1^a Prova - FECD B - 2025/01

Renato Assunção

1. An agricultural researcher is studying the number of insect eggs found on individual leaves of a particular crop. For each sampled leaf, the following two features are recorded: x_1 : the leaf area (in square centimeters) and x_2 : exposure level to sunlight (coded as 0 = shaded, 1 = partial sun, 2 = full sun). Let Y_i denote the number of insect eggs observed on leaf i. The researcher assumes the counts follow a Poisson distribution with mean λ_i, and proposes the following model for the Poisson mean:

$$\lambda_i = \exp(\beta_0 + \beta_1 \cdot \text{Area}_i + \beta_2 \cdot \text{Sunlight}_i)$$

Based on a random sample of n leaves with observed values $\{(y_i, \text{Area}_i, \text{Sunlight}_i), i = 1, \ldots, n\}$, derive the MLE for $\beta_0, \beta_1, \beta_2$ under the model:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{with } \lambda_i = \exp(\beta_0 + \beta_1 \cdot \text{Area}_i + \beta_2 \cdot \text{Sunlight}_i)$$

Write the log-likelihood function, compute the gradient, and describe how the MLEs could be computed in practice using Newton-Raphson.

Solution: We define the linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Then $\lambda_i = \exp(\eta_i)$. The log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\lambda_i) - \lambda_i - \log(y_i!)] = \sum_{i=1}^{n} [y_i \eta_i - \exp(\eta_i)] + \text{const}$$

The gradient (score function) is:

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - \lambda_i) \mathbf{x}_i$$

The Hessian is:

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i'$$

The Newton-Raphson update is:

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\right]^{-1} \left[\frac{\partial \ell}{\partial \beta}\right]$$

In practice, this is implemented in GLM Poisson regression routines. For example, we can use the *statsmodel* package in Python. The partial output is shown in Figure 1. As expected, the fitted coefficients are close to the true values used in simulation: $\beta_0 = 0.5$, $\beta_1 = 0.3$, and $\beta_2 = 0.2$.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Set seed for reproducibility
np.random.seed(42)

# Generate fake data
n = 100
area = np.random.uniform(1, 10, size=n)
sunlight = np.random.choice([0, 1, 2], size=n)

# True model parameters
beta_0 = 0.5
beta_1 = 0.3
beta_2 = 0.2
```

	coef	std err	z	P>|z|
const	0.5021	0.100	5.046	0.000
area	0.3098	0.012	25.587	0.000
sunlight	0.1326	0.033	4.034	0.000

Figura 1: Output of summary function with GLM Poisson model from *statsmodel* package.

```
# Calculate expected counts (lambda)
lin_pred = beta_0 + beta_1 * area + beta_2 * sunlight
lambda_ = np.exp(lin_pred)
y = np.random.poisson(lambda_)

# Create DataFrame
df = pd.DataFrame({'y': y, 'area': area, 'sunlight': sunlight})

# Add constant for intercept
X = sm.add_constant(df[['area', 'sunlight']])
y = df['y']

# Fit Poisson GLM
poisson_model = sm.GLM(y, X, family=sm.families.Poisson())
results = poisson_model.fit()

# Print summary
print(results.summary())
```

---

2. Considere o modelo de regressão linear:

$$Y = X\beta + \varepsilon, \quad \text{com } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

onde $X \in \mathbb{R}^{n\times(p+1)}$ tem posto completo e sua primeira coluna é composta por 1's. Seja $\hat{\beta} = (X'X)^{-1}X'Y$ o estimador de mínimos quadrados e $\hat{Y} = X\hat{\beta}$ o vetor de valores ajustados. A matriz $H = X(X'X)^{-1}X'$ é chamada de matriz *hat*.

   (a) Mostre que H é idempotente e simétrica, ou seja, $H^2 = H$ e $H' = H$.
   Solution: Como $H = X(X'X)^{-1}X'$, então:

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H \Rightarrow \text{H é idempotente}$$

$$H' = (X(X'X)^{-1}X')' = X(X'X)^{-1}X' = H \Rightarrow \text{H é simétrica}$$

   (b) A projeção ortogonal de cada coluna da matriz X no Seja $\mathcal{M}(X)$ o sub-espaço vetorial em $\mathbb{R}^n$ das combinações lineares das colunas da matriz X. Dado um vetor $v \in \mathbb{R}^n$, a sua projeção ortogonal em $\mathcal{M}(X)$ é dada por Hv. Aplicar H à matrix X (isto é, obter HX) retorna uma matriz $n\times(k+1)$ em que cada coluna é a projeção ortogonal da coluna correspondents de X. Mostre que $HX = X$ (ou seja, a projeção de cada coluna é ela mesma). Mostre também que $(I - H)X = 0$.
   Solution:

$$HX = X(X'X)^{-1}X'X = X \Rightarrow HX = X$$

$$(I - H)X = X - HX = X - X = 0$$

---

3. **Variância dos Resíduos e Valores Ajustados** Com as mesmas definições do problema anterior, defina o vetor de resíduos como $r = Y - \hat{Y}$.

(a) Mostre que $\text{Var}(\hat{Y}) = \sigma^2 H$ e $\text{Var}(r) = \sigma^2(I - H)$.
Solution: Como $\hat{Y} = HY$, e $\text{Var}(Y) = \sigma^2 I$, temos:

$$\text{Var}(\hat{Y}) = \text{Var}(HY) = H \cdot \text{Var}(Y) \cdot H' = \sigma^2 HH = \sigma^2 H$$

$$r = (I - H)Y \Rightarrow \text{Var}(r) = (I - H)\sigma^2 I(I - H)' = \sigma^2(I - H)(I - H) = \sigma^2(I - H)$$

(b) Mostre que os vetores $\hat{Y}$ e r são não correlacionados.
Solution:

$$\text{Cov}(\hat{Y}, r) = \text{Cov}(HY, (I - H)Y) = H \cdot \text{Var}(Y) \cdot (I - H)' = \sigma^2 H(I - H) = 0$$

pois $H(I - H) = H - H^2 = H - H = 0$

---

4. **Vício e Variância do Estimador em Regressão Linear** Considere o modelo de regressão simples:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{com } \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

Defina $\hat{\beta}$ como o estimador de mínimos quadrados do vetor $\beta = (\beta_0, \beta_1)$.

(a) Mostre que $\hat{\beta}$ é não-viciado para estimar β.
Solution: Em forma matricial, $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$. Como $\mathbb{E}[\varepsilon] = 0$, então:

$$\mathbb{E}[\hat{\beta}] = \beta + (X'X)^{-1}X' \cdot 0 = \beta$$

(b) Derive a variância de $\hat{\beta}$.
Solution:

$$\text{Var}(\hat{\beta}) = \text{Var}\left((X'X)^{-1}X'\varepsilon\right) = (X'X)^{-1}X' \cdot \sigma^2 I \cdot X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$