Modelos Lineares I

Regressão Linear Simples (RLS):

Bandas de confiança e predição

(10^a, 11^a e 12^a Aulas)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Inferência para reta de regressão:

Introdução:

- Um dos principais objetivos na análise de regressão linear é estimar a média da distribuição de Y, ou seja, E(Y), para um dado valor de X, digamos X_i. O valor médio da variável resposta Y dado X_i, será denotado por $E(Y/X_i)$, ou *alternativamente*, por $E(Y_i)$.
- O estimador de $E(Y/X_i)=E(Y_i)$ é dado por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

2

Inferência para reta de regressão:

Exemplo 1:

- Exemplo da concentração da substância (X) e ganho de peso (Y):

Pergunta: *Qual o ganho médio de peso estimado para bois que receberam uma concentração da substância de:*

a) *4 mg/l ?*

b) *5 mg/l ?*

c) *8 mg/l ?*

3

Qual a média e a variância de $\hat{Y}_i$?

$$E(\hat{Y}_i) = E(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \beta_0 + \beta_1 X_i$$

$$VAR(\hat{Y}_i) = E[\hat{\beta}_0 + \hat{\beta}_1 X_i - (\beta_0 + \beta_1 X_i)]^2 = E[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) X_i]^2$$

$$VAR(\hat{Y}_i) = VAR(\hat{\beta}_0) + VAR(\hat{\beta}_1) X_i^2 + 2\, X_i\, COV(\hat{\beta}_0, \hat{\beta}_1)$$

$$VAR(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) + \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} X_0^2 - 2\, X_i \frac{\bar{X}\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$VAR(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2 + X_i^2 - 2\, X_i \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right)$$

4

Observações:

- ✓ O estimador $\hat{Y}_i$ de $E(Y/X_i)$ é uma função linear das v.a\`s $Y_i$\`s e, portanto, tem distribuição normal com os parâmetros:

$$\hat{Y}_i \sim N\left[\beta_0 + \beta_1 X_i, \quad \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \right]$$

- ✓ $\hat{Y}_i$ é um estimador não tendencioso de $E(Y/X_i) = \beta_0 + \beta_1 X_i$;
- ✓ Demonstra-se que: $E(Y/X_i) = E(Y_i) = E(\hat{Y}/X_i) = E(\hat{Y}_i)$.

5

Intervalo de confiança para $E(Y/X_i)$:

- Dado um valor X_i, pode-se calcular o intervalo de confiança para o valor médio de Y, denotado por $E(Y/X_i)$, ao nível de confiança $100(1-\alpha)\%$, por meio da seguinte estatística:

$$T = \frac{\hat{Y}_i - E(Y/X_i)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \sim T - \text{Student com } (n-2) \text{ g.l\`s}$$

onde:

$$\hat{\sigma}^2 = QMRes = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

6

Intervalo de confiança para $E(Y/X_i)$:

- Portanto o intervalo de confiança para o <u>valor médio de Y</u>, denotado por $E(Y/X_i)$, ao nível de confiança de $100(1-\alpha)\%$, é obtido por:

$$IC_{E(Y/X_i),100(1-\alpha)\%} = \left[\underbrace{\hat{Y}_i - t_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}_{\text{Limite inferior } (L_{inf}) \text{ do intervalo}} \;,\; \underbrace{\hat{Y}_i + t_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}_{\text{Limite superior } (L_{sup}) \text{ do intervalo}}\right]$$

7

Exemplo 1: Dados sobre a concentração da substância X (em mg/l) e ganho de peso Y (em kg) após trinta dias, de n=30 bovinos:

Boi	Conc. Subst. (mg/l)	Ganho de peso (kg)	Boi	Conc. Subst. (mg/l)	Ganho de peso (kg)
1	1,00	9,40	**16**	5,00	14,10
2	3,70	11,40	**17**	5,50	12,50
3	1,00	12,00	**18**	6,00	15,20
4	9,00	16,00	**19**	6,50	14,20
5	2,00	11,00	**20**	7,00	16,50
6	2,25	12,50	**21**	7,50	17,00
7	2,91	10,40	**22**	8,00	14,50
8	2,75	11,50	**23**	8,25	16,00
9	3,00	12,50	**24**	9,40	17,00
10	3,50	14,00	**25**	9,43	14,90
11	3,75	14,50	**26**	8,94	15,00
12	9,45	17,00	**27**	9,20	19,00
13	4,25	13,25	**28**	9,50	17,50
14	7,00	14,80	**29**	8,00	16,00
15	4,75	14,00	**30**	9,00	17,50

8

Revendo os resultados do ajuste do modelo

Descriptive Statistics

	Mean	Std. Deviation	N
Y_ganho_peso	14,3717	2,36708	30
X_conc.subs	5,9177	2,83677	30

Correlations

		Y_ganho_peso	X_conc.subs
Pearson Correlation	Y_ganho_peso	1,000	,877
	X_conc.subs	,877	1,000
Sig. (1-tailed)	Y_ganho_peso	.	,000
	X_conc.subs	,000	.
N	Y_ganho_peso	30	30
	X_conc.subs	30	30

9

Revendo os resultados do ajuste do modelo

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,877[a]	,770	,761	1,15619

a. Predictors: (Constant), X_conc.subs

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	125,059	1	125,059	93,554	,000[a]
	Residual	37,429	28	1,337		
	Total	162,488	29			

a. Predictors: (Constant), X_conc.subs

b. Dependent Variable: Y_ganho_peso

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10,040	,495		20,277	,000	9,025	11,054
	X_conc.subs	,732	,076	,877	9,672	,000	,577	,887

a. Dependent Variable: Y_ganho_peso

10

Exemplo 1 – a) Considerando os dados dos n=30 bovinos, obtenha uma estimativa do valor médio de Y dado que X_i=4 mg/l, e obtenha um IC de 95% para $E(Y/X_i=4)$.

Resp.: $[12{,}443 \text{ kg}; 13{,}493 \text{ kg}] \rightarrow A = 1{,}05 \text{ kg}$

11

Exemplo 1 – b) Considerando os dados dos n=30 bovinos, obtenha uma estimativa do valor médio de Y dado que X_i=5 mg/l, e obtenha um IC de 95% para $E(Y/X_i=5)$.

Resp.: $[13{,}245 \text{ kg}; 14{,}155 \text{ kg}] \rightarrow A = 0{,}91 \text{ kg}$

12

Exemplo 1 – c) Considerando os dados dos n=30 bovinos, obtenha uma estimativa do valor médio de Y dado que X_i=8 mg/l, e obtenha um IC de 95% para E(Y/X_i=8).

Resp.: $[15{,}356 \text{ kg}; 16{,}436 \text{ kg}] \rightarrow A = 1{,}08 \text{ kg}$

13

Intervalo de Confiança para E(Y/X_i) – *Bandas de confiança*:

- Calculando-se intervalos de confiança para E(Y/X_i) considerando diferentes valores de X, pode-se representar no gráfico de dispersão uma região em torno da reta de regressão estimada, indicando os limites superiores e inferiores desses intervalos.

- Essa região recebe o nome de ***bandas de confiança.***

14

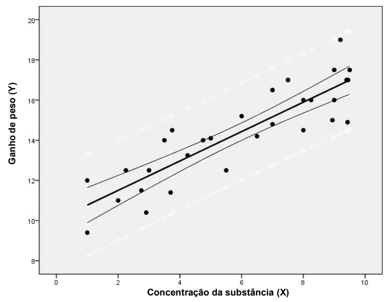
Intervalos de confiança para E(Y/X_i)

(*Bandas de confiança*)

Exemplo 1: Para ilustrar o cálculo das bandas de confiança consideraremos os dados da concentração de substância (X) e ganho de peso (Y) dos n=30 bovinos.

15

Figura 1: Gráfico de dispersão entre a concentração da substância (X) e o ganho de peso (Y), incluindo o modelo de regressão ajustado e as *bandas de confiança*.



16

Predição de Y dado um novo valor de X:

Introdução:

- Uma importante aplicação da análise de regressão linear é a predição de um valor individual da variável resposta Y dado um valor de X_i de interesse, sendo denotado por Y/X_i, ou alternativamente, por Y_i.
- Para obter um IC para um valor individual de Y, dado um X_i, basta determinar a distribuição da diferença $Y_i - \hat{Y}_i$.
- O valor médio de Y dado o valor X_i, será denotado por E(Y/X_i), ou *alternativamente*, por E(Y_i).

17

Inferência para reta de regressão:

Exemplo 2:

- Exemplo da concentração da substância (X) e ganho de peso (Y):

Pergunta: *Qual o ganho de peso estimado para um boi que recebeu uma concentração da substância de:*

a) *4 mg/l ?*

b) *10 mg/l ?*

18

Qual a média e a variância de $Y_i - \hat{Y}_i$?

Supondo que a relação linear permanece quando observamos um novo valor de X, temos que:

(1) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

(2) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Subtraindo-se as equações, obtém-se: (3)= (1) - (2):

(3) $Y_i - \hat{Y}_i = \beta_0 + \beta_1 X_i + \varepsilon_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$

(3) $Y_i - \hat{Y}_i = \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) X_i + \varepsilon_i$

19

Qual a média e a variância de $Y - \hat{Y}_i$? (*continuação*)

$$E\left(Y_i - \hat{Y}_i\right) = 0 \quad \rightarrow \quad E\left(\hat{Y}_i\right) = E\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) = \beta_0 + \beta_1 X_i = E(Y_i)$$

$$VAR\left(Y_i - \hat{Y}_i\right) = VAR(Y_i) + VAR\left(\hat{Y}_i\right)$$

$$VAR\left(Y_i - \hat{Y}_i\right) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

$$VAR\left(Y_i - \hat{Y}_i\right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

20

Observações:

- A estatística $Y_i - \hat{Y}_i$ tem distribuição normal com os parâmetros:

$$Y_i - \hat{Y}_i \sim N\left[0, \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)\right]$$

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ também é um estimador não tendencioso de $E(Y/X_i)=E(Y_i)$, dado um novo valor de X.

21

Intervalo de predição para Y_i:

- Dado o valor X_i, pode-se calcular o intervalo de predição para o valor individual de Y, ao nível de confiança de $100(1-\alpha)\%$, por meio da seguinte estatística:

$$T = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}} \sim T - \text{Student com } (n-2) \text{ g.l`s}$$

onde :

$$\hat{\sigma}^2 = QMRes = \frac{\sum_{i=1}^{n} e_i^2}{n - 2}$$

22

Intervalo de predição para Y_i:

- Portanto o intervalo de predição para o valor individual de Y, dado um X_i, ao nível de confiança de $100(1-\alpha)\%$, é obtido por:

$$IC_{Y_i,100(1-\alpha)\%} = \left[\underbrace{\hat{Y}_i - t_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}_{\text{Limite inferior } (L_{inf}) \text{ do intervalo}} \quad , \quad \underbrace{\hat{Y}_i + t_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}_{\text{Limite superior } (L_{sup}) \text{ do intervalo}}\right]$$

23

Exemplo 2 – a) Considerando o exemplo dos n=30 bovinos, obtenha uma predição de Y para um novo valor de X, digamos X_i=4 mg/l, e construa um intervalo de predição de 95% para o valor de Y/X_i=4.

Resp.: 12,968 kg ; [10,543 kg;15,393 kg]

24

Prof.: José Rodrigo de Moraes: Estatístico (ENCE), Mestre em Estatística Social (ENCE) e Doutor em Saúde Coletiva (IESC/UFRJ)

Exemplo 2 – b) Considerando o exemplo dos n=30 bovinos, obtenha uma predição de Y para um novo valor de X, digamos X_i=10 mg/l, e construa um intervalo de predição de 95% para o valor de Y/X_i=10.

Resp.: 17,36 kg ; [14,871 kg;19,849 kg]

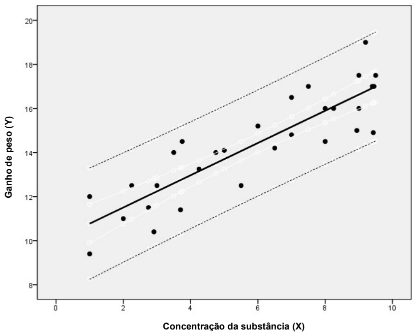
25

Intervalos de predição para Y/X_i - *Bandas de predição:*

- Se calcularmos intervalos de predição para Y considerando diferentes valores de X_i, pode-se representar no gráfico de dispersão uma região, delimitada pelos limites superiores e inferiores desses intervalos.
- Essa região recebe o nome de ***bandas de predição***.

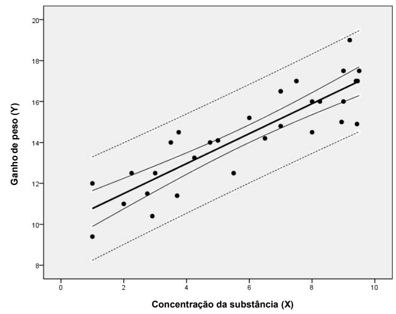
26

Figura 2: Gráfico de dispersão entre a concentração da substância (X) e o ganho de peso (Y), incluindo o modelo de regressão ajustado e as *bandas de predição*.



27

Figura 3: Gráfico de dispersão entre a concentração da substância (X) e o ganho de peso (Y), incluindo o modelo de regressão ajustado e as *bandas de confiança* e *predição*.



Graphs / Legacy Dialogs / Scatter / Overlay Scatter (usar a var. X sempre no eixo das abscissas)

28

Bandas de confiança e predição:

- Tamanho da amostra: n ↑ IC ↓
- Variância de Y: $\hat{\sigma}^2 \downarrow$ IC ↓
- E quanto ao desvio de X_i em relação a sua média ?

29

Aula prática – Exercício 1: Um estudo foi desenvolvido com objetivo de avaliar o efeito da idade no tempo de reação a um certo estímulo.

Os dados referentes as essas duas variáveis para uma amostra de n=20 indivíduos se encontram na tabela 1 a seguir.

a) Escreva a equação do modelo, descrevendo os seus termos e variáveis no contexto do problema.

b) Ajuste o modelo pelo método de mínimos quadrados (MQ) e interprete as estimativas dos parâmetros e o coeficiente de determinação do modelo (contexto).

c) Estime o tempo médio de reação para pessoas de 30 anos de idade, e construa o seu respectivo IC de 95%. E para o grupo de 28 anos?

30

Aula prática – Exercício 1 (*continuação*):

d) Estime o tempo de reação de uma pessoa de 30 anos de idade, e construa o seu respectivo IC de 95%. E para uma pessoa de 28 anos?

e) Obtenha os intervalos de confiança e predição (*usando o R e o SPSS*).

f) Construa um gráfico de dispersão para representar as bandas de confiança e predição, incluindo os dados observados e o modelo de regressão linear ajustado (*usando o R e o SPSS*).

Resp. : b) $\hat{Y}_i = 80,5 + 0,90\ X_i$ c) $IC_{E(Y/X_i=30);\ 95\%} = [104,87;\ 110,13]$ $IC_{E(Y/X_i=28);\ 95\%} = [102,98;\ 108,43]$ d) $IC_{Y/X_i=30;\ 95\%} = [95,46;\ 119,54]$ $IC_{Y/X_i=28;\ 95\%} = [93,64;\ 117,76]$

31

Tabela 1: Informações sobre n=20 indivíduos

Aluno	Idade (X)	Tempo de reação (Y)
1	20	96
2	20	92
3	20	106
4	20	100
5	25	98
6	25	104
7	25	110
8	25	101
9	30	116
10	30	106
11	30	109
12	30	100
13	35	112
14	35	105
15	35	118
16	35	108
17	40	113
18	40	112
19	40	127
20	40	117
Total	600	2.150

$\sum_{i=1}^{n} X_i^2 = 19.000$

$\sum_{i=1}^{n} Y_i^2 = 232.498$

$\sum_{i=1}^{n} X_i Y_i = 65.400$

32

Considerações finais:

- Ressalta-se que os IC`s são mais informativos que as estimativas pontuais;
- Na análise de regressão se deseja prever valores de Y em situações em que o valor de X está fora do intervalo de valores efetivamente observados. Tais previsões (extrapolações), são muito menos confiáveis do que previsões baseadas em valores de X contidos no intervalo de valores previamente observados.

33

Lembrete:

- Se repetirmos o cálculo do IC de E(Y) para diferentes valores X_i obtemos as chamadas *bandas de confiança*.
- Se repetirmos o cálculo do IC de Y para diferentes valores de X obtemos as chamadas *bandas de predição*.
- As estimativas pontuais são as mesmas para um mesmo valor de X, mas não os IC`s.

34