Project 1
A Study on the Efficacy of Nosocomial Infection Control – A closer look at the response and explanatory variables

Statistics 511 – Applied Regression Analysis

February 28th 2020

Efe Umukoro | umukoroe20@apu.eud
Professor: Millie Mao

Table of Contents

Part 1 Introduction……………………………………………..3

Part 2 Interpretation and Parameter Inference……………………………………………………………………….…..……3-4

2.1 Regression Equation………………………………………………………………………………………….….…..3

2.2 Interpretation of regression coefficient……………………………………………………………………...…3

2.3 Scatterplot…………………………………………………………..…………………………………………….3

2.4 Significance of slope coefficient …………………………………………………………………………………...4

2.5 95% Confidence Interval of slope……………………………… ………………………………..…………………….4

Part 3 Point and Interval Estimation……………………………………………..……………………………………..4 -5

3.1 Point estimation…………………………………………………………………………………………….…..4

3.2 Interval estimation………………………………………………………………………………...…..4 and 5

3.2.1 Confidence Interval…………………………………………………………………………..………….4

3.2.2 Prediction Interval………………………………………………………………………………..……5

Part 4 Diagnostics……………………………………………………………………………………………….…... 5 - 7

4.1 Normality Assumption Plots……………………………………………..……...…………..…………………..5

4.1.1 Histogram……………………….………………………………………..………………………………5

4.1.2 Boxplot………………………….………………………………………..………………………………5

4.2 Normality Assumption Test……………………..……………………………………..………………………………..5

4.3 Equal Variance Assumption Plot…………………..……………………………………..………………..……….6

4.3.1 Residual Vs. Predictor……………….…………………………………….……………………………..6

4.3.2 Standardized Residual Vs. Predictor………………………………………………………………………..6

4.5 Equal Variance Assumption Test……………………………………………………………………………….....6

4.6 Lack of Fitness Test……………………………………………………………………………………………….7

4.7 Omitted Variable Age……………………………………………………………………………………...…7

5 Conclusion………7

1. Introduction

The data was provided in the SENIC.Rdata file. The data file contains 113 observations and 12 variables. The 12 variables are ID, LOS, AGE, INFRISK, RCR, RCXR, BEDS, MEDSCH, REGION, AVECENSUS, NURSES and AVAIL. The variables of analysis defined the response variable (Y) to be LOS which measures the length of stay in a hospital (in days) while the Explanatory variable (x) is INFRISK or the infection risk which measures the average estimated probability of acquiring infection in a hospital.

2. Interpretation and Parameter Inference

2.1 Estimated Linear Regression Function

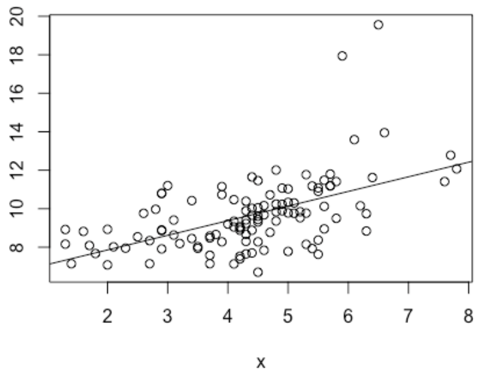
The estimated regression model is defined as

$$\hat{Y} = 6.3368 + 0.7604x_i$$

2.2 Interpretation of the Regression Coefficient

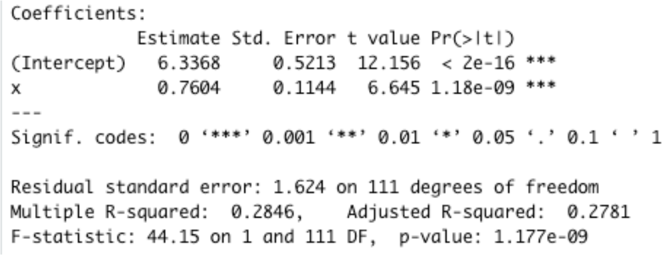
The estimated regression model defines β_0 as the y-intercept and β_1 as the slope. The estimated regression model shows that $\beta_0 = 6.3368$. Further interpretation tells us that when x =0, $\beta_0 = 6.3368$ meaning that when the infection risk is 0, the length of hospital is on average, 6.3368 days. We can further articulate $\beta_1 = 0.7604$ to mean that 1% increase of infection risk results in an estimated probability of 0.7604 days of acquiring infection in a hospital.

2.3 Scatterplot



The scatterplot measures the relationship between the explanatory variable, INFRISK and the response variable, LOS. Since the regression lines is not very steep, it is fair to assume that a weak positive linear relationship exists between the variables INFRISK and LOS. Further observation demonstrates that the infection risk between 5.5 and 7 days shows apparent outliers in the data.

2.4 Coefficient of Determination



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.3368     0.5213  12.156  < 2e-16 ***
x             0.7604     0.1144   6.645 1.18e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.624 on 111 degrees of freedom
Multiple R-squared:  0.2846,	Adjusted R-squared:  0.2781 
F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

The coefficient of determination, $r^2 = 0.2846$, which means that there is a 28.46% variation in the Y variable, LOS, explained by the X variable, INFRISK.

2.5 Significance of Slope Coefficient

The significance of the slope coefficient can be determined through the use of a two-sided hypothesis test:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

The test statistic:

$$t = \frac{b_1}{s(b_1)} = 6.645$$

The decision rule follows

p-value=1.18e-09 < 0.05

Since the p-value is less than the critical value, we reject the null hypotheses and accept the alternative. The slope of the coefficient is significant.

2.6 Confidence Interval of Slope

We are 95% confident that the slope parameter falls between interval [0.5336442, 0.9871976].

3. Point and Interval Estimation

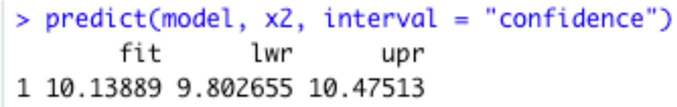
3.1 Point Estimation at x =5

$$\hat{Y} = 6.3368 + 0.7604(5) = 10.1388$$

The fitted value of LOS when infection risk is 5 percent is 10.1388.

3.2 Interval Estimation

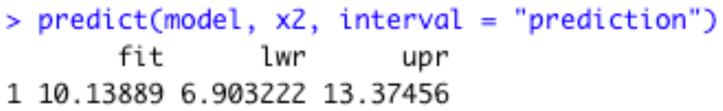
3.2.1 Confidence Interval



```
> predict(model, x2, interval = "confidence")
       fit      lwr      upr
1 10.13889 9.802655 10.47513
```

We are 95% confident that the mean LOS at INFRISK of 5 percent falls between 9.802655 and 10.47513.

3.2.2 Prediction Interval



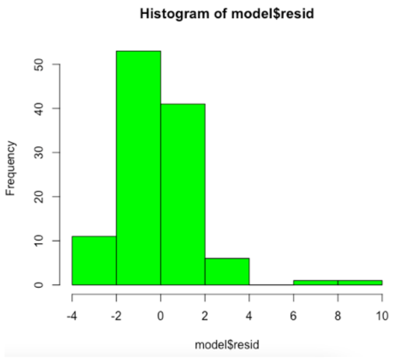
```
> predict(model, x2, interval = "prediction")
       fit      lwr      upr
1 10.13889 6.903222 13.37456
```

We are 95% confident that LOS at INFRISK of 5 percent falls between 6.903222 and 13.37456.

4. Diagnostics

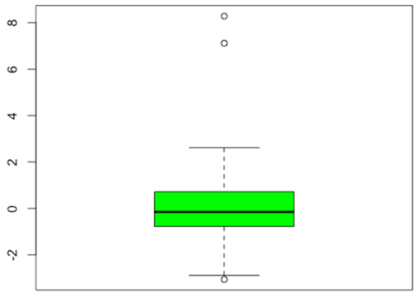
4.1 Normality Assumption Plots

4.1.1 Histogram



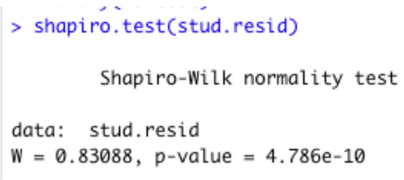
The histogram is skewed right with apparent outliers between the values 6 to 10. The normality assumption has not been satisfied.

4.1.2 Boxplot



The boxplot depicts the lowest value below -2 and the highest value at 3. The IQR shows a median of 0, Q1 of -1 and Q3 of about 1. We also see outliers between 6 and 10. The normality assumption has not been satisfied.

4.2 Normality Assumption Test

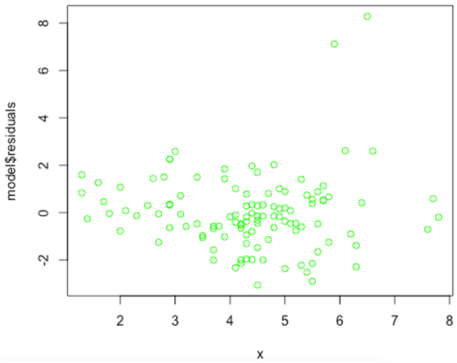


```
> shapiro.test(stud.resid)

	Shapiro-Wilk normality test

data:  stud.resid
W = 0.83088, p-value = 4.786e-10
```

In utilizing the Shapiro-Wilk normality test, the p-value is 4.786e-10. Since the p-value is significantly low, we would reject the null hypothesis. Since 4.786e-10 < 0.05, this further establishes that the data significantly deviates from the normal distribution.

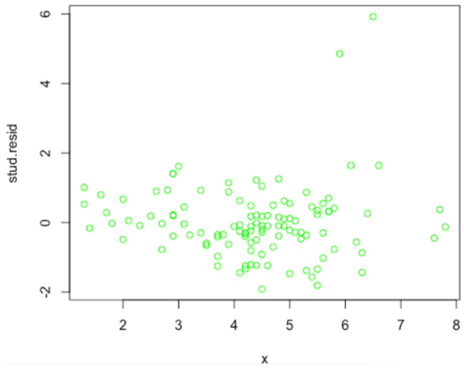
4.3 Equal Variance Assumption Plots

4.3.1 Residual Vs. Predictor



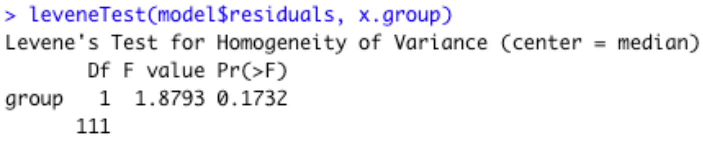
The observations appear to be scattered and do not demonstrate a pattern. It also highlights outliers between 6 and 7. The equal variance assumption appears to be satisfied.

4.3.2 Standardized Residual Vs. Predictor



The observations appear to be scattered and do not demonstrate a pattern. It also highlights outliers between 6 and 7. The equal variance assumption appears to be satisfied.

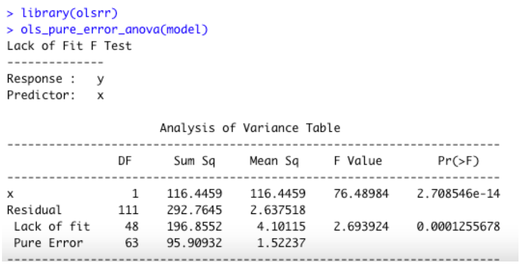
4.4 Equal Variance Assumption Test



```
> leveneTest(model$residuals, x.group)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.8793 0.1732
      111
```

In utilizing Levene's Test of Homogeneity of variance, the p-value is 0.1732. Since, 0.1732 > 0.05, we fail to reject the null hypothesis. This highlights the existence of homogeneity of variance and furthermore tells us that the assumption has been satisfied.

4.5 Lack of Fit Test



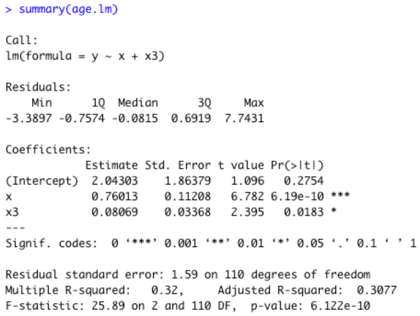
```
> library(olsrr)
> ols_pure_error_anova(model)
Lack of Fit F Test
-------------
Response :   y
Predictor:   x

                   Analysis of Variance Table
-------------------------------------------------------------------
               DF      Sum Sq     Mean Sq     F Value        Pr(>F)
-------------------------------------------------------------------
x               1    116.4459    116.4459    76.48984   2.708546e-14
Residual      111    292.7645    2.637518
 Lack of fit   48    196.8552     4.10115    2.693924   0.0001255678
 Pure Error    63    95.90932     1.52237
-------------------------------------------------------------------
```

The lack of Fit test highlights a p-value of 0.0001255678. We know that since $0.0001255678 < 0.05$, we reject the null hypothesis and accept the alternative. This means that overall, the model is in fact lacking the fitness to establish a relationship between infection risk and length of stay.

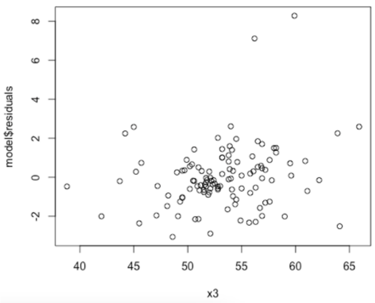
4.5 Omitted Predicted Variable

When including the variable AGE in the estimated regression model,

$$\hat{Y} = 6.3368 + 0.7604x_1 + 0.08069x_2$$



```
> summary(age.lm)

Call:
lm(formula = y ~ x + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3897 -0.7574 -0.0815  0.6919  7.7431 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.04303    1.86379   1.096   0.2754    
x            0.76013    0.11208   6.782 6.19e-10 ***
x3           0.08069    0.03368   2.395   0.0183 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 110 degrees of freedom
Multiple R-squared:   0.32,	Adjusted R-squared:  0.3077 
F-statistic: 25.89 on 2 and 110 DF,  p-value: 6.122e-10
```



The regression model includes age as a variable. The coefficient of determination, The coefficient of determination, $r^2 = 0.32$. This shows that a 32 % variation exists between LOS and both of the response variables AGE and INFRISK. This demonstrates that AGE is a potential omitted variable. We also see that the observations on the scatterplot show a slightly better relationship although values cluster between 50 to 55, we also see that the outliers are also present between 55 to 60. In visually checking the linearity assumption, most of the points are random thus the age variable is a good predictor for the model.

5. Conclusion

The goal of this project was to determine whether the model that contains the response variable INFRISK is an appropriate model. We also go on to test whether the addition of the response variable AGE will help influence the model's appropriateness. Based on our statistical assessment, it is evident that AGE is a potential omitted variable, and that the model's efficacy seems to improve, through the visual observations of its graph. It is clear that based on the linear model, the model that combined both age and infection risk was the best. It is also evident however that the model genuinely establish a weak yet significant relationship or fitness between both the response and the explanatory variable. In order to fully establish an appropriate prediction, it seems that other variables and factors would need to be taken into account. I propose the use of the variables BEDS, REGION and AVECENSUS to aid us in establishing a more accurate prediction model.