Lista 1

Prof. Erica Castilho Rodrigues

Disciplina: Modelos Lineares Generalizados

12 de Maio

Data de entrega: 3 de Junho

1. Para um modelo de regress ao simples,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

onde $\epsilon_i \sim^{iid} N(0, \sigma^2)$. Encontre:

a) $E(Y_i|X_i)$

b) $Var(Y_i|X_i)$

c) $Cov(Y_i, Y_j)$

2. Determine se as seguintes afirmativas são verdadeiras e explique o motivo.

a) Quando pedido para escrever um modelo de regressão linear simples, um estudante escreveu

$$E(Yi) = \beta_0 + \beta_1 X_i + \epsilon_i \ .$$

b) Em um modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

temos que β_0 e β_1 são os únicos parâmetros desconhecidos.

3. Uma análise foi feita para estudar a relação entre a resposta Y e variável exploratória X. Selecionou-se uma amostra de tamanho 10. O modelo de regressão linear

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

onde $\epsilon_i \sim^{iid} N(0, \sigma^2)$ foi ajustado. Os valores dos resíduos do modelo encontram-se na Tabela 1. Encontre os valores correspondentes às caselas marcadas com $*$.

Index	x	y	Resíduo
1	0.0	0.9	-3.46
2	4.1	5.9	-3.82
3	5.1	6.2	-4.83
4	6.1	8.7	-3.63
5	7.1	9.1	-4.54
6	0.0	0.2	-4.16
7	4.8	21.0	10.37
8	3.8	16.7	7.37
9	2.8	12.4	4.38
10	$*$	9.3	$*$

Tabela 1: Tabela referente ao exercício 03.

4. Verifique se as seguintes distribuições pertencem à família exponencial.

   a) $X|\theta \sim U(0, \theta)$ para $\theta > 0$.

   b) $X|\theta \sim \exp(\theta)$ para $\theta > 0$.

   c) $X|\sigma^2 \sim N(\theta, \sigma^2)$ para $\sigma^2 > 0$ (observe que nesse caso θ é conhecido, o único parâmetro desconhecido é σ^2).

5. Para cada um dos exemplos a seguir, identifique qual é a variável resposta e quais são as variáveis explicativas. Escolha qual distribuição mais adequada para variável resposta e justifique sua escolha. Escreva uma função de ligação razoável e especifique como fica a relação entre a média da variável resposta e o preditor linear.

   a) Queremos verificar o efeito da idade, sexo, altura, a média de ingestão diária de alimentos e média diária de gasto de energia no peso de uma pessoa.

   b) Deseja-se verificar as proporções de ratinhos de laboratório, que se tornaram infectados após exposição a bactérias quando cinco níveis de exposição diferentes são usados e 20 ratos são expostos a cada nível.

   c) Desja-se analisar a relação entre o número de idas por semana ao supermercado de uma dona de casa e o número de pessoas na família, a renda familiar e da distância ao supermercado.

6. Considere N variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_n$ tal que

$$P(Y_i = 1) = \pi \quad P(Y_i = 0) = 1 - \pi_i \,.$$

A função de verossimilhança pode ser scrita como

$$f(y_i, \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

onde $y_i = 0$ ou $y_i = 1$.

a) Mostre que essa distribuição de probabilidade pertence à família exponencial.

b) Mostre que o parâmetro canônico é dado por

$$\log\left(\frac{p_i}{1-p_i}\right)$$

c) Utilizando os resultados vistos para família exponencial, mostre que

$$E(Y_i) = \pi_i \ .$$

d) Qual a função de ligação canônica?

7. Suponha que $Y_1, Y_2, \ldots, Y_N$ são variáveis aleatórias independentes com distribuição normal tais que:

$$E(Y_i) = \mu_i = \beta_0 + \log(\beta_1 + \beta_2 x_i) \quad Y_i \sim N(\mu_i, \sigma^2).$$

Esse é um Modelo Linear Generalizado? Justifique sua resposta e encontre a função de ligação.

8. (Este exercício deverá ser entregue separadamente e irá valer 3 pontos na nota final. Alguns alunos serão sorteados para explicar o código para o ajuste do modelo para a turma.) A Tabela 2 mostra o número de casos de AIDS na Austrália de acordo com a data do diagnóstico (ano e trimestre).

a) Faça um gráfico do número de casos y_i em função do período i ($i = 1, 2, \ldots, 20$).

b) Um modelo possível é o modelo Poisson(λ_i) com $\lambda_i = i^\theta$, ou seja

$$Y_i \sim Poisson(\lambda_i) \quad \lambda_i = i^\theta \quad \log(\lambda_i) = \theta \log(i) \ .$$

Faça um gráfico de $\log(y_i)$ e em função de $\log(i)$ e verifique se o modelo parece adequado.

c) Ajuste um Modelo Linear Generalizado para esse conjunto de dados usando a distribuição de poisson e a função de ligação

$$\log(\lambda_i) = \beta_1 + \beta_2 x_i$$

onde $x_i = \log(i)$. O ajuste deve ser feito utilizando o algoritmo de estimação visto em sala de aula. Primeiramente encontre a matriz $\mathbf{W}$ e os outros termos

	Trimestre			
Ano	1	2 3		4
1984	1	6	16	23
1985	27	39	31	30
1986	43	51	63	70
1987	88	97	91	104
1988	110	113	149	159

Tabela 2: Número de casos de AIDS na Austrália.

necessários para obter a equação de interação

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}z.$$

Utilize o software R para fazer os cálculos matriciais necessários.

d) Ajuste o modelo utilizando o R, utilizando diretamente a função *glm*. Compare os resultados obtidos com os anteriores.