# ASSIGNMENT 2

(Chen Liang, Student ID 46275313)

## Question 1

## 1. Introduction

River bank failure can be caused when the gravitational forces acting on a bank exceed the forces which hold the sediment together [1]. Failure depends on sediment type, layering, and moisture content [1].

The Yangtze River is the largest river in China, with a total length of 6300 kilometres, which is located in the third place of the whole world. The basin area of Yangtze River is 1 million 800 thousand square kilometres, accounting for 20% of the total area of China [2]. During summer and autumn, the Yangtze River Basin is always affected by the southeast and southwest monsoon which are prevalent in the Pacific oceans, and precipitation is formed during the confrontation between the monsoon and the cold and warm air flow. Except the Jinsha River basin, there are heavy rains in the other 1 million 500 thousand square kilometres area [2]. Because of the long-term erosion of fast flow, the Yangtze River Basin always has a strong river bank.

However, since 2016, serious rainfall has occurred along Yangtze River for a great number of times, and the highest water level has been constantly refreshed. In the flood season of 2016, 32 million 820 thousand people lived in Yangtze River Basin suffered floods and 56 thousand houses collapsed. Once the river bank of Yangtze River was famous for its strong, there have been 45 places failed till now.

Therefore, for the people staying inside Yangtze River Basin, if it is possible to estimate the performance of river bank according to the geological and river characteristics of the selected address before building their house, it will be helpful to keep the property and personal safety.

The main purpose of this research is to help customers select places with less possibility of riverbank failure along the river to build houses according to the data of survey. A total of 82 families participated in this survey. Some of them lost their homes because of the riverbank failure, and others had lived on the strong bank of the river for decades and had never been affected by heavy rain along Yangtze River. The report studied whether the characteristics of land and river near houses played a role in river bank failure.

## 2.Method

The main tool for this analysis was R Studio, which was a compiler based on R, providing many embedded functions and packages for data analysis and model evaluation.

### 2.1 Data Description and Processing

The data collected 8 variables of 82 observations, however, failure, sediment, meander and landcover were categorical variables (using dim() and names() function to check). Firstly, factor () function was used to convert the categorical variables into factors to make sure the levels of each

variable could perform well when fitting the model. Next, use summary () function to scan the descriptive statistics of each variable.

For categorical variables in data set, bar charts were drawn to analyse the frequency of each level of the variable (using barplot() function). For numeric variables, the side by side boxplots and histograms were used (being realised by boxplot() and hist()) to study the distribution of predictors for the two outcomes of responding variable, which were 'failed' and 'not failed'.

In the dataset, the size of samples between the levels of some categorical variables was quite different, which would reduce the accuracy of judgment of the significance of the level with small sample size for the model. Therefore, in this study, in order to solve this problem, grassy and agriculture which were levels of landcover were combined to achieve the balance of sample size. In addition, if one level of a variable could lead to the failure of riverbank directly, it should be excluded because it would affect the prediction of the significance of other variables in the model.

Also, if a numerical variable in the dataset had a quite poor continuity, when it was added into the model fitting, it would affect the accuracy. Therefore, in this report, the numerical variable in this situation (dredging) was transformed into categorical variable based on the characteristic of its value.

### 2.2 Model Fitting

In the regression process, if the assumption that the dependent variable is normal distribution is not reasonable, generalized linear model could be implemented to fit it. Then, when the responding variable is a categorical variable with two values, using logistic regression, which fits probability of Y belonging to a certain class. Since the corresponding variable (failure) was binary with two values, which were 'failed' and 'not failed', using logistic regression to fit the data. In this study, probability of failure is equal to 1 should be fitted.

$$\boldsymbol{ln}\left(\frac{\boldsymbol{p}(\boldsymbol{failure}=\mathbf{1})}{\mathbf{1}-\boldsymbol{p}(\boldsymbol{failure}=\mathbf{1})}\right)=\boldsymbol{\beta_0}+\boldsymbol{\beta_1 X}$$

For Model 1, all variables were used as independent variables to fit ‘failure’.

**failure ~ sediment + meander + channelwidth + landcover + vegewidth + sinuosity + dredging**

After obtaining the full model, step () was conducted to do the variable selection preliminarily and Model 2 was got.

The p value was used to determine whether the variable was significant in the model (if the p value was too large, the coefficient of that variable was equal to zero could not be rejected). Therefore, the independent variable with the largest P value (which was greater than 0.05 in this study) variable was dropped (using update () function to realise this progress) and Model 2.1 was then got. In this process, summary() and anova() were conducted respectively to check the p-value for each level of the variable and the whole variable.

In terms of statistics, interaction terms were usually added into fitting progress to analyse interactions between variables. Therefore, based on Model 2.1, adding interactions of independent variables to conduct variable selection (using step () function) again.

After getting the model with interaction (Model 3.1), anova () function was used to compare two nested models (Model 3.1 was a subset of Model 2.1). For logistic regression, the parameter 'test' in anova () was set as 'Chisq'. If the Chi square value was not significant, it showed that the model with a small number of independent variables was as good as the original one. At the same time, if the Chi square value was significant, it showed that the dropped variables would affect the accuracy of the model.

### 2.3 Model Selection

In order to see the prediction ability of models, table () function was implemented to generate the confusion matrix. In the validation process, when the 'failure' value of prediction was greater than 0.5, it was considered that the stop bank failed. The stop bank not failed if predicting 'failure' was less than 0.5. Firstly, a vector consisting of 82 'not failed' was created, and then the elements that greater than 0.5 were converted into 'failed'. Next, confusion matrix built by table () could be used to determine how many observations were successfully predicted.

ROC, the receiver operating characteristic curve, are also used to evaluate the logistic model. When two value problems are encountered, the ROC curve is drawn by specificity and sensitivity. Among them, specificity is FPR (false positive rate), which means that a negative class observation is wrongly classified into the positive class. Sensitivity represents TPR, that is, true positive rate, which means that the actual class of the point is positive, but it is classified into the positive class. The area under the ROC curve (AUC) is used to measure the prediction ability of the model. Two methods which were Modelroc() function in package Proc and calculating TPR and FTR directly were both conducted to draw ROC.

When selecting final model, the AIC-value, namely Akaike information criterion (a standard to evaluate the goodness of fit of statistical models) was compared as a reference.

## 3. Result

### 3.1 Data Description and Processing

The summary statistics of the dataset was shown in Table 1. It could be seen from Table 1 that the stop bank failed in half of the observations. About fifty percent of the selected samples were coarse-grain channel fill.

**Table 1 Summary Statistics**

| failure | | sediment | | meander | | Channelwidth (cm) | |
|---|---|---|---|---|---|---|---|
| not failed | 41 | not fill | 40 | inside bend | 20 | Min. | 712.4 |
| failed | 41 | fill | 42 | outside bend | 16 | 1st Qu. | 1533.1 |
| | | | | chute | 24 | Median | 1824.7 |
| | | | | straight | 22 | Mean | 2084 |
| | | | | | | 3rd Qu. | 2562.4 |
| | | | | | | Max. | 4130.2 |
| landcover | | Vegewidth (cm) | | sinuosity | | dredging | |
| open water | 5 | Min. | 0 | Min. | 1.05 | Min. | 0 |
| grassy | 14 | 1st Qu. | 224.6 | 1st Qu. | 1.325 | 1st Qu. | 68032 |
| agriculture | 25 | Median | 427.2 | Median | 1.581 | Median | 98806 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| forest | 38 | Mean | 1218.2 | Mean | 1.689 | Mean | 352759 |
| | | 3rd Qu. | 1460.3 | 3rd Qu. | 2.053 | 3rd Qu. | 755629 |
| | | Max. | 6177.2 | Max. | 2.962 | Max. | 799563 |

The bar charts of variable 'meander' were shown as below. It could be figured out that the sample sizes of different levels were basically balanced and the river bank had a higher possibility to be successful when meander was outside bend and straight. The bar chat of variable 'sediment' was shown as Figure 1 in Appendix 1.

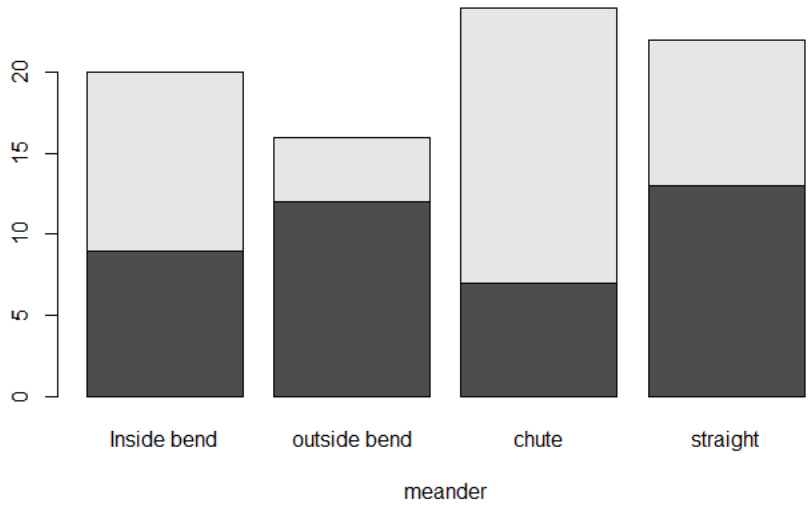
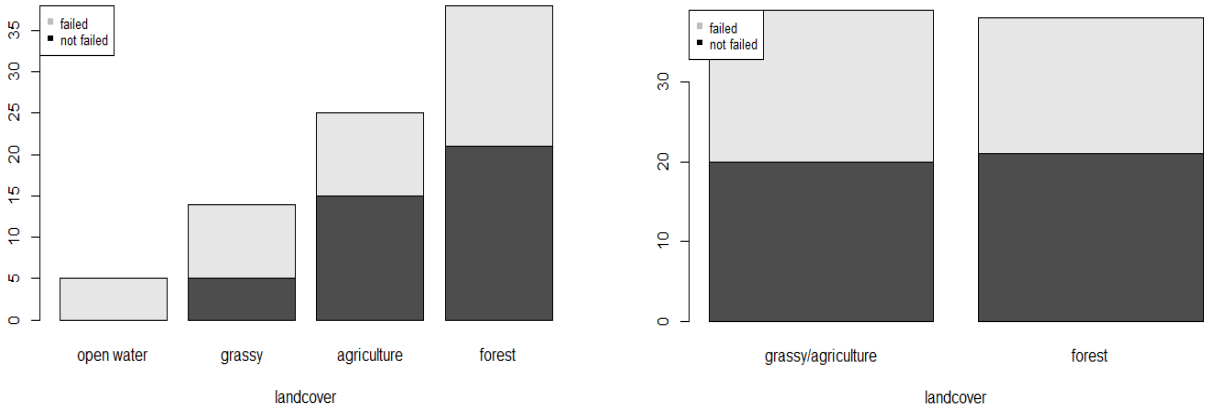



**Figure 1 Bar chart of Meander**

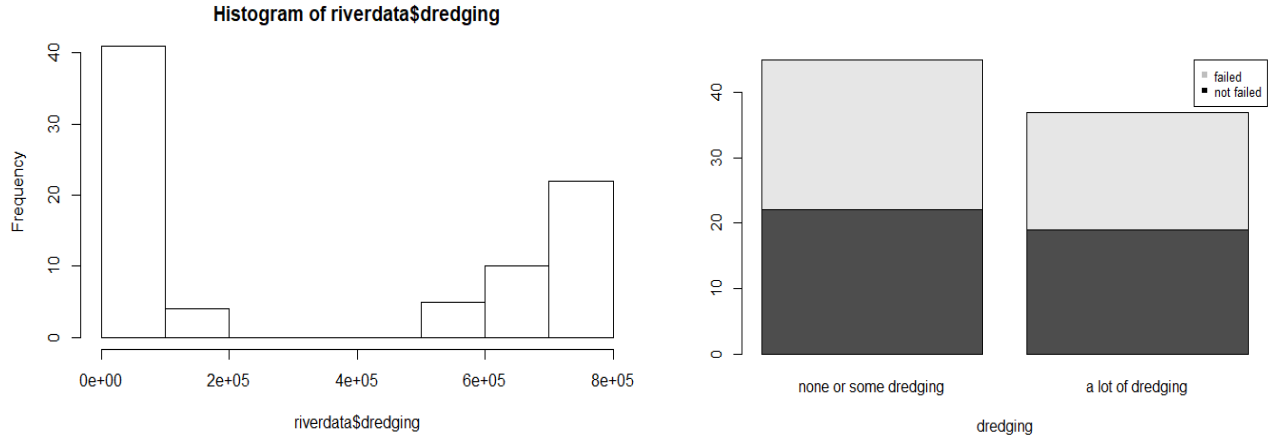
The bar chart of variable 'landcover' was displayed below as (a) in Figure 2, followed by the chart after combination of category 'grassy' and 'agriculture' and the deletion of the level 'open water' which would lead to failure directly. Figure 1 told that when the location was open water, the consequence must be 'failed' in this sample.




**(a)** **(b)**

**Figure 2 Combination the Levels of Landcover**

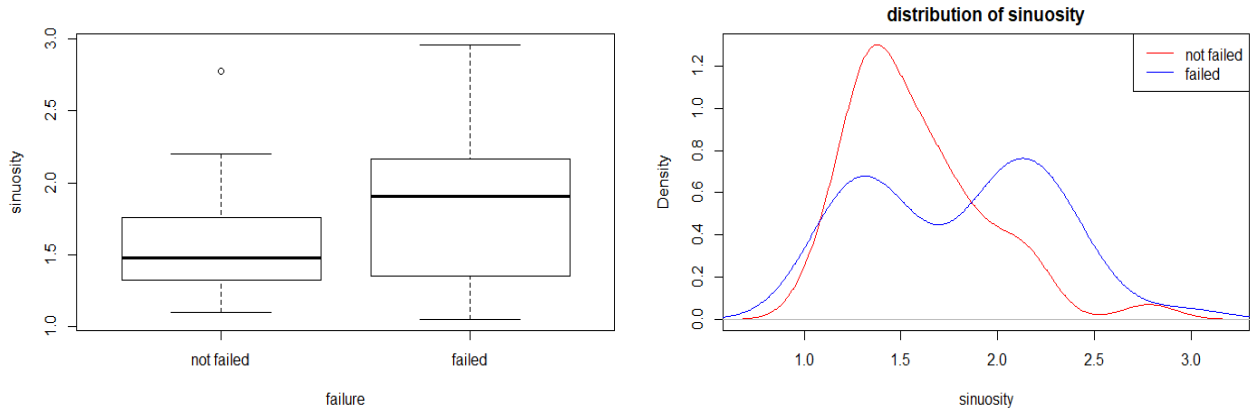
The histogram of variable ‘dredging’ was shown in (a) of Figure 3, which presented a large gap between two valued ranges. (b) was the bar chart of ‘dredging’ after it was transformed into categorical variable.




**(a)** **(b)**

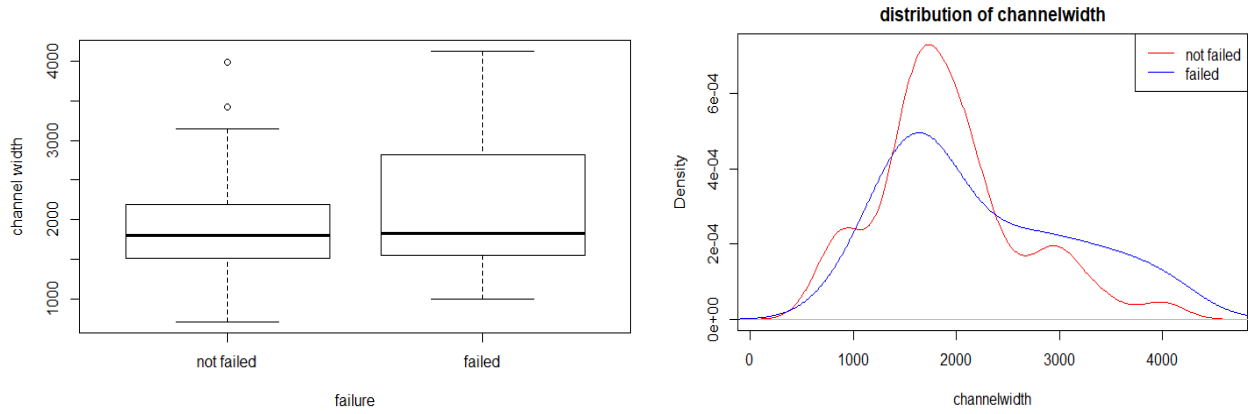
**Figure 3 Comparison between dredging before and after being transformed into categorical variable**

The side by side boxplot and density plot of sinuosity, channel width and vegewidth for two outcomes of ‘failure’ were shown in Figure 4, Figure 5 and Figure6 respectively, which told the difference of the distribution and value range when the river bank failed and not failed.



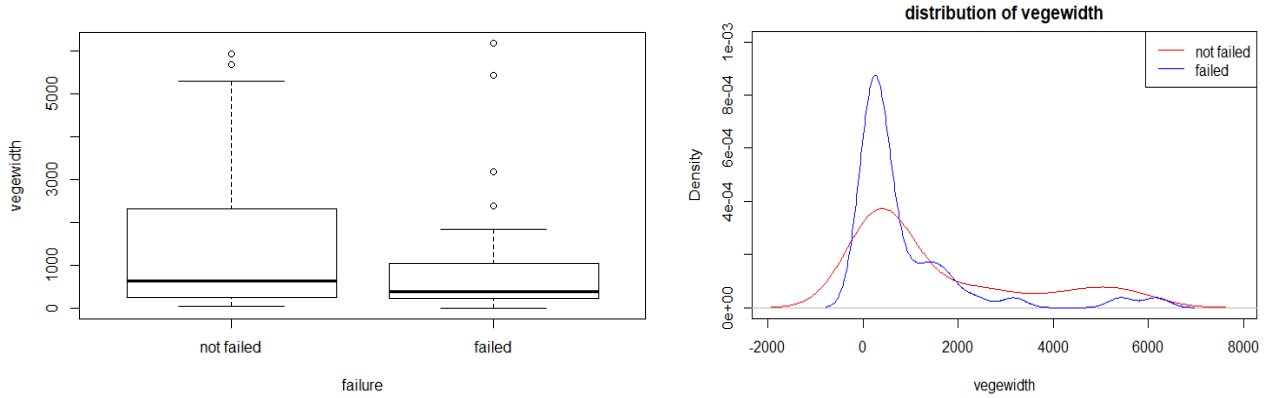

**(a)** **(b)**

**Figure 4 Boxplot and Density Comparison of Sinuosity for ‘failed’ and ‘not failed’**




**(a)** **(b)**

**Figure 5 Boxplot and Density Comparison of Channelwidth for 'failed' and 'not failed'**




**(a)** **(b)**

**Figure 6 Boxplot and Density Comparison of vegewidth for 'failed' and 'not failed'**

### 3.2 Model Fitting

Table2 showed the parameters and statistics of Model 2 obtained from stepwise regression. The predictor with largest variable of p was 'dredging', which had a great p value and low significance for the model.

**failure ~ meander + sinuosity + dredging** **(Model 2)**

**Table2 Coefficients and Statistics for Model 2**

| | Estimate | Pr(>\|z\|) | p-value for the variable | assessment |
|---|---|---|---|---|
| Intercept | -1.5945 | 0.1554 | | |
| meanderoutsidebend | -1.4064 | 0.0743 | 0.03328 | * |
| meanderchute | 0.8178 | 0.3007 | | |
| meanderstraight | -0.4754 | 0.4936 | | |
| sinuosity | 1.1969 | 0.0640 | 0.04719 | * |
| dredging(lots of dredging) | -0.9104 | 0.1262 | 0.11302 | |
| AIC | 103.25 | | | |

Model 2.1 was then got after dropping the variable with the greatest p-value. The coefficients and statistics was shown in Table 3.

**failure ~ meander + sinuosity (Model 2.1)**

**Table3 Coefficients and Statistics for Model 2.1**

| | Estimate | Pr(>\|z\|) | p-value for the variable | assessment |
|---|---|---|---|---|
| Intercept | -1.8417 | 0.0960 | | |
| meanderoutsidebend | -1.4873 | 0.0575 | 0.03328 | * |
| meanderchute | 0.3104 | 0.6534 | | |
| meanderstraight | -0.6047 | 0.3699 | | |
| sinuosity | 1.2312 | 0.0542 | 0.04719 | * |
| AIC | 103.76 | | | |

The interaction terms of predictors in model2.1 were added into the model fitting process, and stepwise regression was implemented to get Model3.1. The coefficients and statistics of model 3.1 were displayed in Table 4.

**failure ~ meander + sinuosity + meander: sinuosity (Model 3.1)**

**Table4 Coefficients and Statistics for Model 3.1**

| | Estimate | Pr(>\|z\|) | p-value for the variable | assessment |
|---|---|---|---|---|
| Intercept | -4.4158 | 0.0457 | | |
| meanderoutsidebend | 6.2280 | 0.0923 | | |
| meanderchute | -0.1651 | 0.9621 | | |
| meanderstraight | 5.8290 | 0.1076 | 0.03328 | * |
| sinuosity | 2.9274 | 0.0447 | 0.04719 | * |
| meanderoutside bend:sinuosity | -4.6434 | 0.0434 | | |
| meanderchute:sinuosity | 0.0075 | 0.9971 | | |
| meanderstraight:sinuosity | -4.3225 | 0.0776 | 0.02917 | * |
| AIC | 100.75 | | | |

## 2.3 Model Selection

Using anova() to compare Model 3.1 with Model 2.1 and getting the result shown as followed. The p-value was less than 0.05, which meant extra item in Model 3.1 was significant.

**Table5 Comparison between Model 3.1 and Model 2.1**

| | Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) | assessment |
|---|---|---|---|---|---|---|
| Model 2.1 | 72 | 93.763 | | | | |
| Model 3.1 | 69 | 84.753 | 3 | 9.0093 | 0.02917 | * |

The confusion matrix was shown as below. It could be seen that the prediction accuracy rate of Model 2.1 was 0.6753 as well as Model 3.1 was 0.7143.

**Table6 Prediction Accuracy of Model 2.1 and Model 3.1**

| Model2.1 | not failed | failed | Model3.1 | not failed | failed |
|---|---|---|---|---|---|
| failed | 9 | 20 | failed | 5 | 19 |
| not failed | 32 | 16 | not failed | 36 | 17 |
| accuracy | 0.6753 | | accuracy | 0.7143 | |

The ROC figures of Model 2.1 and Model 3.1 were shown as followed, the ROC figures drawn by calculating TPR and FTR firstly were put in Appendix1 (Figure 2 and Figure 3). It could be figured out that AUC value of Model 3.1 was greater than that of Model 2.1.

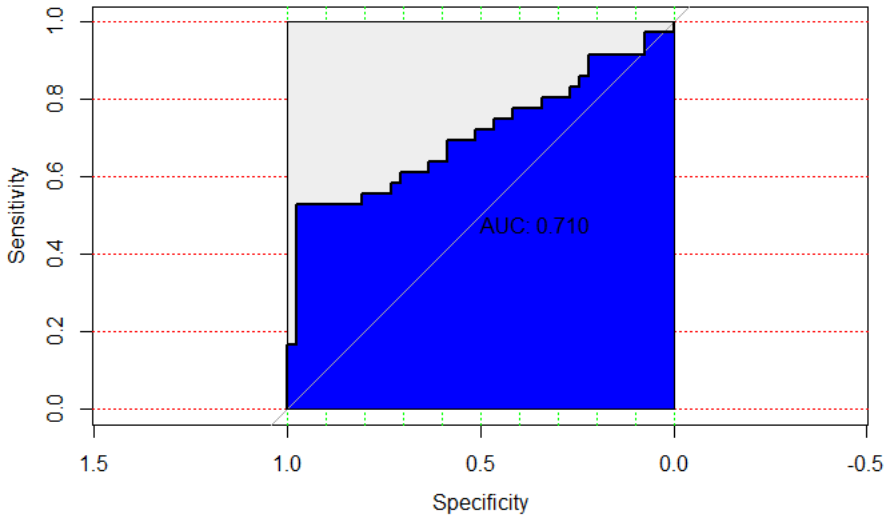



**Figure 7 The ROC of Model2.1**

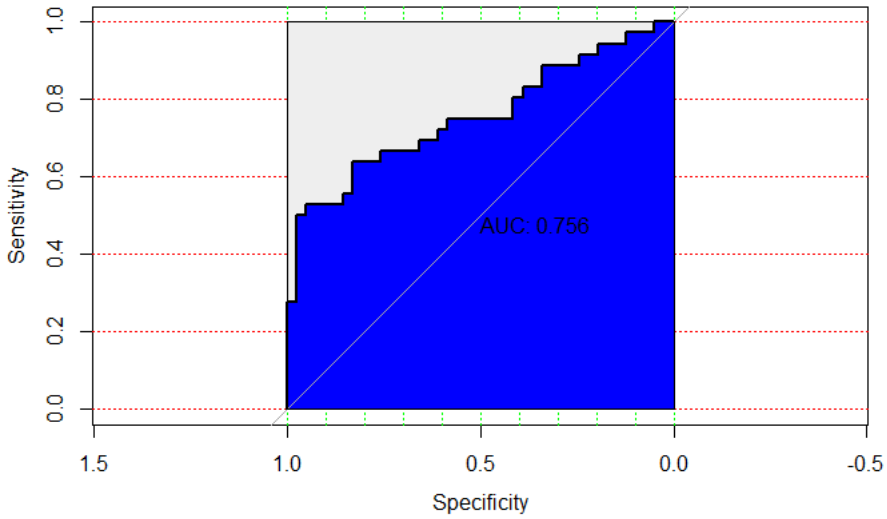



**Figure 8 The ROC of Model3.1**

## 4. Discussion

**· Data Description and Processing**

Figure 1 showed that the sample size of each level of ‘meander’ was relatively balanced. At the same time, each level presented a different contribution to ‘failed’ and ‘not failed’. Therefore, the four levels were all considered when fitting the model.

From Figure 2 (a), it could be seen that all the river bank of samples with level ‘open water’ failed. Although the number of observations with ‘open water’ was small, taking the practical significance into account, it was certain that it was very unscientific to build the house in a place where the cover was open water under normal circumstances. In this situation, when the landcover was ‘open water’, the riverbank was certain to fail no matter what the other predictors were like. Therefore, the observations with ‘open water’ as the landcover was removed. At the same time, because the differences between the sample size of the remaining three levels were too large, level ‘grassy’ and ‘agriculture’ were combined. Figure2 (b) showed that ‘grassy/agriculture’ and ‘forest’, which mainly needed to be investigated, had similar sample sizes after conducting levels combination.

It could be seen in Figure 3 (a) that there was strong polarization in the distribution of the variable ‘dredging’. This meant that if dredging was added as a numerical variable into the model fitting, it would affect the accuracy of the model because of its poor continuity. Therefore, the variable 'dredging' was divided into two categories: 'none or some dredging' and 'a lot of dredging' according to the amount of dredging. However, when Figure 3 (b) was analysed, it showed that approximately half of the observations with ‘none or some dredging’ failed, so did observations with ‘a lot of dredging’, which might present that the variable ‘dredging’ have no effect on river bank. This idea was just a comment on the figure, which needed to be checked when fitting the model.

For numerical variable ‘sinuosity’, it could be figured out from Figure 4 that the river bank of observations with small sinuosity had a lower possibility to fail. In Figure 5(a) and Figure 6(a), it seemed that the river bank of observations with higher channel width and lower vegewidth had a higher failure possibility. However, if the density curves shown in Figure 5(b) and Figure 6(b) in were checked, there was no big difference between the two distributions for ‘failed’ and ‘not failed’, which meant that the differences in boxplots were led by some extreme points.

**· Model Fitting Selection**

Model2.1 and Model3.1, which was with interaction items in it were mainly compared at last. Table 5 displayed that the extra item, which was the interaction was significant for the model since the p-value was low. That meant the accuracy of the model would decrease if the interaction is removed.

When checking the prediction ability, since the dataset contained only 77 observation points after removing the 5 observations with ‘open water’, it was not reasonable to divide data sets into training set and test set. When the data set was small, each observation point would have an impact on the model. In order to increase the accuracy of the model, the whole data set was selected when fitting the model. In this case, the whole data set was the choice of checking the prediction ability of the model.

From the confusion table, it could be seen that although both of Model 2.1 and Model 3.1 had a good

prediction accuracy, the Model 3.1 offered a lower probability of the case which was predicted not to fail but actually fail. This type of prediction error would result in a horrible consequence.

What is more, Table 6 showed that the prediction accuracy of Model3.1 was higher, At the same time, the AIC-value of Model3.1 was lower and the AUC of Model3.1 in ROC was larger (shown in Figure 7 and Figure 8) as well, so Model3.1 was chosen as the finally model.

**· Model Interpretation**

Model 3.1 was the final model, which included variables ' meander', 'sinuosity', and interaction between meander and sinuosity. In Table 4, When 'meander' was chute, the coefficient was negative, which seemed different from the reality in Figure 1. In fact, in the obtained model 3.1, the role of chute was mainly reflected by the interaction with an extremely low coefficient. Sinuosity would enlarge the effect of chute and lead to the failure of river bank.

## 5. Advice

According to this research, in order to ensure the success of river stop bank, which means that P(failure=1) needs to be minimized, the customer need to take sinuosity and meander into account when choosing the location. The advice is trying to select a place where is not covered by open water, on the outside bend of the river or a straight, and the river nearby has a lower sinuosity.

# Question 2

## 1. Introduction

The Galapagos penguin is the only kind of penguins living on the equator on earth. This magical creature lives in the Galapagos Islands, which is an archipelago 1000 kilometres away from the South American continent.

Although the area is less than inland, the biological property of island still has some interesting features. The sea is a barrier for some creatures, but for others, it is a vehicle. At the same time, the creatures on the island will evolve new characteristics in the isolated environment of islands.

For biologists, they need to estimate the species of islands before conducting actual research on islands. The main purpose of this study is to determine which characteristics are related to biological species on islands based on the collected Island instances.

This research took the characteristics of 78 islands and the number of their endemic species into account in total. The area, elevation and isolation of islands which were presented by the distance of the island to the nearest mainland (in km) were measured.

## 2. Method

### 2.1 Data Description and Processing

First, dim () function and summary () function were conducted to help have a general understanding of the whole dataset. Dim () showed the dimension of the dataset, while summary () could provide basic statistics for each variable. Next, because the responding variable needed to be paid more attention, using the hist () function to show the histogram, which was used to display the distribution of the data, and the boxplot () function to offer a boxplot, which could clearly tell the skewness and outliers.

At the same time, in order to learn the distribution of each variables as well as the relationship between them, using ggpairs() function to present the scatter matrix plot of the dataset except the first column which is the name of islands.

### 2.2 Model Fitting and Selection

When a series of continuous and categorical predictors are used to predict a counting variable, Poisson distribution should be conducted. In Bibersmall dataset, the responding variable 'Endemic' recorded the number of endemic species known from the island in historic times. Therefore, glm () function with parameter ‘family’ being set into ‘Poisson’ was implemented to fit the data.

Over dispersion refers to the variance of the responding variable is greater than the variance of the distribution [3]. The main consequence of over dispersion is the poor accuracy of the significance test, which leads to the reduction of the interpretability [3]. Use the qcc package to test the model 2 to figure out if there was over dispersion. If the test results showed that there was over dispersion, the data needed to be fitted again using quasi-Poisson.

In the process of model fitting, meaningful interaction items could help interpret the model better. Therefore, add interaction items to quasi Poisson to fit the data. Then drop the predictors with great p-value one by one and get the model with interaction items.

When selecting the model, use anova() function to compare Model 4 and Model 6 to figure out if it was necessary to add the extra items into the model.

## 3. Result

The dimension and summary statistics were shown as followed in Table 1. At the same time, the histogram and boxplot was displayed in Figure 1. It could be seen that the distribution of 'Endemic' was right skewness and there was an outlier.

**Table1 Dimension and Statistics of Bibersmmall**

| island | | area | | elevation | | isolation | | endemic | |
|---|---|---|---|---|---|---|---|---|---|
| Admiraltys | 1 | Min. | 4 | Min. | 6 | Min. | 50 | Min. | 1 |
| Aldabra | 1 | 1st Qu. | 100.2 | 1st Qu. | 332 | 1st Qu. | 270 | 1st Qu. | 2 |
| Amsterdam | 1 | Median | 695 | Median | 865 | Median | 785 | Median | 4.5 |
| Andamans | 1 | Mean | 7555 | Mean | 1195.8 | Mean | 1533 | Mean | 6.474 |
| Annnobon | 1 | 3rd Qu. | 3875 | 3rd Qu. | 1837.5 | 3rd Qu. | 2400 | 3rd Qu. | 10 |
| Antipodes | 1 | Max. | 150000 | Max. | 4200 | Max. | 5800 | Max. | 26 |
| (other) | 72 | | | | | | | | |
| ncol | 5 | | | nrow | 78 | | | | |

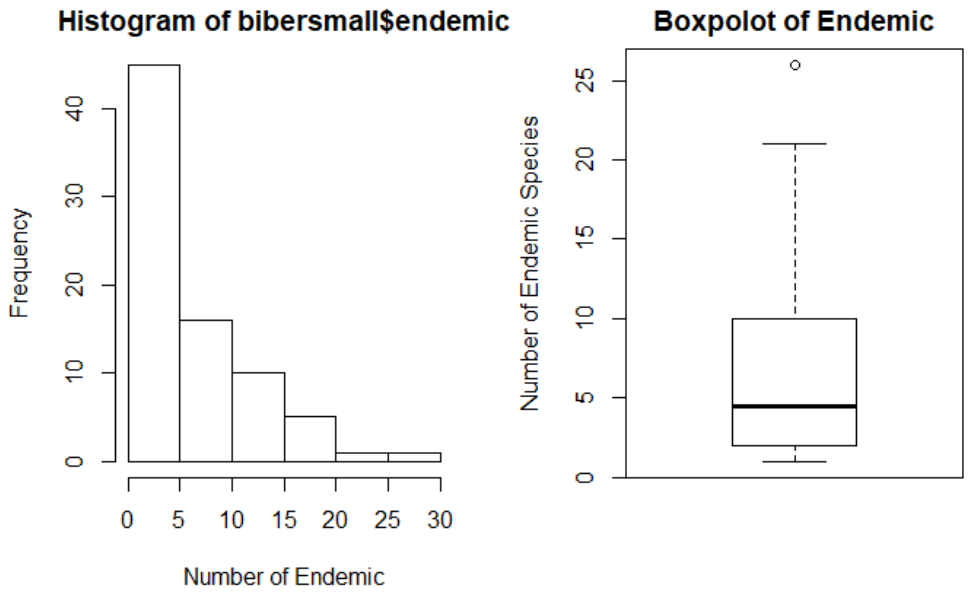



**Figure1 Histogram and Boxplot of Endemic**

The scatter matrix plot of all variables was shown as below in Figure2. The figure told the correlations of every two variables and the distribution of each predictor.

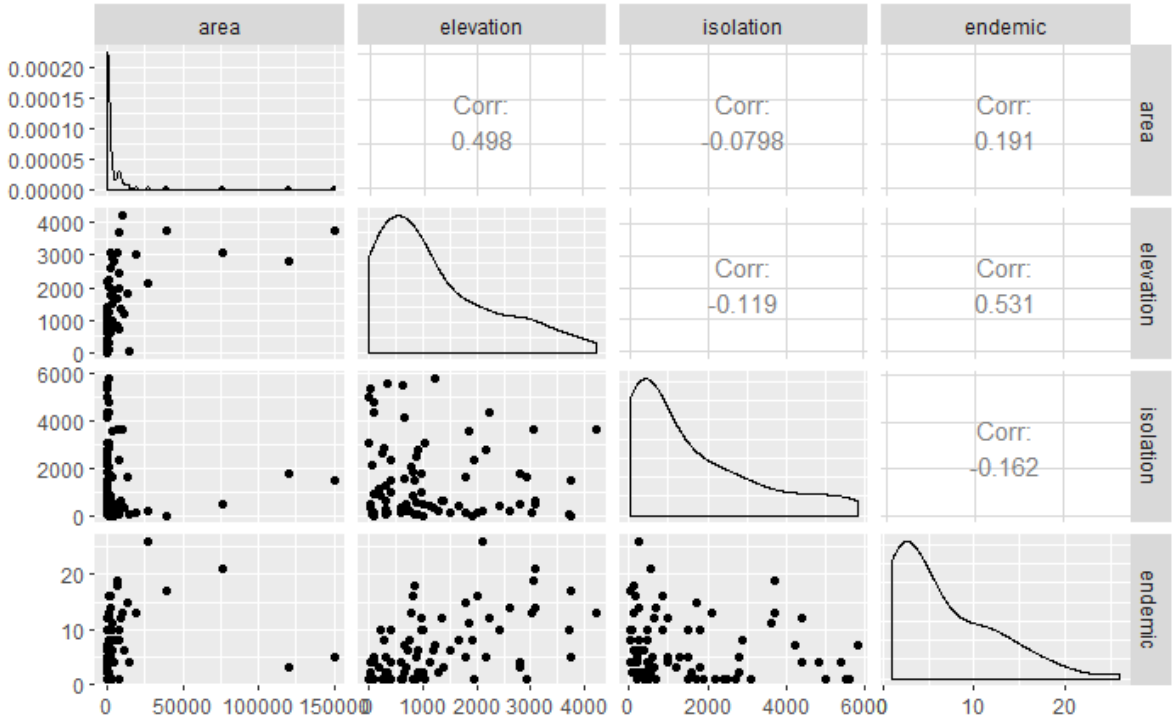



**Figure2 Scatter Matrix Plot**

After Poisson regression, the whole model was got as model 1 and the coefficients and statistics were shown in Table1 in Appendix. We could see that the p value of 'area' is greatest.

**Endemic ~ area + elevation + isolation** **(Model 1)**

The variable with the largest p-value (greater than 0.05) was then dropped and Model 2 was then obtained.

**Endemic ~ elevation + isolation** **(Model 2)**

**Table2 Coefficients and Statistics of Model2**

| | estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 1.456 | 8.74E-02 | 16.65 | <2e-16 |
| elevation | 3.59E-04 | 3.61E-05 | 9.941 | <2e-16 |
| isolation | -7.26E-05 | 3.09E-05 | -2.352 | 0.0187 |
| AIC | 523.15 | | | |

The result of over dispersion checking was shown in Table3, which presented a low p-value.

**Table3 Result of Over Dispersion Checking**

| Over dispersion | test | Obs.Var/Theor.Var | Statostic | p-value |
|---|---|---|---|---|
| Poisson | data | 4.998 | 384.8178 | 0 |

Using quasi Poisson regression to fit the model again and Model 3 was got, which was displayed in Table 2 in Appendix. After dropping the variables with large p-value one by one, Model 4 was obtained and the coefficients and statistics were displayed as followed.

**Table4 Coefficients and Statistics of Model4**

| | estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 1.35E+00 | 1.43E-01 | 9.407 | 2.22E-14 |
| elevation | 3.65E-04 | 6.89E-05 | 5.360 | 8.61E-07 |
| Residual Deviance | | 262.75 | | |

The final model with interaction was Model6, of which coefficients and statistics were shown in Table5.

**Table5 Coefficients and Statistics of Model6**

| | estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 1.22E+00 | 1.49E-01 | 8.205 | 5.30E-12 |
| area | 4.25E-05 | 1.89E-05 | 2.254 | 0.0272 |
| elevation | 4.37E-04 | 7.68E-05 | 5.682 | 2.58E-07 |
| area:elevation | -1.42E-08 | 6.17E-09 | -2.299 | 0.0243 |
| Residual Deviance | | 240.22 | | |

The result of model comparison which was conducted using anova() was shown as followed.

**Table6 Comparison of Model 4 and Model6**

| | Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|---|---|---|---|---|
| Model4 | 76 | 262.75 | | | |
| Model6 | 74 | 240.22 | 2 | 22.534 | 0.03149 |

## 4. Discussion

**· Data Description and Processing**

From the boxplot provided by Figure1, it could be seen that there was an obvious outlier in the distribution of response variable 'endemic'. Since the reason for outliers could not be figured out, and the dataset was not large enough, the outlier was still retained when fitting data. At the same time, the distribution of responding variable was right skewed, which could not meet the assumption of linear regression, therefore, using generalised linear model.

In Figure 2, it could be found out that the number of endemic had a relatively stronger relationship with elevation. What is more, the correlation between area and elevation was not low as well, which might lead to the interaction.

**· Model Fitting Process**

In Model1, elevation and isolation were both significant in the model fitting. However, the over dispersion checking result presented a zero p-value, which meant the over dispersion was extremely serious. Considering the characteristics of the corresponding variables, the reason for the over

dispersion should be the dependence between different species in islands. Generally speaking, the existence of a species could alter the entire biological chain, so the emergence and extinction of all species were related.

The coefficients of Model 3 obtained by quasi Poisson regression was the same as Model 1, but the standard error increased. After considering the over dispersion, isolation was not significant in the model.

The reason why Model 6 was chosen instead of Model 4 was, firstly, Table6 showed a P value, which indicated that interaction and elevation would affect the accuracy of the model if they were removed. In addition, model 6 had a lower residual deviance.

· **Model Interpretation**

In Model 6, the predictors ‘area’ and ‘elevation’ played an important role in the model which was used to predict the number of endemic species. Table 5 told that when area and elevation increases, the number of endemic will increase. The coefficient of the interaction was negative, which meant that the multiple interaction was lower than the product of the effect of those two variables.

## 5. Advices

According to the model obtained by this study, if the island has a large area and a great highest elevation, the number of endemic species on them has a high possibility to be large.

On a 20,000ha island that was 100km from the mainland with a maximum elevation of 1000m, the number of endemic species should be 9 based on Model6.

# Appendix 1 Figures and Tables

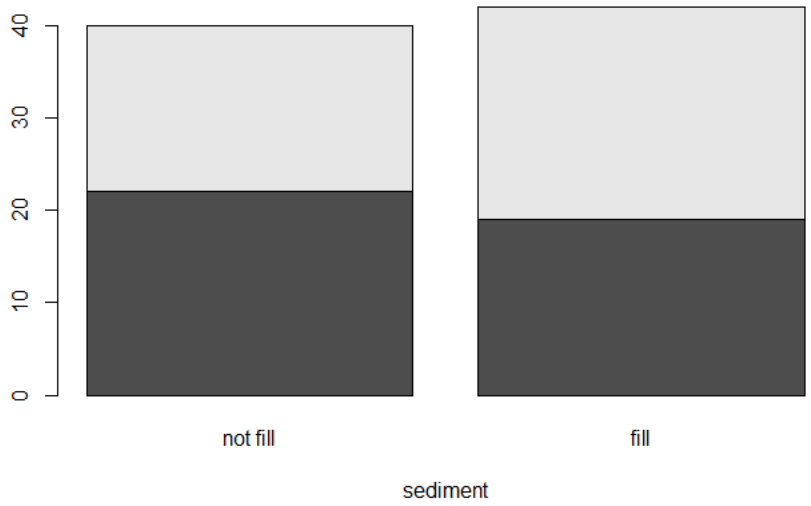
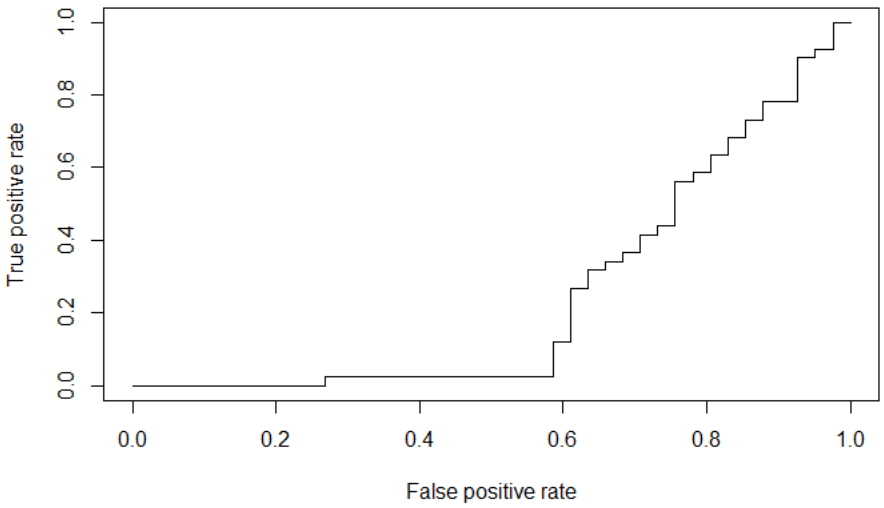



**Figure1 Bar chart of Sediment**




**Figur2 ROC of Model 2.2**

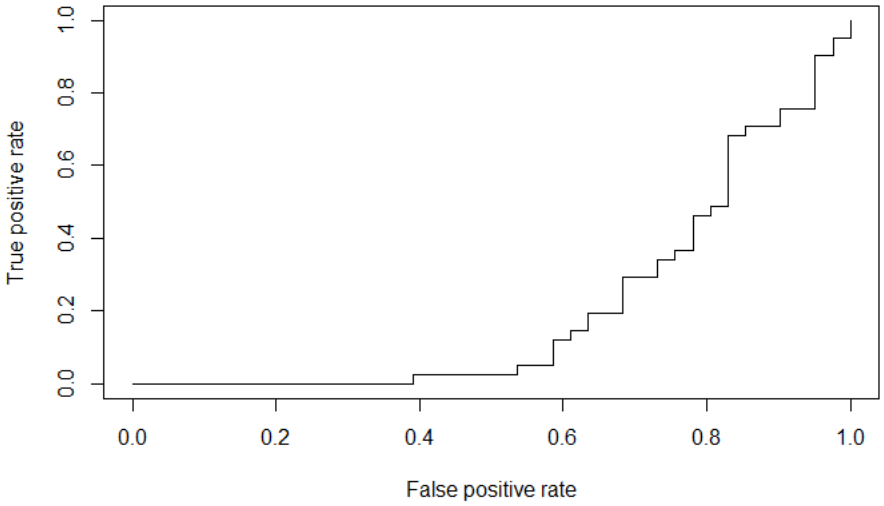



**Figure2 ROC of Model 3.1**

**Table1 Coefficients and Statistics of Model1**

| | estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 1.44E+00 | 8.78E-02 | 16.373 | <2e-16 |
| area | -2.91E-06 | 1.79E-06 | -1.620 | 0.1053 |
| elevation | 3.92E-04 | 4.08E-05 | 9.621 | <2e-16 |
| isolation | -7.42E-05 | 3.07E-05 | -2.419 | 0.0156 |
| AIC | 522.32 | | | |

**Table2 Coefficients and Statistics of Model3**

| | estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 1.44E+00 | 1.67E-01 | 8.629 | 8.37E-13 |
| area | -2.91E-06 | 3.40E-06 | -0.854 | 0.396 |
| elevation | 3.92E-04 | 7.74E-05 | 5.070 | 2.84E-06 |
| isolation | -7.42E-05 | 5.82E-05 | -1.275 | 0.206 |

# Appendix 2 R Code

## R Code for Question1

```{r}
riverdata<-read.table('D:/2018 first semester/linear regression/ASSIGNMENT2/RiverBank.txt',header = TRUE)
names(riverdata)
```
```{r}
dim(riverdata)
summary(riverdata)
head(riverdata)
hist(riverdata$dredging)
```
```{r}
failure_1<-riverdata$failure
riverdata$failure<-factor(failure_1,levels = c(0,1),labels = c('not failed','failed'))
table(riverdata$failure)
sediment_1<-riverdata$sediment
riverdata$sediment<-factor(sediment_1,levels = c(0,1),labels = c('not fill','fill'))
meander_1<-riverdata$meander
riverdata$meander<-factor(meander_1,levels = c(1,2,3,4),labels = c('Inside bend','outside bend','chute','straight'))
landcover_1<-riverdata$landcover
riverdata$landcover<-factor(landcover_1,levels = c(1,2,3,4),labels = c('open water','grassy','agriculture','forest'))
boxplot(riverdata$channelwidth~riverdata$failure,xlab = 'failure',ylab='channel width')
plot(density(riverdata$channelwidth[riverdata$failure=='not failed']),col='red',main='distribution of channelwidth',xlab='channelwidth')
lines(density(riverdata$channelwidth[riverdata$failure=='failed']),col='blue')
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
boxplot(riverdata$vegewidth~riverdata$failure,xlab = 'failure',ylab='vegewidth')
plot(density(riverdata$vegewidth[riverdata$failure=='not failed']),col='red',main='distribution of vegewidth',ylim=c(0,0.001),xlab='vegewidth')
lines(density(riverdata$vegewidth[riverdata$failure=='failed']),col='blue')
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
boxplot(riverdata$sinuosity~riverdata$failure,xlab = 'failure',ylab='sinuosity')
plot(density(riverdata$sinuosity[riverdata$failure=='not failed']),col='red',main='distribution of sinuosity',xlab='sinuosity')
lines(density(riverdata$sinuosity[riverdata$failure=='failed']),col='blue')
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
riverdata$dredging[ (riverdata$dredging <= 2*10^5)] <- '0'
riverdata$dredging[ (riverdata$dredging >= 4*10^5 ) ] <- '1'
library(ggplot2)
ggplot(riverdata,aes(landcover))+geom_bar(stat ='count')
ggplot(riverdata,aes(meander))+geom_bar(stat='count')
riverdata$dredging <- factor(riverdata$dredging,levels=c(0,1),labels=c('none or some dredging','a lot of dredging'))
```

```
barplot(table(riverdata$failure,riverdata$sediment),xlab = 'sediment')
barplot(table(riverdata$failure,riverdata$dredging),xlab = 'dredging')
text.legend=c('failed','not failed')
legend('topright',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
barplot(table(riverdata$failure,riverdata$meander),xlab = 'meander')
table(riverdata$failure,riverdata$meander)
barplot(table(riverdata$failure,riverdata$landcover),xlab = 'landcover')
text.legend=c('failed','not failed')
legend('topleft',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
landcover_1[which(landcover_1==2)]=3
riverdata$landcover<-factor(landcover_1,levels = c(1,3,4),labels = c('open
water','grassy/agriculture','forest'))
barplot(table(riverdata$failure,riverdata$landcover),xlab = 'landcover')
text.legend=c('failed','not failed')
legend('topleft',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
summary(riverdata)
  # delete the observations with open water
which(riverdata$landcover=='open water')
#16 36 38 39 40
riverdata = riverdata[-40,]
riverdata = riverdata[-39,]
riverdata = riverdata[-38,]
riverdata = riverdata[-36,]
riverdata = riverdata[-16,]
which(riverdata$landcover=='open water')
  # factor defined again
failure_1<-riverdata$failure
riverdata$failure<-factor(failure_1,levels = c(0,1),labels = c('not failed','failed'))
table(riverdata$failure)
sediment_1<-riverdata$sediment
riverdata$sediment<-factor(sediment_1,levels = c(0,1),labels = c('not fill','fill'))
meander_1<-riverdata$meander
riverdata$meander<-factor(meander_1,levels = c(1,2,3,4),labels = c('Inside bend','outside
bend','chute','straight'))
landcover_1<-riverdata$landcover
riverdata$landcover<-factor(landcover_1,levels = c(2,3,4),labels = c('grassy','agriculture','forest'))
boxplot(riverdata$channelwidth~riverdata$failure,xlab = 'failure',ylab='channel width')
 #plot data again
plot(density(riverdata$channelwidth[riverdata$failure=='not failed']),col='red',main='distribution of
channelwidth',xlab='channelwidth')
lines(density(riverdata$channelwidth[riverdata$failure=='failed']),col='blue')
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
boxplot(riverdata$vegewidth~riverdata$failure,xlab = 'failure',ylab='vegewidth')
plot(density(riverdata$vegewidth[riverdata$failure=='not failed']),col='red',main='distribution of
vegewidth',ylim=c(0,0.001),xlab='vegewidth')
lines(density(riverdata$vegewidth[riverdata$failure=='failed']),col='blue')
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
boxplot(riverdata$sinuosity~riverdata$failure,xlab = 'failure',ylab='sinuosity')
plot(density(riverdata$sinuosity[riverdata$failure=='not failed']),col='red',main='distribution of
sinuosity',xlab='sinuosity')
lines(density(riverdata$sinuosity[riverdata$failure=='failed']),col='blue')
```

```
legend('topright',lty = c(1,1),legend=c('not failed','failed'),col=c('red','blue'))
riverdata$dredging[ (riverdata$dredging <= 2*10^5)] <- '0'
riverdata$dredging[ (riverdata$dredging >= 4*10^5 ) ] <- '1'
library(ggplot2)
ggplot(riverdata,aes(landcover))+geom_bar(stat ='count')
ggplot(riverdata,aes(meander))+geom_bar(stat='count')
riverdata$dredging <- factor(riverdata$dredging,levels=c(0,1),labels=c('none or some dredging','a lot of
dredging'))
barplot(table(riverdata$failure,riverdata$sediment),xlab = 'sediment')
barplot(table(riverdata$failure,riverdata$dredging),xlab = 'dredging')
text.legend=c('failed','not failed')
legend('topright',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
barplot(table(riverdata$failure,riverdata$meander),xlab = 'meander')
table(riverdata$failure,riverdata$meander)
barplot(table(riverdata$failure,riverdata$landcover),xlab = 'landcover')
text.legend=c('failed','not failed')
legend('topleft',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
landcover_1[which(landcover_1==2)]=3
riverdata$landcover<-factor(landcover_1,levels = c(3,4),labels = c('grassy/agriculture','forest'))
barplot(table(riverdata$failure,riverdata$landcover),xlab = 'landcover')
text.legend=c('failed','not failed')
legend('topleft',pch = c(15,15),legend=text.legend,col=c('grey','black'),cex=0.8)
```

```
 # modeling
model1<-
glm(riverdata$failure~riverdata$sediment+riverdata$meander+riverdata$channelwidth+riverdata$landcove
r+riverdata$vegewidth+riverdata$sinuosity+riverdata$dredging, family = binomial())
summary(model1)
model2<-step(model1)
summary(model2)
anova(model2,test='Chisq')
model2.1<-update(model2,~.-riverdata$dredging)
summary(model2.1)
anova(model2.1,test='Chisq')
anova(model2.1,model2,test='Chisq')
full_model<-glm(riverdata$failure~(riverdata$meander + riverdata$sinuosity +
   riverdata$landcover)^2, family = binomial())
model3.1<-step(full_model, direction = 'backward')
summary(model3.1)
anova(model3.1,test='Chisq')
anova(model2.1,model3.1,test = 'Chisq')
```

```
 # plot the residual and rstudent
```{r}
plot(predict(model2.1,type = 'response'),residuals(model2.1,type = 'deviance'), ylab='Residual of Model 2.1
and Model3.2')
points(predict(model3.1,type = 'response'),residuals(model3.1,type = 'deviance'),col='red')
plot(rstudent(model2.1))
points(rstudent(model3.1),col='red')
   # test if there is over dispersion
model3.1_od<-glm(riverdata$failure~riverdata$meander*riverdata$sinuosity,family = quasibinomial())
```
```

```
pchisq(summary(model3.1_od)$dispersion*model3.1$df.residual,model3.2$df.residual,lower=F)
  # ROC drawing
```{r}

pre2.1=predict(model2.1, type='response')
pre3.1=predict(model3.1, type='response')
library(ROCR)
library(gplots)
pred2.1=prediction(pre2.1,riverdata$failure)
performance(pred2.1,'auc')@y.values
perf2.1=performance(pred2.1,'tpr','fpr')
plot(perf2.1)
pred3.1=prediction(pre3.1,riverdata$failure)
performance(pred3.1,'auc')@y.values
perf3.1=performance(pred3.1,'tpr','fpr')
plot(perf3.1)
library(pROC)
modelroc=roc(riverdata$failure,pre2.1)
plot(modelroc,print.auc=TRUE,auc.polygon=TRUE,grid=c(0.1,0.2),grid.col=c("green","red"),max.auc.polygon
=TRUE,auc.polygon.col="blue")
modelroc=roc(riverdata$failure,pre3.1)
plot(modelroc,print.auc=TRUE,auc.polygon=TRUE,grid=c(0.1,0.2),grid.col=c("green","red"),max.auc.polygon
=TRUE,auc.polygon.col="blue")
`
```

## R Code for Question2

```
bibersmall<-read.table('D:/2018 first semester/linear regression/ASSIGNMENT2/Bibersmall.txt', header =
TRUE)
dim(bibersmall)
summary(bibersmall)
par(mfrow=c(1,2))
hist(bibersmall$endemic,xlab = 'Number of Endemic')
boxplot(bibersmall$endemic, ylab = 'Number of Endemic Species',main='Boxpolot of Endemic')
  # fit the model
fit_endemic<-glm(bibersmall$endemic~bibersmall$area+bibersmall$elevation+bibersmall$isolation, family =
poisson())
summary(fit_endemic)
fit_endemic2<-update(fit_endemic,~.-bibersmall$area)
summary(fit_endemic2)
exp(coef(fit_endemic2))
  # check if there is over dispersion
library(qcc)
qcc.overdispersion.test(bibersmall$endemic, type = 'poisson')
  # fit the quasipoisson model
fit_endemic_qu<-glm(bibersmall$endemic~bibersmall$area+bibersmall$elevation+bibersmall$isolation,
family = quasipoisson())
summary(fit_endemic_qu)
fit_endemic_qu2<-update(fit_endemic_qu, ~.-bibersmall$area)
summary(fit_endemic_qu2)
fit_endemic_qu3<-update(fit_endemic_qu2, ~.-bibersmall$isolation)
summary(fit_endemic_qu3)
```
```

```
exp(coef(fit_endemic_inter2))
  # add interaction to refit
fit_endemic_inter<-glm(bibersmall$endemic~bibersmall$area*bibersmall$elevation*bibersmall$isolation,
family = quasipoisson())
summary(fit_endemic_inter)
fit_endemic_inter1<-
glm(bibersmall$endemic~bibersmall$area+bibersmall$elevation+bibersmall$isolation+bibersmall$area:bibe
rsmall$elevation, family = quasipoisson())
summary(fit_endemic_inter1)
fit_endemic_inter2<-update(fit_endemic_inter1,~.-bibersmall$isolation)
summary(fit_endemic_inter2)
anova(fit_endemic_qu3,fit_endemic_inter2,test= 'Chisq')
 #calculate the number of endemic species
exp(1.219+4.253*10^(-5)*20000+4.365*10^(-4)*1000-1.419*10^(-8)*20000000)
```

## Reference


[1] WIKIPEDIA. https://en.wikipedia.org/wiki/River_bank_failure

[2] BAIDUBAIKE. https://baike.baidu.com/item/%E9%95%BF%E6%B1%9F/388?fr=aladdin

[3] Gareth J, Daniela W, etc. <An Introduction to Statistical Learning with Applications in R>

[4] Yi X, Liping C, etc. <Statistical model and R>