Regresión Poisson

Ciprian C Hdz.

Consideraremos el dataset **dataCar** que corresponde a datos de pólizas de seguro de vehículos de un año de duración suscritas en 2004 a 2005. Esta tabla contiene 67, 856 pólizas por usuario. Las variables/columnas que contiene la tabla son

Feature	Description
veh_value	The value of the vehicle in $10,000s
exposure	Percentage of year of coverage from 0-1
clm	Whether a claim was filed
numclaims	The number of claims filed
claimcst0	Claim amount (including 0 for no claim)
veh_body	vehicle body type
veh_age	1 (youngest), 2, 3, 4
gender	Gender of policyholder
area	Geographic region
agecat	1 (youngest), 2, 3, 4, 5, 6

Estudiamos primero la naturaleza de los datos:

```
#Estudiamos la naturaleza (tipo) de cada variable
str(dataCar)
```

```
'data.frame':	67856 obs. of  11 variables:
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
 $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
 $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
 $ agecat   : int  2 4 2 2 2 4 4 6 3 4 ...
 $ X_OBSTAT_: Factor w/ 1 level "01101    0    0    0": 1 1 1 1 1 1 1 1 1 1 ...
```

Notemos que clm, numclaims, claimcst0 son variables enteras, sin embargo, si observamos más a detalle:

```
unique(dataCar$clm)
```

0 · 1

```
unique(dataCar$agecat)
```

2 · 4 · 6 · 3 · 5 · 1

```
unique(dataCar$veh_age)
```

3 · 2 · 4 · 1

El código unique() nos da los valores únicos de las respectivas columnas, con esto nos damos cuenta que en realidad son variables categóricas, de esta manera, para hacerle saber a R esta información hacemos uso de la función as.factor()

```
dataCar$veh_age=as.factor(dataCar$veh_age)
dataCar$agecat=as.factor(dataCar$agecat)
dataCar$numclaims=as.numeric(dataCar$numclaims)
dataCar$clm=as.factor(dataCar$clm)
```

Existe una variable que no se encuentra reportada, esta es X_OBSTAT_, dado que no tenemos información de ella, la quitamos de la siguiente forma:

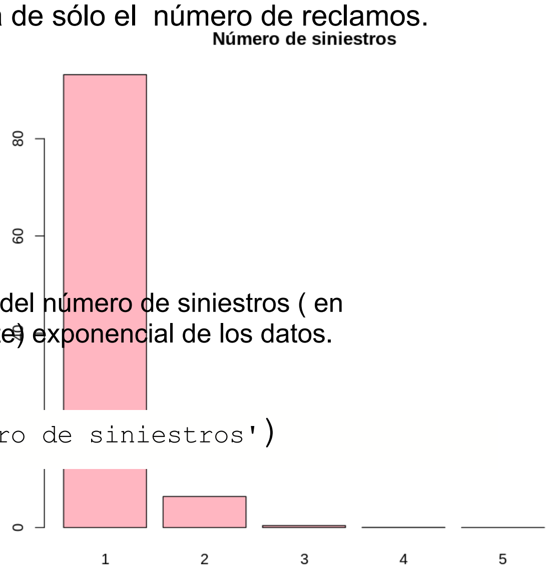
```
dataCar$X_OBSTAT_=NULL
```

Revisamos de nuevo las variables:

```
str(dataCar)
```

```
'data.frame':	67856 obs. of  10 variables:
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ numclaims: num  1 1 1 1 1 1 1 1 1 1 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
 $ veh_age  : Factor w/ 4 levels "1","2","3","4": 3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
 $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
 $ agecat   : Factor w/ 6 levels "1","2","3","4",..: 2 4 2 2 2 4 4 6 3 4 ...
```

Modelación[1] Queremos estimar el número de reclamos (numclaims) respecto al porcentaje de cobertura (exposure), pues podríamos pensar que a mayor cobertura o exposición, es más probable que en ese año la póliza tenga un siniestro. La exposición es un número entre 0 y 1, por ejemplo, si solamente se aseguró 6 meses la exposición es entonces del 50 %. veamos cómo se ve el histograma de sólo el número de reclamos.



[1] La gráfica de barras rosa representa la frecuencia relativa del número de siniestros (en porcentajes), esto ayuda a enfatizar la forma (aparentemente) exponencial de los datos.
El código correspondientes es:

```
tabla=prop.table(table(dataCar$numclaims))
barplot(100*tabla,col='lightpink', main='Número de siniestros')
```

Notemos que su distribución no es (aparentemente) normal, sino más bien una tipo exponencial, aunado a que estamos considerando el conteo del número de reclamos (es decir, una variable entera no negativa). También debemos de considerar el porcentaje de cobertura, esto conlleva a pensar que la variable respuesta es realmente una tasa de cambio: Número de reclamos/ porcentaje de cobertura del seguro. *La forma en que la variable respuesta está representada nos indica que un posible modelo de ajuste de regresión es mediante la regresión Poisson.*

Modelos Lineales Generalizados (MLG). Justamente el caso en que la variable respuesta (número de reclamos en estos casos) no corresponden a una distribución normal, sino a una distribución de la familia exponencial (distribución Poisson), conlleva a hablar de modelos lineales generalizados. En particular, si la variable respuesta[2] se distribuye Poisson hablamos de la regresión Poisson.

Una forma de aportar evidencia sobre la conjetura anterior es la siguiente: sabemos que la distribución Poisson tiene la particularidad que justamente la media y la varianza coinciden.

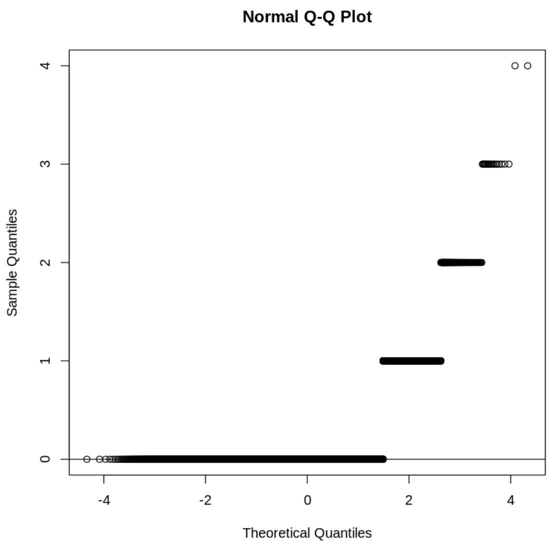
```
var(dataCar$numclaims)
```

0.0773973711246176

```
mean(dataCar$numclaims)
```

0.072757014854987

También podemos intentar ajustar una normal mediante el qqplot.



Matemáticamente:

Si sólo consideramos Y como una variable de conteo (solo toma valores no negativos enteros), entonces

$numClaims_i \sim Poisson(\mu_i)$ donde buscamos estimar μ_i mediante la relacioń

$$\log(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Si además en particular la variable respuesta representa cierta tasa, entonces lo anterior se transforma en

$\frac{numClaims_i}{exposure_i} \sim Poisson(\mu_i)$ donde buscamos estimar μ_i mediante la relación

$$\log(\mu_i) = \log(exposue_i) + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

[2] Que realmente en nuestro caso, el número de reclamos/ el porcentaje es lo que se distribuye Poisson.

Finalmente ajustamos el modelo en R y observamos la salida.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9069  -0.4521  -0.3457  -0.2212   4.5350

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.667803   0.326382  -2.046  0.04075 *
veh_value      0.023980   0.017251   1.390  0.16451
veh_bodyCONVT -1.677911   0.668319  -2.511  0.01205 *
veh_bodyCOUPE -0.514222   0.337512  -1.524  0.12762
veh_bodyHBACK -0.975986   0.318672  -3.063  0.00219 **
veh_bodyHDTOP -0.829563   0.327953  -2.530  0.01142 *
veh_bodyMCARA -0.369072   0.409609  -0.901  0.36757
veh_bodyMIBUS -0.987654   0.350494  -2.818  0.00483 **
veh_bodyPANVN -0.855177   0.339270  -2.521  0.01171 *
veh_bodyRDSTR -0.567799   0.660305  -0.860  0.38984
veh_bodySEDAN -0.923014   0.318063  -2.902  0.00371 **
veh_bodySTNWG -0.906750   0.318239  -2.849  0.00438 **
veh_bodyTRUCK -0.944163   0.328476  -2.874  0.00405 **
veh_bodyUTE   -1.108652   0.322203  -3.441  0.00058 ***
veh_age2       0.054363   0.044623   1.218  0.22312
veh_age3      -0.055849   0.048233  -1.158  0.24691
veh_age4      -0.114908   0.056801  -2.023  0.04307 *
genderM       -0.026181   0.030135  -0.869  0.38495
areaB          0.053158   0.042802   1.242  0.21425
areaC          0.005108   0.038994   0.131  0.89577
areaD         -0.110402   0.052973  -2.084  0.03715 *
areaE         -0.031935   0.057878  -0.552  0.58111
areaF          0.063729   0.066158   0.963  0.33541
agecat2       -0.173250   0.054187  -3.197  0.00139 **
agecat3       -0.230051   0.052896  -4.349 1.37e-05 ***
agecat4       -0.256934   0.052744  -4.871 1.11e-06 ***
agecat5       -0.474475   0.059120  -8.026 1.01e-15 ***
agecat6       -0.453279   0.067686  -6.697 2.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25507  on 67855  degrees of freedom
Residual deviance: 25332  on 67828  degrees of freedom
AIC: 34823

Number of Fisher Scoring iterations: 6
```

Interpretación de la salida: Si las covariables incrementan en una unidad, la variable de conteo o razón incrementara por un factor de exp(covariable).

Es decir, los valores (exponenciados) de cada coeficiente es el **factor multiplicativo** que usamos para calcular el número estimado de siniestros cuando cada variable se incrementa en 1 unidad. En particular, cuando la variable es categórica, el coeficiente exponenciado es el término multiplicativo relativo al nivel base. Por otro lado, como los coeficientes relacionados a agecat son negativos, a medida que aumenta la edad del conductor, el número de reclamos o siniestros disminuye.

Selección de variables

Mediante el uso de la librería leaps, podemos hacer uso del método de selección de variables para observar si podemos reducir el modelo a uno más parsimonioso.

Primero convertimos la tabla de datos en una matriz numérica.

```
%%R
#Matriz numérica
X=dataCar[-c(3,4,5,11)]
y=dataCar[,4]
Xy=cbind(X,y)
```

donde -c(3,4,5,11) corresponde a quitar las columnas 3,4,5 y 11 del dataframe.

```
%%R
subset=bestglm(Xy=Xy, family=poisson(link='log'),IC='BIC', method='exhaustive',offset=log(Xy$exposure))
```

```
R[write to console]: Morgan-Tatar search since family is non-gaussian.

R[write to console]: Note: factors present with more than 2 levels.
```

```
%%R
#Vemos las variables que quedan para el mejor modelo:
subset$BestModel
```

```
Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights, 
    offset = ..1)

Coefficients:
(Intercept)    veh_value     exposure      agecat2      agecat3      agecat4  
   -1.39887      0.04843     -0.47434     -0.16755     -0.22244     -0.24782  
    agecat5      agecat6  
   -0.46184     -0.44367  

Degrees of Freedom: 67855 Total (i.e. Null);  67848 Residual
Null Deviance:      25510 
Residual Deviance: 25330        AIC: 34780
```

Esto nos dice que un posible modelo para utilizar es usando las variables es dejar el intercepto, veh_value y agecat2. Ajustamos el modelo:

```
Call:
glm(formula = numclaims ~ veh_value + agecat, family = poisson(link = "log"),
    data = dataCar, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1275  -0.4541  -0.3481  -0.2228   4.4981

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.69444    0.04811 -35.219  < 2e-16 ***
veh_value    0.04848    0.01064   4.555 5.24e-06 ***
agecat2     -0.16964    0.05390  -3.148  0.00165 **
agecat3     -0.22741    0.05240  -4.340 1.43e-05 ***
agecat4     -0.25190    0.05244  -4.804 1.56e-06 ***
agecat5     -0.47171    0.05872  -8.033 9.50e-16 ***
agecat6     -0.45266    0.06695  -6.761 1.37e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25507  on 67855  degrees of freedom
Residual deviance: 25396  on 67849  degrees of freedom
AIC: 34845

Number of Fisher Scoring iterations: 6
```

Notamos que ahora todas son significativas.

Trabajo Futuro.: Aunque el modelo de selección de variables puede ser una alternativa, es importante realizar un análisis posterior comparando ambos modelos, por ejemplo mediante un análisis ANOVA.

Observación: El resultado de bestglm para el método de mejor subcojunto puede variar dependiendo de la métrica seleccionada.