# CS-324 MACHINE LEARNING

COMPLEX ENGINEERING PROBLEM

**Maha Shoaib Khan (CS-21011)**
**Manahil Siddiqui (CS-21087)**
**Rafay Baig (CS-21060)**

Submitted To: Sir Syed Zafar Qasim

## Objective:

The primary objective of this report is to investigate the effect of gene polymorphism on renal dysfunction in children following liver transplantation. Utilizing the provided hypothetical dataset of 60 pediatric liver transplant recipients using the logistic regression model, the report aims to:

### Step1:

Filling missing values using group members' roll no. In mod 7. The members' roll numbers are 087,011,060 and for the fourth missing value taking the average of all three roll numbers i.e. 52.67 rounded off to 53.

1. 87 mod 7 = 3
2. 11 mod 7 = 4
3. 60 mod 7 = 4
4. 53 mod 7 = 4

(An updated Excel sheet is attached for reference).

### Step2:

Performance of all questioned tasks.

a) Using some suitable software system, produce the binary logistic model from the given training data.

Software Used: Minitab

The given Excel file was uploaded and recoded as follows:

**Summary**

| Original Value | Recoded Value | Number of Rows |
|---|---|---|
| Female | 0 | 30 |
| Male | 1 | 30 |

Recoded data column Sex

**Summary**

| Original Value | Recoded Value | Number of Rows |
|---|---|---|
| Leu/Leu | 0 | 21 |
| Leu/Pro | 1 | 24 |
| Pro/Pro | 2 | 15 |

Recoded data column Type

**Summary**

| Original Value | Recoded Value | Number of Rows |
|---|---|---|
| Absent | 0 | 14 |
| Absent | 0 | 18 |
| Present | 1 | 16 |
| Present | 1 | 12 |

Recoded data column Disease

Then binary logistic model was produced. The following results were obtained.

# Binary Logistic Regression: Disease versus TST, Sex, Type

## Method

| | |
|---|---|
| Link function | Logit |
| Categorical predictor coding | (1, 0) |
| Rows used | 60 |

## Response Information

| Variable | Value | Count | |
|---|---|---|---|
| Disease | 1 | 28 | (Event) |
| | 0 | 32 | |
| | Total | 60 | |

## Regression Equation

P(1) = exp(Y')/(1 +exp(Y'))

| Sex | Type | |
|---|---|---|
| 0 | 0 | Y' = -4.921 + 0.3604 TST |
| 0 | 1 | Y' = -3.031 + 0.3604 TST |
| 0 | 2 | Y' = -3.142 + 0.3604 TST |
| 1 | 0 | Y' = -4.827 + 0.3604 TST |
| 1 | 1 | Y' = -2.937 + 0.3604 TST |
| 1 | 2 | Y' = -3.048 + 0.3604 TST |

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -4.92 | 1.40 | -3.53 | 0.000 | |
| TST | 0.360 | 0.111 | 3.26 | 0.001 | 1.26 |
| Sex | | | | | |
| 1 | 0.094 | 0.638 | 0.15 | 0.883 | 1.05 |
| Type | | | | | |
| 1 | 1.890 | 0.815 | 2.32 | 0.020 | 1.69 |
| 2 | 1.779 | 0.903 | 1.97 | 0.049 | 1.66 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| TST | 1.4338 | (1.1544, 1.7808) |

## Odds Ratios for Categorical Predictors

| Level A | Level B | Odds Ratio | 95% CI |
|---|---|---|---|
| Sex | | | |
| 1 | 0 | 1.0986 | (0.3145, 3.8373) |
| Type | | | |
| 1 | 0 | 6.6179 | (1.3392, 32.7031) |
| 2 | 0 | 5.9215 | (1.0096, 34.7314) |
| 2 | 1 | 0.8948 | (0.2006, 3.9917) |

*Odds ratio for level A relative to level B*

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 25.26% | 20.43% | 71.97 | 73.08 | 82.44 | 0.8270 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 55 | 61.97 | 0.241 |
| Pearson | 55 | 57.74 | 0.374 |
| Hosmer-Lemeshow | 8 | 8.78 | 0.361 |

## Analysis of Variance

| Source | DF | Wald Test Chi-Square | P-Value |
|---|---|---|---|
| Regression | 4 | 12.68 | 0.013 |
| TST | 1 | 10.62 | 0.001 |
| Sex | 1 | 0.02 | 0.883 |
| Type | 2 | 5.87 | 0.053 |

## Fits and Diagnostics for Unusual Observations

| Obs | Observed Probability | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 29 | 1.000 | 0.113 | 2.089 | 2.18 | R |
| 30 | 1.000 | 0.154 | 1.933 | 2.03 | R |

b) Compute the odds ratio for all the predictor variables (using the probability method) and interpret them appropriately.

Using the equation computed by Minitab. Odd ratios are calculated.

$Odds_0$

Sex=0 Type=0 Y' = -4.921 + 0.3604 TST

TST=4

$$Y' = -4.921 + 0.3604(4)$$

$$= -3.4794$$

$$P(1) = \frac{1}{1+e^{y'}} = \frac{1}{1+e^{3.4794}} = 0.0299$$

$$Odds_0 = \frac{P(1)}{1-P(1)} = 0.0308$$

$Odds_1$

Sex=1 Type=0 Y' = -4.827 + 0.3604 TST
TST=4

$$Y' = -4.827 + 0.3604(4)$$

$$= -3.3854$$

$$P(1) = \frac{1}{1+e^{y'}} = \frac{1}{1+e^{3.3854}} = 0.0328$$

$$Odds_1 = \frac{P(1)}{1-P(1)} = 0.339$$

**Odds Ratio $_{\text{sex}}$:**

$$Odds\ Ratio = \frac{Odds_1}{Odds_0}$$

$$Odds\ Ratio_{\ sex} = 1.0995$$

**Interpretation of $\boldsymbol{Odds\ Ratio_{\ sex}}$:**

Males have approximately 9.95% higher odds of renal dysfunction than females, considering other features constant.

$Odds_2$

Sex=0 Type=0 Y' = -4.921 + 0.3604 TST

TST=4

$$Y' = -4.921 + 0.3604(4)$$

$$= -3.4794$$

$$P(1) = \frac{1}{1 + e^{y'}} = \frac{1}{1 + e^{3.4794}} = 0.0299$$

$$Odds_2 = \frac{P(1)}{1 - P(1)} = 0.0308$$

$Odds_3$

Sex=0 Type=1 Y' = -3.031 + 0.3604 TST
TST=4

$$Y' = -3.031 + 0.3604(4)$$

$$= -1.5894$$

$$P(1) = \frac{1}{1 + e^{y'}} = \frac{1}{1 + e^{1.5894}} = 0.1695$$

$$Odds_3 = \frac{P(1)}{1 - P(1)} = 0.2040$$

$\boldsymbol{Odds\ Ratio}_{\boldsymbol{type1vstype0}}$:

$$Odds\ Ratio = \frac{Odds_3}{Odds_2}$$

$$Odds\ Ratio_{type1} = 6.6249$$

**Interpretation of $\boldsymbol{Odds\ Ratio}_{\boldsymbol{type1vstpe0}}$:**

Individuals with the Leu/Pro genotype have approximately 6.62 times higher odds of renal dysfunction compared to those with the Leu/Leu genotype, holding all other features constant.

$Odds_4$

Sex=0 Type=0 Y' = -4.921 + 0.3604 TST

TST=4

$$Y' = -4.921 + 0.3604(4)$$

$$= -3.4794$$

$$P(1) = \frac{1}{1 + e^{y'}} = \frac{1}{1 + e^{3.4794}} = 0.0299$$

$$Odds_2 = \frac{P(1)}{1 - P(1)} = 0.0308$$

$Odds_5$

Sex=0 Type=2 Y' = -3.142 + 0.3604 TST

TST=4

$$Y' = -3.142 + 0.3604(4)$$

$$= -1.7004$$

$$P(1) = \frac{1}{1+e^{y'}} = \frac{1}{1+e^{1.5894}} = 0.1544$$

$$Odds_3 = \frac{P(1)}{1-P(1)} = 0.1826$$

$\boldsymbol{Odds\ Ratio}\ _{\boldsymbol{type2vstpye0}}$

$$Odds\ Ratio = \frac{Odds_5}{Odds_4}$$

$$Odds\ Ratio\ _{type1} = 5.9286$$

**Interpretation of $\boldsymbol{Odds\ Ratio}\ _{\boldsymbol{type1vstpe0}}$:**

Individuals with the Pro/Pro genotype have approximately 5.93 times higher odds of renal dysfunction compared to those with the Leu/Leu genotype, holding all other variables

$Odds_6$

Sex=0 Type=0 Y' = -4.921 + 0.3604 TST

TST=4

$$Y' = -4.921 + 0.3604(4)$$

$$= -3.4794$$

$$P(1) = \frac{1}{1+e^{y'}} = \frac{1}{1+e^{3.4794}} = 0.0299$$

$$Odds_4 = \frac{P(1)}{1-P(1)} = 0.0308$$

$Odds_7$

Sex=0 Type=0 Y' = -4.921 + 0.3604 TST

TST=5

$$Y' = -4.921 + 0.3604(5)$$

$$= -3.1190$$

$$P(1) = \frac{1}{1+e^{y'}} = \frac{1}{1+e^{3.4794}} = 0.0423$$

$$Odds_0 = \frac{P(1)}{1 - P(1)} = 0.0442$$

***Odds Ratio*** $_{TST}$**:**

$$Odds\ Ratio = \frac{Odds_7}{Odds_6}$$

$$Odds\ Ratio\ _{type1} = 1.4351$$

**Interpretation of *Odds Ratio*** $_{TST}$**:**

For each additional year since transplant (TST), the odds of renal dysfunction increase by approximately 43.51%. This means that each extra year post-transplant raises the odds of developing renal dysfunction by 1.4351 times compared to the previous year.

c) Re-compute the odds ratios in sec (b) using the exponential formulas.

$$Odd\ Ratios = e^{bi}$$

$$Odd\ Ratio_{sex} = e^{0.094} = 1.0984$$

$$Odd\ Ratio_{type1vstype0} = e^{1.890} = 6.6194$$

$$Odd\ Ratio_{type2vstype0} = e^{1.779} = 5.9239$$

$$Odd\ Ratio_{TST} = e^{0.360} = 1.4333$$

d) Compare the odds in favor of the patients having three years since transplant with the odds in favor of the patients having seven years since transplant. Also, interpret it properly.

Order Ratio for more than 1 unit change is calculated as,

$$Odd\ Ratios = e^{cbi}$$

$$c = change \qquad bi = coefficient$$

$$TST_0 = 3 \qquad TST_1 = 7$$

$$c = TST_1 - TST_0$$

$$c = 4$$

$$Odd\ Ratio = e^{4(0.360)}$$

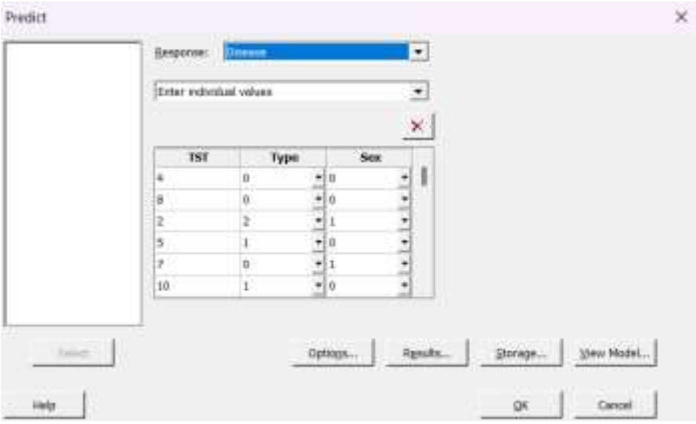
$$Odd\ Ratio = 4.2207$$

## Interpretation:

This means that the odds of renal dysfunction are approximately 4.22 times higher for patients who have been transplanted for 7 years compared to those who have been transplanted for 3 years.

e) Evaluate the performance of the model in (a) from the given test data (see Excel sheet).

The given test data was fed to Minitab for predictions.




The following predictions were made by Minitab

### Prediction for Disease

#### Regression Equation

P(1) = exp(Y')/(1 +exp(Y'))
Y' = -4.92 + 0.360 TST + 0.000000 Sex_0 + 0.094 Sex_1 + 0.000000 Type_0 + 1.890 Type_1
+ 1.779 Type_2

#### Settings

| Variable | Setting |
|---|---|
| TST | 4 |
| Type | 0 |
| Sex | 0 |

#### Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.0298979 | 0.0295554 | (0.0041651, 0.185067) |

#### Settings

| Variable | Setting |
|---|---|
| TST | 8 |
| Type | 0 |
| Sex | 0 |

#### Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.115250 | 0.0735573 | (0.0307066, 0.348799) |

#### Settings

| Variable | Setting |
|---|---|
| TST | 2 |
| Type | 2 |
| Sex | 1 |

#### Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.0888552 | 0.0841557 | (0.0125547, 0.427915) |

## Settings

| Variable | Setting |
|---|---|
| TST | 5 |
| Type | 1 |
| Sex | 0 |

## Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.226273 | 0.118737 | (0.0718405, 0.524928) |

## Settings

| Variable | Setting |
|---|---|
| TST | 7 |
| Type | 0 |
| Sex | 1 |

## Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.0907493 | 0.0748122 | (0.0166011, 0.371101) |

## Settings

| Variable | Setting |
|---|---|
| TST | 10 |
| Type | 1 |
| Sex | 0 |

## Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.639289 | 0.139132 | (0.351999, 0.852560) |

## Settings

| Variable | Setting |
|---|---|
| TST | 5 |
| Type | 0 |
| Sex | 1 |

## Prediction

| Fitted Probability | SE Fit | 95% CI |
|---|---|---|
| 0.0462993 | 0.0470099 | (0.0059887, 0.281190) |

By comparing the given test data and predictions done by the software we have TP=0, TN=3, FN=3, FP=1. Considering the threshold to be 0.5, predictions lesser than the threshold are considered "Absent" and values greater than the threshold are "Present"

***Accuracy***:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{0 + 3}{0 + 3 + 3 + 1}$$

$$Accuracy = 0.4286$$

***Precision:***

$$Precision = \frac{TP}{TP + FP}$$

$$Precision = \frac{0}{0 + 1}$$

$$Precision = 0$$

***Recall*:**

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{0}{0 + 3}$$

$$Recall = 0$$

***Error Rate*:**

$$Error = 1 - Accuracy$$

$$Error = 1 - 0.4286$$

$$Error = 0.5714$$

***Specificity*:**

$$Specificity = \frac{TN}{TN + FP}$$

$$Specificity = \frac{3}{3 + 1}$$

$$Specifity = 0.75$$

***F1 − Score*:**

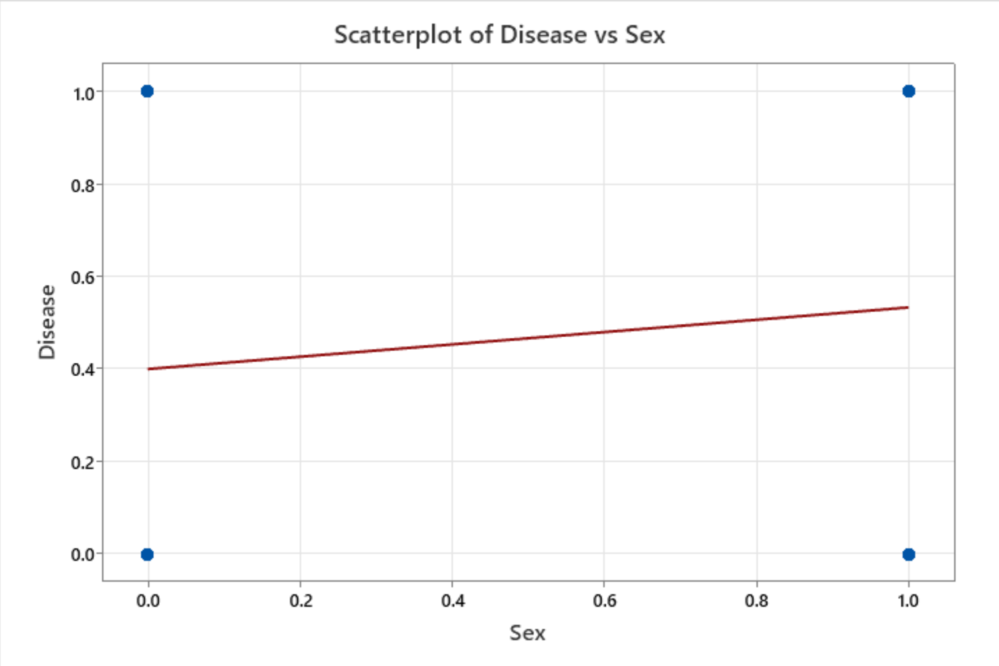
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

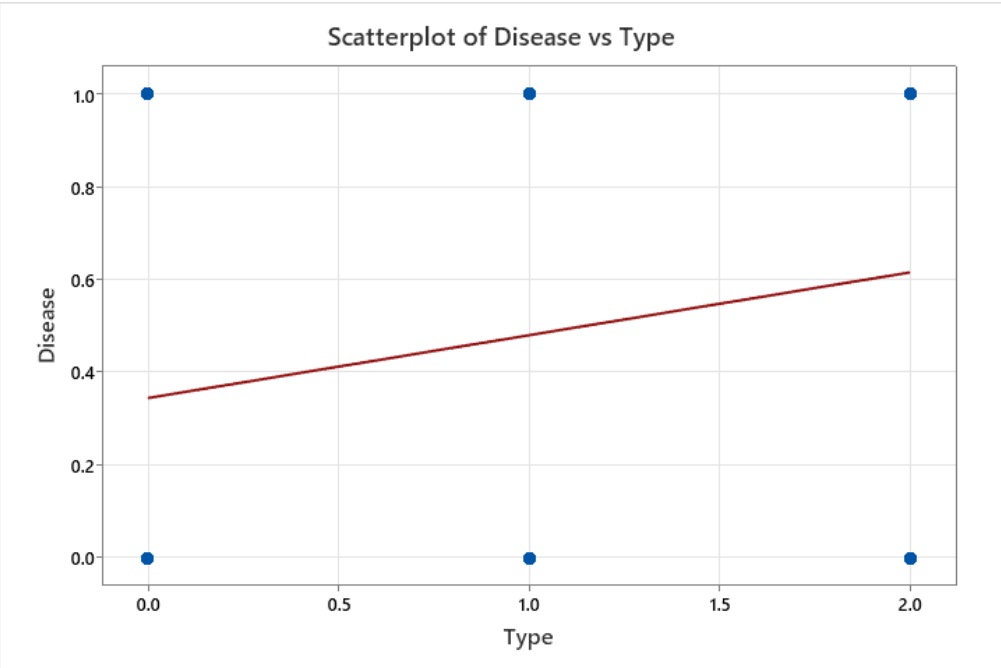
$$Cannot\ be\ computed$$

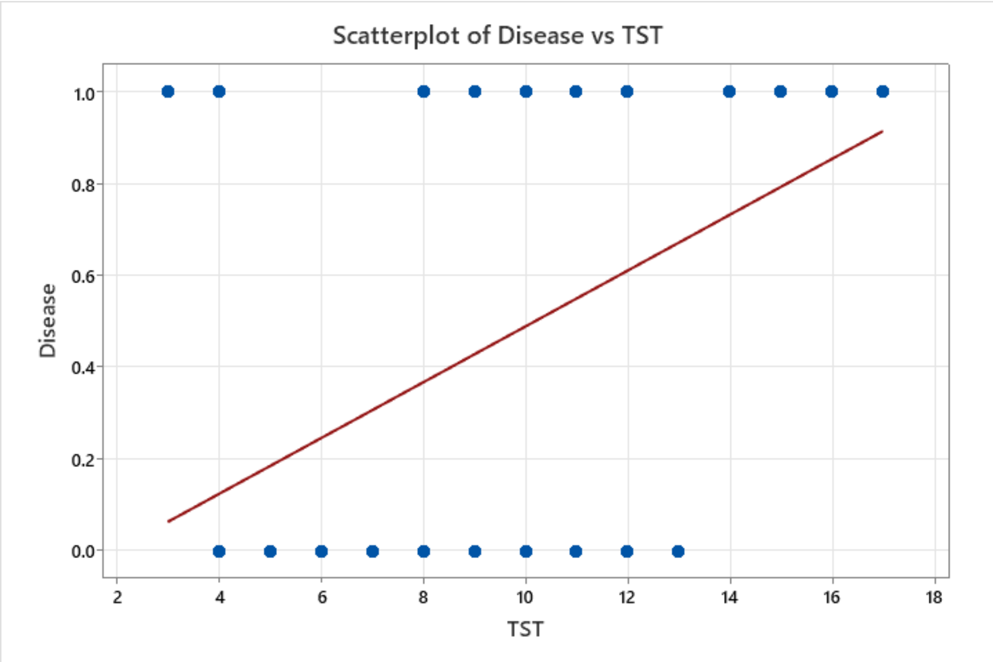
According to the model's performance measures, it can accurately identify diseases that are absent (high specificity), but it has trouble predicting diseases that are present (low sensitivity). This implies that to increase the model's ability to forecast the presence of disease, it might be necessary to change the threshold or take into account new variables.

## Scatterplots to visualize regression trends for the given data model:

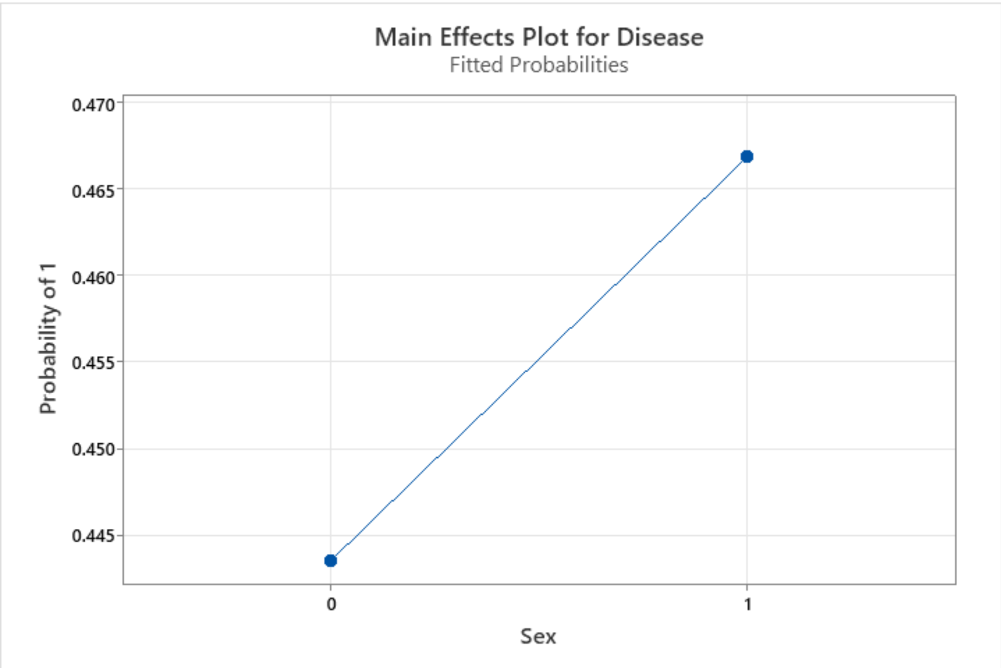
**Scatterplot of Disease vs Sex, Type, TST**

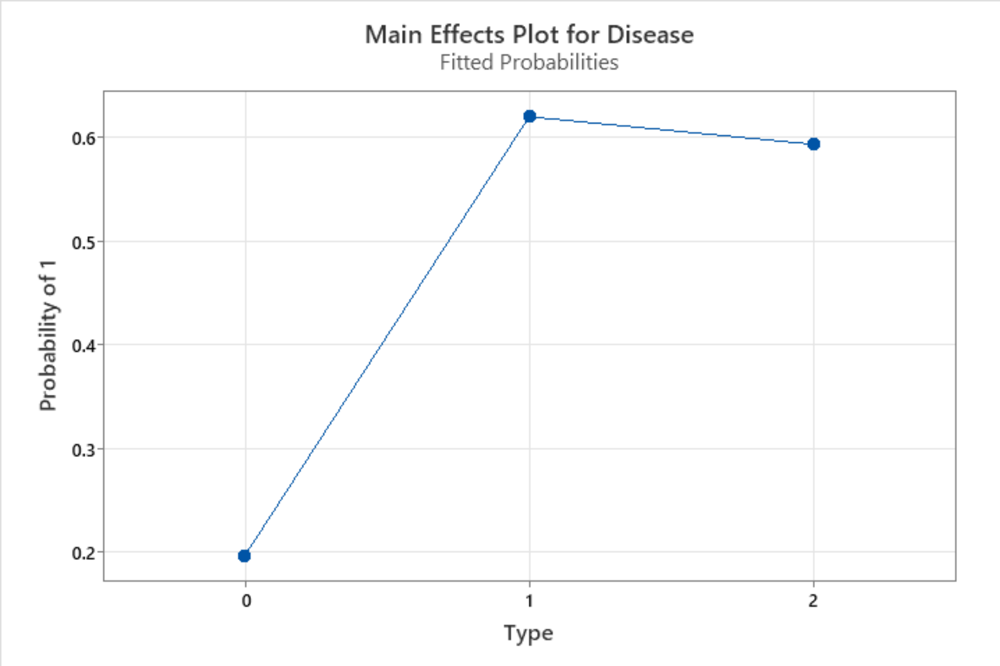

Scatterplot of Disease vs Sex
Disease
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Sex


Scatterplot of Disease vs Type
Disease
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Type

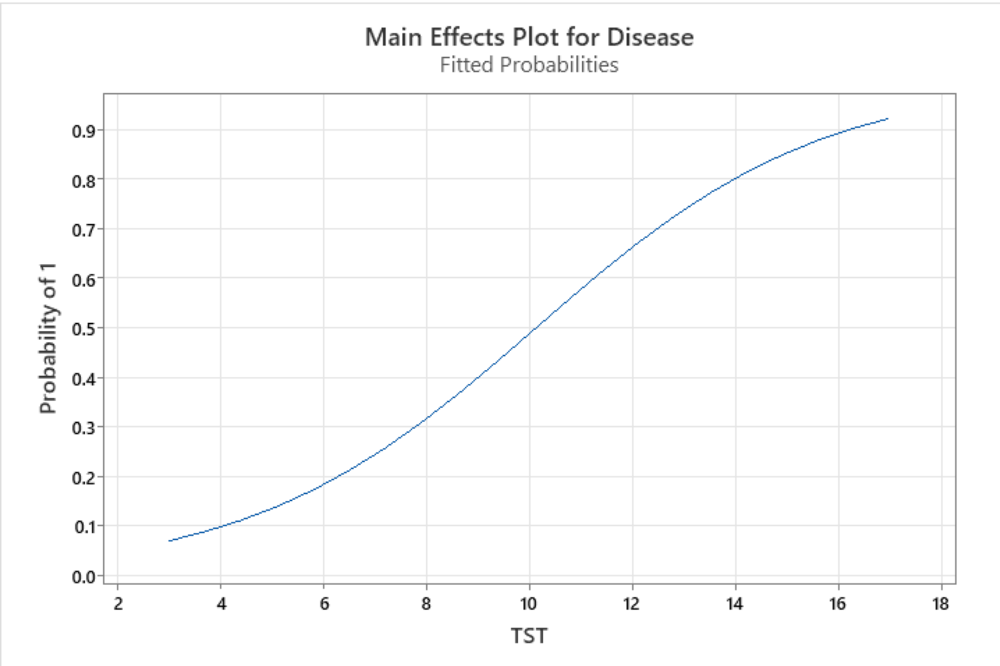



## The diagnostic plot of the model:





Main Effects Plot for Disease
Fitted Probabilities
Probability of 1
0.2
0.3
0.4
0.5
0.6
0
1
2
Type

Main Effects Plot for Disease
Fitted Probabilities
Probability of 1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
2
4
6
8
10
12
14
16
18
TST