamma GAM model summary

Figure 7 shows the gamma GAM model summary. P values of the smooth terms indicated significant, non-linear relationships between the *age* and *los*, and *charges* and *los*. We interpret *diagnosis41011* as an example interpretation for a main effect. The ICD-9 code 41011 is defined to be an "Acute Myocardial Infarction of other anterior wall, initial episode of care." From the summary we see that *diagnosis41011* is associated with an average multiplicative increase of $e^{.197}$ = 1.22 in the estimated length of stay compared to *diagnosis41001* (reference group of diagnosis), holding all other variables constant. The model summary also shows the interaction effect between *sexM* and *age* is a significant predictor of *los*. On average, holding all other variables constant, the multiplicative effect of being male vs. female on the estimated length of stay increases by a factor of $e^{.011}$ = 1.01 with every one-year increase in *age*. Note that this difference is marginal, but significant. In addition, it is interesting that the model also reported interaction effects between *sexF* and *age*, given that sexF is the reference category, and presumably is redundant information.

The effect of *charges.na* is also significant according to the model summary. This implies that, while holding other variables constant, the effect of missing values in *charges* contributes about $e^{.089}$ = 1.09 more beyond the mean value of *charges* to the predicted response. In other words, missing values are significantly positively correlated with longer hospital stays, which could be an indication that there is a systematic reason for the missing values. For example, a hypothetical rationale could be that charges go missing when the hospital stay is free (the minimum of the charge distribution is $3), and patients who don't get charged may stay longer on average.

As a sanity check, one would expect a negative correlation between survival status and length of stay; in particular, we expect those who die to have a shorter length of stay. The model summary supports this speculation in that *drg123*, as an indication of death, had a significant negative relationship with the response variable. However, it is interesting that having cardiovascular complications (*drg122*) does not predict significantly longer stays, as cardiovascular complications, could lead to further medical attention.

Model Type	**Age | Charge EDF**	**Training RMSE**	**Test RMSE**
Poisson	8.44 | 7.78	3.27	3.39
Negative Binomial	3.98 | 7.58	3.28	3.41
Gamma	3.24 | 7.35	3.31	3.49

Table 1. Performance comparison between GAM's.

6. DISCUSSION

Our goal for the project was to investigate the relationship between each provided feature and the response variable – length of stay. Using our knowledge from STAT149, we modeled the response variable with three reasonable distributions. Because the models assumed different response distributions, we could not perform likelihood comparisons, and therefore we used training and test prediction errors to compare model fit. Note that the results do not definitively prove a model's worth, as the test and train accuracies are similar. It is possible that none of these models do a great job of fitting the response variable. Specifically, all of the resulting coefficients appeared to be quite small, regardless of their significance. This may be because there is not a lot of correlation between the set of predictors and the response variable.

We spent a lot of time discussing how to handle the presence of the zero-valued length of stay before fitting the gamma models. After a thorough investigation we outlined the following options:

- Remove the datapoint under the assumption that it was incorrectly entered, using the length of stay distribution for those who died as evidence that same-day discharge is equivalent to a length of stay of 1.
- Remove the datapoint under the assumption that it was correctly entered, effectively imposing a floor on our data and reflecting that in the model inference.
- Shift all the response values by +1 under the assumption that the datapoint is accurate, and there is one person who was discharged and/or had their paperwork completed on the same day. Adjust the model interpretation correspondingly.
- Change the datapoint to be 1, under the assumption that the data was generated using some kind of rounding to the nearest day, and to coincide with what we would likely do with our Gamma model's predictions that fall below 0.5.

We also considered the option of fitting a hurdle model to the data, where the zero population would come from a binomial distribution and the rest of the data would come from a gamma distribution. However, we determined it was not meaningful to fit a model parameter to a single datapoint, thereby electing to forgo this approach. Ultimately, for the reasons mentioned in the Methodology section, we chose to round the datapoint to 1. Note that any treatment of this datapoint would not impact the model fit due to the size of the data and the absence of influential points.

Source Code: https://github.com/phoebewong/Stat149-Final-Project

It is important to qualify the use of our models. If the purpose is to facilitate medical resources, like bed assignments, the desired model should be able to predict patients' length of stay at the time of hospital admittance. However, the information on *charges* would not be available when predictions are made, and therefore, our models are not going to be able to predict length of stay in practice. However, because our objective for the current study was to establish relationships between the predictors and the length of stay, we chose to keep the feature in accordance with stepwise selection results. One interesting idea we discussed was the possibility of normalizing charges by the length of stay. Normalization would uncouple the obvious connection between length of stay and price, but transform *charges* into a proxy for average price differences between hospitals or medical procedures that are more expensive.

Moving forward, we are interested to investigate whether other non-linear transformations of the predictor variables (e.g. log-transformation for right-skewed predictors like *charges*) could improve the predictive power of the model. Additionally, it is possible that the true relationship between the length of stay and the set of features is better fitted by other types of statistical or nonparametric models, such as decision tree regression or neural networks.