Segunda avaliação (A2)

Disciplina: Modelagem Estatística
Instrutor: Luiz Max Carvalho
Monitor: Ezequiel Braga

30 de Junho de 2025

- O tempo para realização da prova é de 3 horas;
- Leia a prova toda com calma antes de começar a responder;
- Responda todas as questões sucintamente;
- Marque a resposta final claramente com um quadrado, círculo ou figura geométrica de sua preferência;
- A prova vale 100 pontos. A pontuação restante é contada como bônus;
- Apenas tente resolver a questão bônus quando tiver resolvido todo o resto;
- Você tem direito a trazer **<u>uma</u> folha de "cola"** tamanho A4 frente e verso, que deverá ser entregue junto com as respostas da prova.

1. You complement me.

Suponha que $Z_i \sim \text{Poisson}(\mu_i)$ para $i = 1, 2, \ldots, n$ são amostras independentes. Suponha ainda que $E[Z_i] = \mu_i = \exp(\boldsymbol{X}_i\boldsymbol{\beta})$. Sejam $Y_i = \mathbb{I}(Z_i > 0)$ e $\theta_i = \Pr(Y_i = 1)$.

a) (5 pontos) Mostre que $Y_i \sim \text{Bernoulli}(\theta_i)$;

b) (10 pontos) Considerando **este modelo**, exiba a função de ligação g que satisfaz $g(\theta_i) = \boldsymbol{X}_i\boldsymbol{\beta}$.

c) (15 pontos) Agora considere um modelo com a seguinte forma:

$$W_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i,$$

onde os erros ϵ_i são i.i.d com densidade comum

$$f_R(\epsilon) = \frac{\exp(-\epsilon)}{[1 + \exp(-\epsilon)]^2}$$

Suponha ainda que observamos apenas Y_i, que é uma variável aleatória definida como

$$Y_i = \begin{cases} 1, & \text{se } W_i > 0, \\ 0, & \text{caso contrário.} \end{cases}$$

Exiba a função de ligação e comente sobre a interpretação dos coeficientes $\boldsymbol{\beta}$ neste caso.

Conceitos trabalhados: modelos lineares generalizados; máxima verossimilhança; função *score*. **Nível de dificuldade**: fácil.

Resolução: Note que

$$Y_i = \begin{cases} 1, & \text{se } Z_i > 0; \\ 0, & \text{c.c.} \end{cases}$$

Logo, $\Pr(Y_i = 1) = \Pr(Z_i > 0) = \theta_i$. Considerando este modelo, queremos a função de ligação canônica. Sabemos que $\Pr(Z_i > 0) = 1 - \Pr(Z_i = 0) = \exp\{-\mu_i\}$. Então, queremos que $g(\exp\{-\mu_i\}) = \log(\mu_i)$, ou seja, $g(t) = \log[-\log(t)]$. Esta é a chamada função de ligação Gompertz ou log-log complementar. Para o último item, observe que para $\eta_i = \boldsymbol{X}_i\boldsymbol{\beta}$:

$$\theta_i = \Pr(Y_i = 1) = \Pr(W_i > 0) = \int_{-\eta_i}^{\infty} \frac{\exp(-\epsilon)}{[1 + \exp(-\epsilon)]^2} d\epsilon = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}},$$

ou seja, a função de ligação é logit. Este modelo é uma representação com variável latente da regressão ordinal com logit para duas categorias, onde β_j representa o aumento na *log-odds* associado ao aumento em uma unidade de x_{ij}. ■

Comentário: Nesta questão vimos algumas representações diferentes de um modelo linear generalizado para uma resposta binária. Como já trabalhado nos exercícios da Lição 7, vimos como representar um modelo com o link Gompertz como uma transformação de uma contagem. Além disso, exploramos a representação da regressão logística através de variáveis latentes, o que se mostra útil em alguns cenários em que é conveniente interpretar os efeitos das covariáveis como efeitos sobre uma variável latente contínua que mede, por exemplo, a aprovação de um candidato.

2. Grit, Love and Passion.

Filipe Luís, um técnico de futebol muito estudioso, preocupado com o número de lesões do seu time na temporada, decidiu elaborar um modelo para prever essa quantidade, baseado na quantidade de jogos e nos estádios nos quais o Flamengo atua. Para isso, ele propôs que o número de lesões do jogador $i = 1, \ldots, n$ no estádio $j = 1, \ldots, m$, Y_{ij}, segue uma distribuição de Poisson tal que

$$\mathbb{E}[Y_{ij}, \mid u_j, x_{ij}] = \exp(u_j + \beta_0 + \beta_1 x_{ij}),$$

onde $u_j \sim \text{Normal}(0, \tau^2)$ é um efeito aleatório por estádio e x_{ij} é o tempo jogado na temporada.

a) (10 pontos) Suponha que Filipinho deseja estimar $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ por máxima verossimilhança. Para isso, ele precisa marginalizar $\boldsymbol{u} = (u_1, \ldots, u_m)^\top$. Ajude-o nesta tarefa, exibindo a verossimilhança marginal conjunta $f_{\boldsymbol{Y}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2)$, onde $\boldsymbol{Y} = (Y_{11}, Y_{12}, \ldots, Y_{nm})^\top$ e $\boldsymbol{X}$ é a matriz de desenho.

b) (20 pontos) Sendo bem esperto, Filipe sabe que a presença dos efeitos aleatórios introduz correlação nas observações dentro de um mesmo grupo. Vamos ajudá-lo a quantificar essa informação.

   i) Determine a esperança marginal de Y_{ij}.

   ii) Calcule a variância marginal de Y_{ij}.

   iii) Por fim, compute $\text{Cov}(Y_{kj}, Y_{lj})$, para $k \neq l$.

Conceitos trabalhados: Modelos lineares generalizados mistos; manipulação de modelos probabilísticos; marginalização e máxima verossimilhança.
Nível de dificuldade: médio .
Resolução: Para o item a, basta usar a lei da probabilidade total,

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2) &= \prod_{j=1}^{m} \int \prod_{i=1}^{n} f_{Y_{ij}}(y_{ij} \mid u_j \boldsymbol{X}, \boldsymbol{\beta}, \tau^2) g_{U_j}(u_j) du_j, \\ &= \prod_{j=1}^{m} \int \prod_{i=1}^{n} \text{Poisson}(y_{ij}; \exp(u_j + \beta_0 + \beta_1 x_{ij}))\, \text{Normal}(u_j; 0, \tau^2) du_j. \end{aligned}$$

Para o tem b, é possível mostrar que se $X \sim \text{Normal}(\mu, \sigma^2)$, então, $\exp(X) \sim \text{Log-Normal}(\mu, \sigma^2)$, cuja média é $\exp(\mu + \sigma^2/2)$ e a variância é $[\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$. Logo, $\mathbb{E}[\exp(u_j)] = \exp(\tau^2/2)$. Assim, pela lei da esperança total, $\mathbb{E}[Y_{ij}] = \exp(\tau^2/2) \exp(\beta_0 + \beta_1 x_{ij})$. Além disso, pela lei da variância total,

$$\begin{aligned} \text{Var}(Y_{ij}) &= \text{Var}(\mathbb{E}[Y_{ij} \mid u_j]) + \mathbb{E}[\text{Var}(Y_{ij} \mid u_j)], \\ &= \text{Var}(\exp(u_j + \beta_0 + \beta_1 x_{ij})) + \mathbb{E}(\exp(u_j + \beta_0 + \beta_1 x_{ij})), \\ &= \exp(2(\beta_0 + \beta_1 x_{ij}))\, \text{Var}(\exp(u_j)) + \exp(\tau^2/2) \exp(\beta_0 + \beta_1 x_{ij}), \\ &= \exp(\beta_0 + \beta_1 x_{ij})[\exp(\beta_0 + \beta_1 x_{ij}) \exp(\tau^2)[\exp(\tau^2) - 1] + \exp(\tau^2/2)]. \end{aligned}$$

Por fim, pela lei da covariância total,

$$
\begin{aligned}
\mathrm{Cov}(Y_{kj}, Y_{lj}) &= \mathbb{E}[\mathrm{Cov}(Y_{kj}, Y_{lj} \mid u_j]) + \mathbb{E}[\mathrm{Cov}(Y_{kj} \mid u_j], \mathbb{E}[Y_{lj} \mid u_j]), \\
&= \mathbb{E}[\mathrm{Cov}(Y_{kj} \mid u_j], \mathbb{E}[Y_{lj} \mid u_j]), \\
&= \mathbb{E}[\exp(u_j + \beta_0 + \beta_1 x_{kj}), \exp(u_j + \beta_0 + \beta_1 x_{lj})]), \\
&= \exp(\beta_0 + \beta_1 x_{kj}) \exp(\beta_0 + \beta_1 x_{lj}) \,\mathrm{Var}(\exp(u_j)).
\end{aligned}
$$

■

Comentário: Nesta questão – que trabalhamos na lista de revisão – nós fizemos alguns cálculos para auxiliar a interpretação e a implementação de um modelo multinível com resposta não-gaussiana, em particular um modelo de contagem com resposta Poisson. Os computações realizadas permitem o cálculo de estimadores de momentos para certas quantidades, além de possibilitar a implementação de rotinas para estimação dos parâmetros por máxima verossimilhança.

3. Houston, we have a problem.

No dia 28 de Janeiro de 1986 o ônibus espacial *Challenger* explodiu 1 minuto e 16 segundos após o lançamento, matando tragicamente todos os sete tripulantes a bordo. Uma profunda investigação posterior, que contou com a participação do famoso físico estadunidense Richard Feynman (1918-1988), revelou que a falha ocorreu numa peça chamada *O-ring*, que era um anel de aproximadamente 6 metros de diâmetro feito de borracha.

A investigação revelou que a borracha dos *O-rings* respondia mal a baixas temperaturas. Nesta questão você vai ajudar a completar a análise do desastre aplicando regressão logística a dados retrospectivos sobre falhas de *O-rings* em testes realizados antes de 1986. Seja Y_i o número de vezes que a peça falhou em n_i tentativas, com $i = 1, 2, \ldots, N$. Suponha também que medimos a temperatura (em Fahrenheit[1]) X_i em que foram realizados os experimentos.

Apresentamos aqui uma parte do *output* do ajuste de dois modelos aos dados:

```
Call:
glm(formula = fail/n ~ temp, family = "binomial",
    data = challenger, weights = n)

Deviance Residuals:
     Min        1Q    Median       3Q      Max
-0.95227 -0.78299 -0.54117 -0.04379  2.65152

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.08498   3.05247   1.666   0.0957 .
temp       -0.11560   0.04702 -2.458   0.0140 *

Null deviance: 24.230 on 22 degrees of freedom
```

[1]Because of course it is.

```
Residual deviance: 18.086 on 21 degrees of freedom

Number of Fisher Scoring iterations: 5
Call:
glm(formula = fail/n ~ 1, family = "binomial",
    data = challenger, weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8996 -0.8996 -0.8996  0.8531  1.9549

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.6626     0.3448  -7.723 1.14e-14 ***

Null deviance: 24.23 on 22 degrees of freedom
Residual deviance: 24.23 on 22 degrees of freedom
```

a) (10 pontos) Explique porque é interessante ajustar estes dois modelos e deduza o valor da verossimilhança do modelo **saturado** a partir do *output* apresentado.

b) (15 pontos) Dick Feynman, como costumava ser chamado, era um físico brilhante, mas é possível que não soubesse muito bem interpretar os resultados de uma regressão logística[2]. Com uma breve dedução matemática, mostre a Feynman como interpretar o coeficiente obtido para a temperatura. Explique o que acontece quando comparamos experimentos em que a temperatura difere por 1 grau ºF.

c) (15 pontos) Para entender o efeito de adicionar uma covariável, muitas vezes é útil computar uma estatística que relacione a *deviance* do modelo que inclui a covariável com um modelo nulo, só com o intercepto:

$$\tilde{R}^2 = 1 - \frac{D}{D_0},$$

onde D é a *deviance* do modelo sob análise e D_0 é a *deviance* do modelo nulo, que contém apenas o intercepto. Calcule, com duas casas decimais de precisão, o $\tilde{R}^2$ a partir do *output* apresentado.

d) (10 pontos, **bônus**) Como devemos interpretar $\tilde{R}^2$? Qual a sua conclusão? Incluir a temperatura melhora a capacidade de prever se a peça vai falhar?

Conceitos trabalhados: modelos lineares generalizados, regressão logística, interpretação dos coeficientes; deviance; comparação de modelos.
Nível de dificuldade: médio.
Resolução: Primeiramente, a análise destes dois modelos permite entender se há melhoria ao incluir covariáveis, com o deviance sendo uma maneira de quantificar essa informação. Além disso, no caso Bernoulli ajustado acima, o

[2] *Statistics is hard*.

modelo saturado tem verossimilhança 1, pois é o modelo "perfeito", onde as probabilidades são substituídas pelas próprias observações, ou seja,

$$\text{loglik(modelo saturado)} = \sum_{i=1}^{n} \log \left[y_i^{y_i} (1 - y_i)^{1-y_i} \right] = 0.$$

Para o item b, sendo β o coeficiente associado a temperatura X, note que

$$\log \left(\frac{\Pr(Y_i = 1 \mid X)}{\Pr(Y_i = 0 \mid X)} \right) = X\beta,$$

ou seja,

$$\log \left(\frac{\Pr(Y_i = 1 \mid X = x + 1)}{\Pr(Y_i = 0 \mid X = x + 1)} \right) - \log \left(\frac{\Pr(Y_i = 1 \mid X = x)}{\Pr(Y_i = 0 \mid X = x)} \right) = (x + 1)\beta - x\beta = \beta.$$

Logo, β representa o efeito no log da razão de chances (*odds-ratio*). Para o cálculo de $\tilde{R}^2$, observe no *output* que $D = 18.086$ e $D_0 = 24.23$, resultando em $\tilde{R}^2 \approx 0.25$. Para interpretar este resultado, note que quando o ajuste é perfeito, isto é, $D = 0$, $\tilde{R}^2 = 1$, enquanto que quando D não acrescenta informação ao modelo nulo, $D = D_0$, ou seja, $\tilde{R}^2 = 0$. Então, quanto mais próximo de 0 o $\tilde{R}^2$ for, isso sugere para a não inclusão das covariáveis, como é o nosso caso. ■

Comentário: A regressão logística é um modelo relativamente simples, mas bastante poderoso. Essa questão trata de uma solução simples para um problema complexo, a saber prever a probabilidade de uma falha catastrófica no lançamento da Challenger (que aconteceu a 31 ºF). Para muito mais, ver a análise de Dalal et al (1989).

Bibliografia

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, 84(408):945–957.