

SME0820 Modelos de Regressão e Aprendizado Supervisionado I: Lista 2

Thomas Peron.

Data de publicação: 23/09/2023. Data da prova: 06/10/2023. Data de entrega exercícios: 12/10/2023

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. Para o modelo de regressão linear

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

onde ε_i é uma variável aleatória satisfazendo $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ e $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, onde δ_{ij} é a função delta de Kronecker, mostre que

(a) $\text{cov}(\hat{\beta}_1, \bar{Y}) = 0$.

(b) $\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{X}\sigma^2/S_{xx}$.

2. Para o modelo da Eq. (1), mostre que $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$, onde

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{n-2},$$

sendo $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ a resposta estimada em $X = x_i$.

3. Para o caso de duas amostras apenas ($n = 2$) na Eq. (1), demonstre as seguintes relações

(a) $(y_1 - \hat{y}_1) = (y_2 - \hat{y}_2) = 0$,

(b) $R^2 = 1$ (Coeficiente de determinação $R^2 = SSR/SST = 1 - SSE/SST$).

4. Num problema de regressão linear simples, como o descrito na Eq. (1), encontre a relação entre o estimador $\hat{\beta}_1$ e o coeficiente de correlação $r_{XY} = \text{cov}(X, Y) / \sqrt{\text{Var}(X)\text{Var}(Y)}$.

5. Há aplicações importantes (e.g., verificações de leis científicas) em que, devido a restrições conhecidas, a linha de regressão *deve passar pela origem* (isto é, o intercepto é zero). Dito de outro modo, o modelo deve ser definido como

$$Y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (2)$$

de maneira que apenas um parâmetro deve ser estimado. Considere que os termos ε_i satisfazem as mesmas condições como no modelo (1).

- (a) Mostre que o estimador da inclinação da reta é dado por

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \quad (3)$$

- (b) Mostre que

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}. \quad (4)$$

- (c) Mostre também que $\hat{\beta}_1$ em (a) é um estimador não-enviesado para β_1 . Isto é, verifique que $\mathbb{E}(\hat{\beta}_1) = \beta_1$.

6. (p -valor é uma variável aleatória) A estatística que calculamos no teste t ,


$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\text{SE}}[\hat{\beta}_1]},$$

possui a propriedade de ser próxima a zero quando a hipótese nula $H_0 : \beta_1 = \beta_1^*$ for verdadeira, e de assumir valores extremos, que podem ser tanto negativos quanto positivos, quando H_0 for falsa. Se a estatística de teste possui essas propriedades, convém sumarizar num único índice como os dados se adéquam à hipótese nula – o chamado p -valor. Seja T^* o valor observado para a estatística T , o p -valor é definido como

$$P = \Pr\{|T| \geq |T^*|\}, \quad (5)$$


que é a probabilidade de que uma variável aleatória de distribuição t -Student seja maior do que o valor observado T^* .

Neste exercício mostraremos que P que é uma variável aleatória uniformemente distribuída sob a hipótese nula. Siga os passos abaixo, considerando que T possua uma distribuição contínua.

- (a) Mostre que se $Q \sim \text{Uniforme}(0, 1)$, então $P = 1 - Q$ possui a mesma distribuição.
 - (b) Seja X uma variável aleatória contínua cuja CDF é F . Mostre que $F(X) \sim \text{Uniforme}(0, 1)$. Dica: a CDF de uma distribuição uniforme é $F_{\text{Unif}(0,1)}(x) = x$.
 - (c) Mostre que P , como definido acima, pode ser escrito como $1 - F_{|T|}(|T^*|)$.
 - (d) Usando os itens anteriores, mostre que $P \sim \text{Uniforme}(0, 1)$.
7.  Suponha, como no exercício 5, que nosso problema é modelado por uma equação linear que passa pela origem, i.e. $\mu_{Y|x} = \beta_1 x$. Dito isso,

- (a) estime a linha de regressão para os seguintes dados:

x	0.5	1.5	3.2	4.2	5.1	6.5
y	1.3	3.4	6.7	8.0	10.0	13.2

- (b) Imagine que não sabemos se a verdadeira linha de regressão deva passar pela origem ou não. Estime o modelo linear $\mu_{Y|x} = \beta_0 + \beta_1 x$, e teste a hipótese de que $\beta_0 = 0$, com um nível de 90% de significância em relação à hipótese alternativa $\beta_0 \neq 0$.
8.  Há um certo tipo de molusco, do gênero *Haliotis*¹, cuja carne é apreciada por várias culturas, podendo esta ser consumida tanto crua quanto cozida. Um problema encontrado por cientistas que estudam esse animal é determinar a idade de indivíduos a partir do tamanho de suas conchas. Essa não é uma tarefa simples de realizar, porque o crescimento das conchas não depende apenas do tempo de vida, e sim também da disponibilidade de alimento. Uma abordagem comumente adotada é retirar uma amostra da concha e analisar, com a ajuda de um microscópio, o número de anéis presentes nela. Imagine você faça parte de um grupo de pesquisa que esteja interessado em utilizar as medidas físicas dos moluscos, especialmente a altura da concha, para prever seus tempos de vida. Acredita-se que um modelo de regressão linear simples com erros normais seja suficiente para descrever a relação entre altura e idade. Em particular, o grupo busca dar suporte à teoria de que conchas maiores correspondem a animais mais velhos.

Os dados que utilizaremos neste exercício estão no arquivo `molusco.csv`. Mais informações sobre essa base podem ser encontradas em <https://archive.ics.uci.edu/ml/datasets/Abalone>.

- (a) Escreva algumas sentenças descrevendo o problema da pesquisa e a hipótese científica que será verificada.
- (b) Examine as duas variáveis da base de dados individualmente. Faça um resumo de suas medidas (média, variância, intervalo de amostragem, etc). Faça essa descrição também por meio de gráficos. Qual é a unidade de Height?
- (c) Faça um gráfico de dispersão dos dados. Descreva as tendências interessantes observadas.
- (d) Ajuste uma linha de regressão aos dados, prevendo o número de anéis na concha utilizando o tempo de vida dos moluscos.

¹<https://en.wikipedia.org/wiki/Haliotis>

- (e) Crie um gráfico de dispersão que mostre os dados e a função de regressão estimada (você pode incluir no gráfico anterior). Descreva o ajuste da reta.
- (f) Forneça um intervalo de 95% de confiança para β_0 e β_1 . Interprete no contexto do problema.
- (g) Há uma relação estatística significativa entre altura e número de anéis (e, portanto, tempo de vida) dos moluscos?
- (h) Faça uma estimativa pontual e encontre o intervalo de 95% de confiança para o número médio de anéis de moluscos com altura 0.128 (na mesma unidade das outras observações).
- (i) Estamos interessados agora em *predizer* o número de anéis de um molusco de tamanho 0.132 (mesma unidade que as anteriores). Encontre o valor predito e o intervalo de predição com 99% de significância.
- (j) Conclua brevemente o estudo.