# Assignment Report

Name: Shagun Mishra
Id no. : 2018A1PS0767H
Dataset: 10

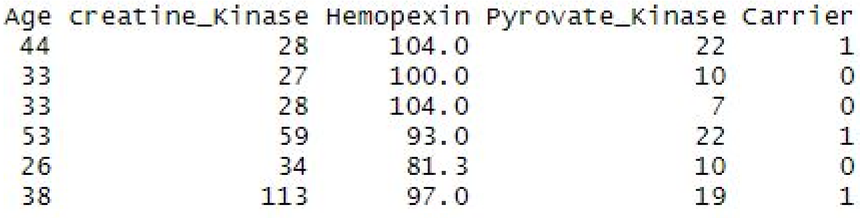
## Objective:

To develop a model to test if a person is a carrier of DMD based on the amount of creatine kinase, pyruvate kinase, Hemopexin levels and the person's age.

## Model :

Our dataset consisted of 5 variables - Age, Creatine kinase levels, Pyruvate Kinase levels, Hemopexin levels and Carrier ( 1- if the person is a carrier of disease, 0 - otherwise)

Sample data:



```
Age creatine_Kinase Hemopexin Pyrovate_Kinase Carrier
 44              28     104.0              22       1
 33              27     100.0              10       0
 33              28     104.0               7       0
 53              59      93.0              22       1
 26              34      81.3              10       0
 38             113      97.0              19       1
```

We have assumed Carrier variable to be dependent variable and Age, Creatine kinase levels, Pyruvate Kinase levels and Hemopexin levels to be independent variables/ predictor variables.

As the Carrier variable only takes two discrete values, Logistic regression is applied to determine the model.Using logistic regression we can estimate the probability that the person is a carrier of DMD disease.

$$\mathbf{E(y\hat{}) = e^{( b0 +b,x1+b2x2+b3x3+b4x4)} / (1 + e^{( b0 +b,x1+b2x2+b3x3+b4x4)})}$$

Where y= Carrier variable
$x_1$= Creatine Kinase variable
$x_2$= Hemopexin Variable
$x_3$= Pyruvate Kinase variable
$x_4$= Age variable

Moreover , the estimate of y essentially means the probability of y being true for given values of dependent variables , i . e. , E(y) = P(y=1 | $x_1$, $x_2$,$x_3$,$x_4$ )

## Assumptions:

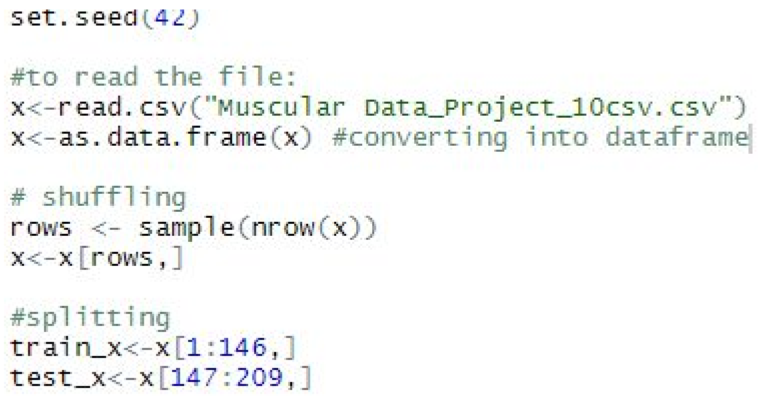
Assumptions made while developing the logistic regression model::

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
- There are no high intercorrelations (i.e. multicollinearity) among the predictors.
- There is a linear relationship between the logit of the outcome and each predictor variable.
- The values of ∈ are independent and the variance of ∈ is the same for all values of x .

## Development & Analysis of model:

The model development and analysis has been carried out by R programming. R packages - ggplot2, dplyr, tidyr have been used.

Preprocessing of data: The data was shuffled and splitted into training set ( 70%) and testing set (30%).



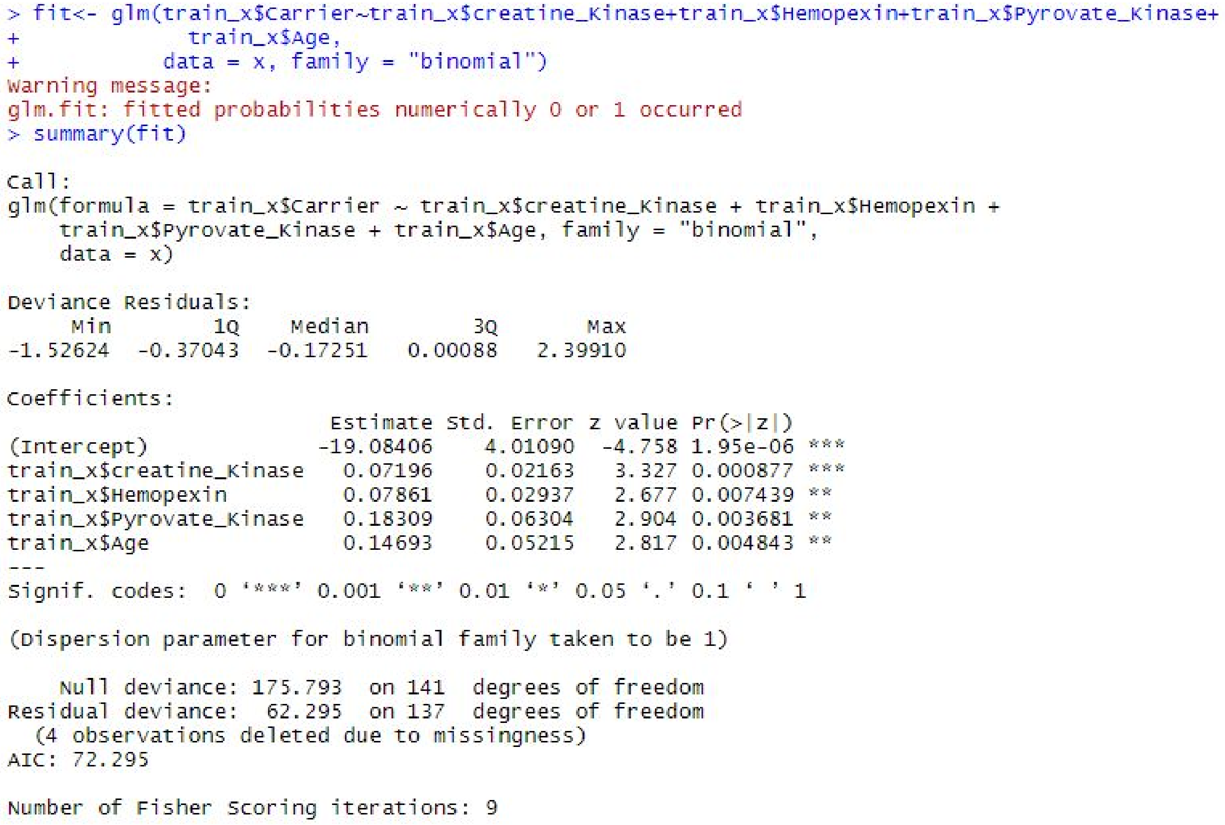
```
set.seed(42)

#to read the file:
x<-read.csv("Muscular Data_Project_10csv.csv")
x<-as.data.frame(x) #converting into dataframe

# shuffling
rows <- sample(nrow(x))
x<-x[rows,]

#splitting
train_x<-x[1:146,]
test_x<-x[147:209,]
```

The estimated model was generated via glm function. Summary of model:



```
> fit<- glm(train_x$Carrier~train_x$creatine_Kinase+train_x$Hemopexin+train_x$Pyrovate_Kinase+
+              train_x$Age,
+            data = x, family = "binomial")
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit)

Call:
glm(formula = train_x$Carrier ~ train_x$creatine_Kinase + train_x$Hemopexin +
    train_x$Pyrovate_Kinase + train_x$Age, family = "binomial",
    data = x)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.52624  -0.37043  -0.17251   0.00088   2.39910

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -19.08406    4.01090  -4.758 1.95e-06 ***
train_x$creatine_Kinase   0.07196    0.02163   3.327 0.000877 ***
train_x$Hemopexin         0.07861    0.02937   2.677 0.007439 **
train_x$Pyrovate_Kinase   0.18309    0.06304   2.904 0.003681 **
train_x$Age               0.14693    0.05215   2.817 0.004843 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.793  on 141  degrees of freedom
Residual deviance:  62.295  on 137  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 72.295

Number of Fisher Scoring iterations: 9
```

Therefore the generated model is :

$$E(y\hat{}) = \frac{e^{( -19.08 + 0.0716x1 + 0.07861x2 + 0.18309x3 + 0.14693x4 )}}{(1 + e^{( -19.08 + 0.0716x1 + 0.07861x2 + 0.18309x3 + 0.14693x4 )})}$$

## Analysis of model:

### Effect of predictor variables

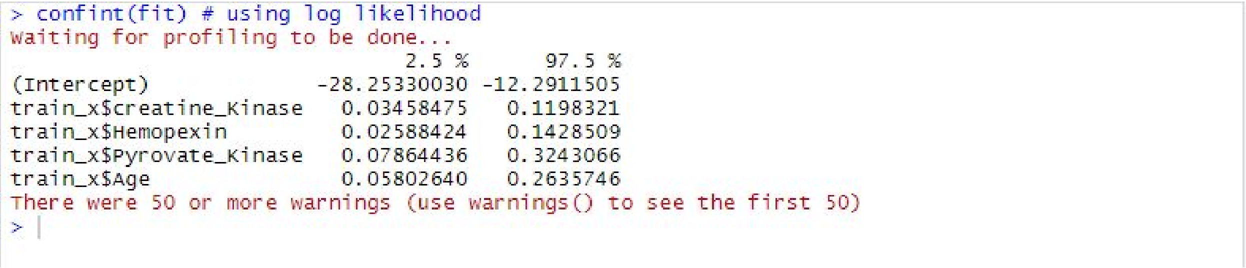
The logistic regression coefficients determine the change in the log odds of the outcome for a unit increase in the predictor variable.
Thus, by looking at model we can conclude :

- For one unit change in creatine kinase level, the log odds of the person being a carrier increases by **0.0716**.
- For one unit change in Hemopexin level, the log odds of the person being a carrier increases by **0.07861**.
- For one unit change in Pyruvate kinase level, the log odds increases by **0.18309**.
- For a unit change in a person's age, the log odds increases by **0.14693.**

Moreover, by going through the p-values of coefficients, it is observed that they are less than 0.05. **Thus it can be inferred that all the variables are statistically significant.**

### Confidence interval :

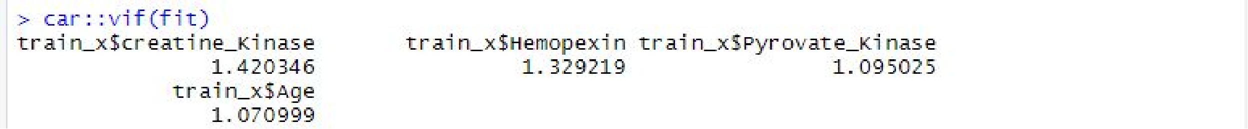
Generating 95% confidence interval using log likelihood:



```
> confint(fit) # using log likelihood
Waiting for profiling to be done...
                              2.5 %      97.5 %
(Intercept)            -28.25330030 -12.2911505
train_x$creatine_Kinase   0.03458475   0.1198321
train_x$Hemopexin         0.02588424   0.1428509
train_x$Pyrovate_Kinase   0.07864436   0.3243066
train_x$Age               0.05802640   0.2635746
There were 50 or more warnings (use warnings() to see the first 50)
> 
```

### Multicollinearity check:

Multicollinearity corresponds to a situation where the data contain highly correlated predictor variables.
Multicollinearity should be fixed by removing the concerned variables.
It can be assessed using the R function vif(), which computes the variance inflation factors:



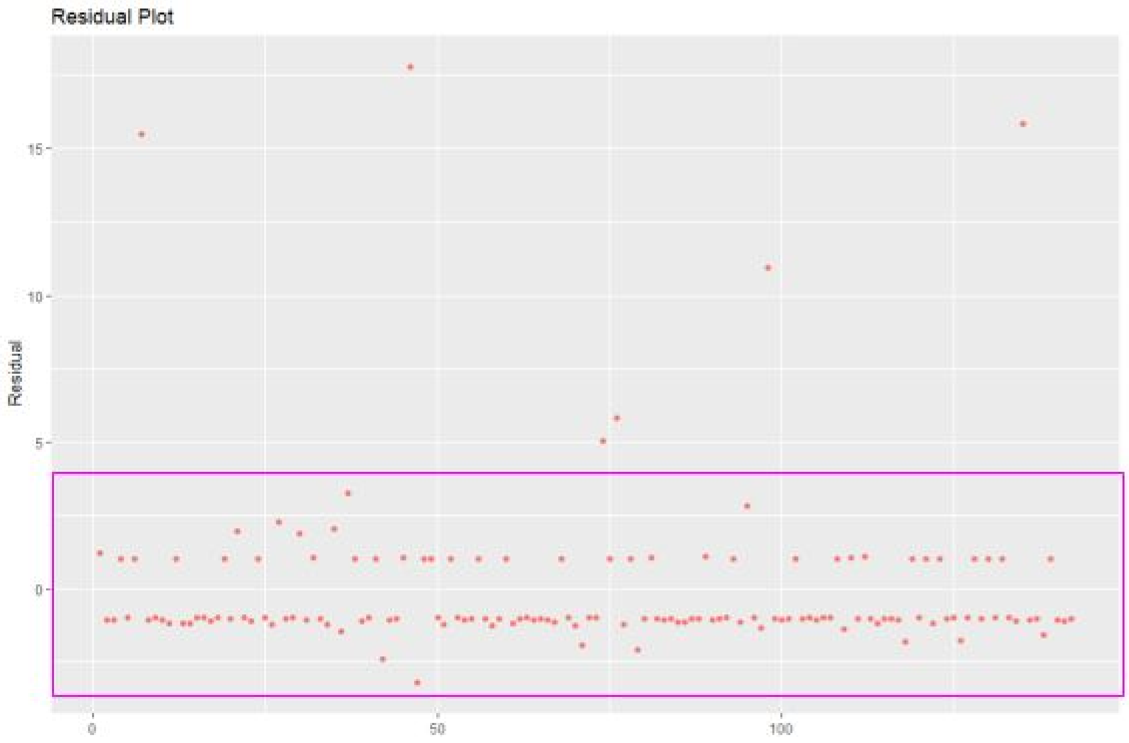
```
> car::vif(fit)
train_x$creatine_Kinase       train_x$Hemopexin train_x$Pyrovate_Kinase 
               1.420346                1.329219                1.095025 
            train_x$Age 
               1.070999 
```

As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a significant amount of collinearity.
But in our model all variables have VIF value well below 5. Thus the variables do not show any collinearity.

### Residual Analysis:

Residual is the error resulting from using the estimated regression equation to predict the value of the dependent variable.
If the assumption that the variance of ∈ is the same for all values of x and the assumed regression model is an adequate representation of the relationship between the variables is valid then the residual plot should give an overall impression of a horizontal band of points

From the summary it can be seen that the Residual deviance is 62.295. This implies that 62.3 % of variation in the data can be explained by the developed model.
To check for the validity of assumptions made for the error term, a residual plot is generated:



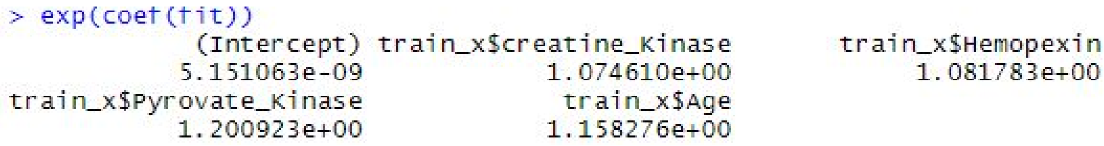
As we can see the residuals fall under a horizontal band. Thus our assumptions are valid.

## Odds ratio:

The odds ratio is the odds that y = 1 given that one of the independent variables has been increased by one unit (odds) divided by the odds that y = 1 given no change in the values for the other independent variables

It can be calculated by the formula:

**Odds Ratio = exp(coefficient)**

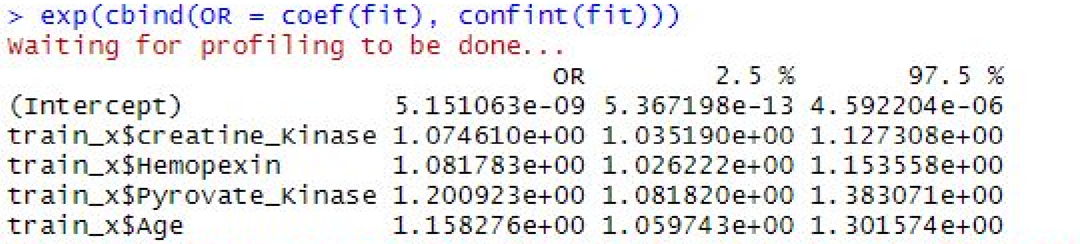


```
> exp(coef(fit))
          (Intercept) train_x$creatine_Kinase        train_x$Hemopexin
         5.151063e-09            1.074610e+00             1.081783e+00
train_x$Pyrovate_Kinase             train_x$Age
         1.200923e+00            1.158276e+00
```

Thus we can infer that :

- ➔ For one unit increase in creatine kinase values, the odds of the person being a carrier of DMD increases by **1.0764.**
- ➔ For a unit increase in pyruvate kinase values, the odds of the person being a carrier of DMD increases by **1.200923**.
- ➔ For one unit increase in age , the odds of the person being a carrier of DMD increases by **1.15827.**
- ➔ For one unit increase in Hemopexin values, the odds of the person being a carrier of DMD increases by **1.0817**

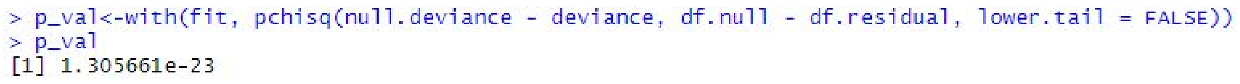
Odds ratio with 95% confidence interval:



```
> exp(cbind(OR = coef(fit), confint(fit)))
Waiting for profiling to be done...
                                  OR        2.5 %       97.5 %
(Intercept)             5.151063e-09 5.367198e-13 4.592204e-06
train_x$creatine_Kinase 1.074610e+00 1.035190e+00 1.127308e+00
train_x$Hemopexin       1.081783e+00 1.026222e+00 1.153558e+00
train_x$Pyrovate_Kinase 1.200923e+00 1.081820e+00 1.383071e+00
train_x$Age             1.158276e+00 1.059743e+00 1.301574e+00
```

## Testing for significance:

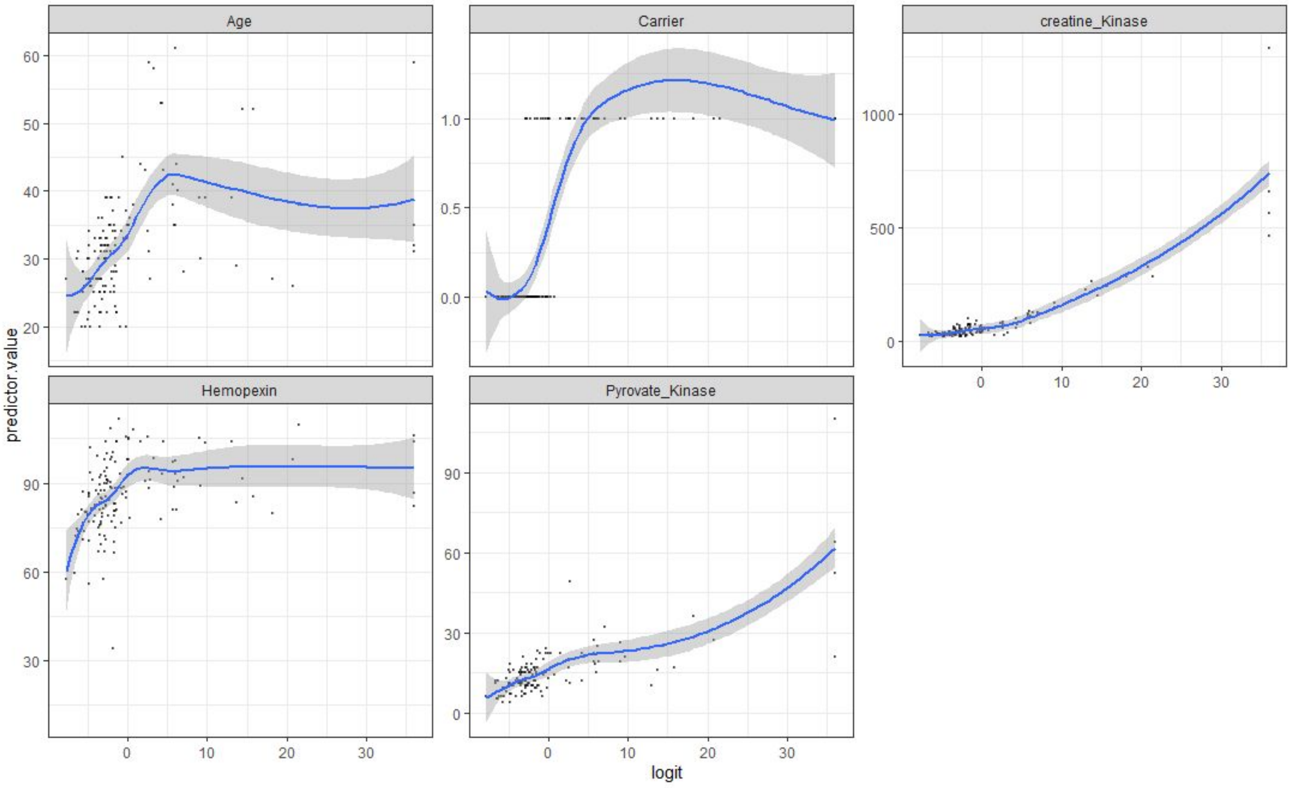
One measure of model fit is the significance of the overall model. This test asks whether the model with predictors fits significantly better than a null model (model with just an intercept).The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model ( i.e. no of extra variables).



```
> p_val<-with(fit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
> p_val
[1] 1.305661e-23
```

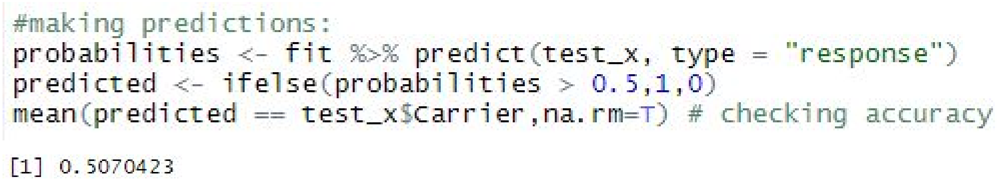
As p-value for the test of significance is way less than the significance level of 0.001. This implies that the model fits better than a null model.

### Linearity Assumption:



As seen from the graphs, the relation between logit of the outcome and predictor variables are nearly linear.

### Testing Accuracy:



```
#making predictions:
probabilities <- fit %>% predict(test_x, type = "response")
predicted <- ifelse(probabilities > 0.5,1,0)
mean(predicted == test_x$Carrier,na.rm=T) # checking accuracy

[1] 0.5070423
```

Thus the model is able to predict with an accuracy of nearly 51%

## Conclusion:

- The developed model is valid within the assumptions made.
- The model can explain 62% of variation in data and prediction accuracy is 51%.
- As the dataset is increased the prediction accuracy and Residual deviation is bound to increase.
- All the variables are statistically significant. This implies that all four variables - Age, Creatine Kinase levels, Pyruvate Kinase Levels and Hemopexin levels influence the dependent variable.
- By looking at the coefficients, it can be said that the pyruvate kinase levels are the most influential variable while creatine kinase levels is the least influential variable.
- There is no collinearity in the variables.
- The test for overall significance gives a p-value less than 0.001 indicating that the model is significant and fits better.