Project 1

Finding the Strongest Relationship:

The average length of stay at hospital relating to infection risk, available facilities and services, and routine chest X-ray ratio

Jasper Dong
Allison Peng

Amy Kim
STA 108

Introduction:

The dataset, SENIC, consists of numerical data that consists of 4 variables. Length is the average length of stay of all patients in hospital in days, Infection is the average estimated probability of acquiring infection in hospital in percent, facility is the percent of 35 potential facilities and services that are provided by the hospital, Xray is the ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100. We will use infection, facility, and x-ray data as the explanatory variable, and the average length of stay as the response variable. We aim to use a simple normal linear regression model to inspect the relationship between each explanatory variable and response variable. By comparing each model, we will determine what explanatory variable has the strongest relationship with the average length of stay of all patients.

Summary:

We will first conduct exploratory data analysis to inspect the individual data types of each variable, as well as the initial relationship between the explanatory variables and response variable. The mean values and the standard deviations are further explored in the table below. It is important to note that the means of Length and Infection are significantly smaller than the means of Facility and Xray. In general, the center of Length is closer to Infection than Facility and Xray. Furthermore, the standard deviations of Facility and Xray are larger than Length and Infection, suggesting a larger spread for the Facility and Xray variables.

We looked at the relationship between each explanatory variable with the Length variable. When comparing Infection with Length, the data appears to follow a linear relationship, however the slope appears to be closer to 0. When comparing Facility with Length, there are clear gaps between the values on the x-axis. This shows that the percentage data is recorded in specific intervals, and can affect the linear model with the length. When comparing Xray and Length, the scatterplot shows a similar relationship to the infection vs length graph. There are two obvious outliers, or points that deviate from the dataset in each of the plots.

We will further inspect the outliers in the next section and determine what to remove from the dataset.

Data Preparation:

We found two outliers in the dataset, in rows 47 and 112 of the dataset. We used the studentized residuals methods to find the outliers: this technique first divides the residual by the estimate of the standard deviation. We then compared each value to a cutoff value determined by the alpha value and t distribution. Values that are larger than the cutoff value are considered outliers, and we removed them from the dataset. This step is necessary to continue to model fitting because the outliers can influence the regression coefficients, thus making the regression line inaccurate to the dataset. After removing the outliers, we can now attempt to fit the model to all three relationships.

Model Fitting:

We aim to use a linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon \qquad i = 1 \ldots n$$

When regressing the response variable to each explanatory variable, we can observe how well the model fits with each dataset. This resulted in three estimated regression lines, and we need to determine the statistically best model. In order to do so, we can compare the R^2 values for each model, or the variance explained by the model divided by the total variance. The highest R^2 value is 0.3019 thus, we will use this dataset for the model diagnostics and interpretation.

When comparing estimated σ^2s, we created ANOVA tables for each model and compared their MSEs, or the average squared difference between observed and predicted values for each model. We found that the linear model of

Facility vs. Length has the highest MSE of 1.898, while Infection vs. Length has the smallest MSE of 1.532 which indicates that on average, predicted values of Infection vs. Length will be fairly close to observed values, while predicted values of Facility vs. Length will be comparably farther away.

After choosing our best model, we move on to model diagnostics to test if the assumptions of Normal linear regression hold.

Model Diagnostics:

Using the relationship between Infection and Length as our best model, we performed model diagnostics to see if the assumptions of the Normal linear regression hold. The assumptions we test for are:

1) Regression function is linear
2) Error terms are independent
3) Error terms are normally distributed
4) Error terms have constant variance

When assessing linearity, we performed a general linear F-test for H_0: $\beta_1 = 0$, and obtained a p-value of 0.0000000004233. Because our p-value is so small, we reject H_0, and conclude that there is a linear relationship between probability of acquiring infection while at the hospital and average length of stay.

When assessing independence of error terms, we created a residual index plot, and observed random scatter of the resulting points. This indicates that the error terms are independent.

When assessing normality of error terms, we conducted a Shapiro-Wilks test to test H_0: the errors are normally distributed vs. H_A: the errors are not normally distributed, and obtained a p-value of 0.662. Since this is larger than any reasonable alpha, we fail to reject H_0, and conclude that the errors are normally distributed.

When assessing constant variance of error terms, we conducted a Fligner-Killeen test to separate the error terms into two groups, and tested H_0: there are equal variances between the upper and lower groups vs. H_A: there are unequal variances between the upper and lower groups, and obtained a p-value of 0.9347. Since this is larger than any reasonable alpha, we fail to reject H_0, and conclude that the error terms have constant variance.

Based on our diagnostics, we conclude that the assumptions of the Normal linear regression hold for the simple linear regression model of Infection vs. Length. After performing model diagnostics, we now move on to interpreting our best model.

Interpretation:

When Infection increases by 1%, we expect Length to increase by 0.60975 days on average. Additionally, when Infection is 0%, we expect Length to be 6.8492 days on average.

Based on our constructed confidence intervals for the parameters, we are 95% confident that when Infection increases by 1%, Length would tend to increase by between 0.4337345 and 0.7857618 days. Additionally, we are 95% confident that when Infection is 0%, we would expect Length to be between 6.0537164 and 7.6447306 days on average.

When calculating R^2 for the model of Infection vs. Length, we obtained a value of 0.3019, meaning that 30.19% of the total variability in Length is explained by its linear regression on Infection.

Conclusion:

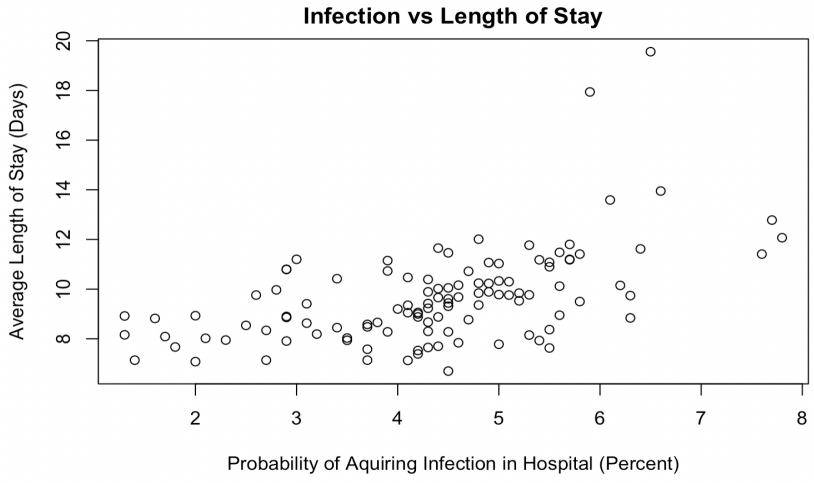
Based on our findings, we found the simple linear model between Infection vs. Length is our statistically best model - compared to Facility vs. Length and Xray vs. Length - because of its relatively high R^2, low estimated σ^2, and because it satisfies the assumptions of Normal linear regression.

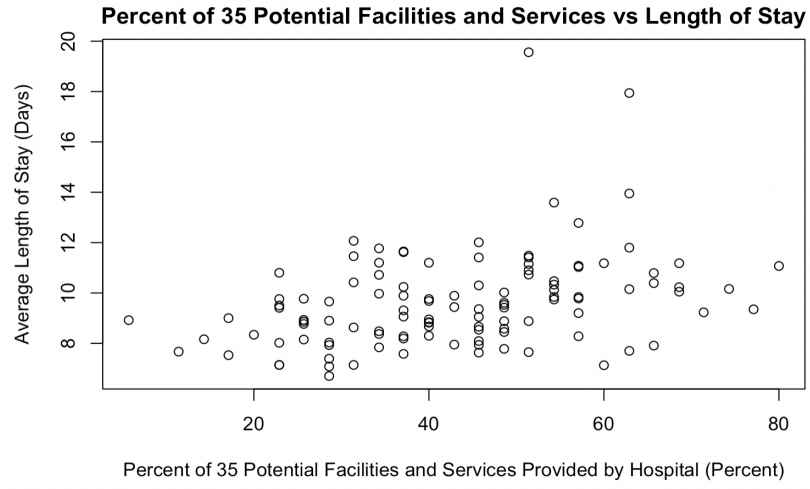
One limitation of our final model (Infection vs. Length) is that our calculated R^2 of 0.3019 is relatively small, indicating a very weak correlation between Infection and Length.

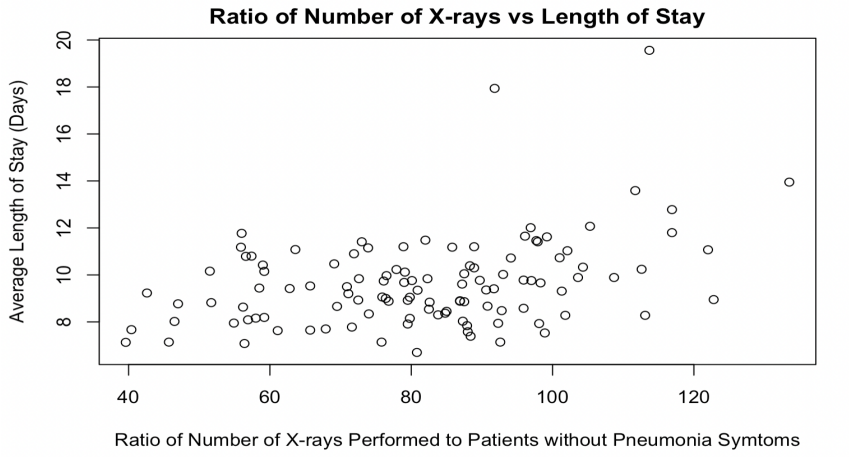
Tables/Plots

I) Data Preparation:

i) Scatterplots:







ii) Number Summaries:

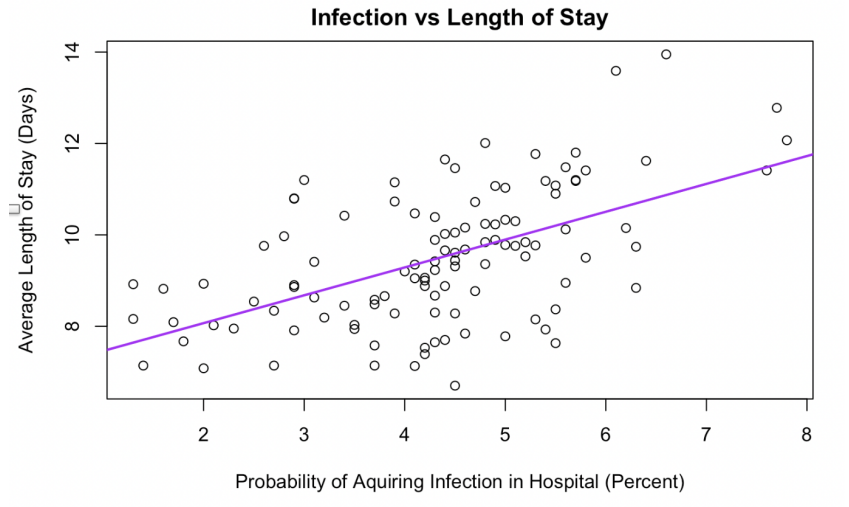
Mean and Standard Deviation for each variable

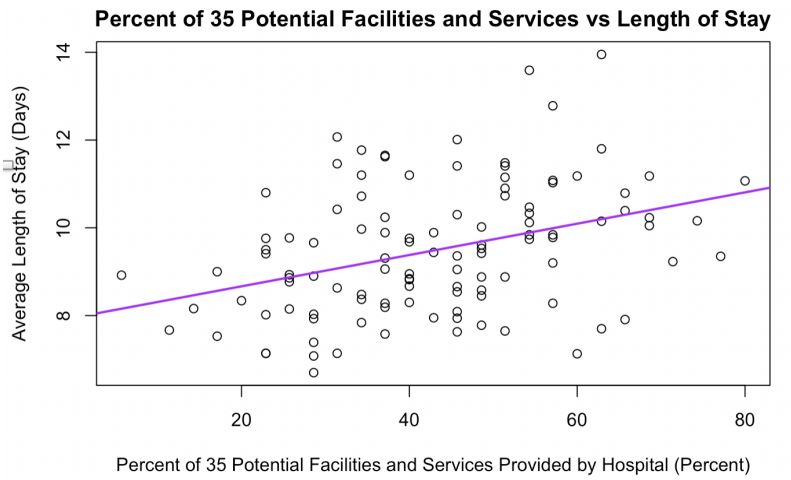
	Length	Infection	Facility	Xray
Mean	9.648319	4.354867	43.15929	81.62832
Standard Deviation	1.911456	1.340908	15.20086	19.36383

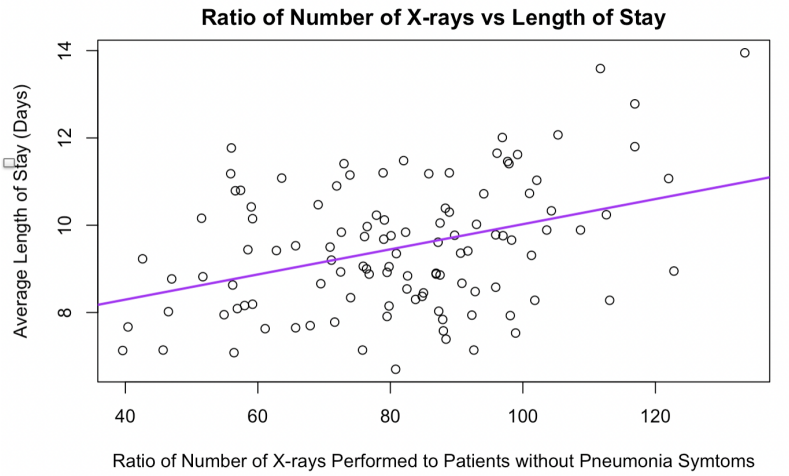
II) Model Fitting

i) Graphs of each explanatory variable vs response variable and estimated regression line

* The purple line indicates the estimated regression line





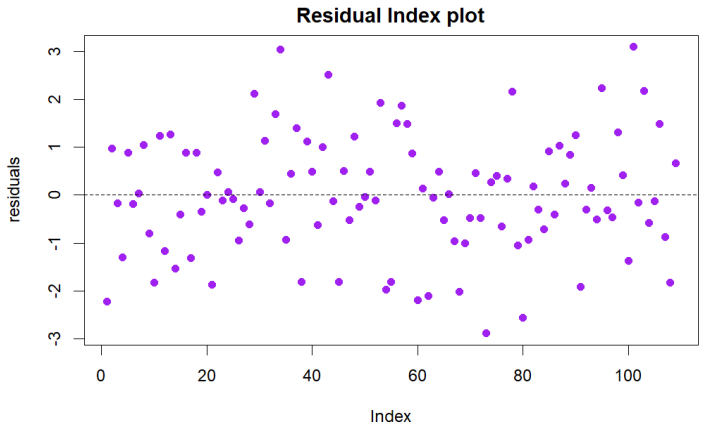


ii) Summary of R^2 and MSE Values

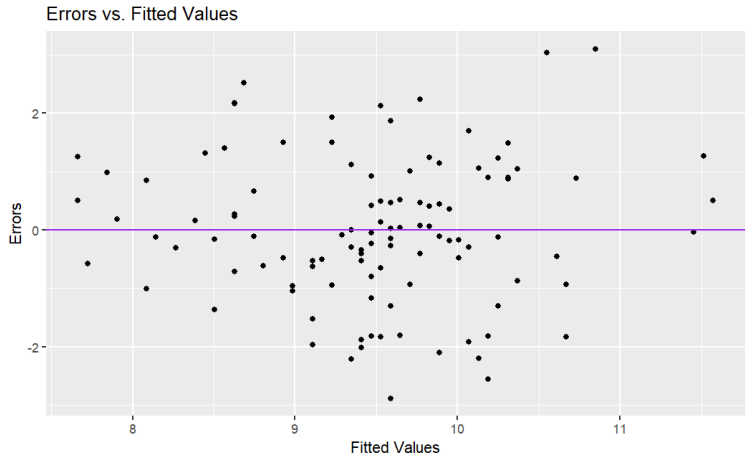
	Infection vs Length	Facility vs Length	Xray vs Length
R^2	0.3019143	0.1353206	0.141625
$E(\sigma^2)$ = MSE	1.532	1.898	1.884

III) Model Diagnostics

i) Assessing Independence



ii) Assessing Constant Variance



iii) Hypothesis Tests for Constant Variance and Normality of Errors

	Fligner-Killeen Test	Shapiro-Wilks Test
P-value	0.9347	0.662

iv) Confidence Intervals for Parameters:

	2.5%	97.5%
Intercept	6.0537164	7.6447306
Infection	0.4337345	0.7857618

STA108 Project I

Jasper Dong, Allison Peng

2023-05-01

R Appendix

```
knitr::opts_chunk$set(echo = FALSE, comment = NA)
options(scipen = 999) #Remove the scientific notation
library(readr)
library(MASS)
library(ggplot2)
SENIC <- read_csv("SENIC (1).csv")
SENIC
# plot exploratory data analysis scatterplots
plot(SENIC$infection, SENIC$length, xlab = "Probability of Aquiring Infection in Hospital (Percent)", yl
plot(SENIC$facility, SENIC$length, xlab = "Percent of 35 Potential Facilities and Services Provided by I
plot(SENIC$Xray, SENIC$length, xlab = "Ratio of Number of X-rays Performed to Patients without Pneumonia

# find mean for each variable
mean(SENIC$length)
mean(SENIC$infection)
mean(SENIC$facility)
mean(SENIC$Xray)

# find SD for each variable
sd(SENIC$length)
sd(SENIC$infection)
sd(SENIC$facility)
sd(SENIC$Xray)

# removing outliers
outliers = c(47, 112)
SENIC1 = SENIC[-outliers,]
SENIC1
##### MODELS #####
infect.model = lm(length ~ infection, data = SENIC1)
facility.model = lm(length ~ facility, data = SENIC1)
xray.model = lm(length ~ Xray, data = SENIC1)
SENIC1$ei1 = infect.model$residuals
SENIC1$yhat1 = infect.model$fitted.values
SENIC1

# plotting the estimated regression line
plot(SENIC1$infection, SENIC1$length, xlab = "Probability of Aquiring Infection in Hospital (Percent)",
abline(new.model1,col = "purple",lwd = 2)
```

```
plot(SENIC1$facility, SENIC1$length, xlab = "Percent of 35 Potential Facilities and Services Provided by
abline(new.model2,col = "purple",lwd = 2)

plot(SENIC1$Xray, SENIC1$length, xlab = "Ratio of Number of X-rays Performed to Patients without Pneumon
abline(new.model3,col = "purple",lwd = 2)

# finding the r^2 value for each model
cor(SENIC1$infection, SENIC1$length)^2
cor(SENIC1$facility, SENIC1$length)^2
cor(SENIC1$Xray, SENIC1$length)^2

 # ANOVA tables for each model
anova(infect.model)
anova(facility.model)
anova(xray.model)
  # Assessing Infection Linearity
options(scipen = 8)
reduced.model = lm(length ~ 1, data = SENIC1)
anova.table = anova(reduced.model, infect.model)
anova.table
  # Assessing Infection Independence
plot(infect.model$residuals, main = "Residual Index plot", xlab = "Index", ylab = "residuals", pch = 19
abline(h = 0, lty = 2)
# Confidence Interval for Parameters
alpha = 0.05
infect.CIs = confint(infect.model, level = 1-alpha)
infect.CIs
 # Shapiro-Wilks Test for Infection Normality
ei = infect.model$residuals
infect.SWtest = shapiro.test(ei)
infect.SWtest
 # Fligner-Killeen test for Infection Homoscedasticity
SENIC$ei1 = infect.model$residuals
Group = rep("Lower",nrow(SENIC1))
Group[SENIC1$length < median(SENIC1$length)] = "Upper"
Group = as.factor(Group)
SENIC1$Group = Group
infect.FKtest= fligner.test(SENIC1$ei1, SENIC1$Group)
infect.FKtest
```