Are Linear Regression Models Capable of Predicting Bilirubin Levels in Patients Being Tested For Hepatitis C Virus?

Celvin Abraham (2139410) & Aayushi Shrivastava (2139461)
Master of Science in Data Analytics, Christ Deemed to be University, Bangalore

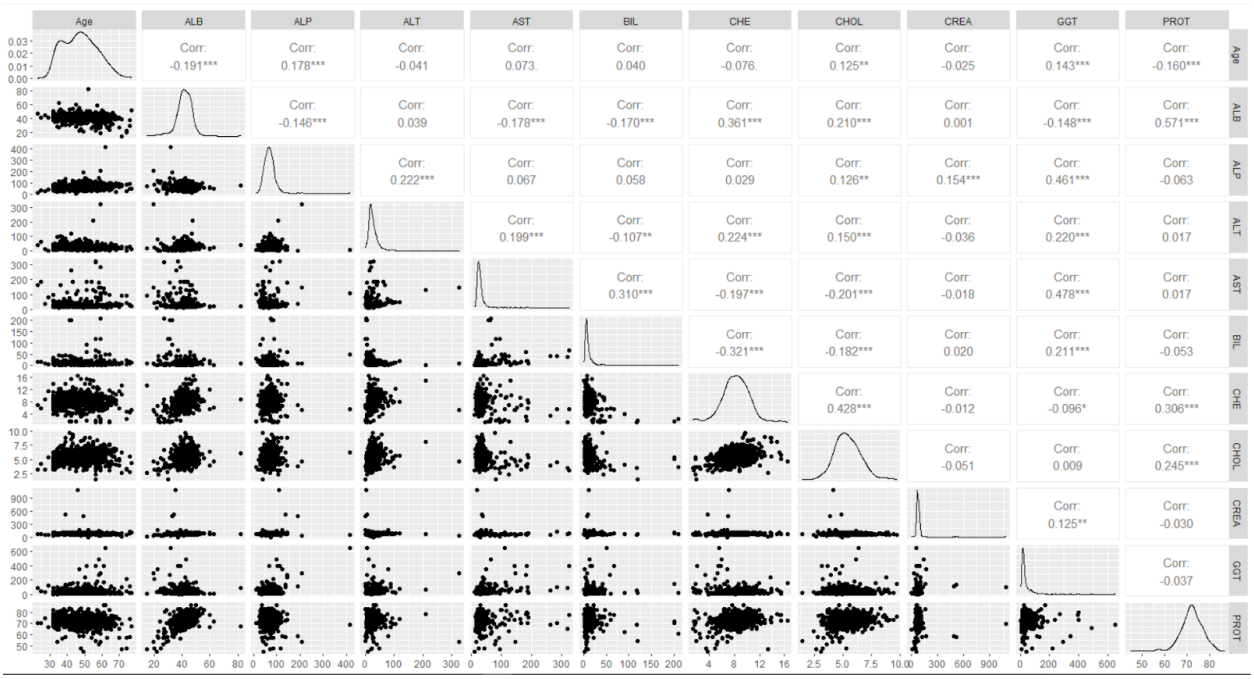
Keywords: Bilirubin, Hepatitis C, Linear Regression, Multiple Linear Regression, Prediction

Abstract: Hepatitis C is a viral infection that causes liver inflammation, sometimes leading to serious liver damage. The hepatitis C virus (HCV) spreads through contaminated blood. Until recently, hepatitis C treatment required weekly injections and oral medications that many HCV-infected people couldn't take because of other health problems or unacceptable side effects. That's changing. Today, chronic HCV is usually curable with oral medications taken every day for two to six months. In this study, we have tried to obtain Simple Linear Regression and Multiple Linear Regression Models in order to truly understand if Linear Regression Models alone are enough to predict Bilirubin levels in patients who need to be tested and diagnosed for Hepatitis C. We have used the method of Hypothesis Testing, Linear Regression Assumption Analysis, Residual Analysis on Simple Linear Regression Models and Multiple Linear Regression Models. We have used also obtained Stepwise, Forward and Backward to understand the best variables that can explain the Bilirubin levels in patients.

Introduction: Hepatitis C is part of a group of hepatitis viruses that attack the liver. It is commonly found in infected blood. It is also rarely found in semen (cum) and vaginal fluids. The virus is usually passed on by using contaminated needles and syringes or other items with infected blood on them. It can also be passed on through unprotected sex, especially when blood is present. It often has no noticeable symptoms. Some people's bodies can clear the infection on their own but others may develop chronic (long-term) hepatitis C and will need to take antiviral treatment to cure the infection and prevent liver damage. When we talk about the liver, we cannot ignore an important chemical Bilirubin that is created in the liver. [1]A bilirubin test measures the levels of bilirubin in your blood. Bilirubin (bil-ih-ROO-bin) is a yellowish pigment that is made during the normal breakdown of red blood cells. Bilirubin passes through the liver and is eventually excreted out of the body. Higher than normal levels of bilirubin may indicate different types of liver or bile duct problems. Occasionally, higher bilirubin levels may be caused by an increased rate of destruction of red blood cells (hemolysis). [2]Bilirubin values of 2.5–3.0 mg/dl or greater establish the presence of the icteric phase of hepatitis. Bilirubin levels in excess of 30 mg/dl suggest hemolysis (overproduction of bilirubin) or renal failure (failure of excretion). Serum bilirubin levels are not always of clinical value.

Methods:

Study Group: A set of 615 patients who were tested and were awaiting their Hepatitis C virus test were considered for our study. The dataset consists of 14 variables. Out of the 14 variables 9 variables refer to laboratory data containing levels for Albumin(ALB), alkaline phosphatase(ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST),Bilirubin(BIL), Serum cholinesterase (ChE),Cholesterol(CHOL),Creatinine(CREA), Gamma-glutamyl transferase (GGT) ,Protein(PROT). The other variables depict the demographics of the patient, like Age and Gender. Below Correlogram depicts the distributions and the correlations between all the variables that we have considered for our study.



As we can see CHE and CHOL are two variables that look Normall Distributed. We also see that there is no strong correlation between any of the variables. The correlation is either weakly negative or weakly positive.

Statistical Analysis: Exhaustive review of laboratory factors and univariate Simple Linear Regression analysis and Multiple Linear Regression with Step-Wise variable selection was performed to identify significant predictors among the collected data. A study on the Linear Regression Assumptions followed with Residual Analysis was performed.

Simple Linear Regression: For our study we have considered 3 of the best models that were better at predicting the Bilirubin levels based on the levels of other laboratory test variables. The Simple Linear Regression Equation is given by:

$$y = \beta 0 + \beta 1.x1$$

Where

y = response variable

x1 = predictor variables

β0 = Intercept

β1 = Regression Coefficients

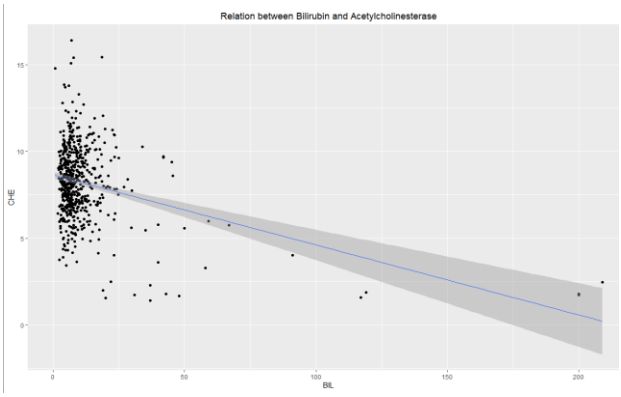
Equation after finding the estimates is given by:

$$\hat{y} = \hat{\beta 0} + \beta 1.\hat{x1}$$

Findings: Below are the findings and analysis that were done for Simple Linear Regression Analysis.

Model 1: Predicting Bilirubin Levels From Acetylcholinesterase

Below Scatter Plot depicts the relationship between the two variables.



From the graph we can see, that the points of the two variables are not scattered around the modeled line. This suggests that there is very little to no correlation between the variables when we look at them visually. The correlation coefficient between the bilirubin and acetylcholinesterase is -0.32. This shows that there is a negative correlation between the bilirubin and acetylcholinesterase.

Bilirubin = β1 (Acetylcholinesterase (CHE)) + β0(c)

The substituted equation is given by **Bilirubin = 31.9 - 2.5(Acetylcholinesterase (CHE))**

Hypothesis Testing: Significance of Regression Coefficients

H_0: The coefficients β_0 and β_1 are insignificant.

H_1: The coefficients β_0 and β_1 are significant.

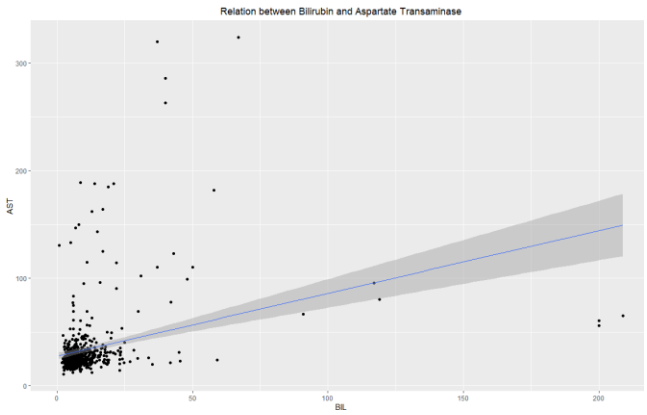
Since the Pr(>|t|) is 2.8927e-30 and 1.479155e-15 are less than α (0.05). So, we reject the Null Hypothesis and say that β0 and β1 are significant.

Now we will see whether the model which we have obtained is a good fit for the data (which we collected) or not. For this test we deploy the ANOVA (which makes use of F-test) to determine if the model is a good fit. The model only predicts 10.1%-10.2% of the records correctly. i.e The accuracy is about 10.1%-10.2%. Since F-statistic value is 67.3 and the p-value is less than 0.05 we can say that the model is not that much of a good fit for the data.

Model 2: Predicting Bilirubin Levels From Aspartate Transaminase

Bilirubin = β_1 (Aspartate Transaminase (AST)) + β_0(c)

Scatterplot depicting the relationship between the variables in consideration



Based on the co-efficient we obtained, our model becomes:

Bilirubin = 5.481 + 0.164 (Aspartate Transaminase (AST))

The above graph does not show the points being scattered around the modeled line and thus we do not see any strong relationship between BIL and AST. The correlation coefficient between the

bilirubin and aspartate transaminases 0.30. This shows that there is a really weak to negligible positive correlation between the bilirubin and aspartate transaminases.

Hypothesis Testing: Significance of Regression Coefficients

H_0: *The coefficients* β_0 *and* β_1 *are insignificant.*

H_1: *The coefficients* β_0 *and* β_1 *are significant.*

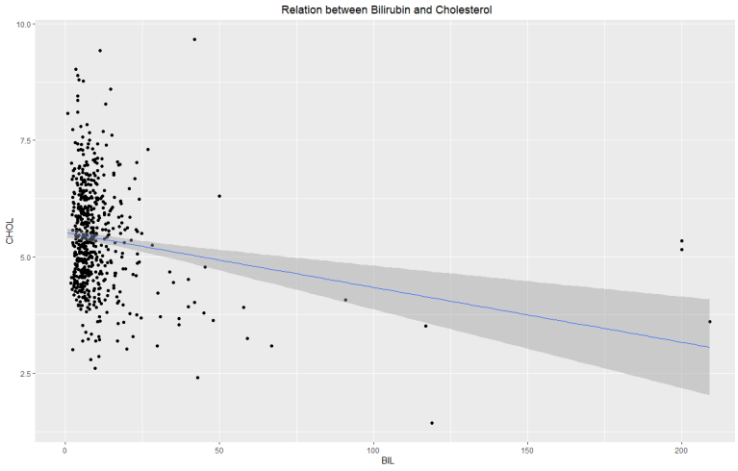
Since the Pr(>|t|) is 3.336248e-08 and 1.506649e-14 are less than α (0.05). So, we reject the Null Hypothesis and say that β0 and β1 are significant.

To confirm if the model we obtained is a good fit for the data, we deploy the ANOVA (which makes use of F-test) to determine if the model is a good fit. From the results we obtained in R, we can say that the model only predicts 9.4%-9.5% of the records correctly. i.e The accuracy is about 9.4%-9.5%. Since F-statistic value is 62.22 and the p-value is less than 0.05 we can say that the model is a somehow fit for the data.

Model 3: Predicting Bilirubin levels based on a patients Cholesterol levels.

$$\text{Bilirubin} = \beta_1 (\text{cholesterol (CHOL)}) + \beta_0(c)$$

Below Scatter Plot depicts the relationship between the two variables.



Based on the Coefficients we obtain Model becomes:

Bilirubin = 26.11 - 2.799 (Cholesterol (CHOL))

We see that the points are not scattered around the modeled line and thus there does not seem to be any correlation between the data points by a lot of points. The correlation coefficient between

the bilirubin and cholesterol -0.18. This shows that there is a negative correlation between bilirubin and cholesterol.

Hypothesis Testing: Significance of Regression Coefficients

H_0: *The coefficients* β_0 *and* β_1 *are insignificant.*

H_1: *The coefficients* β_0 *and* β_1 *are significant.*

Since the Pr(>|t|) is 1.402955e-13 and 9.245018e-06 are less than α (0.05). So, we reject the Null Hypothesis and say that β0 and β1 are significant. Now to see if the model is a good fit and to understand the variability explained by the independent variables for the dependent variable. For this test we deploy the ANOVA (which makes use of F-test) to determine if the model is a good fit.

Multiple Linear Regression: Upon having a look at few Simple Linear Regression Models for the prediction of Bilirubin. To understand the behavior of bilirubin for when we include more than one variables, we deploy a Multiple Linear Regression using Step Wise Variable selection method. The Multiple Linear Regression Equation is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p$$

Where

y = response variable

x1,x2,…..xp = predictor variables

β0 = Intercept

β1, β2,….βp = Regression Coefficients

Equation after finding the estimates is given by:

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} \hat{x_1} + \hat{\beta_2} \hat{x_2} + \ldots\ldots + \hat{\beta_p} \hat{x_p}$$

In order to understand the influence of all the other biological test variables on Bilirubin, we have used Multiple Linear Regression. It is important to note that in our study, we saw that the

Backward, Forward and the Step-Wise method, all returned highly similar results for R-Squared value and Adjusted R-Squared values.
The Multiple Linear Regression Formula obtained is given below:

BIL = ALT + AST + CHE + GGT

From the coefficients obtained, we can write the final equation for the Multiple Linear Regression Model as given below.
Final Multiple Linear Regression Model is then given by:

BIL = (-0.10)ALT + (0.12)AST + (-1.87)CHE + (0.03)GGT + 23.63

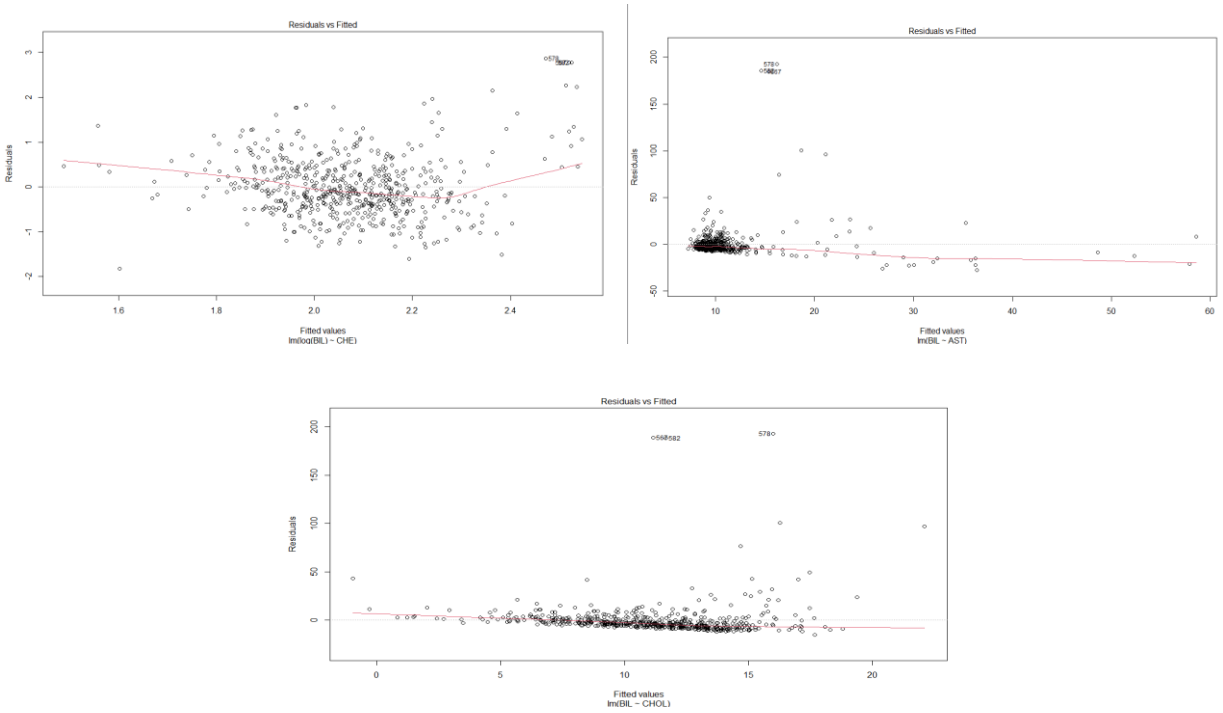
We obtain a p-value which is 2.2e-16, i.e less than the level of significance i.e 0.05, thus we can accept the null hypothesis and conclude that the model is a good fit.
However, in order to look at how much of variability is explained by the set of independent variables for the dependent variable, we need to refer to the R-Squared value, even better, the Adjusted R-Squared value.

We see that the Adjusted R-Squared value is really less i.e 0.1787. This means that only 17% of the variability in the dependent variable is explained by the independent variable for this Multiple Linear Regression. Thus we can say that simple ALT, AST, CHE, GGT levels are not enough to predict the Bilirubin levels with accuracy in patients. In order to understand this relationship better, we might have to opt for Non-Linear Models.

Assumption Testing For Simple Linear Regression

1. Linearity of the data

The linearity assumption can be checked by inspecting the Residuals vs Fitted plots for all the 3 Simple Linear Regression Models used in our study.

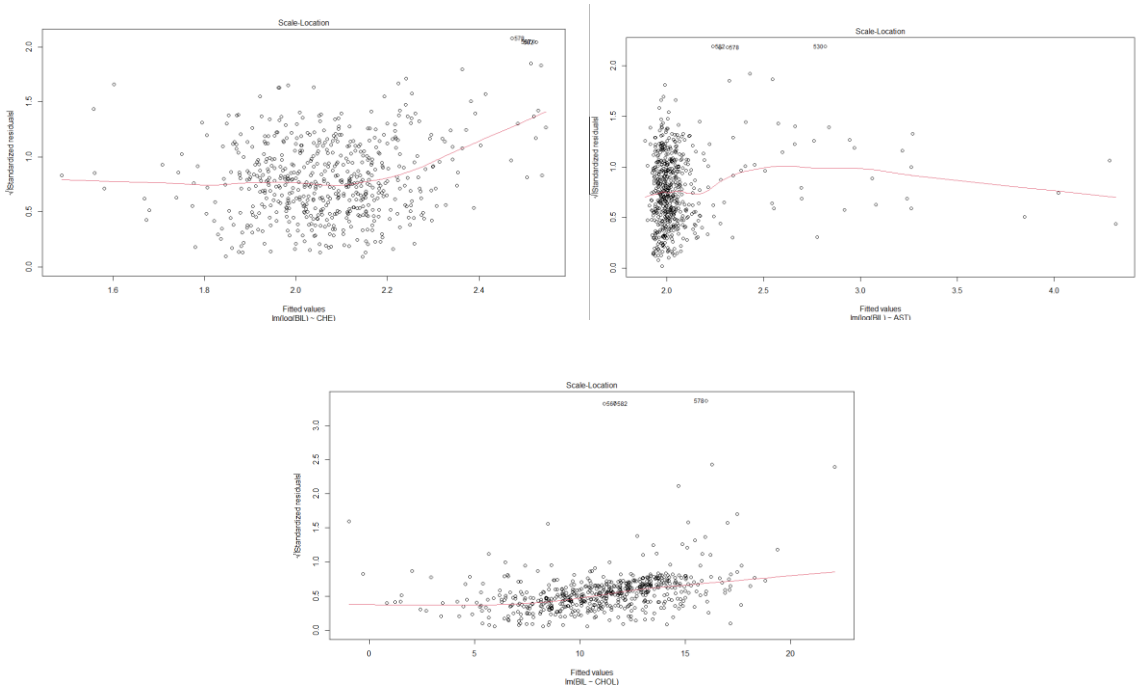


Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.

In the Model 1, there is some pattern in the residual plot and the red line is not that much approximately to the horizontal line. This suggests that we can't assume a linear relationship between the predictors and the outcome variables. In the Model 2, there is no pattern in the residual plot and the red line is not approximately to the horizontal line. So we can't assume a linear relationship between the predictors and the outcome variables. In the Model 3, there is no pattern in the residual plot and the red line is not approximately to the horizontal line. So we can't assume a linear relationship between the predictors and the outcome variables.

2. Homogeneity Of Variance

This assumption can be checked by examining the scale-location plot, also known as the spread-location plot.



These plots shows if residuals are spread equally along the ranges of predictors.

In our Model 1, this is not the case. It can be seen that the variability (variances) of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or heteroscedasticity). So this assumption does not satisfy the Model 1.

In our Model 2, this is not the case. It can be seen that the variability (variances) of the residual points decreases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or heteroscedasticity). So this assumption does not satisfy the Model 2

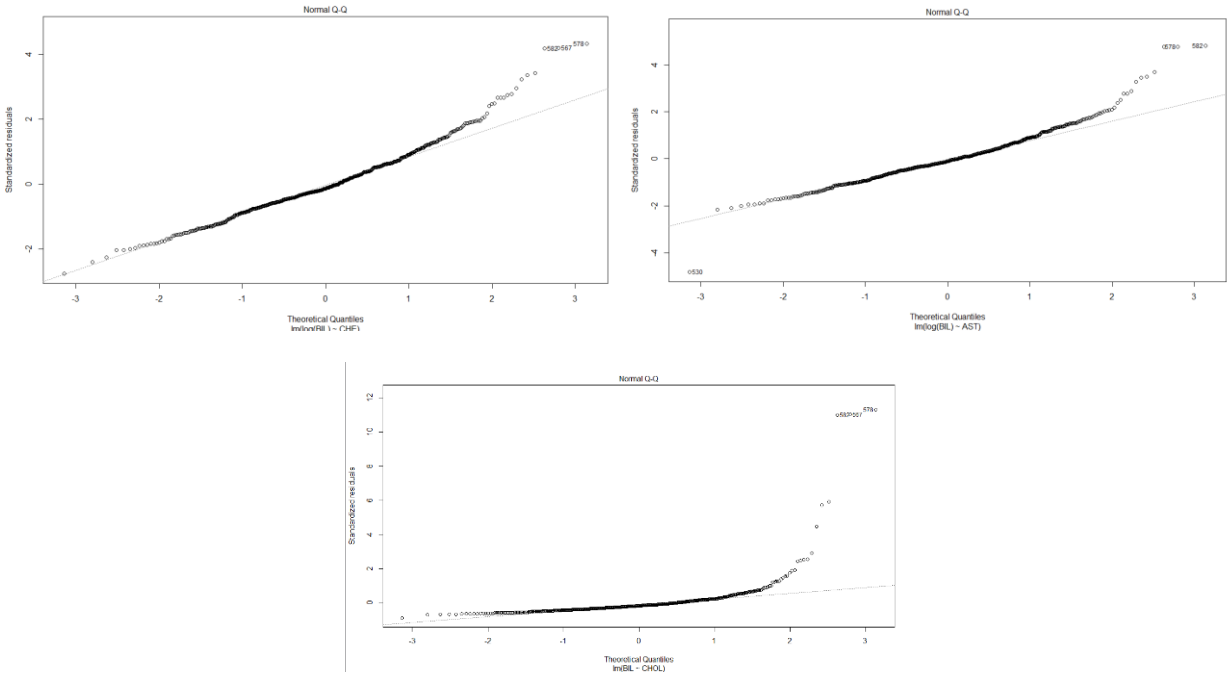
It's good if we see a horizontal line with equally spread points. In our Model 3, this is not the case.

It can be seen that the variability (variances) of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or heteroscedasticity).

So this assumption does not satisfy the Model 3.

3. Normality of Residuals

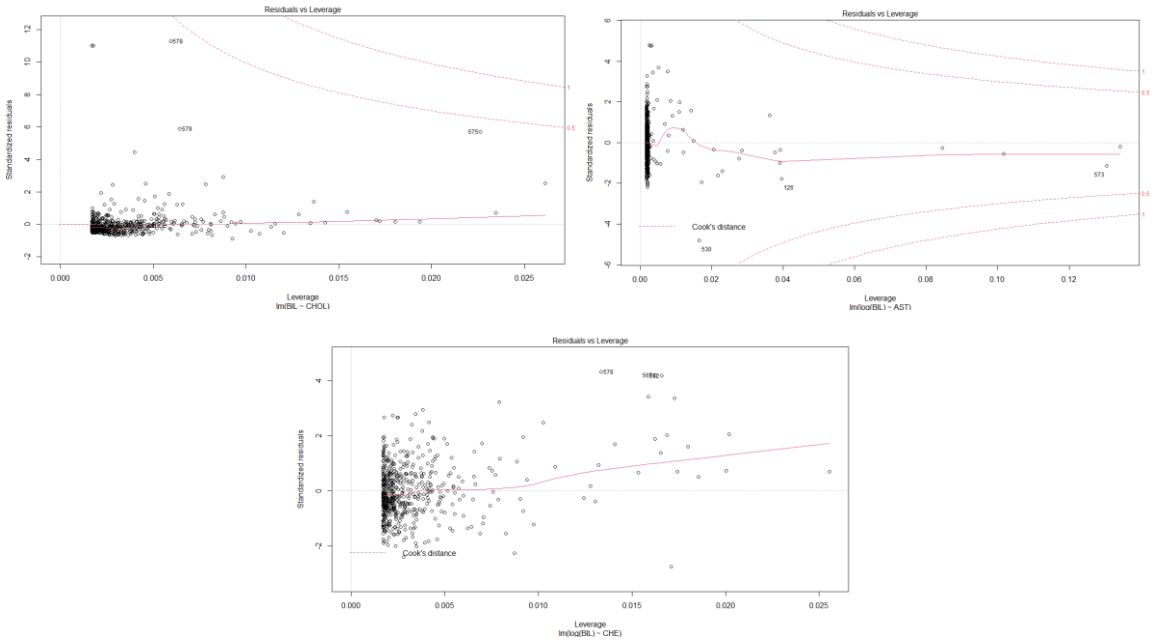
The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.



In the Model 1,Model 2 and Model 3 all the points fall approximately along this reference line, so we can assume normality.

4. Outliers and high leverage points

An outlier is a point that has an extreme outcome variable value. The presence of outliers may affect the interpretation of the model, because it increases the RSE.
Outliers can be identified by examining the standardized residual (or Student's t residual), which is the residual divided by its estimated standard error. Standardized residuals can be interpreted as the number of standard errors away from the regression line.
Observations whose standardized residuals are greater than 3 in absolute value are possible outliers



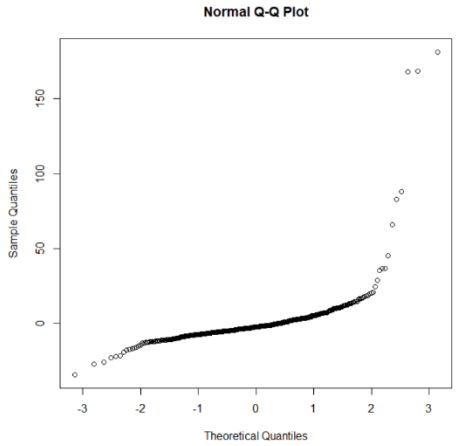
In the model 1, outliers exist. So this assumption does not work for model 1.
Model 1 satisfies only one assumption and the other three assumptions do not satisfy. So we can't use this model for prediction. All the four assumptions have to be satisfied to use the model.
In the model 2, outliers exist. So this assumption does not work for model 2.
Model 2 satisfies only one assumption and the other three assumptions do not satisfy. So we can't use this model for prediction.
In the model 3, outliers exist. So this assumption does not work for model 3.
Model 3 satisfies only one assumption and the other three assumptions do not satisfy. So we can't use this model for prediction.

Residual Analysis

In order to understand our model and variables better, we often opt for Residual Analysis in order to decide if this is the model that we must go ahead with.
There are some Assumptions for Residuals that we need to consider before moving ahead with confirming the model. The assumptions are as stated below:

- The dataset must have some linear relationship
- Multivariate normality - the dataset variables must be statistically Normally Distributed (i.e. resembling a Bell Curve)
- It must have no or little multi-collinearity - this means the independent variables must not be too highly correlated with each other. This can be tested with a Correlation matrix and other tests
- No auto-correlation - Autocorrelation occurs when the residuals are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.
- Homoscedasticity - meaning that the residuals are equally distributed across the regression line i.e. above and below the regression line and the variance of the residuals should be the same for all predicted scores along the regression line.

Normality QQ Plot



Normality Hypothesis Test for Residuals

In order to perform the Normality Test, we will use the Shapiro Wilk test.
The Hypothesis that we have defined is as follows:

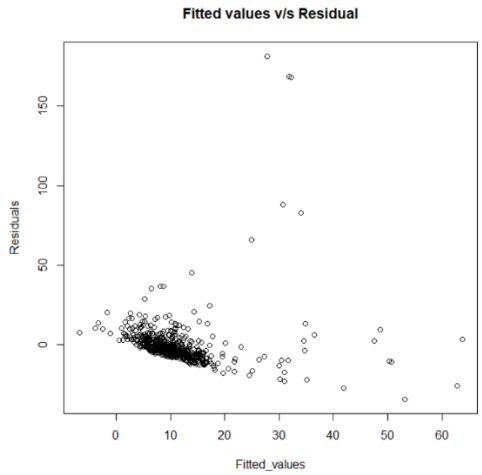
$$H_0: \text{Residuals of the model follow Normal Distribution} \sim (0, \sigma 2)$$

$$H_1: \text{Residuals of the model do not follow Normal Distribution} \sim (0, \sigma 2)$$

The results for the above test states that the Residuals in fact are following Normal Distribution since the p-value is significantly less than the alpha value(0.05) in our case.

Thus we go ahead and accept the Null Hypothesis and state that the Residuals of the model follow Normal Distribution.

Zero Mean and Constant Variance Test



Since the points are not scattered randomly and are present in some pattern constant variance assumption is not satisfied, hence the variance across all points is not constant. We see that there is some trend and a cluster present between the fitted values and the residuals.

Since the points are scattered around 0(in y axis for residuals) we can say that the zero mean assumption is satisfied.

Conclusion: In this study, we have found out that predicting Bilirubin levels in patients using models of Simple Linear Regression and Multiple Linear Regressions are just not enough. There is a lot assumptions of Regressions being violated by these Linear Models which suggests us that probably this dataset is not meant for a linear modeling problem but in fact a non-linear modeling problem. We have proved that the bilirubin levels can neither be predicted using 1 single laboratory variables nor with the help of multiple laboratory variables. This is a strong evidence that we need to opt for models that are either non-linear or perform regression analysis with linear data after performing transformations on the data and run through all the tests yet again. Thus, in conclusion, we can say that the Simple Linear Regression Model and the Multiple Linear Regression Models are not apt in identifying/predicting the Bilirubin levels of patients who are to be classified/diagnosed as Hepatitis C patients. Our R-Squared values were really low which suggested that the variability of the dependent variable is not very well explained for the dependent variable. We also see that a lot of assumptions are being violated which hampers us from moving ahead with the chosen models. We also performed Residual Analysis in order to

understand the behavior of the residuals of our model. We see that to some extent the Residuals are following their assumptions.

References:

[1] Mayo Clinic on Bilirubin Test

[2] Viral Hepatitis A & E Study by John Hopkins University

[3] Predictive Models for Neonatal Follow-Up Serum Bilirubin: Model Development and Validation. Joseph H Chou, MD, PhD, MIT

[4] Method and Model for Jaundice Prediction Through Non-Invasive Bilirubin Detection Technique, Dr Ashok Kumar et al

[5] Prediction of significant hyperbilirubinemia by estimating cord blood bilirubin in neonates with ABO incompatibility, Hemant Jain et al

[6] Use of serum bilirubin/albumin ratio for early prediction of bilirubin induced neurological dysfunction, Dalia Mosullum et al

[7] A clinical prediction rule for acute bilirubin encephalopathy in neonates with extreme hyperbilirubinemia Zhang, Fanhui MD et al.