STA 108: Regression Analysis | Project 1

Doctor Amy Kim

Mahek Bhora & Tyler Le

October 27, 2023

I. Introduction & Summary

The data being used in this project is titled SENIC, n = 113, where each observation is a different hospital, of which we are looking to analyze the effect of the three different predictor variables on the average length of stay in days for all of the patients in a hospital. The predictor variables are: infection, the average estimated probability in percentage of acquiring an infection in the hospital; facility, the percentage of 35 total potential facilities and services that are provided by the hospital; and Xray, percentage of x-rays performed on patients without pneumonia. We did three separate simple linear regressions, one for each predictor variable and found the best predictor of Length to be Infection rate due to it having the highest R^2 value out of the three models. We deduced that the data is highly correlated but that the regression model does not fit the data well. This was due to the outliers as will be explained in the below sections.

II. Model Fitting

The model that we will pick is Model Infection, a model where we analyze the effect of the average estimated probability in percentage of acquiring an infection in the hospital on the variance of the average length of stay in days for all of the patients in a hospital. This is the model that we will pick because it has the highest R^2 value and the lowest MSE value. R^2 indicates the percentage of variation in Length explained by the linear relationship between a predictor variable and a response variable, for our best model, Infection and Length.

$R^2_{Infection} = 0.2846 > R^2_{Xray} = 0.1463 > R^2_{Facility} = 0.1264$

The MSE is the estimator for variance, σ^2, and it tells us the amount of error in Length that is not explained by the variable Infection.

$MSE_{Infection} = 2.638 < MSE_{Xray} = 3.147 < MSE_{Facility} = 3.221$

III. Diagnostics

Understanding Outliers in the Data: The IQR multiplied by 1.5 gives us the exact cutoff for outlier values, and the scatter plot tells us that there are 3 major outliers. These specific values skew the regression line because the least squares method emphasizes the errors of those outliers.

Histogram Analysis: The histogram has a relatively normal shape with the expected outliers falling outside of the main bell curve shape.

QQ Plot Analysis: A majority of the data falls in a linear relationship between X and Y. This tells us that there is a strong linear relationship between the percentile that each residual from the model falls in and the percentile that normally distributed residuals fall in. This relationship indicates an approximately normal distribution of the residuals which is desired.

Running the Shapiro-Wilks test, we find that since our p-value is extremely small, 1.699e-08, we reject the null hypothesis which had stated that the population is normally distributed. This violates our assumption of normally distributed data. However, due to the Central Limit Theorem (CLT), we can still state that the distribution of the sample means follows a normal curve since the sample size is large. Conducting the Fligner-Killeen test, we find that since our p-value is

larger than the significance level of 0.025 (alpha = 0.05/2), we fail to reject the null hypothesis. Thus, we can say that the variances between the two groups are equal at a 5% significance level.

IV. Analysis & Interpretation

Analyzing the confidence interval, increasing the infection rate by 1% can impact the length of the hospital stay by anywhere between 5.3038 days to 7.3697 days. Since the confidence interval is quite large, this implies that there is uncertainty regarding the true value of the population parameter. This is in line with our value of $R^2 = 28.46\%$. This means that only 28.46% of the total variance in the average length of a hospital stay is explained by the regression model including the infection rate. However, the MSE value is smaller than the MSR value, which implies that the infection rate model is explaining more of the variation in the data on average length of hospital stay. This disconnect between a low R\^2 value and a low MSE value implies that the data is highly correlated but that the regression model does not fit the data well. This can be explained by the effect of the outliers on the regression model's fit regarding the data. Additionally, the high t-statistic and f-statistic both lead to the conclusion that the regression coefficients are statistically significant. This essentially means that the relationship between infection rate and average length of hospital stay is not due to chance and that our regression model depicts a statistically significant linear relationship.

V. Prediction Results

Predicting the average length of the hospital stay using the different independent variables returns different values due to the differing relationships. When we used the x-ray model, the length of the stay was the largest as compared to the length of the hospital stay predicted by the percent of facilities in the hospital. The average length of the hospital stay was 9.150344 days when we predicted it using the infection rate, which is close to being in the middle of the values resulting from the other models.

VI. Conclusion

The study analyzed the effect of three predictor variables (infection rate, facility, and X-ray) on the average length of hospital stay in days for all of the patients in a hospital using a dataset of 113 hospitals. The study found that infection rate is the best predictor of length of stay, and while there is a linear correlation between the X and Y variables, the regression model does not fit the data well, likely due to the effect of outliers. The study also found that the relationship between infection rate and average length of hospital stay is statistically significant. One limitation of this study were the outliers that skewed the regression line.

Graphs and Plots

Infection Regression Line ANOVA table, summary table, and scatterplot:

Infection Residuals

Min	1Q	Median	3Q	Max
-3.0587	-0.7776	-0.1487	0.7159	8.2805

Infection Coefficients Table

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.3368	0.5213	12.156	< 2e-16 ***
infection	0.7604	0.1144	6.645	1.18e-09 ***

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

Infection Analysis of Variance Table

	Df	SS	MS	F-Ratio	Pr(>F)
infection	1	116.45	116.446	44.15	1.177e-09 ***
Residuals	111	292.76	2.638		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

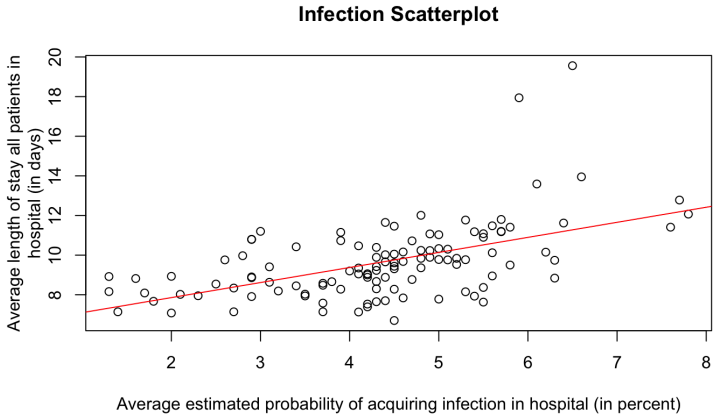

Figure 1

Facility Regression Line ANOVA table, summary table, and scatterplot:

Facility Residuals

Min	1Q	Median	3Q	Max
-3.2712	-1.0716	-0.2816	0.7584	9.5433

Facility Coefficients Table

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	7.71877	0.51020	15.129	< 2e-16 ***
facility	0.04471	0.01116	4.008	0.000111 ***

Residual standard error: 1.795 on 111 degrees of freedom
Multiple R-squared: 0.1264, Adjusted R-squared: 0.1185
F-statistic: 16.06 on 1 and 111 DF, p-value: 0.0001113

Facility Analysis of Variance Table

	Df	SS	MS	F-Ratio	Pr(>F)
facility	1	51.73	51.727	16.061	0.0001113 ***
Residuals	111	357.48	3.221		

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

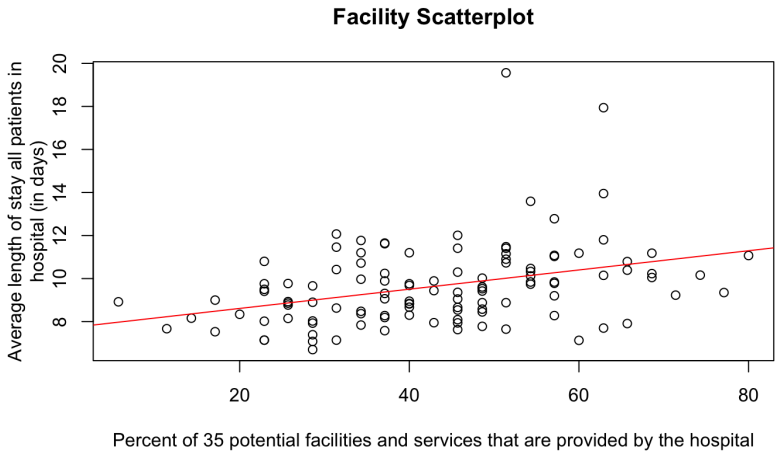

Figure 2

X-ray Regression Line ANOVA table, summary table, and scatterplot:

Xray Residuals

Min	1Q	Median	3Q	Max
-2.9226	-1.0810	-0.1487	0.8200	8.7008

Xray Coefficient Table

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.566373	0.726094	9.043	5.67e-15 ***
Xray	0.037756	0.008657	4.361	2.91e-05 ***

Residual standard error: 1.774 on 111 degrees of freedom
Multiple R-squared: 0.1463, Adjusted R-squared: 0.1386
F-statistic: 19.02 on 1 and 111 DF, p-value: 2.906e-05

Xray Analysis of Variance Table

	Df	SS	MS	F-Ratio	Pr(>F)
Xray	1	59.86	59.864	19.021	2.906e-05 ***
Residuals	111	349.35	3.147		

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

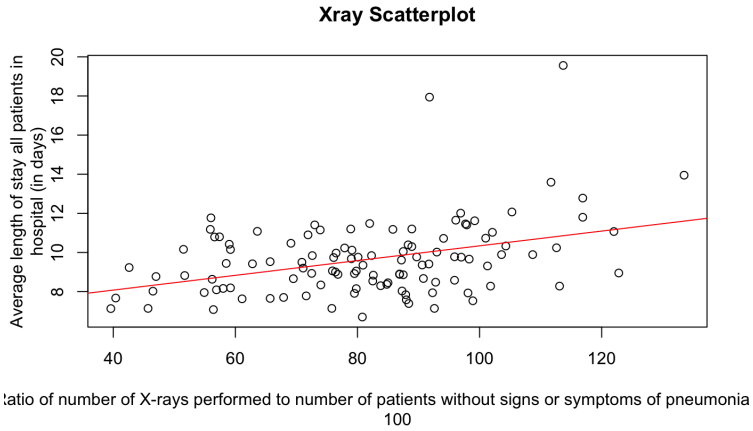

Figure 3

Outliers in Infections graph:

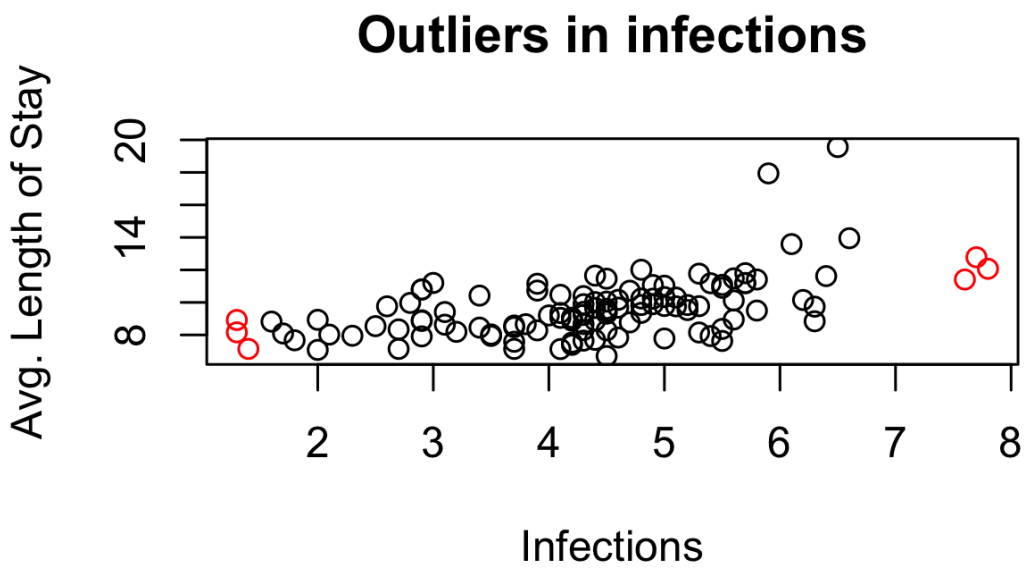

Figure 4

Outliers in Length graph:

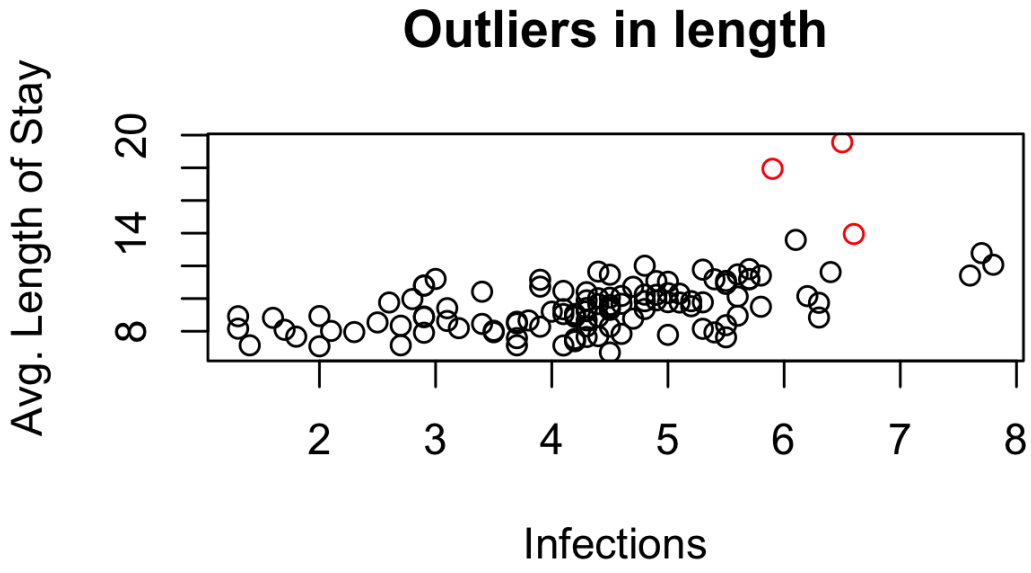

Figure 5

Normality Assessment plots:

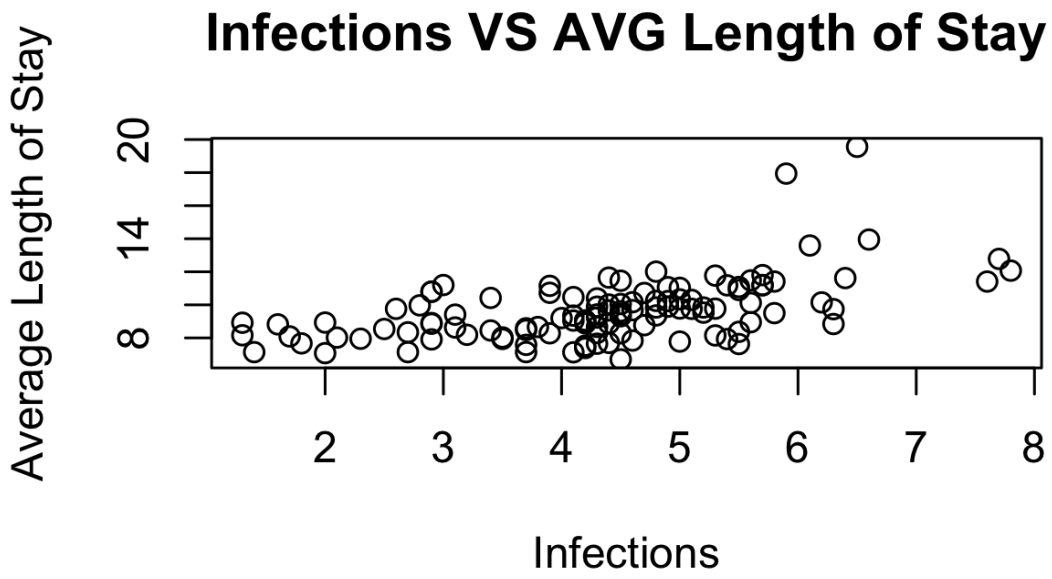

Figure 6

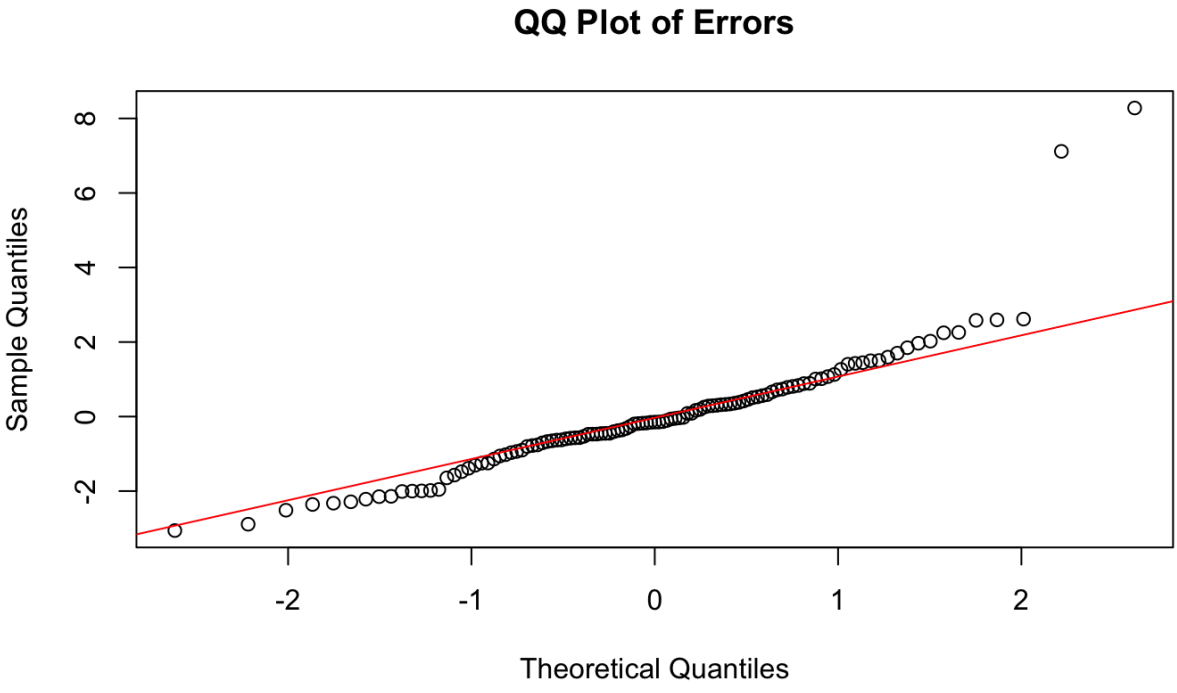

Figure 7

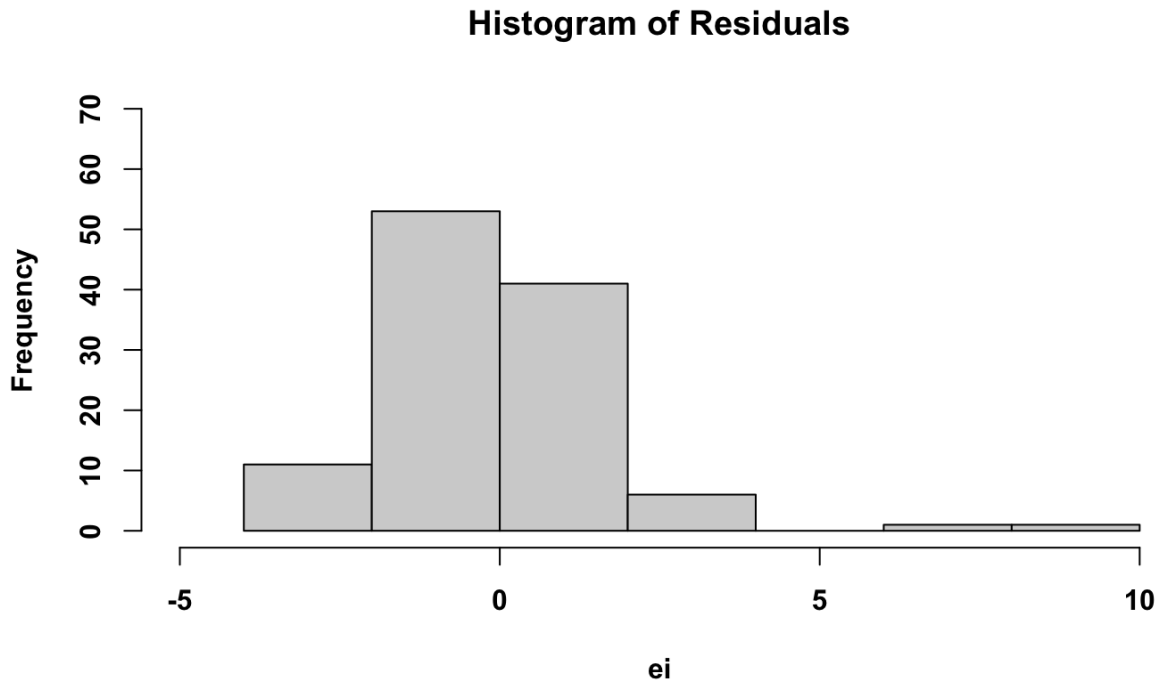

Figure 8

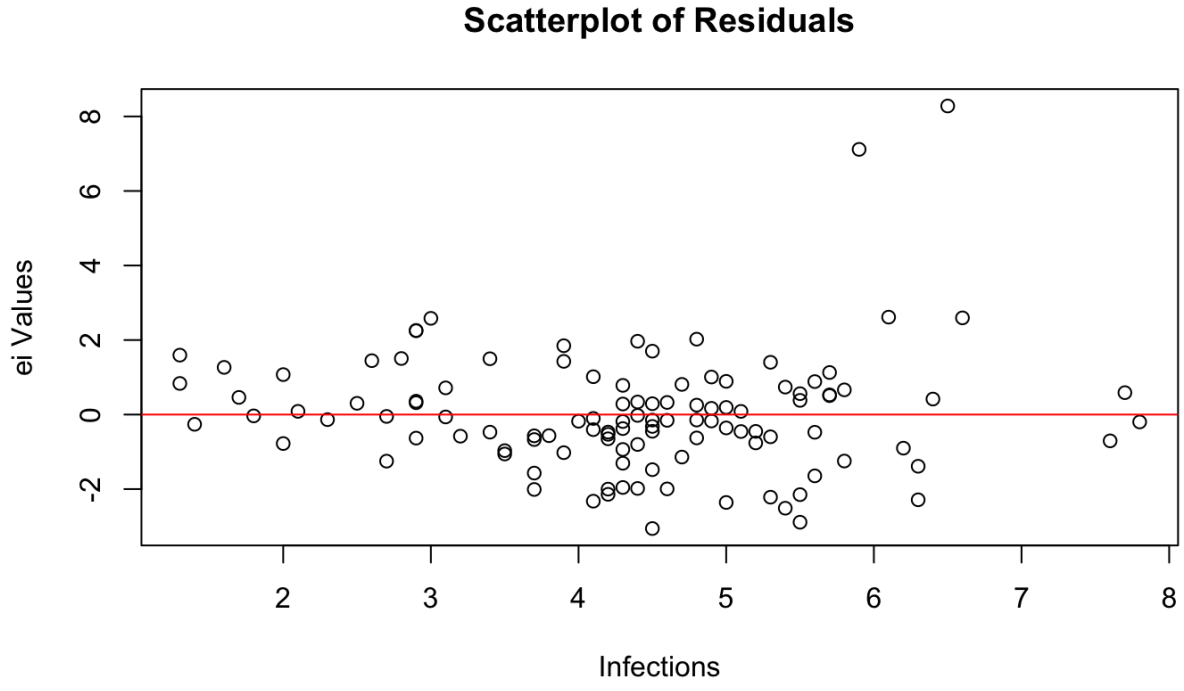

Figure 9

Subset of outliers in residuals

	length <dbl>	infection <dbl>	facility <dbl>	Xray <dbl>	ei <dbl>	yhat <dbl>
47	19.56	6.5	51.4	113.7	8.280478	11.27952
112	17.94	5.9	62.9	91.8	7.116730	10.82327

2 rows

Shapiro-Wilk normality test
data: ei
W = 0.87054,
p-value = 1.699e-08

Fligner-Killeen test of homogeneity of variances
data: senic$ei and senic$Group
Fligner-Killeen:
med chi-squared = 0.052793
df = 1
p-value = 0.8183

Analysis tables (Confidence interval, critical t value, summary containing R^2, ANOVA table containing F-Value, and critical F value)

Confidence interval

	2.5%	97.5%
(Intercept)	5.3038443	7.3697288
infection	0.5336442	0.9871976

Critical t value: 1.981567

Residuals:

Min	1Q	Median	3Q	Max
-3.0587	-0.7776	-0.1487	0.7159	8.2805

Infection Coefficients Table

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	6.3368	0.5213	12.156	< 2e-16 ***
infection	0.7604	0.1144	6.645	1.18e-09 ***

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

Infection Analysis of Variance Table

	Df	SS	MS	F-Ratio	Pr(>F)
infection	1	116.45	116.446	44.15	1.177e-09 ***
Residuals	111	292.76	2.638		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Critical F-Value: 0.051317

Prediction intervals

	Fit	Lower	Upper
Infection = 3.7	**9.150344**	**5.914575**	**12.38611**
Facility = 20	8.612921	5.004609	12.22123
Xray = 90	9.964398	6.430556	13.49824