Project 1

Oliver Hannaoui & Will Khouri, STA 108: Dr. JoAnna C. Whitener, section B01 & section B02

10/24/2019

I. Introduction

For our project we our trying to predict the probability of infections for patients based on the numbers of days they stay in a hostipal. More specifically we are triyng to predict if there is a linear relationship between the number of days they are in the hospital and their probability of infection. To do so we will find the best fit model and calculate the p-value of our test statistic for the hypothesis test for a linear relationship between the two variables.

Determining if the two previously discussed variables in our model have a linear relationship allows to predict the dependent variable, probability of infection, with our indepdendent variable, the amount of days staying in the hospital. In addition, we will be able to predict the probability of infection for people who stay in a hospital for 10 and 40 days as well as predicting the average infection risk for a patients who stay 20 days.

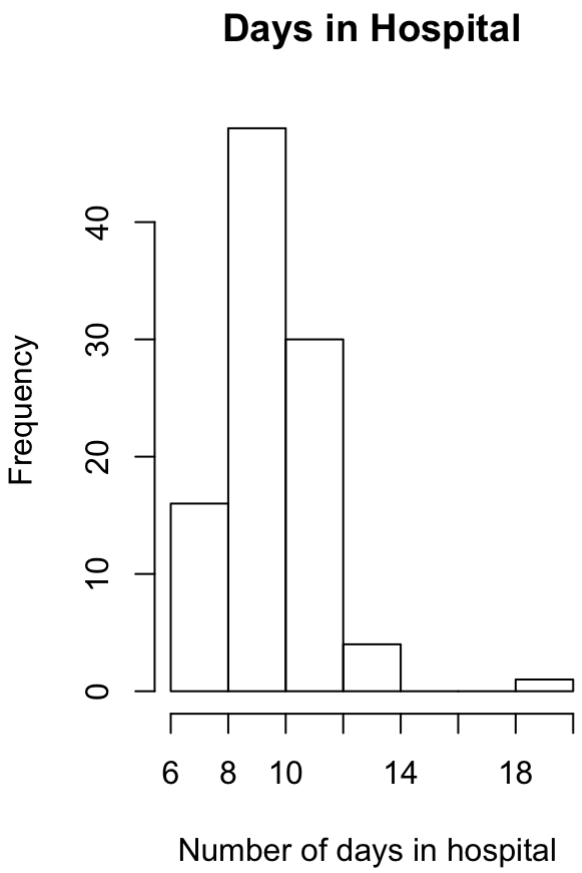
Moreover, we will be using the simple linear regression approach for this particular question and all the model fit, and assumptions that go along with it.

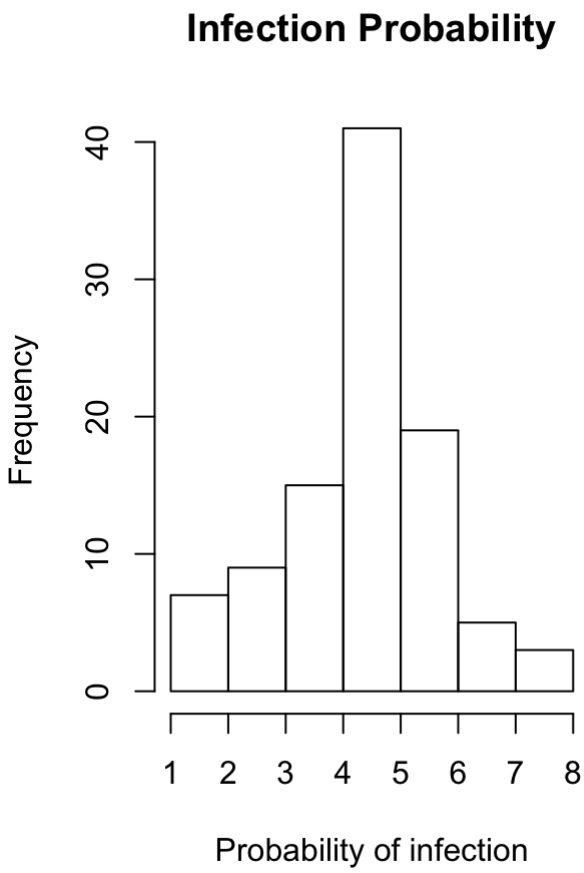
II. Summary

To begin, we will begin by discussing characteristics of our data to make sense of it.

Based on the histogram plot of the values for the probability of infection we can see that our data is distributed approximately normal. Furthermore, the distribtuion of the number of days patients spent in the hospital appears to be approximately skewed right with one noticeable outlier with a number of days spent in the hospital of over

18 days. For the most part the number of days spent in the hospital is most commonly between 6-10 days with patients also staying more than 10 days as well.





Additionally, we calculated the mean, standard deviation, median, and range for the data for number of days in the hospital: The mean number of days in the hospital for the patients is 9.571919.

```
## [1] 9.571919
```

The standard deviation in the context of this problem signifies that the number of days a patient stays in the hospital is typically within approximately 1.7 days of the mean number of days patients stay in the hospital.

```
## [1] 1.737769
```

The median number of days patients spent in the hospital, that is the amount of days 50% of patients stayed less and 50% stayed longer than in the hospital is 9.35 days.

```
## [1] 9.35
```

The range for the number of days patients stayed in the hospital ranges from 6.7 days to 19.56 days.

```
## [1]  6.70 19.56
```

In addition, we calculated the same statistics for the data for the probability of infection: The mean probability of infection for the patients is 4.382828.

```
## [1] 4.382828
```

The standard deviation in the context of this problem signifies that the typical observed value for the probability of infection for the patients is within approximately 1.3 of the mean number of probability of infectoin for patients in the hospital.

```
## [1] 1.29055
```

The median number of the probability of infection, that is the probability of infection that is 50% above and 50% below the probability of infection for all patients in the dataset is 4.4

```
## [1] 4.4
```

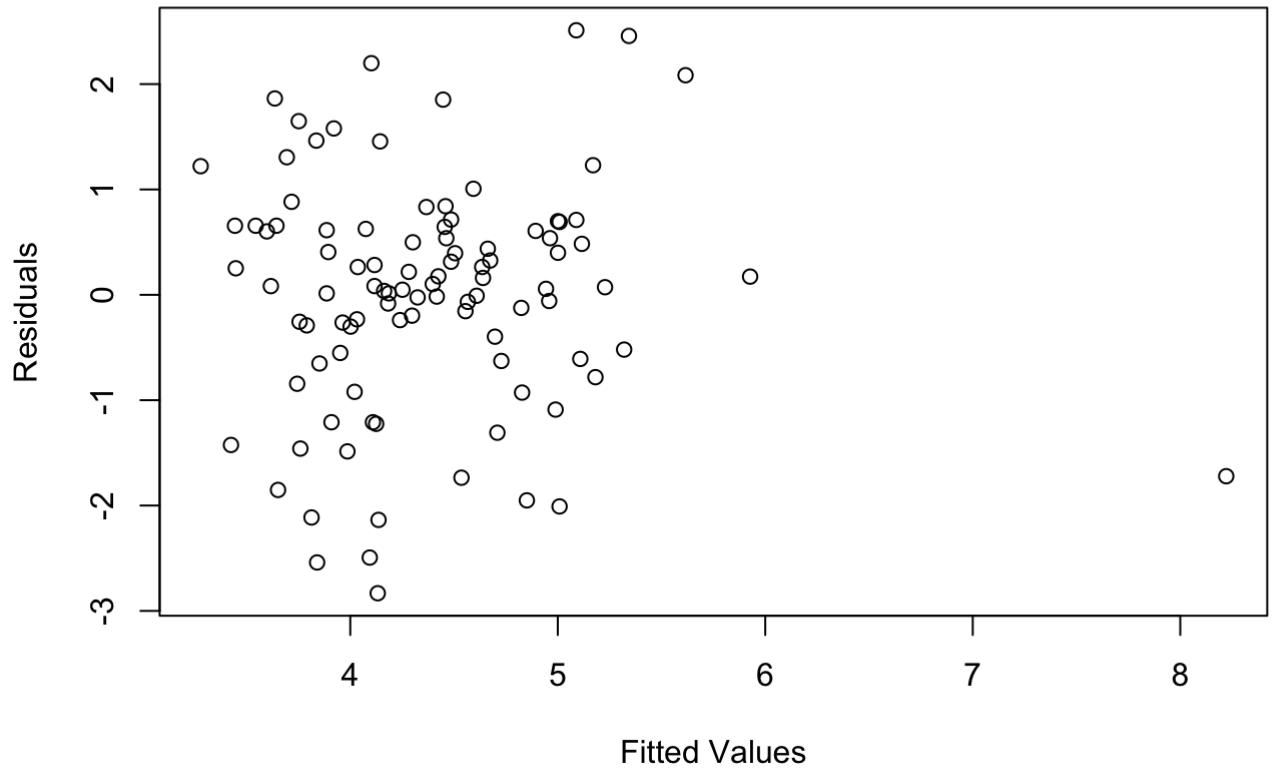
The range for the probability of infection for the patients of the hospital ranges from 1.3 to 7.8.

```
## [1] 1.3 7.8
```

III. Diagnostics

The overall model has some main assumtpions that need to be fullfileed. First off that both X, the number of days stayed in the hospital, and Y, the probability of infection, are independent. However since we did not take this random sample of data assume that X and Y are indepedent.

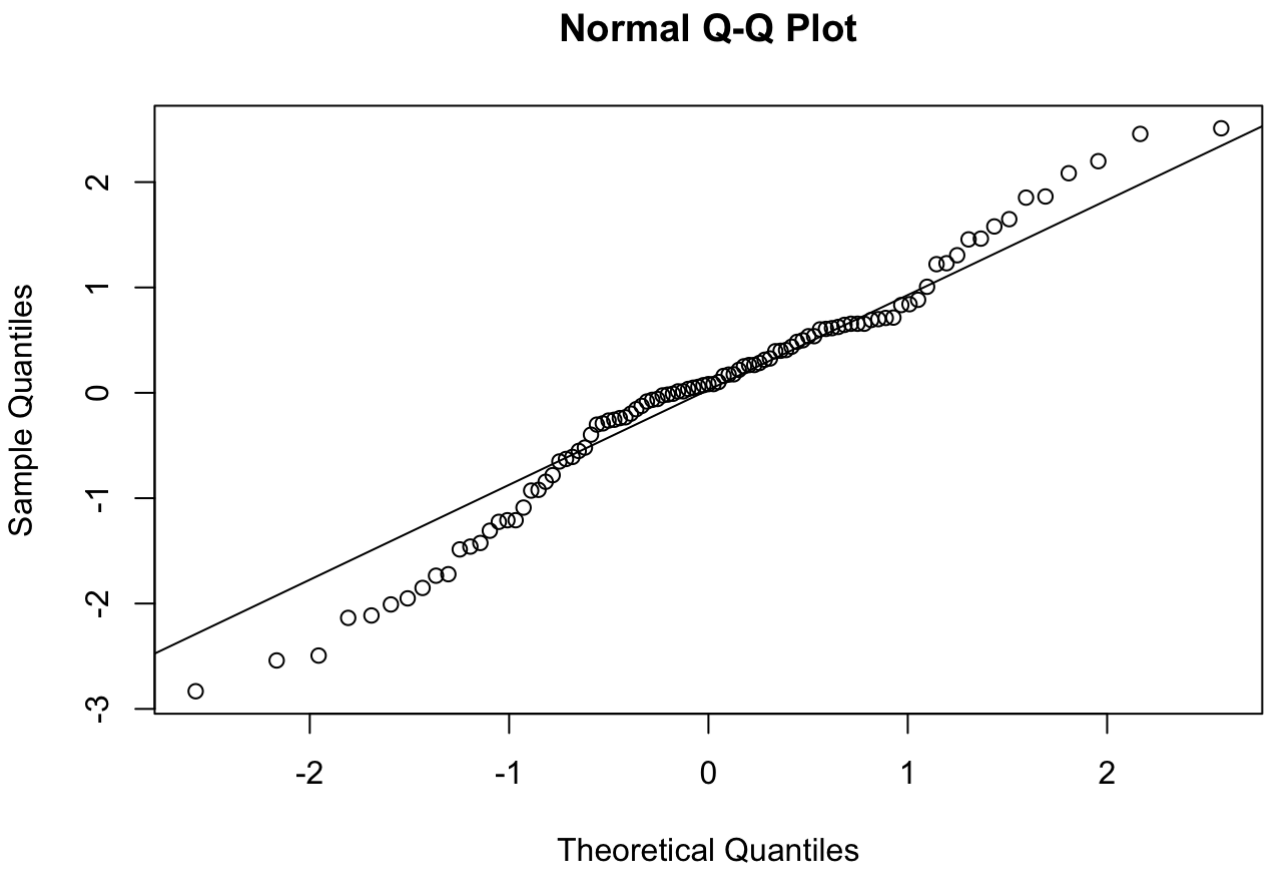
Next off the first assumption we will test is for constant variance of our data.



From the Non-constant variance plots that we graphed, you can see that the pattern, minus the one outlier, follows a

fairly constant rate, with no flairs or patterns visible. Therefore we conclude that the variances of the errors are constant for our data.

Finally, we will prove that our data's errors are distributed normal.



From the QQ plot the theoretical and sample quantiles follow a pattern along the line constant with normally distributed data, which would make us conclude that they are distrubted normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97556, p-value = 0.06198
```

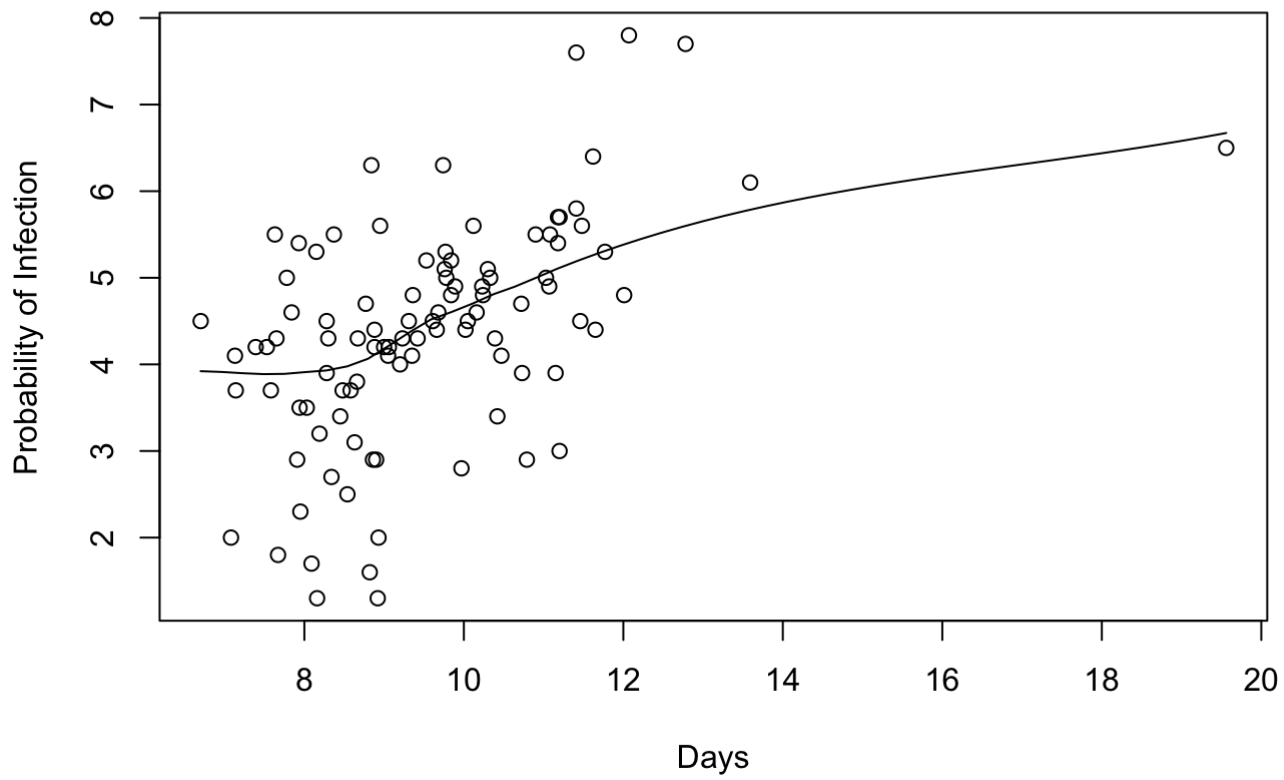
However we looked futher and copnducted a Shapiro-Wilkes test that produced a p-value of 0.06198 which when compared to our alpha value of .05 concludes once again that the errors of our data is distributed normalyl.

IV. Analysis

Here we use the lm function to build our linear model. We find that our intercept is 0.7041 and our slope is 0.3834.

```
##
## Call:
## lm(formula = Infect ~ Days, data = hospital)
##
## Coefficients:
## (Intercept)         Days
##      0.7041       0.3843
```

Using scatter.smooth function to plot the data points along with adding a fitted line for the data.



We use the summary function to obtain the summary statistics of our model.

```
##
## Call:
## lm(formula = Infect ~ Days, data = hospital)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.83228 -0.58006  0.08272  0.63513  2.51075
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.70409    0.62759   1.122    0.265
## Days         0.38433    0.06452   5.957 4.13e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.11 on 97 degrees of freedom
## Multiple R-squared:  0.2678, Adjusted R-squared:  0.2603
## F-statistic: 35.48 on 1 and 97 DF,  p-value: 4.134e-08
```

VI. Interpretation of our Analysis.

To begin we obtained our slope and intercept by finding our model fit. Using the lm function we firstly discovered our intercept, beta zero, 0.7041.

Our intercept describes the expectred probability of infection if we were to stay in the hospital for zero days.

Next we found our slope, beta one, 0.3843. For an increase of one day staying in the hospital we would expect our probability of infection to rise by 0.3843.

We then used the summary function to obtain a variety of important statistics. We obtained p-Values, the model p-Value and the p-Value for our predictor variables beta zero and beta one. Since we obtained p-Values there is also a null and alternate hypothesis associated with them. The Null Hypothesis for our slope is that our beta one equals zero, there's no linear relationship between days stayed in the hospital and probability of infection. The Alternate Hypothesis for our slope is that beta one does not equal zero, there is a linear relationship.

We also obtained a t-value for our beta one. The t-value tells us how far our slope would be from the expected slope if the Null Hypothesis were true. The t-value obtained for our slope is 5.957.

Since our t-value is high, our p-value is significantly lower than our significance level of 0.05. We safely reject our Null Hypothesis that there's no linear relationship between the number of days stayed in the hospital and the probability of infection.

In addition, the model p-value is also significantly below our significance level so we can conclude that our model is statistically significant. The fact that our model is statistically significant allows us to make predictions or estimate the probability of infection with the number of days stayed in the hospital.

We did not construct a confidence interval for the slope of our model since our p-values were very low thanks to our summary statistics and we could conclude our model was statistically significant.

VI. Prediction Results

• Predict the infection risk for a patient who stayed 10 days.

```
##        fit      lwr      upr
## 1 4.547351 2.332586 6.762116
```

Using our model we predict the infection risk for a patient who stayed 10 days to be 4.547351. We are 95% confident that the predicted value of infection risk probabilty for a patient who stays 20 days is between 2.332586 and 6.762116.

• Predict the average infection risk for a patients who stay 20 days.

```
##        fit      lwr      upr
## 1 8.390614 7.036983 9.744244
```

Using our model we predict the average infection risk for a patient who stayed 20 days to be 8.390614 We are 95% confident that the average value infection risk for a patient who stays 20 days is between 7.036983 and 9.74424. Although we calculated these statistics we would advise against using them as 20 days is above our range of data and therefore making inferences on this data would be extrapolation.

• Predict the infection risk for a patient who stayed 40 days.

```
##        fit      lwr      upr
## 1 16.07714 12.17429 19.97999
```

Using our model we predict the infection risk for a patient who stayed 40 days to be 16.07714. We are 95% confident that the predicted value of infection risk for a patient who stays 40 days is between 12.17429 and 19.97999. Although we calculated a prediction and prediction interval for infection probability for patients who

stayed 40 days in the hospital we would also not advise this since 40 days is above our range of data collected.

VII. Conclusion

In conclusion, we calculated 5 main statistics of our data for both the dependent and indepedent variables. This allowed us to get an idea of the description of the data, how spread out the data is, as well as the range of data collected.

We also fulfilled main assumptions in our Diagonistics. We assumed that our dependent and independent variables are independent since we did not take this random sample. Using the a non-constant variance plot we concluded that the variance of the errors are constant for our data. We proved our data's errors are distributed normally using our QQ plot in which the theoretical and sample quantiles follow a pattern along the line constant with the normally distributed data. We also conducted a Shapiro-Wilkes test that produced a p-value of 0.06198 which when compared to our alpha value of .05 concludes once again that the errors of our data are distributed normally.

Using the summary of our model that we created we determine there is a linear relationship between both the dependent and the independent variables. We ran a hypothesis test for the linear relationship and due to the p-value being signficanlty below the .05 significance level, we determined that we would reject the null hypothesis that there is no linear relationship between the amount of days spent in the hospital and the probability of infection.

To close, due to the outlier we aknowledge that the the line would be a better fit for our data if we omitted it and the p-value for our hypothesis test would be even smaller, also making our model more statistically significant overall. Removing the outlier would also narrow our range of values and even lower the standard deviation of our model, as well as alter the mean of the data for number of days stayed in the hospital; our independent variable.

We also built two prediction intervals and one confidence interval for three seperate values of our indepndent variable. We also made predictions of those three values based off our model.

We advised against making inferences on data from values outside the scope of our sampled data to avoid extrapolition in light of the fact that we still made the calculations.

Appendix:

```
# Loading the data
hospital <- read.csv("~/Desktop/hospital.csv")
par(mfrow=c(1, 2)) # Displaying both plots adjacently
hist(hospital$Days, main = 'Days in Hospital',xlab = "Number of days in hospital ", ylab
= 'Frequency') # Hospital Days  Histogram
hist(hospital$Infect, main = "Infection Probability" ,  xlab = "Probability of infectio
n", ylab='Frequency') #Infection Prob. Histogram
mean(hospital$Days) # mean of Days in Hospital
sd(hospital$Days) # standard deviation of Days in Hospital
median(hospital$Days) # median of Days in Hospital
range(hospital$Days) # range of Days in Hospital
mean(hospital$Infect) # mean of Infection probability
sd(hospital$Infect) # standard deviation of Infection probability
median(hospital$Infect) # median of Infection probability

range(hospital$Infect) # range of Infection probability
LM = lm(Infect ~ Days, data = hospital) # creating the model
res = LM$residuals # storing residuals from model in res variable
plot(LM$fitted.values, LM$residuals, xlab = "Fitted Values", ylab = "Residuals") # check
ing for constant variance
qqline(LM$res) # use res for QQ Plot

qqnorm(LM$res) # creating normal Q-Q Plot
qqline(LM$res) # Q-Q line
shapiro.test(res) # Shapiro-Wilkes test
LM <- lm(Infect ~ Days, data=hospital) # creating the linear model
LM
scatter.smooth(x=hospital$Days,xlab= 'Days',ylab='Probability of Infection', y=hospital
$Infect) # plotting points with fitted line
summary(LM) # summary statistics of our model
pred.value <- data.frame(Days=10)
predict(LM, newdata = pred.value, interval = "prediction") # constructing our prediction
interval
pred.value <- data.frame(Days=20)
predict(LM, newdata = pred.value, interval = "confidence")
# 95% default CI
pred.value <- data.frame(Days=40)
predict(LM, newdata = pred.value, interval = "confidence")
# 95% prediction interval for 40 days
```