MSC1090-Fall 2022
Yuchen Zhou
Assignment 8

Introduction

My research is to use gene editing tools to genetically modify stem cells. Then the stem cells will be differentiated into immune cells to treat cancer. This kind of therapy for cancer is called immunotherapy so I am very interested in data that are related to cancer. I choose esoph data set from the R data sets. The dataset comes from a case-control study of esophageal caner in Ille-et-Vilaine, France (Breslow and Day, 1980). The dataset has 88 records with different age/alcohol/tobacco combination. For each combination, the record has seven columns (averageAge, alcgp, tobgp, ncases, ncontrols, probability, lowrisk). “averageAge” is the average age of a group of participants. “alcgp” is the average alcohol consumption (gram) per day. “tobgp” is the average tobacco consumption (gram) per day. “ncases” is the number of participants who have esophageal cancer. “ncontrol” is the number of participants who do not have esophageal cancer. “probability” is the probability of getting esophageal cancer for a particular age/alcohol/tobacco combination. “lowrisk” indicates whether a particular age/alcohol/tobacco combination has a low risk of developing cancer (TRUE for low risk of developing cancer). Low risk means that the probability of getting esophageal cancer is 0% (Breslow and Day, 1980).

Method

Multiple linear regression and polynomial regression was used to create models that fit the data. Features will be averageAge, alcgp, and tobgp. The target is probability. I use multiple linear regression because there is a relationship between two or more independent variables and one dependent variable (Bevans, 2020). By checking the dataset, I realize that high averageAge, alcgp, and tobgp are correlated with high probability of developing esophageal cancer. Therefore, there are multiple independent variables that can affect the dependent variable (probability). Thus, multiple linear regression is being used. Features are averageAge, alcgp, and tobgp. Label is lowrisk.

Decision tree and logistic regression was used to classify whether a patient with a particular age/alcohol/tobacco combination has a low risk of developing esophageal cancer.

Implementation

The function for each analysis is programmed and stored in Utilities.R. DriverScript.R is used to receive command line argument and call the functions in Utilities.R to do the analysis. The function load.data is used to load the esoph dataset into a dataframe. Then all the functions will use this dataframe to do the analysis. The function multiple.linear.regression will perform multiple linear regression on the data. It uses

```
lm(probability ~ averageAge+alcgp+tobgp,data=mydataframe)
```

to build the model. Then residual standard error (RSE) of the model is calculated. Furthermore, residuals are plotted via different ways (including histogram and Q-Q plot).

The function poly.regression will perform polynomial regression on the data. It uses
```
lm(probability ~
poly(averageAge,5)+poly(alcgp,3)+poly(tobgp,3)+alcgp:tobgp+averageAge:
alcgp+averageAge:tobgp+0,data=mydataframe)
```
to build the model. The order for the polynomial is the highest possible order for each independent variable. Then residual standard error (RSE) of the model is calculated. Furthermore, residuals are plotted via different ways (including histogram and Q-Q plot).

The function decision.tree will build decision tree and classify whether a patient has a low risk or a higher risk of developing esophageal cancer. The data are being split into training set and test set (70%-30% split). The formula that is used to build decision is shown below:
```
f <- lowrisk ~ averageAge+alcgp+tobgp
 mydataframe.tree <- rpart(f,data=train.data,method="class")
```

The accuracy and confusion matrix are calculated using both training set and test set to see the results of the model.

The function log.regression will perform logistic regression on the data to classify whether a patient has a low risk or a higher risk of developing esophageal cancer.. The data are being split into training set and test set.
```
model <- glm(lowrisk ~
averageAge+alcgp+tobgp,family="binomial",data=train.data)
```

is used to build the model. The accuracy of the model is calculated by predicting the labels for the test set. Furthermore, ROC curve was plotted, and the AUC of ROC was calculated to examine the results of the model.

Results

Multiple linear regression

The function multiple.linear.regression output many result. Firstly, it will output the summary of the multiple linear regression model and the RSE of the model.

```
Call:
lm(formula = probability ~ averageAge + alcgp + tobgp, data =
mydataframe)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35102 -0.06570  0.00741  0.06443  0.35162

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.4082239  0.0574546  -7.105 3.62e-10 ***
averageAge   0.0073827  0.0008061   9.158 2.85e-14 ***
alcgp        0.0025214  0.0003239   7.784 1.64e-11 ***
tobgp        0.0021702  0.0012984   1.671   0.0984 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1167 on 84 degrees of freedom
Multiple R-squared:  0.6308,  Adjusted R-squared:  0.6176
F-statistic: 47.84 on 3 and 84 DF,  p-value: < 2.2e-16

the residual standard error of the model is 0.116742 .
```

From the output, we can tell that the RSE of the model is 0.116742, which is not very low. It shows that the model is not a perfect fit for the data. If the model is a good one, the RSE should be as close to 0 as possible. We also see that the R-squared value is equal to 0.6308. R-squared value shows the proportion of variation in the response can be explained by the linear model. The R-squared value should be as close to 1 as possible. However, our R-squared value is only 0.6308, which indicated there are shortcoming in the models or noises. The function also plots several graphs. The first graph is a plot of residuals against index, each independent variable, and the dependent variable (**Figure 1)**. If the model is a good fit of the data, we should see a snowstorm. However, for the five plots shown below, we could still see come clumps, which asl indicates the models do not fit well to the data.

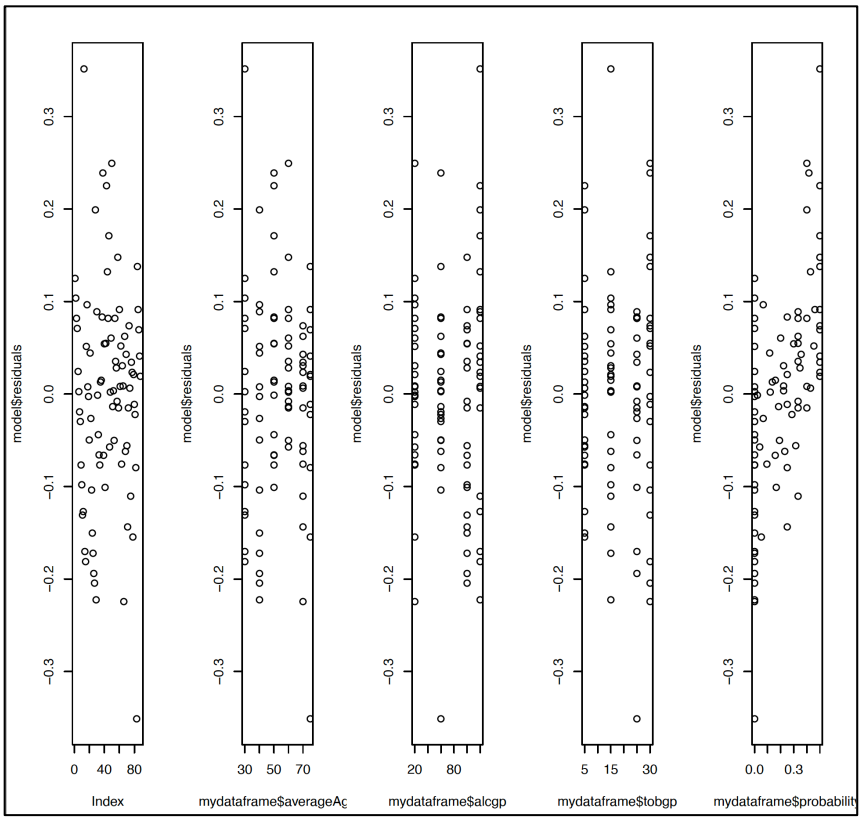

Figure 1. plot of residuals against index, averageAge, alcgp, tobgp, probability for multiple.linear.regression

The second figure is a histogram of the residuals (**Figure 2**). The mean of the plot should be zero, and this histogram shows that the residuals are centered on zero. However, the distribution is not exactly symmetric, which means that the model is biased. Some structures in that data are not being captured by the multiple linear regression.

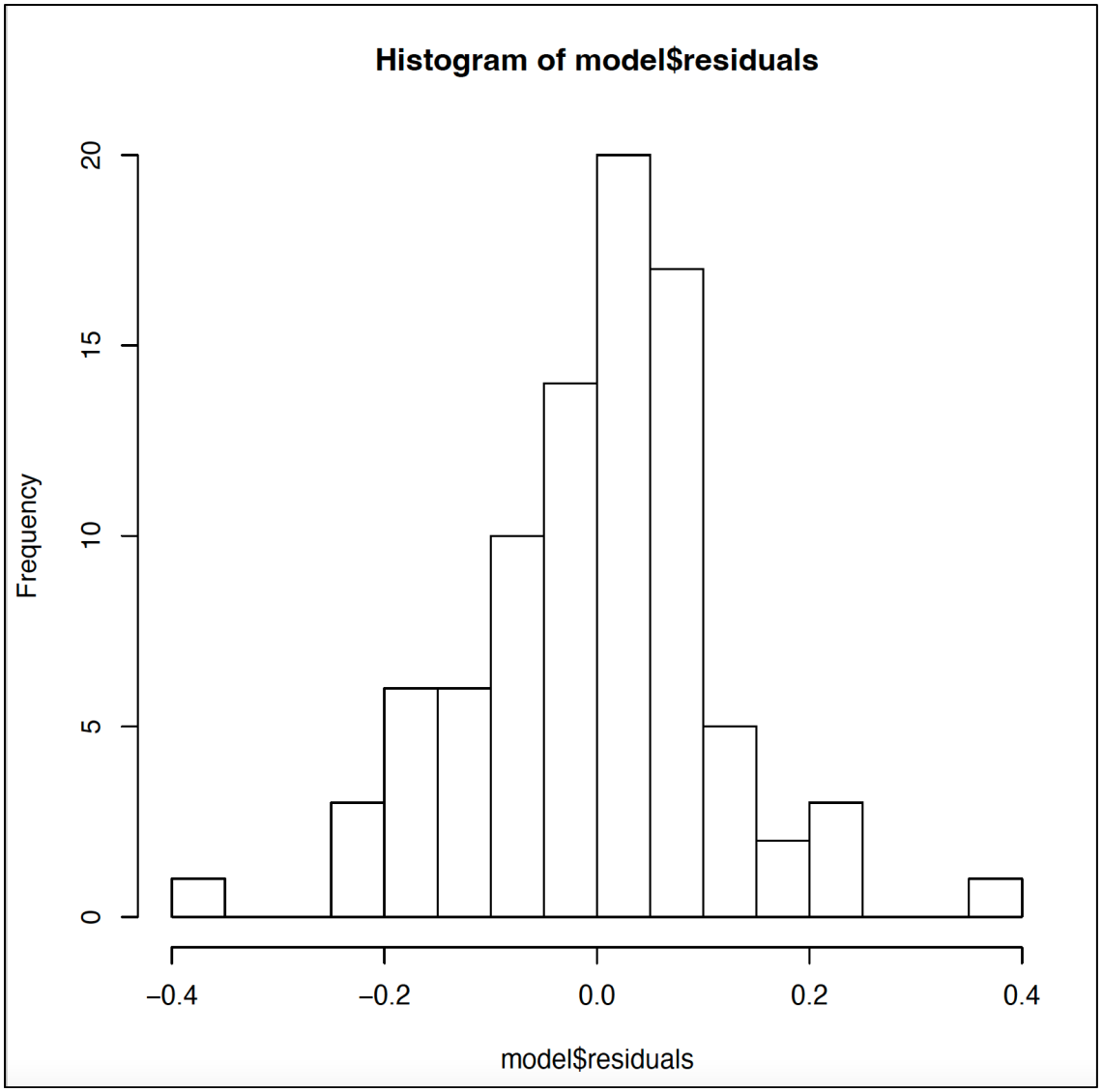

Figure 2. a histogram of residuals for multiple.linear.regression

The third figure is a Q-Q plot of the residuals (**Figure 3**). A Q-Q plot graphically demonstrate the normality of the residuals. The best case should be that the residuals are normally distributed. However, in the Q-Q plot, we can see that most points are on the line. However, at the two ends of the plot, the points are not on the line. This also indicates that the model is not a perfect fit for the data.

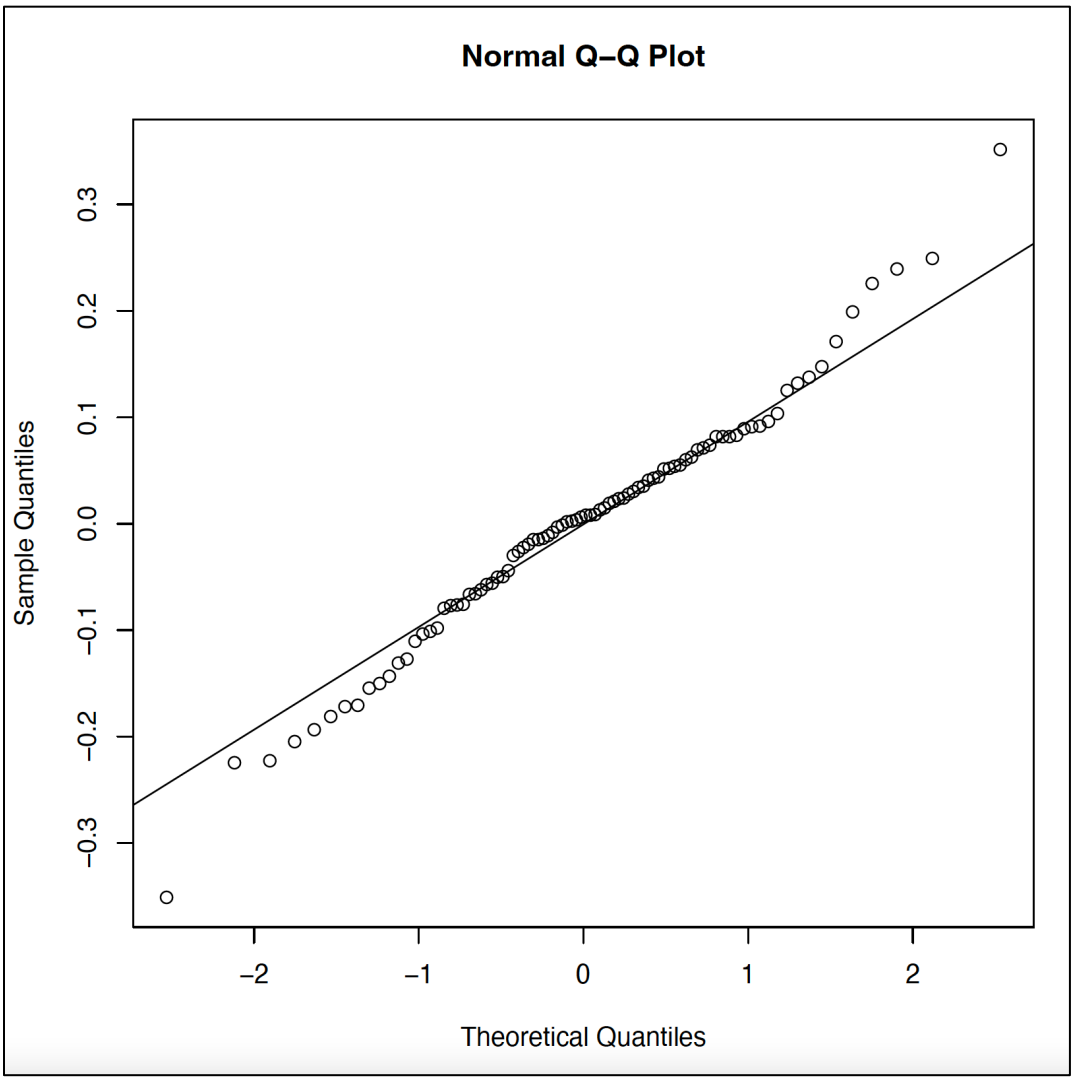

Figure 3. the Q-Q plot of residuals for multiple.linear.regression

Polynomial regression

The function poly.regression also outputs many result and figures. Firstly, it will output the summary of the polynomial regression model.

```
Call:
lm(formula = probability ~ poly(averageAge, 5) + poly(alcgp,
    3) + poly(tobgp, 3) + alcgp:tobgp + averageAge:alcgp +
averageAge:tobgp +
    0, data = mydataframe)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30774 -0.05323 -0.00483  0.05826  0.35618

Coefficients:
```

```
                      Estimate Std. Error t value Pr(>|t|)
poly(averageAge, 5)1  4.212e-01  1.510e-01   2.790  0.00670 **
poly(averageAge, 5)2 -1.869e-01  1.091e-01  -1.713  0.09088 .
poly(averageAge, 5)3 -1.239e-01  1.088e-01  -1.139  0.25848
poly(averageAge, 5)4  2.144e-01  1.088e-01   1.971  0.05242 .
poly(averageAge, 5)5 -1.134e-02  1.084e-01  -0.105  0.91693
poly(alcgp, 3)1       2.459e-01  2.917e-01   0.843  0.40198
poly(alcgp, 3)2       1.452e-01  1.085e-01   1.338  0.18499
poly(alcgp, 3)3       2.026e-01  1.088e-01   1.863  0.06641 .
poly(tobgp, 3)1      -3.404e-02  2.958e-01  -0.115  0.90870
poly(tobgp, 3)2       4.859e-02  1.083e-01   0.448  0.65516
poly(tobgp, 3)3       1.429e-01  1.090e-01   1.310  0.19410
alcgp:tobgp          -2.487e-05  2.875e-05  -0.865  0.38973
alcgp:averageAge      4.276e-05  1.386e-05   3.085  0.00286 **
tobgp:averageAge      7.923e-05  5.750e-05   1.378  0.17238
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1082 on 74 degrees of freedom
Multiple R-squared:  0.8753,  Adjusted R-squared:  0.8517
F-statistic:  37.1 on 14 and 74 DF,  p-value: < 2.2e-16
```

You can see that the R-squared value is now 0.8753, which is much higher than the R-square value of multiple linear regression model. This indicates that the polynomial regression model can explain more variance in the data. As a result, polynomial model is a better fit for our dataset. However, it is still not a perfect fit. The R-squared needs to be as close to 1 as possible.

The function also plots several graphs. The first graph is a plot of residuals against index, each independent variable, and the dependent variable (**Figure 4)**. If the model is a good fit of the data, we should see a snowstorm. However, for the five plots shown below, we could still see come clumps, which asl indicates the models do not fit perfectly to the data.

The second figure is a histogram of the residuals (**Figure 5**). The mean of the plot should be zero, and this histogram shows that the residuals are centered on zero. However, the distribution is not exactly symmetric, but it is more symmetric than Figure 2. This shows that the polynomial model is less biased in comparison to multiple linear model. However, some structures in that data are not being captured by the polynomial regression model.

The third figure is a Q-Q plot of the residuals (**Figure 6**). A Q-Q plot graphically demonstrate the normality of the residuals. The best case should be that the residuals are normally distributed. However, in the Q-Q plot, we can see that most points are on the line. However, at the two ends of the plot, the points are not on the line. This also indicates that the model is not a perfect fit for the data.

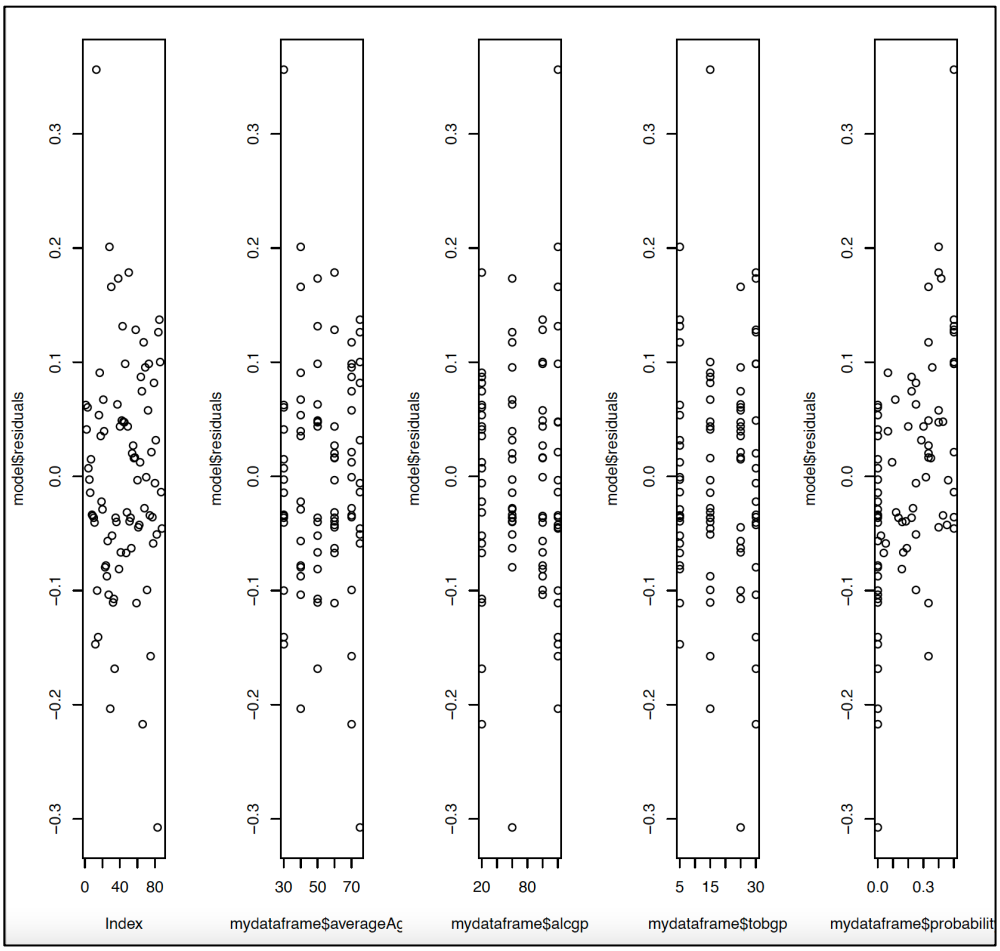

Figure 4. plot of residuals against index, averageAge, alcgp, tobgp, probability for multiple.linear.regression for poly.regression

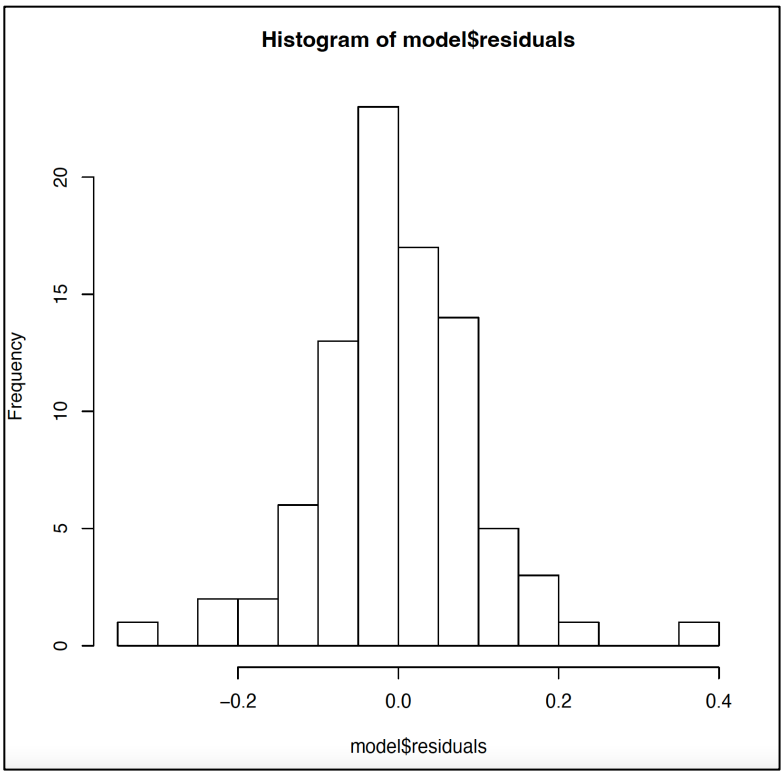

Figure 5. a histogram of residuals for poly.regression

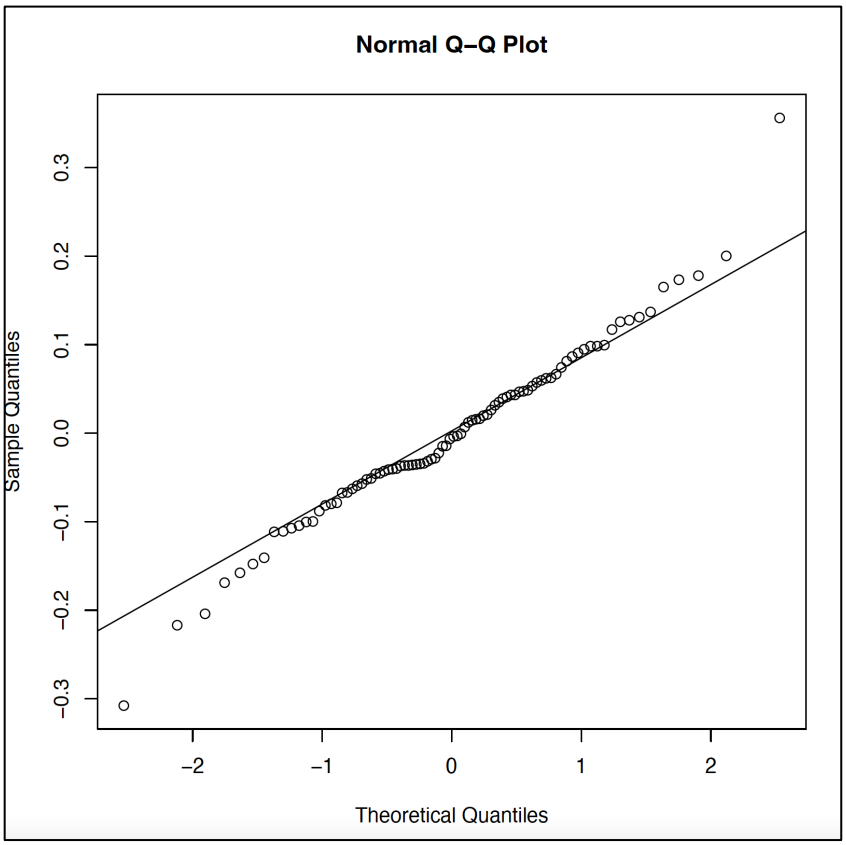

Figure 6. the Q-Q plot of residuals for poly.regression

Decision Tree

The function decision.tree output many result. Firstly, it will plot the results of the decision tree (**Figure 7**).

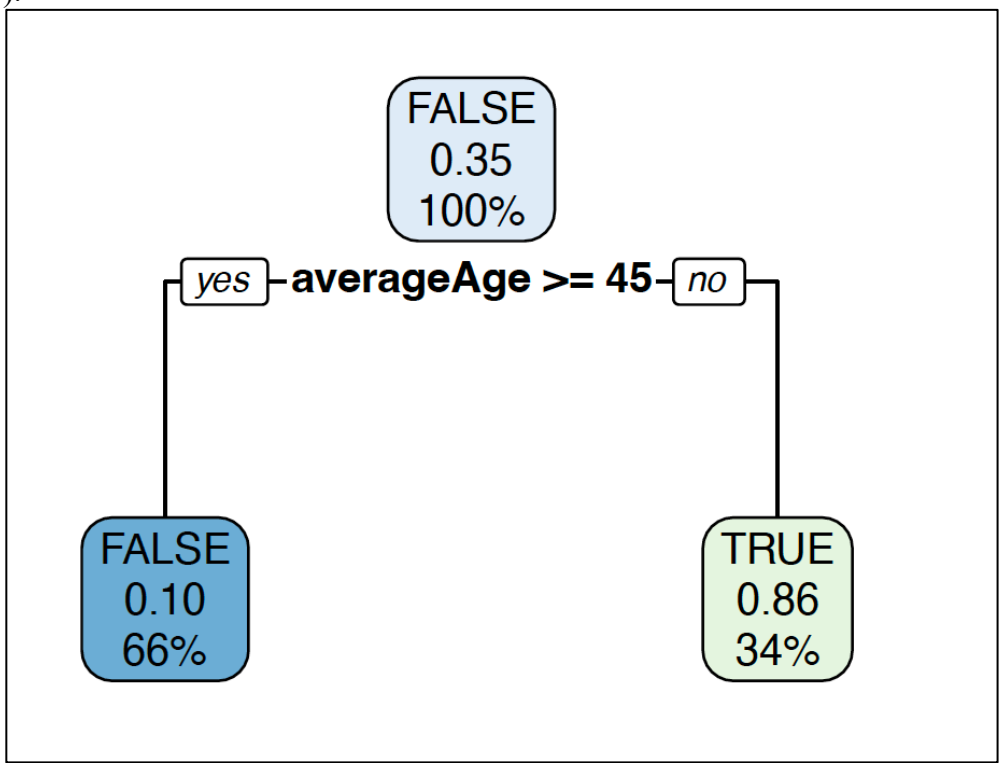

Figure 7. Results of decision tree for classifying lowrisk=TRUE or lowrisk=FALSE

Then the function will calculate and ouput the accuracy and confusion matrix for the training set and test set. The results are shown below:

```
The accuracy of decision tree for classifying training data is
0.8870968 .
       pred
        FALSE TRUE
  FALSE    37    3
  TRUE      4   18
The accuracy of decision tree for classifying testing data is
0.8461538 .
       testPred
        FALSE TRUE
  FALSE    16    3
  TRUE      1    6
```

The accuracy and confusion matrix tells us that the decision tree could classify most of the training data and test data correctly. However, there are still 10—15% of the data cannot be

classified into correct category. Also the model does worse on test set in comparison to training set.

Logistic regression

The function log.regression output many result. Firstly, it will output the summary of the model. Secondly, it will test the model on test set. The function calculates and outputs the accuracy and the confusion matrix of the model. The result is shown below:

```
      FALSE TRUE
  FALSE    17    3
  TRUE      0    9
The accuracy of logistic regression model for classifying test data is
0.8965517 .
```

As you can see the model classifies all the true cases as true. Only 3 cases are true but they are classified as false (false negative). The accuracy is 0.896, which is relatively high for a model. Thus, this logistic model is a good classifier for the data. Furthermore, the function plots ROC curve (**Figure 8)** and calculates the area under curve (AUC). The programs shows that `The AUC of ROC curve is 0.9555556.`

AUC of ROC curve could be used to check the quality of the classifier. A good classifier will have an AUC that is close to 1. Our AUC value is relatively close 1, so the logistic regression model is a good classifier for our data set.

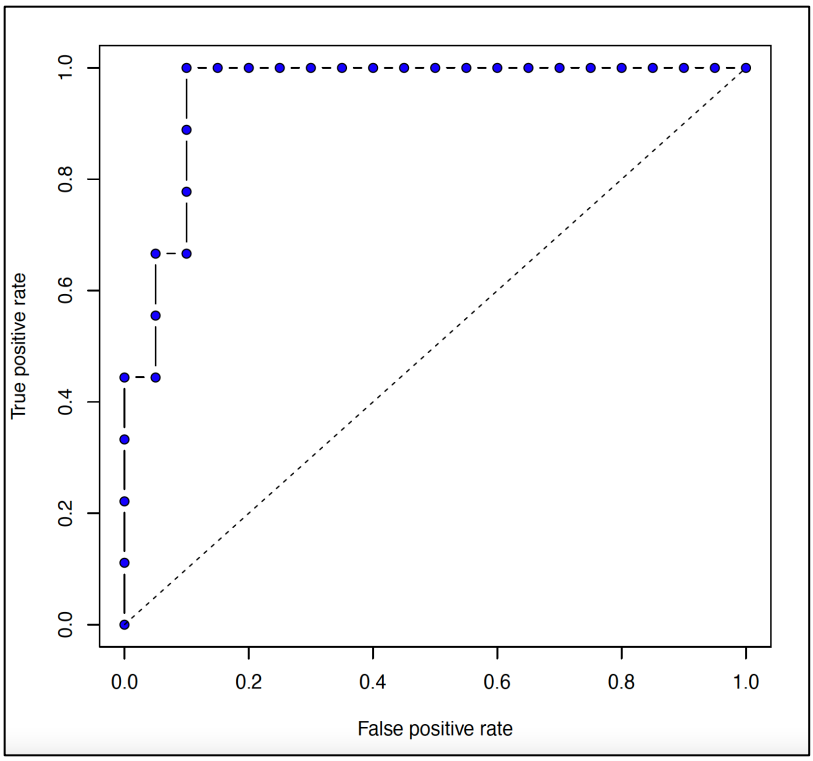

Figure 8. ROC curve for the logistic regression model

Discussion

The R language has many advantages for doing analysis. First of all, R has many modules that contain different kinds of statical analysis methods. Thus, you do not need to program by yourself. However, in Excel and SPSS, some of the statistical analyses (such as decision tree) are not available. Thus, you might need to write your own algorithm and program to perform those analysis. Secondly, R language is more flexible than other software such as Excel. For R language, you can change the parameters for analyses. You can do some further analysis on your models. However, Excel just outputs the model, and you are not able to change some parameters of the analyses and check whether the model is a good one by further analysis such as Q-Q plot.

However, R languages also have many drawbacks. R language is different from user-friendly software which has an interface that is easy to use. You need to know the modules and the function in the modules to perform analysis correctly.

References

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies.* IARC Lyon / Oxford University Press.

Bevans, R. (2022, June 01). *Multiple Linear Regression | A Quick Guide (Examples).* Scribbr. Retrieved November 11, 2022, from https://www.scribbr.com/statistics/multiple-linear-regression/