Report II - SF2930 Regression Analysis

Chen Hanmo, 990904-T072

10 March, 2019

Contents

1 Introduction

1.1 Background

In Sweden, most of the vehicles are required by law to have a third part liability insurance. Many tractor owners complement this legally required insurance with an insurance covering vehicle damage to their own tractor.

And the insurance company must find a way of the insurance pricing because different tractors usually have different risks, thus leading to different insurance prices.

1.2 Goals

In general, we are tring to develop a price model like this

$$\text{price}_i = \gamma_0 \prod_{k=0}^{M} \gamma_{k,i} \tag{1}$$

in which γ_0 is the base price level and $\gamma_{k,i}, k = 1, 2, \cdots, M, i = 1, 2, \cdots, n$ are the risk factors corresponding to variable number k and variable group number i.

What we need to do is

1. **Grouping and risk differentiation**
    - Seperate the tractors into several groups. For each group, the risk can be seen to be **homogeneous**.
    - Perform **Likelihood Ratio Test** to the division of groups.
    - Use GLM in different groups to generate the specifit risk factor $\gamma_{k,i}$
2. **Calculate the basic level** γ_0
    - Use the history data to estimate the expected claim cost.
    - Use the specifit risk factor $\gamma_{k,i}$ to calculate the total risk factor
    - Use the expected claim cost and total risk factor to find the basic level γ_0.

The whole model-builing process is illstrated in the follwing flow chart[1].

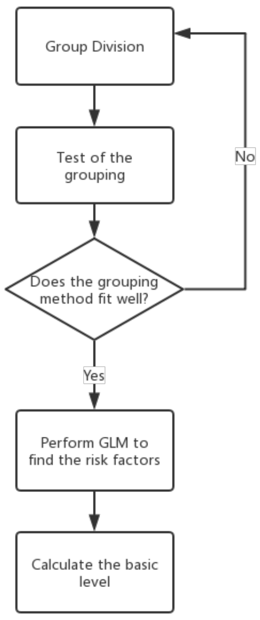

Group Division
Test of the grouping
No
Does the grouping method fit well?
Yes
Perform GLM to find the risk factors
Calculate the basic level

Figure 1: the model-building process

2 Model Development

2.1 The division of Groups

First, we need to determine the number of groups we want to divide into. There are two considerations:

- If we divide them into too many groups, there must be some groups containing less than 5 (maybe even none) individuals, which is not approriate for the following GLM regression because it may increase the error.
- If we divide them into too few groups, the risk factor in each group will differ greatly and would violate the ”risk homogeneous” condition.

After these two considerations, we think $4 \sim 8$ groups of each variable is preferable as a trade-off.

Then we should determine what variables to use. The dataset contains *RiskYear, VehicleAge, Weight, Climate, ActivityCode, Duration, NoOfClaims, ClaimCost.*

- We first rule out the *RiskYear* variable under the assumption that the claim costs don’t vary much with time. If we should take time as a factor, it may be concerned with the subject of Time Series Analysis, which isn’t our focus in this project.
- The *NoOfClaims* variable is also a factor to be considered. We have tried different ways of handles this, like dividing the claim costs by the number of claims to get a mean claim cost, or to just view one single tractors with several claims as several tractors which have only one claim.
- The *Duration* variable is similar with the *NoOfClaims*. They will certainly influence the claim costs, however, we think it should be considered in the final pricing model but not the original factor model.

Due to those reasons, We finally use the *VehicleAge, Weight, Climate, ActivityCode* variable as our factors.

Then we will discuss different methods of division.

2.1.1 The original method

The original code gives a simple method of dividing.

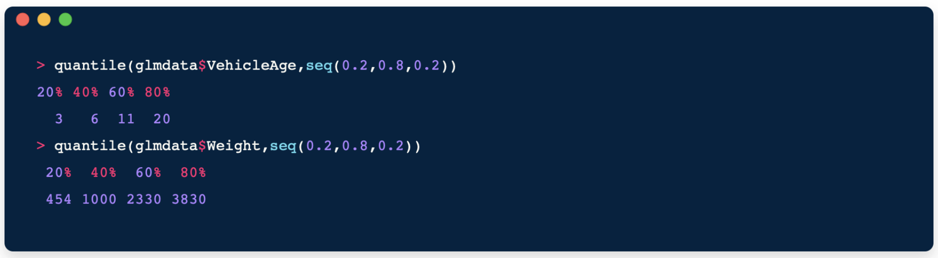
- **4 VehicleAge Groups**: 0-2 years; 3-5 years; 6-14 years;$\geqslant$ 15 years.
- **6 Weight Groups**: 0-999 kg; 1000-1999 kg; 2000-2999 kg; 3000-3999 kg; 4000-4999 kg; $\geqslant$ 5000 kg.
- **Climate and ActivityCode Groups** is divided in the natural way.

It is simple and just works out fine.

2.1.2 Using the $k-$quantiles

The basic idea of this method is to split the tractors into k groups with (almost) the same size, that’s why we use the $k-$ quantiles.

For example, if you want to divide the tractors into 5 groups by vehicle age or weight,



```
> quantile(glmdata$VehicleAge,seq(0.2,0.8,0.2))
20% 40% 60% 80%
  3   6  11  20
> quantile(glmdata$Weight,seq(0.2,0.8,0.2))
 20%  40%  60%  80%
 454 1000 2330 3830
```

Figure 2: Divided by $k-$ quantiles

Thus the division is,

- **5 VehicleAge Groups**: 0-2 years; 3-5 years; 6-10 years;11-19 years;$\geqslant$ 20 years.
- **5 Weight Groups**: 0-453 kg; 454-999 kg; 1000-2329 kg; 2330-3829 kg; $\geqslant$ 3800 kg.
- **Climate and ActivityCode Groups** is divided in the natural way.

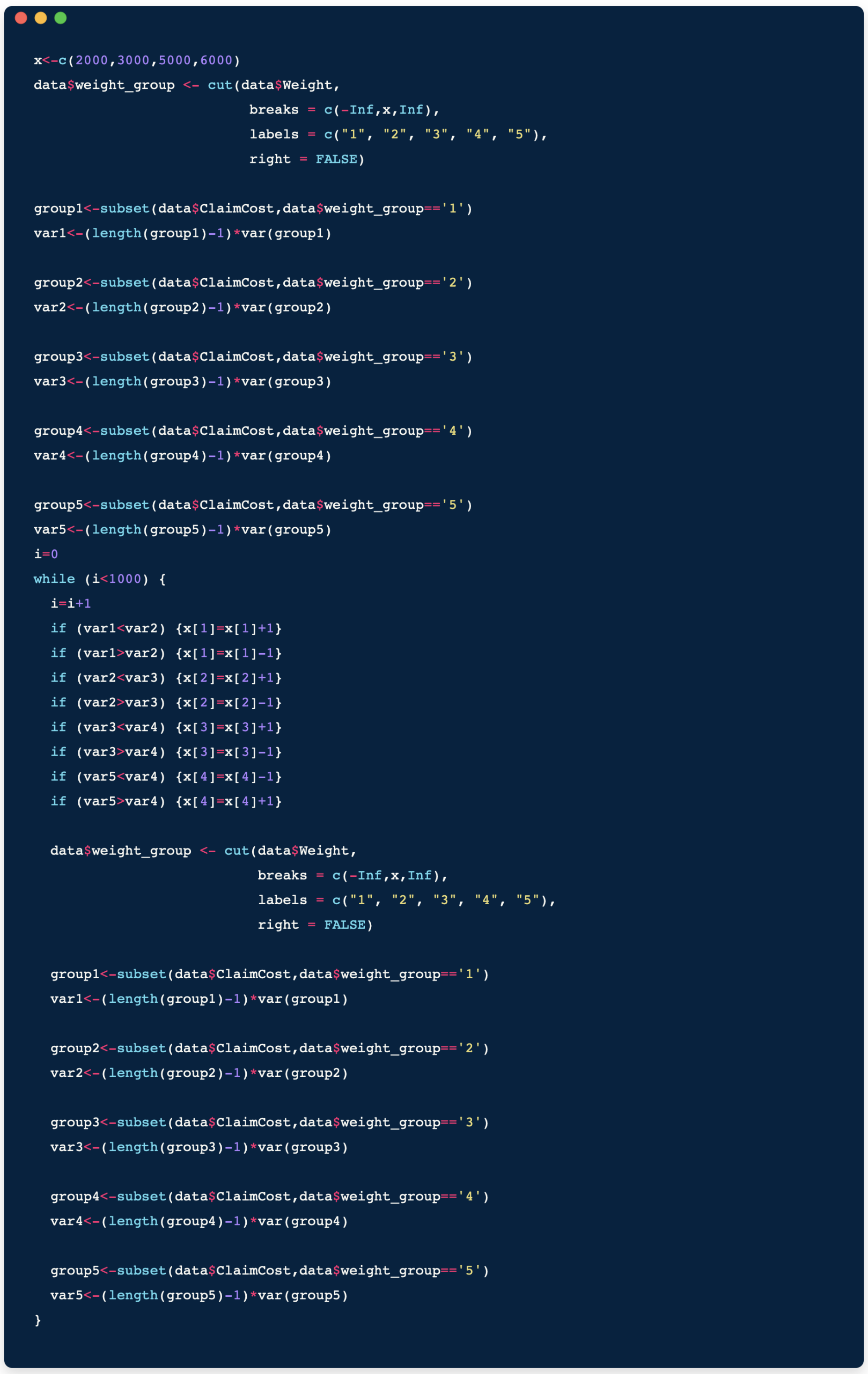
It improves slightly from the original method, however, although it ensures that it is equally split in one dimention, when divided by different variables, there would be groups that contain few and even none tractors, which is a problem.

2.1.3 The minmax theorm

The last method concentrates on controlling the size of groups, and this and the following method will concentrate on how to make individuals inside each group **risk homogeneous**.

One indicator of the difference between individuals inside the group is the **variance** or **the sum of squared errors(SST)**. So our job is to control the SST of each group as low as possible, or to control the maximum SST within the group of all groups as low as possible. Using the min-max theorm, to achieve this, we can set all groups' SST within the group to be equal, given the group number k.

For example, if we want to divide them into 5 groups by weight,



```
x<-c(2000,3000,5000,6000)
data$weight_group <- cut(data$Weight,
                          breaks = c(-Inf,x,Inf),
                          labels = c("1", "2", "3", "4", "5"),
                          right = FALSE)

group1<-subset(data$ClaimCost,data$weight_group=='1')
var1<-(length(group1)-1)*var(group1)

group2<-subset(data$ClaimCost,data$weight_group=='2')
var2<-(length(group2)-1)*var(group2)

group3<-subset(data$ClaimCost,data$weight_group=='3')
var3<-(length(group3)-1)*var(group3)

group4<-subset(data$ClaimCost,data$weight_group=='4')
var4<-(length(group4)-1)*var(group4)

group5<-subset(data$ClaimCost,data$weight_group=='5')
var5<-(length(group5)-1)*var(group5)
i=0
while (i<1000) {
  i=i+1
  if (var1<var2) {x[1]=x[1]+1}
  if (var1>var2) {x[1]=x[1]-1}
  if (var2<var3) {x[2]=x[2]+1}
  if (var2>var3) {x[2]=x[2]-1}
  if (var3<var4) {x[3]=x[3]+1}
  if (var3>var4) {x[3]=x[3]-1}
  if (var5<var4) {x[4]=x[4]-1}
  if (var5>var4) {x[4]=x[4]+1}

  data$weight_group <- cut(data$Weight,
                           breaks = c(-Inf,x,Inf),
                           labels = c("1", "2", "3", "4", "5"),
                           right = FALSE)

  group1<-subset(data$ClaimCost,data$weight_group=='1')
  var1<-(length(group1)-1)*var(group1)

  group2<-subset(data$ClaimCost,data$weight_group=='2')
  var2<-(length(group2)-1)*var(group2)

  group3<-subset(data$ClaimCost,data$weight_group=='3')
  var3<-(length(group3)-1)*var(group3)

  group4<-subset(data$ClaimCost,data$weight_group=='4')
  var4<-(length(group4)-1)*var(group4)

  group5<-subset(data$ClaimCost,data$weight_group=='5')
  var5<-(length(group5)-1)*var(group5)
}
```

Figure 3: the minmax method

and it is the same working with *VehicleAge*. Using this method, we get division like this:

- **5 VehicleAge Groups**: 0-2 years; 3-5 years; 6-8 years;9-10 years;$\geqslant$ 11 years.
- **5 Weight Groups**: 0-2601 kg;2602-3829 kg;3830-4169kg; 4170-5829;$\geqslant$ 5830 kg.
- **Climate and ActivityCode Groups** is divided in the natural way.

2.1.4 The Otsu's method

This method comes from the digital image processing, is used to automatically perform clustering-based image thresholding. However, in this project, we can use this method to find the threshold of our .

In Otsu's method, we use the exhausive search for the threshold that minimizes the intra-class variance (the variance within the class), defined as a weighted sum of variances of the two classes:

$$\sigma_w^2 = \omega_0 \sigma_0^2 + \omega_1 \sigma_1^2 \tag{2}$$

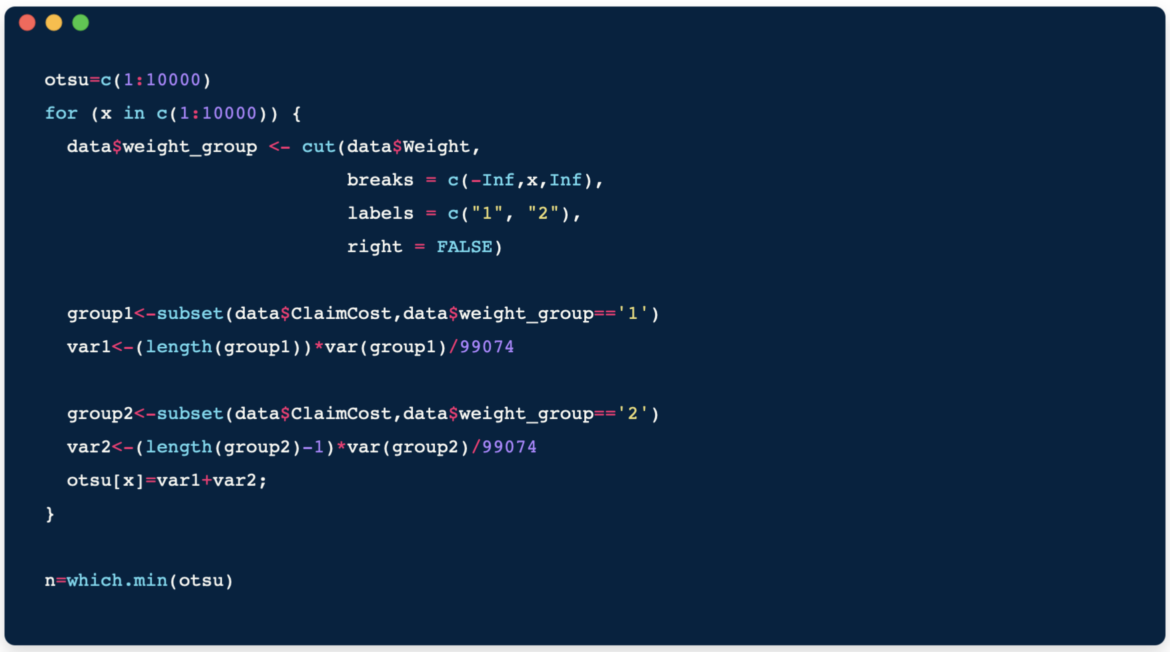
where weights ω_0 and ω_1 are the probabilities of the two classes separated by a threshold t ,and σ_0^2 and σ_1^2 are variances of these two classes.

Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance:

$$\sigma_b^2 = \sigma_T^2 - \sigma_w^2 = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \tag{3}$$

where σ_b^2 is the inter-class variance and σ_t^2 is the total variance in the whole group.

Following this method, we can divide our group into two subgroups to minimize the intra-class variance(or) maximizing inter-class variance.But the limit of the Otsu's method is that it can only divide the group into two subgroups. But we can apply this to the subgroup again to get 4,8,and more groups.



```
otsu=c(1:10000)
for (x in c(1:10000)) {
  data$weight_group <- cut(data$Weight,
                           breaks = c(-Inf,x,Inf),
                           labels = c("1", "2"),
                           right = FALSE)

  group1<-subset(data$ClaimCost,data$weight_group=='1')
  var1<-(length(group1))*var(group1)/99074

  group2<-subset(data$ClaimCost,data$weight_group=='2')
  var2<-(length(group2)-1)*var(group2)/99074
  otsu[x]=var1+var2;
}

n=which.min(otsu)
```

Figure 4: The Otsu's method

And we perform two more times like this, we can get four groups divided.We can also do the same to the *VehicleAge*.

- **4 VehicleAge Groups**: 0-4 years; 5-11 years; 12-18 years;$\geqslant$ 19 years.
- **4 Weight Groups**: 0-999 kg;1000-3680 kg;3681-6150kg;$\geqslant$ 6151 kg.
- **Climate and ActivityCode Groups** is divided in the natural way.

2.2 The test of division methods

After the division, we got 4 models for different division methods. here we are going to test how good does the model fits the data. We named the 4 model as model, model2, model3 and model4, and then using different criteria to test them.

2.2.1 Log Likelihood

Using `logLik()` function in R, we found the log likehood of 4 models in Table 1

From this perspective, we can say that the Otsu's method (method 4) does improve the goodness the model fits the data.

Methods	log Likelihood
Model	-608.2747
Model2	-578.8223
Model3	-601.0832
Model4	-462.6472

Table 1: Log likelihood of 4 models

2.2.2 Akaike information criterion

We can also use the Akaike Information Criterion (AIC) to test if our model should be simplified.

$$\mathrm{AIC} = 2k - 2\log(\hat{\mathcal{L}}) \tag{4}$$

where k is the number of parameters in the model and $\mathcal{L}$ is the maximum likelihood (ML) estimate of the GLM.

To calculate this, we use the `AIC()` function in R. The result is listed in Table 2.

Methods	AIC value
Model	1258.549
Model2	1199.645
Model3	1244.166
Model4	963.2945

Table 2: AIC values of 4 models

Because the smaller AIC is, the better the model is, we can say that model 4 perform better.

2.2.3 Bayesian information criterion

Because method 4 performs better in AIC test, we can continue with the Bayesian information criterion(BIC) test.

$$\mathrm{AIC} = \log(n) \cdot k - 2\log(\hat{\mathcal{L}}) \tag{5}$$

where n is the number of observations,k is the number of parameters in the model and $\mathcal{L}$ is the maximum likelihood (ML) estimate of the GLM.

Methods	AIC value
Model	1354.917
Model2	1297.137
Model3	1340.389
Model4	1042.91

Table 3: BIC values of 4 models

Considering the log likelihood, the AIC and BIC values, the Otsu's method is a good way of dividing. And our final model is based on that method.

2.3 The test of variable selection

After the test of dividing method, we should test the variable selection. In this section, we try to remove some variables to see which model is better.

2.3.1 Remove the ActivityCode

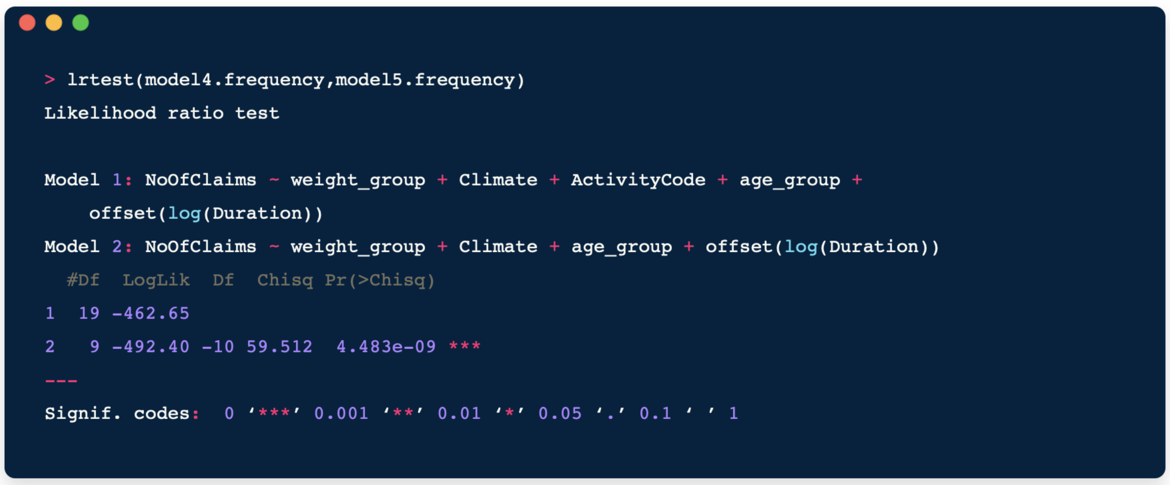
We try to remove the ActivityCode and get the Reduced Model(RM, the model5), and compare it to the full model(FM,the model4) using likelihood ratio test.

As the result[5,6] shows, because the χ^2 statistic is large than the given we can safely use the full model.

2.3.2 Remove the Climate

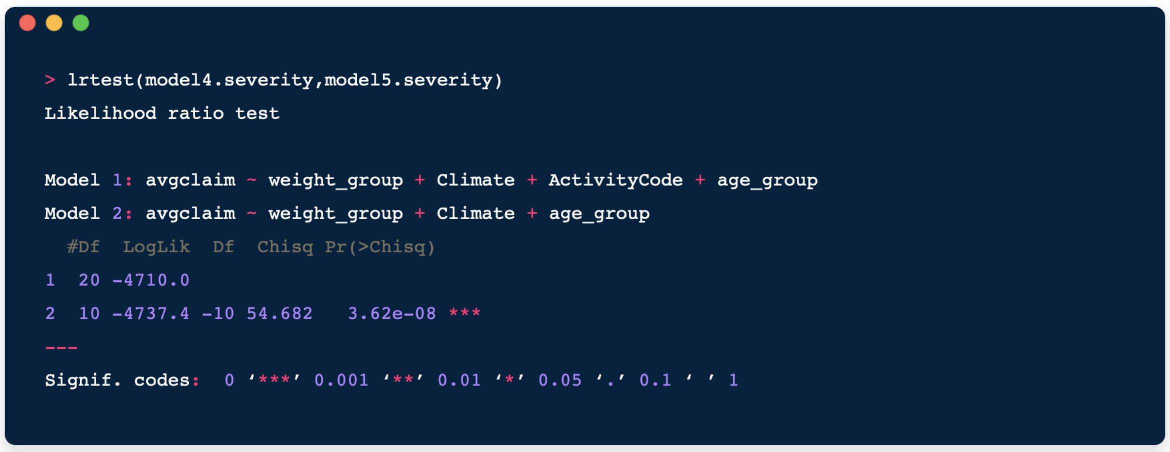
We try to remove the ActivityCode and get the Reduced Model(RM, the model6), and compare it to the full model(FM,the model4) using likelihood ratio test.

As the result[7] shows, the climate's effect on the number of claims is not significant but it has significant influence on the severity, so we have to keep it as a factor.



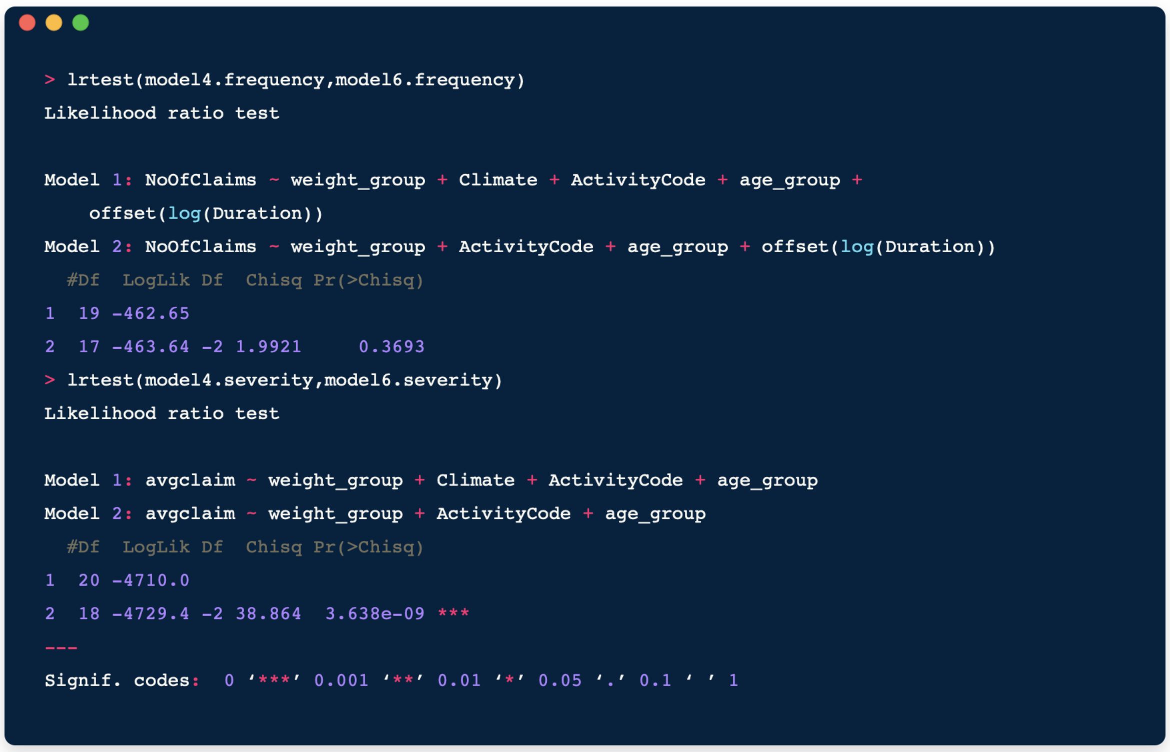
```
> lrtest(model4.frequency,model5.frequency)
Likelihood ratio test

Model 1: NoOfClaims ~ weight_group + Climate + ActivityCode + age_group +
    offset(log(Duration))
Model 2: NoOfClaims ~ weight_group + Climate + age_group + offset(log(Duration))
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  19 -462.65
2   9 -492.40 -10 59.512  4.483e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Frequency test



```
> lrtest(model4.severity,model5.severity)
Likelihood ratio test

Model 1: avgclaim ~ weight_group + Climate + ActivityCode + age_group
Model 2: avgclaim ~ weight_group + Climate + age_group
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  20 -4710.0
2  10 -4737.4 -10 54.682   3.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Severity test



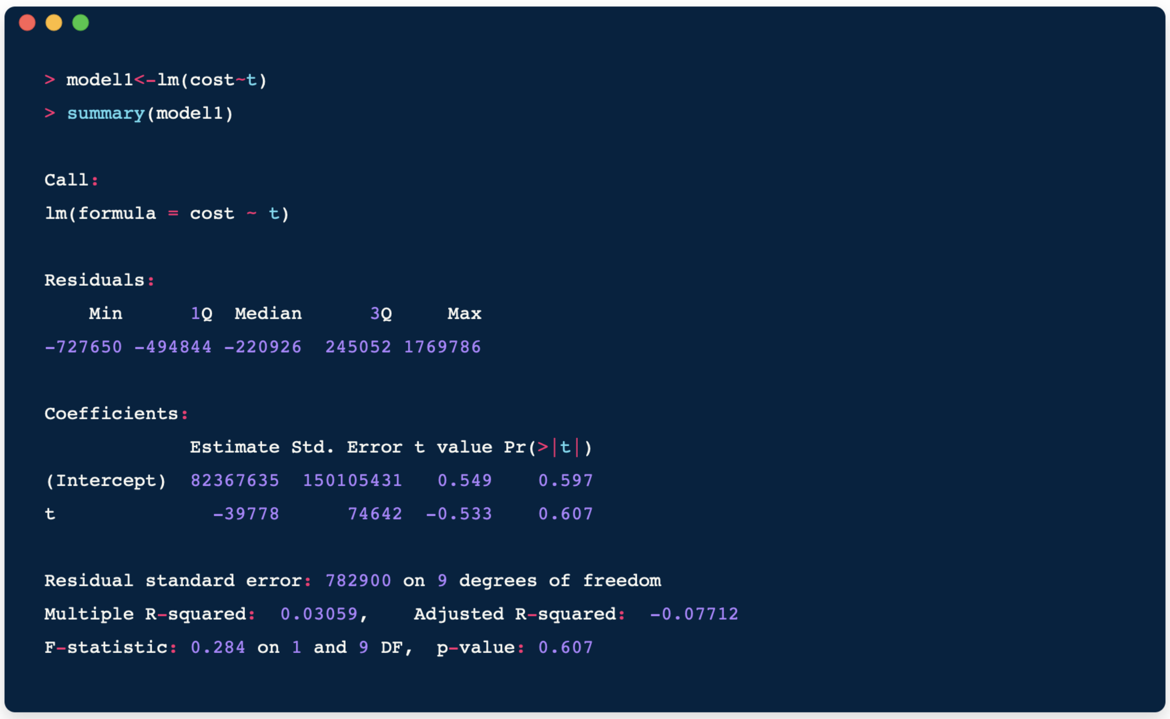
```
> lrtest(model4.frequency,model6.frequency)
Likelihood ratio test

Model 1: NoOfClaims ~ weight_group + Climate + ActivityCode + age_group +
    offset(log(Duration))
Model 2: NoOfClaims ~ weight_group + ActivityCode + age_group + offset(log(Duration))
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  19 -462.65
2  17 -463.64 -2 1.9921     0.3693
> lrtest(model4.severity,model6.severity)
Likelihood ratio test

Model 1: avgclaim ~ weight_group + Climate + ActivityCode + age_group
Model 2: avgclaim ~ weight_group + ActivityCode + age_group
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  20 -4710.0
2  18 -4729.4 -2 38.864  3.638e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: FM V.S. RM

2.4 Determine the basic level

2.4.1 Estimate the expected claim cost

First, we get claim costs divided by the duration to get a whole year's estimate cost.
Then we sum every year's estimate claim cost and use a simple linear regression between costs and time.

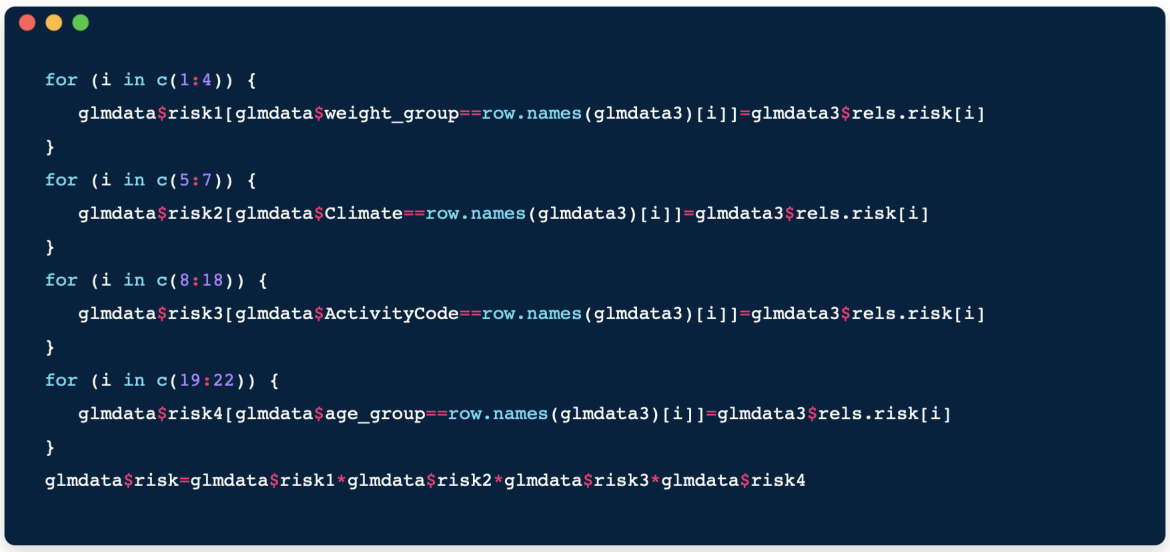


```
> model1<-lm(cost~t)
> summary(model1)

Call:
lm(formula = cost ~ t)

Residuals:
    Min      1Q  Median      3Q     Max 
-727650 -494844 -220926  245052 1769786 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  82367635  150105431   0.549    0.597
t              -39778      74642  -0.533    0.607

Residual standard error: 782900 on 9 degrees of freedom
Multiple R-squared:  0.03059,    Adjusted R-squared:  -0.07712 
F-statistic: 0.284 on 1 and 9 DF,  p-value: 0.607
```

Figure 8: The linear model between time and costs

It seems that the linear regresstion model is poor so we just use the mean of the past 11 years claim cost as 2017's estimate claim cost, and the total sum of tractors' premium should be 2375151/0.9 = 2639057

2.4.2 Find the total risk factor



```
for (i in c(1:4)) {
   glmdata$risk1[glmdata$weight_group==row.names(glmdata3)[i]]=glmdata3$rels.risk[i]
}
for (i in c(5:7)) {
   glmdata$risk2[glmdata$Climate==row.names(glmdata3)[i]]=glmdata3$rels.risk[i]
}
for (i in c(8:18)) {
   glmdata$risk3[glmdata$ActivityCode==row.names(glmdata3)[i]]=glmdata3$rels.risk[i]
}
for (i in c(19:22)) {
   glmdata$risk4[glmdata$age_group==row.names(glmdata3)[i]]=glmdata3$rels.risk[i]
}
glmdata$risk=glmdata$risk1*glmdata$risk2*glmdata$risk3*glmdata$risk4
```

Figure 9: Calculate the total risk factor

And we calculate every year's sum of total risk factors and etimate 2017's sum of total risk to be 37048.

3 Results

3.1 The factors

The risk factors we obtain are listed in Table 4

Groups	Risk Factors
01_ >1000kg	1
02_1000-3680years	3.226021
03_3681-6150years	10.89426
04_ >6151years	22.18261
Middle	0.674267
North	1.086456
South	1
A - Agriculture, Hunting and Forestry	1.025265
C - Mining and quarrying	1.06995
F - Construction	2.895707
G - Wholesale & retail trade; repair of motor vehicles, household	2.61727
H - Hotels and restaurants	1.382193
I - Transport, storage and communication	0.60966
L - Public administration and defence; compulsory social security	1.223695
M - Education	0.60734
Missing	1
N - Health and social work	2.550046
Other	1.182008
01_<5years	1.533029
02_5-11years	1
03_12-18years	0.659524
04_>=19years	0.328926

Table 4: Risk Factors

3.2 The basic level

The basic level is

$$\gamma_0 = 2639057 \div 37048 = 71.233453898 \tag{6}$$

4 Conclusion

In this project, we try to develop a model for risk management and pricing. The core of the pricing model is to use GLM for estimating the risk factors.

We develop 4 methods of dividing, and then use log likelihood, AIC and BIC to choose the best model. The Otsu's method we adopt comes from the image processing field, and it is interesting that it works well in this project,too.

Then, to verify our variable selection, we use likelihood ratio test for the full model and reduced model, which proves that the full model is favorable.

Finally we do some work of calculation to find the basic level which completes the model.