

Segunda avaliação (A2)

Disciplina: Inferência Estatística
Instrutor: Professor Carvalho

03 de Dezembro de 2020

- Por favor, entregue um único arquivo PDF;
- O tempo para realização da prova é de 3 horas, mais vinte minutos para upload do documento para o e-class;
- Leia a prova toda com calma antes de começar a responder;
- Responda todas as questões sucintamente;
- Marque a resposta final claramente com um quadrado, círculo ou figura geométrica de sua preferência;
- A prova vale 80 pontos. A pontuação restante é contada como bônus;
- Apenas tente resolver a questão bônus quando tiver resolvido todo o resto;
- Lembre-se de consultar o catálogo de fórmulas no fim deste documento.

Dicas

- Em uma regressão linear simples, temos:

$$\hat{\beta}_0 \sim \text{Normal} \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right) \right),$$

$$\hat{\beta}_1 \sim \text{Normal} \left(\beta_1, \frac{\sigma^2}{s_x^2} \right),$$

$$\text{Cov} \left(\hat{\beta}_0, \hat{\beta}_1 \right) = -\frac{\bar{x}\sigma^2}{s_x^2},$$

onde $s_x = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$ e $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de máxima verossimilhança dos coeficientes.

- Um processo de Poisson com taxa λ por unidade de tempo é um processo estocástico que satisfaz:
 - O número de chegadas em um intervalo de tempo Δ_t tem distribuição Poisson com média $\lambda\Delta_t$.
 - Os números de chegadas em qualquer coleção de intervalos disjuntos são independentes.
- Se X tem distribuição Poisson com média λ , então

$$\Pr(X \leq x) = Q(\lfloor x + 1 \rfloor, \lambda),$$

onde

$$Q(x, s) = \frac{\Gamma(x, s)}{\Gamma(x)}$$

é a função Gama regularizada superior e $\lfloor y \rfloor$ é maior inteiro menor ou igual a y – também chamado de *floor*. Ademais, temos

$$\frac{\partial}{\partial s} Q(x, s) = -\frac{e^{-s} s^{x-1}}{\Gamma(x)},$$

onde $\Gamma(x) = (x-1)!$ é a função Gamma.

1. Catando poesia.

*Eu te vejo sumir por aí
Te avisei que a cidade era um vão
Dá tua mão, olha pra mim
Não faz assim, não vai lá, não
Os letreiros a te colorir
Embaraçam a minha visão
Eu te vi suspirar de aflição
E sair da sessão frouxa de rir
Já te vejo brincando gostando de ser
Tua sombra a se multiplicar
Nos teus olhos também posso ver
As vitrines te vendo passar
Na galeria, cada clarão
É como um dia depois de outro dia
Abrindo um salão
Passas em exposição
Passas sem ver teu vigia
Catando a poesia
Que entornas no chão*

As Vitrines (Almanaque, 1981) de Chico Buarque (1944-).

O eu-lírico da canção, que vamos chamar aqui de Ivo, pensa em seu amado, Adão. Adão é poeta, e tem a estranha mania de deixar cair seus poemas ao passear pelo shopping. Ivo, muito solícito e perdidamente apaixonado, corre atrás do companheiro catando os papéis que o desastrado deixa cair. Sendo estatístico, Ivo sabe que pode modelar o processo de queda dos poemas como um processo de Poisson com média θ . Ivo quer saber se será capaz de acompanhar Adão na sua jornada sem perder nenhum poema. Para isso, julga que se $\theta \leq \theta_0$, ele será capaz de catar toda a poesia deixada por Adão antes de ser carregada pelo vento.

Suponha que Ivo observa o processo de queda dos poemas em n intervalos de exatamente t unidades de tempo e toma nota dos números Y_1, Y_2, \dots, Y_n de poemas caídos em cada intervalo. Ivo considera a estatística de teste $S = \sum_{i=1}^n Y_i$ e constrói o teste δ_c de modo que, se $S \geq c$, ele rejeita a hipótese $H_0 : \theta \leq \theta_0$.

- (10 pontos) Encontre a função poder do teste de Ivo.
- (10 pontos) Mostre que a função poder do item anterior é **não-decrescente** em θ ;
- (2,5 pontos) Encontre uma expressão para o tamanho α_0 do teste δ_c ;
- (2,5 pontos) O teste em questão é não-viesado? Justifique;
- (5 pontos) Discuta se é possível atingir qualquer tamanho para δ_c e o que fazer se queremos um tamanho de, por exemplo, $\alpha_0 = 0.01$.

Conceitos trabalhados: Testes de hipótese; poder; tamanho.

Nível de dificuldade: fácil.

Resolução: A partir das dicas, sabemos que $Y_i \sim \text{Poisson}(\theta t)$, de modo que $S \sim \text{Poisson}(n\theta t)$. Portanto, a probabilidade de rejeitar H_0 é:

$$\pi(\theta \mid \delta_c) := \Pr(S \geq c \mid \theta) = \sum_{k=c}^{\infty} \frac{e^{-n\theta t} (n\theta t)^k}{k!} = 1 - Q(\lfloor c+1 \rfloor, n\theta t).$$

Isto responde a). Utilizando a outra dica dada no começo da prova, vemos que a derivada

$$\begin{aligned} \frac{d\pi(\theta \mid \delta_c)}{d\theta} &= \frac{d}{d\theta} [1 - Q(\lfloor c+1 \rfloor, n\theta t)], \\ &= - \left(-nt \frac{e^{-n\theta t} (n\theta t)^{\lfloor c+1 \rfloor - 1}}{\Gamma(\lfloor c+1 \rfloor)} \right) = nt \frac{e^{-n\theta t} (n\theta t)^c}{\Gamma(\lfloor c+1 \rfloor)} \end{aligned}$$

é não-negativa para todo $\theta \in (0, \infty)$, o que responde b). A solução para c) é que sim, o teste é não-viesado: como a função poder é não-decrescente em θ temos que $\pi(\theta \mid \delta_c) \leq \pi(\theta' \mid \delta_c)$ para todo par (θ, θ') tal que $\theta \leq \theta_0$ e $\theta' > \theta_0 \geq \theta$, isto é, $\theta \in \Omega_0$ e $\theta' \in \Omega_1$. Para responder d), vamos nos lembrar que

$$\text{tamanho}(\delta) := \sup_{\theta \in \Omega_0} \pi(\theta \mid \delta).$$

Como $\pi(\theta \mid \delta_c)$ é não decrescente e $\Omega_0 = (0, \theta_0]$, temos:

$$\text{tamanho}(\delta_c) = \sup_{\theta \leq \theta_0} \pi(\theta \mid \delta_c) = 1 - Q(\lfloor c+1 \rfloor, n\theta_0 t) =: \alpha_0.$$

Olhando para a expressão que acabamos de encontrar, podemos ver que não será sempre possível inverter a relação encontrada para, a partir de um par (α_0, θ_0) , obter c que satisfaça a expressão exatamente. Para construir um teste com $\alpha_0 = 0.01$, precisamos encontrar c de modo que $\pi(\theta \mid \delta_c) \leq 0.01$, o que pode levar a tamanhos efetivos bem menores que o especificado. Com isso, respondemos e). Como extra, poderíamos citar a aleatorização aprendida no Trabalho IV como uma maneira de atingir α_0 exatamente. ■

2. Temos que pegar!

Além de apaixonados um pelo outro, Joelinton e Valcicléia também amam Pokémon. Os dois jogam competitivamente na Liga Brasileira de Pokémon (LBP). Há, contudo, um pequeno inconveniente: Joelinton é *Team Magma* enquanto Valcicléia é *Team Aqua*. Durante uma conversa acalorada, Joelinton afirma que o *Team Magma* é melhor, em termos de *pokescores* médios, que o *Team Aqua*. Valcicléia propõe consultar o site da LBP para obter dados sobre o assunto. Ao consultar o site, eles obtêm m valores de *pokescores* de integrantes do *Team Magma* e n valores de integrantes do *Team Aqua*.

Suponha que modelamos os *pokescores* de cada jogador(a) em cada time como variáveis aleatórias normais com médias μ_M e μ_A e variâncias σ_M^2 e σ_A^2 , respectivamente. Nos itens a seguir, **enuncie claramente** qual é a hipótese nula – e a hipótese alternativa – em cada caso, qual é a estatística de teste e qual o procedimento de teste.

- a) (2,5 pontos) A partir do desenho experimental descrito, encontre quantidades pivotais para μ_M e μ_A , supondo σ_M^2 e σ_A^2 **desconhecidas**. Justifique;
- b) (2,5 pontos) Utilize as quantidades do item anterior para construir intervalos de confiança exatos de 99% para μ_A e μ_M ;
- c) (5 pontos) Suponha que, no calor do momento, Valcicléia afirme que o *Team Magma* é tão ruim que não tem pokescore médio suficiente nem para competir na Liga Regional de Pokemon (LRP). Sabendo que o pokescore médio necessário para admissão na LRP é μ_0 , mostre a Joelinton como utilizar o intervalo de confiança obtido no item anterior para testar a hipótese levantada por sua amada;
- d) (10 pontos) Nossa dupla dinâmica está interessada em comparar as médias supondo que as variâncias são iguais. Proponha um teste de tamanho α_0 para avaliar a premissa de homogeneidade (variâncias iguais);
- e) (10 pontos) Suponha que o teste do item anterior falhou em rejeitar H_0 . Proponha um teste de tamanho α_0 para testar a hipótese inicial de Joelinton;

Conceitos trabalhados: Quantidades pivotais; intervalos de confiança; comparação de médias; comparação de variâncias.

Nível de dificuldade: médio.

Resolução: Vamos começar definindo $\mathbf{X} = \{X_1, X_2, \dots, X_m\}$ como sendo os pokescores do *Team Magma* e $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$ como os pokescores do *Team Aqua*. Como vimos em aula, as quantidades pivotais aqui são

$$U_M = \frac{\bar{X}_m - \mu_M}{\hat{\sigma}'_M} \sim \text{Student}(m-1),$$

$$U_A = \frac{\bar{Y}_n - \mu_A}{\hat{\sigma}'_A} \sim \text{Student}(n-1),$$

onde $\hat{\sigma}'_M$ e $\hat{\sigma}'_A$ são dados pelas fórmulas no catálogo. A partir destas quantidades (estatísticas) pivotais, conseguimos responder b): os intervalos

$$I_M = \bar{X}_m \pm T^{-1}(0,995; m-1) \frac{\hat{\sigma}'_M}{\sqrt{m}},$$

$$I_A = \bar{Y}_n \pm T^{-1}(0,995; n-1) \frac{\hat{\sigma}'_A}{\sqrt{n}},$$

são intervalos de confiança exatos para μ_M e μ_A , respectivamente. Depois de respirar fundo e beber um copo d'água, Joelinton pode resolver c) e testar a hipótese aventada por Valcicléia simplesmente avaliando se

$$\bar{X}_m + T^{-1}(0,995; m-1) \frac{\hat{\sigma}'_M}{\sqrt{m}} < \mu_0.$$

Caso afirmativo, podemos rejeitar a hipótese $H_0 : \mu_M \geq \mu_0$ ao nível $\alpha = 0.01$. Para responder d) precisamos lembrar que o teste para a premissa de homogeneidade é o teste F. Nossa hipótese nula é $H_0 : \sigma_M^2 = \sigma_A^2$ e nossa estatística de teste é

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}.$$

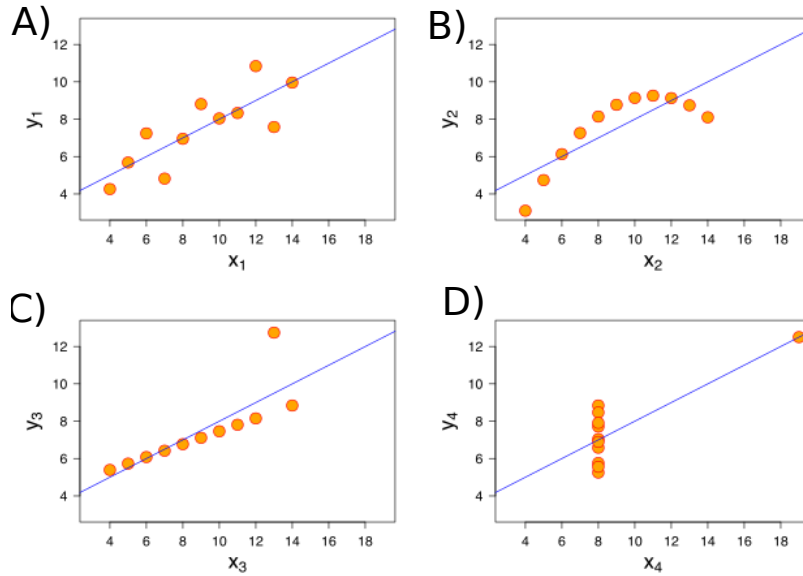
Sabemos que sob H_0 , $V \sim F(m-1, n-1)$. Vamos agora construir um teste de tamanho α_0 . Como nossa hipótese alternativa $H_0 : \sigma_M^2 \neq \sigma_A^2$ é bilateral, definimos uma região de rejeição $R = (0, c_1) \cup (c_2, \infty)$ para V , onde $c_1 = F^{-1}(\alpha_0/2; m-1, n-1)$ e $c_2 = F^{-1}(1-\alpha_0/2; m-1, n-1)$ são os quantis apropriados de uma distribuição F. Nosso procedimento de teste é “se $V \in R$, rejeitamos H_0 , caso contrário falhamos em rejeitar H_0 ”. Para finalizar, vamos responder e). Se não rejeitamos a hipótese de diferença nas variâncias, podemos empregar um teste t de Student unilateral para testar $H_0 : \mu_M \leq \mu_A$ sob a premissa de variâncias iguais. Para isso, computamos

$$U = \frac{\sqrt{m+n-2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{(\frac{1}{m} + \frac{1}{n})(S_X^2 + S_Y^2)}},$$

que, sob H_0 , tem distribuição t de Student com $m+n-2$ graus de liberdade. Nosso procedimento de teste é rejeitar H_0 se $U \geq c$, onde $c = T^{-1}((1-\alpha_0)/2; m+n-2)$ é o quantil apropriado de uma distribuição t de Student com $m+n-2$ graus de liberdade. ■

3. Regressão linear: o melhor modelo ruim que você já viu.

Considere a figura a seguir:



Em todos os painéis, $\bar{x} = 9$, $\bar{y} = 7,5$, $s_x^2 = 110$ e $\text{Cor}(X, Y) = 0,816$. Isto é, todas as estatísticas sumárias relevantes atingem os mesmos valores. Disto resulta que $\hat{\beta}_0 = 3$ e $\hat{\beta}_1 = 0,5$ para todos os painéis.

- a) (15 pontos) Comente sobre quais premissas básicas – ou nenhuma – da regressão linear aparentam estar sendo violadas em cada painel. Justifique.

- b) (5 pontos) Os estimadores de máxima verossimilhança para os coeficientes no modelo

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

são

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{y})(X_i - \bar{x})}{\sum_{i=1}^n (X_i - \bar{x})^2}.\end{aligned}$$

Tais estimadores são viesados? Justifique.

Dica: Pode ser conveniente escrever

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{x}) Y_i}{s_x^2}.$$

Conceitos trabalhados: Regressão linear; premissas da regressão linear; viés.
Nível de dificuldade: médio.

Resolução: Vamos responder a) por painel:

- No painel A) não é possível identificar visualmente nenhuma violação clara das premissas do modelo de regressão;
- Já no painel B), podemos notar que a relação entre Y e X é não-linear, violando a premissa de linearidade, $E[Y] = \beta_0 + \beta_1 X$;
- Em C) vemos um ponto aberrante, o que sugere violação da premissa de dados condicionalmente independentes e identicamente distribuídos (i.i.d);
- Finalmente em D) vemos que as premissas de linearidade e i.i.d. não parecem ser atendidas pois o ponto extremo à direita (chamado comumente de “alavanca”) cria uma relação linear artificial entre Y e X , que desapareceria se excluíssemos esse ponto.

Para responder b) temos dois caminhos: o mais simples é utilizar a dica dada no começo da prova e argumentar que como as distribuições de $\hat{\beta}_0$ e $\hat{\beta}_1$ são normais com média β_0 e β_1 , os estimadores considerados são de fato não-viesados. O segundo caminho, e mais complicado, é utilizar manipulações de esperanças, o que faremos a seguir. Lembrando que $E[Y_i] = \beta_0 + \beta_1 X_i$, usando a dica dada na própria questão e passando o operador de esperança, temos

$$\begin{aligned}E[\hat{\beta}_1] &= \frac{\sum_{i=1}^n (X_i - \bar{x}) E[Y_i]}{s_x^2}, \\ &= \frac{\sum_{i=1}^n (X_i - \bar{x}) (\beta_0 + \beta_1 X_i)}{s_x^2}, \\ &= \frac{\beta_0 \sum_{i=1}^n (X_i - \bar{x}) + \beta_1 \sum_{i=1}^n X_i (X_i - \bar{x})}{\sum_{i=1}^n (X_i - \bar{x})^2}.\end{aligned}$$

Como $\sum_{i=1}^n (X_i - \bar{x}) = 0$, temos que

$$\begin{aligned}E[\hat{\beta}_1] &= \frac{\beta_1 \sum_{i=1}^n X_i (X_i - \bar{x})}{\sum_{i=1}^n (X_i - \bar{x})^2}, \\ &= \beta_1,\end{aligned}$$

porque¹ $\sum_{i=1}^n X_i (X_i - \bar{x}) = \sum_{i=1}^n (X_i - \bar{x})^2$. Para $\hat{\beta}_0$, temos

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{y}] - \beta_1 \bar{x}, \\ &= \frac{\sum_{i=1}^n \beta_0 + \beta_1 X_i}{n} - \beta_1 \bar{x}, \\ &= \beta_0 + \frac{\sum_{i=1}^n \beta_1 X_i}{n} - \beta_1 \bar{x}, \\ &= \beta_0. \end{aligned}$$

Nota: A questão b) foi retirada *ipsis literis* dos exercícios 2 e 3 da seção 11.2 de DeGroot (recomendados!).

Questão Bônus: uma transformação útil.

Muitas vezes na aplicação de modelos de regressão é conveniente aplicar uma transformação à(s) variável(is) independente(s) de modo a facilitar a computação e/ou a interpretação das estimativas.

- a) (10 pontos) Considere uma regressão linear simples. Encontre uma transformação $X' = f(X)$ da variável independente de modo que $\hat{\beta}'_0$ e $\hat{\beta}'_1$ sejam independentes.
- b) (5 pontos) Encontre o valor de $\hat{\beta}'_0$ e $\hat{\beta}'_1$ sob a transformação do item anterior.
- c) (5 pontos) Como essa transformação muda a interpretação dos coeficientes estimados?

Conceitos trabalhados: Transformação de covariáveis; interpretação dos coeficientes.

Nível de dificuldade: difícil.

Resolução: Resposta de a): a partir da dica dada no começo da prova, sabemos que

$$\text{Cov}(\hat{\beta}'_0, \hat{\beta}'_1) = -\frac{\bar{x}'\sigma^2}{s_x^2}.$$

Portanto, podemos considerar a transformação $X'_i = X_i - \bar{x}$, de modo que $\bar{x}' = 0$. Chamamos este procedimento de “centrar” (em inglês, *centering*) o preditor ou variável independente. Disto, temos que $\text{Cov}(\hat{\beta}'_0, \hat{\beta}'_1) = 0$ e, portanto, $\hat{\beta}'_0$ e $\hat{\beta}'_1$ são não-correlacionadas. Como a distribuição conjunta de $\hat{\beta}'_0$ e $\hat{\beta}'_1$ é bivariada normal (fato deduzido a partir dica), segue que $\hat{\beta}'_0$ e $\hat{\beta}'_1$ são independentes. Note que essa última conclusão se aplica somente a variáveis aleatórias normais, pois **neste caso**, correlação zero implica independência. Para responder b) a respeito de $\hat{\beta}'_0$ basta aplicar a fórmula dada para encontrar

¹Prove isto, se quiser.

$\hat{\beta}'_0 = \bar{y}$. Já para $\hat{\beta}'_1$, precisamos notar que como $X'_i = X_i - \bar{x}$,

$$\begin{aligned}\hat{\beta}'_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{y})(X'_i - \bar{x}')}{\sum_{i=1}^n (X'_i - \bar{x}')^2}, \\ &= \frac{\sum_{i=1}^n (Y_i - \bar{y})([X_i - \bar{x}] - 0)}{\sum_{i=1}^n ([X_i - \bar{x}] - 0)^2}, \\ &= \hat{\beta}_1,\end{aligned}$$

ou seja, a estimativa do coeficiente angular não muda!

Sobre c) vimos que $\hat{\beta}'_0 = \bar{y}$, isto é, o intercepto é a média da variável dependente. Isso significa que β_0 pode ser entendido como a média da variável dependente quando a variável independente (a original!) atinge sua média. Isso facilita bastante a interpretação, especialmente em situações em que X nunca atinge zero em sua escala natural (por exemplo, quando X é a altura de um indivíduo). A interpretação de β_1 muda muito pouco, já que X' mantém as mesmas unidades que X . Além disso, como vimos em b), $\beta'_1 = \hat{\beta}_1$. ■

Fórmulas úteis

Como usar este catálogo: as fórmulas dadas aqui estão propositalmente privadas do seu contexto. O objetivo desta coleção é ajudar você a lembrar das expressões. Entretanto, saber quais expressões são utilizadas em que contexto é sua tarefa.

- $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i;$
- $\hat{\sigma}' = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2};$
- $S_X^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2;$
- $S_Y^2 = \sum_{j=1}^n (Y_j - \bar{Y}_n)^2;$
- $U = \frac{\sqrt{m+n-2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{(\frac{1}{m} + \frac{1}{n})(S_X^2 + S_Y^2)}};$
- $V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)};$
- $\bar{x} = (1/n) \sum_{i=1}^n X_i;$
- $\bar{y} = (1/n) \sum_{i=1}^n Y_i.$