Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
Gustavo L. Gilardoni

Lista de Exercícios 3

1. Para ajustar o modelo $y_i = \alpha + \beta\, x_i + \epsilon_i$ $(i = 1, \ldots, 20)$ foi calculado que $\sum_1^{20} x_i = -2.52$, $\sum_1^{20} x_i^2 = 20.86$, $\sum_1^{20} y_i = 51.59$, $\sum_1^{20} y_i^2 = 742.48$ e $\sum_1^{20} x_i\, y_i = 9.10$.

(i) Calcule as estimativas de MQ de α e β e a estimativa de σ^2; **(ii)** Assumindo normalidade, construa intervalos com confiança 95% para α, β e σ^2; **(iii)** Teste ao nível de significância 5% a hipótese nula que $\beta = 0$ contra a alternativa que $\beta \neq 0$ (calcule o p-valor); **(iv)** construa a tabela de análise de variância e calcule o coeficiente de determinação R^2.

2. A seguinte tabela de análise de variância foi construída a partir do ajuste de um modelo de regressão linear simples para duas variáveis y_i e x_i.

Fonte de variação	**gl**	**Soma de quadrados**	**Quadrados médios**	F_{obs}
Regressão	a	b	c	d
Resíduos	e	1368.00	f	
Total	21	1474.74		

(i) Calcule a, b, c, d, e e f; **(ii)** Quantas indivíduos foram observados no experimento?; **(iii)** Com base nessa tabela, existe evidência para concluir que o coeficiente da regressão β é diferente do que 0? (Use $\alpha = 0.10$ e calcule também o p-valor do teste); **(iv)** Calcule o coeficiente de determinação R^2; **(v)** Sabendo que $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 5.67$, calcule o valor absoluto da estimativa de MQ $\hat{\beta}$ (sugestão: qual é a relação entre o F_{obs} e o estatístico t_{obs} para testar se $\beta \neq 0$?)

3. A tabela seguinte mostra as notas no vestibular e o IRA durante o primeiro ano de $n = 20$ alunos do curso de estatística.

aluno	1	2	3	4	5	6	7	8	9	10
vest.	66	62	66	56	93	66	91	53	60	81
IRA	63.5	53.7	55.3	79.7	73.4	72.6	84.0	55.0	67.7	60.4

aluno	11	12	13	14	15	16	17	18	19	20
vest.	57	58	55	65	72	75	64	67	67	73
IRA	58.3	38.9	71.3	70.0	71.4	81.9	68.8	72.6	70.3	77.1

Considere a seguir o modelo linear $\text{IRA}_i = \alpha + \beta\,(\text{vestibular})_i + \epsilon_i$, com os erros ϵ_i independentes e normalmente distribuídos com média 0 e variância σ^2.

(i) Faça um gráfico de dispersão das variáveis `vestibular` e `IRA`. Esse gráfico sugere que um modelo linear poderia explicar o rendimento dos alunos durante o primeiro ano? **(ii)** Calcule as estimativas de MQ de α e β e a estimativa de σ^2; **(iii)** Assumindo normalidade, construa intervalos com confiança 95% para α, β e σ^2; **(iv)** Teste ao nível de significância 5% a hipótese nula que $\beta = 0$ contra a alternativa que $\beta \neq 0$ (calcule o p-valor); **(v)** construa a tabela de análise de variância; **(vi)** construa um IC com nível 95% para o IRA médio dos alunos que obtêm nota 70 no vestibular; **(vii)** construa um intervalo de previsão para o IRA de um único aluno que obtém nota 70 no vestibular. **(viii)** qual dos intervalos (v) e (vi) é mais comprido? Explique; **(ix)** Faça gráficos dos resíduos do ajuste que permitam avaliar se os supostos de homocedasticidade e de normalidade são adequados para estes dados e interprete.

4. Considere o modelo de regressão sem intercepto $y_i = \beta\, x_i + \epsilon_i$ $(i = 1, \ldots, n)$, onde os erros ϵ_i são independentes e seguem uma distribuição Normal com média 0 e variância σ^2. Mostre que: **(i)** O estimador de β que minimiza a soma de quadrados $S(\beta) = \sum_{i=1}^n (y_i - \beta\, x_i)^2$ é $\hat\beta = \sum_{i=1}^n x_i\, y_i / \sum_{i=1}^n x_i$; **(ii)** $\mathrm{E}(\hat\beta) = \beta$ (i.é. esse estimador é não-viesado); **(iii)** $\mathrm{Var}(\hat\beta) = \sigma^2 / \sum_{i=1}^n x_i^2$; **(iv)** $\hat\sigma^2 = \sum_{i=1}^n (y_i - \hat y_i)^2/(n-1) = \sum_{i=1}^n (y_i - \hat\beta\, x_i)^2/(n-1)$ é um estimador não-viesado de σ^2; **(v)** (a) $Z = \sqrt{\sum_{i=1}^n x_i^2}\,(\hat\beta - \beta)/\sigma \sim \mathrm{Normal}(0,1)$, (b) $U = (n-1)\,\hat\sigma^2 \sim \chi^2_{n-1}$, (c) Z e U são independentes e portanto (d) $\sqrt{\sum_{i=1}^n x_i^2}\,(\hat\beta - \beta)/\hat\sigma$ segue uma distribuição t de Student com $(n-1)$ graus de liberdade.

Observação: As partes **(iv)** e **(v)** são opcionais, mas esses resultados serão usados no Exercício 5

5. (Conjunto de dados de Edwin Hubble). A seguinte tabela mostra a distância do sistema solar x_i (em megaparsecs, 1 megaparsec $\approx 3.3 \times 10^6$ anos luz) e a velocidade de recessão y_i (em km/segundo) para $n = 24$ nebulosas.

Nebulosa	1	2	3	4	5	6	7	8
distância	0.032	0.034	0.214	0.263	0.275	0.275	0.450	0.500
velocidade	170	290	-130	-70	-185	-220	200	290

Nebulosa	9	10	11	12	13	14	15	16
distância	0.500	0.630	0.800	0.900	0.900	0.900	0.900	1.000
velocidade	270	200	300	-30	650	150	500	920

Nebulosa	17	18	19	20	21	22	23	24
distância	1.100	1.100	1.400	1.700	2.000	2.000	2.000	2.000
velocidade	450	500	500	960	500	850	800	1090

A *Lei de Hubble* afirma que a velocidade deve ser proporcional à distância ($V = H_0\, D$, com a constante de proporcionalidade H_0, a *constante de Hubble*, desempenhando um rol fundamental na cosmologia moderna). Assim, faz sentido ajustar os dados acima usando um modelo de regressão sem intercepto $v_i = \beta\, d_i + \epsilon_i$. **(i)** Faça um gráfico de dispersão. Visualmente, a hipótese de Hubble faria sentido para essas observações;

(ii) Calcule as estimativas para o coeficiente β e para a variância σ^2 do erro; **(iii)** Calcule um IC com confiança 90% para β; **(iv)** Faça gráficos dos resíduos que permitam avaliar se os supostos de homocedasticidade e normalidade são razoáveis; **(v)** A constante de Hubble H_0 é usualmente expressa em $mpc/(km/s)$ (megaparsecs / quilômetros por segundo), uma unidade que é recíproca do tempo. O seu recíproco, H_0^{-1} é medido em unidades de tempo e na cosmologia moderna corresponde à idade do universo (tempo desde o *big bang*. Usando que 1mpc $\approx 3.26 \times 10^6$ anos luz e que a velocidade da luz é aproximadamente 299000km/s, qual seria a estimativa da idade do universo de acordo a esses dados? Construa um IC com confiança 90%. (As medidas mais modernas situam a idade do universo em 13.8 bilhões de anos; um modelo não pode ser melhor que os dados usados para estima-lo!)

6. (Adaptado de Bussab e Morettin). Os dados abaixo correspondem às variáveis `renda familiar` e `gasto com alimentação` numa amostra de dez familias, as duas medidas em salários mínimos.

Família (i)	1	2	3	4	5	6	7	8	9	10
Renda (x_i)	3	5	10	20	30	50	70	100	150	200
Gasto (y_i)	1,5	2,0	6,0	10,0	15,0	20,0	25,0	40,0	60,0	80,0

(i) Faça um gráfico de dispersão; **(ii)** Obtenha as estimativas dos parâmetros da regressão (α, β e σ^2; **(iii)** Existe suficiente evidência para concluir que a renda é importante para explicar o gasto com alimentação? Use nível de significância 10%; **(iv)** Qual é a proporção da variabilidade do gasto com alimentação que é explicado pela variável renda? **(v)** Obtenha um IC com confiança 90% para o gasto médio de famílias com renda de 20 salários mínimos; **(vi)** Obtenha um intervalo de previsão para o gasto com alimentação de uma única família com renda de 20 salários mínimos; **(vii)** Qual á previsão do gasto com alimentação para uma família com renda de 1000 salários mínimos? Você acha esse valor razoável? Explique.

7. (Adaptado de Montgomery et al). Um artigo no Journal of Sound and Vibration (Vo. 151, 1991, p. 383-394) descreve um estudo para investigar a relação entre exposição ao ruído e hipertensão. Os dados na seguinte tabela são representativos dos descritos no estudo.

y	1	0	1	2	5	1	4	6	2	3
x	60	63	65	70	70	70	80	90	80	80

y	5	4	6	8	4	5	7	9	7	6
x	85	89	90	90	90	90	94	100	100	100

(i) Faça um gráfico de dispersão de y (aumento da pressão arterial) versus x (ruído em decibéis). De acordo ao gráfico, um modelo de regressão simples seria razoável? **(ii)** Calcule as estimativas dos parâmetros do modelo; **(iii)** Calcule um IC com confiança 90% para o aumento médio da pressão arterial associado com um nível de ruído de 85 decibéis; **(iv)** Calcule um intervalo de previsão com 90% de confiança para o aumento

de pressão arterial de uma observação futura com nível de ruído de 85 decibéis. **(v)** Construa a tabela ANOVA e calcule o coeficiente R^2; **(vi)** Faça gráficos dos resíduos que permitam avaliar se os supostos de homocedasticidade e normalidade são adequados para esses dados.

Algumas soluções

1. **(i)** $\hat{\beta} \doteq 0.744$, $\hat{\alpha} \doteq 2.546$, $\hat{\sigma}^2 \doteq 31.809$; **(ii)** $0.170 < \beta < 1.318$; **(iii)** $t_{obs} \doteq 2.711$, p-valor $\doteq 0.014$; **(iv)** $SQ_{tot} \doteq 615.741$, $SQ_{reg} \doteq 11.373$, $R^2 \doteq 0.0185$ ou 1.85%.

2. **(i)** $a = 1$ (sempre), $e = 20$, $b = 106.74$, $c = b = 106.74$, $f = 68.4$, $d = 1.561$; **(ii)** $n = 22$; **(iii)** $F_{obs} \doteq 1.561$, $F_{1,20;0.10} \doteq 2.975$, logo a esse nível não existe evidência que $\beta \neq 0$. O p-valor é $P(F_{1,20} > F_{obs}) \doteq 0.226$ ou 22.6%; **(iv)** $R^2 \doteq 0.072$ ou 7.2%; **(v)** $|\hat{\beta}| \doteq 4.339$ $[F_{obs} = (t_{obs})^2 = \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2/\hat{\sigma}^2 = (n-2)\,\hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2/SQ_{res}]$

3. Veja o código **R** a seguir.

```
> vest<-c(66,62,66,56,93,66,91,53,60,81,57,58,55,
+         65,72,75,64,67,67,73)
> ira<-c(63.5,53.7,55.3,79.7,73.4,72.6,84.0,55.0,67.7,60.4,58.3,38.9,71.3,
+        70.0,71.4,81.9,68.8,72.6,70.3,77.1)
> n<-length(ira)
> conf<-0.95
> tail<-(1+conf)/2
>
> # Grafico de dispersão, mostra que ajuste linear pode ser razoavel
> par(mfrow=c(1,1))
> plot(vest,ira)
> ajuste<-lm(ira~1+vest)
> abline(ajuste) # inclua a reta de minimos quadrados no grafico
> summary(ajuste) # estimativas de alfa (36.3444), beta (0.4595) e sigma2 (10.08^2)

Call:
lm(formula = ira ~ 1 + vest)

Residuals:
    Min      1Q  Median      3Q     Max
-24.098  -5.687   3.105   5.859  17.621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.3444    14.3320   2.536   0.0207 *
vest          0.4595     0.2101   2.187   0.0422 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.08 on 18 degrees of freedom
Multiple R-squared:  0.2099,Adjusted R-squared:  0.166
F-statistic: 4.782 on 1 and 18 DF,  p-value: 0.0422

> (a<-ajuste$coefficients[1])  # estimativa de alfa
```

```
(Intercept)
   36.34436
> (b<-ajuste$coefficients[2])  # estimativa de beta
     vest
0.4595492
> (s2<-sum((ira-a-b*vest)^2)/(n-2))  # estimativa de sigma2
[1] 101.6858
> (sa<-sqrt(s2*sum(vest^2)/(n*sum((vest-mean(vest))^2)))) # estimativa do erro de a (eq. 9)
[1] 14.33197
> (sb<-sqrt(s2/sum((vest-mean(vest))^2))) # estimativa do erro de b (eq. 8)
[1] 0.2101482
> a+qt(c(1-tail,tail),n-2)*sa # IC para alfa
[1]  6.234017 66.454707
> b+qt(c(1-tail,tail),n-2)*sb # IC para beta
[1] 0.01804426 0.90105413
> s2/qchisq(c(tail,1-tail),n-2) # IC para sigma2
[1]  3.22542 12.35438
>
> 2*(1-pt(2.187,n-2))  # p-valor do teste que beta = 0 (bilateral)
[1] 0.0421863
>
> anova(ajuste) # tabela de analise da variancia
Analysis of Variance Table

Response: ira
          Df  Sum Sq Mean Sq F value Pr(>F)
vest       1  486.27  486.27   4.782 0.0422 *
Residuals 18 1830.34  101.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> 1-pf(4.782,1,n-2)  # outra forma de calcular o p-valor
[1] 0.04220488
>
> xc<-70  # valor de vest para a previsao
> # calculamos primeiro o erro da estimacao de muc (eq. 12 das notas)
> (erro<-sqrt(s2)*sqrt(((xc-mean(vest))^2)/sum((vest-mean(vest))^2)+1/n))
[1] 2.322589
> (mc<-a+b*xc)   # estimativa pontual para a media correspondente a xc
(Intercept)
   68.51281
> mc+qt(c(1-tail,tail),n-2)*erro # IC para a media mu_c
[1] 63.63323 73.39238
> # Agora o erro da previsao (eq. 15)
> (erro.prev<-(erro<-sqrt(s2)*sqrt(1+((xc-mean(vest))^2)/sum((vest-mean(vest))^2)+1/n)))
```

```
[1] 10.34796
> mc+qt(c(1-tail,tail),n-2)*erro.prev # Int. previsao
[1] 46.77255 90.25306
> par(mfrow=c(2,2))
> plot(ajuste)
> # Dos dois primeiros gráficos pareceria que os supostos de
> # homoscedasticidade e normalidade são razoaveis
```