Faculdade de Ciências e Tecnologias

Universidade NOVA de Lisboa
Estatística Numérica Computacional

PROJETO 3

Ana Breia - 61877
Gonçalo Santos - 55585
João Funenga - 61635
Mário Miranda - 62286

2021

Conteúdo

Introdução **1**

Família Exponencial **1**

Forma Canónica 1

Exercício 1 1

a) 2

b) 2

c) 2

Exercício 2 4

Modelos Lineares Generalizados **5**

IRWLS 6

Exercício 1 7

Referências **15**

Introdução

Família Exponencial

Seja $X \sim F(\theta)$, com $\theta \in \Theta$ desconhecido. Se a f.d.p. for da forma

$$f(x;\theta,\phi) = exp\{\frac{\eta(\theta)T(x) - A(\theta)}{b(\phi)} + c(x,\phi)\},$$

com ϕ um escalar, $A()$, $b()$ e $c()$ funções reais conhecidas, então a v.a. X diz-se pertencer à **família exponencial**.

Nestas condições,

1. $b(\phi)$ diz-se o parâmetro de incômodo (*nuisance*)/dispersão (geralmente conhecido);
2. $\eta(\theta)$ diz-se o parâmetro natural (quando $\eta(\theta) = \theta$ diz-se que a família exponencial se encontra na forma canónica);
3. $T(X)$ diz-se a estatística suficiente natural (T é uma estatística suficiente para $n = 1$ se não depende de θ);
4. $A(\theta)$ diz-se a constante *log-normalization* e assume-se que é duas vezes diferenciável.

Forma Canónica

Suponha-se que $\eta(\theta) = \theta$. Neste caso, como já foi referido, diz-se que a família exponencial se encontra na forma canónica e η diz-se o *canonical link*.

Assim, a p.d.f. simplifica-se para

$$f(x;\theta,\phi) = exp\{\frac{\theta T(x) - A(\theta)}{b(\phi)} + c(x,\phi)\}.$$

No caso de $\eta(\theta) \neq \theta$, podemos utilizar o *canonical link* por forma a chegar à forma canónica, isto é,

$$f(x;\eta,\phi) = exp\{\frac{\eta T(x) - A(\eta)}{b(\phi)} + c(x,\phi)\},$$

com $\eta = \eta(\theta)$.

Para este caso, temos as seguintes propriedades:

1. $E(T(X)) = A'(\eta)|_{\eta=\eta(\theta)}$ e $V(T(X)) = A''(\eta)|_{\eta=\eta(\theta)}\, b(\phi)$
2. A função *score*, para $n = 1$, é dada por
$$S(\theta) = S(\eta)|_{\eta=\eta(\theta)}\, \frac{d\eta}{d\theta}, \quad S(\eta) = \frac{T(x) - A'(\eta)}{b(\phi)}$$
3. o estimador ML de θ, para $n = 1$, é a solução de $S(\eta) = 0|_{\eta=\eta(\theta)}$
4. A informação de Fisher, para $n = 1$, é dada por
$$I(\theta) = I(\eta)|_{\eta=\eta(\theta)}\, (\frac{d\eta}{d\theta})^2, \quad I(\eta) = \frac{A''(\eta)}{b(\phi)}$$

Exercício 1

Let $X \sim Weibull(\alpha,\beta)$, with β known and α unknown, which has p.d.f.

$$f(x;\alpha,\beta) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1} e^{-(\frac{x}{\alpha})^{\beta}}, \quad \alpha,\beta > 0, \quad x \in \mathbb{R}^+$$

a)

Show that this distribution belongs to the exponential family.

$$\begin{aligned} f(x;\alpha,\beta) &= exp\{log(\frac{\beta}{\alpha})(\frac{x}{\alpha})^{\beta-1}) - (\frac{x}{\alpha})^{\beta}\} = \\ &= exp\{log(\frac{\beta}{\alpha}) + (\beta-1)log(\frac{x}{\alpha}) - (\frac{x}{\alpha})^{\beta}\} = \\ &= exp\{log(\frac{\beta}{\alpha}) + (\beta-1)(log(x) - log(\alpha)) - (\frac{x}{\alpha})^{\beta}\} = \\ &= exp\{log(\frac{\beta}{\alpha}) + (\beta-1)log(x) - (\beta-1)log(\alpha) - (\frac{x}{\alpha})^{\beta}\} = \\ &= exp\{-\alpha^{-\beta}x^{\beta} + log(\frac{\beta}{\alpha}) - (\beta-1)log(\alpha) + (\beta-1)log(x)\} \end{aligned}$$

$\implies \eta(\alpha) = -\alpha^{-\beta}$

$\implies T(X) = x^{\beta}$ (β é conhecido)

$\implies A(\alpha) = (\beta-1)log(\alpha) - log(\frac{\beta}{\alpha})$

$\implies b(\beta) = 1, \quad c(x;\beta) = (\beta-1)log(x).$

b)

Clearly identify the canonical link and the sufficient statistic. Do you already have the canonical form? If not, write it down.

Neste caso, o 'canonical link' é $\eta = -\alpha^{-\beta}$ e o 'sufficient statistic' é $T(X) = x^{\beta}$. Uma vez que $\eta(\theta) \neq \theta$, temos que a forma deduzida na alínea anterior ainda não se encontra na forma canónica. Assim, a forma canónica é dada por

$$f(x;\eta,\beta) = exp\{\eta x^{\beta} + log(-\frac{\beta}{\eta^{-\frac{1}{\beta}}}) - (\beta-1)log(-\eta^{-\frac{1}{\beta}}) + (\beta-1)log(x)\},$$

com $\eta = -\alpha^{-\beta}$.

c)

Use the canonical form to

i. compute $E(X^{\beta})$ and $V(X^{\beta})$.

$$E(X^{\beta}) = E(T(X)) = A'(\eta)|_{\eta=-\alpha^{-\beta}}$$

Através da forma canónica, temos que

$$\begin{aligned} A(\eta) &= (\beta-1)log(-\eta^{-\frac{1}{\beta}}) - log(-\frac{\beta}{\eta^{-\frac{1}{\beta}}}) \\ &= (\beta-1)log(-\eta^{-\frac{1}{\beta}}) - log(-\beta\eta^{\frac{1}{\beta}}), \end{aligned}$$

então

$$\begin{aligned} A'(\eta) &= (\beta-1)\frac{\frac{1}{\beta}\eta^{-\frac{1}{\beta}-1}}{-\eta^{-\frac{1}{\beta}}} - \frac{-\eta^{\frac{1}{\beta}-1}}{-\beta\eta^{\frac{1}{\beta}}} \\ &= -\frac{1}{\beta}(\beta-1)\eta^{-\frac{1}{\beta}-\frac{\beta}{\beta}+\frac{1}{\beta}} - \frac{1}{\beta}\eta^{\frac{1}{\beta}-\frac{\beta}{\beta}-\frac{1}{\beta}} \\ &= -\frac{1}{\beta}\eta^{-1}(\beta-1+1) \\ &= -\eta^{-1} \end{aligned}$$

Assim, $E(X^\beta) = A'(\eta)|_{\eta=-\alpha^{-\beta}} = \alpha^\beta$.

Agora, temos que

$$\begin{aligned} V(X^\beta) = V(T(X)) &= A''(\eta)|_{\eta=-\alpha^{-\beta}}\, b(\beta) \\ &= A''(\eta)|_{\eta=-\alpha^{-\beta}} \\ &= \eta^{-2}\big|_{\eta=-\alpha^{-\beta}} \\ &= -\alpha^{2\beta} \end{aligned}$$

ii. write the score function $S_n(\alpha)$ and see if it is possible to analytically derive the maximum likelihood estimator of α, α_{MLE}

Temos que

$$\begin{aligned} S(\alpha) &= S(\eta)|_{\eta=-\alpha^{-\beta}}\,\frac{d\eta}{d\alpha} \\ &= (x^\beta + \eta^{-1}\big|_{\eta=-\alpha^{-\beta}}) \times \beta\alpha^{-\beta-1} \\ &= (x^\beta - \alpha^\beta) \times \beta\alpha^{-\beta-1} \end{aligned}$$

E assim,

$$S_n(\alpha) = \beta\alpha^{-\beta-1}(\sum_{i=1}^{n} x_i^\beta - n\alpha^\beta).$$

O estimador ML de α é a solução de $S_n(\eta) = 0|_{\eta=-\alpha^{-\beta}}$, e temos que

$$S_n(\eta) = \sum_{i=1}^{n} x_i^\beta + n\eta^{-1}.$$

Então,

$$\begin{aligned} S_n(\eta) = 0 &\iff \sum_{i=1}^{n} x_i^\beta + n\eta^{-1}\Big|_{\eta=-\alpha^{-\beta}} = 0 \\ &\iff \sum_{i=1}^{n} x_i^\beta - n\alpha^\beta = 0 \\ &\iff \overline{y} - \alpha^\beta = 0 \\ &\implies \alpha = -\overline{y}^{\frac{1}{\beta}} \vee \alpha = \overline{y}^{\frac{1}{\beta}}, \end{aligned}$$

com $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad y_i = x_i^\beta, \quad i \in \{1, ...n\}$. Assim, não é possível determinar analiticamente α_{MLE}.

iii. compute the Fisher Information $I_n(\alpha)$

Para $n = 1$, temos que

$$\begin{aligned} I(\alpha) &= I(\eta)|_{\eta=-\alpha^{-\beta}} \times (\frac{d\eta}{d\alpha})^2 \\ &= \frac{A''(\eta)}{b(\beta)}\Big|_{\eta=-\alpha^{-\beta}} \times (\frac{d\eta}{d\alpha})^2 \\ &= -\alpha^{2\beta} \times (-\beta(\beta+1)\alpha^{-\beta-2}) \\ &= \beta(\beta+1)\alpha^{\beta-2} \end{aligned}$$

$$\implies I_n(\alpha) = nI(\alpha) = n\beta(\beta+1)\alpha^{\beta-2}$$

iv. report the asymptotic variance of the maximum likelihood estimator α_{MLE}

Exercício 2

Say some distribution depending on unknown parameters $(\alpha, \beta) \in R+ \times R+$ has p.d.f. such that

$$f(x;\alpha,\beta) = exp\Big\{\sum_{i=1}^{2} \eta_i(\alpha,\beta)T_i(x) - A(\alpha,\beta) + c(x)\Big\}$$

with

- $\eta(\alpha, beta) = (\eta_1(\alpha,\beta), \eta_2(\alpha,\beta)) = (\alpha, -\beta)$
- $(T_1(x), T_2(x)) = (log\ x, x)$
- $A(\alpha,\beta) = -\alpha\ log\ \beta + log\ \Gamma(\alpha)$
- $c(x) = log\ x$

Use the canonical form of the p.d.f. to compute $E(X)$, $V(X)$ and $I_n(\alpha,\beta)$. Report the asymptotic variance of the maximum likelihood estimator $(\alpha_{MLE}, \beta_{MLE})$.

Para calcular estes valores, precisamos primeiro de passar a função para a forma canónica.

Sabendo que $(\eta_1(\alpha,\beta) = \alpha$ e $\eta_2(\alpha,\beta) = -\beta$ a função fica:

$$f(x;\eta_1,\eta_2) = exp\Big\{\eta_1 log\ x + \eta_2\ x + \eta_1\ log(-\eta_2) - log\ \Gamma(\eta_1) - log\ x\Big\}$$

Agora vamos calcular $E(X)$:

$$\begin{aligned} E(X) = E(T_2(X)) &= \frac{\partial A(\eta)}{\partial\, \eta_2}\Big|_{\eta_2=-\beta} \\ &= \frac{\partial}{\partial\, \eta_2}\Big[\eta_1 log(-\eta_2) - log\,\Gamma(\eta_1)\Big] \\ &= -\frac{\eta_1}{\eta_2}, \end{aligned}$$

E assim chegamos a

$$-\frac{\eta_1}{\eta_2}\Big|_{\eta_2=-\beta,\eta_1=\alpha} = \frac{\alpha}{\beta}$$

Para $V(X)$ temos que $b(\phi) = 1$ tendo em conta que a p.d.f. é escrita da seguinte forma:

$$f(x;\theta,\phi) = exp\Big\{\frac{\sum_{j=1}^{k}\eta_j(\theta)T_j(x) - A(\theta)}{b(\phi)} + c(x,\phi)\Big\}$$

Logo, para calcular $V(X)$:

$$\begin{aligned} V(X) = V(T_2(X)) &= \frac{\partial^2 A(\eta)}{\partial^2 \eta_2}\Big|_{\eta_2=-\beta,} b(\phi) \\ &= \frac{\partial}{\partial \eta_2}\Big[-\frac{\eta_1}{\eta_2}\Big]\Big|_{\eta_2=-\beta,\eta_1=\alpha} \\ &= \frac{\eta_1}{\eta_2^2}\Big|_{\eta_2=-\beta,\eta_1=\alpha} \\ &= \frac{\alpha}{\beta^2} \end{aligned}$$

Logo, tendo em conta que $I(\eta) = \frac{A''(\eta)}{b(\phi)}$ temos que

$$I_n(\alpha,\beta) = nI(\alpha,\beta) = I(\eta)|_{\eta=\eta(\alpha,\beta)}\, b(\phi) = \frac{\alpha}{\beta^2}$$

E finalmente, para a variância assintótica do MLE (α,β) (seja $Var(X)$ a representação da variância assintótica):

$$Var(X) = \frac{1}{I_n(\alpha,\beta)} = \frac{\beta^2}{\alpha}$$

Modelos Lineares Generalizados

Este tipo de modelos consiste na variável de resposta Y_i ser escrita como uma combinação linear de um conjunto de p variáveis $x_{1i},...,X_{pi}$ mais um termo aleatório relativo ao erro ξ_i, ficando assim

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \xi_i, \quad i = 1,...,n$$

tal que ξ_i é independente e identicamente distribuido com $E(\xi_i) = 0$ e $V(\xi_i) = \sigma^2$. Dizemos que é generalizado porque assumimos que $p \geq 2$. Para podermos fazer inferência, também assumimos que ξ_i é independente e identicamente distribuido com uma $N(0,\sigma^2)$. Este modelo também ser forumado usando uma matriz no lugar das p variaveis ficando com

$$Y = X\beta + \xi$$

em que $X = (1, X_1, ..., X_p)$ corresponde à matriz de interceção mais as p covariáveis, $\beta = (\beta_0, ..., \beta_p)$ ao vetor de parâmetros fixos desconhecidos com p+1 parâmetros e $\xi \sim N(0,\sigma^2 I)$ ao termo representante do erro aleatório.

Como assumimos que $\xi \sim N(0,\sigma^2 I)$ então temos que $\mu = E(Y|X) = X\beta$ em que μ corresponde à média da população. Existem situações em que não é apropriado o uso de um modelo linear como o usado no projeto anterior, por exemplo quando o domínio da variável de resposta é binário ou à contagem de algo, ou quando a variância da variavel de resposta está dependente da média. Para combater este tipo de problemas temos então os modelos lineares generalizados que são uma extensão do modelo linear e que consideram que

- $g(\mu) = X\beta$, com g sendo uma função monótona e diferenciável, g é a função link tal que o inverso desta função $g^{-1}(\eta)$ exista e desta maneira podemos escrever o polinómio de Taylor de primeira ordem de g como
$$g(y) = g(\mu) + (y-\mu)g'(\mu).$$
- Que a distribuição da variável de resposta Y pertence à familia exponencial de um único parametro.

O modelo linear simples, que utilizámos no segundo projeto e assume uma distribuicao normal para os erros e para a resposta, também pode ser visto como um GLM só que com a função link como $g(\mu) = \mu$.

As funções link faladas dependem do tipo de resposta e do tipo de estudo que estamos a fazer. Por exemplo, neste trabalho iremos analisar o número de caso de AIDs sobre um determinado intervalo de tempo e por isso usaremos a função link usada com uma distribuição de Poisson (log) [1].

Embora a função link canónica seja derivada diretamente da função de densidade de probabilidade de uma dada familia de GLM, existem alguns modelos GLM que são usados com outras funções link, estas são chamadas de funções link não canónicas. Visto que funções link diferentes levam a interpretações diferentes das estimações dos parametros significa que estas devem ser escolhidas não por conveniência das simplificações algébricas mas sim do problema em si que estamos a resolver. Neste projeto, por consistir em contagem de dados iremos utilizar o Poisson Regression Model.

IRWLS

Os GLM's são normalmente estimados usando um algoritmo iterativo usando os quadrados mínimos, que ajusta os pesos (coeficientes) da expressão de modo a maximizar a verosimilhança (IRWLS).

Os parâmetros que nos interessam e queremos estimar são os $\beta = (\beta_0, ..., \beta_p)$ e estes serão estimados pela máxima verosimilhança. Relativamente ao parâmetro de dispersão (ϕ), este é estimado pelo método dos momentos. O critério de convergência deste algoritmo é baseado na mudança ou no desvio padrão ou na log-verosimilhança.

O método IRWLS é baseado no método de scoring de Fisher e pode ser implementado da seguinte forma

1. Inicializar a resposta esperada $\mu = E(Y|X)$ e a função link $\eta = g(\mu)$
2. Calcular os pesos
$$W^{-1} = Vg'(\mu)^2 = V\left(\frac{d\eta}{d\mu}\right)^2,$$
onde $g'(\mu)$ é a derivada da função link, V é a variância definida pela segunda derivada do CUMULATIVO ?? ($V = A''(\eta)$), ficando assim neste caso
$$W^{-1} = A''(\eta(\mu))\left(\frac{d\eta}{d\mu}\right)^2$$
3. Calcular a pseudodata, isto é, a linearização de um único termo da série de Taylor da função de log-verosimilhança com forma geral dada por
$$z = \eta + (y - \mu)g'(\mu)$$
4. Regredir ?? z sobre os estimadores $X_1, ..., X_p$ com os pesos W de modo a obter atualizações no vetor de parâmetros a serem estimados (β).
$$\beta_r = (X'WX)^{-}1X'Wz.$$
5. Calcular η (ou estimador linear $X\beta$) baseando-nos nas estimativas de regressão.
6. Calcular μ (ou $E(Y)$) como $g^{-1}(\eta)$
7. Calcular o desvio (normalmente como $-2l_{fitted}$) mas podemos utilizar outros critérios de paragem.
8. Iterar entre os passos 2 a 7 até a mudança no desvio entre duas iterações seja inferior a um determinado limiar de tolerância declarado e reportar a estimação da máxima verosimilhança de $\widehat{\beta}_{MLE}$

Este método de estimação pode ser utilizado para qualquer membro da família GLM.

Exercício 1

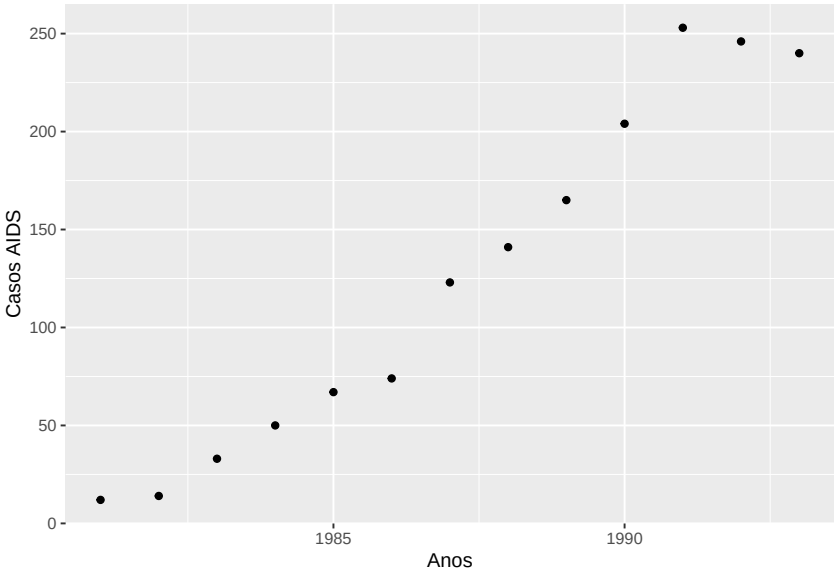
The following data refers to the number of new AIDs cases (y) each year (x) in Belgium, from 1981 to 1993 (Venables and Ripley, Modern Applied Statistics with S)

12 14 33 50 67 74 123 141 165 204 253 246 240

The question here is whether these data support evidence that the increase in the rate of new case generation is slowing down.

a) **Fit a Poisson regression model to the data** (Y, x)**, i.e., to** $Y \sim x$**. Report the model fit residual deviance and AIC, plot the residual plots for the fitted model and comment on model fit adequacy.**

Nesta pergunta temos como objetivo fazer o *fit* de modelo de regressão de *Poisson* aos nossos dados com a fórmula respetiva de Y x. Primeiramente, este exercício deve ser modelado usando a distribuição de *Poisson* pois estamos a lidar com contagens de acontecimentos num determinado período de tempo, sendo este indicado para tais casos. Começaremos por criar os vetores respetivos correspondentes tanto aos anos (período de análise), como o para os casos de *AIDS*. Para uma melhor visualização dos dados ao longo do tempo faremos então o plot dos dados.



Nesta pergunta, por os novos casos expectados mudarem em função do tempo, esta tem a forma de

$$\mu_i = \gamma^{\beta_1 t_i}$$

em que γ e β_1 são parâmetros desconhecidos. Caso façamos o logaritmo da expressão de modo a passá-la para a escala logarítmica temos que

$$log(\mu_i) = log(\gamma) + \beta_1 t_i$$

Como pedido, iremos agora fazer fit deste modelo usando a função do R, *glm* e analisar os valores resultantes da chamada do *summary* para ver como se adequa este aos nossos dados.

```
##
## Call:
## glm(formula = y ~ x, family = poisson(link = "log"), data = dataMatrix1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -4.6784  -1.5013  -0.2636   2.1760   2.7306
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.971e+02  1.546e+01  -25.68   <2e-16 ***
## x            2.021e-01  7.771e-03   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.206  on 12  degrees of freedom
## Residual deviance:  80.686  on 11  degrees of freedom
## AIC: 166.37
##
## Number of Fisher Scoring iterations: 4
```

Como podemos verificar pelo output acima, o *Intercept Estimate* diz-nos o valor estimado da variável de resposta quando as variáveis explicativas são 0 (neste caso, o ano). Temos também os coeficientes que descrevem o declive da relação que estabelecemos e a partir destes, conseguimos ver que este é positivo, ou seja, com o passar do tempo, os casos de *AIDS* têm tendência a subir, o que representa o contrário da hipótese postulada no enunciado.

Relativamente à *Residual Deviance* temos que esta apresenta um valor de 80.686 com 11 graus de liberdade enquanto que a *Null Deviance* tem 872.206 com 12 graus de liberdade.

Os desvios residuais representam a qualidade do *fit* de um determinado modelo. Caso um modelo tenha um bom *fit*, este valor será baixo, caso contrário será elevado. Estes medem a medida de variabilidade nas variáveis de resposta que não são explicadas pelo modelo proposto. Relativamente aos resíduos nulos, caso este valor seja baixo, os nossos dados conseguem ser bem modelados usando apenas o *Intercept*. Caso seja alto (o nosso caso), deveremos usar mais atributos para criarmos um modelo que se adapte bem aos nossos dados.

Outro aspeto que representa a qualidade do *fit* é a relação entre os desvios residuais e os graus de liberdade usados. Isto é, caso o valor dos desvios resíduais seja semelhante aos graus de liberdade, o nosso modelo está bem ajustado aos nossos dados. Ao analisarmos o nosso caso percebemos que temos um valor de 80.686 para a medida referida com 11 graus de liberdade logo estes valores divergem bastante, dizendo-nos assim que este modelo não é de todo adequado para os nossos dados. Isto pode ser resultado de sobre-dispersão em que a variância é maior que a prevista pelo modelo.

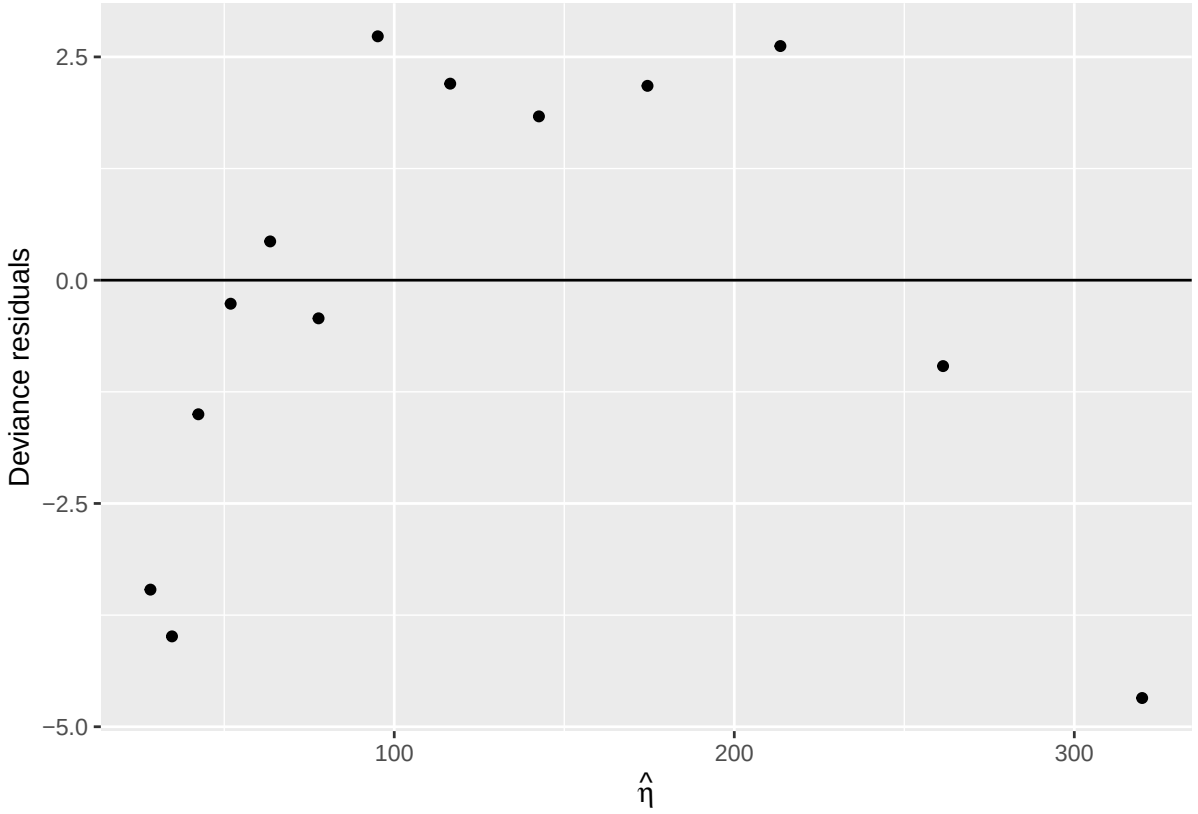
A métrica AIC (Akaike Information Criterion) descreve-nos a qualidade do modelo e é definida por

$$AIC = 2p - 2ln(\hat{L})$$

em que p é o número de parâmetros do modelo e $\hat{L}$ é o máximo da função de verosimilhança. Para a interpretação deste valor, é melhor fazer a comparação deste entre dois modelos visto que não tem grande importância ser analisado sozinho. No entanto, regra geral, um valor baixo é caracterizado por uma baixa complexidade (minimizar p, que representa diminuir o número de parâmetros usados) e um bom *fit* (maximizar o valor máximo da função de verosimilhança $\hat{L}$).

$$\textit{Residual Deviance} = 80.686$$
$$AIC = 166.37$$

Por último iremos fazer o *plot* dos resíduos sobre os valores *fitted*, sendo este um gráfico bastante importante na análise do ajustamento do modelo aos dados. Este gráfico serve para verificar se existe evidência de não-linearidade entre os resíduos e os valores *fitted*.



Para analisar o gráfico acima [2] temos de ter em conta que a distância à linha do 0 representa o quão má foi a estimativa para aquele valor, ou seja, quanto maior a distância, mais longe a estimativa foi do valor que deveria ter sido visto que os resíduos são dados por $Observdos - Previstos$. Valores acima da linha do 0 representam estimativas muito baixas e o contrário para estimativas muito altas. No entanto, caso as estimativas sejam acertadas, o ponto ficaria sobre a linha do 0. Pelo gráfico resultante, conseguimos perceber que a grande maioria das estimativas foram longe das acertadas ficando assim aquém de um modelo apropriado para os dados.

Inicialmente íamos utilizar o teste de*Kolmogorov-Smirnov* [3] ao invés da observação gráfica cujo objetivo é o mesmo mas infelizmente apenas pode ser utilizado para distribuições contínuas, o que não é o caso. Por este motivo decidimos utilizar o teste de *Hosmer and Lemeshow goodness of fit*. Este tem o mesmo propósito mas permite ser aplicado a distribuições discretas [5].

```
## Warning: package 'ResourceSelection' was built under R version 4.1.2
```

```
## ResourceSelection 0.3-5   2019-07-22
```

Hosmer and Lemeshow goodness of fit
X-squared = -0.99223, $df = 8$, *p-value = 1*

Como podemos observar, o *p-value* é superior a um $\alpha = 0.05$, pelo que não podemos rejeitar a hipótese de linearidade dos dados ao contrário do que pareceu pela análise gráfica.

b) **Fit a Poisson regression model again for the relationship $Y \sim x + x^2$. Report the model fit residual deviance and AIC, plot the residual plots for the fitted model and comment on model fit adequacy when compared to the previous model fit.**

Agora realizarmos a mesma análise que na pergunta anterior mas com uma diferença, alteraremos a fórmula relativa ao modelo utilizado para ajustarmos aos nossos dados. Anteriormente usámos $Y \sim x$ enquanto que agora será $Y \sim x + x^2$. À primeira vista, parece que este modelo se comportará melhor

porque estamos a realizar uma transformação das variáveis que temos disponíveis e, possivelmente, levará a que o modelo se adapte melhor aos dados. No caso anterior, o modelo apenas funcionaria bem caso os dados tivessem dispostos de uma forma estritamente linear, enquanto que agora, pela reta que fará *fit* aos dados ser uma parábola, esta adaptar-se-á melhor a dados que não estão organizados da forma referida. Agora, com este possivel melhoramento do modelo temos que a expressão que o caracteriza é

$$log(\mu_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$$

Faremos agora a seguinte alteração em R e observaremos o output gerado pela mesma função que anteriormente.

```
##
## Call:
## glm(formula = y ~ x + I(x^2), family = poisson(link = "log"),
##     data = dataMatrix1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45903  -0.64491   0.08927   0.67117   1.54596
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.478e+04  1.051e+04  -8.066 7.29e-16 ***
## x            8.509e+01  1.057e+01   8.048 8.45e-16 ***
## I(x^2)      -2.135e-02  2.659e-03  -8.029 9.82e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.2058  on 12  degrees of freedom
## Residual deviance:   9.2402  on 10  degrees of freedom
## AIC: 96.924
##
## Number of Fisher Scoring iterations: 4
```

Analisando o output acima, relativamente à *Residual Deviance* temos que esta apresenta um valor de 9.2402 com 10 graus de liberdade enquanto que a *Null Deviance* tem 872.206 com 12 graus de liberdade.

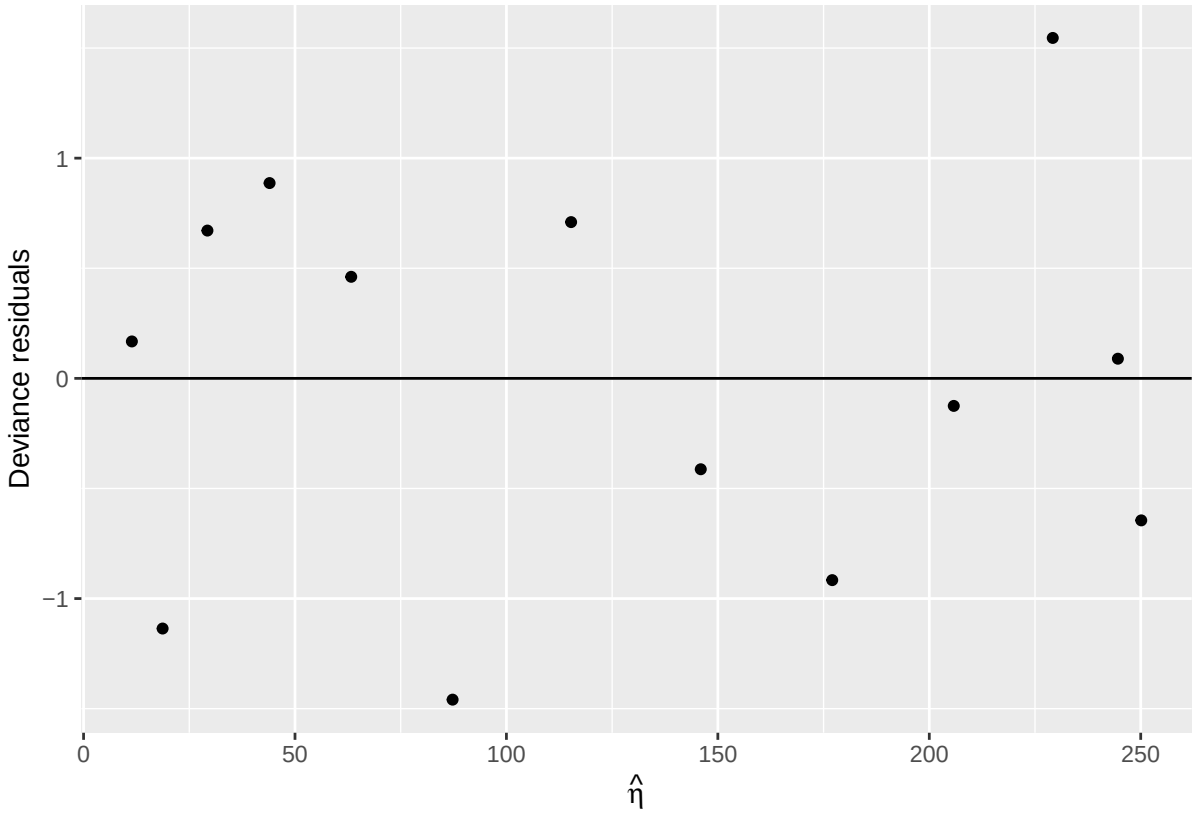
Como explicado na pergunta anterior, os desvios residuais representam a qualidade do *fit* de um determinado modelo. Caso um modelo tenha um bom *fit*, este valor será baixo, caso contrário será elevado. Neste caso, a diferença deste valor para este modelo comparativamente ao anterior é abismal o que nos diz que este se adaptou muito melhor aos dados.

Comentando sobre a relação entre os desvios residuais e os graus de liberdade usados, o valor dado pelos desvios resíduais foi aproximadamente igual ao dos graus de liberadade usados (um pouco inferior até). Este facto diz-nos que não houve de todo sobre-dispersão, ao contrário do primeiro modelo dando assim mais evidências que este modelo se comporta significativamente melhor que o anterior [4].

Para o valor de AIC, este foi de 96.924, significativamente mais baixo que os 166.37 do modelo anterior. Como referimos na última pergunta, este valor tem mais significado caso o comparemos entre dois modelos e quanto mais baixo, melhor o modelo é. No entanto, esta métrica também tem em conta e penaliza o valor relativamente a uma maior complexidade do modelo (número de parâmetros usados) e no modelo apresentado nesta pergunta, este é de facto mais complexo que o anterior. Esta penalização tem como objetivo prevenir de incluirmos parâmetros extra que não sejam relevantes ou que de pouco adicionem para o ajuste aos dados.

$$\textit{Residual Deviance} = 9.2402$$
$$AIC = 96.924$$

Por último faremos o *plot* dos resíduos sobre os valores *fitted*.



Tendo em conta a explicação anterior de como analisar este gráfico, percebemos que os valores se encontram todos muito mais próximos da linha do 0, representando assim muito melhores estimativas sobre os valores previstos tendo em conta a realidade. Conseguimos perceber imediatamente que a escala do gráfico (no Y) é muito mais reduzida e que os valores se encontram homogeneamente dispersos em torno do 0 havendo até vários que praticamente coincidem com a linha, falhando apenas por um pequeno desvio. Comparativamente ao anterior, neste as estimativas que mais se afastam da realidade encontram-se afastadas por cerca de 1.5 unidades comparativamente ao real enquanto que no anterior existiam exemplos a 4 unidades de distância.

Como para o primeiro modelo, o resultado do teste de *Hosmer-Lemeshow* revelou também que o *p-value* se encontra acima de um $\alpha = 0.05$ pelo que também verificamos o ajuste do modelo aos dados reais.

Hosmer and Lemeshow goodness of fit
X-squared = -0.098135, *df* = 8, *p-value = 1*

Tendo estes fatores em conta, o segundo modelo aparenta ajustar-se muito melhor aos dados que o primeiro sendo assim uma melhor escolha para realizar previsões acerca dos casos de *AIDs* na Bélgica.

c) **One now wishes to compare the previous two models by means of an analysis of variance table. Describe model selection via the ANOVA table. Use the R built-in anova() function to compare both models. Which model better fits the data?**

Para compararmos os dois modelos abordados, podemos recorrer a uma tabela de análise de variância que relaciona ambos [6]. A função *anova()* recebe como parâmetros dois modelos e retorna uma tabela comparando os dois e que, a partir desta, podemos verificar se de facto o modelo mais complexo (alínea b)) é significativamente melhor a ajustar-se aos dados comparativamente ao mais simples (alíena a)). Caso o p-value enunciado na tabela seja abaixo de um determinado limiar que nós assumimos (podemos considerar $\alpha = 0.01$), podemos concluir que de facto o modelo mais complexo é significativamente melhor que o outro e assim, é o que devemos utilizar mesmo tendo em conta a complexidade acrescentada. Caso

o p-value nao seja suficientemente baixo (abaixo do limiar que falámos anteriormente), não podemos assumir que a complexidade adicional valha a pena e devemos optar pelo mais simples [7].

```
anovaTest = anova(fit1,fit2, test= "Chisq"); anovaTest
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + I(x^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        11     80.686
## 2        10      9.240  1   71.446 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anovaTest)
```

```
##    Resid. Df       Resid. Dev          Df       Deviance        Pr(>Chi)
##  Min.   :10.00   Min.   : 9.24   Min.   :1   Min.   :71.45   Min.   :0
##  1st Qu.:10.25   1st Qu.:27.10   1st Qu.:1   1st Qu.:71.45   1st Qu.:0
##  Median :10.50   Median :44.96   Median :1   Median :71.45   Median :0
##  Mean   :10.50   Mean   :44.96   Mean   :1   Mean   :71.45   Mean   :0
##  3rd Qu.:10.75   3rd Qu.:62.82   3rd Qu.:1   3rd Qu.:71.45   3rd Qu.:0
##  Max.   :11.00   Max.   :80.69   Max.   :1   Max.   :71.45   Max.   :0
##                                  NA's   :1   NA's   :1       NA's   :1
```

Como podemos ver pelo output, o segundo modelo (mais complexo) tem mais um grau de liberdade que o primeiro (indicando assim que tem um parâmetro adicional) e um p-value muito baixo, inferior ao α que considerámos ($pvalue < alpha$). Isto diz-nos que ao adicionarmos o parâmetro extra, mesmo aumentando a complexidade do modelo, esta mudança revelou-se útil e significativamente melhor. Para além deste factor, também percebemos que a redução de varância do primeiro para o segundo modelo é drástica (redução de 71.446) o que representa que as estimativas foram muito mais próximas dos valores reais. Depois da análise desta tabela, podemos concluir que o segundo modelo tem um melhor fit aos nossos dados e assim deveríamos escolhê-lo em detrimento do primeiro.

d) **Provide model summary, confidence intervals for the fixed parameters and model interpretation for the selected model. Plot the data. Predict 100 values from the fitted model (use the R built-in predict() function). Plot the data versus the fitted line. Add confidence bands for the fitted line at ±2se. Make an attempt at answering the main question.**

Pela análise da pergunta anterior, concluímos que o melhor modelo que se ajustou aos casos de AIDS por ano na Bélgica foi o segundo.

```
##
## Call:
## glm(formula = cases ~ year + I(year^2), family = poisson(link = log),
##     data = belg.aids)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45903  -0.64491   0.08927   0.67117   1.54596
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.901459   0.186877  10.175  < 2e-16 ***
```

```
## year         0.556003   0.045780  12.145  < 2e-16 ***
## I(year^2)   -0.021346   0.002659  -8.029 9.82e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.2058  on 12  degrees of freedom
## Residual deviance:   9.2402  on 10  degrees of freedom
## AIC: 96.924
##
## Number of Fisher Scoring iterations: 4
```

Este é caracterizado pela seguinte equação

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$$

em que substituindo pelos coeficientes e interceção com o eixo das abcissas dado pela função *summary()* ficamos com

$$Y_i = 1.901 + 0.556 t_i - 0.021 t_i^2$$

Para os parâmetros em específico substituídos na expressão, iremos também calcular os seus intervalos de confiança. Para isto recorreremos à função *confint()*.

```
confint(fit2Aux)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept)  1.5255558  2.25840743
## year         0.4678224  0.64734015
## I(year^2)   -0.0266260 -0.01620133
```

Esta função calcula os intervalos de confiança para $\alpha = 0.05$ para todos os parâmetros que constituem o nosso modelo, embora os valores das colunas que aparecem no output sejam diferentes. Um intervalo de 95% de confiança consiste de dois pontos extremos. Poderíamos calcular a limiar minimo de 1% que representaria que ao repetir a experiência muitas vezes, o valor verdadeiro será inferior a este limiar em apenas 1% das vezes e o contrário para o limite superior (valor verdadeiro ser superior a este limite em 4% das vezes. No entanto, existe simetria logo o intervalo de confiança "default" consiste num limiar inferior de 2.5% e um superior de 97.5% (100-2.5). Aqui, caso a experiência seja repetida muitas vezes, o valor verdadeiro em 2.5% das vezes será inferior e outras 2.5% superior ao do intervalo de confiança de 95%.

Após verificarmos o output conseguimos dizer com 95% de confiança que os valores dos seguintes parâmetros se encontram entre os respetivos valores.

$$\beta_0 \text{ (Intercept)} = [1.526, 2.258]$$

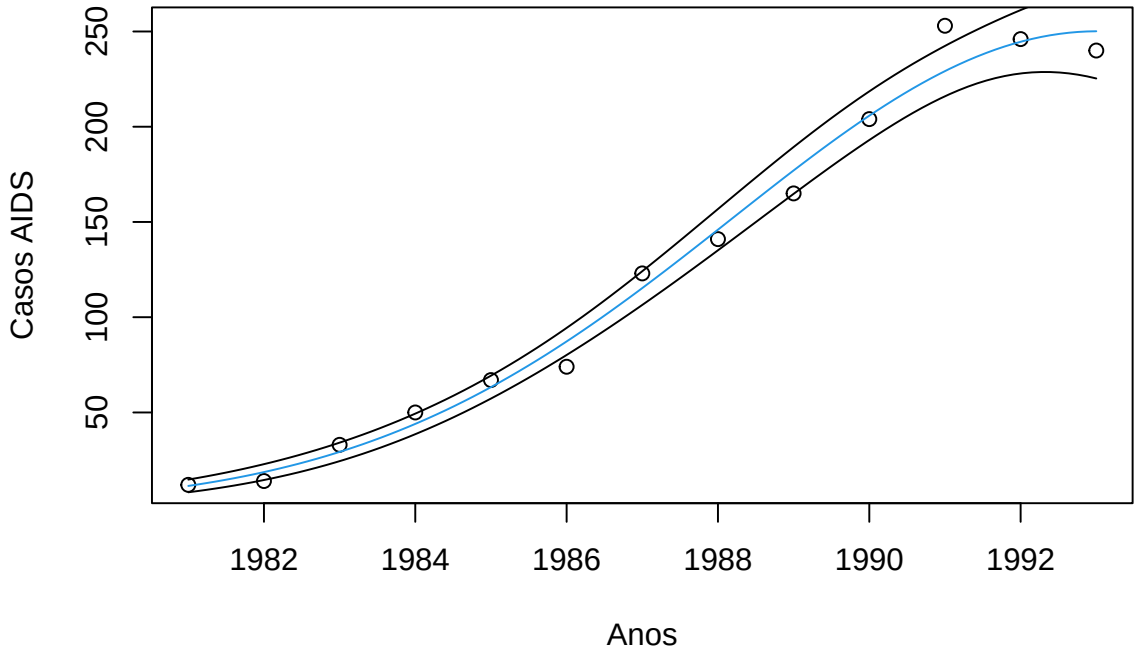
$$year = [0.468, 0.647]$$

$$year^2 = [\text{-}0.0266, \text{-}0.0162]$$

De seguida, queremos prever 100 valores a partir do modelo a que chegámos. Isto é, com o modelo linear generalizado analisado nas últimas perguntas, iremos prever o número de casos de AIDS para 100 observações ao longo dos anos. Para além do plot dos dados previstos, também mostraremos a correspondente reta ajustada ao modelo bem como duas outras cuja área entre elas representa o intervalo para qual o limite superior e inferior vêm da seguinte expressão: $+-2se$. Isto diz-nos que caso os valores

previstos se encontrem dentro destas margens, o valor previsto nunca se afastou mais do que duas vezes o desvio padrão, que por sua vez será representativo da qualidade das previsões. Caso os valores previstos não se encontrem dentro deste intervalo, conseguimos perceber que para esses pontos, o modelo não teve capacidade de os prever. Este tipo de situações pode ocorrer caso se tente prever mais pontos fora deste intervalo de tempo.

Tal acontece porque, por este seguir uma função de segundo grau, esta pode-se adaptar demasiado aos dados e não ter grande capacidade de generalização devido às oscilações da função. Isto poderia ser mitigado com algum tipo de regularização, no entanto, este tema não foi abordado nesta unidade curricular.



Como podemos verificar, a grande maioria dos pontos que foram previstos encontra-se dentro das margens faladas acima, embora exista 1 ponto em que isto não acontece. Voltando à questão principal do problema, sendo esta "Será que a taxa de aumento em casos de AIDS está a diminuir?", por análise do gráfico e conhecermos o comportamento da função, embora esta estar a atingir um máximo no final do ano de 1993, percebemos facilmente que o declive da *fitted line* está a diminuir drasticamente ao longo dos anos. Com esta análise, conseguimos concluir que o aparecimento do número de casos em 1993, está a aumentar a uma taxa muito mais lenta que no final da década de 80.

Referências

[1] Jabeen, H. (2019, 27 de Fevereiro). *Poisson Regression in R.*
Acedido a 7/1/2022, em https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/

[2] Qualtrics *Interpreting Residual Plots to Improve Your Regression.*
Acedido a 7/1/2022, em https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

[3] Engineering Statistics Handbook (2003, 6 de Janeiro) *Kolmogorov- Smirnov test .*
Acedido a 7/1/2022, em https://www.itl.nist.gov/div898/handbook/prc/section2/prc212.htm

[4] Lillis, D. (2017, 3 de Agosto) *Generalized Linear Models in R.*
Acedido a 7/1/2022, em https://www.theanalysisfactor.com/r-glm-model-fit/

[5] Stephanie (2016, 28 de Agosto) *Hosmer-Lemeshow Test: Definition.*
Acedido a 7/1/2022, em https://www.statisticshowto.com/hosmer-lemeshow-test/

[6] Carey, G. *Theory: The General Linear Model .*
Acedido a 7/1/2022, em http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/GLM%20Theory.pdf

[7] Nathaniel D. Phillips (2018, 22 de Janeiro) *Comparing regression models with anova() .*
Acedido a 7/1/2022, em https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html