Universidade Federal do Paraná - Departamento de Estatística

CE310 - Modelos de Regressão Linear

Prof. Cesar Augusto Taconeli

Lista de exercícios

1. Defina o modelo de regressão linear simples. Especifique cada um de seus componentes e as suposições assumidas para os erros.

2. Qual o princípio da estimação por mínimos quadrados? Quais as principais propriedades dos estimadores de mínimos quadrados dos parâmetros de um modelo de regressão linear?

3. Qual o princípio da estimação por máxima verossimilhança? Qual a relação dos estimadores de mínimos quadrados e de máxima verossimilhança dos parâmetros do modelo de regressão linear se assumirmos que os erros são normalmente distribuídos?

4. Considere o modelo de regressão linear simples com $\beta_0 = 10$, $\beta_1 = 5$ e $\sigma = 4$. Assuma distribuição normal para os erros.

a) Apresente gráficos da distribuição de y condicional a (i) $x = 3$; (ii) $x = 5$ e (iii) $x = 10$;

b) Descreva o significado de β_0 e β_1. Suponha que $x = 0$ pertença ao escopo do modelo;

c) Calcule $P(20 < y < 30)$ para: (i) $x = 3$; $x = 5$.

5. Considere o modelo de regressão linear simples com $\beta_0 = 10$; $\beta_1 = 0.5$ e $\sigma = 1$. Assuma que os erros sejam normalmente distribuídos.

a) Simule uma amostra de $n = 10$ observações para y locadas em x = -2, -1, 0, 1 e 2 (duas observações para cada valor de x);

b) Ajuste um modelo de regressão linear simples aos dados simulados no item a. Extraia as estimativas de β_0 e β_1.

c) Repita os itens a e c 5000 vezes. Armazene as estimativas obtidas numa matriz com duas colunas (uma referente a cada parâmetro);

d) Construa histogramas e calcule média e variância das estimativas produzidas para cada parâmetro. Compare os resultados obtidos na simulação aos apresentados no item b;

e) Para cada uma das 5000 simulações, obtenha os intervalos de confiança 95% para β_0 e β_1. Qual proporção dos intervalos contêm os valores fixados para os respectivos parâmetros?

6. Considere o modelo de regressão linear simples sem intercepto:

$$y = \beta x + \epsilon,$$

com as suposições usuais para os erros para o modelo de regressão linear.

a) Mencione uma situação prática em que o modelo de regressão linear passando pela origem possa ser considerado;

b) Determine o estimador de mínimos quadrados de β;

c) Obtenha esperança e variância para o estimador deduzido no item b.

7. Neste exercício consideramos transformações lineares de x e y. Em todos os itens, considere o modelo de regressão linear simples conforme especificado em sala de aula. Sejam β_0, β_1, SQE e r os parâmetros do modelo, a soma de quadrados dos erros e o coeficiente de correlação, respectivamente.

a) Suponha que cada valor de x seja transformado usando $x' = x - 10$ e a regressão linear simples de y em x'. Como ficam β_0', β_1', $\mathrm{SQ}'_{\mathrm{Res}}$ e r'? O que acontece com essas quantidades quando $x' = 10x$? E quando $x' = 10(x-1) = 10x - 10$?

b) Agora, suponha que os valores da variável resposta sejam transformados para $y' = y + 10$ e considere a regressão de y'. em x. Como ficam β_0', β_1', $\mathrm{SQ}'_{\mathrm{Res}}$ e r'? O que acontece com essas quantidades quando $y' = 5y$? E quando $y' = 5(y+2) = 5y + 10$?

c) Em geral, como os resultados da regressão linear simples ficam afetados por transformações lineares em x e em y?

8. Solicitado a especificar o modelo de regressão linear simples, um aluno escreveu o seguinte:

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x + \epsilon.$$

Você concorda com essa especificação? Justifique.

9. Qual o impacto da ausência de normalidade dos erros nas propriedades dos estimadores de mínimos quadrados do modelo de regressão linear?

10. Mostre que:

a) $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$;

b) $\sum_{i=1}^n r_i = 0$;

c) Para $x = \bar{x}$ tem-se $\hat{y} = \bar{y}$;

d) $\sum_{i=1}^n (x_i - \bar{x}) = 0$;

e) $\sum_{i=1}^n (x_i - \bar{x}) y_i = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$;

f) $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$;

g) $\sum_{i=1}^n x_i r_i = 0$;

h) $\sum_{i=1}^{n} \hat{y}_i r_i = 0$.

11. Um estudo foi conduzido para avaliar o efeito da temperatura na produção química de um processo. Os seguintes dados foram coletados:

Temp.	-5	-4	-3	-2	-1	0	1	2	3	4	5
Prod.	1	5	4	7	10	8	9	13	14	13	18

Vamos proceder a análise dos dados usando o modelo de regressão linear simples.

a) Determine as estimativas de mínimos quadrados de β_0 e β_1 e apresente a equação do modelo ajustado.

b) Apresente um intervalo de confiança (95%) para β_1;

c) Teste a hipótese $H_0 : \beta_1 = 0$ ao nível de significância de 5%;

d) Apresente os limites de confiança (95%) para a resposta média quando a temperatura é igual a 3;

e) Apresente limites de confiança (95%) para a diferença nas resposta média quando a temperatura é igual a 3 em relação à resposta média sob temperatura -2;

f) Sob qual temperatura se estima produção igual a 12?

12. A base de dados `Prestige` do pacote `car` apresenta dados referentes à percepção da população canadense quanto a 102 diferentes profissões. Vamos considerar, para ajuste de um modelo de regressão linear simples, as seguintes variáveis:

- `education`: Educação média dos profissionais (em anos de estudo);
- `prestige`: Escore de prestígio da profissão segundo a resposta dos entrevistados.

Considere o prestígio da profissão como a resposta e a escolaridade média como a variável explicativa.

a) Ajuste o modelo de regressão linear simples aos dados apresentados e apresente a equação do modelo ajustado;

b) Construa o diagrama de dispersão e adicione a reta de regressão ajustada. A reta obtida parece se ajustar bem aos dados?

c) Qual a predição para o escore de prestígio para uma profissão com escolaridade média de 12.5 anos?

d) Qual o valor ajustado pelo modelo para o prestígio dos administradores públicos (primeira linha da base)? Qual o correspondente resíduo?

e) Em quanto se estima a variação esperada no escore de prestígio para um ano a mais de escolaridade média entre os profissionais? E para três anos a mais?

f) O intercepto tem alguma interpretação prática nesta análise?

g) Apresente uma estimativa para σ^2;

h) Teste a hipótese $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$ ao nível de significância de 5% e apresente suas conclusões;

i) Teste a hipótese $H_0 : \beta_1 = 6$ *vs* $H_1 : \beta_1 \neq 6$ ao nível de significância de 5% e apresente suas conclusões;

j) Apresente intervalos de confiança (95%) para os parâmetros do modelo;

k) Apresente intervalos de confiança para a média e de predição considerando (i) $x = 9$; (ii) $x = 15$; (iii) $x = \bar{x}$;

l) Adicione ao diagrama de dispersão as bandas de confiança e de predição (95%);

m) Apresente o quadro de análise de variância e o teste F. Compare o p-valor desse teste ao teste da hipótese $H_0 : \beta_1 = 0$;

n) Calcule o valor de R^2 e interprete-o.

13. Neste exercício vamos analisar os dados de velocidade (x, em milhas por horas) e consumo de combustível (y, em milhas por galão) para $n = 28$ automóveis de certa marca. Os dados estão disponíveis na página da disciplina no Moodle. O objetivo é ajustar um modelo de regressão linear simples para explicar o consumo de combustível em função da velocidade sem usar a função `lm` e suas dependências.

Dados:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -1184.39; \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 7316.96; \quad \sum_{i=1}^{n}(y_i - \hat{(y)}_i)^2 = 11.20 \quad \bar{x} = 75.54; \quad \bar{y} = 12.89$$

a) Usando o R, faça o gráfico de dispersão;

b) Calcule as estimativas de mínimos quadrados para β_0 e β_1. Interprete-as.

c) Qual a variação estimada no consumo de combustível para 15mph a mais de velocidade?

d) Escreva a expressão do modelo ajustado. Calcule o consumo de combustível estimado sob velocidade $x = 75mph$;

e) Calcule os erros padrões de $\hat{\beta}_0$ e $\hat{\beta}_1$;

f) Estime as variâncias para (i) o consumo médio sob velocidade $x = 75mph$; (ii) o consumo predito para um particular automóvel sob velocidade $x = 75mph$. Compare os resultados;

g) Idêntico ao item anterior, mas para velocidade $x = 50mph$. Compare com os resultados do item anterior;

h) Apresente intervalos de confiança 95% para β_0 e para β_1;

i) Apresente intervalos de confiança 95% para o consumo médio e o consumo predito para um novo automóvel sob velocidades: (i) $x = 75mph$; (ii) $x = 50mph$;

j) Teste a significância do modelo de regressão, ou seja, teste a hipótese $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$;

k) Um especialista afirma que o consumo de combustível altera, em média, em $-0.18mpg$ para cada unidade a mais de velocidade (em mph). Teste essa hipótese.

Nota: Para as questões envolvendo testes de hipóteses, o seguinte procedimento deve ser aplicado:

(I) Formulação das hipóteses nula e alternativa;
(II) Apresentação e cálculo da estatística teste;
(III) Definição da regra de decisão para os níveis de significância de 5% e 1%;
(IV) Conclusão do problema baseada nas regras de decisão descritas no passo anterior;
(V) Cálculo do nível descritivo (p-valor) do teste.

14. Sejam y_1 e y_2 variáveis aleatórias com distribuição conjunta normal bivariada, conforme definido na sequência:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \text{Normal}\left(\boldsymbol{\mu} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}\right) \tag{1}$$

a) Quais as distribuições marginais de y_1 e y_2;

$y_1 \sim \mathrm{N}(6, 1)$ e $y_2 \sim \mathrm{N}(2, 0.5)$

b) Usando o R, faça gráficos das distribuições (funções densidade de probabilidade) de y_1, y_2 e da conjunta de y_1 e y_2;

c) Qual a distribuição de probabilidades de:

i) $z_1 = y_1 + y_2$;

ii) $z_2 = \frac{y_1+y_2}{2}$;

iii) $z_3 = y_1 - y_2$;

iv) $z_4 = 0.875y_1 - 2.278y_2$?

15. Sejam $y_1, y_2, ..., y_{30}$ variáveis aleatórias independentes com distribuição $y_i \sim \mathrm{N}(i, i^2)$, $i = 1, 2, ..., 30$. Qual a distribuição de $z = y_1 + y_2 + ... + y_{30}$?