

## SME0820 Modelos de Regressão e Aprendizado Supervisionado I: Lista 3

Thomas Peron



Data de entrega dos exercícios 📅 : 17/11/2023

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

1. (Não haverá prova teórica sobre esta lista; utilize-a como treino para a P2) Considere o modelo de regressão linear múltipla dado por

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{p-1,i} + \varepsilon_i \quad (i = 1, \dots, n). \quad (1)$$

- (a) Escreva a forma matricial da Eq. (1), incluindo as suposições acerca dos erros  $\varepsilon_i$  discutidas em sala. Indique as dimensões das matrizes relevantes. Considere  $p < n$ .
  - (b) Obtenha as equações normais pelo método dos mínimos quadrados e a partir delas encontre o vetor dos coeficientes ajustados  $\hat{\beta}$ . Mostre que  $\hat{\beta}$  é um estimador não viciado e calcule a sua matriz de variância.
  - (c) Calcule o valor esperado e a matriz de variância da resposta ajustada,  $\hat{Y} = X\hat{\beta}$ , onde  $X$  é a matriz com os valores das covariáveis, como definido nas aulas.
  - (d) Explique como é definida a matriz  $H$ . Mostre que  $H$  é idempotente e simétrica.
  - (e) Seja  $e = Y - \hat{Y}$  o vetor de resíduos. Expresse  $e$  em termos da matriz  $H$ . Calcule seu valor esperado e sua matriz de covariância.
  - (f) Calcule o valor esperado e a matriz de covariância da resposta média  $\hat{Y}_a$  em  $X_a = [1 \ X_{1a} \ X_{2a} \ \dots \ X_{p-1,a}]^T$ . Encontre o intervalo de  $100(1 - \alpha)\%$  de confiança de  $\hat{Y}_a$  em termos de  $X_a$ ,  $X$  e  $MSE$ .
  - (g) Seja  $Y_a$  uma nova observação feita para  $X_a = [1 \ X_{1a} \ X_{2a} \ \dots \ X_{p-1,a}]^T$ . Calcule o intervalo de predição de  $Y_a$ , com  $100(1 - \alpha)\%$  de confiança, em termos de  $X_a$ ,  $X$  e  $MSE$ .
2. 📁 Este exercício utilizará os dados do projeto SENIC (SENIC.csv), que está descrito no Apêndice C.1 do livro texto [Applied Linear Statistical Models (5 ed.), J. Neter, W. Wasserman, e M. Kutner]. O objetivo é explicar o tempo médio ( $Y$ ) de permanência de um paciente num hospital em termos das covariáveis fornecidas (veja o livro para as descrições detalhadas). Suponha que o modelo 1 utilize como preditores a idade do paciente ( $X_1$ ), probabilidade do risco de infecção ( $X_2$ ) e porcentagem de serviços fornecidos ( $X_3$ ). O modelo 2 utilizará número de leitos ( $X_1$ ), probabilidade do risco de infecção ( $X_2$ ) e porcentagem de serviços fornecidos ( $X_3$ ).
- (a) Para cada um dos modelos, crie *scatter plots* das covariáveis e visualize as correlações entre elas. Descreva esses resultados brevemente.
  - (b) Para cada um dos modelos, ajuste um modelo de regressão linear múltipla [Eq. (1)] com três variáveis preditoras.

- (c) Calcule  $R^2$  para cada modelo. Algum modelo é preferível em termos dessa medida?
- (d) Novamente para cada modelo, obtenha os resíduos e visualize-os em função da resposta ajustada  $\hat{Y}$  e em função de cada variável preditora (você pode agrupar a dispersão de  $X_i \times e_i$  em um mesmo gráfico, visualizando os pontos de cada covariável de maneira diferente).
- (e) Prepare um *QQ plot* para cada um dos modelos ajustados. Analise os resultados. Algum modelo é mais apropriado em termos da análise dos resíduos?
3.  Para este exercício, utilize novamente os dados do projeto SENIC. Considere agora as quatro covariáveis do exercício acima: idade ( $X_1$ ), número de leitos ( $X_2$ ), risco de infecção ( $X_3$ ) e porcentagem de serviços prestados ( $X_4$ ).
- (a) Crie uma matriz de correlações entre as covariáveis.
- (b) Ajuste um modelo de regressão para explicar o tempo médio de permanência no hospital apenas em termos da porcentagem de serviços prestados. Qual é o valor do coeficiente e seu erro padrão?
- (c) Ajuste um modelo linear agora com as quatro covariáveis e enuncie a função de regressão estimada.
- (d) Qual a diferença entre o coeficiente da covariável de serviços prestados do modelo ajustado em (b) e (c)? Ambos são significativos com nível de 95% de confiança? Se encontrar alguma diferença, forneça uma explicação baseada na sua observação feita em (a).
- (e) Forneça uma interpretação para o coeficiente do risco de infecção. Teste, utilizando um  $\alpha$  apropriado, a hipótese de que o coeficiente relacionando risco de infecção e tempo de permanência no hospital seja diferente de zero. Forneça suas conclusões no contexto das variáveis deste problema.
- (f) Qual é o valor da média dos quadrados dos resíduos? Explique se o modelo se ajusta aos dados de maneira satisfatória.
4.  Para este exercício, utilize novamente os dados do projeto SENIC.
- (a) Para cada região geográfica, ajuste um modelo de regressão ( $Y$ ) com as covariáveis idade ( $X_1$ ), taxa de culturas coletadas de sinais ou sintomas de terem adquirido infecção ( $X_2$ ), número médio de pacientes no hospital (census,  $X_3$ ) e número de serviços disponibilizados ( $X_4$ ). Enuncie as funções de regressão encontradas.
- (b) Os modelos ajustados das quatro regiões são similares? Discuta.
- (c) Calcule o  $MSE$  e o  $R^2$  para cada região. Discuta como a variância dos dados de cada região é explicada por cada modelo.
- (d) Obtenha os resíduos e crie *QQ plots* para cada modelo ajustado. Interprete e discuta os resultados.