2ª Lista de Exercícios - EST507 - Modelos de Regressão - 18/9/2024 - Entrega sugerida 8/10/2024

Exercício 1. Mostre que no modelo de regressão linear simples, os componentes do vetor de estimadores mínimos quadrados, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ coincidem com os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ apresentados nos livros de estatística básica.

Exercício 2. Seja $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma I_n)$ com $\mathbf{X}$ de posto completo $(k+1)$. Prove que

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \frac{X'X}{\sigma} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2_{k+1}.$$

Exercício 3. Sejam

$$\begin{aligned} y_1 &= \theta + \epsilon_1 \\ y_2 &= 2\theta - \gamma + \epsilon_2 \\ y_3 &= \theta + 2\gamma + \epsilon_3 \end{aligned}$$

onde $\mathrm{E}(\epsilon_i) = 0$, $i = 1, \ldots, 3$. Expresse este modelo como um modelo linear geral e determine os estimadores de mínimos quadrados de θ e γ.

Exercício 4. No modelo de regressão linear simples com as suposições usuais, determine o estimador de máxima verossimilhança de β_0 e β_1 sob a hipótese $H_0 : \beta_0 = 2$.

Exercício 5. Considere o modelo linear geral, em que $\boldsymbol{\epsilon} \sim N_n(0, \sigma\mathbf{I})$, $y_i = \beta_0 + \beta_1 x^\star_{i1} + \beta_2[x^\star_{i1}]^2 + \epsilon_i$ e $x^\star_{i1} = x_{i1} - \sum_{i=1}^n x_{i1}/n$.

(a) Determine o estimador de mínimos quadrados de β_0 e β_1 sob a hipótese $H_0 : \beta_2 = 1$.

(b) Determine o estimador de máxima verossimilhança de β_1 e β_2 sob a hipótese $H_0 : \beta_0 = 1$, assumindo que $\sigma^2 = 1$.

Exercício 6. Considere o modelo linear geral de posto completo com intercepto, em que $\boldsymbol{\epsilon} \sim N_n(0, \sigma\mathbf{I})$. Mostre que $\sum_{i=1}^n \epsilon_i^2 = \text{SQRes } + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Exercício 7. Deseja-se comparar dois métodos de ensino A e B e com esse objetivo, aplicou-se uma prova a n_1 alunos sujeitos ao método de ensino A e n_2 alunos sujeitos ao método de ensino B. Admitindo o modelo

$$y_i = \beta_0 + \beta_1 w_i + \epsilon_i,\ i = 1, \ldots, n$$

com ϵ_i e ϵ_j independentes para $i \neq j$, $\epsilon_i \sim N(0, \sigma)$ e

$$w_i = \begin{cases} 0 \text{ para o método A e} \\ 1 \text{ para o método B} \end{cases}$$

e y_i a nota da prova do i-ésimo aluno.

(a) Qual o significado de β_0 e β_1?

(b) Obtenha o ENVVUM da diferença das notas médias para alunos sujeitos aos métodos A e B, justificando sua resposta.

(c) Construa um teste de hipóteses para a igualdade das notas médias seguindo os dois métodos, contra a hipótese alternativa de que as médias são diferentes. Utilize a forma geral $\mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

(d) Prove que o teste do item (c) é equivalente ao teste t de igualdade de médias para amostras independentes.

Exercício 8. Considere o modelo

$$\begin{aligned} y_1 &= \theta_0 + \epsilon_1 \\ y_2 &= \theta_0 + \theta_1 + \epsilon_2 \\ y_3 &= \theta_0 + \theta_2 + \epsilon_3 \\ y_4 &= \theta_0 + \theta_1 + \theta_2 + \epsilon_4 \end{aligned}$$

com $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)' \sim N(0, \sigma \mathbf{I})$.
Obtenha o teste para avaliar se $H_0 : \theta_1 = \theta_2$ contra $H_1 : \theta_1 \neq \theta_2$.

Exercício 9. Seja $\mathbf{X} = (, \mathbf{X}_1)$ e $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\top)^\top$. Mostre que no modelo linear geral de posto completo, em que $\boldsymbol{\epsilon} \sim N_n(0, \sigma \mathbf{I})$, então

$$\text{E(SQReg}/k) = \sigma^2 + (1/k)\boldsymbol{\beta}_1^\top \mathbf{X}_c^\top \mathbf{X}_c \boldsymbol{\beta}_1$$

em que $\mathbf{X}_c = (\mathbf{I} - (1/n)^\top)\mathbf{X}_1$.

Exercício 10. Sendo $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, demonstre o teorema que apresenta a distribuição da estatística do teste sob H_0 e sob H_1 para o teste que tem como hipóteses $\mathrm{H}_0 : \boldsymbol{\beta}_2 = 0$ e $\mathrm{H}_1 : \boldsymbol{\beta}_2 \neq 0$ no modelo linear geral de posto completo com intercepto em que $\boldsymbol{\epsilon} \sim N_n(0, \sigma \mathbf{I})$.

Exercício 11. Considere o modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ com $\boldsymbol{\epsilon} \sim N_n(0, \sigma \mathbf{I})$ e $\mathbf{X}$ de posto completo $k + 1$. Se $\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, prove que

(i) $\text{Cov}(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) = 0$,
(ii) $\text{Cov}(\boldsymbol{\epsilon}, \widehat{\boldsymbol{\beta}}) = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\sigma$,
(iii) $\text{Cov}(\widehat{\boldsymbol{\epsilon}}, \mathbf{y}) = [\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top]\sigma$,
(iv) $\text{Cov}(\widehat{\boldsymbol{\epsilon}}, \widehat{\mathbf{y}}) = 0$.

Com base nos itens (iii) e (iv), discuta por que é preferível conduzir análise de diagnóstico do modelo a partir da análise gráfica de $\widehat{\boldsymbol{\epsilon}} \times \widehat{\mathbf{y}}$ em relação à análise gráfica de $\widehat{\boldsymbol{\epsilon}} \times \mathbf{y}$.

Exercício 12. Quando a gasolina é colocada em um tanque de um carro, vapores são lançados na atmosfera. Um experimento foi conduzido para verificar se a quantidade de vapor lançada na atmosfera (y) pode ser explicada usando as seguintes 4 variáveis relacionadas com o tanque e a gasolina: x_1: temperatura do tanque (^oF), x_2: temperatura da gasolina (^oF), x_3: pressão do vapor no tanque (psi), x_4: pressão da gasolina no tanque (psi). Os dados estão disponíveis no arquivo DadRegLis2.xls. Assuma que o modelo linear geral com $\boldsymbol{\epsilon} \sim N_n(0, \sigma \mathbf{I})$ e $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$ é adequado para esses dados. Nos itens que são necessários a realização de teste de hipóteses ou intervalo de confiança, assuma $\alpha = 0,05$

(a) Construa diagramas de dispersão envolvendo cada uma das variáveis preditoras e a variável resposta.
(b) Obtenha estimativas de $\boldsymbol{\beta}$ e σ^2 e interprete as estimativas dos componentes de $\boldsymbol{\beta}$ em termos do problema em questão. As estimativas estão de acordo com o que era esperado pela análise dos diagramas de dispersão? Caso não estejam, qual é o motivo disso?
(c) Obtenha estimativa de $\text{cov}(\hat{\beta})$.
(d) Encontre o coeficiente de determinação do modelo e interprete-o.
(e) Sendo $\beta = (\beta_0^\top, \boldsymbol{\beta}_1^\top)^\top$, teste $\mathrm{H}_0 : \boldsymbol{\beta}_1 = 0 \times \mathrm{H}_1 : \boldsymbol{\beta}_1 \neq 0$. Qual a conclusão do teste em termos do problema?

(f) Sendo $\beta = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ e $\boldsymbol{\beta}_2 = (\beta_1, \beta_3)$, teste $\mathrm{H}_0 : \boldsymbol{\beta}_2 = 0 \times \mathrm{H}_1 : \boldsymbol{\beta}_2 \neq 0$. Qual a conclusão do teste em termos do problema?

(g) Para cada $j = 1, 2, 3, 4$, teste $\mathrm{H}_0 : \beta_j = 0 \times \mathrm{H}_1 : \beta_j \neq 0$. Qual a conclusão dos testes em termos do problema? Repita os testes considerando algum método de controle do nível de significância global do teste. As conclusões são as mesmas?

(h) Obtenha o poder do teste $\mathrm{H}_0 : \beta_2 = 0 \times \mathrm{H}_1 : \beta_2 \neq 0$, para $\beta_2 = 0{,}5$, supondo $\sigma^2 = 7{,}5$. O que significa o valor obtido em termos do problema em questão?

(i) Teste $\mathrm{H}_0 : \beta_2 = 0 \times \mathrm{H}_1 : \beta_2 > 0$. Qual a conclusão dos testes em termos do problema?

(j) Conduza testes de hipóteses com as seguintes hipóteses nulas: $\mathrm{H}_{01} : \beta_1 = \beta_2 = 12\beta_3 = 12\beta_4$, $\mathrm{H}_{02} : \beta_1 = \beta_2$, $\mathrm{H}_{03} : \beta_2 = 12\beta_3$, $\mathrm{H}_{04} : \beta_3 = \beta_4$ e $\mathrm{H}_{05} : \beta_1 = \beta_2$ e $\beta_3 = \beta_4$.

(k) Para cada $j = 1, 2, 3, 4$, construa intervalos de confiança para β_j usando algum método de controle do coeficiente de confiança global.

(l) Obtenha intervalos de confiança para $\mathrm{E}(y_0)$ e para uma nova observação y_0 se $\mathbf{x}_0 = (1, 60, 60, 4, 4)^\top$. O que significam esses intervalos em termos do problema em questão?

(m) Proponha um modelo para ajustar a quantidade de vapor lançada na atmosfera (não incluindo necessariamente todas as variáveis preditoras disponíveis), interprete as estimativas dos parâmetros e realize adequados testes de hipóteses para esse modelo.

(n) Realize análise de diagnóstico do modelo que você propôs no item anterior. O modelo parece ser adequado para ajustar a quantidade de vapor lançada na atmosfera?

Exercício 13. Em um experimento para investigar o efeito da cor do papel na taxa de resposta de questionários colocado em parabrisas de carro estacionados em supermercados, 15 estacionamentos de supermercados foram escolhidos em uma área metropolitana e cada cor foi atribuída aleatoriamente para cinco estacionamentos. Os dados estão disponíveis no arquivo DadRegLis2.xls.

(a) Proponha um modelo que assuma que o efeito do tamanho do estacionamento na taxa média de resposta dos questionários varia de acordo com a cor do papel e obtenha as estimativas dos parâmetros desse modelo.

(b) No modelo proposto no item anterior, teste a hipótese de que o efeito do tamanho do estacionamento na taxa média de resposta dos questionários varia de acordo com a cor do papel. Assuma $\alpha = 5\%$.

(c) Ajuste um modelo que assuma que o efeito do tamanho do estacionamento na taxa média de resposta dos questionários não varia de acordo com a cor do papel e interprete as estimativas dos parâmetros.

(d) Baseado no modelo do item anterior e considerando $\alpha = 5\%$, há evidências de que a taxa média de resposta dos questionários varia em função do tamanho do estacionamento? E em função da da cor do papel?

(f) Você escolheria para ajustar a taxa média de resposta dos questionários, o modelo do item (a) ou o modelo do item (c)? Justifique sua resposta.

(g) Baseado no modelo do item (c), teste a hipótese de que a taxa média de resposta dos questionários é a mesma se a cor é laranja ou verde? Assuma $\alpha = 5\%$.

(h) Obtenha um intervalo de confiança para a taxa média de resposta e um intervalo de previsão para uma nova observação quando a cor do papel é laranja e o tamanho do estacionamento é 300. Assuma $1 - \alpha = 95\%$

(i) Realize uma análise de diagnóstico completa do modelo por você escolhido no item (f) e discuta as conclusões.

Exercício 14. Está sendo estudado o consumo de gasolina de um determinado carro em função da velocidade. O consumo do carro foi avaliado em um determinada viagem realizada 12 vezes com diferentes velocidades médias. Os dados estão disponíveis no arquivo DadRegLis2.xls, em que são apresentadas a velocidade média em milhas e o número de milhas percorridas com cada galão de gasolina.

(a) Faça um diagrama de dispersão entre a variável resposta e a variável preditora e proponha um modelo para este problema.
(b) Estime os parâmetros do modelo proposto, calcule o coeficiente de determinação do modelo e interprete-o.
(c) Teste se há ou não relação entre número de milhas percorridas com cada galão de gasolina e a velocidade média. Assuma $\alpha = 1\%$.
(c) Teste se algum dos parâmetros da curva de regressão pode ser excluído do modelo. Assuma $\alpha = 10\%$.
(d) Construa intervalos de confiança para os parâmetros do modelo. Assuma $1 - \alpha = 90\%$.
(e) Obtenha um intervalo de previsão para uma nova observação da variável resposta em que a velocidade média será de 100 milhas. O intervalo obtido é razoável? Por que não?
(f) Realize uma análise de diagnóstico completa do modelo proposto e discuta as conclusões.

Exercício 15. Um pesquisador está estudando a quantidade de chumbo no solo de determinada região. O objetivo dele é verificar se a quantidade média de chumbo no solo varia de acordo com o tipo de área da região: industrial, residencial e rural. Ele então coletou 300 gramas de solo em 5 diferentes locais de cada um dos tipos de áreas da região. Os dados estão disponíveis no arquivo DadRegLis2.xls.

(a) Calcule medidas descritivas da quantidade de chumbo por tipo de área. O que a análise descritiva parece sugerir?
(b) Proponha um modelo, obtenha as estimativas dos parâmetros desse modelo e interprete-as?
(c) Teste se a quantidade média de chumbo varia em função do tipo de área. Assuma $\alpha = 1\%$.
(d) Caso você tenha rejeitado a hipótese nula no item anterior, realize comparações múltiplas e discuta as conclusões considerando um nível de significância global de 1%.
(e) Construa intervalos de confiança com coeficiente de confiança global de 99% para as diferenças entre as áreas (duas a duas) em relação à quantidade média de chumbo no solo.
(f) Realize uma análise de diagnóstico completa do modelo proposto e discuta as conclusões.

Exercício 16. Um pesquisador está estudando o efeito do tipo de pimentão (nos níveis amarelo, verde e vermelho) e do tipo de cultura (nos níveis orgânico e convencional) na quantidade média de alguns nutrientes. Um dos nutrientes analisados foi a vitamina C. Para cada combinação entre um nível do tipo de pimentão e um nível do tipo de cultura, ele coletou entre 4 e 5 pimentões e mediu a quantidade de vitamina C presente no mesmo.

(a) Proponha um modelo, obtenha as estimativas dos parâmetros desse modelo e interprete-as?
(b) Conduza os testes de hipóteses que seriam de interesse do pesquisador controlando, quando necessário, o nível de significância global em 5%. Discuta os resultados obtidos.

(c) Obtenha um intervalo de confiança para a diferença entre o tipo de cultura orgânico e convencional quando o pimentão é verde em relação a quantidade média de vitamina C. Assuma $1 - \alpha = 95\%$.

(d) Realize uma análise de diagnóstico completa do modelo proposto e discuta as conclusões.