SME0820 Modelos de Regressão e Aprendizado Supervisionado I: Lista 1

Thomas Peron.

Data de publicação: 28/08/2023. Data da provinha: $\boxed{15/09/2023}$. Data de entrega exercício 💻: $\boxed{15/09/2023}$

Resolva os exercícios computacionais (💻) da maneira que quiser: com o software ou linguagem de sua preferência (`R`, `Python`, `C`, `Fortran`, etc.), manualmente, ou ambos.

1. Demonstre as seguintes relações:

   (a)
$$\frac{\sum_i X_i Y_i - \left[(\sum_i X_i)(\sum_i Y_i)\right]/n}{\sum_i X_i^2 - (\sum_i X_i)^2/n} = \frac{\sum_i (X_i - \overline{X})(Y_i - \bar{Y})}{\sum_i (X_i - \overline{X})^2};$$

   (b)
$$S_{XY} = \sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_i X_i Y_i - n\overline{X}\overline{Y} = \sum_i (X_i - \overline{X}) Y_i = \sum_i X_i (Y_i - \overline{Y});$$

   (c)
$$S_{XX} = \sum_i (X_i - \overline{X})^2 = \sum_i X_i^2 - n\overline{X}^2 = \sum_i (X_i - \overline{X}) X_i.$$

2. Considere o modelo de regressão linear simples
$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$
   onde $\mathbb{E}[\varepsilon] = 0$ e $\mathrm{Var}[\varepsilon] = \sigma^2$. Utilizando os resultados dos parâmetros ótimos β_0 e β_1, mostre que
$$\mathbb{E}\left[(Y - (\beta_0 + \beta_1 X))^2\right] = \sigma^2. \tag{2}$$

3. Para o modelo de regressão linear
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n, \tag{3}$$
   onde ϵ_i é uma variável aleatória satisfazendo $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ e $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0\ \forall i, j$, calcule as seguintes quantidades relacionadas aos estimadores $\hat{\beta}_1$ e $\hat{\beta}_0$ obtidos pelo métodos dos mínimos quadrados:

   (a) $\mathbb{E}[\hat{\beta}_0]$ e $\mathrm{Var}[\hat{\beta}_0]$.

   (b) $\mathbb{E}[\hat{\beta}_1]$ e $\mathrm{Var}[\hat{\beta}_1]$.

   (c) $\mathbb{E}[\hat{Y}]$ e $\mathrm{Var}[\hat{Y}]$, onde $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{m}(x)$ é o modelo ajustado. Em particular, para o cálculo desta variância, mostre que
$$\mathrm{Var}\left[\frac{1}{n}\sum_i \left(1 + (x - \overline{X})\frac{(X_i - \overline{X})}{S_{XX}}\right)\varepsilon_i\right] = \frac{\sigma^2}{n}\left[1 + \frac{(x - \overline{X})^2}{S_{XX}}\right]$$

4. Ainda para o modelo (3), mostre que

   (a) $\sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$,

   (b) $\sum_i X_i e_i = 0$,

   (c) $\sum_i \hat{Y}_i e_i = 0$,

   sendo que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são os parâmetros estimados pelo método dos mínimos quadrados.

5. 💻(*Verificando flutuações nos parâmetros*) Repita o experimento que vimos em sala[1]. Suponha que a relação verdadeira entre as variáveis é dada por
$$Y = \beta_0 + \beta_1 X + \varepsilon,\ \text{with } \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{4}$$
   com $\beta_0 = 3$, $\beta_1 = 5$, e $\sigma^2 = 4$, por exemplo (você pode testar outros valores também). Agora

[1]“You don’t understand your model until you can simulate from it.” (Gelman, Hill, e Vehtari em *Regression and Other Stories*)

(a) realize N_{exp} experimentos (e.g., $N_{\text{exp}} = 50, 100, ...$), e, para cada experimento, construa uma tabela $\{(X_i, Y_i)\}_{i=1}^{N}$ com N pontos amostrados de Eq. (4). Ajuste o modelo usando os estimadores $\hat{\beta}_{0,1}$ obtidos pelo métodos de mínimos quadrados. Calcule a média empírica e desvios padrão de $\hat{\beta}_{0,1}$ e os compare com os valores verdadeiros ($\beta_0 = 3$, $\beta_1 = 5$). Visualize as flutuações no modelo ajustado como vimos em sala.

(b) (*Dependência com o tamanho da amostra*) Como em (a), estime os coeficientes $\beta_{0,1}$ usando mínimos quadrados, mas agora, em cada experimento, varie o tamanho da amostra, N. Visualize os parâmetros estimados $\hat{\beta}_{0,1}$ como função de N. Discuta os resultados. *Dica:* Para obter resultados mais robustos, faça várias realizações para um tamanho fixo da amostra.

(c) Para um tamanho fixo de amostra, crie uma visualização para os dados $\{(X_i, Y_i)\}_{i=1}^{N}$, o modelo ajustado $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ e seu erro padrão $\mathrm{SE}[\hat{m}(x)]$.

(d) Também para um tamanho fixo de amostra, realize vários experimentos aumentando o intervalo de amostragem de X. Visualize $\mathrm{Var}[\hat{\beta}_0]$ e $\mathrm{Var}[\hat{\beta}_1]$ em função de S_{XX}, e compare as curvas com os valores teóricos obtidos no exercício 3.