SME0820 Modelos de Regressão e Aprendizado Supervisionado I: Lista 5

Thomas Peron

Não haverá provinha e tampouco entrega de exercícios desta lista; utilize-a apenas como treino para a P2.

1. (Sim, repita o exercício 1 da Lista 3. Certifique-se de que compreendeu todas passagens) Considere o modelo de regressão linear múltipla dado por

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

   (a) Escreva a forma matricial da Eq. (1), incluindo as suposições acerca dos erros ε_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 $\hat{\boldsymbol{\beta}}$. Mostre que $\hat{\boldsymbol{\beta}}$ é um estimador não viesado e calcule a sua matriz de variância.

   (c) Calcule o valor esperado e a matriz de variância da resposta ajustada, $\hat{\mathbf{Y}} = \mathbb{X}\hat{\boldsymbol{\beta}}$, onde $\mathbb{X}$ é a matriz com os valores das covariáveis, como definido nas aulas.

   (d) Explique como é definida a matriz $\mathbf{H}$. Mostre que $\mathbf{H}$ é idempotente e simétrica.

   (e) Seja $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ o vetor de resíduos. Expresse $\mathbf{e}$ em termos da matriz $\mathbf{H}$. 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 $\mathbf{X}_a = [1 \ X_{1a} \ X_{2a} \ \cdots X_{p-1,a}]^T$. Encontre o intervalo de $100(1-\alpha)\%$ de confiança de $\hat{Y}_a$ em termos de $\mathbf{X}_a$, $\mathbb{X}$ e MSE.

   (g) Seja Y_a uma nova observação feita para $\mathbf{X}_a = [1 \ X_{1a} \ X_{2a} \ \cdots X_{p-1,a}]^T$. Calcule o intervalo de predição de Y_a, com $100(1-\alpha)\%$ de confiança, em termos de $\mathbf{X}_a$, $\mathbb{X}$ e MSE.

2. Suponha que num problema de regressão linear múltipla haja duas variáveis preditoras mais o intercepto. Considere a matriz

$$\mathbb{X} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix},$$

   onde X_{ij} é a i-ésima amostra da j-ésima variável preditora. Escreva a matriz $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ em termos de $\overline{X_1^2} = \frac{1}{n}\sum_j^n X_{j1}^2$, $\overline{X_2^2} = \frac{1}{n}\sum_j^n X_{j2}^2$ e $\overline{X_1 X_2} = \frac{1}{n}\sum_j^n X_{j1}X_{j2}$.

3. Mostre que $\text{cov}[Y_i, \hat{Y}_i] = \sigma^2 H_{ii}$.

4. Você deve ajustar o modelo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$ pelo método dos mínimos quadrados para a situação em que β_2 é conhecido e dado por $\beta_2 = 5$. Como deve ser feito esse ajuste utilizando as ferramentas computacionais que discutimos em sala?

5. Para um problema de regressão linear múltipla com cinco variáveis preditoras, qual é a soma extra de quadrados relevante para testar se $\beta_5 = 0$? E para testar $\beta_2 = \beta_4 = 0$? Enuncie o teste F a ser feito para cada hipótese nula.

6. Mostre que

   (a) $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3)$;

   (b) $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3)$.

7. A tabela abaixo contém a comparação entre as métricas de todos os modelos lineares possíveis para um determinado conjunto de dados com quatro variáveis preditoras.

Num. de covars.	p	Variáveis no modelo	SSE	R^2	R^2_{adj}	MSE	C_p
0	1	Nenhuma	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.4578	122.7073	198.1
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.5
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.888	138.23
2	3	x_3x_4	175.738	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.5
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5

Tendo como base o que discutimos em sala, qual seria o modelo mais apropriado para ser utilizado numa aplicação com os dados que geraram a tabela?