STK 802 Assigment3
Mixture Rergession Model

Henry Wandera 17253129

September 2018

1 Introduction

In this assignment, I critically evaluate the properties and performance of a mixture regression model on a standardized claims data set containing information relating to the claims history of branches within an insurance company. The records are aggregated to the branch level and it has the following variables:

1. Y = Claims - Total claims cost
2. X1 = Beneficiaries - Number of beneficiaries in the branch
3. X2 = Age - Average age of members
4. X3 = Pens_Ratio - Ratio of pensioners to the total number of members
5. X4 = Female_Ratio - Ratio of female beneficiaries to the total number of members
6. X5 = Depen_Ratio - Average dependency ratio per policy

2 Exploratory Data Analysis (EDA)

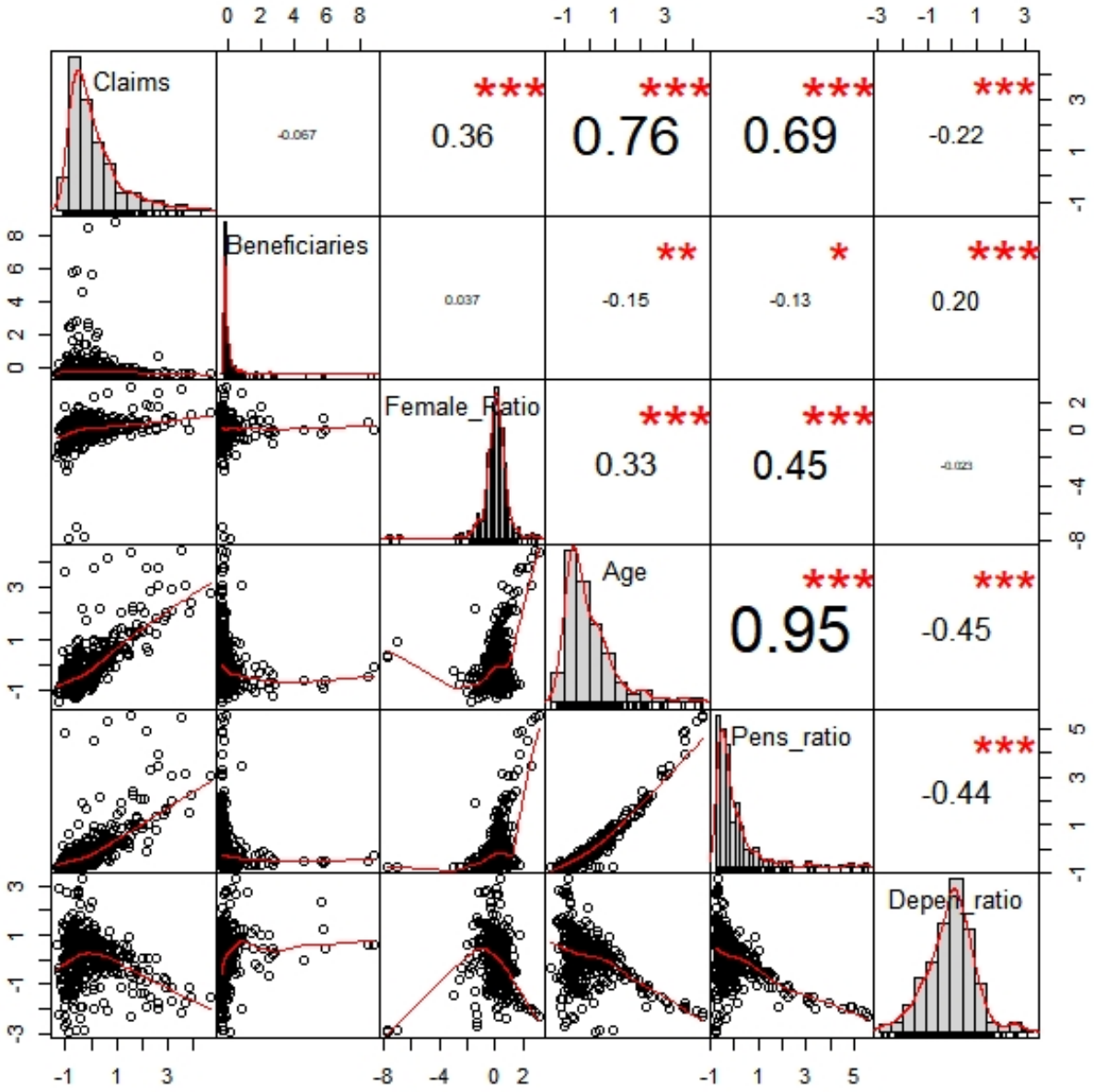
2.1 Correlations Matrix

	Claims	Beneficiaries	Female_Ratio	Age	Pens_ratio	Depen_ratio
Claims	1					
Beneficiaries	-0.067	1				
Female_Ratio	0.36	0.037	1			
Age	0.758	-0.152	0.331	1		
Pens_ratio	0.692	-0.13	0.452	0.954	1	
Depen_ratio	-0.223	0.203	-0.023	-0.448	-0.438	1

Observation

- There is a high correlation between Claims and Age (0.758), and Claims and Pens_Ratio (0.692). However, there is a very high Pearson's correlation of 0.954 between Age and Pens_Ratio which suggests that these two variables are somewhat bounded together. For instance, young people rarely request for pension as compared to old people.

2.2 Correlation Chart



Observations

- The histograms in the diagonal represent the distributions of every variable. Most of the variables are skewed to left or right expect the Depen_ratio which does not violate normality.
- Scatter plots under the diagonal indicating the relationships between 2 variables intersecting at that cell.There are outliers in all the independent variables in relation with the dependent variable. Mixture regression models are sensitive to outliers and assumes normality in the distribution [1]. The data set had 88 observations as outliers which I didn't remove because it would reduce the number of observations for training the model.
- Variable Beneficiaries is indicated insignificant in predicting Claims because it has no codes (red stars). The level of significance is shown by the number of stars in the upper side of the diagonal. P-value and Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. The smaller the P-value the better independent variable is in predicting the dependent variable.

- Pearson's correlation coefficients between variables are indicated above the diagonal section. The font size of the coefficients is proportional to the size of the coefficient between variables.

3 Regression Model

I trained the model using `flexmix package` in r. Below are the estimated parameters of the model.

'log Lik.' -247.9067 (df=23)
AIC: 541.8134
BIC: 632.1327

Due to the outliers, I consider the performance of the model moderate. I expect the AIC, BIC values to reduce if outliers are removed for the model to be considered good.

	comp.1	Pr(> |z|)	comp.2	Pr(> |z|)	comp.3	Pr(> |z|)
(Intercept)	0.378094	4.755e-08 ***	0.039000	0.5271622	-0.523255	2.2e-16 ***
Beneficiaries	0.041812	0.347394	0.011121	0.8177345	0.034009	0.467534
Female_Ratio	0.163901	0.005082 **	0.136653	0.0019711 **	0.237938	8.085e-05 ***
Age	1.295036	9.723e-08 ***	0.940110	1.207e-08 ***	0.370325	0.008925 **
Pens_ratio	-0.058321	0.837482	-0.145088	0.3635163	-0.197149	0.144297
Depen_ratio	0.047065	0.510508	0.174217	0.0005032 ***	0.044668	0.328210
sigma	0.4105568	-	0.3153245	-	0.3096758	-
prior	0.3066276	-	0.4030854	-	0.2902870	-
Cluster sizes	65	-	211	-	99	-

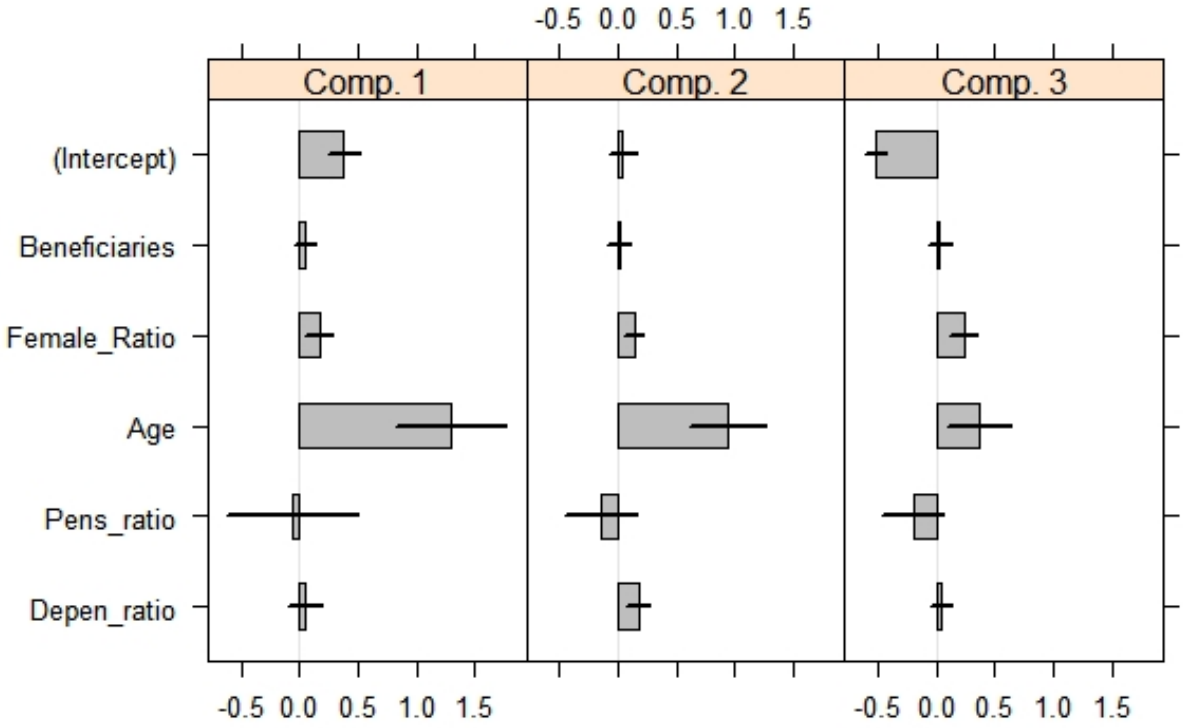
3.1 Interpretation

- Component 2 was the biggest with 211 observations, followed by component 3 with 99 and lastly component 1 with 65 observations. `Prior` also shows the size of the clusters or components on a scale of 0 - 1.
- For component 1, Age was significant in determining the claims followed by Female Ratio. Other variables were not indicated significant.
- For component 2, Age and Depen_ratio were indicated most relevant, followed by Female Ratio. Other variables were not significant.
- For component 3, Female Ration was most significant, followed by age and other variables were not significant.
- The estimated total claim cost for component 1, 2, 3 when all independent variables are zero is 0.378094, 0.039000 and -0.52325 respectively. These are the Intercept values.
- An increase of 1 year in the average age while adjusting or controlling other variables is associated with an increase of 1.295036 of claim cost for component 1, 0.940110 for component 2, and 0.370325 for component 3.
- Component 1 had the oldest group with the median age of 0.4027, followed by component 3 with -0.248770 and lastly component 2 with -0.4881. These were standardized values.
- An increase in the number of beneficiaries by 1 in every group, would increase the total claim cost by 0.041812 for component 1, 0.034009 for component 3, and 0.011121 for component 2. Therefore, members in component 1 claim for high amounts than any other group.

3.2 Visualization of the values in the table above

The rootogram below shows the size of the coefficients as given in the table. For instance:

- Age had the highest coefficient or slope. As mentioned earlier, Component 1 still leads.
- Female ratio is more significant for component 3.
- Component 2 had the leading Depen_ratio coefficient.



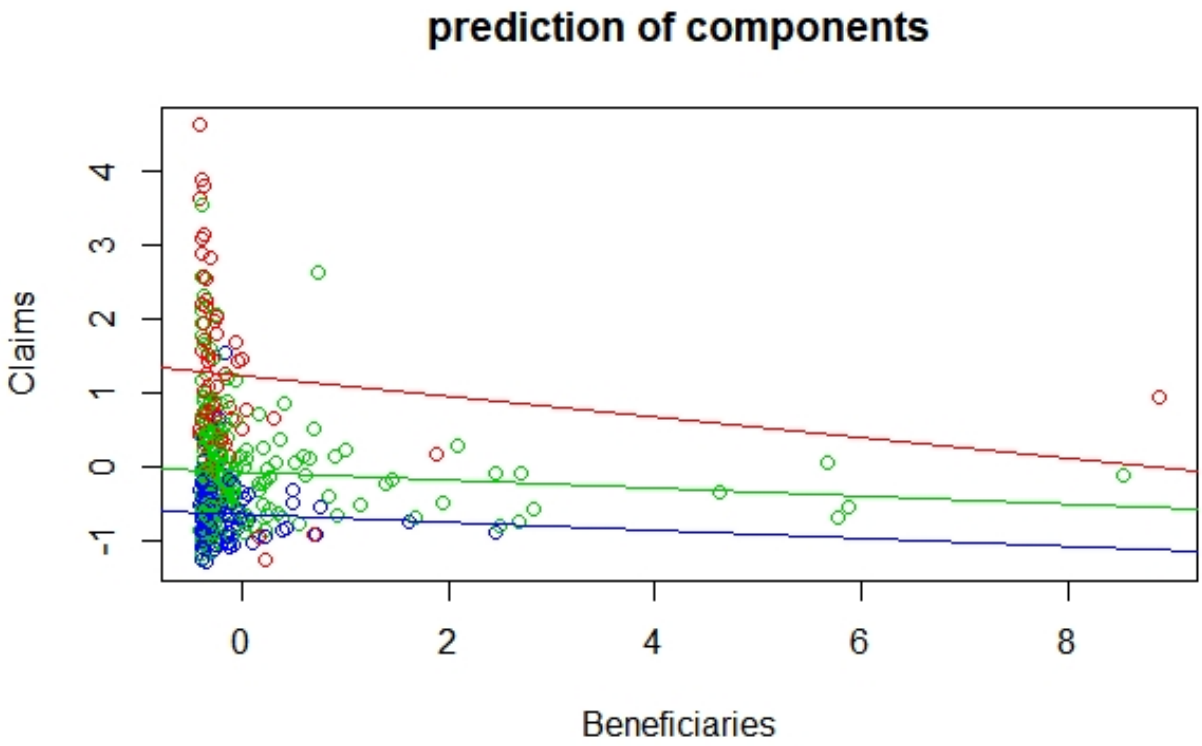
4 Graphs of the mixture regression model on the observed data

Note:

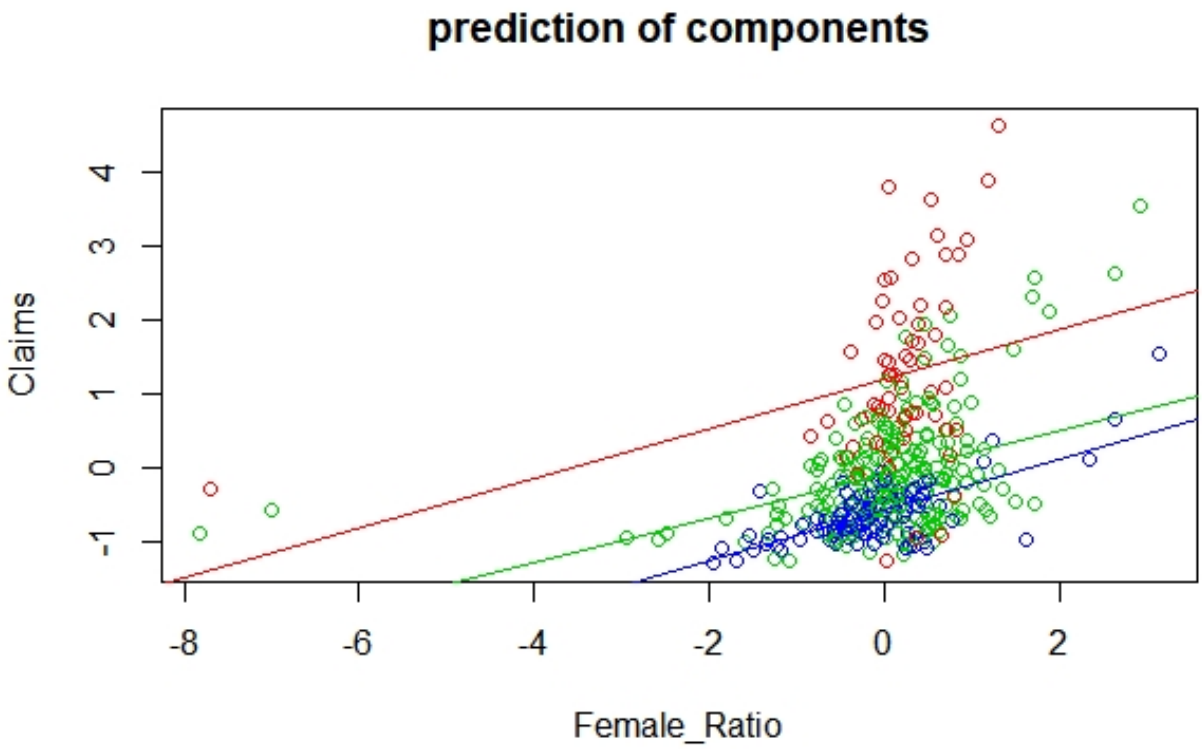
- Red represents component 1
- Green represents component 2
- Blue represents component 3
- The regression lines correspond with the colors of the data points.
- Data points were colored depending on the clusters of the mixture model.

- The graphs are different representing each independent variable against the total claim costs. However, the model was trained in form of multiple regression where all variables were combined and trained against the cost, like: "`model <- flexmix(Claims` $\sim (Beneficiaries + Female\_Ratio + Age + Pens\_ratio + Depen\_ratio), k = 3$"
  In all graphs, component 1 had the highest claims (at the top), followed by component 2 (in the middle) and lastly component 3 for every independent variable plotted against the dependent variable.

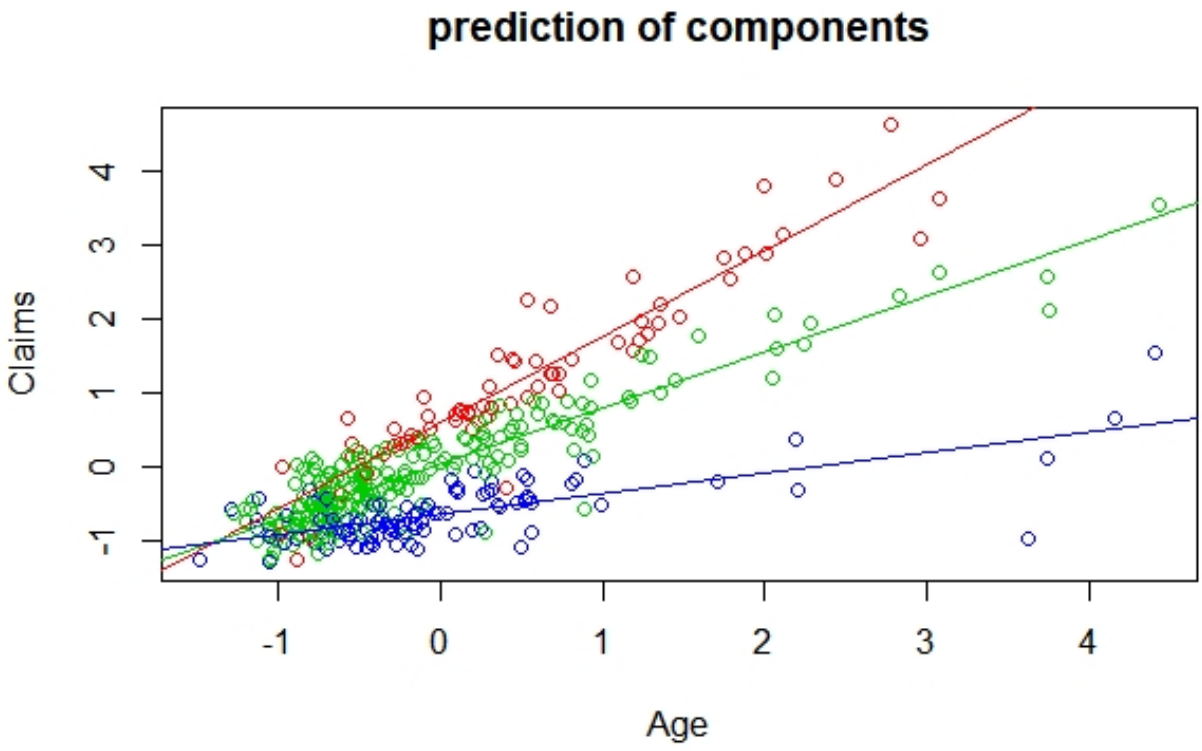
4.1 Number of beneficiaries Vs Claims



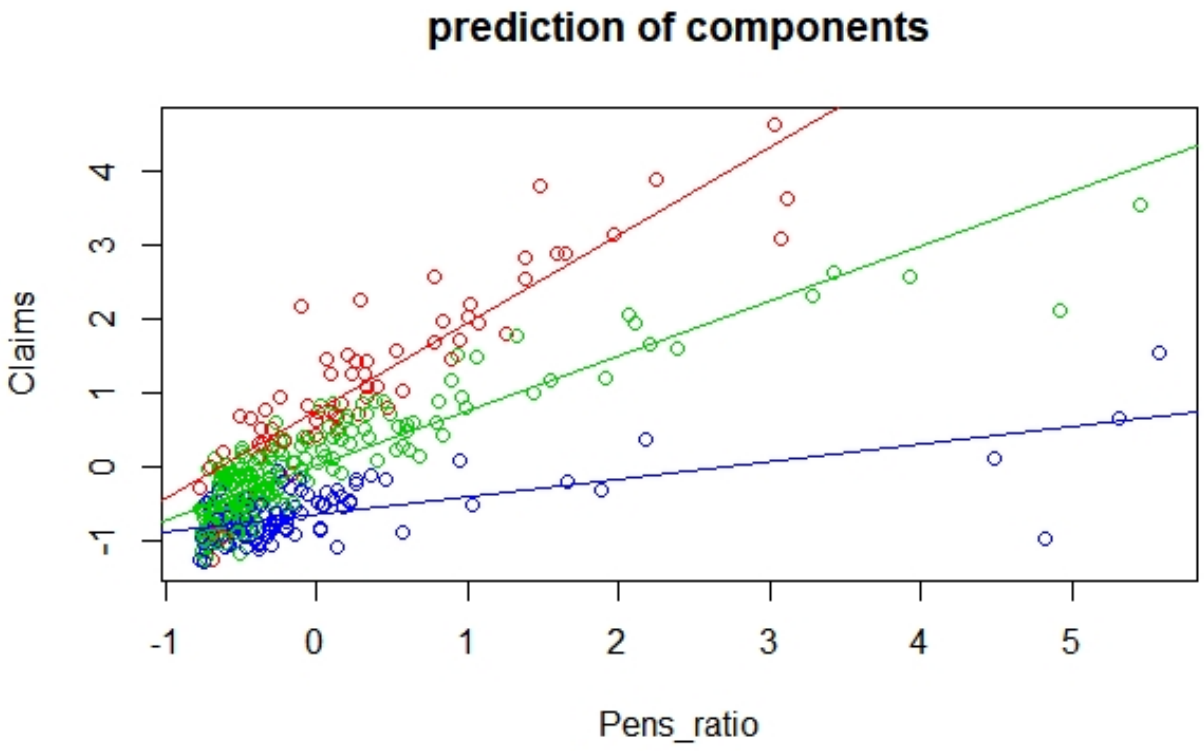
4.2 Ratio of female beneficiaries to the total number of members Vs Claims



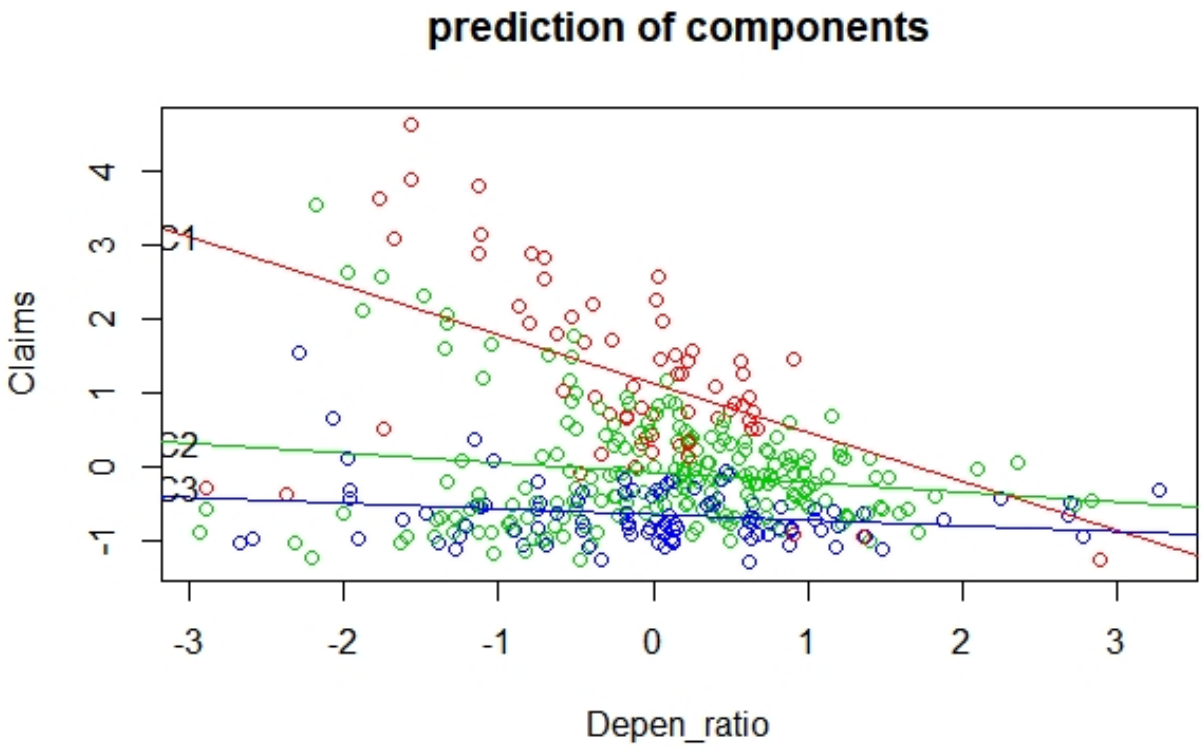
4.3 Average age of members Vs Claims



4.4 Ratio of pensioners to the total number of members Vs Claims



4.5 Average dependency ratio per policy Vs Claims



5 Conclusion

Some data points were overlapping. This is effect can be reduced if the data set is cleaned for example by removing outliers. Component 2 had the biggest number of observations but the claims in component 1 were seen to be the highest.

References

[1] C. B. Zeller, C. R. Cabral, and V. H. Lachos, "Robust mixture regression modeling based on scale mixtures of skew-normal distributions," *Test*, vol. 25, no. 2, pp. 375–396, 2016.