Cole Yant
Andrew Huang
Ryan DeStefano

STAT 334 Final Project

The Initial Model

The initial model consisted of data from 113 US hospitals. The response variable, LenStay, denoted the average length of stay for a patient (in days). There were 10 variables that were used to predict the value of LenStay (Table 1a). They were as follows:

1) **Age** – average patient age (in years)
2) **InfRisk** – percent chance of acquiring an infection while in the hospital (in percent)
3) **Culture** – number of cultures performed per 100 patients without signs of hospital acquired infection
4) **XRay** – number of x-rays performed per 100 patients without signs of pneumonia
5) **School** – Is the hospital associated with a medical school (yes or no)
6) **Region** – region of the country (North East, North Central, South, West)
7) **Beds** – number of beds at the hospital
8) **Census** – average daily number of patients at the hospital
9) **Nurses** – average number of nurses at the hospital
10) **Services** – percentage of a list of medical services that are provided at the hospital (in percent)

We discovered that the Beds variable and the Census variable had extremely high VIFs (35.699204 and 34.211423, respectively), whereas the Nurses variable had a moderately high VIF of 7.055523 (Table 1b). Furthermore, we found that observations 8 and 112 had high leverages, while observation 47 was an outlier (Plot 1c).

Fixing the Initial Model

We found no violations of the ANOVA assumptions in the initial model (Plots 1a/1b). However, using the cor(), function, we determined that there was a 0.981 correlation between Beds and Census (Table 1b). Therefore, we created a new model without the Beds variable. In this new model, we found moderate multicollinearity between Census and Nurses (Table 2a), so we transformed the variables. This was done by dividing the Nurses variable by the Census variable, helping us to determine the average number of nurses per patient. This created a new variable, NursePerson, which replaced Nurses, thereby resolving the issues with multicollinearity (Table 3a).

Because observations 8 and 112 were the only observations with high leverages, we decided that their effects on the accuracy of the model were negligible. The same reasoning was applied to the outlier, observation 47.

Choosing the Final Model

When selecting our final model, we used the best subsets method (Table 3b). Thus, we chose the options that have the highest Adj. R-square values while also having the lowest Cp, AIC, and SBC. There were two models that we deemed satisfactory. The first contained the variables LenStay, Age, InfRisk, XRay, Region, Census, and NursePerson. The second contained the same variables in addition to the Services variable. After looking at the ANOVA tables for both models, we found that Services was not significant, so we proceeded with the first model (Table 3c).

While checking the assumptions, we found that there was a slight curve off the diagonal in the Quantile to Quantile plot. However, only 4 of the 113 observations trailed off, and so their effects on the overall Normality of the model were negligible (Plot 4a). All other assumptions (linearity, equal variance, and independence) were satisfied (Plot 4a).

We then used a stepwise regression with all possible interactions, this resulted with the potential best model being the one with only the interaction between InfRisk and Census being added and the predictor age being omitted from the model (Table 4a). From there we tried adding more interactions to the model with no success as described in the next section.

Possible Interactions

Based on differences between regions in the United States, we figured there might be significant interactions between the Region variable and other predictors. One of the interactions that we tried adding to the model was Region interacting with NursePerson. This interaction was significant (p-value = 0.014), but its addition caused the variable XRay to become insignificant (p-value = 0.07). Therefore, we decided not to add this interaction. We also tried to include an interaction between Region and Census, but the results of this addition were very similar to that of the previous interaction. The interaction was significant (p-value = 0.04) but caused the XRay predictor to become insignificant (p-value = 0.07). Thus, we decided to go with our smaller model. This pattern happened again when trying to add an interaction between InfRisk and Region. Interaction term p-value, 0.03, and XRay p-value, 0.08. (Results discussed here may be seen in detail in Table 4b.)

Fixing the Final Model

With our final model we did not find any issues with ANOVA assumptions (Plot 5a), but we did find that there was extreme multicollinearity between the interaction and Census(Table 5a). To fix this, we centered InfRisk and Census because they are contained in the interaction. We also centered XRay and NursePerson since the two variables are negatively affected by each other (see the section on “Possible Interactions”). Additionally, we checked for influential observations. Results showed there to be none exceeding the Cook’s Distance limit of 0.94 (Plot 6a).

We compared this revised final model (reduced model) to a test model (full model), which contained all the predictors from the original model as well as all the predictors that were added to our final model. This comparison was done with the anova command, and the test resulted in a p-value of 0.341, meaning that our final model wsa more accurate than the full model (Table 6b).

<u>Interpretation of the Coefficients</u>

1) **Intercept –** When all other predictors are average, in the North Central region, the mean average length of stay at the hospital is 9.393 days.
2) **cInfRisk –** After adjusting for all other predictors, each increase in 1% chance of acquiring an infection while in the hospital is associated with a 0.624 increase in the mean average length of stay in the hospital.
3) **cXRay –** After adjusting for all other predictors, each additional x-ray performed per 100 patients without signs of pneumonia is associated with a 0.016 increase in the mean average length of stay.
4) **RegionNE –** After adjusting for all other predictors, states in the East region are associated with 0.957 more mean average days spent in the hospital for patients compared to states in the North Central region.
5) **RegionS –** After adjusting for all other predictors, states in the South region are associated with 0.298 less mean average days spent in the hospital for patients compared to states in the North Central region.
6) **RegionW –** After adjusting for all other predictors, states in the West region are associated with 0.788 less mean average days spent in the hospital for patients compared to states in the North Central region.
7) **cCensus –** After adjusting for all other predictors, each increase in 1 of average daily number of patients at the hospital is associated with a 0.002 increase in the mean average length of stay in the hospital.
8) **cNursePerson –** Each increase of one percentage point in the ratio of average number of nurses to average daily number of hospital patients is associated with a decrease of 1.059 days in the mean average length of hospital stay of patients after adjusting for the other predictors.
9) **cInfRisk:cCensus: –** After adjusting for other predictors, when there is an average daily number of patients in the hospital, each increase in 1% chance of acquiring an infection while in the hospital is associated with an increase of 0.003 days in the mean average length of stay in the hospital.
10) **Adj R-Squared –** 59.69% of the variation in average length of stay at the hospital can be accounted for by the regression of average length of stay at the hospital vs all other predictors.