A Regression Analysis on The Efects of The Risk of Infection and on The Length of Stay in Hospitals

Rumil Legaspi, Rumil.legaspi@gmail.com Mei Leng Lao, Email

28 February 2021

Contents

Purpose

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

Our Data

Quick background on Dataset and variable

```
#Setting up our work environment
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Project 1/Project

#Loading in packages
library(nortest)
library(car)
library(lmtest)
library(formatR)

#Loading in the data
load(file = "SENIC.rdata")

Infection_data <- data.frame("SENIC.rdata")

#Defining and renaming our Explanatory(X) and Response(Y) variables
infection_risk = SENIC$INFRISK #X
length_of_stay = SENIC$LOS #Y
age = SENIC$AGE #Z
```

(delete this) some interpretations:

- Length of stay is explained by the average estimated probability of acquiring infection in hospital.
- As the risk of infection increases the average length of stay in the hospital also increases.

Part 1: Interpretation and Parameter Inference

Estimated Linear Regression Function

```
# Generating our Linear Model using lm() then summarizing
infection_lm = lm(length_of_stay ~ infection_risk, data = Infection_data)
summary(infection_lm)
```

```
##
## Call:
## lm(formula = length_of_stay ~ infection_risk, data = Infection_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.3368     0.5213  12.156  < 2e-16 ***
## infection_risk   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
```

From summarizing our Linear Regression model we can see:

$\beta_0 = \mathbf{6.3368}$ ***(intercept)***

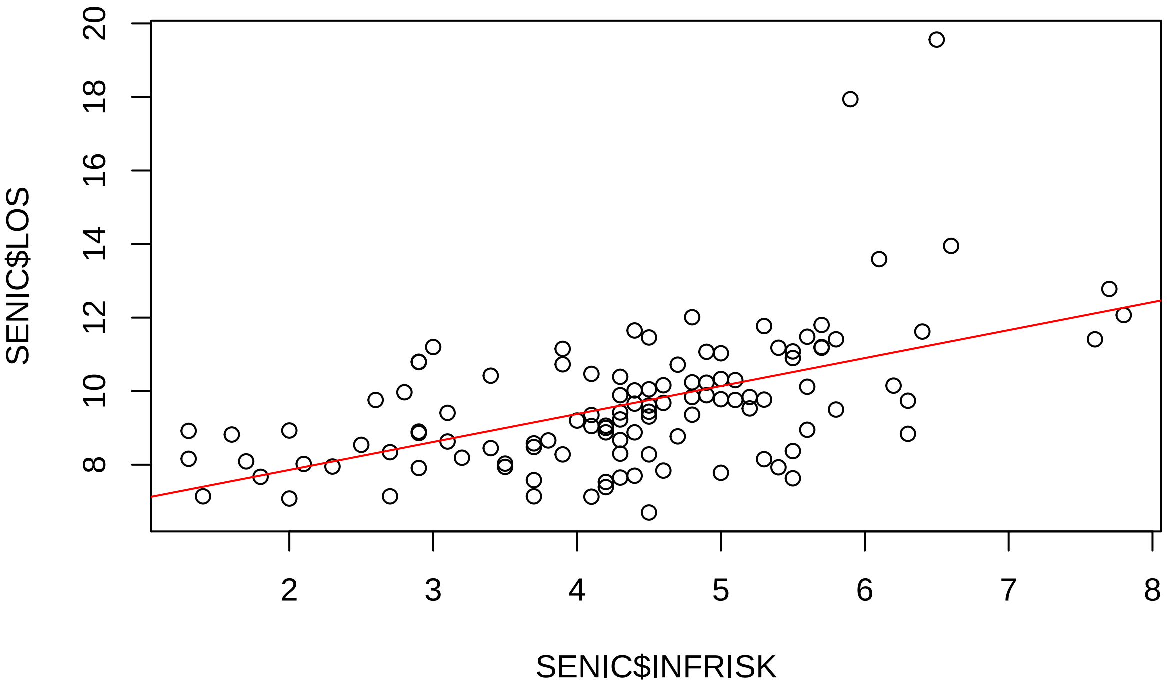
$\beta_1 = \mathbf{0.7604}$ ***(slope)***

and the estimated regression equation to be:

$\widehat{Y} = 6.3368 + 0.7604X$

Fitting on Scatterplots

```
plot(SENIC$INFRISK, SENIC$LOS)
abline(infection_lm, col = "red")
```



Interpretting Regression Coefficients & R^2

From our model we derive that our intercept, β_0 = 6.3368:

This indicates where our response output lies when there is no input or when X is 0. 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 truth in a patient having a prolonged length of stay in a hospital.

Our slope, $\hat{\beta}_1 = 0.7604$:

Indicates as 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.

Our $R^2 = 0.2846$:

***The R squared found at 0.2846 indicates that the risk of infection (input variable) helps explain close to 28% of the variability in our response variable, length of stay. In other words, **our model explains a small amount of the variable of length of stay (response variable)**.

Hypothesis Testing on our Slope to Test Significance

Our null hypothesis is that there is no linear relationship

Null Hypothesis: H_0: $\beta_1 = 0$ (slope is horizontal/ no relationship), in other words there is no linear relationship between risk of infection and length of stay

Alternative Hypothesis: H_1: $\beta_1 \neq 0$ (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 (1.177e-09) is very close to 0 which is less than our $\alpha = 0.05$, this indicates that we can reject the null hypothesis and conclude with the alternative hypothesis that the slope coefficient and the linear relationship are both significant.

Finding the 95% Confidence Interval of the Slope

```
#alpha at 0.05
alpha <- 0.05

#constructing our 95% confidence interval
confint(infection_lm, level = 1 - alpha)
```

```
##                    2.5 %    97.5 %
## (Intercept)    5.3038443 7.3697288
## infection_risk 0.5336442 0.9871976
```

Interpretation This output reads that within our confidence interval 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 we repeat this experiment many times, we are 95% confident that our interval captures the true population parameter of our slope β_1 will be between the interval 0.5336442 and 0.9871976 with and α (accepted error) of 5%.

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

Part 2: Point and interval estimation

Conducting 95% Confidence Interval when Length of Stay (Input Variable) is 5 for the Mean Length of Stay (Response Variable)

```r
#Creating a new single observation where risk is 5
new_infection_data <- data.frame(infection_risk = 5)

#Constructing our prediction interval
ci_infection_5 <- predict(infection_lm, new_infection_data,
                          interval = "confidence", level = 1 - alpha )

ci_infection_5
```

```
##        fit      lwr      upr
## 1 10.13889 9.802655 10.47513
```

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 ***.___

This 95% confidence interval when risk of infection **is at 5% is from 9.802 to 10.475.**

In other words, when the risk of infection is 5%, with 95% confidence we can expect our confidence interval to capture the **average(true mean)** of the length of stay (response variable) **which is roughly 9 to 10 and a half days.**

Constructing a Prediction Interval

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

```r
#Constructing prediction interval when INFRISK is 5
pi_infection_5 <- predict(infection_lm, new_infection_data,  interval = "prediction", level = 1 - alpha

pi_infection_5
```

```
##        fit      lwr      upr
## 1 10.13889 6.903222 13.37456
```

Prediction Interval Interpretation

From the results we can predict with 95% confidence that when a patient has a risk of infection at 5%, the length of stay will fall somewhere between 6.903 and 13.37 days or about 7 days to 13 days.

Part 3: Diagnostics

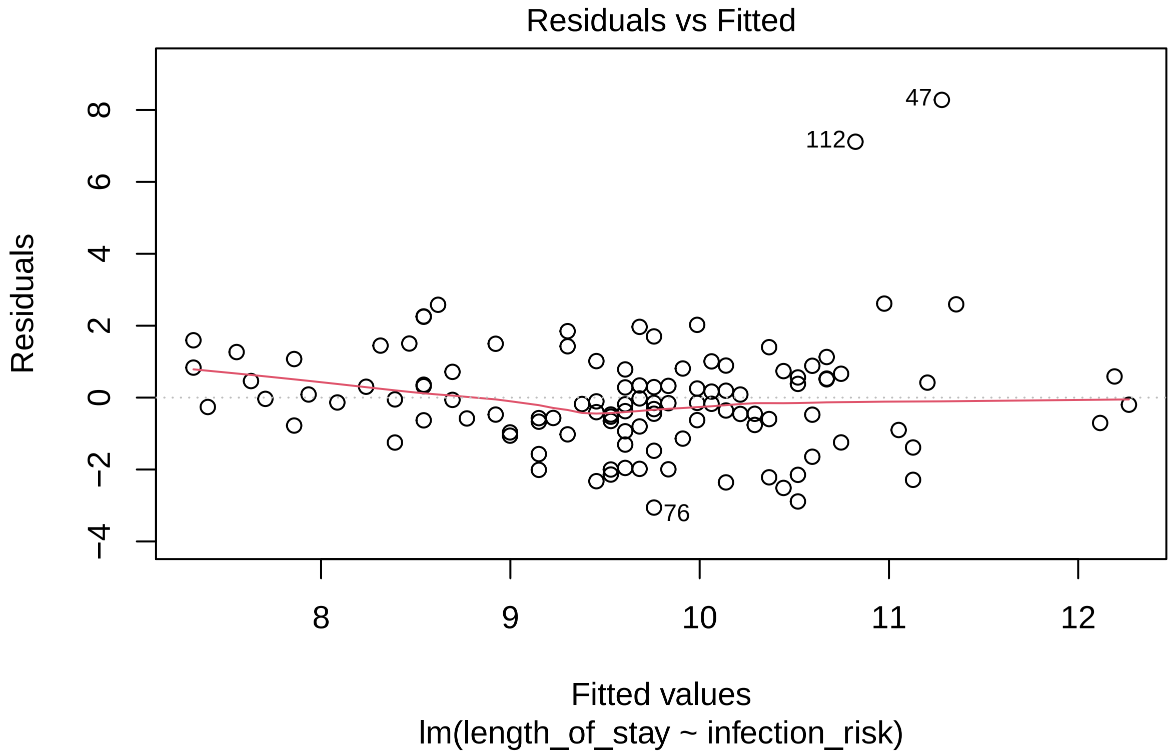
Our Assumptions

To ensure our model is still within the bounds of our made assumptions for a linear regression model lets plot them using different plotting methods.

lets recall our made assumptions for a linear regression model:

*** **L** inearity - **I** ndependence - **N** ormality of the errors - **E** qual error variance for all values of X (homoskedasticity)

```
#plotting scatterplots to check assumption
plot(infection_lm , which = c(1))
```



### ***Testing Linearity

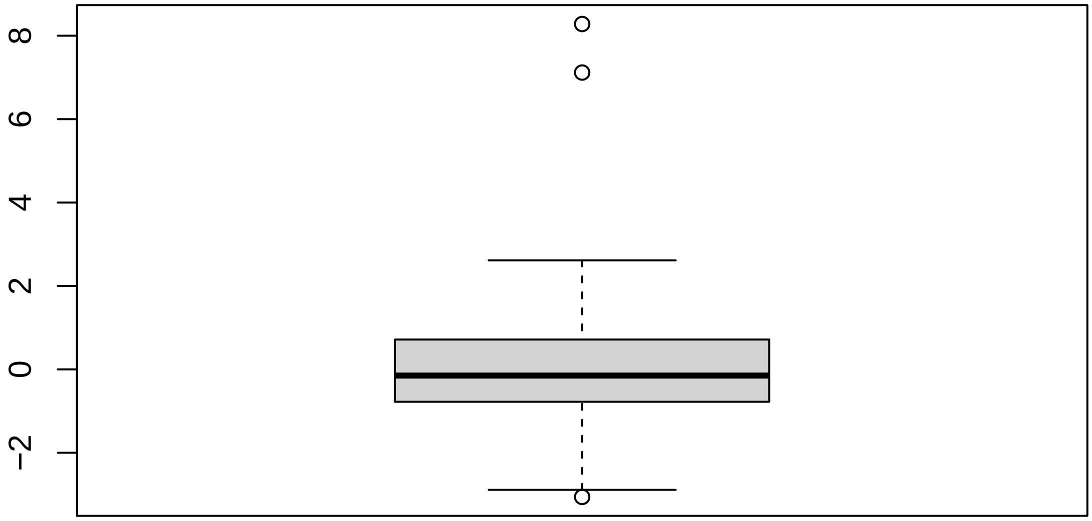
We cannot see a clear violation of linearity assumption in our **residual vs fitted plot** since we do not see a systematic pattern.

```
#Checking to independence

#install.packages("MASS")
#library(MASS)
#infection.resid = studres(infection_lm)

#Studentized residuals vs. predictor
#plot(SENIC$infection_risk, infection.resid)

#Residuals vs. Order
#data.order = c(1:25)
#plot(data.order, )
```

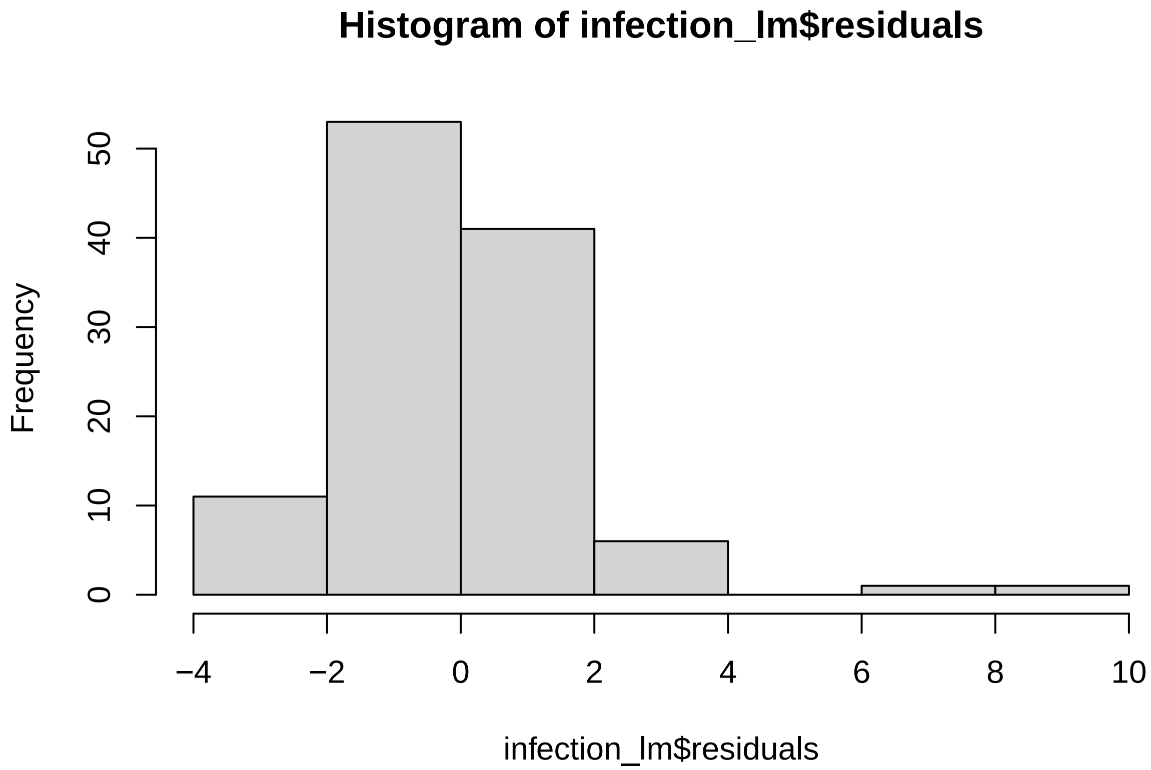
Testing Normality

Plotting to Check for Normality and Equal Variance Assumptions

```
#Plotting a boxplot and #histogram
boxplot(infection_lm$residuals)
```



```
hist(infection_lm$residuals)
```



Boxplot and Histogram Interpretation

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

Stating our Hypothesis

Null Hypothesis: H_0: The data **is** from a normal distribution

Alternative Hypothesis: H_1: The data is **NOT** from a normal distribution

Testing our Hypothesis

To test these we can use several normality tests using…

- 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. **Are main goal is to see if our data truly follows a normal distribution.**

```r
#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
```

```r
#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
```

```r
#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. Therefore we reject our NULL hypothesis and that **there is an issue with a violation of our normality assumption.**

Plotting to Check for Equal Variance Assumptions

Testing Equal Variance Assumptions

Stating our Hypothesis

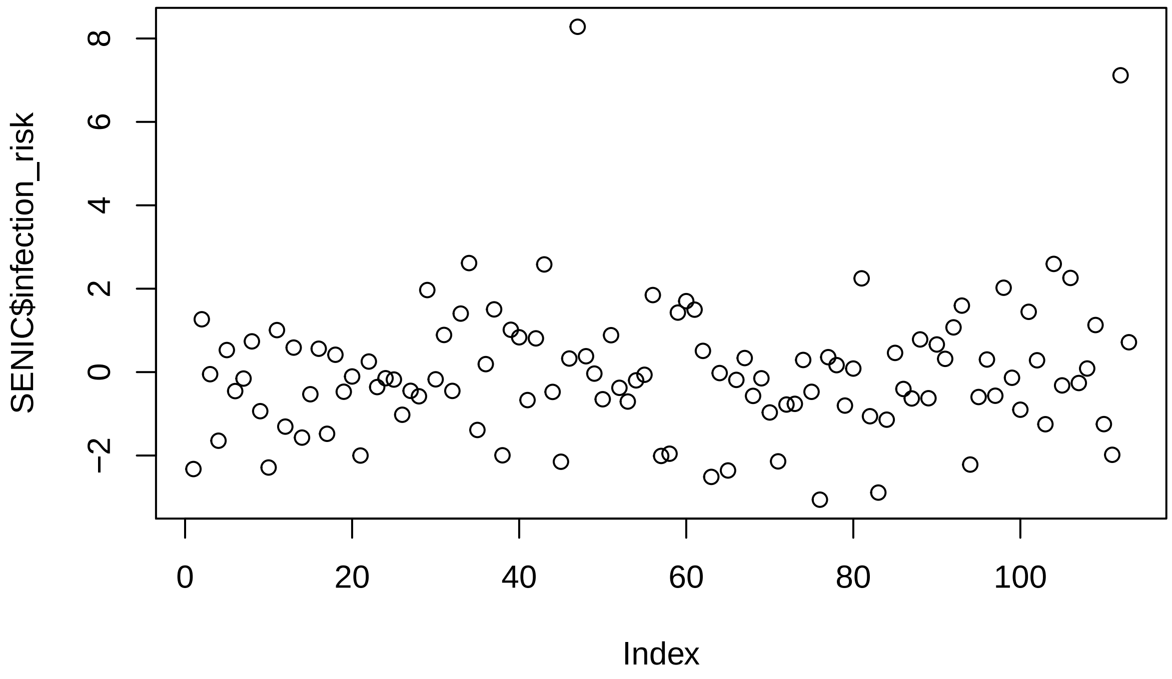
Null Hypothesis: H_0: The variances in the data **is** equal **Alternative Hypothesis**: H_1: The variances in the data are **NOT** equal

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

We dont want to see fan shapes otherwise they violate equal variance assumption

we want equal and random spread of our scatterplots

```r
#Residuals vs predictor variable
plot(infection_lm$residuals, SENIC$infection_risk)
```



```r
#Conducting Levene Test splitting into two groups
#bf.test(infection_lm, data = SENIC)

lmtest::bptest(infection_lm, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  infection_lm
## BP = 23.437, df = 1, p-value = 1.291e-06
```

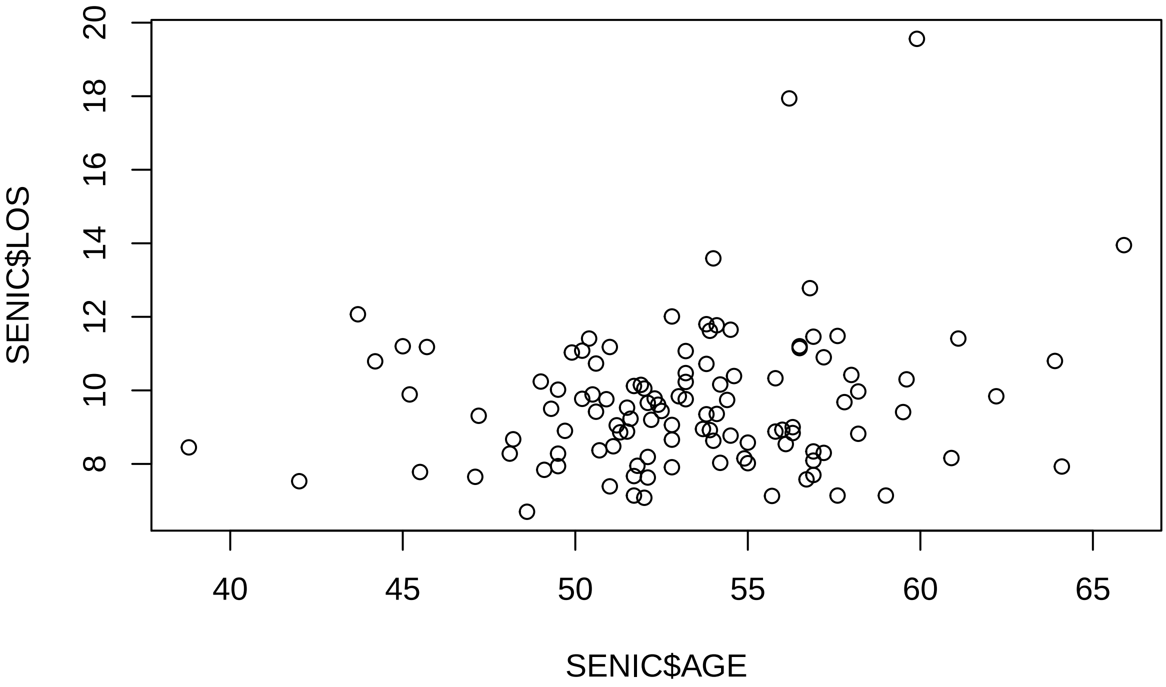
Residuals vs predictor variable plot indicates equal spread therefore equal variances of error (homoskedasticity) is not violated.

Breusch-Pagan test

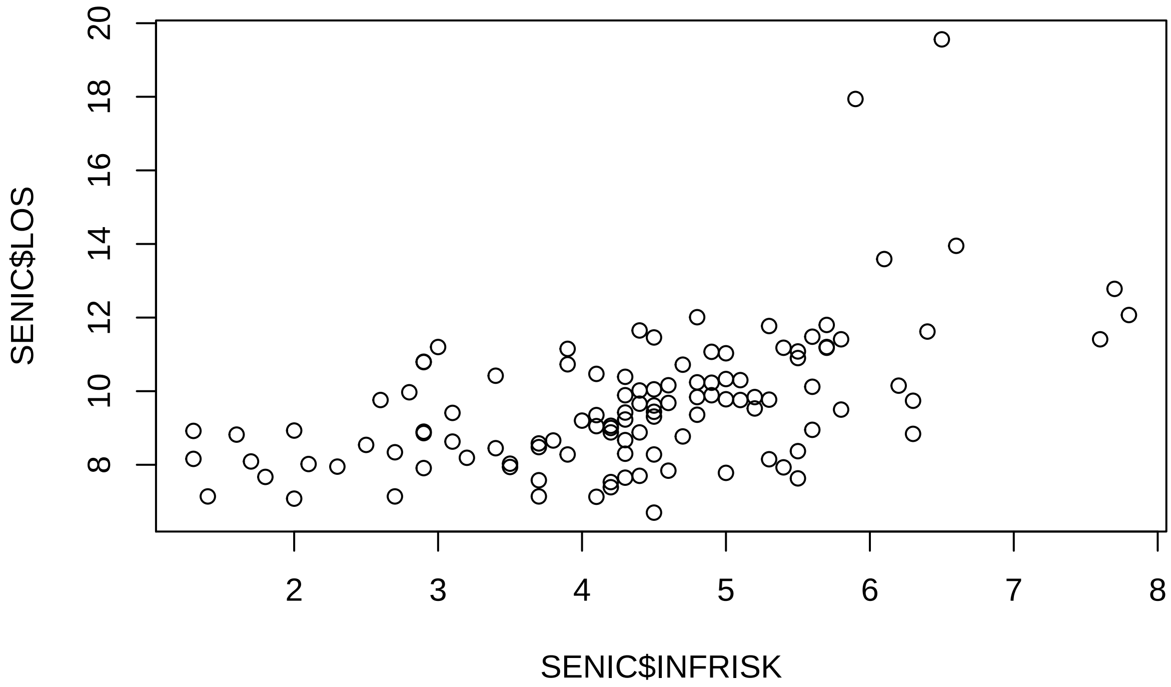
***Also the Breusch-Pagan test gives us a low p value which means we can reject the null hypothesis and that there is an issue with our equal variance assumption

Checking for omitted predictors

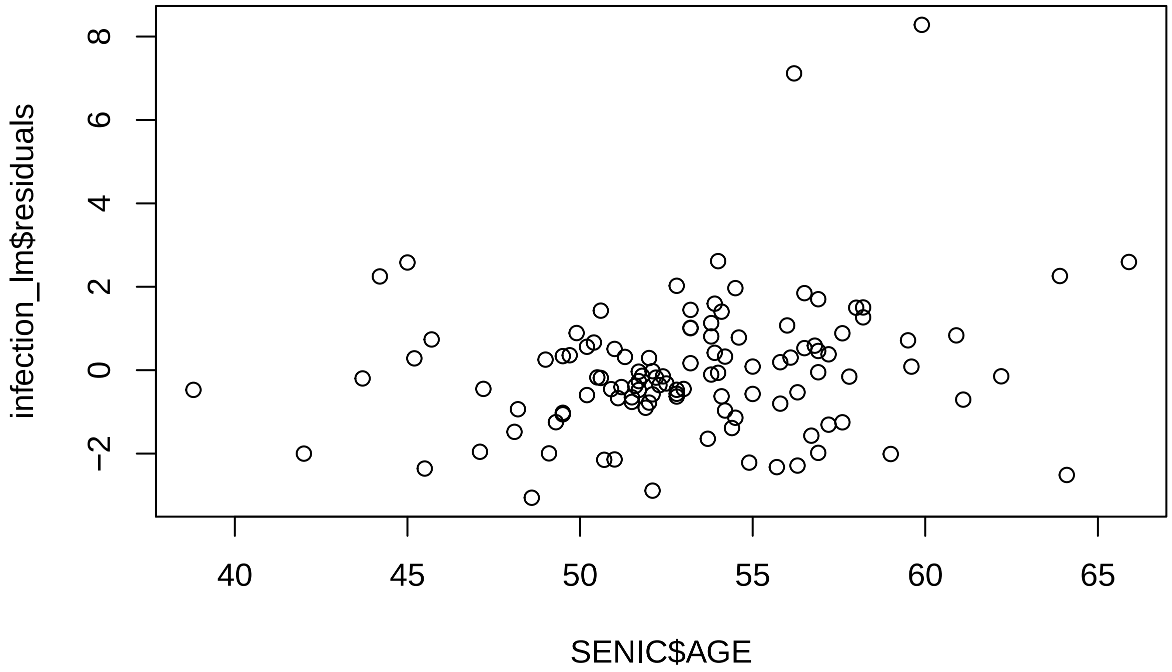
```
#scatterplot
plot(SENIC$AGE, SENIC$LOS) #plotting the potential omitted variable agaisnt our response variable
```



```
plot(SENIC$INFRISK, SENIC$LOS)#plotting our first predictor risk of infection agaisnt our response vari
```



```
#Plotting infection_lm model residuals vs Age
plot(SENIC$AGE, infection_lm$residuals)
```



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

```
age_lm <- lm(age ~ infection_risk, data = Infection_data)
summary(age_lm)
```

```
##
## 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
```

```r
qt(0.975, 111)
```

```
## [1] 1.981567
```

Because the absolute value of our critical value is less than our T-value we fial to reject our NULL hypothesis. Age is not a potentially omitted variable.