PROJECT 2: REPORT

March 11, 2019

ASMI Souhail
FOUFA Mastafa
SF2930 Regression analysis VT2019

Contents

Chapter 1

Variable selection

Prior to statistical analysis of the data, we decided to do some analysis of the data on python. We noticed the large number of missing values for the feature "Activity Code". This is pictured in figure 1.1.

The goal is to know whether the feature Activity Code presenting a lot of null values should be removed or not. We proceed with different tests: likelihood ratio test, AIC test and BIC test. Also we will focus on the severity model to test significance of such feature.

- H0: parameter coefficient for this feature is null
- H1: parameter coefficient for this feature is not null

From the tests summarized in the figures 1.4 - 1.6, we can conclude that it is better to keep the variable "activity code". Indeed, we have AIC(FM) $<$ AIC(RM). According to the lecture, if AIC(Large model) $>$ AIC(Small model), then we might consider excluding the variable removed in the small model.
BIC is supposed to punish additional variables even more. So if we have BIC(Large model) $<$ BIC(Small model), we can feel safe about adding the additional variable. Here, we have BIC(Small Model) $<$ BIC(Large Model). So a priori this means that the small model is better. However, the difference is very slight.
Also, we decided to test significance of the feature "Activity Code" by focusing on the severity model. We fit the reduced model (without such feature) and the full model and do a comparison. Such comparison is summarized in figure 1.2.

In conclusion, we decide to include the feature "Activity Code" in the rest of the analysis.

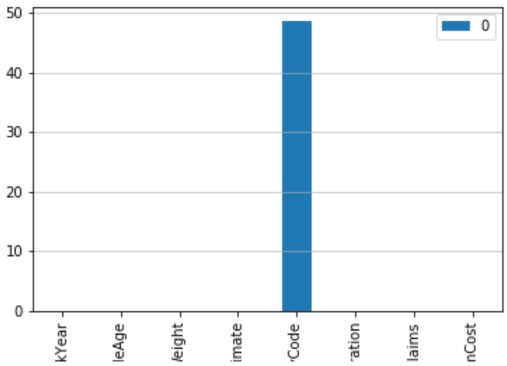


Figure 1.1: Percentage of missing values in each column

```
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  25 -4280.5
2  15 -4303.7 -10 46.491  1.169e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.2: Likelihood ratio test on severity model

```
> AIC.test_severity
                       df      AIC
model.severity         25 8610.990
model.severity_reduced 15 8637.481
```

Figure 1.3: AIC test with severity model

```
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  24 -651.96
2  14 -680.91 -10 57.905  9.007e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.4: Likelihood ratio test

```
> AIC_test
                        df      AIC
model.frequency         24 1351.913
model.frequency_reduced 14 1389.818
```

Figure 1.5: AIC test

```
                        df      BIC
model.frequency         24 1471.278
model.frequency_reduced 14 1459.448
```

Figure 1.6: BIC test

Chapter 2

Grouping and risk differentiation

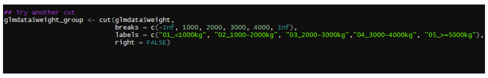
Perform a GLM analysis to figure out how best to describe the risk for the tractors.Use the template GLM.R. The outcome should be a multiplicative GLM model, as described in Eq. 1, that model claims frequency and claim severity separately. Use the same variables and variable groups in both models, and propose the final risk factorÎşk,i, where the final risk factor is the product of the claim frequency and the claim severity.In order to perform your GLM analysis, you will have to group some of the variables. Consider, for example, the tractorsâĂŹ weights. These cover a very wide range, astractors can be both very small and light, and extremely big and heavy. Thus, it wouldbe impossible to analyze each individual weight alone; it is necessary to group them.When grouping a variable, there are two things to consider:

- Make each group "Risk homogeneous", meaning that you believe that the risk does not vary much within the group, with regard to the particular variable.2
- Create groups with enough data to get a stable GLM analysis for each group. What is "enough" has no clear answer, bur varies, depending among other things on how many variables you use in your analysis.

Weight test – Grouping: We tested several grouping schemes different from the one provided originally. Only two of them are summarized in figures 2.1 - 2.4. We end up choosing the one with the lowest AIC.

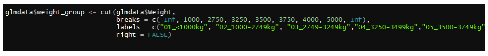
Based on AIC test, we finally decide to choose a simple grouping with two groups for the age of the vehicle:

- Group 1: vehicle with age < 20 years
- Group 2: vehicle with age > 20 years



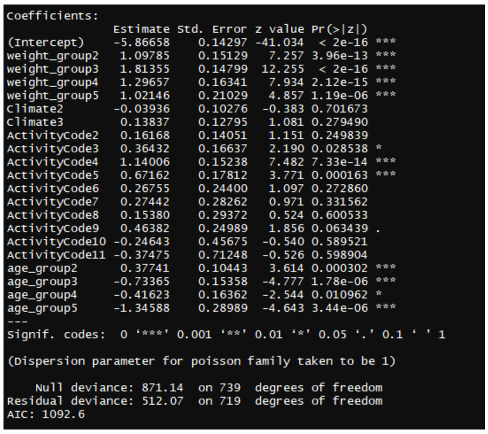
```
## Try another cut
glmdata$weight_group <- cut(glmdata$Weight,
                            breaks = c(-Inf, 1000, 2000, 3000, 4000, Inf),
                            labels = c("01_<1000kg", "02_1000-2000kg", "03_2000-3000kg","04_3000-4000kg", "05_>=5000kg"),
                            right = FALSE)
```

Figure 2.1: Test of cut (selected one) - example.1



```
glmdata$weight_group <- cut(glmdata$Weight,
                            breaks = c(-Inf, 1000, 2750, 3250, 3500, 3750, 4000, 5000, Inf),
                            labels = c("01_<1000kg", "02_1000-2749kg", "03_2749-3249kg","04_3250-3499kg","05_3500-3749kg"
                            right = FALSE)
```

Figure 2.3: Test of cut - example.2

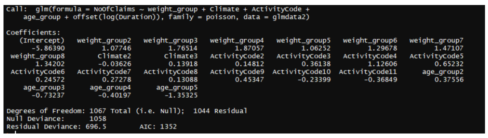


```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.86658    0.14297 -41.034  < 2e-16 ***
weight_group2   1.09785    0.15129   7.257 3.96e-13 ***
weight_group3   1.81355    0.14799  12.255  < 2e-16 ***
weight_group4   1.29657    0.16341   7.934 2.12e-15 ***
weight_group5   1.02146    0.21029   4.857 1.19e-06 ***
Climate2       -0.03936    0.10276  -0.383 0.701673
Climate3        0.13837    0.12795   1.081 0.279490
ActivityCode2   0.16168    0.14051   1.151 0.249839
ActivityCode3   0.36432    0.16637   2.190 0.028538 *
ActivityCode4   1.14006    0.15238   7.482 7.33e-14 ***
ActivityCode5   0.67162    0.17812   3.771 0.000163 ***
ActivityCode6   0.26755    0.24400   1.097 0.272860
ActivityCode7   0.27442    0.28262   0.971 0.331562
ActivityCode8   0.15380    0.29372   0.524 0.600533
ActivityCode9   0.46382    0.24989   1.856 0.063439 .
ActivityCode10 -0.24643    0.45675  -0.540 0.589521
ActivityCode11 -0.37475    0.71248  -0.526 0.598904
age_group2      0.37741    0.10443   3.614 0.000302 ***
age_group3     -0.73365    0.15358  -4.777 1.78e-06 ***
age_group4     -0.41623    0.16362  -2.544 0.010962 *
age_group5     -1.34588    0.28989  -4.643 3.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 871.14  on 739  degrees of freedom
Residual deviance: 512.07  on 719  degrees of freedom
AIC: 1092.6
```

Figure 2.2: Test summary - example.1



```
Call:  glm(formula = NoOfClaims ~ weight_group + Climate + ActivityCode +
    age_group + offset(log(Duration)), family = poisson, data = glmdata2)

Coefficients:
   (Intercept)   weight_group2   weight_group3   weight_group4   weight_group5   weight_group6   weight_group7
      -5.86390         1.07746         1.76514         1.87057         1.06252         1.29678         1.47107
 weight_group8        Climate2        Climate3   ActivityCode2   ActivityCode3   ActivityCode4   ActivityCode5
       1.34202        -0.03626         0.13918         0.14812         0.36138         1.12606         0.65232
 ActivityCode6   ActivityCode7   ActivityCode8   ActivityCode9  ActivityCode10  ActivityCode11      age_group2
       0.24572         0.27278         0.13088         0.45347        -0.23399        -0.36849         0.37556
    age_group3      age_group4      age_group5
      -0.73237        -0.40197        -1.35325

Degrees of Freedom: 1067 Total (i.e. Null);  1044 Residual
Null Deviance:      1058
Residual Deviance: 696.5        AIC: 1352
```

Figure 2.4: Test summary - example.2

```
glmdata$age_group <- cut(glmdata$VehicleAge,
                         breaks = c(-Inf, 4, 13, 18, 30, Inf),
                         labels = c("01_<4years", "02_4-13years", "03_13-18years", "04_18-30years", "05_>30years"),
                         right = FALSE)
```

Figure 2.5: Test of cut - example.1

```
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -5.86390    0.14295 -41.020  < 2e-16 ***
weight_group2    1.07746    0.14574   7.393 1.43e-13 ***
weight_group3    1.76514    0.16782  10.518  < 2e-16 ***
weight_group4    1.87057    0.16460  11.365  < 2e-16 ***
weight_group5    1.06252    0.23333   4.554 5.27e-06 ***
weight_group6    1.29678    0.21753   5.961 2.50e-09 ***
weight_group7    1.47107    0.23830   6.173 6.69e-10 ***
weight_group8    1.34202    0.24566   5.463 4.68e-08 ***
Climate2        -0.03626    0.10284  -0.353  0.72444    
Climate3         0.13918    0.12796   1.088  0.27671    
ActivityCode2    0.14812    0.14079   1.052  0.29275    
ActivityCode3    0.36138    0.16637   2.172  0.02984 *  
ActivityCode4    1.12606    0.15275   7.372 1.68e-13 ***
ActivityCode5    0.65232    0.17863   3.652  0.00026 ***
ActivityCode6    0.24572    0.24501   1.003  0.31591    
ActivityCode7    0.27278    0.28267   0.965  0.33454    
ActivityCode8    0.13088    0.29434   0.445  0.65658    
ActivityCode9    0.45347    0.25004   1.814  0.06973 .  
ActivityCode10 -0.23399    0.45704  -0.512  0.60868    
ActivityCode11 -0.36849    0.71245  -0.517  0.60501    
age_group2       0.37556    0.10459   3.591  0.00033 ***
age_group3      -0.73237    0.15364  -4.767 1.87e-06 ***
age_group4      -0.40197    0.16400  -2.451  0.01424 *  
age_group5      -1.35325    0.28912  -4.681 2.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1057.80  on 1067  degrees of freedom
Residual deviance:  696.55  on 1044  degrees of freedom
AIC: 1351.9

Number of Fisher Scoring iterations: 6
```

Figure 2.6: Test summary - example.1

```
## Try another cut on age
glmdata$age_group <- cut(glmdata$VehicleAge,
                         breaks = c(-Inf, 20, Inf),
                         labels = c("01", "2"),
                         right = FALSE)
```

Figure 2.7: Test of cut (selected one)- example.2

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -5.7585     0.1336 -43.117  < 2e-16 ***
weight_group2    1.0110     0.1455   6.947 3.72e-12 ***
weight_group3    1.6808     0.1669  10.073  < 2e-16 ***
weight_group4    1.8254     0.1642  11.119  < 2e-16 ***
weight_group5    0.9309     0.2317   4.018 5.87e-05 ***
weight_group6    1.2752     0.2171   5.873 4.27e-09 ***
weight_group7    1.4182     0.2378   5.964 2.46e-09 ***
weight_group8    1.3052     0.2448   5.332 9.71e-08 ***
Climate2        -0.0549     0.1026  -0.535 0.592495    
Climate3         0.1056     0.1277   0.827 0.408304    
ActivityCode2    0.1419     0.1401   1.013 0.311101    
ActivityCode3    0.3662     0.1663   2.202 0.027661 *  
ActivityCode4    1.1734     0.1524   7.699 1.37e-14 ***
ActivityCode5    0.6693     0.1783   3.755 0.000174 ***
ActivityCode6    0.2920     0.2442   1.196 0.231702    
ActivityCode7    0.2905     0.2822   1.029 0.303252    
ActivityCode8    0.1635     0.2938   0.556 0.577872    
ActivityCode9    0.4527     0.2499   1.812 0.070003 .  
ActivityCode10  -0.1978     0.4569  -0.433 0.665080    
ActivityCode11  -0.3318     0.7123  -0.466 0.641336    
age_group1      -1.0428     0.1506  -6.924 4.38e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 715.01  on 484  degrees of freedom
Residual deviance: 384.15  on 464  degrees of freedom
AIC: 859.31

Number of Fisher Scoring iterations: 6
```

Figure 2.8: Test summary - example.2

Chapter 3

Leveling

After we have found the risk factors $\gamma_{k,i}$ we proceeded to the determination of the base level corresponding to 2017, and we did it in the three following steps :

3.0.1 Estimation of the expected claim cost of 2017

Before beginning with the estimation we can see that our data doesn't contain any insurance contract in the year 2017, all the contacts available go from 2006 to 2016 . We begin by creating cells , each cell corresponds to one year. In each cell we computed the average claim cost and we got the graph in figure 3.1.

We can see from figure 3.1 that there is not a flashing linear relation between the two variables. So we used the Null model (using only the intercept as predictor), the estimated total cost of 2017 is the average of all means values of each year. For this approach we used the table summarized in figure 3.2 and computed the average cost for the year 2017 as merely the average of the costs in the past years.

Also, we can notice in figure 3.3 that there is no clear linear relationships between the years and the number of claims. So, the simple model with only and intercept can be an acceptable solution for such problem.

3.0.2 Computing the total premium and base level γ_0

For our simple model only using a model with an intercept we found the following base level:

$$\gamma_0 = 639.3 \tag{3.1}$$

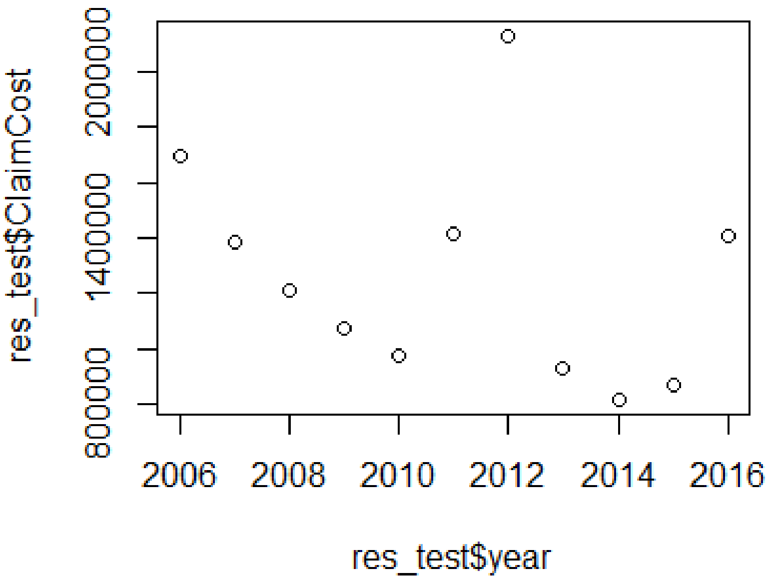


Figure 3.1: summary after fitting all the regressors

Now, we can use a more complex model with GLM. Assume the response variables describing now the number of claims follow a Poisson distribution. We use the model defined in R in figure 3.4.

Out of the model described in figure 3.4 we can predict the number of claims for the year 2017 as:

$$NbClaims2017 = exp(2017 * (-0.0098) + 23.58) \Rightarrow NbClaims2017 = 41.8 \tag{3.2}$$

Now for the expected claim cost, we also use a GLM assuming that the claim cost follow a Gamma distribution from the exponential family. The results are summarized in table plot in figure 3.5. We can first notice that the p value for the t-test on the single regressor is quite high. This means that the regressor is not really significant. Also, we have the following result for the expected claim cost in the test observation for the year 2017:

$$ECC2017 = exp(2017 * (-0.0267) + 67.812) \Rightarrow ECC2017 = 1108083 \tag{3.3}$$

year	NoOfClaims	ClaimCost	AVG_cost
2006	56	1697212.9	30307.37
2007	49	1383131.8	28227.18
2008	41	1211450.5	29547.57
2009	46	1071715.5	23298.16
2010	31	972082.8	31357.51
2011	37	1418261.9	38331.40
2012	56	2127615.5	37993.13
2013	36	929759.6	25826.65
2014	44	817945.1	18589.66
2015	46	866741.5	18842.21
2016	46	1407425.2	30596.20

Figure 3.2: Number of claims and average cost for each year

Let us now define the total cost as:

$$totalCost2017 = ECC2017 * NOC2017 = 46317869 \tag{3.4}$$

And

$$priceEstimate = \frac{totalCost2017}{0.9} = 51464299 \tag{3.5}$$

Hence,

$$\gamma_0 = \frac{priceEstimate}{\prod_{i=1}^{n} \gamma_i} = 23461.11 \tag{3.6}$$

We can summarize the results by comparing the simple model with the more complex one with GLM:

$$\gamma_0^{Simple} = 639.3$$
$$\gamma_0^{GLM} = 23461.11$$

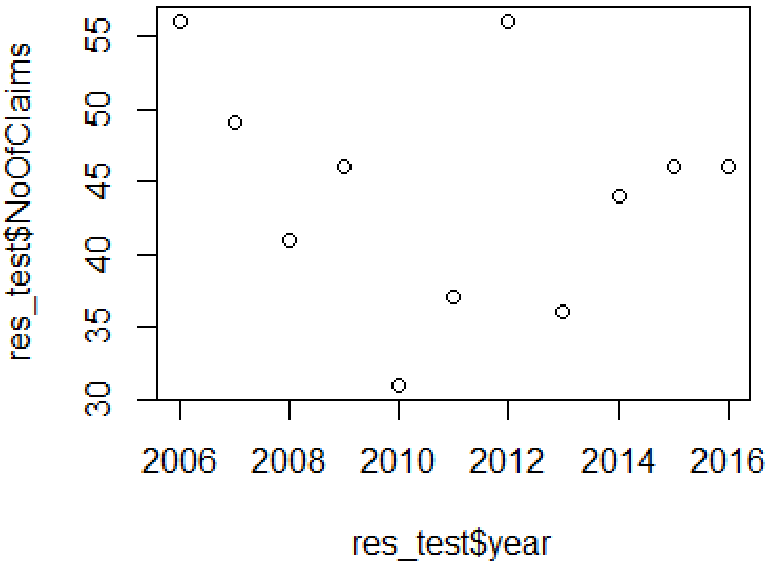
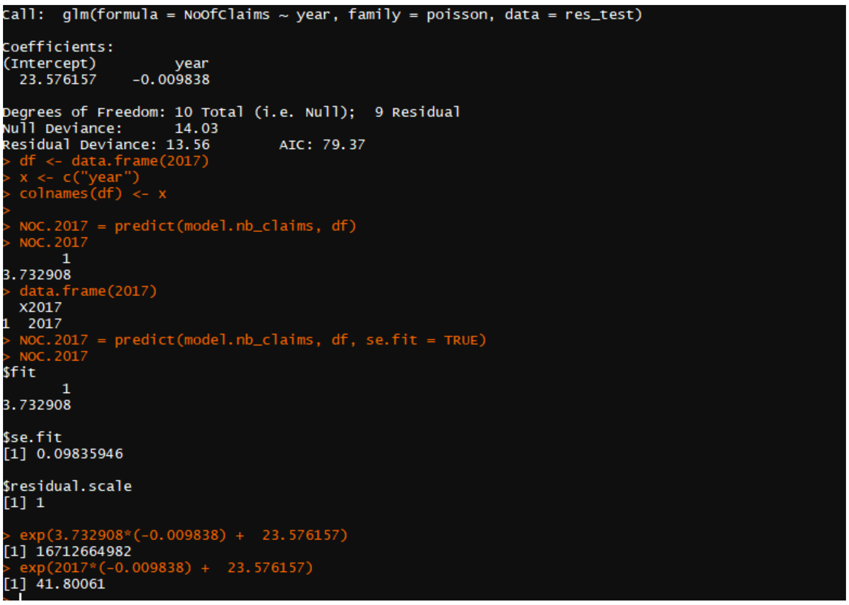


Figure 3.3: Number of claims vs year



```
Call:  glm(formula = NoofClaims ~ year, family = poisson, data = res_test)

Coefficients:
(Intercept)         year
  23.576157    -0.009838

Degrees of Freedom: 10 Total (i.e. Null);  9 Residual
Null Deviance:      14.03
Residual Deviance: 13.56        AIC: 79.37
> df <- data.frame(2017)
> x <- c("year")
> colnames(df) <- x
>
> NOC.2017 = predict(model.nb_claims, df)
> NOC.2017
       1
3.732908
> data.frame(2017)
  X2017
1  2017
> NOC.2017 = predict(model.nb_claims, df, se.fit = TRUE)
> NOC.2017
$fit
       1
3.732908

$se.fit
[1] 0.09835946

$residual.scale
[1] 1

> exp(3.732908*(-0.009838) +  23.576157)
[1] 16712664982
> exp(2017*(-0.009838) +  23.576157)
[1] 41.80061
```

Figure 3.4: Number of claims and average cost for each year

```
> summary(model.claim_cost)

Call:
glm(formula = ClaimCost ~ year, family = Gamma("log"), data = res_test,
    weights = NoOfClaims)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3807  -1.6391  -0.9326   0.7878   4.2670

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.81382   62.22689   1.090    0.304
year        -0.02672    0.03094  -0.864    0.410

(Dispersion parameter for Gamma family taken to be 5.022755)

    Null deviance: 43.848  on 10  degrees of freedom
Residual deviance: 40.081  on  9  degrees of freedom
AIC: 13879

Number of Fisher Scoring iterations: 4
```

Figure 3.5: Number of claims and average cost for each year