A Regression Analysis to Study the Relationship between Risk of Infection and The Average Length of Stay in Hospitals

**Azusa Pacific University**

**STAT 512, Prof. Millie**

Rumil Legaspi, [Rumil.legaspi@gmail.com](mailto:Rumil.legaspi@gmail.com)

Mei Leng Iao, Lenginnet@gmail.com

28 February 2021

Table of Contents

[Introduction 2](#_Toc65351565)

[Part 1: Interpretation and Parameter Inference 2](#_Toc65351566)

[Estimated Linear Regression Function 2](#_Toc65351567)

[Scatterplots with Regression Line 2](#_Toc65351568)

[Hypothesis Testing on the Slope Coefficient 3](#_Toc65351569)

[Testing Using the p-value 3](#_Toc65351570)

[Finding the 95% Confidence Interval of the Slope 4](#_Toc65351571)

[Part 2: Point and interval estimation 4](#_Toc65351572)

[Confidence Interval Interpretation when INFRISK = 5 4](#_Toc65351573)

[Constructing a Prediction Interval 4](#_Toc65351574)

[Prediction Interval Interpretation 4](#_Toc65351576)

[Part 3: Diagnostics 5](#_Toc65351577)

[Regression Model Assumptions 5](#_Toc65351578)

[Testing Normality 5](#_Toc65351579)

[Plotting to Check for Normality and Equal Variance Assumptions 5](#_Toc65351580)

[Boxplot and Histogram Interpretation 6](#_Toc65351581)

[Stating our Hypothesis 6](#_Toc65351582)

[Testing our Hypothesis 6](#_Toc65351583)

[Interpretation of Normality Tests 6](#_Toc65351584)

[Testing Equal Variance Assumptions 7](#_Toc65351585)

[Hypothesis: 7](#_Toc65351586)

[Testing our Hypothesis by plotting and Breusch-Pagan test 7](#_Toc65351587)

[Breusch-Pagan test 7](#_Toc65351588)

[Checking for omitted predictors 8](#_Toc65351589)

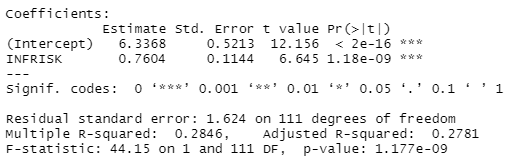
# Introduction

We are conducting a simple linear regression model using the SENIC dataset containing 113 observations, to analyze the relationship between the explanatory variable, infection of risk (INFRISK) and the response variable, length of stay (LOS).

# Part 1: Interpretation and Parameter Inference

## Estimated Linear Regression Function

From summarizing our linear regression model we can see:

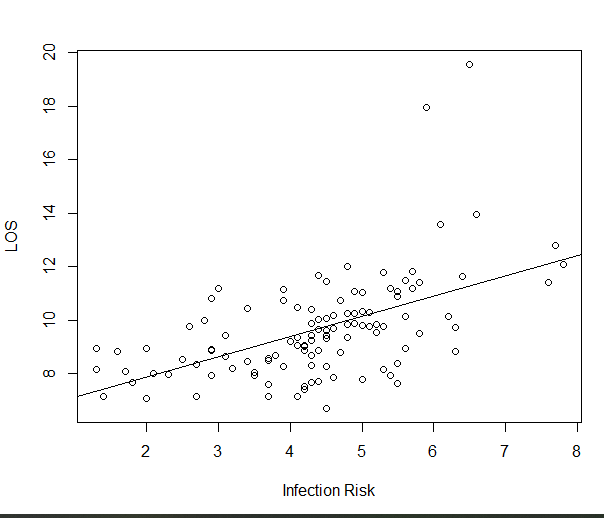


**= 6.3368 *(intercept)***

**= 0.7604 *(slope)***

and the **estimated regression equation** to be:

## Scatterplots with Regression Line



**From our model we derive that the intercept, = 6.3368**:

This indicates where our response output lies when there is no input or when . In other words, when risk of infection (explanatory variable) is at 0, the average length of stay of patients in a hospital is roughly 6 days.

Analyzing the intercept on its own might be confusing and at times misleading. In understanding the context of our data, we can see that despite patients having an average estimated probability of acquiring an infection in a hospital be 0%, we know that this is impossible. Additionally, we know that it is possible for patients to be in the hospital for roughly 6 days for other medical reasons.

In other words, although a bit misleading at first glance, when risk of infection is close to zero and almost nonexistent, there is still some truth in a patient having a prolonged length of stay in a hospital.

**The Slope:** :

Indicates when the risk of infection increases by 1 unit, the average length of stay increases by 0.74 days. This can also be thought of as when the risk of infection increases by 1% the average length of stay in a hospital increases by about 18 hours.

**The R-Squared:**

The R squared indicates that the risk of infection (input variable) helps explain close to 28% of the variability in our response variable, average length of stay. In other words, **our model explains a small amount of the variability in our response variable, length of stay.**

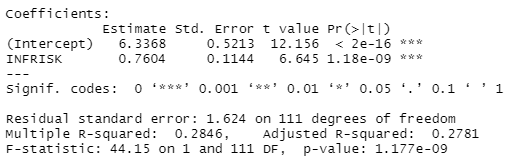
## Hypothesis Testing on the Slope Coefficient

α = 0.05 (Accepted error of 5%)

**Null Hypothesis**: : (slope is horizontal/ no relationship), in other words there is no linear relationship between risk of infection and length of stay

**Alternative Hypothesis**: : (slope exists/ relationship exists), there is linear relationship either positive or negative between risk of infection and length of stay.

### Testing Using the P-value



The slope indicates a positive relationship and the p-value (1.18e-09) is very close to 0 which is less than our , we then reject our null hypothesis and conclude with the alternative hypothesis that the slope coefficient and the linear relationship between our variables are both significant.

### Finding the 95% Confidence Interval of the Slope



#### Interpretation

This output reads that within our confidence interval of 95% from 2.5% (the lower limit of our interval) to 97.5% (the upper limit of our interval), our **intercept** and **slope** are both found within the listed intervals.

In this case, if this experiment is conducted many times, we are 95% confident that our interval captures the true population parameter of our slope , within the **interval 0.533 and 0.987** with an accepted error of 5%.

0 is not included in our interval. We are interested in this because if zero was included in our confidence interval then that would indicate (that there is a chance that) no change/linear relationship exists and would make risk of infection (INRFRISK) a bad predictor for length of stay (LOS). Thus, in this case, since 0 is not included, we can conclude that there is a significant linear relationship.

# Part 2: Point and Interval Estimation



## 95% Confidence Interval Interpretation when INFRISK = 5

The fitted value of the length of stay variable when the risk of infection is at 5% is **10.13889 days**.

The 95% confidence interval when risk of infection **=** 5% is **(9.802 to 10.475)**

We are 95% confident that when the infection risk = 5, the **true mean** of the length of stay (response variable) **will be within the interval of 9.8 to 10.5 days.**

## Constructing a Prediction Interval

We can use a prediction interval when trying to find where an individual observation will fall. We constructed a prediction interval given risk of infection is at 5 percent.

## 

## Prediction Interval Interpretation

From the results we are 95% confident that when a patient has a risk of infection at 5%, the predicted length of stay response will fall between 6.903 and 13.37 days or about 7 to 13 days.

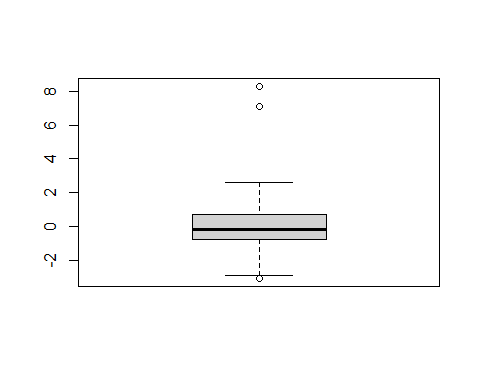
# Part 3: Diagnostics

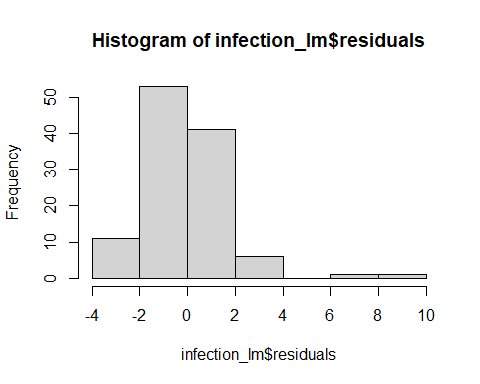
## Regression Model Assumptions

A simple linear regression model is appropriate when the linearity, independent, normality and equal error variance are satisfied for all our X values (homoskedasticity). In this report we will **only** be testing for **normality** and **homoskedasticity**.

## Testing Normality

### Plotting to Check for Normality and Equal Variance Assumptions with Boxplot & Histogram





## Boxplot and Histogram Interpretation

We can see that our boxplot is not symmetrical due to outliers, and our histogram shows our residuals as a being right skewed. Therefore, from these visualizations our assumption of normality is violated.

### Stating our Hypothesis

**Null Hypothesis**: : The data **IS** from a normal distribution

**Alternative Hypothesis**: : The data is **NOT** from a normal distribution

### Testing our Hypothesis

**To test these, we can use several normality tests:**

* Shapiro-Wilk normality test
* Shapiro-Francia normality test
* Anderson-Darling normality test

These tests focus mainly on the usage of regression residuals with a p-value as an output which is useful for hypothesis testing. **Our main goal is to see if our data truly follows a normal distribution.**

*#Shapiro-Wilk normality test*  
**shapiro.test**(infection\_lm**$**residuals)

Shapiro-Wilk normality test  
data: infection\_lm$residuals  
W = 0.87054, p-value = 1.699e-08

*#Shapiro-Francia normality test*  
nortest**::sf.test**(infection\_lm**$**residuals)

Shapiro-Francia normality test  
data: infection\_lm$residuals  
W = 0.86188, p-value = 7.85e-08

*#Anderson-Darling normality test*  
nortest**::ad.test**(infection\_lm**$**residuals)Anderson-Darling normality test  
data: infection\_lm$residuals  
A = 2.008, p-value = 3.823e-05

## Interpretation of Normality Tests

Looking at the results of these three tests we can see that the p-values are smaller than our alpha. Hence, we reject our NULL hypothesis and **there is a violation of our normality assumption.**

## 

## Testing Equal Variance Assumptions

### Hypothesis:

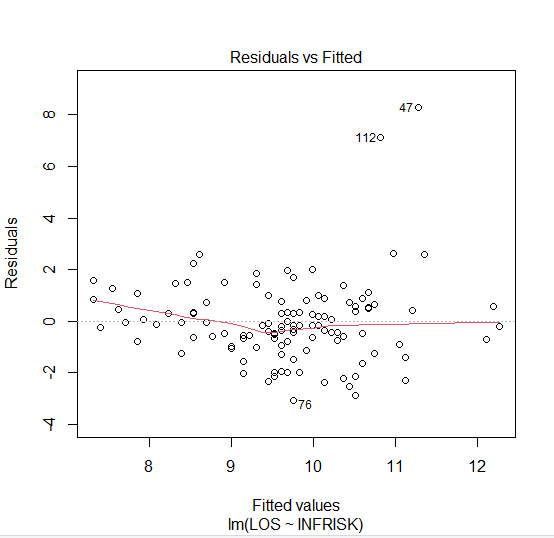
**Null Hypothesis**: : The error variances in the data **ARE** equal

**Alternative Hypothesis**: : The error variances in the data are **NOT** equal

### Testing our Hypothesis by plotting and Breusch-Pagan test

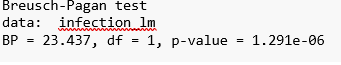
We don’t want to see fan shapes otherwise they violate equal variance assumption

we want equal and random spread of our scatterplots

Residuals vs predictor variable plot: 

Residuals vs Fitted values plot shows concentration of data in certain area, **thus equal variances of error (homoskedasticity) is violated.**

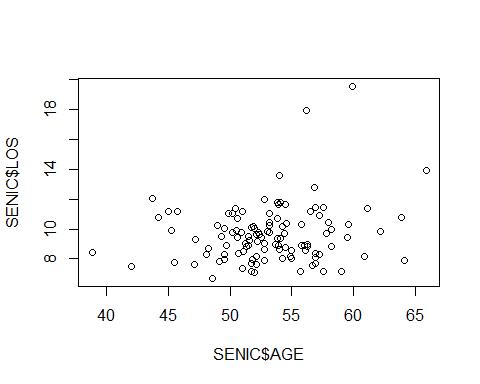
### Breusch-Pagan test



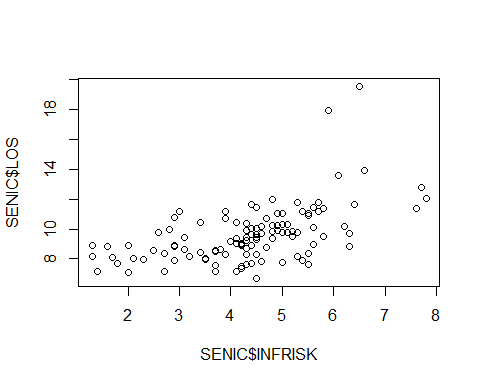
The Breusch-Pagan test results a low p-value (1.291e-06) therefore, we reject the null hypothesis and that there is an issue with our equal variance assumption

## Checking for Omitted Predictors

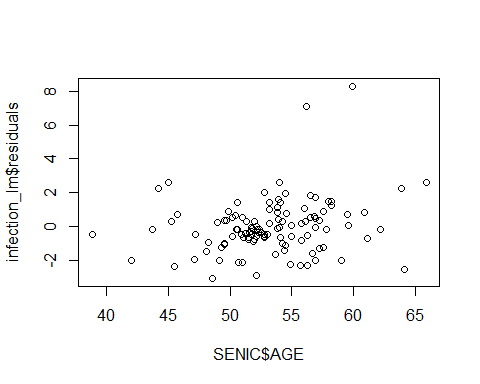
**Plotting of potentially omitted variable (Age) against the response variable (LOS)**



**Plotting first predictor risk of infection (INFRISK) against the response variable (LOS)**



**Plotting our Linear Model’s residuals against Age (potentially omitted variable)**



In the residuals vs. the potentially omitted variable (Age) the plots are randomly scattered and show no kind of relation between the residuals and Age.

Call:  
lm(formula = age ~ infection\_risk, data = Infection\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14.4284 -2.3346 -0.0338 2.9625 12.6600   
   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 53.216018 1.438494 36.994 <2e-16 \*\*\*  
infection\_risk 0.003637 0.315813 0.012 0.991   
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.482 on 111 degrees of freedom  
Multiple R-squared: 1.195e-06, Adjusted R-squared: -0.009008   
F-statistic: 0.0001326 on 1 and 111 DF, p-value: 0.9908

**qt**(0.975, 111)

1.981567

After constructing a linear model using Age as our new predictor and looking at our summary, we see that because the absolute value of our critical value is less than our T-value, we **fail to reject our NULL hypothesis**. Age is **NOT** a potentially omitted variable.

## Conclusion

Due to our model violating normality and homoskedasticity assumptions, remedial measures should be considered so that having a linear regression model is appropriate.