STAT3500 Assignment 1

Name: Chee Kitt Win

Student Number: 45589140

Importing Libraries and EDA

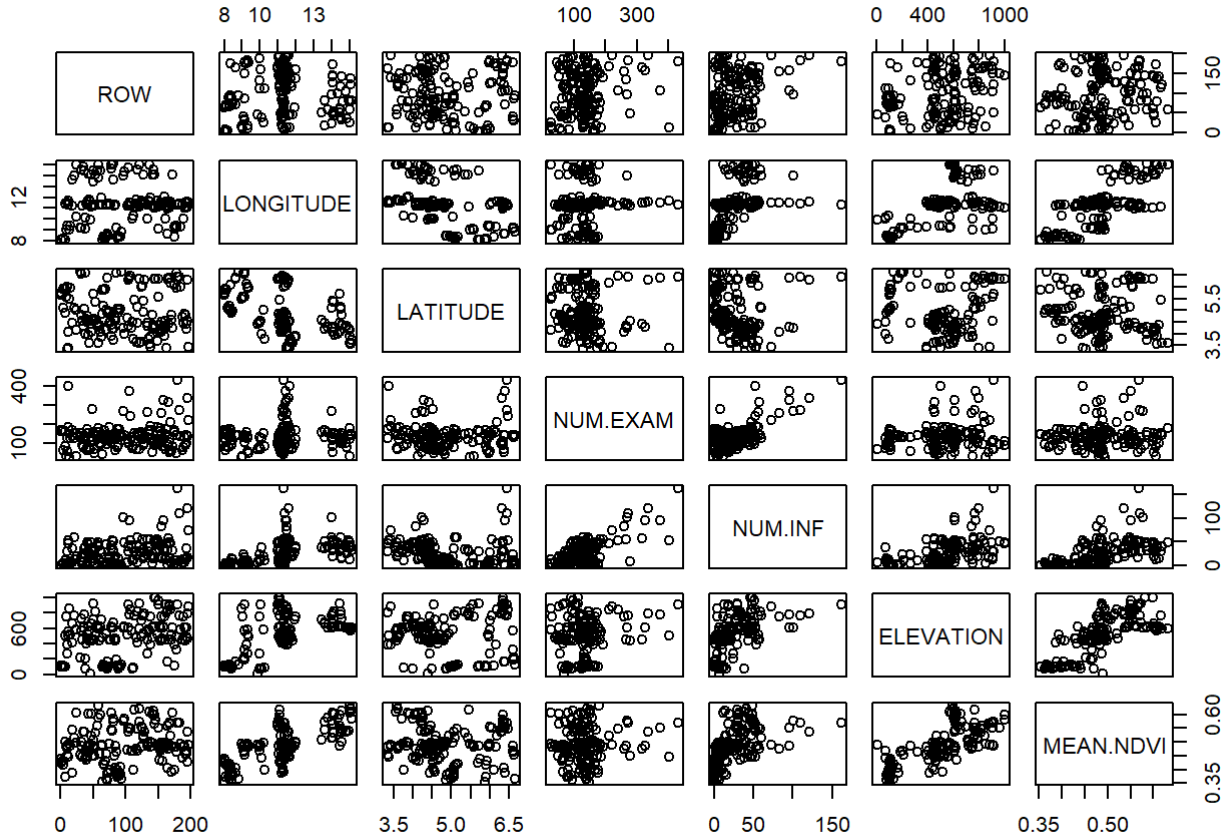
```
parasite = read.csv("C:\\Users\\Owner\\Desktop\\UQ Year 3 Sem 2 Courses\\STAT3500\\Assignment
1\\parasite.csv")
cholest = read.csv("C:\\Users\\Owner\\Desktop\\UQ Year 3 Sem 2 Courses\\STAT3500\\Assignment
 1\\cholest.csv")
library(lattice)
summary(parasite)
```

```
##       ROW           LONGITUDE         LATITUDE        NUM.EXAM
##  Min.   :  1.00   Min.   : 8.004   Min.   :3.350   Min.   : 26.0
##  1st Qu.: 51.25   1st Qu.: 9.969   1st Qu.:4.254   1st Qu.: 95.0
##  Median : 93.50   Median :11.303   Median :4.609   Median :132.0
##  Mean   : 99.39   Mean   :11.268   Mean   :4.898   Mean   :136.2
##  3rd Qu.:150.75   3rd Qu.:11.684   3rd Qu.:5.705   3rd Qu.:155.5
##  Max.   :197.00   Max.   :15.136   Max.   :6.650   Max.   :432.0
##     NUM.INF        ELEVATION        MEAN.NDVI
##  Min.   :  0.0   Min.   :   4.0   Min.   :0.3535
##  1st Qu.:  7.0   1st Qu.: 411.2   1st Qu.:0.4497
##  Median : 20.0   Median : 531.0   Median :0.4848
##  Mean   : 27.2   Mean   : 512.3   Mean   :0.4893
##  3rd Qu.: 42.5   3rd Qu.: 664.5   3rd Qu.:0.5378
##  Max.   :162.0   Max.   :1006.0   Max.   :0.6327
```

```
head(parasite)
```

```
##   ROW LONGITUDE LATITUDE NUM.EXAM NUM.INF ELEVATION MEAN.NDVI
## 1   1   8.04186  5.73675      162       0       108 0.4389815
## 2   2   8.00433  5.68028      167       1        99 0.4258333
## 3   4   8.10072  5.91742       62       5       104 0.4324074
## 4   5   8.18251  5.10454      167       3       109 0.4150000
## 5   7  11.36000  4.88500      163      11       503 0.5019444
## 6   8   8.06749  5.89780       83       0       103 0.3731481
```

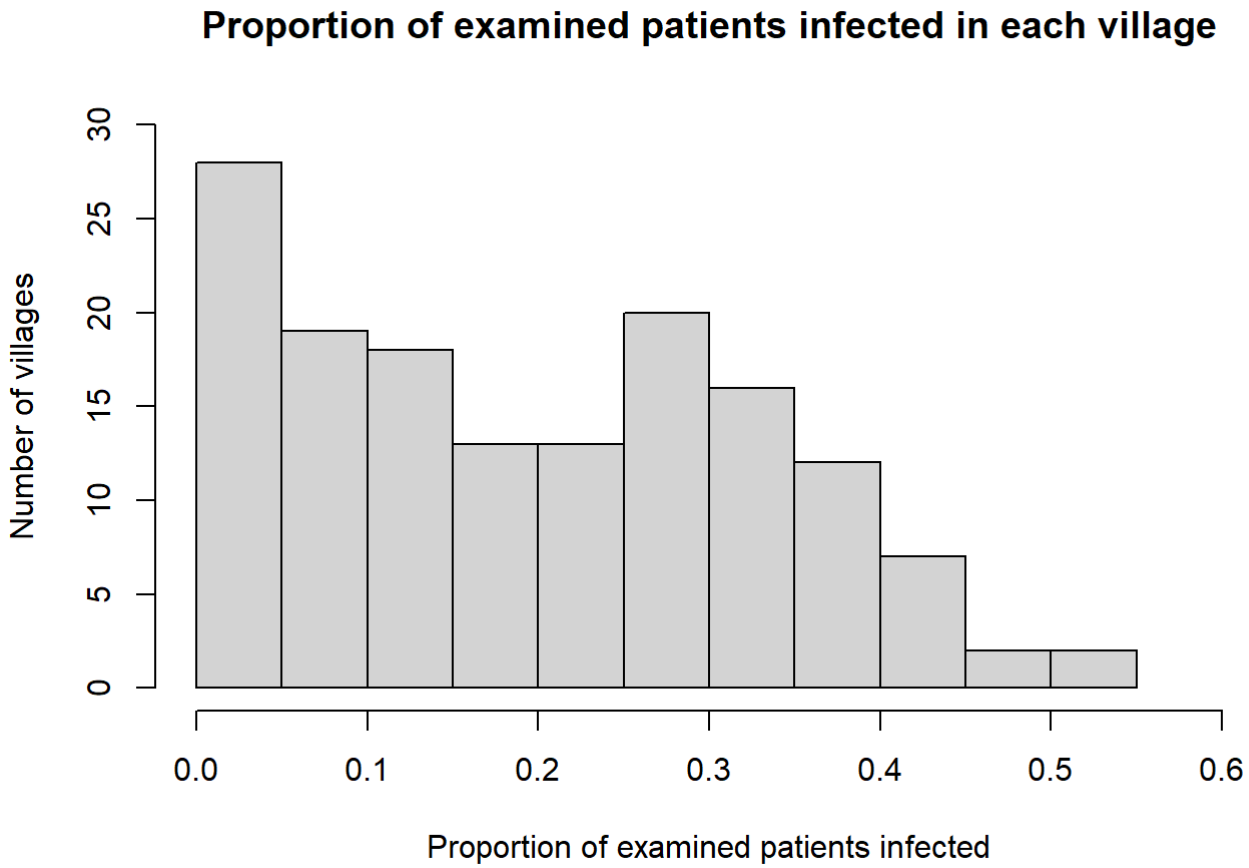
```
plot(parasite)
```



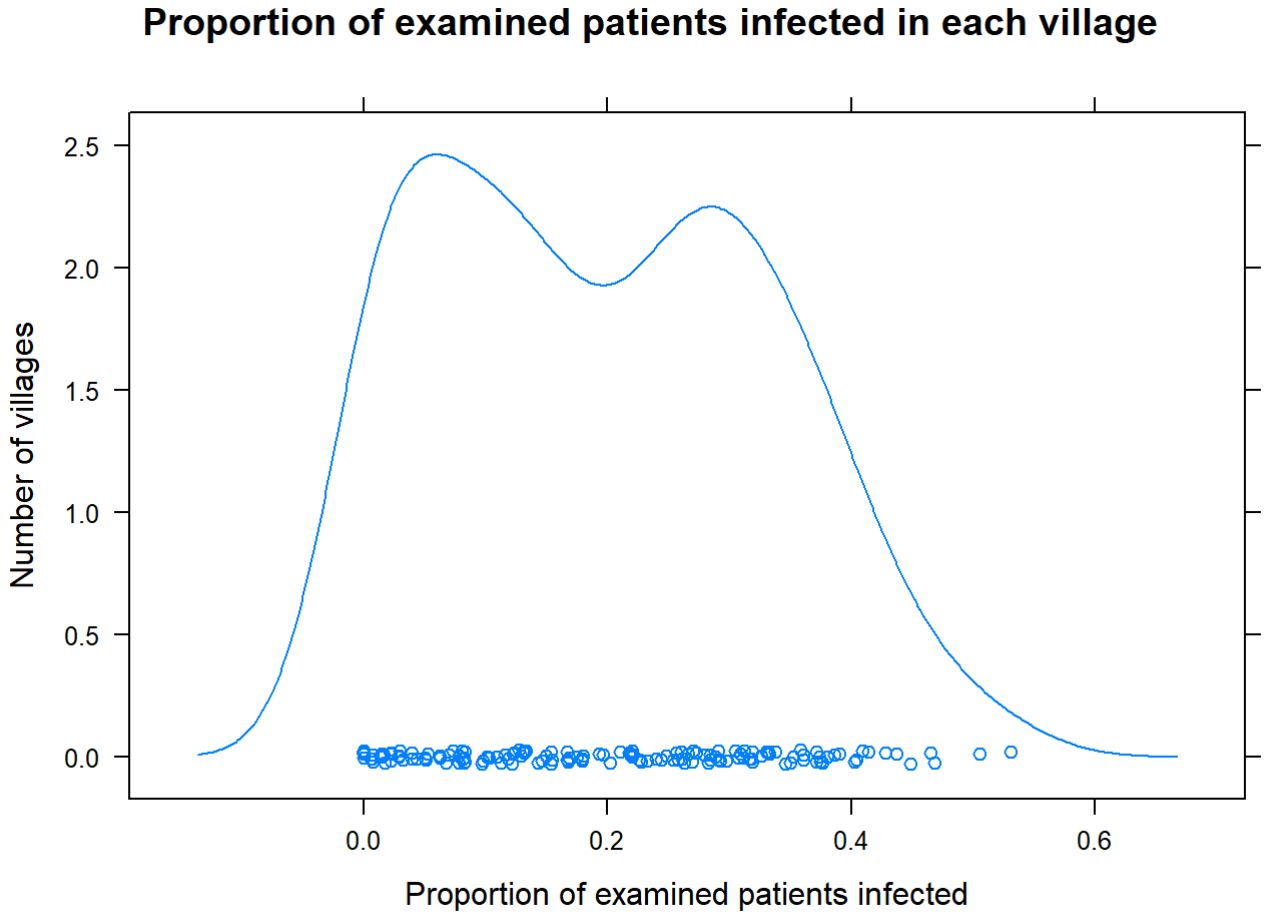
4.a)

Below are some graphical and numerical summaries of the new variable I named INF_PROPORTION

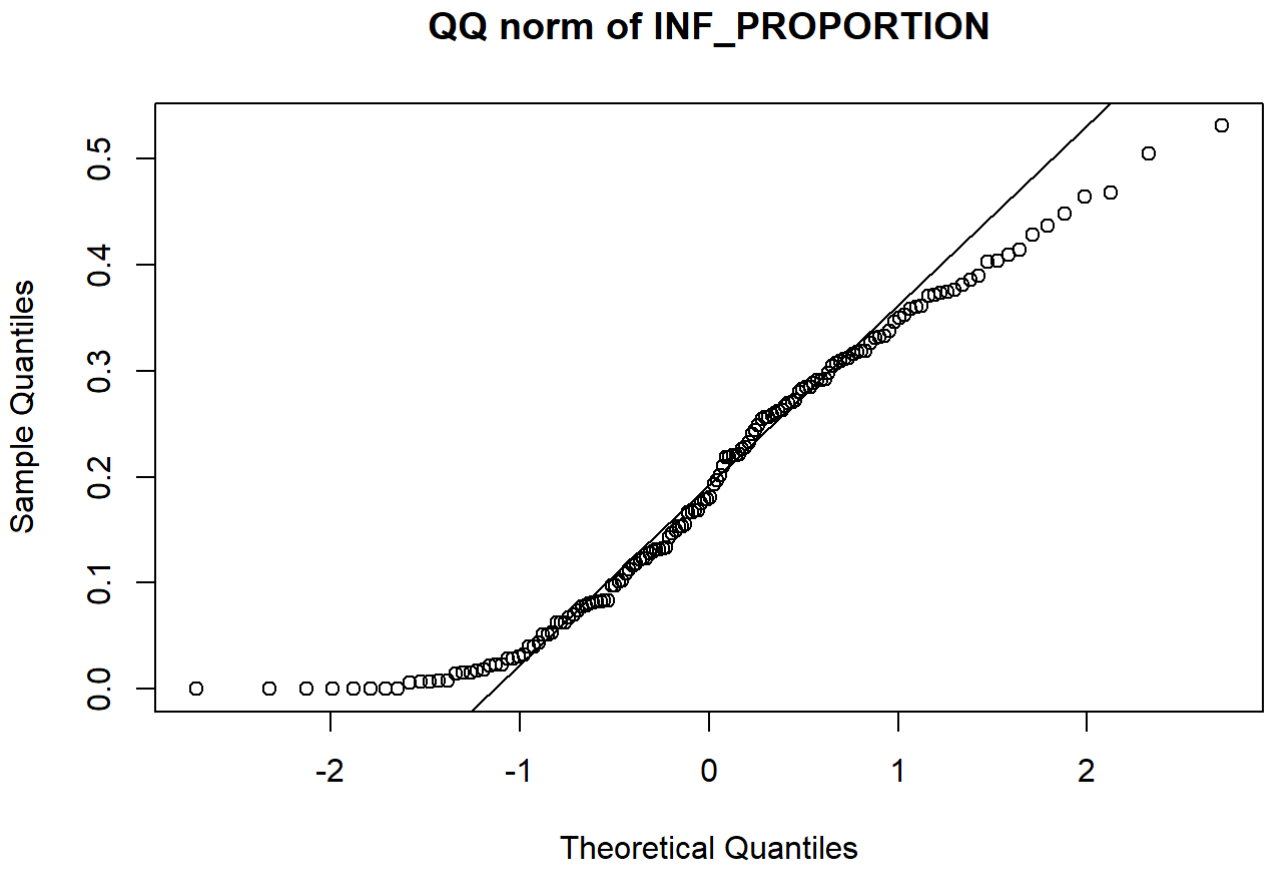
```
# Create the new variable

parasite$INF_PROPORTION = parasite$NUM.INF/parasite$NUM.EXAM

# Graphical summaries of INF_PROPORTION

hist(parasite$INF_PROPORTION, xlab = "Proportion of examined patients infected", ylab = "Numb
er of villages", main ="Proportion of examined patients infected in each village",ylim = c(0,
30), xlim = c(0,0.6))
```



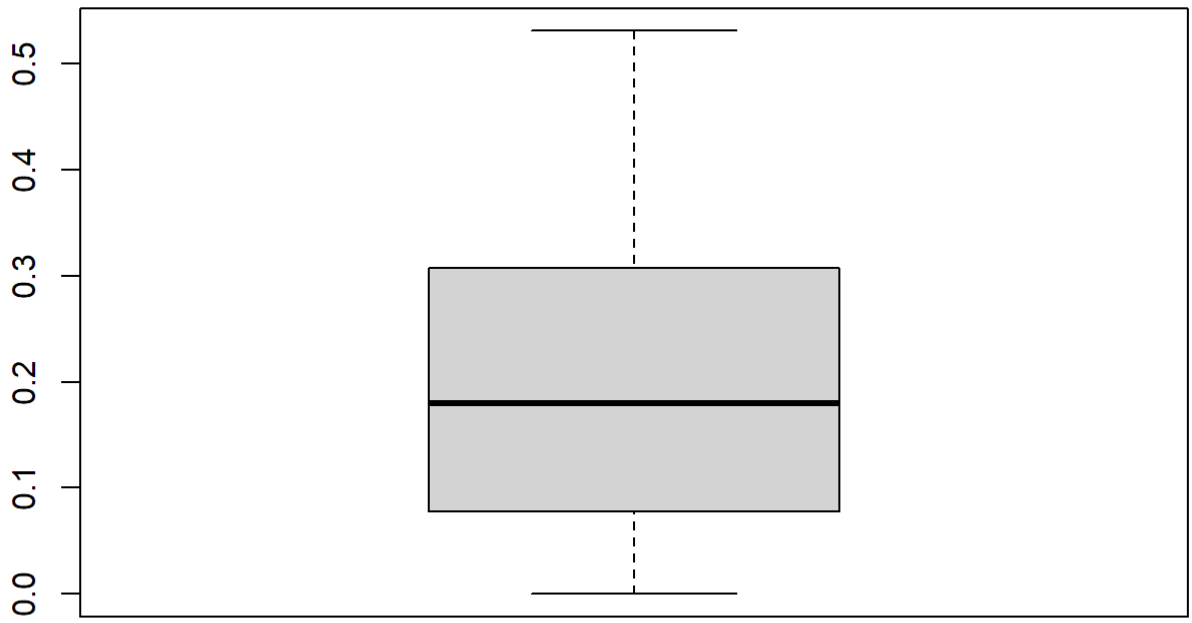

```
densityplot(parasite$INF_PROPORTION, xlab = "Proportion of examined patients infected", ylab
 = "Number of villages", main ="Proportion of examined patients infected in each village")
```



```
qqnorm(parasite$INF_PROPORTION, main="QQ norm of INF_PROPORTION");
qqline(parasite$INF_PROPORTION)
```




```
boxplot(parasite$INF_PROPORTION, main="Boxplot of INF_PROPORTION")
```



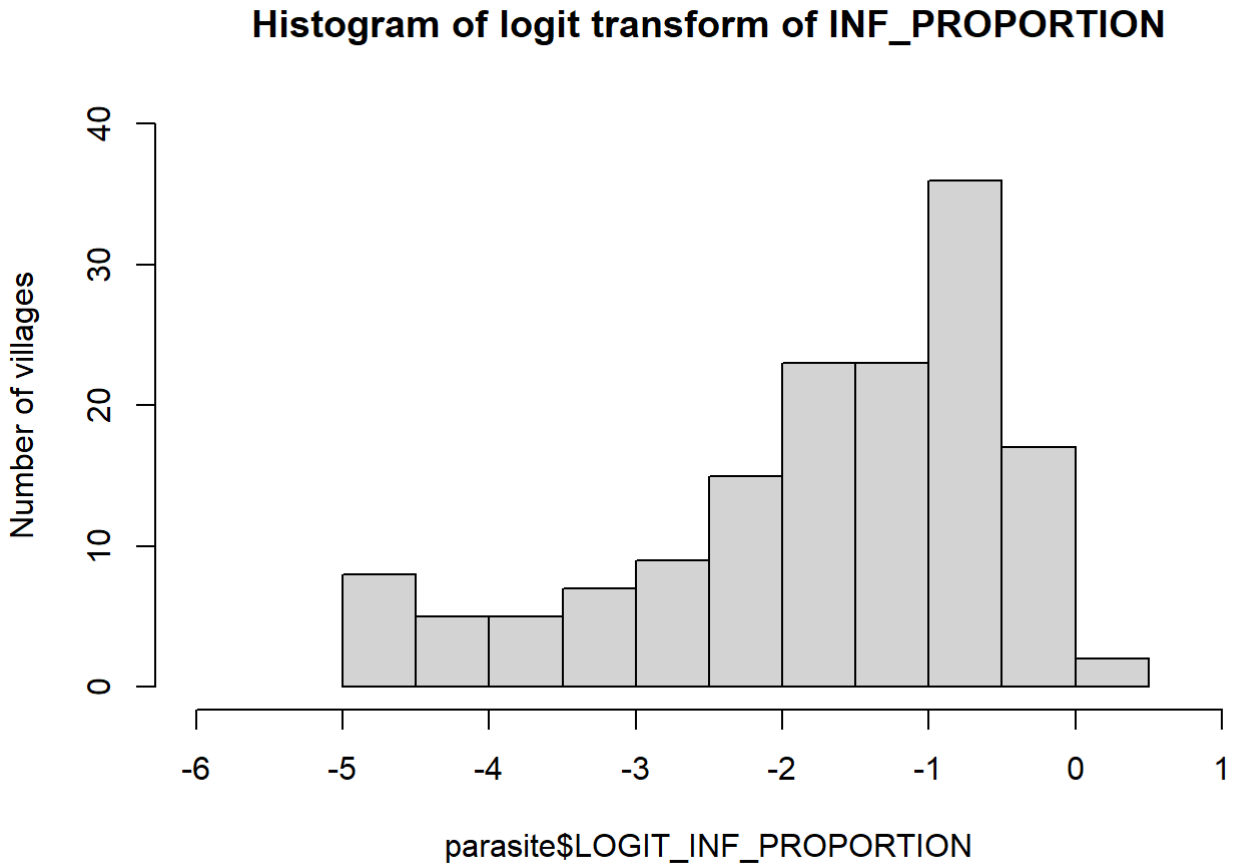

```
# Numerical summary of INF_PROPORTION

summary(parasite$INF_PROPORTION)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.07817 0.17985 0.19525 0.30704 0.53125
```

From the barchart, we can see that the data is skewed to the right. The data seem to be centred around 2 “peaks” at 0 and 0.3, and this is more pronounced when looking at the density plot. As expected, the data doesn’t look very normal from the QQ plot, especially at the tails.The boxplot and numerical summaries tell us that the mean is 0.19525, which is somewhere between the 2 “peaks” as expected.

I decided to use a logit transformation on INF_PROPORTION to make sure that it is bounded between 0 and 1 since it is a proportion.Since some of the INF_PROPORTION values are 0, I added a small value of 0.01 before transforming. Below is the corresponding barchart.

```
parasite$LOGIT_INF_PROPORTION = log((parasite$INF_PROPORTION+0.01)/(1-(parasite$INF_PROPORTIO
N+0.01)))
hist(parasite$LOGIT_INF_PROPORTION, ylab = "Number of villages", main ="Histogram of logit tr
ansform of INF_PROPORTION", xlim = c(-6,1), ylim = c(0,40))
```



4.b)

Fitting the kitchen sink linear model, we see that the R squared and adjusted R squared values are almost 0.7 which indicates high linear correlation, so our approximate linear relationship assumption for linear regression is satisfied. The p values of all the corresponding explanatory variables excluding longitude are very small which indicates that after performing a T test on the corresponding coefficients, there is very strong evidence to reject the null hypothesis (that the relevant coefficient is 0), indicating that there is a relationship between the corresponding explanatory variable and the response variable. The residual error is 0.7003 which intuitively is very small when you look at the histogram of LOGIT_INF_PROPORTION which has values ranging from -5 to 1. Overall, all these facts indicate a good fit to the model.

```
parasite.lm.full = lm(LOGIT_INF_PROPORTION ~ ELEVATION + LATITUDE + LONGITUDE + MEAN.NDVI, da
ta=parasite)
summary(parasite.lm.full)
```

```
##
## Call:
## lm(formula = LOGIT_INF_PROPORTION ~ ELEVATION + LATITUDE + LONGITUDE +
##     MEAN.NDVI, data = parasite)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4306 -0.4319 -0.0358  0.4859  1.9980
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.6594015  0.7444305  -6.259 4.13e-09 ***
## ELEVATION    0.0022288  0.0003396   6.563 8.76e-10 ***
## LATITUDE    -0.3247672  0.0852953  -3.808 0.000207 ***
## LONGITUDE    0.0365728  0.0585867   0.624 0.533443
## MEAN.NDVI    6.0795968  1.6109702   3.774 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7003 on 145 degrees of freedom
## Multiple R-squared:  0.6801, Adjusted R-squared:  0.6712
## F-statistic: 77.06 on 4 and 145 DF,  p-value: < 2.2e-16
```

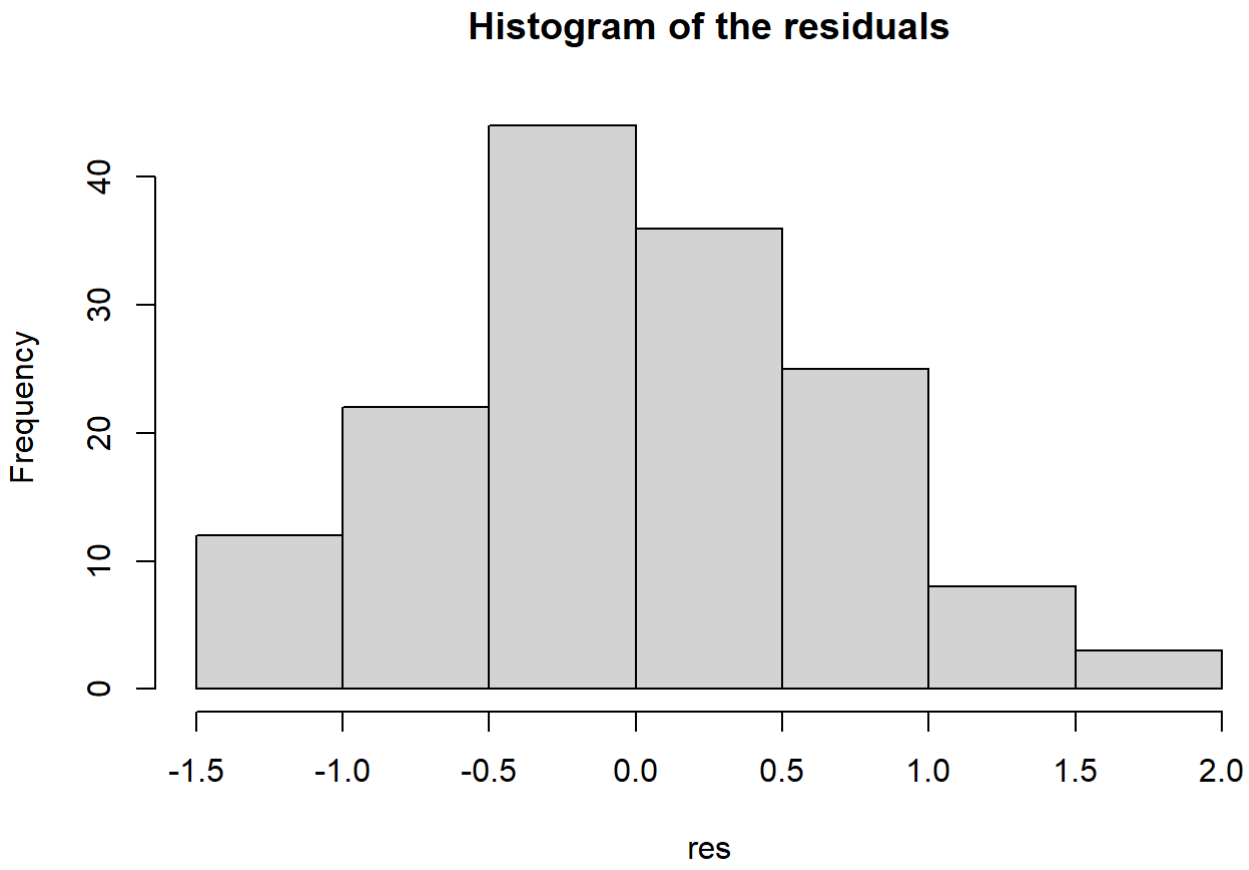
4.c)

The insignificant covariate here is LONGITUDE. After removing it and refitting the model, we find that the R squared values and residual standard errors have been almost unchanged. What has changed is that the p-values have become even smaller which shows even stronger evidence that they affect the response. As with before, the approximate linear relationship assumption is satisfied. From the bar chart, we see that our normally distributed residuals assumption is also satisfied. This can also be seen from the QQ plot where the line is pretty straight. The constant variance for the residuals is the only condition that is slightly questionable, but it does not seem to be that bad from the residual plot, so overall, the fit is good.

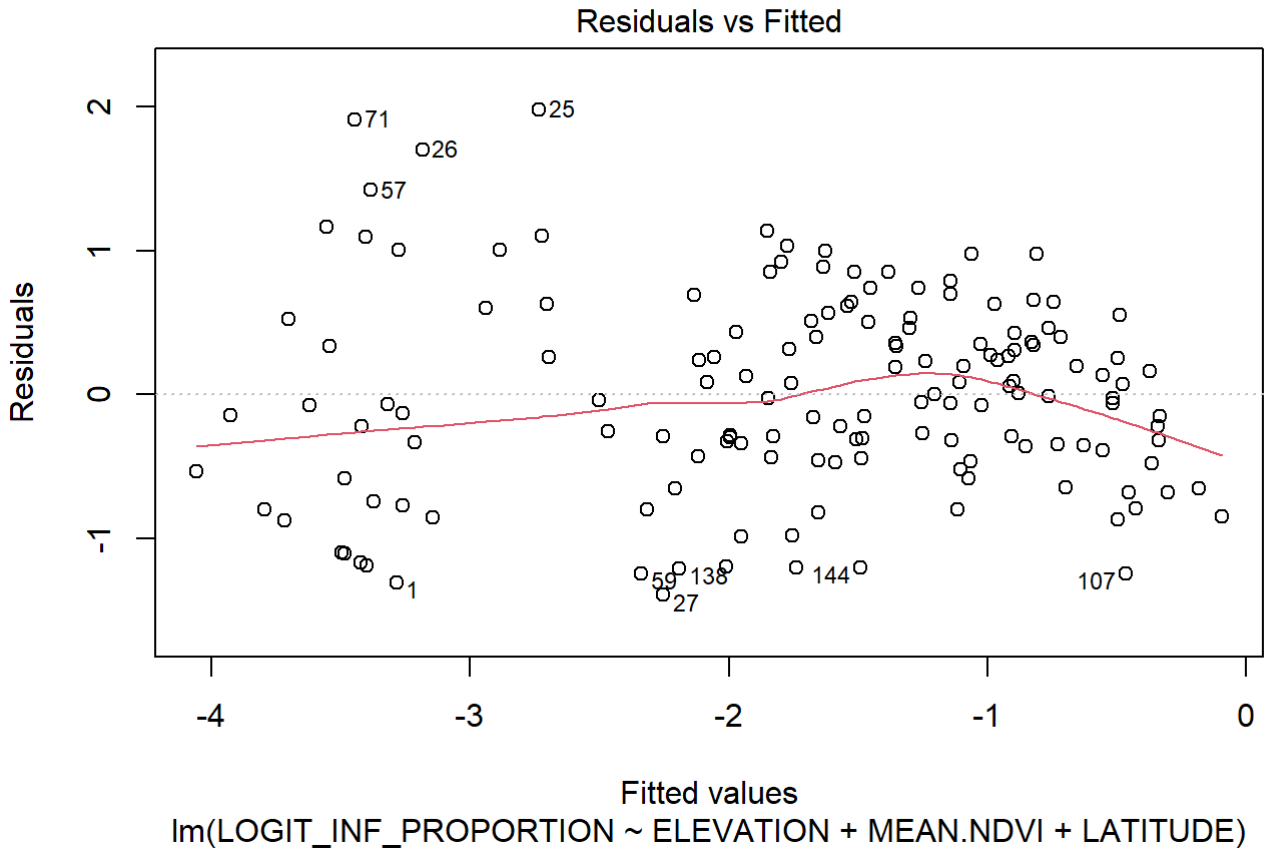
```
# 4(c) Remove LONGITUDE and refit model
parasite.lm.refit = lm(LOGIT_INF_PROPORTION ~ ELEVATION + MEAN.NDVI + LATITUDE, data=parasit
e)
summary(parasite.lm.refit)
```

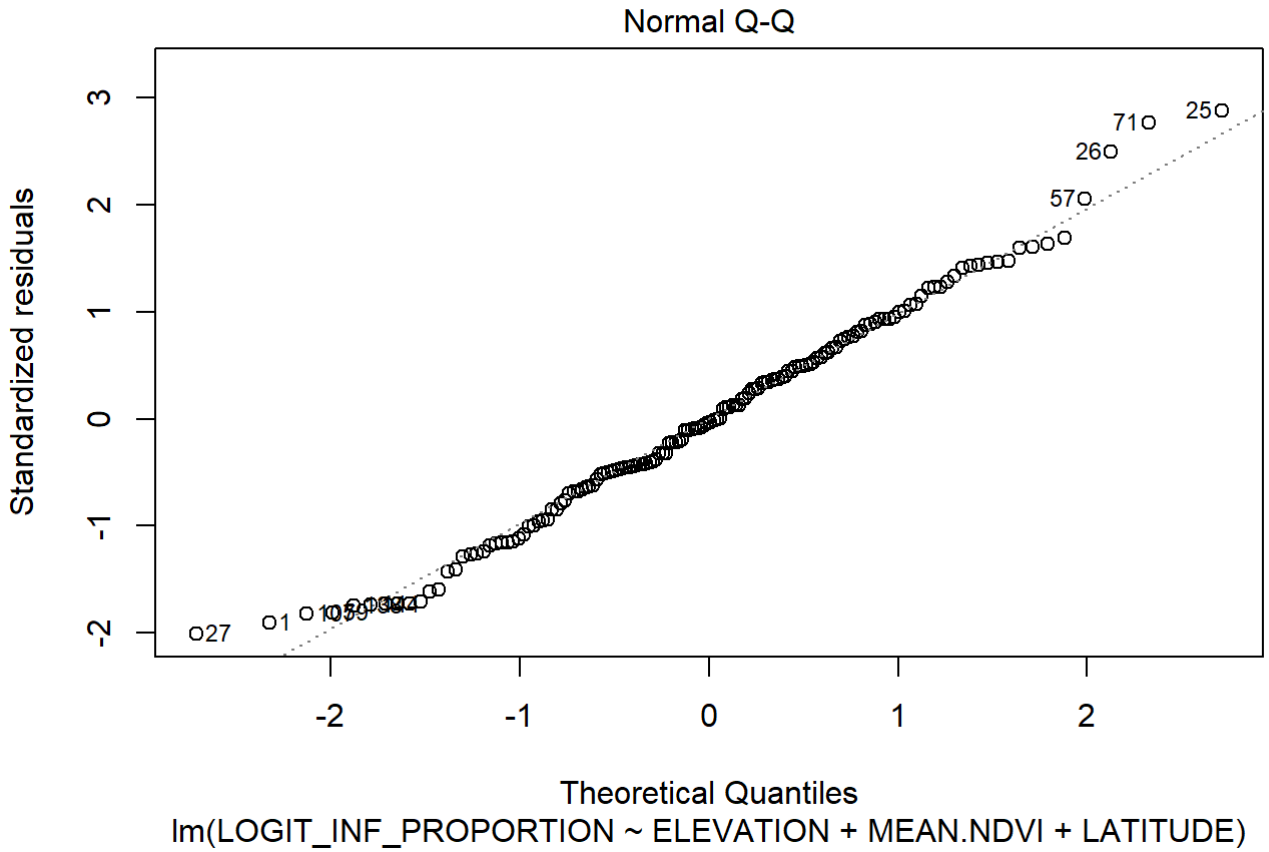
```
##
## Call:
## lm(formula = LOGIT_INF_PROPORTION ~ ELEVATION + MEAN.NDVI + LATITUDE,
##     data = parasite)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39526 -0.45491 -0.02526  0.46022  1.97825
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.4090421  0.6258494  -7.045 6.78e-11 ***
## ELEVATION    0.0022852  0.0003267   6.995 8.88e-11 ***
## MEAN.NDVI    6.7084199  1.2545597   5.347 3.37e-07 ***
## LATITUDE    -0.3604603  0.0631609  -5.707 6.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6988 on 146 degrees of freedom
## Multiple R-squared:  0.6792, Adjusted R-squared:  0.6726
## F-statistic:   103 on 3 and 146 DF,  p-value: < 2.2e-16
```

```
# Residual analysis
res = residuals(parasite.lm.refit)
hist(res, main = "Histogram of the residuals")
```




```
plot(parasite.lm.refit, which=c(1,2), id.n=10)
```


Residuals vs Fitted
Residuals
Fitted values
lm(LOGIT_INF_PROPORTION ~ ELEVATION + MEAN.NDVI + LATITUDE)
71
25
26
57
1
59
138
27
144
107
-4
-3
-2
-1
0
2
1
0
-1


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(LOGIT_INF_PROPORTION ~ ELEVATION + MEAN.NDVI + LATITUDE)
71
25
26
57
27
1
-2
-1
0
1
2
3

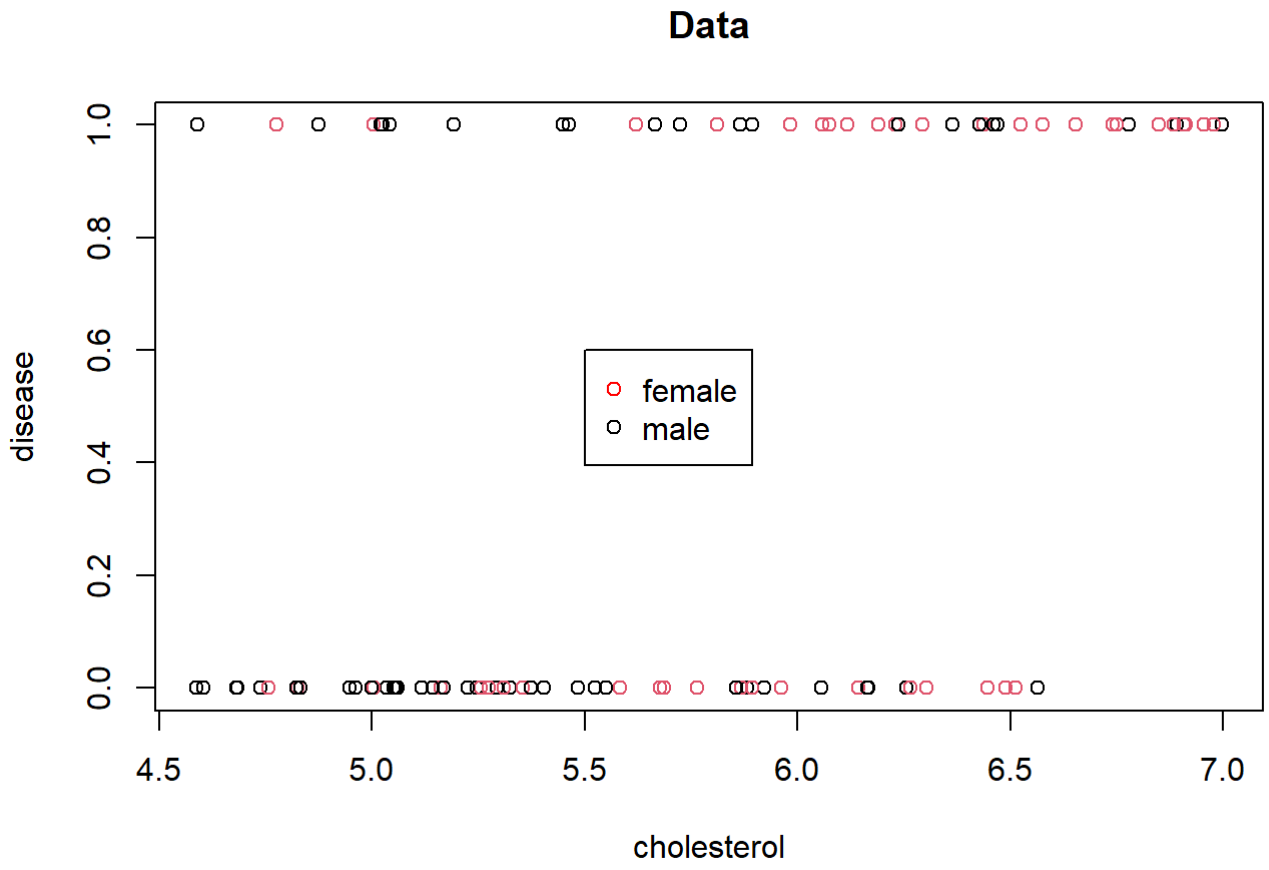
5.a)

Plot of data with gender identified

```
# Summary of cholest data

summary(cholest)
```

```
##        X           cholesterol        gender       genderS
##  Min.   :  1.00   Min.   :4.587   Min.   :0.00   Length:100
##  1st Qu.: 25.75   1st Qu.:5.137   1st Qu.:0.00   Class :character
##  Median : 50.50   Median :5.743   Median :0.00   Mode  :character
##  Mean   : 50.50   Mean   :5.759   Mean   :0.45
##  3rd Qu.: 75.25   3rd Qu.:6.319   3rd Qu.:1.00
##  Max.   :100.00   Max.   :6.997   Max.   :1.00
##     disease
##  Min.   :0.00
##  1st Qu.:0.00
##  Median :0.00
##  Mean   :0.45
##  3rd Qu.:1.00
##  Max.   :1.00
```

```
# 5(a) Plot data with gender identified

plot(disease~cholesterol, data = cholest, col = factor(cholest$genderS), main = "Data")
legend(5.5, 0.6, legend=c("female","male"), col=c("red","black"), pch=1:1)
```



5.b)

The fitted interaction term has a corresponding p value of 0.51208 which shows that it is insignificant and suggests that we should retain the null hypothesis (that the coefficient is 0). In other words, different combinations of genders and cholesterol levels do not significantly affect the likelihood of disease (apart from the independent contributions from each of those variables).

```
# 5(b) Logistic model for gender and cholesterol levels

cholest.logit <- glm(disease ~ cholesterol + genderS + cholesterol:genderS, data = cholest, f
amily = binomial)
summary(cholest.logit)
```

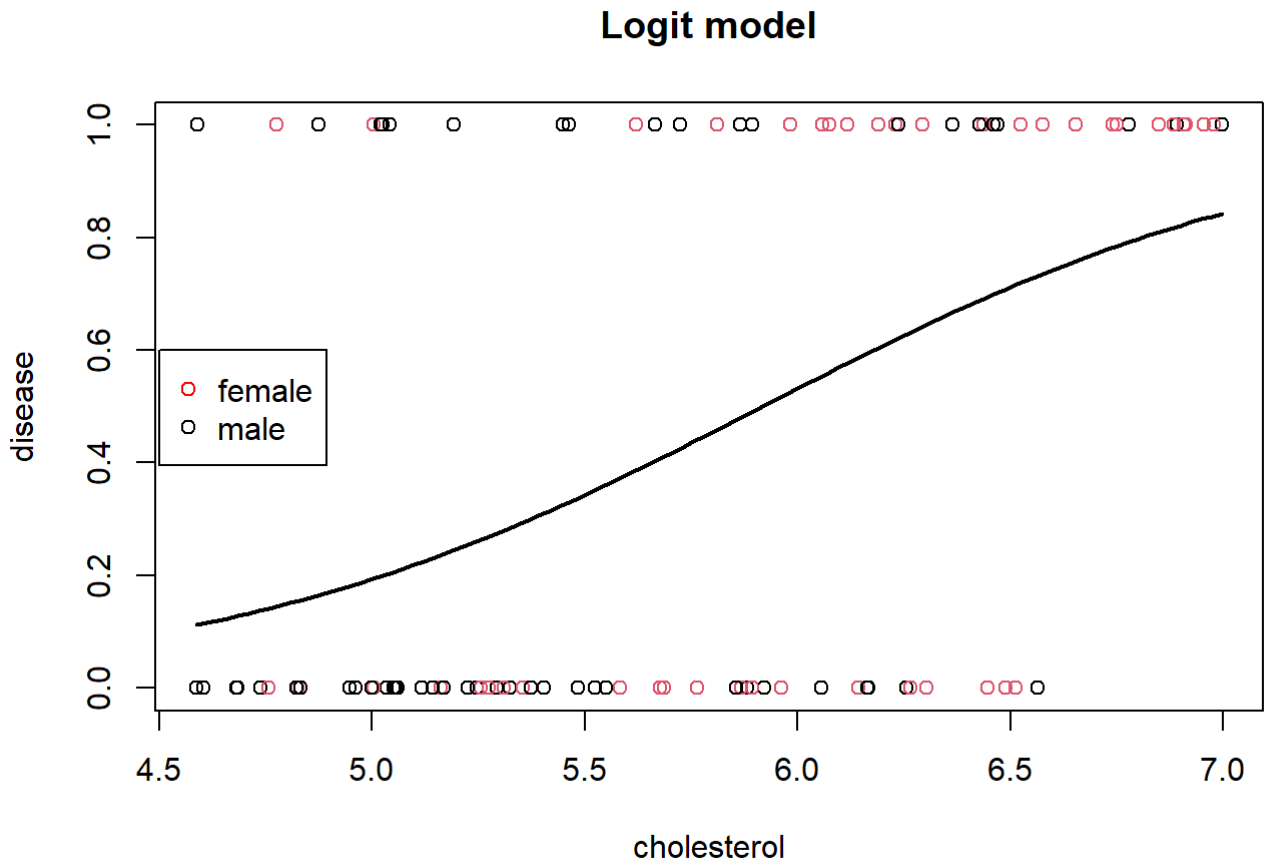
```
##
## Call:
## glm(formula = disease ~ cholesterol + genderS + cholesterol:genderS,
##     family = binomial, data = cholest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6356  -0.8787  -0.5629   0.9091   2.1716
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -8.1310     2.7444  -2.963  0.00305 **
## cholesterol            1.3719     0.4870   2.817  0.00485 **
## genderSm              -3.1692     4.7059  -0.673  0.50065
## cholesterol:genderSm   0.5216     0.7955   0.656  0.51208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137.63  on 99  degrees of freedom
## Residual deviance: 113.55  on 96  degrees of freedom
## AIC: 121.55
##
## Number of Fisher Scoring iterations: 4
```

Since both the interaction term and gender have p values of approximately 0.5, they are insignificant, and after removing them and fitting the model, we get the following plot shown below.

```
# Logistic model without gender

cholest.logit.refit <- glm(disease ~ cholesterol, cholest, family=binomial)
summary(cholest.logit.refit)
```

```
##
## Call:
## glm(formula = disease ~ cholesterol, family = binomial, data = cholest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6236  -0.8488  -0.5481   0.8726   2.0896
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.2012     2.0992  -4.383 1.17e-05 ***
## cholesterol   1.5549     0.3592   4.329 1.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137.63  on 99  degrees of freedom
## Residual deviance: 114.04  on 98  degrees of freedom
## AIC: 118.04
##
## Number of Fisher Scoring iterations: 4
```

```
# Add the fitted line to plot in a)

newdata = data.frame(cholesterol = seq(min(cholest$cholesterol),max(cholest$cholesterol), len
=100))
newdata$logit = predict(cholest.logit.refit, newdata, type = "response")
plot(disease~cholesterol, data = cholest, col = factor(cholest$genderS), main = "Logit model"
)
lines(newdata$cholesterol, newdata$logit, lwd=2)
legend(4.5, 0.6, legend=c("female","male"), col=c("red","black"), pch=1:1)
```



5. c)

Expression for the fitted probit model:
$\Phi^{-1}(disease) = -5.5537 + 0.9411 * cholesterol.$

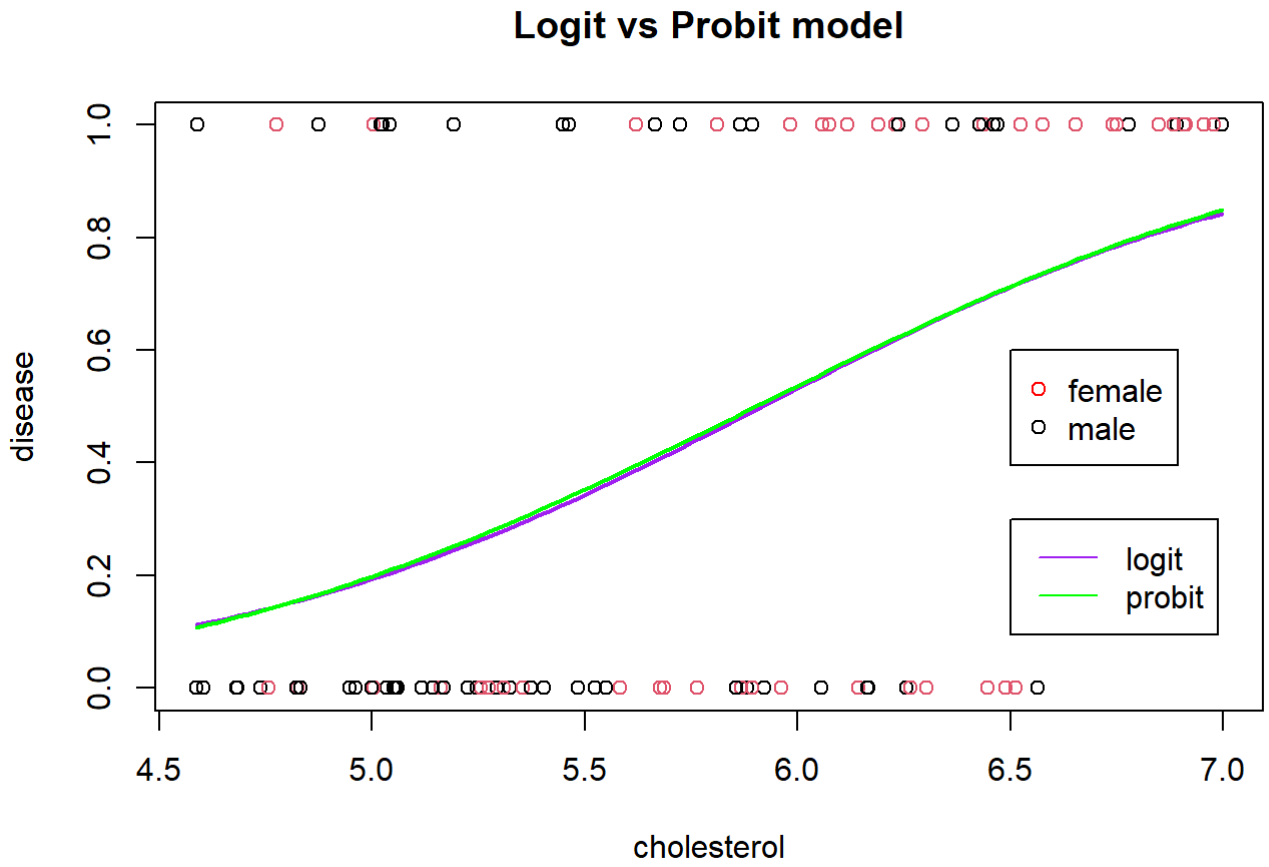
```
# 5(c) Fit probit model

cholest.probit <- glm(disease ~ cholesterol, data=cholest, family=binomial(link="probit"))
summary(cholest.probit)
```

```
##
## Call:
## glm(formula = disease ~ cholesterol, family = binomial(link = "probit"),
##     data = cholest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8645  -0.5464   0.8705   2.1071
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.5537     1.1963  -4.642 3.44e-06 ***
## cholesterol   0.9411     0.2053   4.583 4.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137.63  on 99  degrees of freedom
## Residual deviance: 114.12  on 98  degrees of freedom
## AIC: 118.12
##
## Number of Fisher Scoring iterations: 4
```

5. d)

The following graph shows that the probit and logit models are very similar in this instance and there is no significant difference between them.

```
# 5(d) Add fitted probit mean curve to the graph

newdata = data.frame(cholesterol=seq(min(cholest$cholesterol),max(cholest$cholesterol), len=1
00))
newdata$logit = predict(cholest.logit.refit, newdata, type="response")
newdata1 = data.frame(cholesterol=seq(min(cholest$cholesterol),max(cholest$cholesterol), len=
100))
newdata1$probit = predict(cholest.probit, newdata, type="response")
plot(disease~cholesterol, data=cholest, col=factor(cholest$genderS), main = "Logit vs Probit
 model")
lines(newdata$cholesterol, newdata$logit, lwd=2, col="purple")
lines(newdata$cholesterol, newdata1$probit, lwd=2, col="green")
legend(6.5, 0.6, legend=c("female","male"), col=c("red","black"), pch=1:1)
legend(6.5, 0.3, legend=c("logit","probit"), col=c("purple","green"), lty=1:1)
```


Logit vs Probit model
disease
0.0
0.2
0.4
0.6
0.8
1.0
4.5
5.0
5.5
6.0
6.5
7.0
cholesterol
female
male
logit
probit