Poisson Regression to Predict Frequency of Claims

Andrew vanderWilden

October 11, 2019

Abstract

This report uses data on automobile insurance claims for regions near Waltham, Massachusetts. In the report a Poisson regression model is used to predict the frequency of claims. We found Gender, Age, Vehicle Type, Vehicle Use, and Region to be significant predictors of frequency of claims. Additionally we found Age and Gender to have an interactive affect.

Introduction

Orientation Material

Insurance companies seek to offer policies to customers that are a fair reflection of the riskiness of their expected driving behavior. A number of variables affect the expected risk that any particular policyholder has.

Key Aspects

This report attempts to fit a model using a Poisson regression model to accurately predict the frequency of claims that a policyholder is expected to have in a given year.

Plan for the Rest of the Report

The outline for the remainder of the report is as follows. In section 3, we present the most important characteristics of the data. In section 4, the model selection process and the following interpretation will be discussed. Concluding remarks can be found in section 5 with details to follow in the Appendix.

3. Data Characteristics

The data are cross-sectional and describe automobile insurance claims for 0 regions near Waltham, Massachusetts over an unspecified length of time. The data set includes 10000 observations with information on 11 variables. The variables are:

Item	Variable	Definition
1	Claims	The number of claims a policy holder has had
2	Claim	Indicator if a claim occurred or not
3	Exposure	The fraction of the year the policy holder was exposed to risk

Item	Variable	Definition
4	Age	The age of the policy holder
5	Gender	Gender of the policy holder
6	Marital Status	Civil status of policy holder
7	Education	Number of years of education
8	Region	Geographical region where the vehicle is garaged
9	Vehicle Age	Age of vehicle in years. Zero is brand new
10	Vehicle Body	Type of vehicle
11	Vehicle Use	Principal type of use

The variable `Claims` describes the number of claims a policyholder has had. The variable `Claim` is an indicator variable. A value of 1 indicates there has been a claim while a value of 0 indicates no claim has occurred.

The variable of interest is the frequency of accidents. To compute frequency we take the sum of all `Claims` and divide by the sum of the entire `Exposure` to account for the length of time an individual drives under a policy. In the entire data set, the frequency of claims is 11.29%.

The following table provides summary statistics on all of the numerical variables in the data set:

```
              Mean Median Standard Deviation     Minimum Maximum
Claims      0.1014      0          0.3678744  0.00000000       5
Exposure    0.8981      1          0.2269346  0.08333333       1
Age        36.2299     33         13.9833687 17.00000000      75
Education  16.9766     16          3.6006595 12.00000000      22
Vehicle Age 7.3477      6          6.0823601  0.00000000      25
```

The variable `Region` includes 0 distinct regions. Their names are .

The following table shows the claim frequency by `Region`.

```
 Banks Square     Bleachery     Highlands      Lakeview  Piety Corner
        0.092         0.087         0.128         0.129         0.127
The Chemistry     The Lanes    Warrendale
        0.113         0.122         0.089
```

Note that Banks Square, Bleachery, and Warrendale all have relatively low frequencies when compared to the other regions, while Highlands, Lakeview, Piety Corner, and The Lanes all have relatively high frequencies, suggesting grouping is appropriate. This indicates `Region` may be an important predictor of frequency.

The variable `Vehicle Use` has 0 distinct values: .

It is worth noting there are far more Commute and Private vehicles in this data set than Business vehicles.

```
   Length     Class      Mode
    10000 character character
```

The following table shows the claim frequency by `Vehicle Use`.

```
Business  Commute  Private
   0.140    0.138    0.069
```

Note that Business and Commute vehicles have nearly identical frequencies and are nearly double the frequency for Private vehicles. A collapsed grouping of Business and Commute vehicles may be appropriate for model building. This suggests `Vehicle Use` would be an important predictor of frequency.

The variable `Vehicle Body` has 0 distinct values: .

The following table shows the claim frequency by `Vehicle Body`.

```
    Hatchback       Minibus     Panel Van      Roadster         Sedan
        0.114         0.086         0.149         0.143         0.113
Station Wagon           SUV         Truck
        0.065         0.115         0.144
```

Note that Station Wagons and Minibuses, traditionally thought of as family vehicles, have significantly lower frequencies than other types of vehicles. Also of note, Hatchbacks, Suvs, and Roadsters have nearly identical frequencies while Panel Vans and Trucks stand out with significantly higher values. A derived grouping variable may be appropriate. This suggests `Vehicle Body` would be an important predictor of frequency.

Similarly, the below table shows the claim frequency by `Gender`.

```
Female   Male
 0.096  0.137
```

Males clearly have a higher frequency than females, indicating `Gender` is an important predictor of frequency.

Finally, we felt it would be appropriate to break `Age` up into buckets.

The below table shows the claim frequency by `Age`.

```
17-20 21-24 25-30 31-40 41-50 51-60 61-75
0.188 0.156 0.136 0.092 0.089 0.053 0.095
```

It is clear that `Age` is a significant predictor of frequency.

Additionally, we found that `Age` affects frequency differently when accounting for `Gender`:

```
      Female  Male
17-20  0.172 0.214
21-24  0.118 0.210
25-30  0.114 0.169
31-40  0.082 0.106
41-50  0.074 0.110
51-60  0.031 0.084
61-75  0.104 0.082
```

This information suggests an interaction term may be appropriate between `Age` and `Gender`.

4. Model Selection and Interpretation

Based on the above Data Characteristics section, it has been established there are clear correlations and patterns between the frequency of claims, and many of the predictor variables.

In this section we summarize these relationships using regression modeling. We also explain the ways in which we manipulated the data during our selection process.

Based on our investigation of the data, we recommend a Poisson regression model using a logarithmic link function to estimate the mean frequency. The variables used to create the regression model are: `Gender`, `Age`, `Vehicle Use`, `Region`, and `Vehicle Body`. Other than `Gender`, all other variables were transformed into derived grouped variables.

The model was built using a randomly generated subset of 6,500 observations from the data set with 3,500 observations set aside to use for testing the accuracy of the model.

Our final model includes an interaction term between `Gender` and `Age`. An offset equal to the logarithm of `Exposure` is necessary to control for the fact that not all policy holders are exposed to risk for the same amount of time.

The model was fit using an iteratively weighted least squares algorithm and the following table shows the value of the estimated coefficients and their standard errors.

	Estimate	Std. Error
(Intercept)	-2.153	0.223
genderMale	0.253	0.309
age.cat21-24	-0.369	0.265
age.cat25-30	-0.150	0.241
age.cat31-40	-0.507	0.239
age.cat41-50	-0.596	0.292
age.cat51-60	-1.097	0.389
age.cat61-75	-0.166	0.289
use.colpsB&C	0.698	0.114
reg.grpBSBW	-0.474	0.123
reg.grpChem	-0.071	0.141
v.bodFML	-0.558	0.148
v.bodHST	-0.007	0.120
genderMale:age.cat21-24	0.694	0.378
genderMale:age.cat25-30	-0.187	0.367
genderMale:age.cat31-40	-0.037	0.363
genderMale:age.cat41-50	0.130	0.435
genderMale:age.cat51-60	0.164	0.554
genderMale:age.cat61-75	-0.753	0.489

Note that all predictor variables are categorical. The first level of each variable is taken to be the base level for the regression. Also note that all estimated coefficients are on a logarithmic scale because we used a logarithmic link function to build the regression. To convert them back to the scale of the response variable, we need to exponentiate the coefficients.

The following table illustrates the calculation of the expected frequency for a policyholder with the following characteristics:

- age 22
- Male
- living in Highlands
- driving a Truck
- primarily using the vehicle for Business

Variable	Level	Coeff	exp Coeff
Intercept		-2.153	0.116
gender	Male	0.253	1.288
age	21-24	-0.369	0.691

Variable	Level	Coeff	exp Coeff
region	HLPC	0.000	1.000
vehicle body	HST	-0.007	0.993
vehicle use	B&C	0.698	2.010
Age*Gender	Male(21-24)	0.694	2.002
Mean Frequency			**0.412**

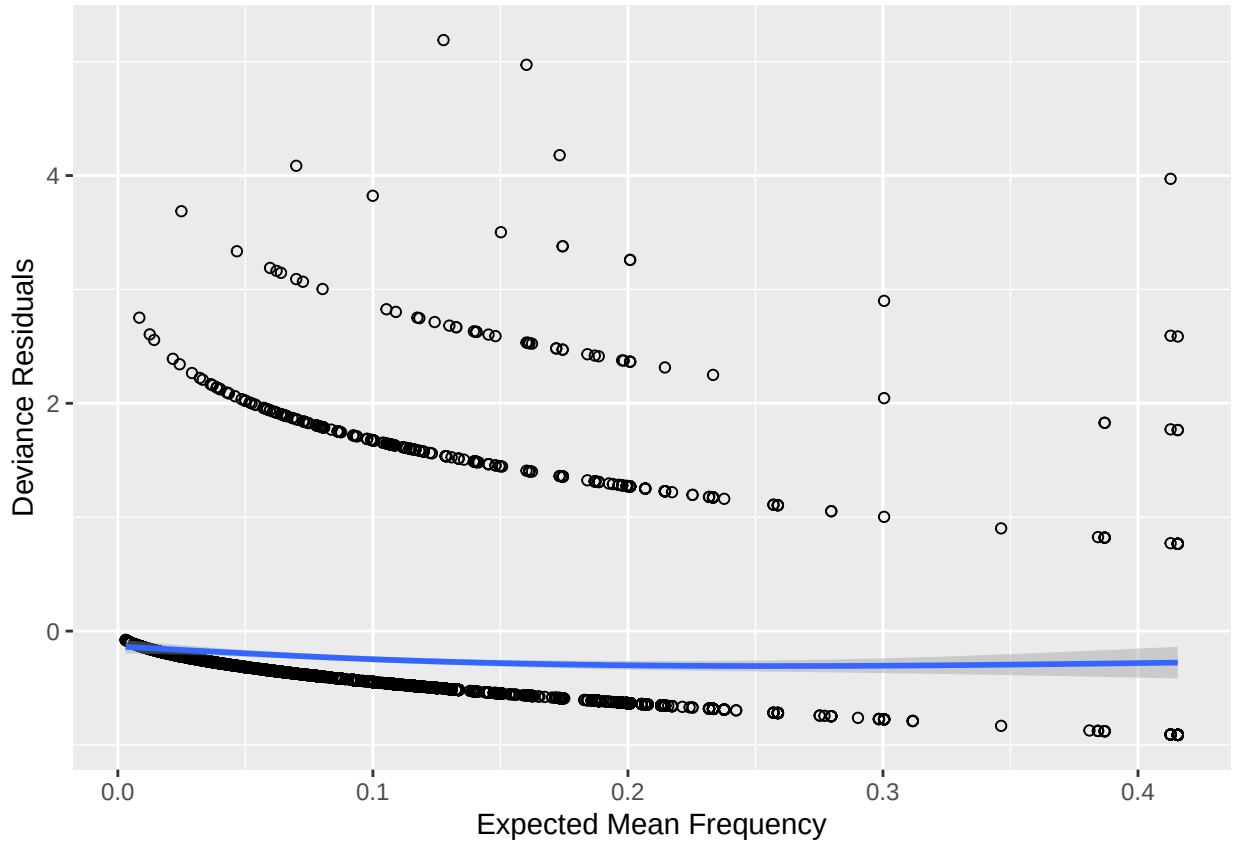
Thus this policyholder has as estimated annual mean frequency equal to 41.2%, the product of exponentiated coefficients

The probability having zero claims would be $\exp(-0.412) = 0.662$, the probability of incurring one claim would be $\exp(-0.412) \cdot 0.412 = 0.273$, and of having two claims it would be $\exp(-0.412) \cdot 0.412^2/2 = 0.056$

Discussion of Model

The residuals for our recommended model did not show significant patterns. The following graph shows the deviance residuals against the expected mean frequency

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



This is a typical plot for count models. The blue line shows the overall estimate of the pattern of residuals as the mean frequency increases. The line is reasonably close to flat indicating no issues with the residuals. Towards the right hand side the grey area widens to indicate the increasing uncertainty however this is to be expected as we have fewer observations with which to estimate the line.

A number of competing models were considered. We began fitting a model with similar variables without adjusting them into the final groupings that were used. This yielded a model with a high AIC and few significant coefficients.

Through the examination described in the above Data Characteristics section, we were able to find groupings within variables that made sense from both a statistical standpoint (similar frequency) as well as from a observational standpoint (family type vehicles).

Our second best model was the same as our final model without the addition of the interaction term between `Gender` and `Age`.

Adding the interaction term decreased the Akaike Information Criterion from 2877.7 to 2873.7, a difference of 4, indicating the interaction term improved the model.

Of all the models we tested, our final model had the lowest MSEP.

```
[1] 0.149027
```

The below table shows the performance of the model using the mean and standard deviations of the average maximum absolute difference over a repeated simulation.

```
       Build    Test
Mean 0.02263 0.02654
SD   0.00458 0.00632
```

Our model performed similarly on both the build and test data sets indicating an accurate model.

Summary and Concluding Remarks

In trying to predict frequency of insurance claims, we found the variables `Age`, `Gender`, `Vehicle Body`, `Vehicle Use`, and `Region` to be significant predictors in our Poisson regression model. Additionally we found an interaction term between `Age` and `Gender` to be useful in making predictions. In the future it would be interesting to examine if these predictions can be used in other areas or if they are solely useful within the region from which the data comes from.

Score Function

```
score <- function(newdata) {
  df <- newdata

  df <- df %>%
    mutate(reg.grp = case_when(
      region == "Highlands" ~ "HLPC",
      region == "Lakeview" ~ "HLPC",
      region == "Piety Corner" ~ "HLPC",
      region == "Bleachery" ~ "BBW",
      region == "Banks Square" ~ "BBW",
      region == "Warrendale" ~ "BSBW",
      region == "The Lanes" ~ "HLPC",
      region == "The Chemistry" ~ "Chem",
      TRUE ~ "Other"))
```

```
  df$reg.grp <- factor(df$reg.grp,
                       levels = c("HLPC", "BSBW", "Chem"))


    df <- df %>%
    mutate(use.colps = case_when(
      vehicle.use == "Business" ~ "B&C",
      vehicle.use == "Commute" ~ "B&C",
      vehicle.use == "Private" ~ "Prv",
      TRUE ~ "Other"))
  df$use.colps <- factor(df$use.colps,
                       levels = c("Prv", "B&C"))

  df$v.bod <- fct_collapse(df$vehicle.body,
                           FML = c("Minibus", "Station Wagon"),
                           MED = c("Hatchback", "SUV", "Sedan"),
                           HST = c("Truck", "Panel Van", "Roadster"))


  bks <- c(16, 20, 24, 30, 40, 50, 60, 76)


  lbs <- c("17-20", "21-24", "25-30", "31-40",
         "41-50", "51-60", "61-75")

  df$age.cat <- cut(df$age, breaks = bks, labels = lbs)

    ans <- predict(fin_mod, newdata = df, type = "response")
  return(ans)
}
```

Appendix

```
mod13 = glm(claims~gender + age.cat + use.colps + reg.grp + v.bod ,
            data = df,subset = btv == "B",
            family = poisson(link = "log"),
            offset = log(exposure))

smod_13 <- summary(mod13)
```

```
mod12 = glm(claims~gender + age.cat + use.colps + reg.grp + v.bod + education,
            data = df,subset = btv == "B",
            family = poisson(link = "log"),
            offset = log(exposure))

smod_12 <- summary(mod12)
```

```
mod11 = glm(claims~gender + age.cat + reg.grp + v.bod + education, data = df,
            subset = btv == "B",
            family = poisson(link = "log"),
```

```
            offset = log(exposure))

smod_11 <- summary(mod11)
```

```
mod10 = glm(claims~gender + poly(age,degree = 2) + use.colps + reg.grp + v.bod + education,
            data = df,subset = btv == "B",
            family = poisson(link = "log"),
            offset = log(exposure))

smod_10 <- summary(mod10)
```

```
mod9 = glm(claims~age.cat + use.colps + reg.grp + v.bod + education,
           data = df,subset = btv == "B",
           family = poisson(link = "log"),
           offset = log(exposure))

smod_9 <- summary(mod9)
```