Universidade Federal do Paraná
Departamento de Estatística

Regressão para Dados de Contagem - Segurança e privatização ferroviária na Grã-Bretanha

CE225 - Modelos Lineares Generalizados

Francielle Przibiciem de Mattos GRR20124686
Guilherme Henrique Stadler de Mello GRR20124678
Simone Matsubara GRR20124663

Curitiba, 24 de Novembro de 2017

Sumário

1 Introdução

A Grã-Bretanha, antiga Albion, é uma das muitas Ilhas Britânicas da Europa que abrange a maior parte do Reino Unido. Nesta ilha estão três das quatro nações britânicas: Escócia, na parte norte; Inglaterra, no sul; e País de Gales, a oeste.

Por ser uma ilha relativamente pequena e contar com uma malha ferroviária extensa, viajar de trem em distâncias médias e longas é fácil e confortável.

A rede ferroviária da Grã-Bretanha está composta por 24 companhias particulares, todas regulamentadas pela National Rail, onde atendem mais de 2.500 estações. De Leste a Oeste, de Norte a Sul, a maioria das viagens se realiza em trens de Alta Velocidade.

De uma forma geral, o transporte ferroviário de passageiros é considerado bastante seguro para os seus utilizadores. No âmbito ferroviário, um acidente é sinônimo de um acontecimento súbito, indesejado ou involuntário, ou de uma cadeia de acontecimentos dessa natureza com consequências danosas. Considera-se como acidente grave qualquer colisão ou descarrilamento de comboios que tenha por consequência, no mínimo, um morto, ou cinco ou mais feridos graves, ou danos significativos no material circulante, na infraestrutura ou no ambiente e qualquer outro acidente semelhante com impacte manifesto na regulamentação de segurança ferroviária ou na gestão da segurança.

Os dados provém da fonte: Andrew W. Evans (2007). "Segurança ferroviária e privatização ferroviária na Grã-Bretanha", Análise e Prevenção de Acidentes, Vol. 39, pp. 510-523.

2 Material e Métodos

2.1 Material

O estudo foi realizado com 57 observações, com o objetivo de modelar a probabilidade de morte por colisão de trem entre os anos de 1946 e 2002; Os dados do último ano da pesquisa, de 2003, foram desconsiderados, pois estavam incompletos, possivelmente foi interrompido na metade do ano, ou antes. As variáveis presentes neste estudo são:

Ano: Ano observado
Km: Trilho da linha principal em milhões de Quilômetros
Inv: Número de acidentes de invasão
Fat: Números de acidentes fatais
Coli: Número de acidentes por colisão
Mort: Mortes por colisão de trem/veículo rodoviário (variável resposta)
Acit: Número total de acidentes com o trem em movimento e com o trem parado
Acif: Número total de acidentes fatais com o trem em movimento e com o trem parado.

Tabela 1: Primeiras 6 observações do conjunto de dados.

Ano	Km	Inv	Fat	Coli	Mort	Acit	Acif
1946	600	9	38	7	16	370	378
1947	571	10	99	9	20	298	304
1948	589	10	48	11	20	273	273
1949	613	4	5	5	5	278	281
1950	618	5	13	14	19	247	250
1951	605	7	47	7	8	231	233

2.2 Métodos

Como temos uma contagem de eventos em unidades de tempo/espaço com ocorrências dos eventos sendo independentes e identicamente distribuídos, primeiramente utilizaremos a distribuição Poisson com função de ligação logarítmica, pois é o modelo mais comum usado para estas situações.

O modelo de Poisson é apresentado por:

$$f(y, \mu) = \frac{e^{-\mu}\mu^{y}}{y!}$$

E seguindo a função de ligação logarítmica, o modelo log-linear é:

$$y_i|x_i \sim Poisson(\mu_i)$$

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

3 Análise Descritiva

A fim de conhecer melhor os dados com os quais estamos estudando, e prever possíveis problemas e sugestões de modelagem, mostraremos a seguir uma análise descritiva e exploratória dos dados.

```
##       Ano             Km             Inv             Fat        
##  Min.   :1946   Min.   :372.0   Min.   : 0.000   Min.   :  0.00  
##  1st Qu.:1960   1st Qu.:430.0   1st Qu.: 1.000   1st Qu.:  2.00  
##  Median :1974   Median :452.0   Median : 2.000   Median :  7.00  
##  Mean   :1974   Mean   :492.9   Mean   : 3.421   Mean   : 16.74  
##  3rd Qu.:1988   3rd Qu.:584.0   3rd Qu.: 5.000   3rd Qu.: 17.00  
##  Max.   :2002   Max.   :618.0   Max.   :10.000   Max.   :120.00  
##       Coli            Mort            Acit            Acif       
##  Min.   : 1.000   Min.   : 1.000   Min.   : 16.0   Min.   : 16.0  
##  1st Qu.: 3.000   1st Qu.: 4.000   1st Qu.: 48.0   1st Qu.: 49.0  
##  Median : 6.000   Median : 7.000   Median : 73.0   Median : 73.0  
##  Mean   : 6.474   Mean   : 9.088   Mean   :114.4   Mean   :116.4  
##  3rd Qu.:10.000   3rd Qu.:13.000   3rd Qu.:183.0   3rd Qu.:184.0  
##  Max.   :14.000   Max.   :31.000   Max.   :370.0   Max.   :378.0
```

Devido a amplitude assimétrica observada em algumas covariáveis, exploraremos outros métodos de analisar estas assimetrias.

O histograma é uma alternativa para observar a forma da distribuição dos dados. Na figura 1, podemos observar uma predominância de assimetria a esquerda, naquelas mesmas covariáveis que observamos ter uma amplitude assimétrica no summary acima.

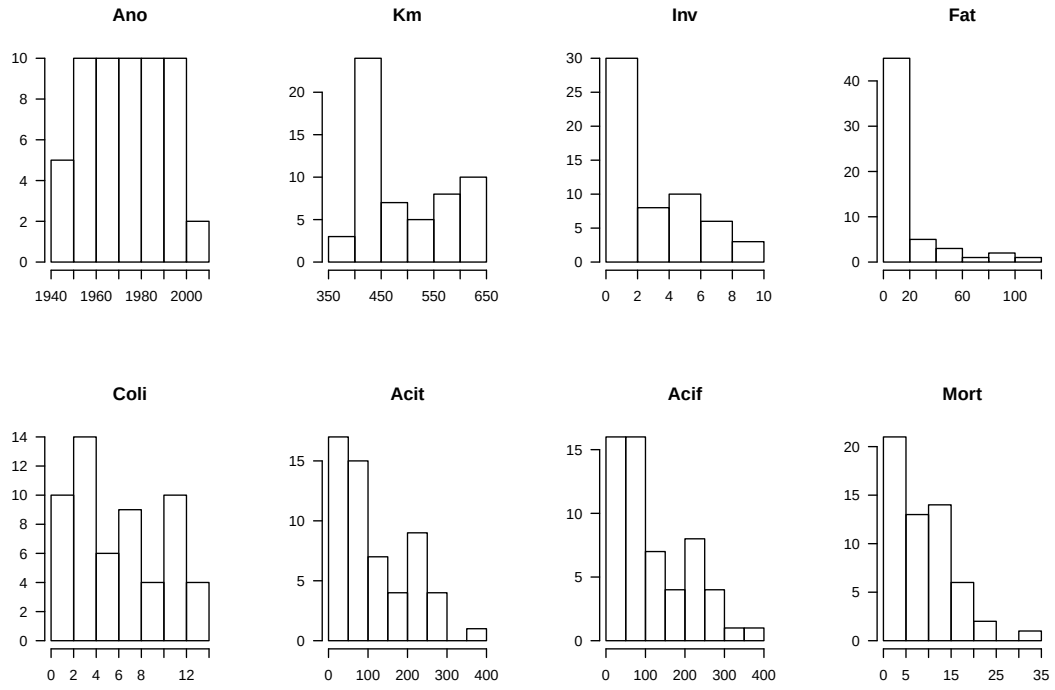
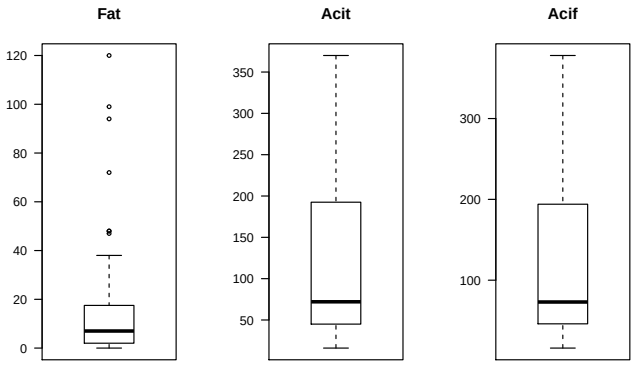


Figura 1: Histogramas

Para algumas das variáveis, notamos que o terceiro quartil, é menor que a metade do valor máximo das observações, o que pode ser devido à uma assimetria à esquerda, decidimos então, observar estas variáveis em um gráfico de box-plot.



Decidimos por exibir, o gráfico construído das correlações, para mostrar a correlação perfeita existente entre duas covariáveis, **Acit** e **Acif**, que respectivamente são: Número total de acidentes com o trem em movimento/trem parado e Número total de acidentes fatais com o trem em movimento/trem parado.

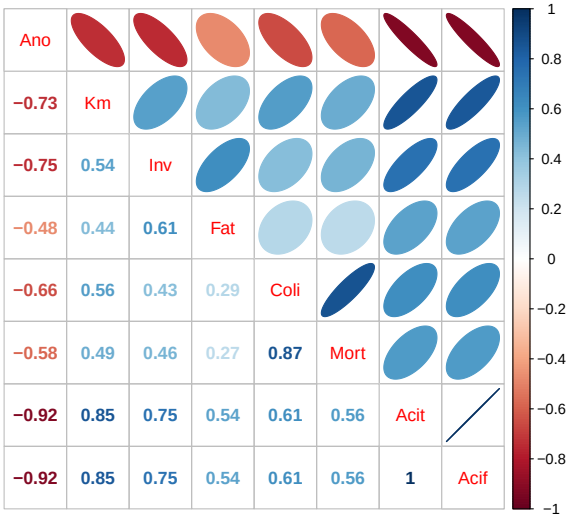


Figura 2: Gráfico de Correlação

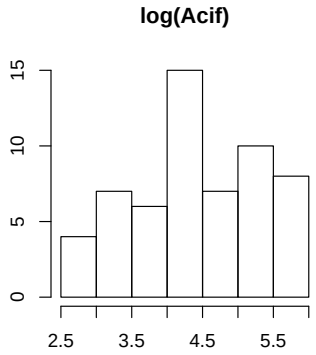
Apresentada na tabela a seguir, estão algumas das observações destas covariáveis:

Ano	Acit	Acif
1946	370	378
1947	298	304
1948	273	273
1949	278	281
1950	247	250
1951	231	233
1952	249	254
1953	278	286

Estas duas variáveis, explicam a mesma coisa e não faria sentido manter as duas no estudo, por conta desse ocorrido, optamos pela remoção da covariável **Acit**, dado que, pelo o que pudemos observar, todas as observações desta covariável, estão também contidas na covariável **Acif**, com a diferença de que nesta, estão incluídas as vítimas fatais dos acidentes.

3.1 Transformando as variáveis

Dado que na análise descritiva dos dados, obtivemos algumas variáveis assimétricas, aplicaremos uma transformação logarítmica na variável Acif, para simetrizar a distribuição da mesma, e para as outras duas (Inv e Fat), manteremos como estão, pois elas contém zeros, o que gera erro ao criar o modelo.



4 Ajuste dos Modelos de Regressão

Como citado no início deste relatório, vamos primeiramente considerar o modelo GLM, com resposta Poisson:

$$log(\mu_i) = \beta_0 + \beta_1 Ano_i + \beta_2 Km_i + \beta_3 Inv_i + \beta_4 Fat_i + \beta_5 Coli_i + \beta_6 log(Acif_i)$$

E também vamos ajustar um M.L.G com distribuição Binomial Negativa e função de ligação logaritimica.

$$y_i | x_i \sim BinomialNegativa(\mu_i, k)$$

$$log(\mu_i) = \beta_0 + \beta_1 Ano_i + \beta_2 Km_i + \beta_3 Inv_i + \beta_4 Fat_i + \beta_5 Coli_i + \beta_6 log(Acif_i)$$

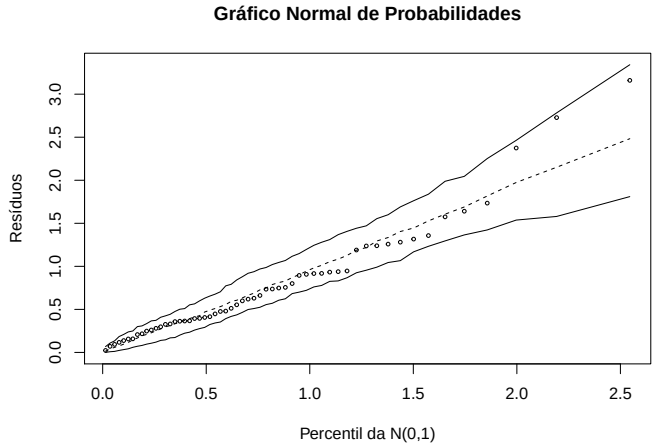
5 Escolha do Modelo

Para selecionarmos um dos modelos, utilizaremos o Critério de informação AIC e a verossimilhança dos modelos.

```
#         ajuste      aic    logLik
#1       poisson 285.7006 -135.8503
#2 bin. negativa 286.9916 -135.4958
```

Apesar de os dois modelos estarem com as verossimilhanças muito similares, o modelo pelo qual optaremos, será o de Poisson, pois apresentou menor AIC e uma maior verossimilhança.

Para verificar a adequação do modelo ajustado, usaremos um gráfico Normal de probabilidades com envelopes simulados.



Podemos verificar a adequação do modelo ajustado, pois os resíduos estão dispersos aleatoriamente dentro do envelope.

Para analisar quais são as variáveis explicativas, foram testados os métodos de seleção de variáveis Forward, Backward e Stepwise, todos obtiveram o mesmo resultado, tendo que apenas as variáveis Inv e Coli são significativas.

```
##
## Call:  glm(formula = Mort ~ Inv + Coli, family = "poisson", data = dados1)
##
## Coefficients:
## (Intercept)          Inv         Coli
##     0.90038      0.03403      0.15320
##
## Degrees of Freedom: 56 Total (i.e. Null);  54 Residual
## Null Deviance:     257.9
## Residual Deviance: 57.61  AIC: 279.3
```

Deixando o modelo da seguinte maneira:

$$log(\mu_i) = 0,90038 + 0,03403Inv_i + 0,15320Coli_i$$

6 Análise de Diagnóstico

Vamos primeiramente avaliar a qualidade do nosso ajuste.

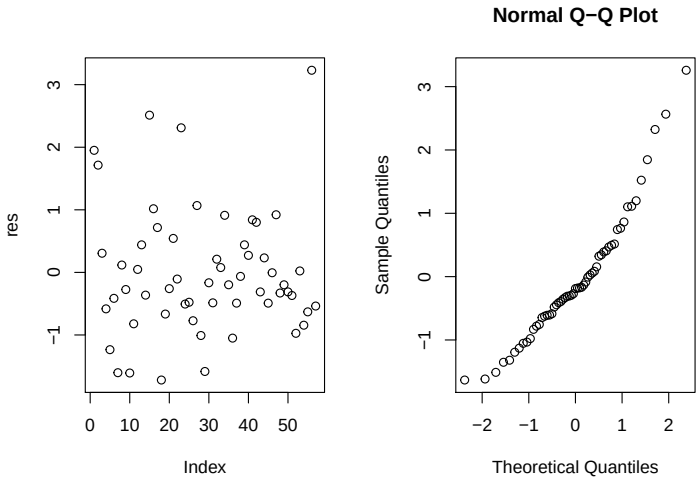


Figura 3: Resíduos Quantílicos Aleatorizados

Como vemos na Figura 3, os resíduos estão dispersos dentro do intervalo aceitável e em torno de zero. No gráfico da direita, os resíduos aparentam ter uma aderência razoável à distribuição Normal, apesar de um desvio na cauda superior.

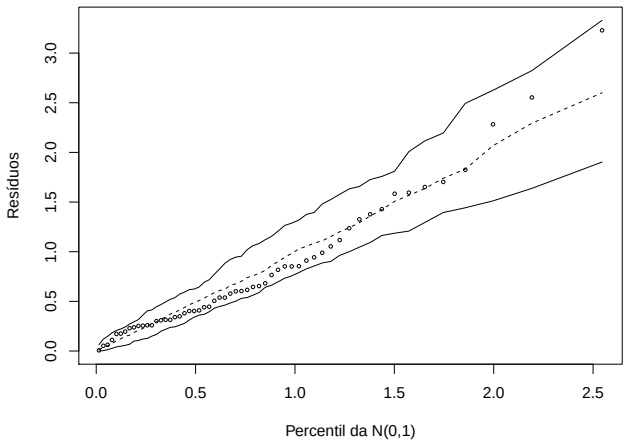


Figura 4: Gráfico Normal de Probabilidades

Devido ao desvio apresentado no Q-Q plot, figura 3, foi verificada a adequação do modelo à distribuição Normal, por meio de um envelope simulado, onde os resíduos se apresentaram dispersos dentro do intervalo esperado. Portanto assim, podemos concluir que o ajuste está correto.

Em uma análise de resíduos do modelo, foi observado que aparentemente não existem candidatos a outliers e que o modelo atende os pressupostos.

6.1 Quase-verossimilhança

Pode ser usado como um modelo alternativo, quando se tem dados com superdisperção.

```
modqv <- glm(Mort ~ Ano + Km + Inv + Fat + Coli + lAcif, data = dados1, family = 'quasipoisson')
```

Portanto, testamos um modelo de quase-verossimilhança para vermos se derrepente ele mostra um melhor ajuste que o modelo de Poisson escolhido anteriormente, mas realizadas as análises de diagnóstico, constatamos que ele praticamente não diferiu com as análises que obtivemos com o modelo de Poisson.

7 Considerações finais

Pudemos concluir que das seis variáveis inicialmente presentes no estudo, apenas duas conseguiram explicar o principal motivo de mortes por colisão ocorridos nas redes ferroviárias na Grã Betanha, que foram por invasão (de trilho) e pela quantidade de acidentes que acorrem todo ano, e que por vezes causam mortes.