**F. Linear Regression**

Regression Analysis is a way of predicting the value of one variable from another. It is a hypothetical model of the relationship between two variables. The models we have used are linear ones. Therefore, we describe the relationship using the equation of a straight line. A model is used to predict an outcome (Y) based on set of predictors (X). For both the Simple and Multiple Regression Model we are considering the entire sample of the dataset which is 100k samples.

*Simple Linear Regression model*:

We want to see how the value of time\_in\_hospital (in days) varies for patients who have been re-admitted in hospital within 30 days based on the number of medications administered.

Outcome variable is time\_in\_hospital ().

Predictor variable is Num\_medications ().

*Equation of the model: = +*

By performing preliminary assessments using scatter plot we determine the correlation between the Outcome and predictor variables. In our scatter plot as depicted in Figure 9a, we cannot depict the relationship between the two variables time\_in\_hospital and num\_medications as the variables are measured on a discrete scale, this presents us with a challenge. However, there could still be a linear relationship difficult to deduce. So, by using the heat map R scatter plot along with some base R functions such as Smooth scatter and contour functions which are based on defining colors to regions based on the density of data instances present. Figure 9b depicts the heatmap R scatter plot.

A computer screen capture

Description automatically generated with low confidenceChart, table

Description automatically generated

|  |  |
| --- | --- |
| *Figure 9a*: Scatter Plot to depict the correlation between outcome and predictor variables. | *Figure 9b*: Heat map for depicting the correlation between the outcome and predictor variables. |
|  |  |

Text

Description automatically generatedText

Description automatically generatedFrom the heat map it seems to be a positive correlation, but we cannot determine the relationship strength. So, Pearson’s and spearman correlation tests helped us to determine the relationship.

Figure 10: Shows the Pearson’ s and spearman’s correlation test values.

The correlation tests display estimate value with -1 to +1 so anu estimate that is close to -1 is strong negative and the closer it gets to zero the weaker it gets and vice versa for the positive estimate. In our case we got both the test estimates to be around 0.48 which indicate a moderate positive correlation.

Equation with Intercept and Slope Coefficients: *time\_in\_hospital = 0.17929 (num\_medications) + 1.73771*

Text

Description automatically generated

Figure 11: Summarized Output of the Simple Linear Model

Interpretation of Figure 11 is as follows:

* The num\_medications coefficient suggests that for every 10-count increase in the count of medications of the patient, we can expect increase in length of stay by 0.17929 \* 10 = 1.7929 days, on average.
* Is the Slope statistically significant? Yes. It is because the value of slope () must be significantly different from 0, and in our case (0.17929 - 0) is different from zero.
* Also, if the coefficient is large compared to the standard error, then statistically our coefficient will not be zero.

Observation from 95% confidence interval test, we can say with 95% confidence that slope lies between 0.1732 and 0.1853.

Text

Description automatically generated with medium confidence

Figure 12: Confidence Interval for Slope

Hypothesis test for the Slope coefficient is as follows:

*H0: β=0*

*H1: β≠0*

H0: No Linear relationship between time\_in\_hospital and num\_medications

H1: There is Linear relationship between time\_in\_hospital and num\_medications

* Figure 11 also shows that num\_medications coefficient is 58.21 standard errors away from zero and it is far from zero. The larger our t-statistic is the more certain we can be that the coefficient in not zero.
* The p-value is calculated is using t-statistic from the t distribution and it also helps us understand how significant our coefficient is is to the model. In practical terms any p-value below 0.05 is significant. In our model, we can see that *Intercept* and *num\_medications* have p-value of 2e-16 which is extremely small, and it is even below 0.001. We can conclude that the coefficients in this model are not zero.
* The residual Standard error is a measure of how well the model fits the data. From our model summary, we can see that on average, the actual values are 2.658 Days away from regression line.
* The F-statistic and overall p-value help us determine the result of our Hypothesis test. It is common for the F-statistic to be close to 1 if we have lot of predictors. However, for smaller models, a larger F-statistic and a small p-value generally indicates that null hypothesis should be rejected, and it clearly indicates that the coefficient in the model isn’t zero.

*Multiple Linear Regression Model:*

For Multiple Linear Regression model, we want to see how the value of time\_in\_hospital (in Days) varies for patients who have been re-admitted in hospital within 30 days based on the Number of Medications they have also used Number of lab procedures they have undergone.

Outcome variable is time\_in\_hospital ().

Predictor variables are Num\_medications (), Num\_lab\_procedures ().

*Equation of the model: =*

From the Simple linear regression preliminary assessment, we already know that time\_in\_hospital has a moderate positive correlation with num\_procedures, but we don’t know what the relation between the other variables is.

Diagram

Description automatically generated

Figure 13: Multi Collinearity Scatter plot

But with function (*cor*) we can identify correlation between multiple variables.

Text

Description automatically generated

Figure 14: Correlation Matrix

Equation with Intercept and Slope Coefficients:

*0.1578 (Num\_medications) + 0.0299 (Num\_lab\_procedures) + 0.7758*

Text

Description automatically generated

Figure 15: Summary of the Multiple Linear Regression Model

Interpretation of the Multiple linear regression model is as follows:

* In this case the p-value of F-statistic is <2.2e-16, which is highly significant, this means at least one of the predictor variables is significantly related to the outcome variable.
* Change in number of num\_medications and num\_lab\_procedures, the time\_in\_hospital of a patient is associated.
* For instance, for 10 count increase in the number of medications taken by the patient, we can expect an increase of 0.1578 \* 10 = 1.578 days of patient staying in hospital (when the num\_lab\_procedures are constant).
* Confidence Interval for the model coefficient is as follows:

Graphical user interface, text

Description automatically generated with medium confidence

Figure 16: Confidence Interval for the model coefficients

* What about the goodness of fit for the Model? The adjusted R-squared value for multiple regression model is 0.2627, meaning that 26.27% of the variance in measure of days can be predicted by num\_medications and num\_lab\_procedures number count.
* This model is better than the simple linear model with only num\_medications which had an adjusted R-squared of 0.2297.
* The RSE gives us a measure of error of prediction. Multiple linear regression model gives an error rate of 54%, which is better than the simple linear regression model where the RSE was 0.558 (i.e., 55.8% error rate).

**G. Analysis of Residuals**

For Analysis of Residuals, we verify the Gauss Markov Theorem conditions, we also try to explain and diagnose what multicollinearity is and explain how violations to the theorem affect the model.

*Verifying conditions of Gauss Markov Theorem for the Simple linear regression Model:*

Chart, scatter chart

Description automatically generated

Figure 17: Residual Plot for the Simple Regression Model

From Simple linear regression Model, we used in regression analysis we compute the mean value of residuals, i.e., -3.671336. Therefore, the first condition of Gauss Markov Theorem that errors have expectation zero is satisfied.

Next, we test the residuals are not correlated among themselves, for this we do the Durbin Watson test,

Text

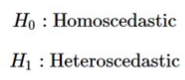
Description automatically generated

Text

Description automatically generated

From the above Durbin Watson test we can see that with a low p-value we reject the null hypothesis. We conclude that the residuals are correlated with each other.

We then perform Breusch-Pagan test to see if the Homoscedastic condition is satisfied or not satisfied.



Text

Description automatically generated

Based on the small p-value we reject the hypothesis that residuals are homoscedastic. We conclude that the errors do not have equal variances.

We test for normality of the residuals by using Histogram and QQ-plot. We observe from the figure 18 that the distribution of errors is right skewed. Therefore, normality condition is not satisfied.​ ​

Chart, histogram

Description automatically generated

Figure 18: Histogram to test normality

Chart, line chart

Description automatically generated

Figure 19: QQ-plot for the residuals

From figure 19, the qq-plot shows that not many samples lie within the confidence band and by this we can state that the normality conditions are not satisfied.

We try to test for normality Shapiro-Wilk test, if the residuals are Gaussian Distributed or not,

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Since we have too many samples, we were not able to test for Shapiro-Wilk test. Therefore, from the above test we can say that there is correlation between the variables that we took into consideration for the simple linear regression model.

*Verifying conditions of Gauss Markov Theorem for the Multiple linear regression Model:*

Chart, scatter chart

Description automatically generated

Figure 20: Residual plot for predictor variable Num\_medications

Chart, scatter chart

Description automatically generated

Figure 21: Residual Plot for predictor num\_lab\_procedures

For the Multiple model as the simple model, we computed the mean of residuals, i.e. 5.223827e-16, therefore the first condition of gauss Markov theorem that errors have expectation zero is satisfied.​

Next, we test to see if the residuals are correlated with each other or not, by using Durbin Watson test,

Text

Description automatically generated

Text, letter

Description automatically generated

Based on the Durbin Watson test since we have a low p-value we reject null hypothesis. ​ We conclude that the residuals are correlated with each other.​ Then We perform Breusch-Pagan test to see if the homoscedasticity condition is satisfied or not.​

Text

Description automatically generated

Text

Description automatically generated

From the above Breusch-Pagan test we reject the hypothesis as in the residuals show homoscedasticity. Due to the p value being very small and we conclude that the errors do not have equal variances. By using Histogram, we determine the normality of the residuals to be right skewed, and from the QQ-plot we conclude that the samples do not lie in the confidence band which results in not being normal.

Chart, histogram

Description automatically generated

Figure 22: Histogram of multiple linear regression model’s residuals to check normality

Chart, line chart

Description automatically generated

Figure 23: QQ-plot of multiple linear regression model’s residuals to check normality

Like the simple linear regression model the Shapiro-Wilk test doesn’t work for the multiple linear regression model as the number of samples we have been working for the model is 100k.

We then checked for the multicollinearity between the variables, which is the predictors must not be correlated. If the predictors are correlated and the estimation of regression coefficients gets affected which indicates multicollinearity!​

Diagram

Description automatically generated

Figure 24: Multicollinearity Scatter plots for all the 3 variables.

From figure 24, we can’t describe the collinearity between the 3 variables, but by using the correlation function, we find the correlation to be positive as the estimate are greater than 0.

Text, letter

Description automatically generated

Figure 25: Shows the Correlation Matrix for the Multiple linear regression model

As displayed above, we observe that the predictors number of medications and number of lab procedures are correlated. The square root of the VIF says how much larger the standard error of bi is, as compared to the case of predictors that are uncorrelated. So, we calculated the VIF for the model,



There are few thumb rules for VIF, and our model has VIF > 1 which indicates there is no reason for concern, and it is possible to say that there is correlation between the two predictors that we are using.

Looking at the results of both the model it is safe to say that both of our regression models 1 and 2 satisfy only one of the four conditions of the Gauss-Markov theorem conditions, i.e., expectation of errors is zero​. From the tests conducted it is visible that residuals are correlated and display heteroscedasticity and are not normally distributed. ​ Since the number of samples is greater than 5000, the formal Shapiro-Wilk test cannot be performed on our data.​ There is evidence of multicollinearity among the predictors used in regression model 2.​

**H. Analysis of Variance**

With the results of both regression testing and analysis of residuals only analyzing models with just continuous variables with numerical data is not a good modelling approach and with the methodologies used in Analysis of variance we are able to make use of the categorical variables, we have and create model by factoring the categorical variables and then adding then to the linear models.

By performing Analysis of variance (ANOVA) we can determine up to what percent of the samples have been explored, and how much are yet to be extracted and explained, and instead of creating multiple t-tests we ANOVA groups differences by comparing means of each group. Like a t-test, we use ANOVA test with a null hypothesis that the means are the same.

Since we can know the method that will let us use categorical variable in our linear models, we are designing new models with random 500 samples from the total 100k samples.

We are considering the same simple linear regression model as in the before section, i.e.,

Outcome variable is time\_in\_hospital (). Predictor variable is num\_medications ().

*Equation of the model: = +*

Calendar

Description automatically generated

Figure 25: Scatter plot to verify linearity between the variables

From figure 25, we cannot describe what kind of correlation exists between the variables, so we use the R programming base functions such as Smooth scatter and contour to plot a density heat scatter plot to show the correlation.

Surface chart

Description automatically generated with medium confidence

Figure 26: Heat scatter plot for displaying correlation between the variables

The figure 26, doesn’t show the kind of correlation that exists, so we work on finding the Pearson’s and spearman’s correlation coefficient.

Graphical user interface, text, application

Description automatically generated

We got the Pearson’s coefficient and it indicated there exists a positive correlation between the variables, the spearman’s correlation test was not working for this random sample that we were working on, so we are not showing any estimate values.

*Text, letter

Description automatically generated*

A picture containing text

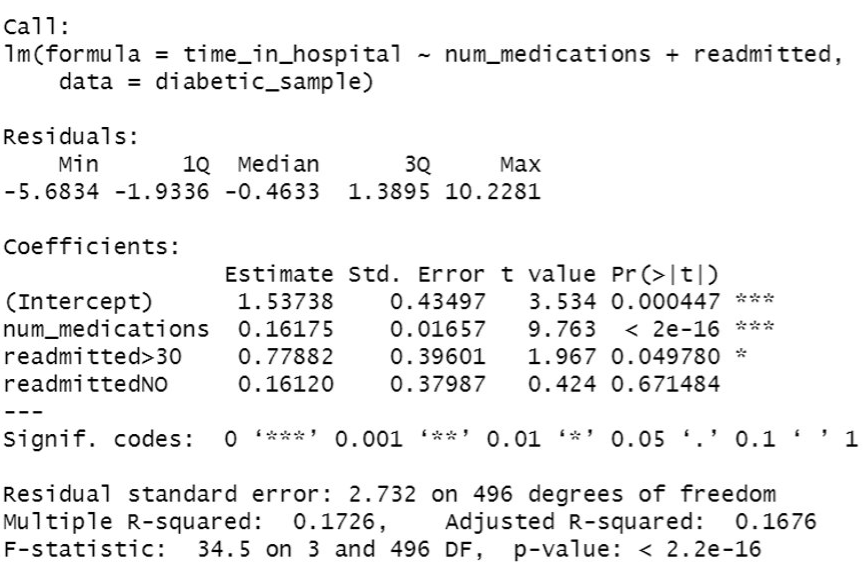
Description automatically generatedFrom the summary above, we can understand that the slope is significant since it is not equal to zero, and t value is not small along with that the p-value is very significant when we consider the alpha to be 95% which is 0. 005.When considering the 95% confidence interval the values of intercept and slope lie within, the below values,

The goodness of fit of the model is 0.6204048, and with all the values and interpretation we can assume the model to be a good one.

For the multiple linear regression model the outcome variable is same as the previous model i.e., time\_in\_hospital, but this time in the predictor variables instead of just having numerical variable we have taken an addition of a categorical variable.

Outcome variable time\_in\_hospital (). Predictors variables are num\_medications () & readmitted (X2).

*Equation of the model: = + +*



From the above summary of the new multiple model, we can see that both the predictor variables are significant, since the estimate coefficients are not equal to zero, but when looking value at the t value the readmittedNO is closer to zero since it is 0.424, and the rest 3 are 2 and above. The p values are significant, but it varies based on the alpha we take since if the alpha we take is 0.05 only readmitted<30 and readmitted>30 is significant not readmittedNO. When comparing the multiple models with linear model the adjusted R-squared is better as 0.1676 > 0.1598. Also, the residual standard error is less for the multiple model when compared to the simple model. So, we can safely assume that the new Multiple linear regression model is better than the simple model.

Chart, scatter chart

Description automatically generatedWe wanted to group the scatter plot of the outcome and predictor variable based on the categorical variable.

Figure 27: Scatter plot based grouped based on the categorical variable.

Chart, scatter chart

Description automatically generated

Figure 28: Linear regression analysis paradigm predicted lines

Timeline

Description automatically generated

Figure 29: Histogram to view the time\_in\_hospital distribution based on categorical variable

Chart, box and whisker chart

Description automatically generated

Figure 30: Visual comparison using boxplots to account for variability of observations within each group