Apresentação Prova 2 Regressão

Moisés Sales

Abril 2025

Primeira Questão

Considere uma aplicação de regressão logística em análise de sobrevivência. Seja $\pi_i(t)$ a probabilidade de um equipamento do tipo i falhar no intervalo $I_t = (t-1, t]$ dado que o mesmo não falhou até o tempo $t-1$. Seja Y_{it} o número de falhas no intervalo I_t e seja n_{it} o número de equipamentos que não falharam até o tempo $t-1$ no i-ésimo grupo.

Primeira Questão

Assumir que $Y_{it} \sim \text{B}(n_{it}, \pi_i(t))$, e que as falhas são independentes. Ajustar um modelo logístico do tipo

$$\log\left(\frac{\pi_i(t)}{1-\pi_i(t)}\right) = \alpha + \beta_i t + \gamma_i t^2$$

Primeira Questão

Assumir que $Y_{it} \sim \text{B}(n_{it}, \pi_i(t))$, e que as falhas são independentes. Ajustar um modelo logístico do tipo

$$\log\left(\frac{\pi_i(t)}{1-\pi_i(t)}\right) = \alpha + \beta_i t + \gamma_i t^2$$

Apresente o gráfico com as curvas ajustadas e os valores observados. Tente selecionar um submodelo apropriado. Verifique a adequação do modelo adotado através de gráficos de resíduos. Interprete os resultados.

Primeira Questão

Os dados estão apresentados da seguinte forma:

	tempo	tipo	n	y
1	1	1	42	4
2	2	1	38	3
3	3	1	35	3
4	4	1	31	5
5	5	1	26	6
6	1	2	50	6

Primeira Questão

Ajustaremos um modelo logístico da seguinte forma:

$$\log\left(\frac{\pi_i(t)}{1-\pi_i(t)}\right) = \alpha + \beta_i t$$

com $i = 1, 2, 3$ e $t = 1, 2, 3, 4, 5$, referente ao tipo de equipamento A, B e C, respectivamente.

Primeira Questão

Temos, então

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.2656	0.3893	-8.39	0.0000
tipo2	1.1326	0.3083	3.67	0.0002
tipo3	1.6358	0.3172	5.16	0.0000
tempo	0.4201	0.0915	4.59	0.0000
Desvio	2.9159	11 g.l.		

Em que podemos notar um valor de desvio que indica um ajuste adequado, realizando uma análise de diagnóstico, temos:

Primeira Questão

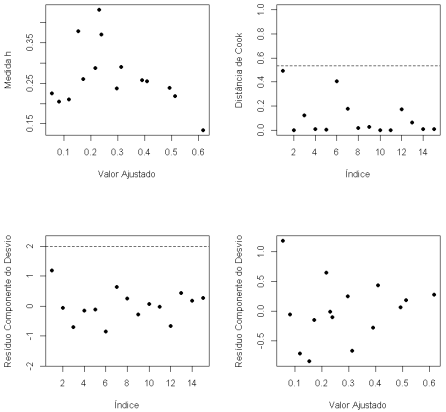

Figura: Gráficos de diagnósticos referentes a regressão logística.

Primeira Questão

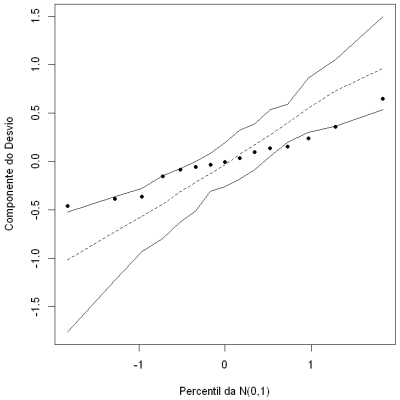

Figura: Gráfico normal de probabilidade referente a regressão logística.

Primeira Questão

A probabilidade de um produto do tipo A ($i = 1$) falhar no tempo 3 é dada por:

$$\log\left(\frac{\hat{\pi}_1(1)}{1-\hat{\pi}_1(1)}\right) = -3.2656 + 0.4201$$

$$\frac{\hat{\pi}_1(1)}{1-\hat{\pi}_1(1)} = \exp(-3.2656 + 0.4201)$$

$$\hat{\pi}_1(1) = \frac{\exp(-2.8455)}{1+\exp(-2.8455)}$$

$$\hat{\pi}_1(1) = 0.054 = 5.4\%$$

que é a menor probabilidade entre todos os tipos de materiais e todos os tempos.

Primeira Questão

Já a probabilidade de um equipamento do tipo A quebrar no tempo 5, é dada por:

$$\hat{\pi}_1(5) = 0.237 = 23.7\%$$

E o material que tem a maior probabilidade de quebrar é o material do tipo C no tempo 5, com

$$\hat{\pi}_3(5) = 0.615 = 61.5\%$$

Primeira Questão

A razão de chances entre o material do tipo A no tempo 3 e no tempo 1 é dada por:

$$\frac{\exp(-3.2656 + (0.4201 * 3))}{\exp(-3.2656 + 0.4201)} = 2.31683$$

ou seja, a probabilidade de um equipamento do tipo A quebrar no tempo 3 é 2.31 vezes maior, comparada a probabilidade de quebrar no tempo 1.

Segunda Questão

No arquivo são descritos os resultados de um estudo desenvolvido em 1990 com recrutas americanos referentes a associação entre o número de infecções de ouvido e alguns fatores. Os dados são apresentados na seguinte ordem: hábito de nadar (ocasional ou frequente), local onde costuma nadar (piscina ou praia), faixa etária ($15-19, 20-25$ ou $25-29$), sexo (masculino ou feminino) e número de infecções de ouvido diagnosticadas pelo próprio recruta.

Segunda Questão

	nadar	local	idade	sexo	num_infec
1	Occas	NonBeach	15-19	Male	0
2	Occas	NonBeach	15-19	Male	0
3	Occas	NonBeach	15-19	Male	0
4	Occas	NonBeach	15-19	Male	0
5	Occas	NonBeach	15-19	Male	0
6	Occas	NonBeach	15-19	Male	0

Segunda Questão

Verifique qual dos modelos, log-linear de Poisson, quase-verossimilhança ou log-linear binomial negativo, se ajusta melhor aos dados. Utilize métodos de diagnóstico como critério.

Segunda Questão

Denotamos por Y_{ijkl} a quantidade de infecções de ouvido do l-ésimo recruta, que possui o i-ésimo hábito de nadar, no j-ésimo local e está na k-ésima faixa etária. Supondo que $Y_{ijkl} \sim \text{P}(\mu_{ijkl})$, o modelo utilizado possui parte sistemática dada por:

Segunda Questão

Denotamos por Y_{ijkl} a quantidade de infecções de ouvido do l-ésimo recruta, que possui o i-ésimo hábito de nadar, no j-ésimo local e está na k-ésima faixa etária. Supondo que $Y_{ijkl} \sim \text{P}(\mu_{ijkl})$, o modelo utilizado possui parte sistemática dada por:

$$\log \mu_{ijkl} = \alpha + \beta_i \text{nadar}_i + \theta_j \text{local}_j + \gamma_k \text{faixa}_k$$

com as restrições $\beta_1 = \theta_1 = \gamma_1 = 0$, para $i = 1, 2$; $j = 1, 2$ e $k = 1, 2, 3$.

Segunda Questão

Os níveis $20-25$ e $25-29$, da variável faixa etária, foram unidos, visto que o nível $25-29$ foi o único nível não significativo para o modelo.

Segunda Questão

Os níveis $20-25$ e $25-29$, da variável faixa etária, foram unidos, visto que o nível $25-29$ foi o único nível não significativo para o modelo. O seguinte modelo resulta em:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1699	0.1157	-1.47	0.1422
nadarOccas	0.6136	0.1050	5.84	0.0000
localNonBeach	0.4982	0.1029	4.84	0.0000
idade20-29	-0.2742	0.1011	-2.71	0.0067
Desvio	757.23	283 g.l.		

Segunda Questão

Os níveis $20-25$ e $25-29$, da variável faixa etária, foram unidos, visto que o nível $25-29$ foi o único nível não significativo para o modelo. O seguinte modelo resulta em:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1699	0.1157	-1.47	0.1422
nadarOccas	0.6136	0.1050	5.84	0.0000
localNonBeach	0.4982	0.1029	4.84	0.0000
idade20-29	-0.2742	0.1011	-2.71	0.0067
Desvio	757.23	283 g.l.		

Podemos notar um alto valor do desvio, indicando que o modelo não está adequadamente ajustado, sendo provável a existência de sobredispersão.

Segunda Questão

Realizando uma análise de diagnóstico, temos:

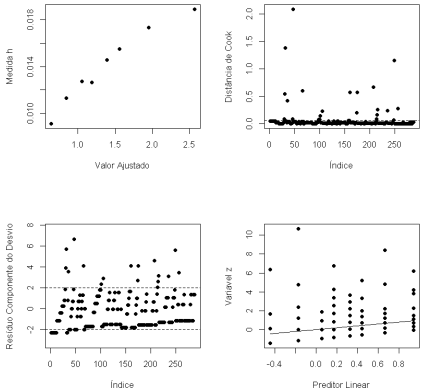

Figura: Gráficos de diagnóstico referentes ao modelo log-linear.

Segunda Questão

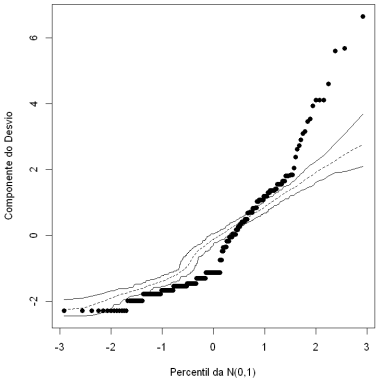

Figura: Gráfico normal de probabilidade referente ao modelo log-linear.

Segunda Questão

Utilizando, agora, o método de quase-verossimilhança, calculamos

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} / (n - p)$$

para obtermos a estimativa do parâmetro de dispersão, que, nesse caso, é dada por:

$$\hat{\sigma}^2 = 3.348$$

que é um valor maior que um, mais uma vez indicando uma sobredispersão.

Segunda Questão

Dessa forma, corrigimos o desvio do modelo, que agora é dado por:

$$\mathrm{D}^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \frac{757.23}{3.348} = 226.173 \quad \text{com } 283 \text{ g.l.}$$

indicando um ajuste mais adequado.

Segunda Questão

Dessa forma, corrigimos o desvio do modelo, que agora é dado por:

$$\mathrm{D}^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \frac{757.23}{3.348} = 226.173 \quad \text{com } 283 \text{ g.l.}$$

indicando um ajuste mais adequado. O resíduo componente desvio corrigido é dado por:

$$t^*_{D_i} = \pm d_i/\hat{\sigma}\sqrt{1-\hat{h}_{ii}}$$

Segunda Questão

As novas estimativas corrigidas são dadas por:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1699	0.2118	-0.80	0.4232
nadarOccas	0.6136	0.1921	3.19	0.0016
localNonBeach	0.4982	0.1883	2.65	0.0086
idade20-29	-0.2742	0.1849	-1.48	0.1392
Desvio	226.173	283 g.l.		

Segunda Questão

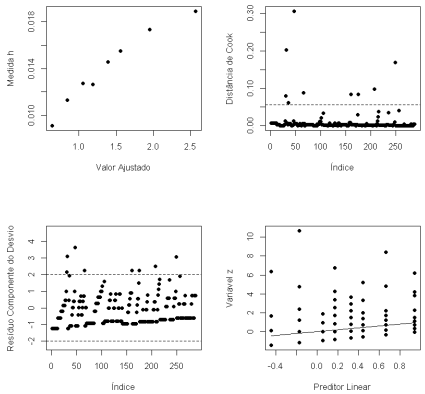

Figura: Gráficos de diagnóstico referente ao modelo quase-verossimilhança.

Segunda Questão

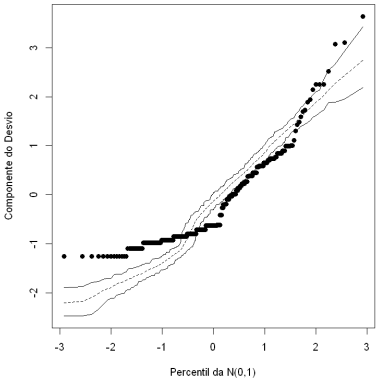

Figura: Gráfico normal de probabilidade referente ao modelo quase-verossimilhança.

Segunda Questão

Supondo, agora, que $Y_{ijkl} \sim \text{BN}(\mu_{ijkl}; \phi)$, cuja parte sistemática do modelo é a mesma, temos:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1322	0.1975	-0.67	0.5033
nadarOccas	0.6117	0.1898	3.22	0.0013
localNonBeach	0.4838	0.1893	2.56	0.0106
idade20-29	-0.3349	0.1890	-1.77	0.0764
ϕ	0.572			
Desvio	269.10	283 g.l.		

Segunda Questão

O desvio do modelo indica que é um ajuste adequado, e a análise de diagnóstico resulta em:

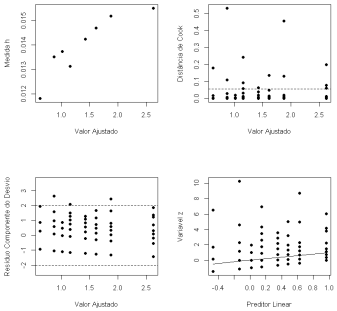

Figura: Gráficos de diagnóstico para o modelo log-linear binomial negativo.

Segunda Questão

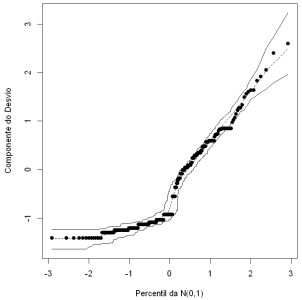

Figura: Gráfico normal de probabilidade par ao modelo log-linear binomial negativo.

Segunda Questão

Notamos que com esse modelo, todos os pontos estão contidos no envelope de confiança da figura 8. Logo, em comparação com os outros apresentados, este é o modelo que parece mais se ajustar aos dados, portanto é o escolhido.

$$\log \mu_{ijkl} = -0.1322 + \beta_i \text{nadar}_i + \theta_j \text{local}_j + \gamma_k \text{faixa}_k$$