Estatística Numérica Computacional
Trabalho nº3
Grupo II

Marta Paz nº49861
Rafael Almeida nº49788
Rafael Gameiro nº50677
Ricardo Pinto nº49811

November, 2018

Utilize o ficheiro de dados "Dados.T3.G2.1819.txt". Com o objetivo de tentar estabelecer fatores associados a temperaturas baixas em certo local, deve analisar os dados deste ficheiro que contém informação sobre temperaturas médias anuais e possíveis fatores condicionantes, em períodos de 5 anos num século de dados (20 períodos). O ficheiro compreende as seguintes variáveis:

- *temp.b* : Número de anos com temperaturas média anual inferior a 15°, em cada período - a variável resposta Y;
- *prec.b* : Precipitação média no período (mm) - covariável x_1;
- *co2.b* : Níveis médios de dióxido de carbono na atmosfera no período (ppm) - covariável x_2;
- *ozono.b* : Níveis média de ozono na atmosfera no período (ppm) - covariável x_3;

Objetivo:

Ajustar um **modelo de regressão Poisson com ligação identidade** a este conjunto de dados. Siga os passos seguintes:

Alínea 1

Faça uma breve análise preliminar dos dados, com as usuais medidas descritivas e gráficos adequados.

Resolução

Para a análise preliminar dos dados, calculámos diversas medidas descritivas para cada dado fornecido, nomeadamente: a média, a variância, o desvio padrão e o coeficiente de variação. A média de uma variável é o valor obtido fazendo a divisão entre a soma de todos os valores e o número total de valores. Ou seja,

$$E(X) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Assim, temos:

- $E(temp.b) = 1.5$
- $E(prec.b) = 1363.314$
- $E(co2.b) = 395.127$
- $E(ozono.b) = 0.05451$

A variância de uma variável aleatória é uma medida de dispersão que indica o "quão longe" cada valor está do valor esperado. Para calcular a variância, temos a seguinte fórmula:

$$V(X) = E(X^2) - E(X)^2$$

E, consequentemente, temos que:

- $V(temp.b) = 1.315789$
- $V(prec.b) = 5840.207$
- $V(co2.b) = 319.5272$
- $V(ozono.b) = 6.313579e - 06$

O desvio padrão é a raiz quadrada da variância. Ou seja,

$$\sigma = \sqrt{V(X)}$$

Assim, para cada valor dos dados temos que:

- $V(temp.b) = 1.147079$
- $V(prec.b) = 76.42125$
- $V(co2.b) = 17.87532$
- $V(ozono.b) = 0.002512684$

Através do calculo da variância e do desvio padrão podemos concluir que os valores de temperatura e ozono são próximos da média de cada um uma vez que os valores de variância e desvio padrão são próximos de 0. Pelo contrário, relativamente aos valores da precipitação e níveis de dióxido de carbono, estes têm uma grande diferença entre os valores e a média uma vez que a variância e o desvio padrão são elevados.

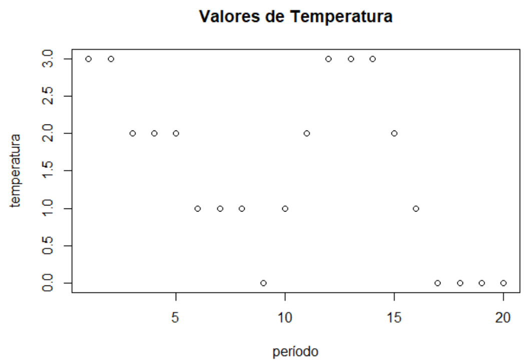
Por último, temos o coeficiente de variação. Este mede o desvio em relação à média e é calculado através da seguinte fórmula:

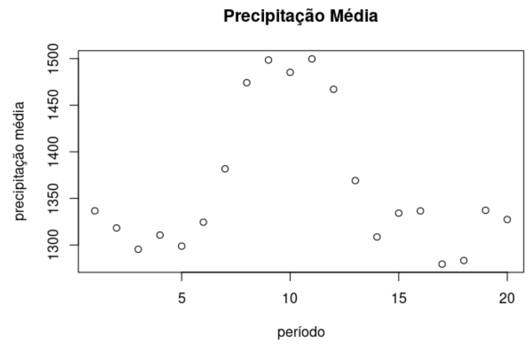
$$CV = \frac{\sigma}{\mu}$$

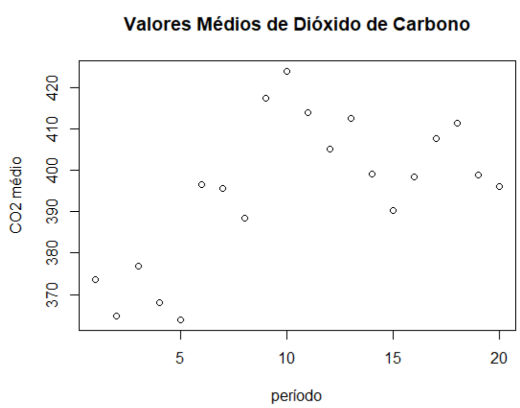
Desta forma, para cada um dos dados temos que:

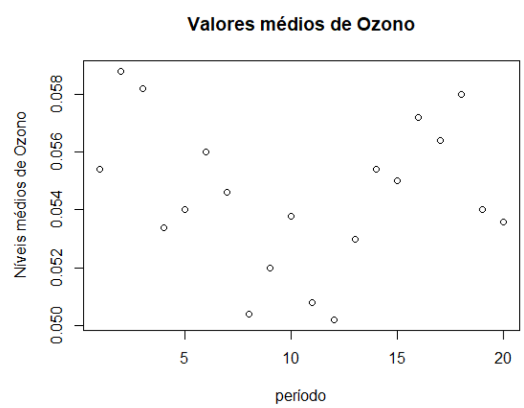
- $V(temp.b) = 0.7647191$
- $V(prec.b) = 0.0560555$
- $V(co2.b) = 0.04523944$
- $V(ozono.b) = 0.04609583$

Para além destes calculos, também obtivémos os seguintes gráficos para os diferentes dados:


Valores de Temperatura
temperatura
período


Precipitação Média
precipitação média
período


Valores Médios de Dióxido de Carbono
CO2 médio
período


Valores médios de Ozono
Níveis médios de Ozono
período

Alínea 2

Descreva o modelo, identificando na família exponencial a distribuição da variável resposta Y - Poisson(λ) - e identificando a função ligação entre a média de Y e o preditor linear.

Resolução

Uma vez que estamos a ajustar um modelo de regressão Poisson com ligação identidade ao conjunto de dados fornecido, temos que, para a função de ligação entre a média Y e o preditor linear, a seguinte função:

$$g(\mu) = \mu$$

Também sabemos que o preditor linear(η) tem a seguinte fórmula:

$$\eta = \beta_0 z_{i1} + \beta_1 z_{i2} + ... + \beta_k z_{i(k+1)} <=> \eta = z_i' \beta$$

e que $\mu = \eta$. Pelo que, ficamos com a função de ligação igual a:

$$g(\mu) = \eta$$

Alínea 3

Escreva a função de log-verosimilhança dos dados e as equações de verosimilhança para estimar os coeficientes de regressão β do modelo.

Resolução

A f.m.p da distribuição de Poisson é:

$$f(y|\lambda) = \frac{(e^{-\lambda} * \lambda^y)}{y!}$$

A função verosimilhança pode ser calculada da seguinte forma:

$$\begin{aligned} L(\lambda) &= f((y_i, ..., y_n)|\lambda) \\ &= \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \frac{(e^{-\lambda} * \lambda^{y_i})}{y_i!} = \prod_{i=1}^{n} \lambda^{y_i} * \frac{e^{-\lambda}}{y_i!} \\ &= e^{-n\lambda} * \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \end{aligned}$$

Sabendo que a função log-verosimilhança corresponde a $ln(L(\lambda))$ temos que:

$$
\begin{aligned}
ln((L(\lambda))) = ln\left(e^{-n\lambda} * \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}\right) = \\
= ln(e^{-n\lambda}) + ln\left(\frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}\right) = \\
= -n\lambda + \left(\sum_{i=1}^{n} y_i\right) ln\lambda - ln\left(\prod_{i=1}^{n} y_i!\right)
\end{aligned}
$$

Segundo o enunciado, temos que ajustar um modelo de regressão Poisson com ligação identidade. Como tal, temos que $\mu = \lambda = \eta$. Também sabemos que $\eta = z_i'\beta$, para as duas equações anteriores, temos que:

$$L(\beta) = e^{-nz_i'\beta} \frac{(z_i'\beta)^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

$$l(\beta) = ln(L(\beta)) = -nz_i'\beta + \left(\sum_{i=1}^{n} y_i\right) ln\,(z_i'\beta) - ln\left(\prod_{i=1}^{n} y_i!\right)$$

Por último sabemos que as equações de verosimilhança correspondem à derivada parcial da função log-verosimilhança em função do β_j, sendo j um dos parâmetros em estudo:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad , j = 1, ..., p$$

Alínea 4

Obtenha a função score e a matriz de informação de Fisher do modelo.

Resolução

A função score pode ser calculada através da derivação da função log-verosimilhança em função do β , ou seja,

$$S(\beta) = \frac{\partial l(\beta)}{\partial \beta},$$

Temos que:

$$
\begin{aligned}
S(\beta) = \frac{\partial l(\beta)}{\partial \beta} &= \frac{\partial\left(-nz_i'\beta + \left(\sum_{i=1}^{n} y_i\right) ln(z_i'\beta) - ln\left(\prod_{i=1}^{n} y_i!\right)\right)}{\partial \beta} \\
&= -nz_i' + \sum_{i=1}^{n} y_i \frac{(z_i'\beta)'}{z_i'\beta} - 0 = -nz_i' + \sum_{i=1}^{n} y_i \frac{z_i'}{z_i'\beta}
\end{aligned}
$$

Usando a função score, podemos calcular a matriz hessiana a partir da seguinte forma:

$$
\begin{aligned}
H(\beta) = l(\beta)'' = S(\beta)' &= \left(-nz_i' + \sum_{i=1}^{n} y_i \frac{1}{\beta} \right)' \\
&= 0 + \left(\sum_{i=1}^{n} y_i \right)' \frac{1}{z_i'\beta} + \sum_{i=1}^{n} y_i \left(\frac{1}{\beta} \right)' \\
&= 0 + 0 + \sum_{i=1}^{n} y_i \left(\frac{1' * (\beta) - 1 * (\beta)'}{(\beta)^2} \right) \\
&= \sum_{i=1}^{n} y_i \left(\frac{-1}{\beta^2} \right)
\end{aligned}
$$

A partir da fórmula obtida anteriormente, podemos obter a matriz de Informação de Fisher, fazendo:

$$
\Im(\beta) = E[-H(\beta)] = E\left[-\sum_{i=1}^{n} y_i * \left(\frac{-1}{\beta^2} \right) \right] = E\left[\sum_{i=1}^{n} y_i * \left(\frac{1}{\beta^2} \right) \right]
$$

Como não depende de $x_1, ..., x_n$, temos que

$$
\Im(\eta) = E[-H(\eta)] = \sum_{i=1}^{n} y_i * \left(\frac{1}{\beta^2} \right)
$$

Outra alternativa ao cálculo da matriz é através do produto matricial, dado por:

$$
\Im(\eta) = \mathbf{Z'WZ}, \quad W = diag\left(w_i * = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2}{Var(Y_i)} \right)
$$

Sendo Z a matriz das covariáveis onde a primeira coluna são 1's e as restantes os valores das covariáveis x_1, x_2 e x_3.

$$
\begin{bmatrix}
1 & z_{11} & z_{12} & ... & z_{1p} \\
1 & z_{21} & z_{22} & ... & z_{2p} \\
\vdots & \vdots & \vdots & & \vdots \\
1 & z_{n1} & z_{n2} & ... & z_{np}
\end{bmatrix}
$$

Alínea 5

Construa e programe no **R** o algoritmo iterativo de estimação dos coeficientes de regressão que se baseia no método dos scores de Fisher. Estime os coeficientes do seu modelo, apresente-os e comente. Apresente em gráfico igualmente os resíduos do modelo e comente.

Resolução

Neste momento, podemos escrever a recursão que nos dará os valores de β através estimação dos coeficientes de regressão que se baseia no método dos scores de Fisher. Utilizando a expressão podemos obter o próximo β e assim sucessivamente até chegarmos ao melhor valor.

$$\beta^{k+1} = \left(\mathbf{Z'W}^{(k)}\mathbf{Z}\right)^{-1}\mathbf{Z'W}^{(k)}\mathbf{u}^{(k)}$$

Sendo k o número da iteração currente, $u^{(k)}$ um vetor de elementos genéricos dado por:

$$\begin{aligned} u_i^{(k)} &= \sum_{j=1}^{p} z_{ij}\beta_j^k + (y_i - \mu_i^{(k)})\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} = \eta_i^{(k)} + (y_i - \mu_i^{(k)})\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \\ &= \eta_i^{(k)} + (y_i - \eta_i^{(k)})\frac{\partial \eta_i^{(k)}}{\partial \eta_i^{(k)}} = \eta_i^{(k)} + y_i - \eta_i^{(k)} = y_i \end{aligned}$$

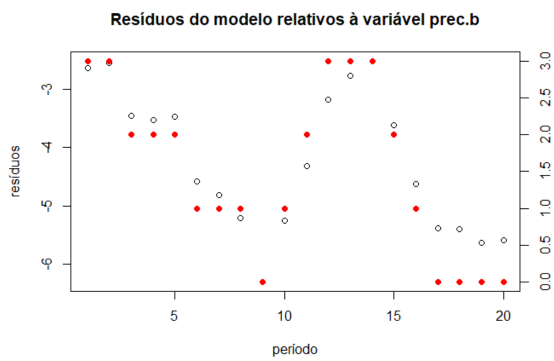
```r
#EXERCICIO 5

# Carregar a biblioteca stats4, para otimização numérica
library(stats4)

prevZ <-data.frame(dados$prec.b, dados$co2.b, dados$ozono.b)
Z <- cbind(1, as.matrix(prevZ))
ZT <- t(Z)
beta0 <- c(1,1,1,1)

calculateError <- function(x, y) {
  d1 <- (x[1] - y[1])^2
  d2 <- (x[2] - y[2])^2
  d3 <- (x[3] - y[3])^2
  d4 <- (x[4] - y[4])^2

  error <- sqrt(d1 + d2 + d3 + d4)
  return(error)
}

metScoresPoi <- function(beta0=c(1,1,1,1),tolerancia=0.0000001,x){

  contador=1
  erro <- 1
  W <- diag(x=20)
  diag(W) <- 1/(Z%*%beta0)
  u <- dados$temp.b
  beta.antes <- beta0
  n = length(x)

  beta.depois <- 1:4

  while(erro > tolerancia){
    beta.depois <- solve(ZT%*%W%*%Z)%*%(ZT%*%W%*%u)
    erro <- calculateError(beta.antes, beta.depois)
    beta.antes <- t(beta.depois)

    cat( "Iteration",contador, "\n")
    contador <- contador + 1
    diag(W) <- 1/(Z%*%beta.depois)
  }
  return(t(beta.depois))
}

betasFinais <- metScoresPoi(beta0=beta0, x=dados[,1])

glm(temp.b ~ prec.b + co2.b + ozono.b, family = poisson(link = "identity"), data = dados, start = beta0, maxit = 300)


#Calcular os residuos
residualTemp <- c(NA, length(dados[,1]))
residualPrec <- c(NA, length(dados[,1]))
residualCO2 <- c(NA, length(dados[,1]))
residualOzono <- c(NA, length(dados[,1]))

for (i in 1:length(dados[,1])){
  residualTemp[i] <- dados[i,1] - Z[i,1]*betasFinais[1]
  residualPrec[i] <- dados[i,1] - Z[i,2]*betasFinais[2]
  residualCO2[i] <- dados[i,1] - Z[i,3]*betasFinais[3]
  residualOzono[i] <- dados[i,1] - Z[i,4]*betasFinais[4]
}

plot(residualPrec, main = "Resíduos do modelo relativos à variável prec.b", xlab = "período", ylab = "resíduos", col="black")
par(new=TRUE)
plot(dados[,1], axes=FALSE, ann=FALSE, pch=16, col = 2)
axis(4)

plot(residualCO2, main = "Resíduos do modelo relativos à variável co2.b", xlab = "período", ylab = "resíduos")
par(new=TRUE)
plot(dados[,1], axes=FALSE, ann=FALSE, pch=16, col = 2)
axis(4)

plot(residualOzono, main = "Resíduos do modelo relativos à variável ozono.b", xlab = "período", ylab = "resíduos")
par(new=TRUE)
plot(dados[,1], axes=FALSE, ann=FALSE, pch=16, col = 2)
axis(4)
```

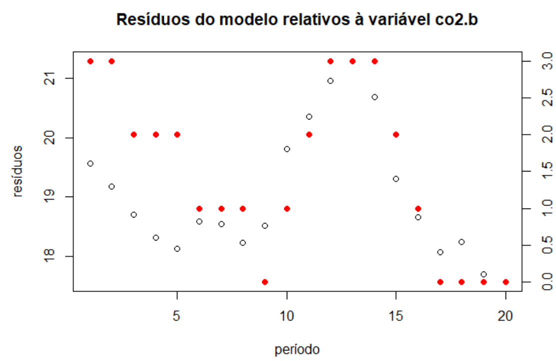
Análise

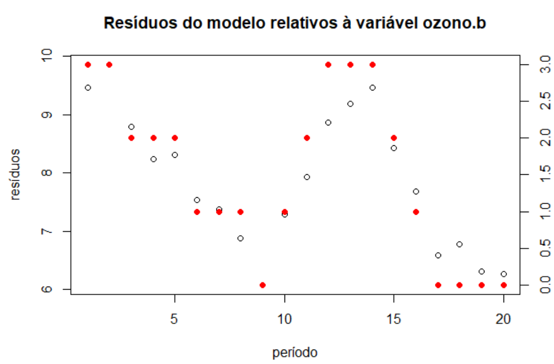
A fórmula usada para o calculo dos resíduos foi:

$$\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta}$$

Que nos permitiu obter os seguintes gráficos, que compara os residuos de cada covariável com a variável de resposta.







Observando os gráficos obtidos, podemos concluir que, comparando a variável de resposta Y (temperatura) com o cálculo dos resíduos para as restantes variáveis, a variável dióxido de carbono é a que, em termos médios se afasta mais dos valores de Y. Por outro lado, as variáveis ozono e precipitação, parecem ser semelhantes nesta comparação. No entanto, consideramos que a primeira destas duas se encontra mais próxima da variável de resposta e, por isso, podemos concluir que o ozono é a variável que mais afeta as baixas temperaturas num dado local.