

## SME0820 Modelos de Regressão e Aprendizado Supervisionado I: Lista 4

Thomas Peron

**Data de entrega dos exercícios** 📅 : 10/12/2023


*Resolva os exercícios que acompanham dados da maneira que quiser: com o software de sua preferência, manualmente, ou ambos.*

*Não haverá provinha teórica sobre esta lista; utilize-a como treino para a P2*

- Escreva as seguintes somas extras de quadrados e enuncia seus graus de liberdade:
  - $SSR(X_5|X_1)$ ;
  - $SSR(X_3, X_4|X_1)$ ;
  - $SSR(X_4|X_1, X_2, X_3)$ .
- 📄 O arquivo `dados_temperatura_pressao.csv` contém os dados da variação da pressão  $Y$  em função da temperatura  $X$  em um certo sistema mecânico.
  - Ajuste um modelo linear de primeira ordem. Visualize o gráfico do modelo ajustado juntamente com o gráfico de dispersão dos dados. Discuta a adequabilidade do modelo.
  - Faça um gráfico dos resíduos em função da resposta predita,  $\hat{Y}$ , e discuta novamente adequabilidade do modelo.
  - Ajuste agora um modelo de segunda-ordem aos dados. Há alguma evidência de que o termo quadrático é estatisticamente significativo?
  - Repita os itens (a) e (b) para o modelo de segunda ordem. Há evidência de que o modelo de segunda ordem fornece um melhor ajuste aos dados?
- 📄 Considere o arquivo `dados_genericos.csv` contendo a variável resposta  $Y$  em função das covariáveis  $X_1$ ,  $X_2$  e  $X_3$ .
  - Ajuste um modelo quadrático aos dados, incluindo termos cruzados; isto é, considere o modelo
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{1,1} X_1^2 + \beta_{2,2} X_2^2 + \beta_{3,3} X_3^2 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \beta_{2,3} X_2 X_3 + \varepsilon,$$
onde  $\varepsilon$  possui as propriedades usuais:  $\mathbb{E}(\varepsilon) = 0$  e  $\text{Var}(\varepsilon) = \sigma^2$ . *Dica:* use o `PolynomialFeatures`<sup>1</sup> para criar a tabela com as covariáveis  $\{1, X_1, X_2, X_3, X_1^2, \dots, X_2 X_3\}$ .
  - Teste a significância da regressão, e construa a estatística  $t$  para cada coeficiente do modelo. Discuta os resultados.

<sup>1</sup><https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html>

- (c) Visualize os resíduos e comente sobre a adequabilidade do modelo.
- (d) Utilize a soma extra dos quadrados para avaliar a contribuição de todos os termos de segunda ordem do modelo.

4.  O arquivo `dados_comunidades_mobilidade.csv` contém as seguintes informações sobre 729 cidades dos Estados Unidos:

**Mobility:** A probabilidade de que uma criança nascida entre 1980-1982 e pertencente ao quantil mais baixo (20%) de renda familiar chegue ao quantil mais alto aos 30 anos de idade. A cada indivíduo é atribuída a sua cidade de origem.

**Commute:** Fração de trabalhadores que levam menos de 15 minutos para chegar ao local de trabalho.

**Longitude:** Coordenada geográfica do centro da cidade.

**Latitude:** Idem.

**Name:** Nome da cidade.

**State:** Estado a que pertence a cidade.

Neste problema iremos prever a mobilidade econômica ( $Y$ ) de uma fração da população através da variável `Commute` ( $X_i$ ), considerando os seguintes modelos

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

$$Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i, \quad (2)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i. \quad (3)$$

- (a) Ajuste os três modelos e explore suas propriedades. Qual modelo você escolheria para descrever os dados? Por quê?
- (b) Calcule a correlação entre  $X_i$  e  $X_i^2$ . Em seguida, centre os valores de  $X_i$  em torno da média, i.e. crie  $Z_i = X_i - \bar{X}$ , e recalcule a correlação entre  $Z_i$  e  $Z_i^2$ . Há alguma mudança nas correlações?
- (c) Ajuste novamente o modelo (3) utilizando  $Z_i$  e  $Z_i^2$ . Compare os coeficientes estimados e os erros padrão dos modelos definidos em termos de  $X_i$  e  $Z_i$ . Comente se a transformação de variável é útil neste problema.
- (d) Os estados dos Sul e Norte dos EUA possuem características políticas e socioculturais significativamente distintas, e é bastante provável que essas diferenças influenciem as estatísticas de mobilidade social. Com as informações do arquivo `dados_comunidades_mobilidade.csv`, construa uma variável categórica que divida os estados em dois grupos, Sul e o resto do país<sup>2</sup>. Ajuste o modelo (3) com a variável categórica, e discuta se ela é relevante para explicar a mobilidade social.

---

<sup>2</sup>Defina o Sul como sendo composto pelos estados Virginia, Florida, Georgia, Alabama, Arkansas, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee e Texas ("VA", "FL", "GA", "AR", "AL", "LA", "MS", "NC", "SC", "TN", "TX").