Modelos Lineares I

Regressão Linear Simples (RLS):

ANOVA

(7ª, 8ª e 9ª Aulas)

Professor: Dr. José Rodrigo de Moraes

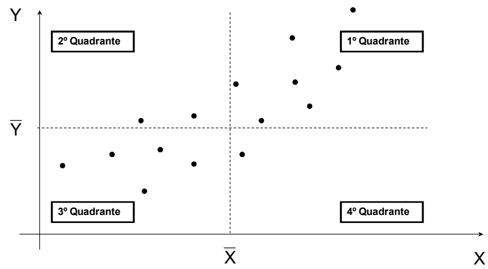
Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Análise de Correlação no Modelo de RLS:

Suponha que uma amostra de observações (X_i,Y_i), $\forall$ i=1,2,...,n; seja representada pelo gráfico de dispersão, dividido em 4 quadrantes definidos pelas médias das variáveis X e Y:



2

Análise de Correlação no Modelo de RLS:

- A soma de todos os produtos $A = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ é uma medida de associação linear entre as variáveis X e Y.
  - **Maioria dos pontos (X_i,Y_i) situados nos quadrantes ímpares** → A será positiva (relação positiva ou crescente).
  - **Maioria dos pontos (X_i,Y_i) situados nos quadrantes pares** → A será negativa (relação negativa ou decrescente).
  - **Pontos situados predominantemente nos quatro quadrantes** → nuvem de pontos sem tendência crescente ou decrescente (ausência de relação).

3

Análise de Correlação no Modelo de RLS:

- Da forma como a medida A foi definida apresenta alguns inconvenientes:
  - *Pode ser influenciada pelas unidades de medida de X e Y;*
  - *Pode ser aumentada pelo simples acréscimo de novas observações.*
- Para resolver tais inconvenientes, divide-se a medida A pelo produto entre os desvios-padrão de X e Y e o número de observações da amostra, obtendo assim o chamado **coeficiente de correlação linear de pearson**:

$$R = \frac{A}{n \cdot S_X \cdot S_Y}, \quad \text{onde:} \quad S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} \quad \text{e} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}$$

4

Coeficiente de Correlação no Modelo de RLS:

$$R = \frac{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Fórmula ramificada

$$R = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2}\sqrt{\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2}}$$

Intervalo de variação: $-1 \le R \le 1$

5

Coeficiente de Correlação Linear entre X e Y:

- Valores de R próximos de -1 ou +1 indicam correlação forte.
- *Sentido da relação:*
  - $R > 0$ → relação positiva (ou crescente) entre X e Y.
  - $R < 0$ → relação negativa (ou decrescente) entre X e Y.
- Se:
  - $R = -1$ → relação linear negativa perfeita entre X e Y.
  - $R = 0$ → ausência de relação linear entre X e Y.
  - $R = +1$ → relação linear positiva perfeita entre X e Y.

6

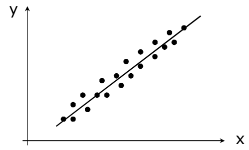
Coeficiente de Correlação Linear entre *X* e *Y*:

- *Grau da relação:*
  - **0 < |** R **| ≤ 0,30** → fraca relação linear entre X e Y.
  - **0,30 < |** R **| ≤ 0,70** → moderada relação linear entre X e Y.
  - **|** R **| > 0,70** → forte relação linear entre X e Y.

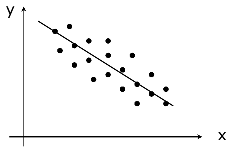
7

Gráfico de Dispersão (esboço): Grau e sentido da correlação

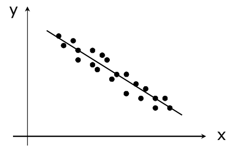
R>0: relação linear **positiva fraca**



R>0: relação linear **positiva forte**



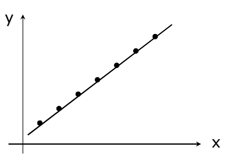
R<0: relação linear **negativa fraca**



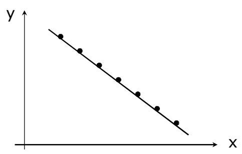
R<0: relação linear **negativa forte**

8

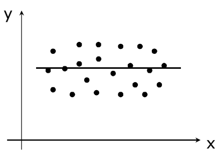
Gráfico de Dispersão (esboço): Grau e sentido da correlação



R =+1: relação linear **positiva perfeita**



R =-1: relação linear **negativa perfeita**



R=0: **ausência de relação linear** entre X e Y

9

Comentários sobre coeficiente de correlação linear:

- O coeficiente de correlação linear de Pearson (R) mede o quanto os pontos num gráfico de dispersão se aproximam de uma linha reta.
- Quanto mais próximo o valor de R estiver de 1 ou -1 mais forte a correlação linear; e quanto mais próximo o valor de R estiver de 0 mais fraca a correlação linear.

10

Exemplo: Dados de n=30 bovinos sobre a concentração da substância *X* (mg/L) e ganho de peso *Y* (kg):

Coeficiente de correlação linear de Pearson (R):

Analyse /Regression / Linear

Qual a interpretação do *R* no contexto do problema ?

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

Coeficiente de correlação linear de Pearson (*R*)

11

Cálculo do coeficiente de Correlação Linear no Modelo de RLS usando os dados dos n=30 bovinos. Use 4 casas decimais (ou mais) e a fórmula abaixo:

$$R = \frac{\sum_{i=1}^{n} X_i Y_i - n\,\bar{X}\,\bar{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}\,\sqrt{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}}$$

12

Relação entre o estimador de β_1 e o coef. de correlação R:

- Uma relação básica importante é a relação estabelecida entre o estimador de β_1 e o coeficiente de correlação linear de pearson R, como mostrada a seguir:

Obtendo os desvios: $x_i = X_i - \bar{X}$ e $y_i = Y_i - \bar{Y}$

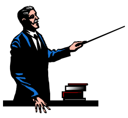
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \rightarrow \sum_{i=1}^{n} x_i y_i = n\hat{\beta}_1 S_X^2 \quad (1)$$

$$R = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{nS_X S_y} \quad (2)$$

13

Relação entre o estimador de β_1 e o coef. de correlação R:

- Substituindo (2) em (1):



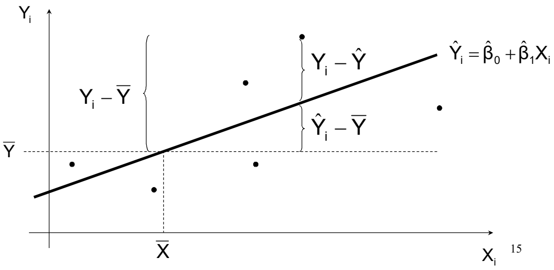
$$R = \frac{\sum_{i=1}^{n} x_i y_i}{nS_X S_y} = \frac{n\hat{\beta}_1 S_X^2}{nS_X S_y} = \hat{\beta}_1 \cdot \frac{S_X}{S_y} \rightarrow \hat{\beta}_1 = R \cdot \frac{S_y}{S_X}$$

Que conclusões podemos extrair dessa relação ?

14

Análise de Variância no Modelo de RLS:

É um método utilizado para testar a significância da relação linear entre X e Y. Consiste em decompor a variação total em duas componentes como ilustra a figura abaixo:



15

Decomposição da variação total:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}\left[(\hat{Y}_i-\bar{Y})+(Y_i-\hat{Y}_i)\right]^2$$

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y})^2 + 2\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})(Y_i-\hat{Y})}$$

Qual o valor dessa quantidade ?

16

Decomposição da variação total:

Logo a variação total pode ser descomposta da seguinte forma:

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SQT} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{SQReg} + \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}_{SQRes}$$

SQT$\rightarrow$ Soma dos quadrados do total (mede a variação total).

SQReg $\rightarrow$ Soma dos quadrados da regressão (variação explicada pelo modelo de regressão ajustado).

SQRes $\rightarrow$ Soma dos quadrados dos resíduos (variação não explicada pelo modelo).

17

Para calcular as somas dos quadrados (SQ`s) recomenda-se o seguinte procedimento:

- Calcula-se primeiramente as somas dos quadrados total (SQT) e dos resíduos (SQRes):

$$SQT = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

$$SQRes = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Em seguida, calcula-se a soma dos quadrados da regressão (SQReg) por diferença como mostrado abaixo:

$$SQReg = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = \sum_{i=1}^{n}\hat{Y}_i^2 - n\bar{Y}^2 \rightarrow SQM = SQT - SQRes$$

18

❑ Tabela de Análise de Variância (ANOVA):

- Com base nas somas dos quadrados definidos constrói-se a chamada Tabela de Análise de Variância.
- Dividindo-se a soma dos quadrados (*SQ`s*) pelos respectivos graus de liberdade (*gl`s*), obtém-se o que define-se de quadrado médio. Assim, o quadrado médio dos resíduos (QMRes) e da regressão (QMReg) são dados, respectivamente, por:

QMRes = SQRes /n-2 e QMReg = SQReg/1

19

Tabela de Análise de Variância (ANOVA) - RLS:

Fontes de variação	Soma dos quadrados	gl	Quadrado médio	Estatística de teste
Modelo	$SQM = \sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2$	1	$QMReg = \frac{SQReg}{1}$	$F = \frac{QMReg}{QMRes} \sim F_{1,n-2}$
Resíduos	$SQRes = \sum_{i=1}^{n} e_i^2$	n - 2	$QMRes = \frac{SQRes}{n-2}$	
Total	$SQT = \sum_{i=1}^{n} Y_i^2 - n\,\bar{Y}^2$	n -1		

20

Voltando ao exemplo dos n=30 gados

[Adaptado de Magalhães & Lima (2003)]:

❑ Em uma dada região acredita-se que o gado alimentado em determinado pasto tem ganho de peso maior que o normal. Estudos de laboratório detectaram uma substância no pasto e deseja-se obter evidências de que tal substância pode ser utilizada para melhorar o ganho de peso dos bovinos. Foram selecionados 15 bois de mesma raça e idade, e cada animal recebeu uma determinada concentração da substância X (em mg/l). O ganho de peso (Y) após 30 dias, foi medido e os dados estão apresentados na tabela abaixo (em kg):

21

Tabela: 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

22

Tabela de Análise da Variância (ANOVA):

Analyse /Regression / Linear

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

Soma dos quadrados (*SQ*) e quadrados médios (*QM*).

23

Teste de Hipóteses com base na Tabela ANOVA:

❑ A partir das v.a`s abaixo:

$$\frac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2} \sim \chi^2_1 \qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

obtemos uma nova v.a F conforme mostrado a seguir:

$$F = \frac{\dfrac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2 \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2} \Big/ 1}{\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \Big/ n-2} = \frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\hat{\sigma}^2} = \frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n} e_i^2 \Big/ n-2}$$

OBS: A v.a F tem distribuição de F-Snedecor com 1 e (n-2) graus de liberdade.

24

Teste de Hipóteses com base na Tabela ANOVA:

- Para o modelo de regressão linear simples, a análise de variância se resume na construção do teste estatístico:

❑ Hipóteses a serem testadas:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

❑ Estatística de Teste:

$$F = \frac{QMReg}{QMRes} \sim F_{1,n-2}$$

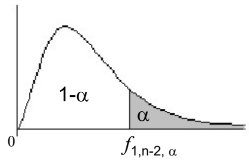
- A estatística F tem distribuição F de Snedecor com 1 e (n-2) graus de liberdade.

25

Testes de Hipóteses com base na Tabela ANOVA:

❑ Região crítica:

$$RC = \{ f \in \Re \;\; / \;\; f \geq f_{1,n-2,\,\alpha} \}$$



❑ Tomada de Decisão:

- Se $f_{obs} \in$ RC rejeita-se H_0:β_1=0 ao nível de significância α, e conclui-se que existe relação linear estatisticamente significante entre X e Y.
- Se $f_{obs} \notin$ RC não há evidências para rejeitar H_0:β_1=0 ao nível de significância α, e conclui-se que não existe relação linear estatisticamente significante entre X e Y.

26

Tabela de Análise de Variância (ANOVA):

<u>**Observações:**</u>

- No modelo de regressão linear simples (RLS) o teste F realizado com base na tabela de análise de variância corresponde ao teste T de significância individual realizado para o parâmetro β_1.
- O mesmo não acontece no caso do modelo de regressão linear múltipla (RLM).

27

Tabela de Análise da Variância (ANOVA):

Analyse /Regression / Linear

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

Teste F **com base na tabela ANOVA, ao nível de 5%**

Arred.: 125,059 ≅ 125,05901915408795   37,429 ≅37,42939751257866

162,488 ≅ 162,4884166666666

28

Medida de Qualidade do Ajuste:
Coeficiente de Determinação do Modelo

- O <u>coeficiente de determinação simples</u>, denotado por R^2, é dado por:

$$R^2 = \frac{SQReg}{SQT} \;, \text{ ou alternativamente: } \quad R^2 = 1 - \frac{SQRes}{SQT}$$

- O R^2 mede o quanto (em termos %) da variação total dos valores da variável resposta Y é explicado pelo modelo ajustado.
- **Intervalo de variação:** $0 \leq R^2 \leq 1$

29

Medida de Qualidade do Ajuste:
Coeficiente de Determinação do Modelo

Observações:

- Se R^2=1 → $SQRes$= 0. Neste caso, todos os resíduos e_i\`s são nulos e, portanto, os pontos estarão sobre a reta de regressão (relação perfeita entre X e Y).
- Se R^2=0 → $SQRes$=SQT. Logo X não contribui para explicar a variação dos valores de Y.

30

Coeficiente de Determinação do modelo de RLS:

$$R^2 = (R)^2$$

$$R^2 = \left[\frac{\sum_{i=1}^{n} X_i Y_i - n\,\bar{X}\,\bar{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2}\,\sqrt{\sum_{i=1}^{n} Y_i^2 - n\,\bar{Y}^2}} \right]^2$$

31

Coeficiente de determinação do modelo (R^2): medida de qualidade do ajuste

Analyse /Regression / Linear

Qual a interpretação do R^2 no contexto do problema ?

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

Coeficiente de determinação simples (R^2=77,0%)

32

Cálculo do coeficiente de Determinação do modelo de RLS:

$$R^2 = (0{,}877)^2 \cong 0{,}77 \quad \rightarrow \quad 77\%$$

Ou alternativamente:

$$R^2 = 1 - \frac{SQRes}{SQT} = 1 - \underline{\qquad\qquad\qquad} \cong$$

33

Medida de Qualidade do Ajuste:
Coeficiente de determinação do modelo ajustado(R^2_{aj})

- O coeficiente de determinação do modelo ajustado, denotado por R^2_{aj}, é dado por:

$$R^2_{aj} = 1 - \frac{SQRes/_{n-2}}{SQT/_{n-1}} = 1 - \left(\frac{n-1}{n-2}\right) \cdot \frac{SQRes}{SQT}$$

OBS: Tanto R^2 quanto R^2_{aj} são medidas da qualidade global do modelo.

34

Coeficiente de determinação do modelo ajustado (R^2_{aj}): medida de qualidade do ajuste:

Analyse /Regression / Linear

Qual a interpretação do R^2_{aj} no contexto do problema ?

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

Coeficiente de determinação ajustado (R^2_{aj} =76,1%).

35