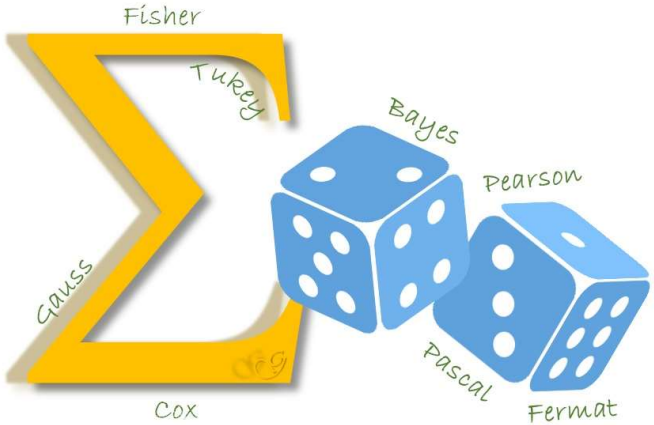


BIOESTATÍSTICA

Modelos de Regressão Aplicados em Epidemiologia

Profª. Drª. Maria do Rosário D O Latorre

Profª. Drª. Gleice M S Conceição

Programa

1. Introdução à análise de regressão
2. Noções de covariância e correlação
3. Modelo de regressão linear simples e múltipla
    - estimação dos parâmetros
    - tabela de análise de variância (ANOVA)
    - distribuições de probabilidades: Normal, t-Student, F-Snedecor e χ^2
    - interpretação dos coeficientes
    - análise dos resíduos
    - teste F-parcial
    - correlação parcial e múltipla
    - variáveis indicadoras
    - confusão e interação
    - escolha do melhor modelo

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Programa

4. Modelo de regressão polinomial
5. Análise de tendência em séries históricas usando modelos de regressão;
6. Modelo de regressão logística simples e múltipla:
    - o modelo logístico
    - estimação dos parâmetros
    - interpretação dos coeficientes
    - medidas de ajuste do modelo
    - confusão e interação
    - escolha do melhor modelo
    - análise de resíduos

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Bibliografia recomendada

1. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. **Applied regression analysis and other multivariable methods**. 3rd edition. Brooks/Cole Pub Co, Boston, 1997.
2. Curns AT, Mizam A. **Student solutions manual for Kleimbaum, Kupper, Muller and Nizam´s Applied regression analysis and other multivariable methods.** Brooks/Cole Pub Co, Boston, 1998.
3. Kutner MH, Christopher J. Nachtsheim CJ, Neter J, Li W. **Applied Linear Statistical Models. 5ª ed.** McGraw-Hill/Irwin, Boston, 2004.
4. Draper NR, Smith H. **Applied Regression Analysis.** John Wiley and Sons, 3rd edition. New York, 1998.
5. Kleinbaum DG, Klein M. **Logistic regression. A self-learning text**. 2nd edition. Springer-Verlag, New York, 2002.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Bibliografia recomendada

6. Hosmer DW, Lemeshow S. **Applied logistic regression**. John Wiley and Sons, 2nd edition. New York, 2000.
7. Pereira MG. **Epidemiologia Teoria e Prática**. Rio de Janeiro: Editora Guanabara Koogan, 1999.
8. Laporta GZ, Latorre, MRDO. **Epidemiologia Aplicada via ambiente R.** Publicação independente, 2019.
9. Crawley MJ. **The R Book.** John Wiley & Sons Inc. Hoboken, 2012.
10. Wickham H, Grolemund G. **R for Data Science.** O'Reilly Media. Sebastopol, CA, 2017.

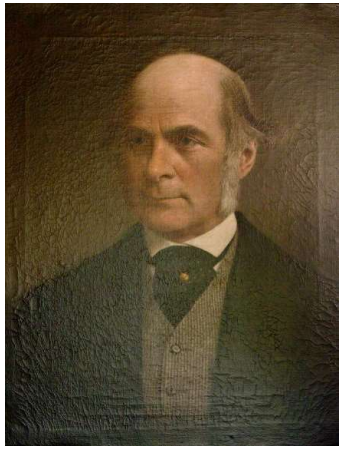
    Existe uma versão traduzida para o português:

    Wickham H, Grolemund G. **R para data science: Importe, arrume, transforme, visualize e modele dados.**

    Alta Books. Brasil, 2019.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Johann Carl Friedrich Gauss (1777-1855), matemático, astrônomo e físico alemão, conhecido popularmente como o “príncipe dos matemáticos”, trouxe grandes contribuições em diversas áreas da ciência, entre leras, a estatística. Desenvolveu o método dos mínimos quadrados.



Sir Francis Galton (1822-1911) , antropólogo, meteorologista, matemático e estatístico inglês. Desenvolveu uma descrição matemática da tendência, que chamou de regressão e que foi o precursor dos modelos de regressão atuais.

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

História natural dos alunos do curso de Regressão...

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Introdução à análise de regressão

Na prática há diversas situações em que a análise de regressão é apropriada:

1. Quando se deseja caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (Xi), ié, avaliar a extensão, direção e força da relação (associação).
2. Procurar uma função matemática ou equação para descrever a variável dependente (Y) como função da variáveis independentes (Xi), ié, predizer Y em função dos Xi; determinando o melhor modelo estatístico que descreva essa relação.
3. Descrever quantitativa e/ou qualitativamente a relação entre os Xi e Y, controlando o efeito de outras variáveis (Ci).

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Introdução à análise de regressão

Na prática há diversas situações em que a análise de regressão é apropriada:

4. Verificar o efeito interativo de 2 ou mais variáveis independentes às quais se relacionam com a variável dependente.
5. Determinar quais das muitas variáveis independentes são importantes para descrever ou predizer a variável dependente. Ordenar as variáveis independentes em sua ordem de importância em relação à variável dependente.
6. Comparar múltiplos relacionamentos derivados da análise de regressão.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Introdução à análise de regressão

- ✓ É importante ser cauteloso sobre os resultados obtidos em uma análise de regressão, ou, de uma maneira mais geral, em qualquer análise utilizando técnicas estatísticas que procurem quantificar uma associação entre 2 ou mais variáveis.
- ✓ A análise estatística pode estar correta, porém os dados podem estar viciados e/ou incompletos (vícios no delineamento, na amostragem, nas medidas, na escolha das variáveis e outros)
- ✓ O achado de uma associação estatística significativa em um particular estudo não estabelece uma relação causal.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Introdução à análise de regressão

Questões básicas

- ✓ Qual a função matemática mais apropriada a ser utilizada? (Em outras palavras: os dados se ajustam melhor a uma reta? A uma parábola? A uma função logística?)
- ✓ Como determinar o melhor modelo que se ajuste aos dados?
- ✓ Qual a validade e a precisão da(s) estimativa(s) do(s) coeficiente(s) de regressão?
- ✓ A presença, no modelo, de determinada variável independente melhora a precisão do mesmo?
- ✓ Dado um modelo específico, o que ele significa?

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Introdução à análise de regressão

ESTRATÉGIAS (*stepwise*):

MODELO MAIS COMPLEXO ➡ MAIS SIMPLES

(BACKWARD SELECTION)

MODELO MAIS SIMPLES ➡ MAIS COMPLEXO

(FORWARD SELECTION)

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Análise de Regressão Simples

- ✓Duas variáveis quantitativas
- ✓Descrever a relação entre elas
- ✓Eventualmente, prever o valor de uma delas para um determinado indivíduo quando só conhecemos o valor da outra

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Análise de Regressão Simples

Exemplos

- ✓Tempo de reação a um estímulo (segundos) e idade
- ✓Peso e idade
- ✓Peso e altura
- ✓Número de óbitos e concentração de um determinado poluente

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Análise de Regressão Simples

✓ Variável resposta, dependente ou preditiva

A variável que está sendo afetada pela outra ou outras, que acreditamos depender das outras, que pode ser explicada ou prevista pelas outras.

✓ Variável explicativa, independente ou preditora

A variável que afeta a outra, que pode ajudar a explicar a variabilidade da outra e a prever a outra.

Quando o modelo envolve apenas uma variável explicativa, será chamado de modelo de Regressão Linear Simples.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Exemplo 1

Amostra de crianças (dados hipotéticos)

Criança	Peso (libras)	Idade (anos)	Altura (pés)
1	64	8	57
2	71	10	59
3	53	6	49
4	67	11	62
5	55	8	51
6	58	7	50
7	77	10	55
8	57	9	48
9	56	10	42
10	51	6	42
11	76	12	61
12	68	9	57

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Exercício 1

Inicialmente, vamos estudar a relação entre idade e peso na amostra de crianças.

a) Identifique a variável resposta (ou dependente) e a explicativa (ou independente).

b) Construa o diagrama de dispersão, com a variável dependente no eixo y e a independente no eixo x. Interprete-o.

c) Calcule a média, a variância e o desvio padrão de ambas as variáveis.

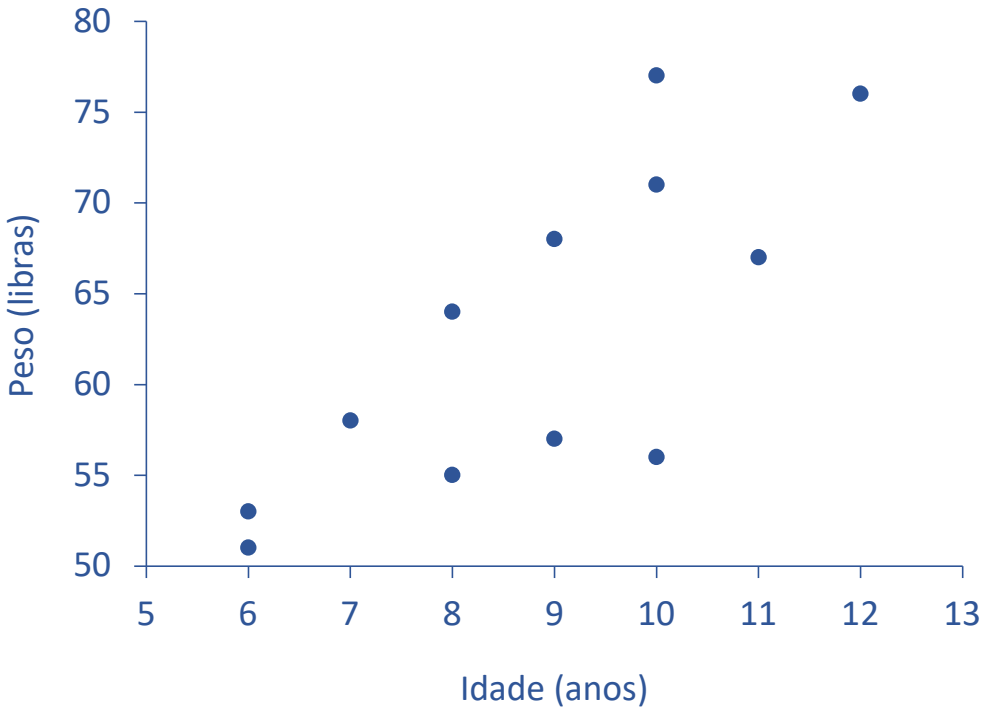
Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Diagrama de dispersão

Diagrama de dispersão entre o peso e a altura em uma amostra de crianças.



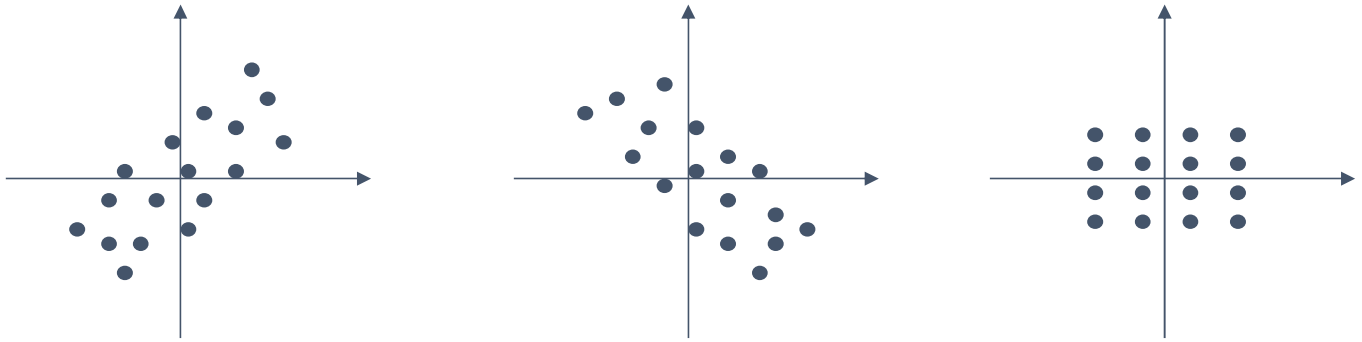
Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Tipos de associação entre variáveis

(a) (b) (c)



É possível obter uma medida de associação?

$$\sum X \cdot Y$$

- ✓ Número positivo grande
  ⟹ associação direta
- ✓ Número negativo grande
  ⟹ associação inversa

Mas, convém levar em conta o número de observações. Então:

$$\sum \frac{X \cdot Y}{n}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de correlação linear de Pearson (r)

O coeficiente de correlação linear de Pearson pode ser escrito como

$$r = corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i - \bar{X})}{S_X}\frac{(Y_i - \bar{Y})}{S_Y} \qquad (1)$$

Isto é, o coeficiente de correlação (r) é a média dos produtos dos valores padronizados das variáveis X e Y.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de correlação linear de Pearson (r)

$$r = corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{X})}{S_X}\frac{(y_i - \bar{Y})}{S_Y}$$

A definição formal para o coeficiente de correlação linear de Pearson é:

$$r = corr(X,Y) = \frac{cov(X,Y)}{S_x S_y}$$

onde $cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de correlação linear de Pearson (r)

Propriedades

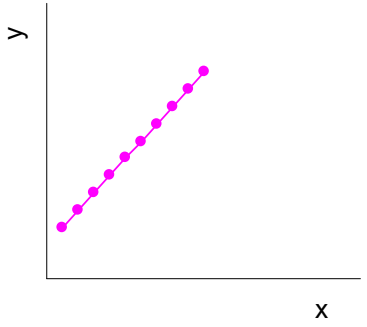
- ✓ $-1 \leq corr(X,Y) \leq 1$
- ✓ Valores próximos de 1 ou –1 indicam uma associação forte
- ✓ Valores próximos de zero quando não existe associação

O coeficiente de correlação linear mede:

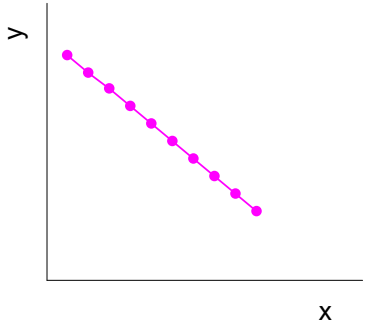
- Presença de associação linear
- Força de uma associação linear

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

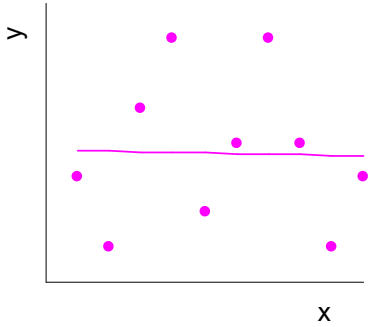
Coeficiente de correlação linear de Pearson (r)



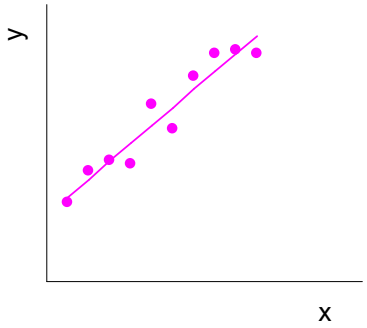
r = 1



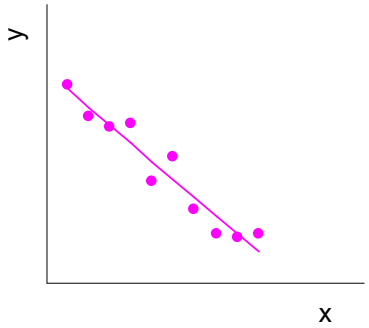
r = -1



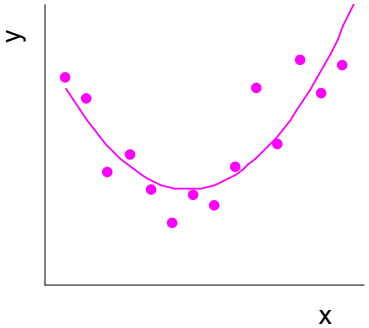
r próximo de 0



r próximo de 1



r próximo de -1



r próximo de 0

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de correlação linear de Pearson (r)

Alguns autores sugerem avaliar a presença e a força de uma associação linear a partir do coeficiente de correlação do seguinte modo:

- ✓de 0,10 a 0,39 - fraca
- ✓de 0,40 a 0,69 - moderada
- ✓de 0,70 até 1 - forte

Mas não há, de fato, uma norma rígida sobre isto.

Deve-se levar em conta o contexto , o tamanho da amostra, e sempre avaliar a associação observando conjuntamente o coeficiente de correlação e o diagrama de dispersão.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Exercício 1

Vamos estudar a relação entre idade e peso em uma amostra de crianças. Os dados estão na tabela ao lado.

d) Calcule o coeficiente de correlação linear de Pearson entre idade e peso. Interprete-o.

Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	8	64					
2	10	71					
3	6	53					
4	11	67					
5	8	55					
6	7	58					
7	10	77					
8	9	57					
9	10	56					
10	6	51					
11	12	76					
12	9	68					
Soma							
Média			----		----		----

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	8	64					
2	10	71					
3	6	53					
4	11	67					
5	8	55					
6	7	58					
7	10	77					
8	9	57					
9	10	56					
10	6	51					
11	12	76					
12	9	68					
Soma	106	753					
Média	8,833	62,750		----		----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	8	64	-0,833				
2	10	71	1,167				
3	6	53	-2,833				
4	11	67	2,167				
5	8	55	-0,833				
6	7	58	-1,833				
7	10	77	1,167				
8	9	57	0,167				
9	10	56	1,167				
10	6	51	-2,833				
11	12	76	3,167				
12	9	68	0,167				
Soma	106	753	0,000				
Média	8,833	62,750	0,000	----		----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	8	64	-0,833	0,694			
2	10	71	1,167	1,361			
3	6	53	-2,833	8,028			
4	11	67	2,167	4,694			
5	8	55	-0,833	0,694			
6	7	58	-1,833	3,361			
7	10	77	1,167	1,361			
8	9	57	0,167	0,028			
9	10	56	1,167	1,361			
10	6	51	-2,833	8,028			
11	12	76	3,167	10,028			
12	9	68	0,167	0,028			
Soma	106	753	0,000	39,667			
Média	8,833	62,750	0,000	----		----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	8	64	-0,833	0,694	1,250	1,563	
2	10	71	1,167	1,361	8,250	68,063	
3	6	53	-2,833	8,028	-9,750	95,063	
4	11	67	2,167	4,694	4,250	18,063	
5	8	55	-0,833	0,694	-7,750	60,063	
6	7	58	-1,833	3,361	-4,750	22,563	
7	10	77	1,167	1,361	14,250	203,063	
8	9	57	0,167	0,028	-5,750	33,063	
9	10	56	1,167	1,361	-6,750	45,563	
10	6	51	-2,833	8,028	-11,750	138,063	
11	12	76	3,167	10,028	13,250	175,563	
12	9	68	0,167	0,028	5,250	27,563	
Soma	106	753	0,000	39,667	0,000	888,250	
Média	8,833	62,750	0,000	----	0,000	----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	8	64	-0,833	0,694	1,250	1,563	-1,042
2	10	71	1,167	1,361	8,250	68,063	9,625
3	6	53	-2,833	8,028	-9,750	95,063	27,625
4	11	67	2,167	4,694	4,250	18,063	9,208
5	8	55	-0,833	0,694	-7,750	60,063	6,458
6	7	58	-1,833	3,361	-4,750	22,563	8,708
7	10	77	1,167	1,361	14,250	203,063	16,625
8	9	57	0,167	0,028	-5,750	33,063	-0,958
9	10	56	1,167	1,361	-6,750	45,563	-7,875
10	6	51	-2,833	8,028	-11,750	138,063	33,292
11	12	76	3,167	10,028	13,250	175,563	41,958
12	9	68	0,167	0,028	5,250	27,563	0,875
Soma	106	753	0,000	39,667	0,000	888,250	144,500
Média	8,833	62,750	0,000	----	0,000	----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Idade X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	8	64	-0,833	0,694	1,250	1,563	-1,042
2	10	71	1,167	1,361	8,250	68,063	9,625
3	6	53	-2,833	8,028	-9,750	95,063	27,625
4	11	67	2,167	4,694	4,250	18,063	9,208
5	8	55	-0,833	0,694	-7,750	60,063	6,458
6	7	58	-1,833	3,361	-4,750	22,563	8,708
7	10	77	1,167	1,361	14,250	203,063	16,625
8	9	57	0,167	0,028	-5,750	33,063	-0,958
9	10	56	1,167	1,361	-6,750	45,563	-7,875
10	6	51	-2,833	8,028	-11,750	138,063	33,292
11	12	76	3,167	10,028	13,250	175,563	41,958
12	9	68	0,167	0,028	5,250	27,563	0,875
Soma	106	753	0,000	39,667	0,000	888,250	144,500
Média	8,833	62,750	0,000	----	0,000	----	----

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{106}{12} = 8{,}833$$

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{39{,}667}{11} = 3{,}606$$

$$S_X = \sqrt{S_X^2} = \sqrt{3{,}606} = 1{,}899$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{753}{12} = 62{,}750$$

$$S_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{888{,}250}{11} = 80{,}750$$

$$S_Y = \sqrt{S_X^2} = \sqrt{80{,}75} = 8{,}986$$

$$cov(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{144{,}500}{11} = 13{,}136$$

$$r = \frac{cov(X,Y)}{S_x S_y} = \frac{13{,}136}{1{,}899 \cdot 8{,}986} = 0{,}76982$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Altura X_i	Peso Y_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	57	64	4,250	18,063	1,250	1,563	5,313
2	59	71	6,250	39,063	8,250	68,063	51,563
3	49	53	-3,750	14,063	-9,750	95,063	36,563
4	62	67	9,250	85,563	4,250	18,063	39,313
5	51	55	-1,750	3,063	-7,750	60,063	13,563
6	50	58	-2,750	7,563	-4,750	22,563	13,063
7	55	77	2,250	5,063	14,250	203,063	32,063
8	48	57	-4,750	22,563	-5,750	33,063	27,313
9	42	56	-10,750	115,563	-6,750	45,563	72,563
10	42	51	-10,750	115,563	-11,750	138,063	126,313
11	61	76	8,250	68,063	13,250	175,563	109,313
12	57	68	4,250	18,063	5,250	27,563	22,313
Soma	**633**	**753**	**0,000**	**512,250**	**0,000**	**888,250**	**549,250**
Média	**52,750**	**62,750**	**0,000**	----	**0,000**	----	----

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando o coeficiente de correlação linear de Pearson (r)

Criança	Altura X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	57	64	4,250	18,063	1,250	1,563	5,313
2	59	71	6,250	39,063	8,250	68,063	51,563
3	49	53	-3,750	14,063	-9,750	95,063	36,563
4	62	67	9,250	85,563	4,250	18,063	39,313
5	51	55	-1,750	3,063	-7,750	60,063	13,563
6	50	58	-2,750	7,563	-4,750	22,563	13,063
7	55	77	2,250	5,063	14,250	203,063	32,063
8	48	57	-4,750	22,563	-5,750	33,063	27,313
9	42	56	-10,750	115,563	-6,750	45,563	72,563
10	42	51	-10,750	115,563	-11,750	138,063	126,313
11	61	76	8,250	68,063	13,250	175,563	109,313
12	57	68	4,250	18,063	5,250	27,563	22,313
Soma	**633**	**753**	**0,000**	**512,250**	**0,000**	**888,250**	**549,250**
Média	**52,750**	**62,750**	**0,000**	----	**0,000**	----	----

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{633}{12} = 52{,}750$$

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{512{,}250}{11} = 46{,}568$$

$$S_X = \sqrt{S_X^2} = \sqrt{46{,}568} = 6{,}8240$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{735}{12} = 62{,}750$$

$$S_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{888{,}250}{11} = 80{,}750$$

$$S_Y = \sqrt{S_X^2} = \sqrt{80{,}75} = 8{,}986$$

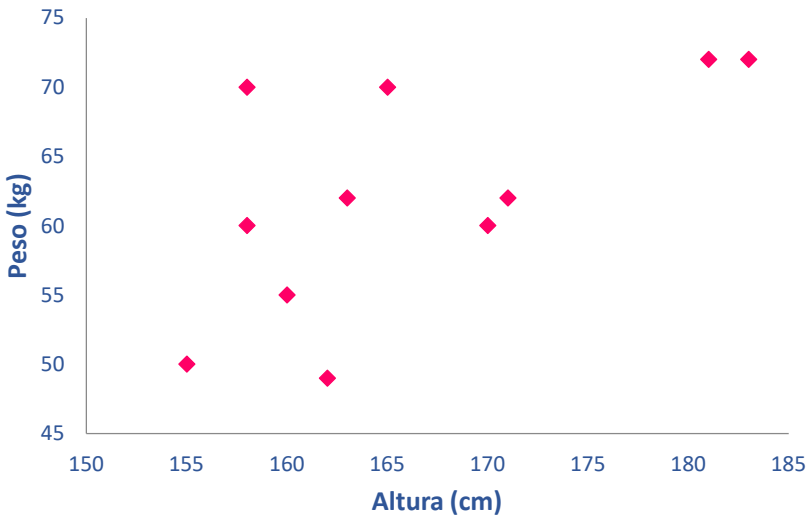
$$cov(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{549{,}250}{11} = 49{,}932$$

$$r = \frac{cov(X,Y)}{S_x S_y} = \frac{49{,}932}{6{,}824 \cdot 8{,}986} = 0{,}8143$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

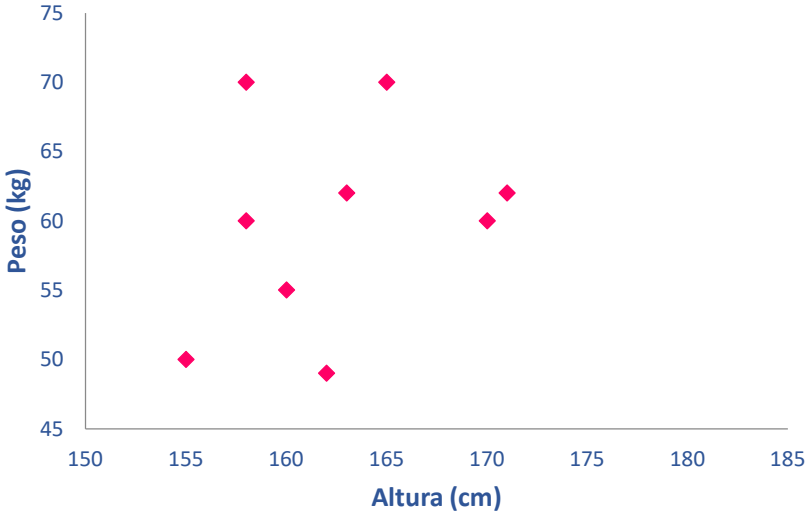
Coeficiente de correlação linear de Pearson (r)

Diagrama de dispersão do peso em função da altura em uma amostra de alunos da Farmácia



r=0,62

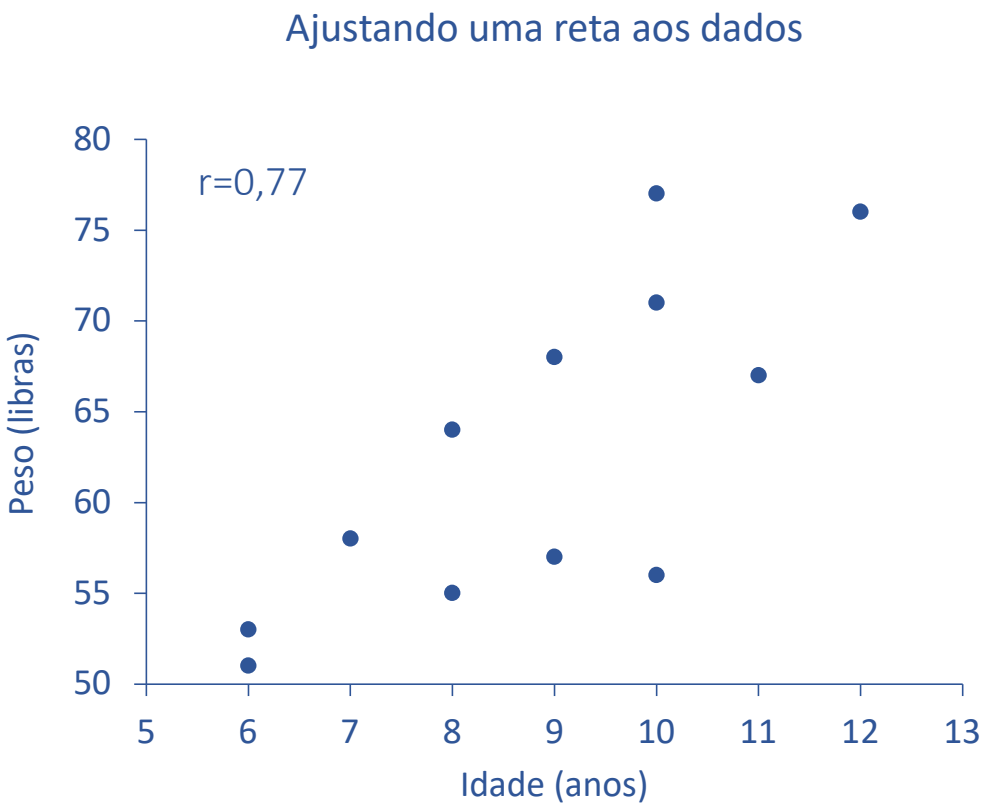
Diagrama de dispersão do peso em função da altura em uma amostra de alunos da Farmácia



r=0,28

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

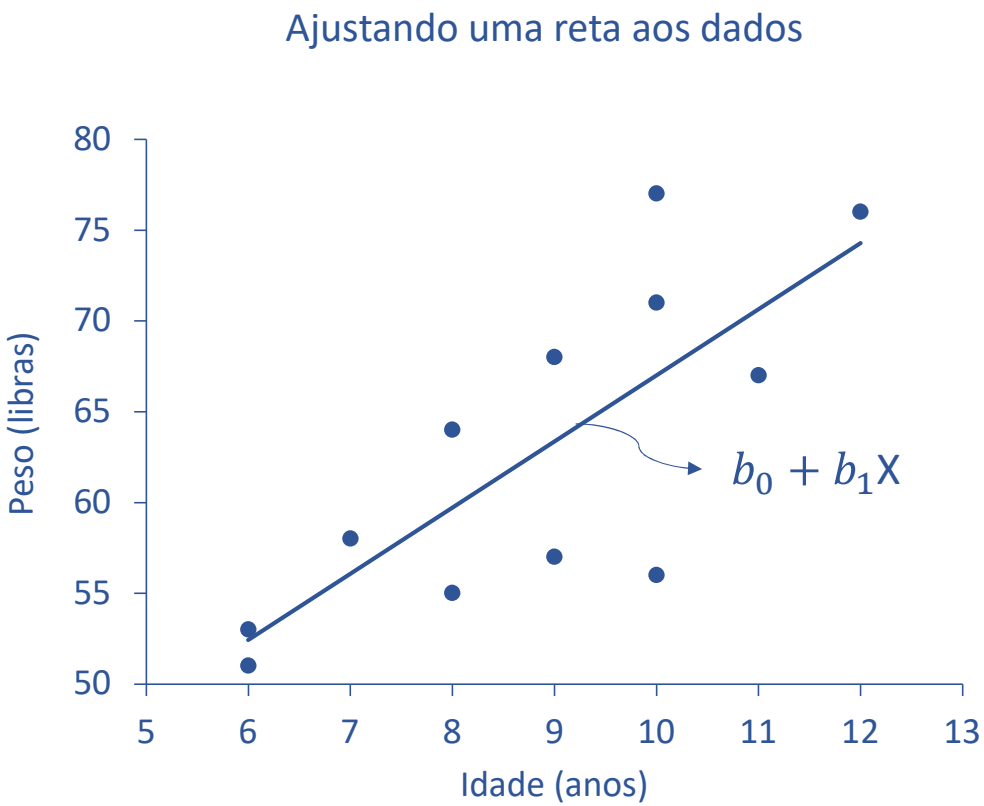
Análise de Regressão Simples



Criança	Idade (anos)	Peso (libras)
1	8	64
2	10	71
3	6	53
4	11	67
5	8	55
6	7	58
7	10	77
8	9	57
9	10	56
10	6	51
11	12	76
12	9	68

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

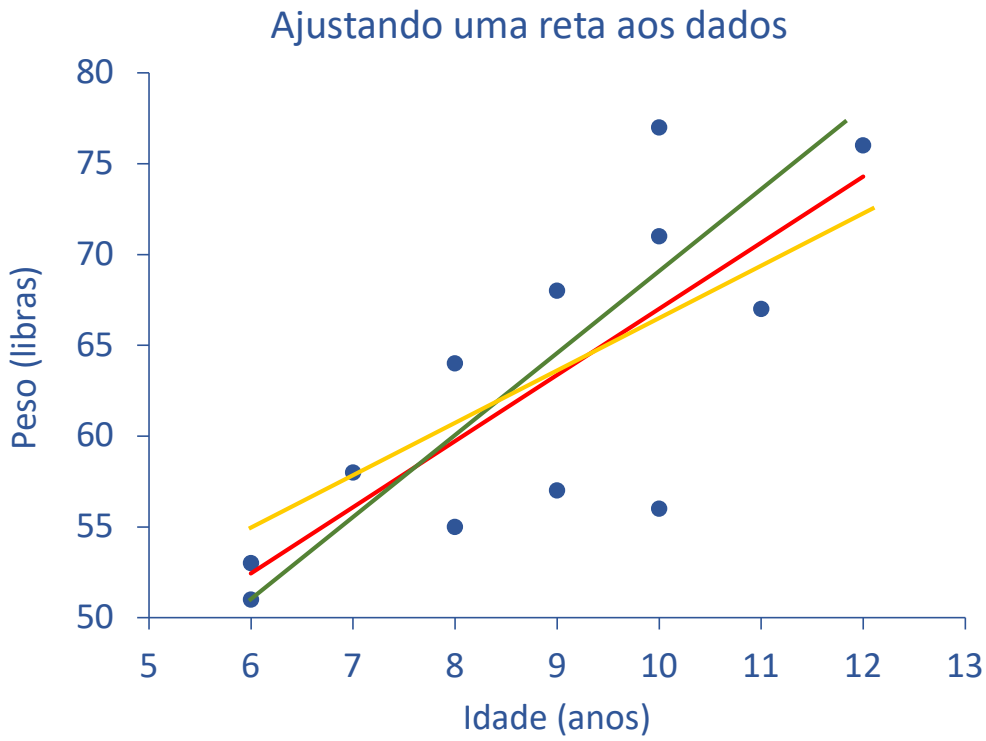
Análise de Regressão Simples



Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Análise de Regressão Simples



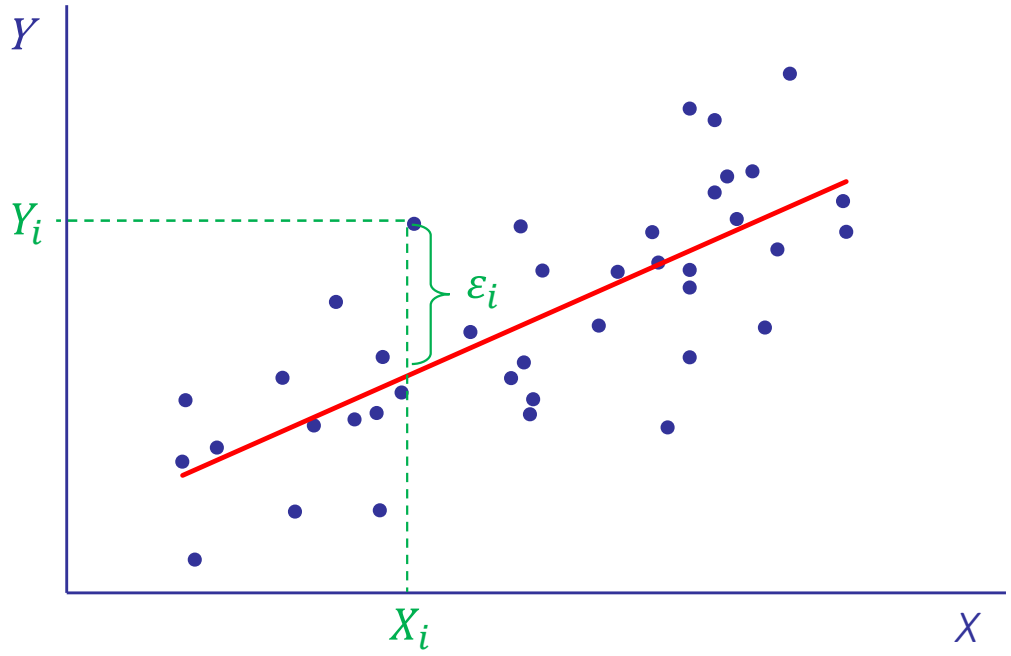
Como escolher a melhor reta?

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Na população $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$; independentes

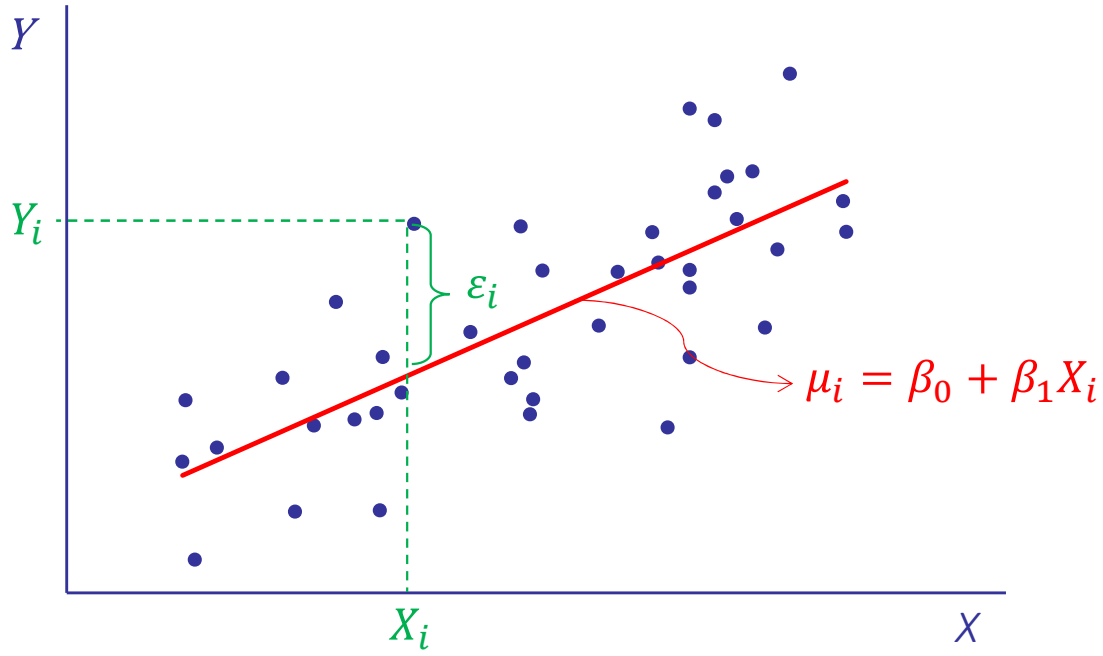


GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Na população $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$; independentes



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

O modelo de regressão linear simples pode ser escrito como:

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

onde

Y_i é o valor da variável resposta na i-ésima observação;

β_0 *e* β_1 são parâmetros;

X_i é uma constante conhecida, o valor da variável preditora (ou explicativa) na i-ésima observação;

ε_i é um erro aleatório não observável

$\varepsilon_i \sim N(0, \sigma^2)$; $(\varepsilon_i, \varepsilon_j)$ são independentes para todo *i, j*.

i = 1, ..., n

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Alguns aspectos importantes:

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

Este modelo é

- ✓ simples
- ✓ linear nos parâmetros
- ✓ linear na variável preditora

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Alguns aspectos importantes:

O modelo é escrito como

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

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independentes}$$

Mas também pode ser escrito como

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

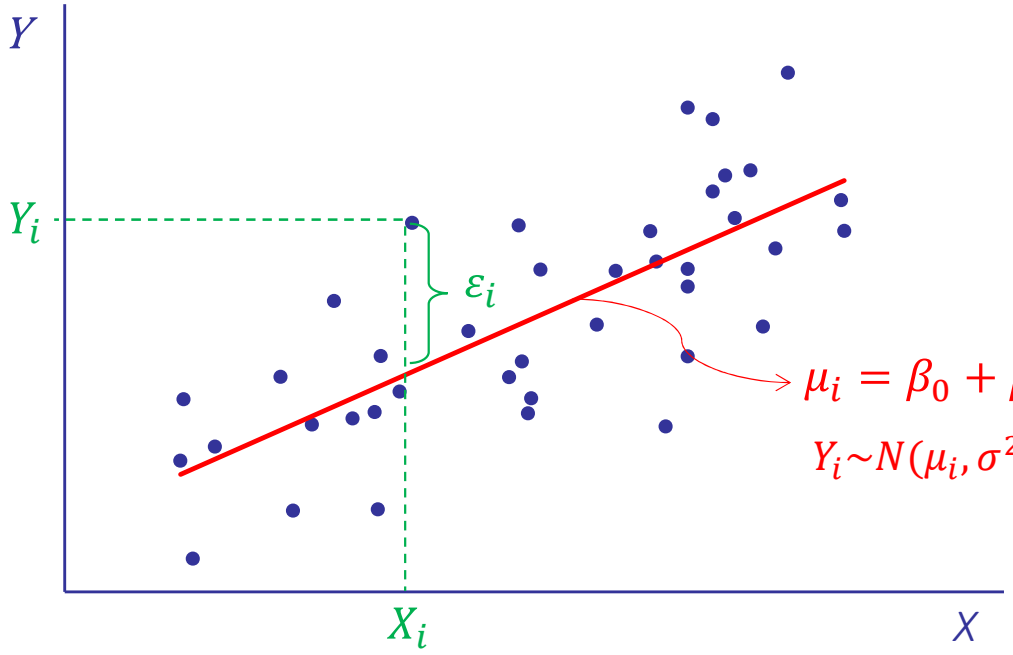
$$onde\ Y_i \sim N(\mu_i;\ \sigma^2), \text{independentes}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Na população $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$; independentes



$\mu_i = \beta_0 + \beta_1 X_i$

$Y_i \sim N(\mu_i, \sigma^2)$; independentes

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Entendendo isto:

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

$\varepsilon_i \sim N(0, \sigma^2)$, independentes

✓ Obtendo a esperança e a variância de Y:

$$E(Y_i) = \mu_i = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 X_i) + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$$

$$Var(Y_i) = Var(\beta_0 + \beta_1 X_i + \varepsilon_i) = Var(\beta_0) + Var(\beta_1 X_i) + Var(\varepsilon_i) = \sigma^2$$

✓ Como Y_i é a soma de uma constante $(\beta_0 + \beta_1 X_i)$ e uma variável aleatória (ε_i) com distribuição Normal, onde $(\varepsilon_i, \varepsilon_j)$ são independentes para todo *i, j*, então, Y_i~Normal com (Y_i, Y_j) independentes para todo *i, j*, .

✓ Assim, o modelo equivale a

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

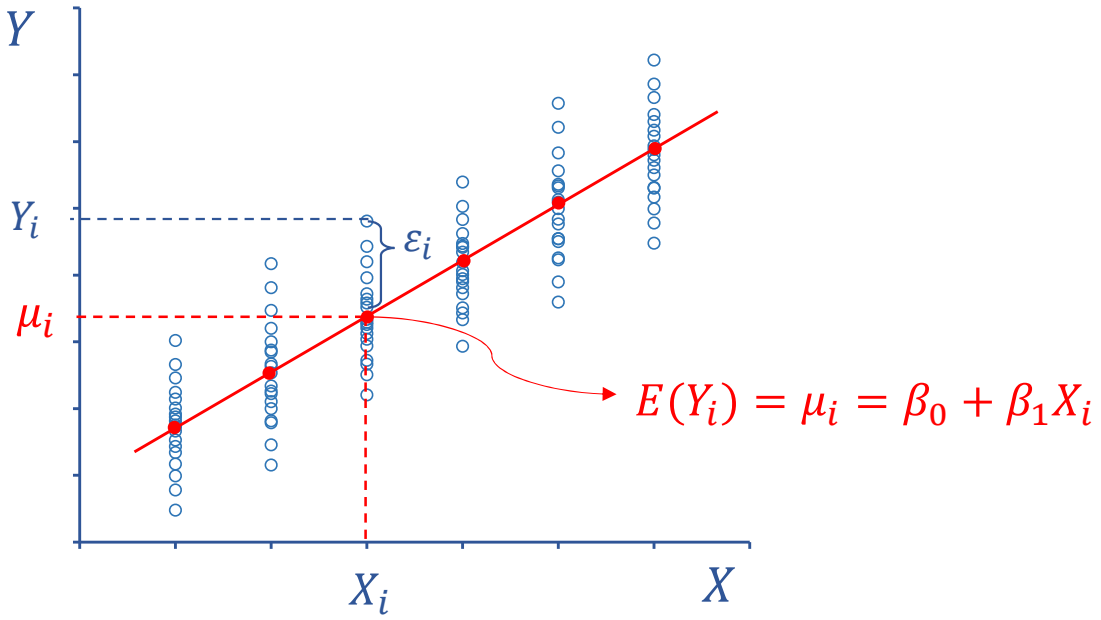
$onde\ Y_i \sim N(\mu_i;\ \sigma^2)$, independentes

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Na população

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2);\ \text{independentes}$$



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Alguns aspectos importantes:

✓ O Modelo de Regressão Linear Simples é dados por:

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

$\varepsilon_i \sim N(0, \sigma^2)$, independentes

✓ Este modelo modelo equivale a

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

$onde\ Y_i \sim N(\mu_i;\ \sigma^2)$, independentes

✓ Como X_i é constante, também é comum utilizar a notação

$$E(Y_i|X_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (3)$$

$onde\ Y_i/X_i \sim N(\mu_i;\ \sigma^2)$, independentes

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

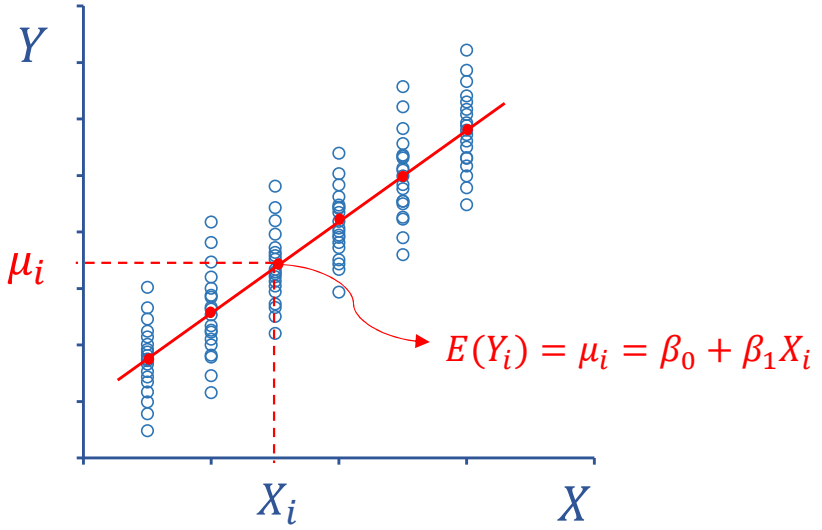
Entendendo as suposições do modelo

1. Distribuição Normal para a variável resposta

Para um valor fixo de X, Y é uma v.a. com distribuição normal.

Então, há uma distribuição Normal para Y em cada nível de X.

$$Y_i/X_i \sim N(\mu_i;\ \sigma^2)$$



2. Os valores de Y são independentes uns dos outros.

(Y_i, Y_j) são independentes para todo *i, j*.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Entendendo as suposições do modelo

3. Linearidade

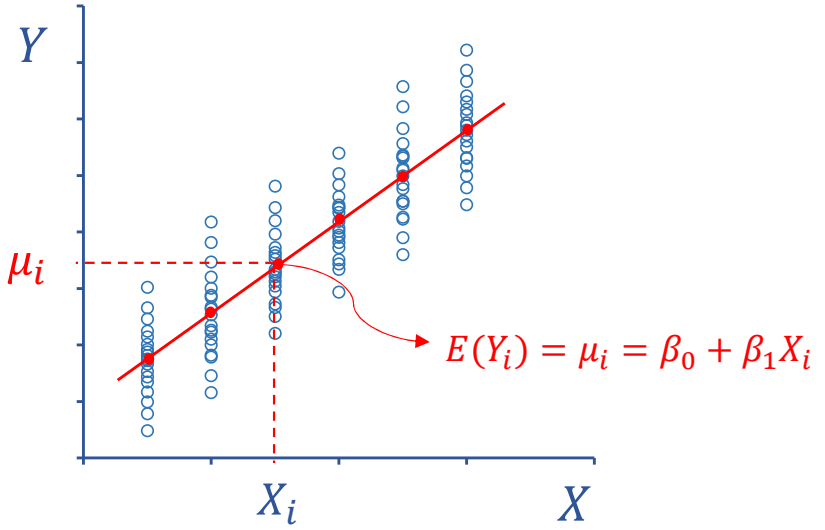
O valor esperado ou médio de Y_i/X_i , que é μ_i,

é uma função de linha reta sobre os X_i.

4. Homocedasticidade

A variância de Y_i/X_i é a mesma, qualquer que seja X_i.

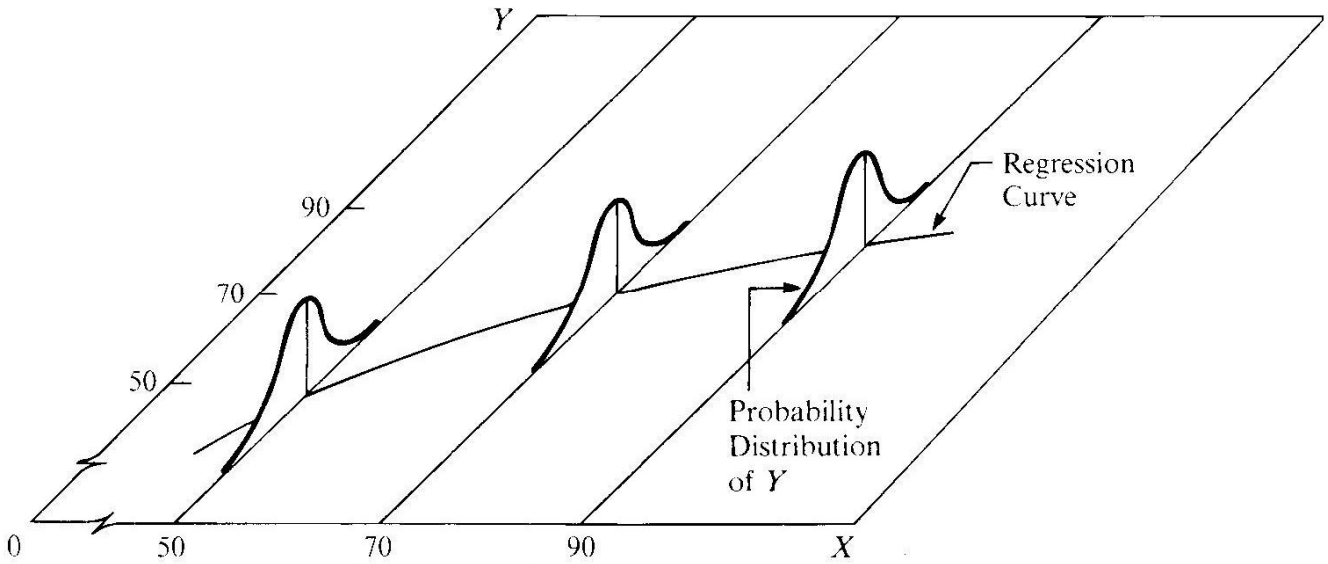
$Var(Y_i/X_i) = \sigma^2$, constante.



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

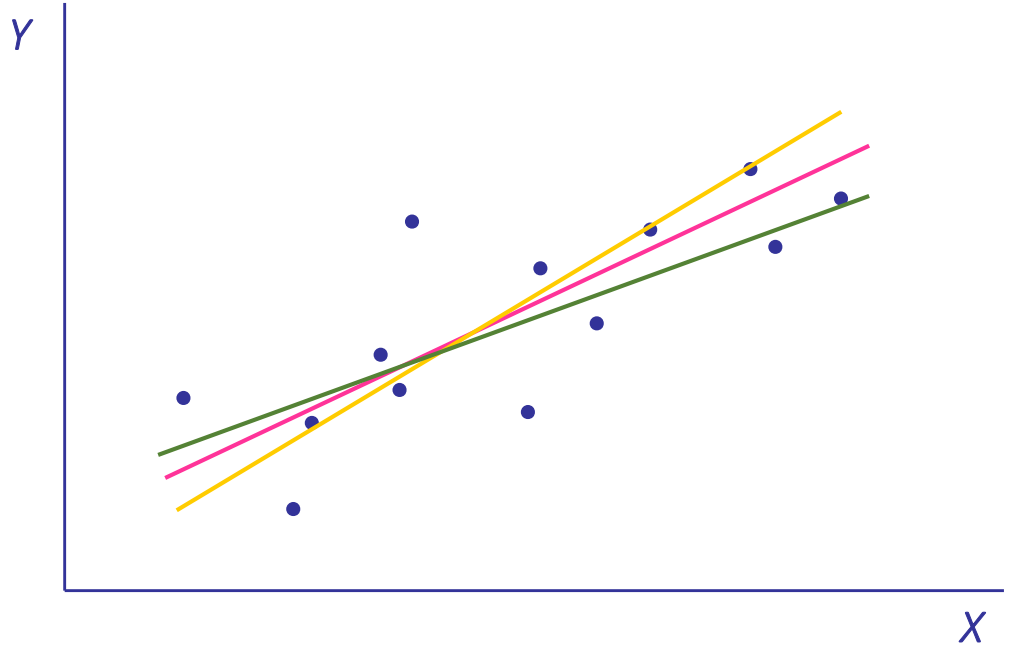
Entendendo as suposições do modelo



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimando a reta de regressão

Na amostra



Como escolher a melhor reta?

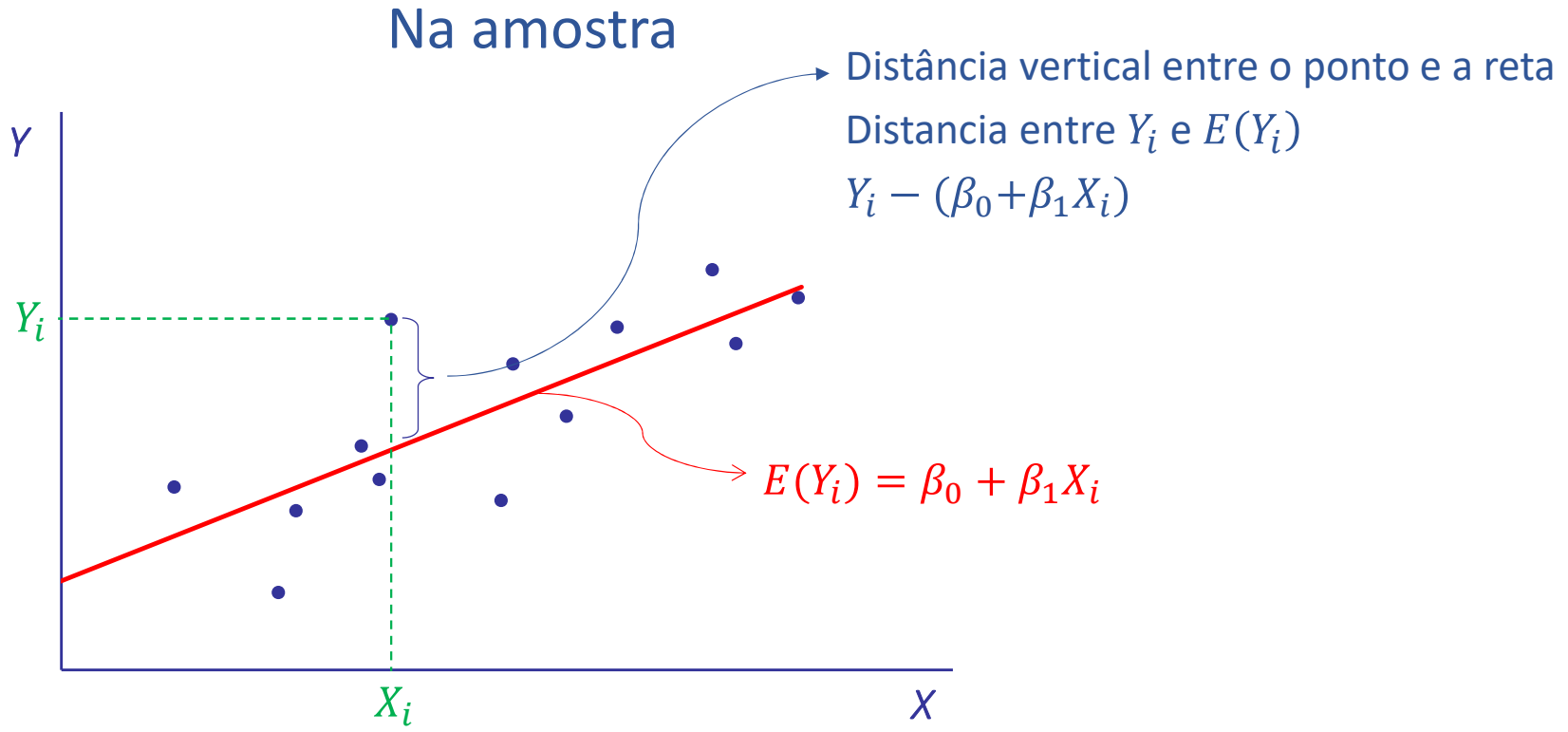
GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimando a reta de regressão

- ✓ Como escolher a melhor reta?
- ✓ Como estimar os valores de β_0 e β_1?
- ✓ Métodos de estimação
  - Mínimos Quadrados
  - Máxima Verossimilhança

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimando a reta de regressão



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Método de Mínimos Quadrados

Distância vertical entre o ponto e a reta

Distancia entre Y_i e $E(Y_i)$

Para cada ponto: $Y_i - (\beta_0 + \beta_1 X_i)$

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)\,]^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i\,]^2$$

Para encontrar os valores de β_0 e β_1 que tornam Q a mínima possível:

$$\frac{\partial Q}{\partial \beta_0} = 0 \qquad \frac{\partial Q}{\partial \beta_1} = 0$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Método de Mínimos Quadrados

Chamaremos esses valores de b_0 e b_1:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

A reta de regressão estimada será:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (4)$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Calculando b_0 e b_1

Criança	Idade X_i	Peso Y_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	8	64	-0,833	0,694	1,250	1,563	-1,042
2	10	71	1,167	1,361	8,250	68,063	9,625
3	6	53	-2,833	8,028	-9,750	95,063	27,625
4	11	67	2,167	4,694	4,250	18,063	9,208
5	8	55	-0,833	0,694	-7,750	60,063	6,458
6	7	58	-1,833	3,361	-4,750	22,563	8,708
7	10	77	1,167	1,361	14,250	203,063	16,625
8	9	57	0,167	0,028	-5,750	33,063	-0,958
9	10	56	1,167	1,361	-6,750	45,563	-7,875
10	6	51	-2,833	8,028	-11,750	138,063	33,292
11	12	76	3,167	10,028	13,250	175,563	41,958
12	9	68	0,167	0,028	5,250	27,563	0,875
Soma	106	753	0,000	39,667	0,000	888,250	144,500
Média	8,833	62,750	0,000	----	0,000	----	----

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_1 = \frac{144{,}5000}{39{,}667} = 3{,}64$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 62{,}750 - 3{,}64 * 8{,}833 = 30{,}57$$

$$\hat{Y}_i = 30{,}5714 + 3{,}6429X_i$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Interpretação dos coeficientes do modelo ajustado

$$\hat{Y} = b_0 + b_1 X$$

Para interpretar os parâmetros do modelo, variamos X e vemos o que acontece com $\hat{Y}$

✓ Fazendo X=0:

$$X = 0 \Rightarrow \hat{Y}_{(X=0)} = b_0 + b_1 * 0 \ = b_0$$

$\Rightarrow b_0$ é o valor de $\hat{Y}$ quando $X = 0$

✓ Aumentando X de uma unidade:

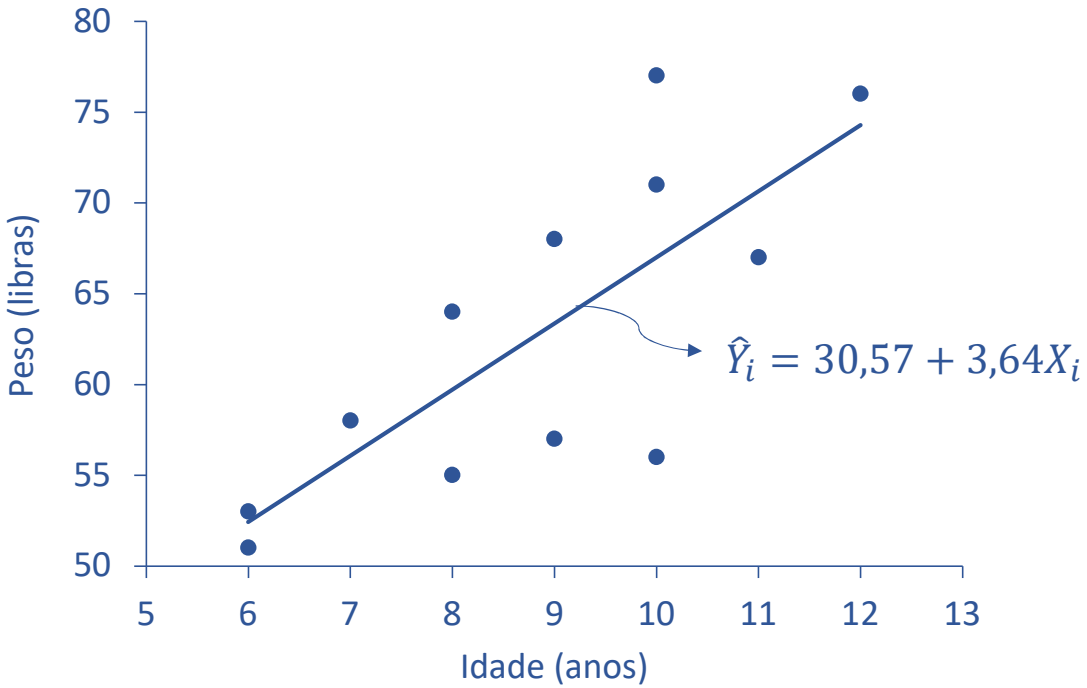
$$\hat{Y}_{(X+1)} = b_0 + b_1(X + 1) \ = b_0 + b_1 X + b_1 \ = \hat{Y} + b_1$$

$\Rightarrow b_1$ é o aumento médio ou esperado em $\hat{Y}$ quando aumentamos X de uma unidade.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Modelo ajustado



Interpretação dos coeficientes ajustados

- ✓ Neste caso, b_0 não tem interpretação, já que não temos crianças com idade zero.
- ✓ A cada ano que passa, espera-se um aumento de o 3,64 libras no peso, em média.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Interpretação dos coeficientes ajustados

b_0 e b_1 são as "estimativas" de β_0 e β_1

Na população: $Y = \beta_0 + \beta_1 X + \varepsilon$ ou

$\mu = \beta_0 + \beta_1 X$

Na amostra: $\hat{Y} = b_0 + b_1 X$

- ✓ b_0 é a **estimativa** do valor esperado ou médio de Y quando $X = 0$
- ✓ b_1 é a **estimativa** do aumento esperado ou médio em Y quando aumentamos X de uma unidade
- ✓ Em geral, omitimos a palavra "estimativa" na interpretação, já que isto está implícito

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Interpretação dos parâmetros de modelo

$$E(Y) = \mu = \beta_0 + \beta_1 X$$

Para interpretar os parâmetros do modelo, variamos X e vemos o que acontece com μ.

✓ Fazendo X=0:

$$X = 0 \Rightarrow \mu_0 = \beta_0 + \beta_1 * 0 = \beta_0$$

$\Rightarrow \beta_0$ é o valor de μ quando $X = 0$

✓ Aumentando X de uma unidade:

$$\mu_{(X+1)} = \beta_0 + \beta_1(X + 1) = \beta_0 + \beta_1 X + \beta_1 = \mu + \beta_1$$

$\Rightarrow \beta_1$ é o aumento em μ (isto é, o aumento esperado ou médio em Y) quando aumentamos X de uma unidade.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

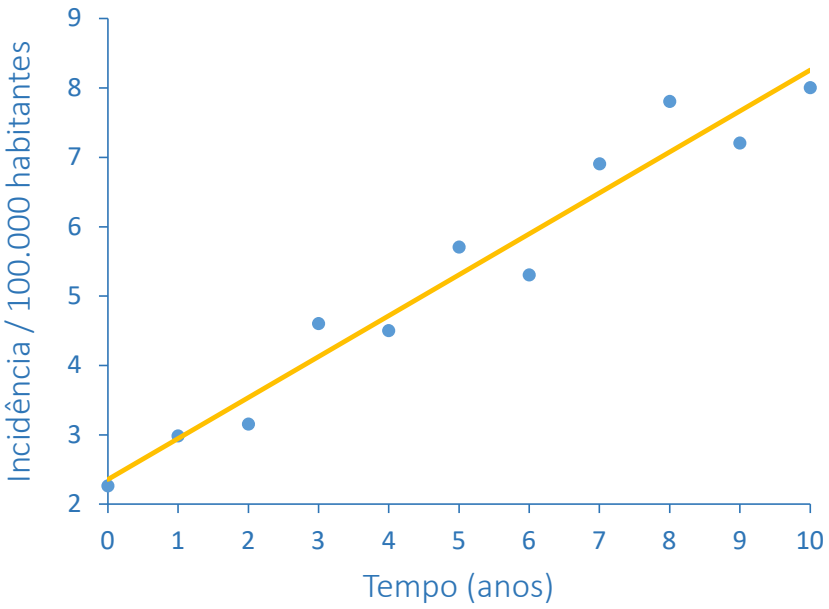
Interpretação dos coeficientes ajustados

Por exemplo, ao investigar a incidência de uma doença (Y) ao longo dos anos (X), o modelo ajustado foi:

$$\hat{Y}_i = 2{,}18 + 1{,}3Ano$$

⇒ A estimativa da incidência média no ano zero (início do estudo) é 2,18

⇒ Estima-se que a cada ano que passa a incidência aumenta, em média, 1,3.



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Exercício 1

Vamos estudar a relação entre idade e peso em uma amostra de crianças. Os dados estão na tabela ao lado.

e) Obtenha a reta de regressão do peso em função da idade.

f) Interprete os coeficientes da reta de regressão ajustada.

g) Desenhe a reta de regressão ajustada no diagrama de dispersão.

h) Obtenha as estimativas para o peso esperado (ou médio) em crianças com 8 e 11 anos.

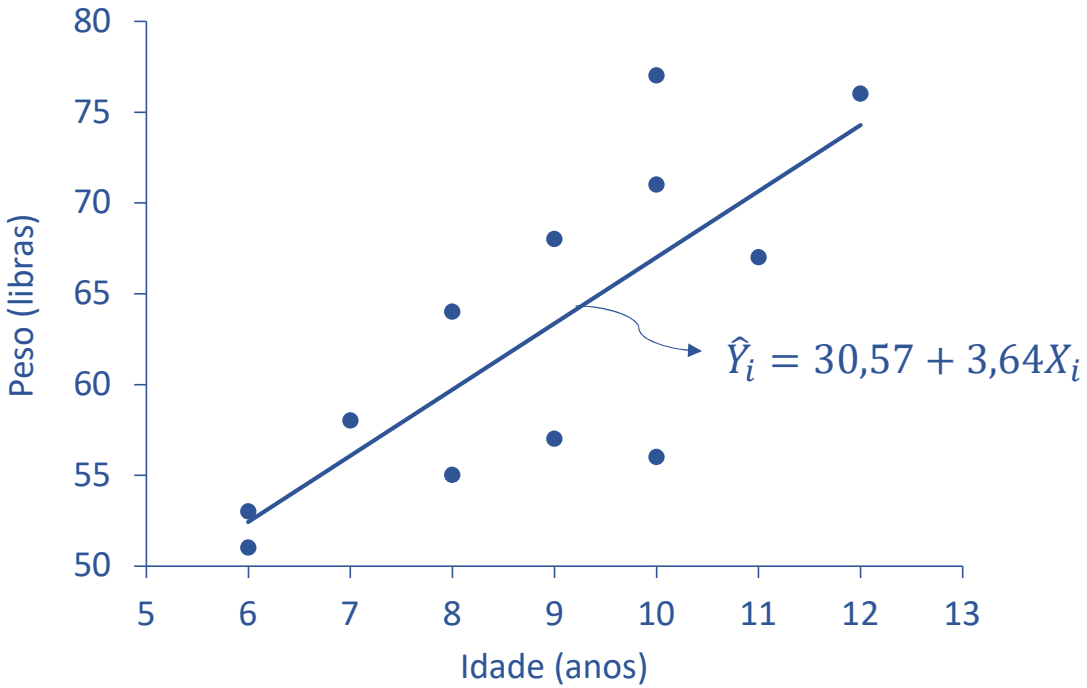
Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Modelo ajustado



Para traçar a reta no gráfico:

Para $X_i = 7$

$\hat{Y}_i = 30{,}57 + 3{,}64*7 = 56{,}07$

Para $X_i = 10$

$\hat{Y}_i = 30{,}57 + 3{,}64*10 = 66{,}97$

(7; 56,07)

(10; 66,97)

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Algumas propriedades

$$\hat{Y}_i = b_0 + b_1 X_i \qquad \begin{cases} b_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ b_0 = \bar{Y} - b_1 \bar{X} \end{cases}$$

Note que, substituindo-se o valor de b_0 na equação acima, temos:

$$\hat{Y}_i = \bar{Y} - b_1 \bar{X} + b_1 X_i$$

$$\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X})$$

Isto singifica que quando

$$X_i \to \bar{X} \;\Rightarrow\; \hat{Y}_i \to \bar{Y}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Resíduo

O modelo ajustado é:

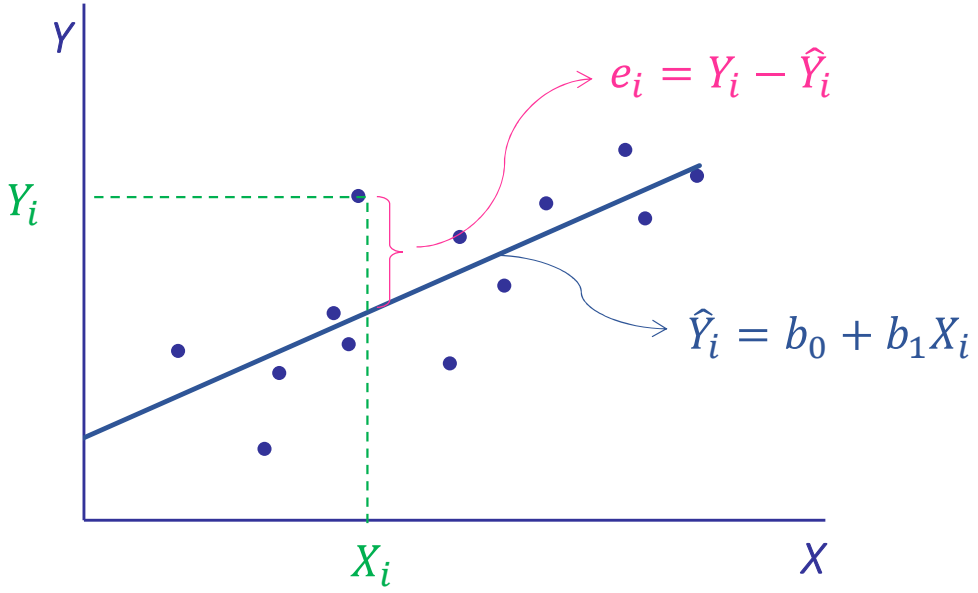
$$\hat{Y}_i = b_0 + b_1 X_i \qquad (4)$$

O resíduo é definido como a distância entre o valor observado Y_i e o valor ajustado pelo modelo, $\hat{Y}_i$.

Assim, os resíduos do modelo são dados por:

$$e_i = Y_i - \hat{Y}_i \quad i = 1, \ldots, n \qquad (5)$$

Note que $e_i \neq \varepsilon_i$, falaremos sobre isto mais adiante.



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências sobre o modelo de regressão

Inferências sobre o modelo de regressão

- ✓Tabela de ANOVA
- ✓Teste F

Inferências acerca de β_1

- ✓Teste de hipóteses para β_1
- ✓Intervalo de confianca para β_1

Inferências acerca de β_0

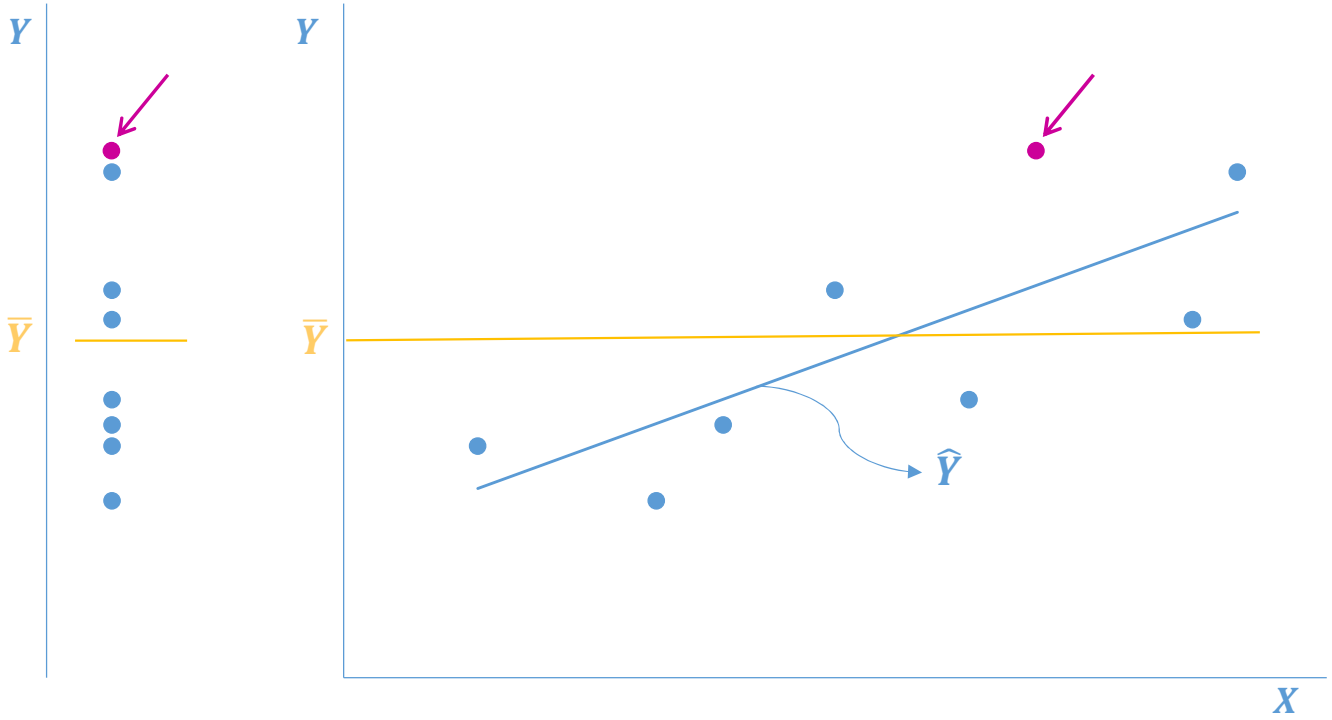
- ✓Teste de hipóteses para β_0
- ✓Intervalo de confianca para β_0

Estimação e Predição

- ✓Intervalo de Estimação
- ✓Intervalo de predição
- ✓Bandas de confiança para a reta de regressão

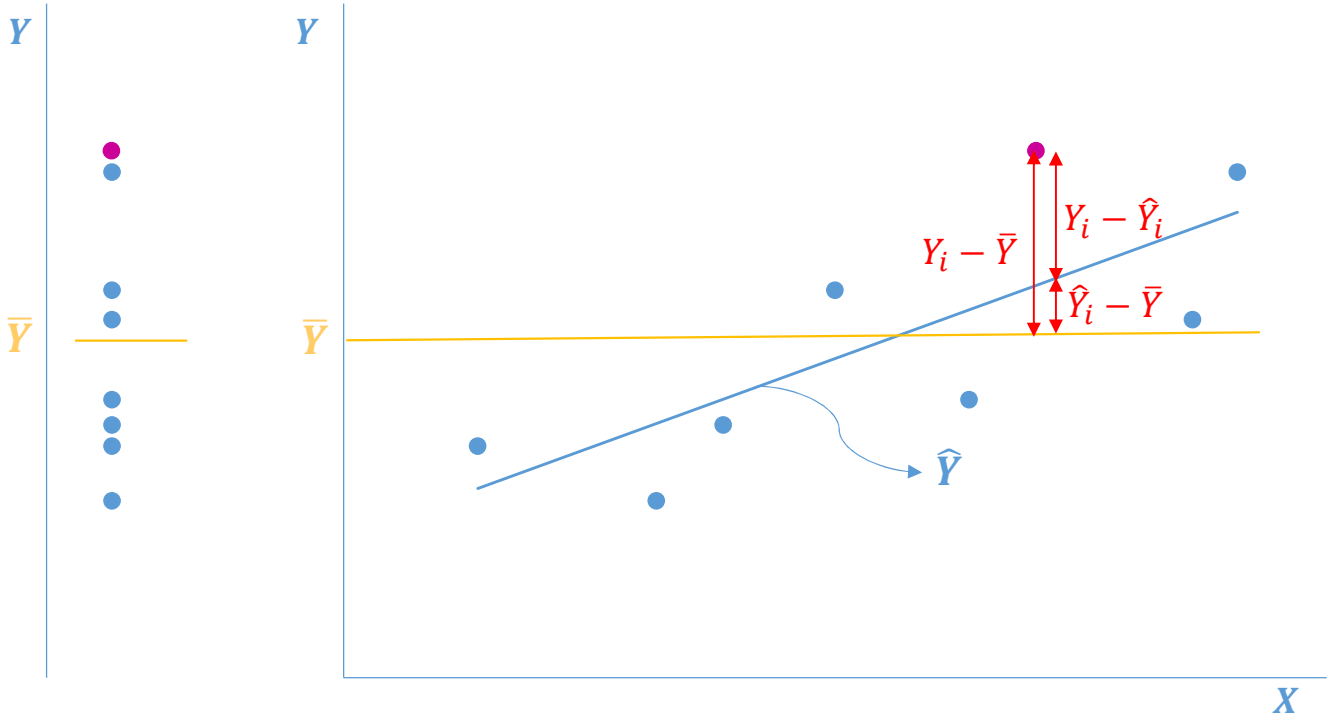
GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Fontes de variabilidade



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Fontes de variabilidade



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

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Partição da Soma de Quadrados

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right) \qquad (6)$$

Elevando-se ao quadrado os 2 lados da igualdade acima e fazendo-se a soma de todas as n equações (i=1,2, ...,n), obtem-se:

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

Após trabalhar o lado direito da equação, chegamos a:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \qquad (7)$$

$$SQT = SQM + SQR$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Partição da Soma de Quadrados

A Soma de Quadrados Total (SQT) é dada por

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

e fornece uma medida da variabilidade total das observações em relação à média geral.

O número de graus de liberdade (g.l.) associado à SQT é (n-1).

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Partição da Soma de Quadrados

A Soma de Quadrados do Modelo de Regressão (SQM) é dada por

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

e fornece uma medida da variabilidade entre a média geral e a média estimada pela reta de regressão. Quanto mais distante a reta de regressão estiver da média geral, maior será a contribuição do modelo para explicar a variabilidade de Y e maior será SQM.

O número de graus de liberdade (g.l.) associado à SQM é 1 (o número de parâmetros menos 1).

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Partição da Soma de Quadrados

A Soma de Quadrados do Resíduo (SQR) é dada por

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

e fornece uma medida da variabilidade das observações em torno da reta de reta de regressão.

Quanto mais próximas as observações estiverem da reta, menor será SQR.

O número de graus de liberdade (g.l.) associado à SQR é (n-2).

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Equação Fundamental da Regressão

A equação (7) é chamada Equação Fundamental da Regressão e postula que:

$$SQT = SQM + SQR$$

- ✓ Isto é, a soma dos quadrados sobre a média (SQT) = soma de quadrados devida à regressão (SQM) + soma de quadrados sobre a regressão (SQR).
- ✓ Isso significa que a variação total dos Y's sobre sua média pode ser explicada, em parte, pela linha de regressão e, em parte, pelos resíduos. Se todos os Y's caíssem sempre na linha de regressão a SQR seria zero!!
- ✓ Portanto, quanto mais a SQM for próxima da SQT melhor.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Quadrados Médios

Dividindo SQM e SQR pelos correspondentes graus de liberdade obtemos, respectivamente, o quadrado médio da regressão (QMM) e o quadrado médio dos resíduos (QMR), isto é:

$$QMM = \frac{SQM}{1} \qquad QMR = \frac{SQR}{n-2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Teste de hipóteses

H_0: $\beta_1 = 0$ (Não existe associção entre X e Y)

H_1: $\beta_1 \neq 0$ (Existe associção entre X e Y)

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estatística do teste

Pode-se demonstrar que

$$E(QMM) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$E(QMR) = \sigma^2$$

- ✓ Se H_0 for verdadeira, β_1 será igual a zero (ou seja, não existe associação entre X e Y), E(QMM) e E(QMR) serão iguais.
- ✓ Neste caso, espera-se que o quociente $\frac{QMM}{QMR}$ seja próximo de 1.
- ✓ Um valor observado para $\frac{QMM}{QMR}$ próximo de 1 é uma indicação de que Ho é verdadeira.
- ✓ Um valor grande para esse quociente é uma indicação de que Ho é falsa.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estatística do teste

Pode-se demonstrar que

$$E(QMM) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$E(QMR) = \sigma^2$$

✓ O que é um valor grande para $\frac{QMM}{QMR}$?

✓ A estatística $\frac{QMM}{QMR} \sim F_{(1,n-2)}$, basta usar a tabela da F.

✓ Note que QMR é um estimador não viesado da variância σ^2.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Quadro de Análise de Variância

Fonte de variação	*g.l.*	*SQ*	*QM*	*E(QM)*	F_0	*p-valor*
Regressão	*1*	*SQM*	*QMM*	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$\frac{QMM}{QMR} \sim F_{(1,n-2)}$	
Resíduo	*n-2*	*SQR*	*QMR*	σ^2		
Total	*n-1*	*SQT*				

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Exercício 1

Vamos estudar a relação entre idade e peso em uma amostra de crianças. Os dados estão na tabela ao lado.

h) Obtenha a reta de regressão do peso em função da idade.

i) Interprete os coeficientes da reta de regressão ajustada.

j) Desenhe a reta de regressão ajustada no diagrama de dispersão.

k) Obtenha as estimativas para o peso esperado (ou médio) em crianças com 8 e 11 anos.

Criança	Peso (libras)	Idade (anos)
1	64	8
2	71	10
3	53	6
4	67	11
5	55	8
6	58	7
7	77	10
8	57	9
9	56	10
10	51	6
11	76	12
12	68	9

BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Somas de Quadrados

Criança	Idade X_i	Peso Y_i	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$\hat{Y}_i$	$\hat{Y}_i - \overline{Y}$	$(\hat{Y}_i - \overline{Y})^2$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
1	8	64	1,250	1,563	59,714	-3,036	9,216	4,286	18,367
2	10	71	8,250	68,063	67,000	4,250	18,063	4,000	16,000
3	6	53	-9,750	95,063	52,429	-10,321	106,532	0,571	0,327
4	11	67	4,250	18,063	70,643	7,893	62,297	-3,643	13,270
5	8	55	-7,750	60,063	59,714	-3,036	9,216	-4,714	22,224
6	7	58	-4,750	22,563	56,071	-6,679	44,603	1,929	3,719
7	10	77	14,250	203,063	67,000	4,250	18,063	10,000	100,000
8	9	57	-5,750	33,063	63,357	0,607	0,369	-6,357	40,413
9	10	56	-6,750	45,563	67,000	4,250	18,063	-11,000	121,000
10	6	51	-11,750	138,063	52,429	-10,321	106,532	-1,429	2,041
11	12	76	13,250	175,563	74,286	11,536	133,073	1,714	2,939
12	9	68	5,250	27,563	63,357	0,607	0,369	4,643	21,556
Soma	**106**	**753**	**0,000**	**888,250**	**753,000**	**0,000**	**526,393**	**0,000**	**361,857**
Média	**8,833**	**62,750**	**0,000**						

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

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

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

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Construindo o Quadro de Análise de Variância

Fonte de variação	*g.l.*	*SQ*	*QM*	F_0	*p-valor*
Regressão	1	526,4	526,4	14,5	0,003
Resíduo	10	361,9	36,2		
Total	11	888,3			

Hipóteses: H_0: $\beta_1 = 0$ (Não existe associção entre X e Y)

H_1: $\beta_1 \neq 0$ (Existe associção entre X e Y)

$$R^2 = \frac{526,4}{888,3} = 0,5926$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de determinação do Modelo

$$R^2 = \frac{SQM}{SQT}$$

- ✓ O R^2 Mede a proporção da variabilidade total que é explicada pelo modelo adotado.
- ✓ $0 \leq R^2 \leq 1$
- ✓ Quanto mais próximo de 1 estiver o R^2, mais *X* contribui para explicar a variabilidade de *Y* e para prever *Y*.
- ✓ Se $R^2 = 1$ e aceitarmos que $\beta_1 \neq 0$, todos os pontos cairiam em cima da reta e o ajuste seria perfeito (altamente improvável!!!).

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de determinação do Modelo

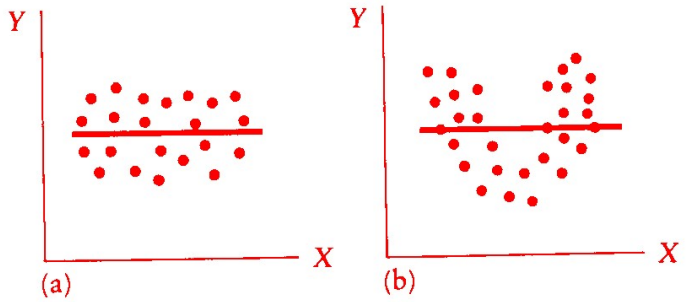
$$R^2 = \frac{SQM}{SQT}$$

- ✓ Por outro lado, se R^2 está próximo de 0 e aceitarmos que $\beta_1 = 0$, *X* não contribui para explicar a variabilidade de *Y,* ou para prever *Y*.
- ✓ No modelo de Regressão Linear Simples, o R^2 é igual ao coeficiente de correlação *(r)* ao quadrado, isto é, $R^2 = r^2$.
- ✓ Sua interpretação do exige cautela. Da mesma forma que o coeficiente de correlação, deve ser observado em conjunto com outras ferramentas, como o diagrama de dispersão, por exemplo.

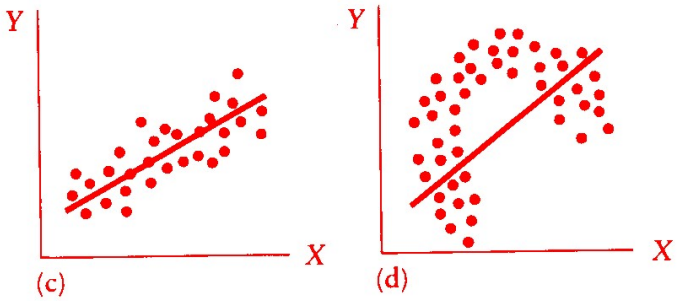
GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de determinação do Modelo

quando r2 é baixo



Examples when r^2 is high



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Coeficiente de determinação do Modelo

$$R^2 = \frac{SQM}{SQT}$$

O que R^2 não mede:

- ✓ a magnitude da inclinação de uma reta de regressão
- ✓ a linearidade da relação entre *Y* e *X*
- ✓ se o modelo está bem ajustado (quem faz isto é a análise de resíduos)

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

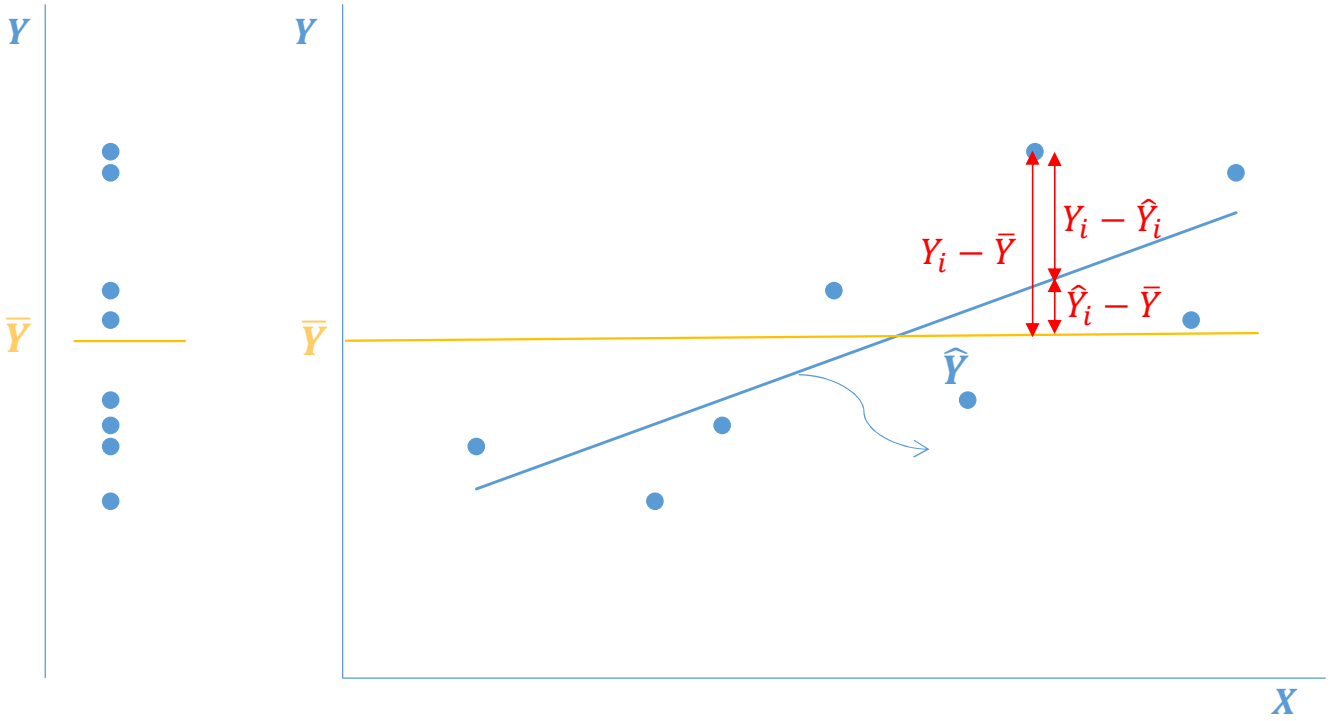
Estimador para σ^2

(1)

$$Y_i = \mu_i + \varepsilon_i$$
$$E(Y_i) = \mu_i$$

(2)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$$



A figura ao lado mostra a representação de dois modelos distintos para Y.

- ✓ No primeiro modelo, explicamos Y apenas pela sua média.
- ✓ No segundo, explicamos Y levando em conta a idade.
- ✓ Qual deles explica mais o comportamento de Y?
- ✓ Qual deles apresenta a menor variabilidade em torno da média?

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

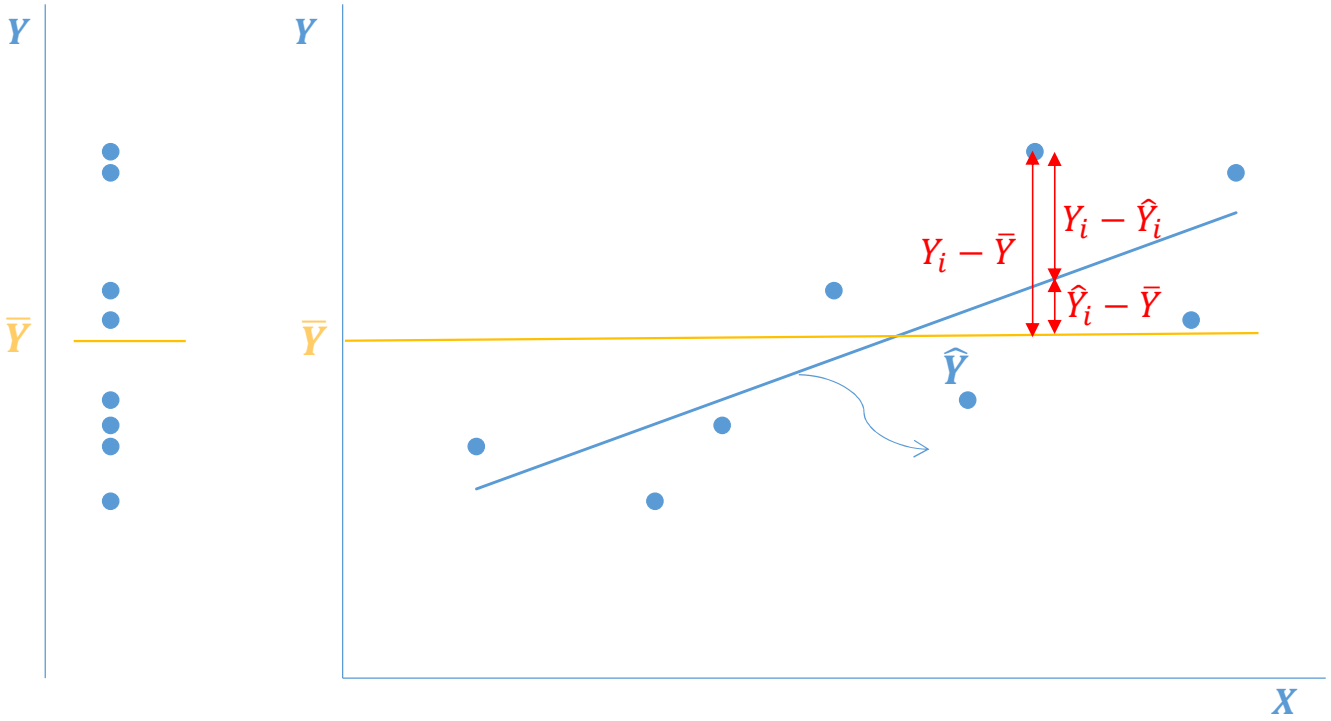
Estimador para σ^2

(1)

$$Y_i = \mu_i + \varepsilon_i$$
$$E(Y_i) = \mu_i$$

(2)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$$



BIOESTATÍSTICA

- ✓ No primeiro modelo, a variabilidade de *Y* em torno de sua média, é estimada por

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

que é um estimador não viesado para $VAR(Y)$.

- ✓ No segundo modelo, a variabilidade de *Y* em torno de sua média (que é a reta), é estimada por

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2} = QMR$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

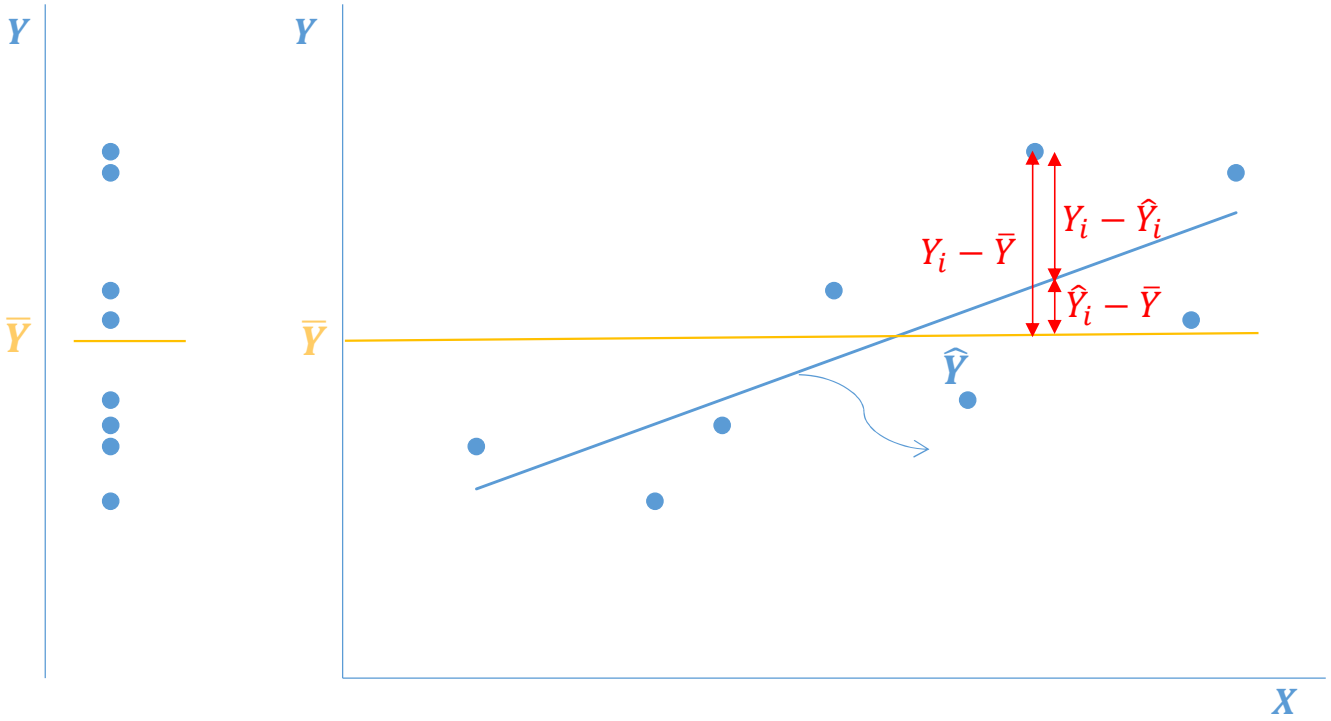
Estimador para σ^2

(1)

$$Y_i = \mu_i + \varepsilon_i$$
$$E(Y_i) = \mu_i$$

(2)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$$



BIOESTATÍSTICA

De fato, se o modelo estiver bem ajustado, o quadrado médio do resíduo (QMR) que é um estimador não viesado para $VAR(Y) = \sigma^2$, isto é:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2} = QMR$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_1

Teste de hipóteses para β_1

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

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

H_0: $\beta_1 = 0$ (Não existe associção entre X e Y)

H_1: $\beta_1 \neq 0$ (Existe associção entre X e Y)

Note que, se $\beta_1 = 0$, a reta de regressão é paralela ao eixo X e o modelo fica

$$Y_i = \beta_0 + \varepsilon_i$$

$$E(Y_i) = \mu_i = \beta_0$$

O estimador para β_1 é

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Pode-se demonstar que b_1 tem distribuição Normal com

$$E(b_1) = \beta_1$$

$$VAR(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_1

Teste de hipóteses para β_1

Como não conhecemos σ^2, nós o substituímos por $S_Y^2 = QMR$, e o estimador para a $VAR(b_1)$ fica

$$S_{b_1}^2 = \frac{QMR}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

De modo que

$$\frac{b_1 - \beta_1}{S_{b_1}} \sim t_{n-2}$$

Como, sob H_0, $\beta_1 = 0$,

a estatística do teste será

$$\frac{b_1}{S_{b_1}} \sim t_{n-2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_1

Intervalo de confiaça para β_1

Lembrando que o formato usual para um intervalo de confiança de um parâmetro é:

Estimador do parâmetro $\mp$ Quantil de uma distribuição X Desvio padrão do estimador

O intervalo de confiança para β_1, com coeficiente de confiança $(1-\alpha)$, será:

$$IC(\beta_1; 1-\alpha) = b_1 \mp t_{(1-\frac{\alpha}{2};\, n-2)} S_{b_1}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_0

Teste de hipóteses para β_0

Existem poucas situações nas quais desejamos fazer inferências sobre β_0 . Elas ocorrem, basicamente, as situações nas quais faz sentido que X assuma o valor zero e, preferencialmente, quando a faixa de valores observados de X inclui o valor zero.

H_0: $\beta_0 = 0$

H_1: $\beta_0 \neq 0$

O estimador para β_0 é

$$b_0 = \bar{Y} - b_1\bar{X}$$

Pode-se demonstar que b_0 tem distribuição Normal com

$$E(b_0) = \beta_0$$

$$VAR(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_0

Teste de hipóteses para β_0

Como não conhecemos σ^2, nós o substituímos por $S_Y^2 = QMR$, e o estimador para a $VAR(\beta_0)$ fica

$$S_{b_0}^2 = QMR\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

De modo que

$$\frac{b_0 - \beta_0}{S_b} \sim t_{n-2}$$

Como, sob H_0, $\beta_0 = 0$,

a estatística do teste será

$$\frac{b_0}{S_{b_0}} \sim t_{n-2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Inferências acerca de β_0

Intervalo de confiaça para β_0

Lembrando que o formato usual para um intervalo de confiança de um parâmetro é:

Estimador do parâmetro $\mp$ Quantil de uma distribuição X Desvio padrão do estimador

O intervalo de confiança para β_0, com coeficiente de confiança $(1-\alpha)$, será:

$$IC(\beta_1; 1-\alpha) = b_0 \mp t_{(1-\alpha;\, n-2)} S_{b_0}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimação e Predição

- ✓ Os principais objetivos da análise de regressão são a descrição e a previsão.
- ✓ A reta de regressão ajustada $(\hat{Y}_i = b_0 + b_1 X_i)$ fornece a descrição da relação linear entre a variável resposta e a explicativa e quantifica, por meio dos coeficientes estimados, a velocidade com a qual Y varia a partir de X.
- ✓ É possível fazer previsões sobre o valor de *Y* para um dado valor de *X* a partir da reta de regressão ajustada.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimação e Predição

Existem duas situações de interesse para as quais desejamos fazer previsões e obter intervalos de confiança:

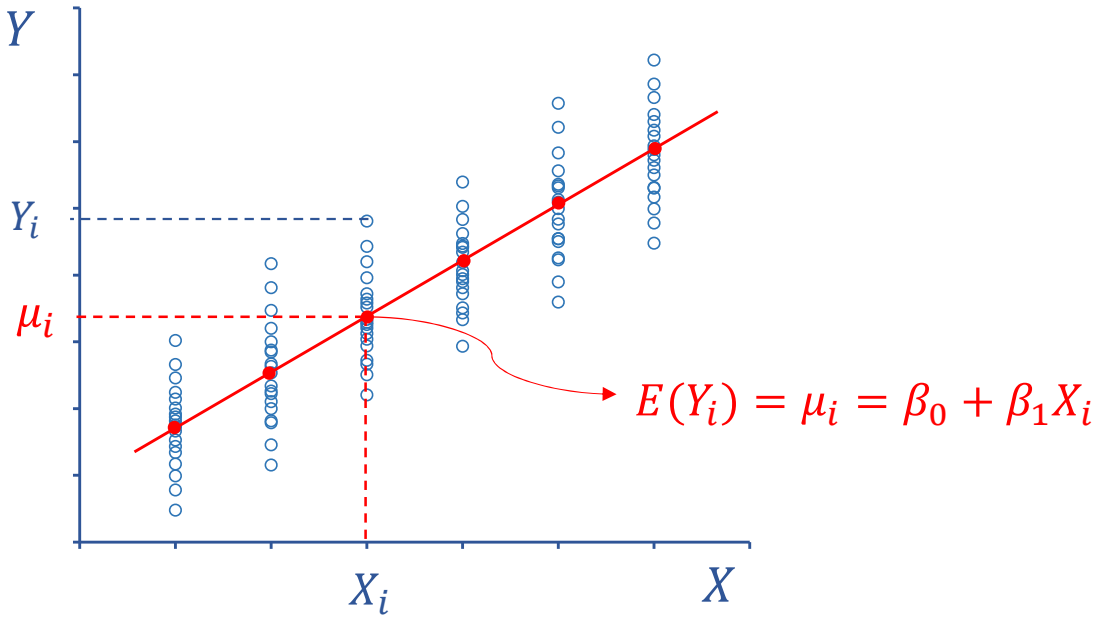
- ✓ A primeira envolve a previsão do valor médio de *Y* para um dado nível de *X*, isto é, a previsão de $E(Y/X)$ ou μ, dada por $\hat{Y}$, que pertence à reta de regressão estimada.
- ✓ A segunda envolve a previsão dos possíveis valores de *Y* que podem ser observados (e não a sua média) em um dado nível de *X*, isto envolve a distribuição de probabilidades de *Y* em torno da média μ.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O modelo de regressão linear simples

Na população

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2);\ \text{independentes}$$



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Estimação

Vamos chamar de X_h o valor de *X* para o qual queremos estimar a resposta média μ_h, ou $E(Y_h)$.

X_h pode ser um valor observado de *X* na amostra ou qualquer outro valor de *X* (observado ou não), preferencialmente dentro da faixa dos valores de *X* na amostra.

O estimador pontual de μ_h é, obviamente,

$$\hat{Y}_h = b_0 + b_1 X_h.$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Intervalo de estimação da média μ_h

O estimador por intervalo de μ_h é chamado de Intervalo de Estimação e é dado por:

$$IC(\mu_h; 1-\alpha) = \hat{Y}_h \mp t_{(1-\frac{\alpha}{2};n-2)} S_{\hat{Y}h}$$

$$S^2_{\hat{Y}_h} = QMR\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

A probabilidade de este intervalo conter a média μ_h é $1-\alpha$.

- ✓ Note que a variância de $\hat{Y}_h$ é proporcional a $X_h - \bar{X}$, que é a distância de X_h em relação à média de *X*.
- ✓ Ou seja, quanto mais próximo X_h estiver de $\bar{X}$, menor será a variância do estimador $\hat{Y}_h$ e menor será a a amplitude do seu intervalo de confiança.
- ✓ Analogamente, quanto mais distante X_h estiver de $\bar{X}$, maior será a variância do estimador $\hat{Y}_h$ e maior a amplitude do seu intervalo de confiança.
- ✓ Isto significa que as estimativas de $\hat{Y}$ na reta de regressão vão se tornando menos precisas à medida que *X* se distancia de $\bar{X}$.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Teste de hipóteses para a média μ_h

O teste de hipóteses para μ_h é da forma

$$H_0: \mu_h = \mu_0$$

$$H_1: \mu_h \neq \mu_0$$

A estatística do teste será:

$$t = \frac{\hat{Y}_h - \mu_0}{S_{\hat{Y}_h}} \sim t_{n-2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Predição

Agora estamos interessados na predição de um possível valor de *Y* que possa ser observado (e não a sua média) em um dado nível de *X*.

Novamente, seja X_h o valor de *X* para o qual desejamos prever um possível valor para *Y*. Este valor deve pertencer à distribuição de Y/X_h. Vamos denotar este novo valor por $Y_{h(novo)}$.

O estimador pontual de $Y_{h(novo)}$ é o mesmo:

$$\hat{Y}_{h(novo)} = b_0 + b_1 X_h.$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Intervalo de predição da observação Y_h

O intervalo que delimita, com $(1-\alpha)\%$ de confiança, onde os possíveis valores de *Y* pertencentes à distribuição Normal com média μ_h podem ocorrer é chamado

Intervalo de Predição e é dado por:

$$IC(Y_h; 1-\alpha) = \hat{Y}_h \mp t_{(1-\frac{\alpha}{2};n-2)} S_{pred}$$

$$S^2_{pred} = QMR\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

- ✓ Note que este intervalo é maior do que o anterior, já que pretende englobar não só a média μ_h, mas a distribuição dos valores de *Y* em torno da média.
- ✓ Novamente, quanto mais próximo X_h estiver de $\bar{X}$, menor será a variância do estimador $\hat{Y}_h$ e menor será a a amplitude do seu intervalo de confiança, e vice-versa.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Teste de hipóteses para $Y_{h(novo)}$

O teste de hipóteses para $Y_{h(novo)}$ é da forma

$$H_0: Y_{h(novo)} = Y_{h0}$$

$$H_1: Y_{h(novo)} \neq Y_{h0}$$

A estatística do teste será:

$$t = \frac{\hat{Y}_{h(novo)} - Y_{h0}}{S_{pred}} \sim t_{n-2}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Bandas de confiança para a reta de regressão

Frequentemente, estamos interessados em obter bandas de confiança para toda a reta de regressão. Tais bandas permitem visualizar toda uma região onde a verdadeira reta de regressão $(E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i)$ poderia estar, isto é, a região que com probabilidade $(1 - \alpha)$ contem a verdadeira reta.

Estas bandas são dadas por

$$\hat{Y}_h \mp W S_{\hat{Y}_h}$$

$$W^2 = 2F_{(1-\alpha;2,n-2)}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Bandas de confiança para a reta de regressão

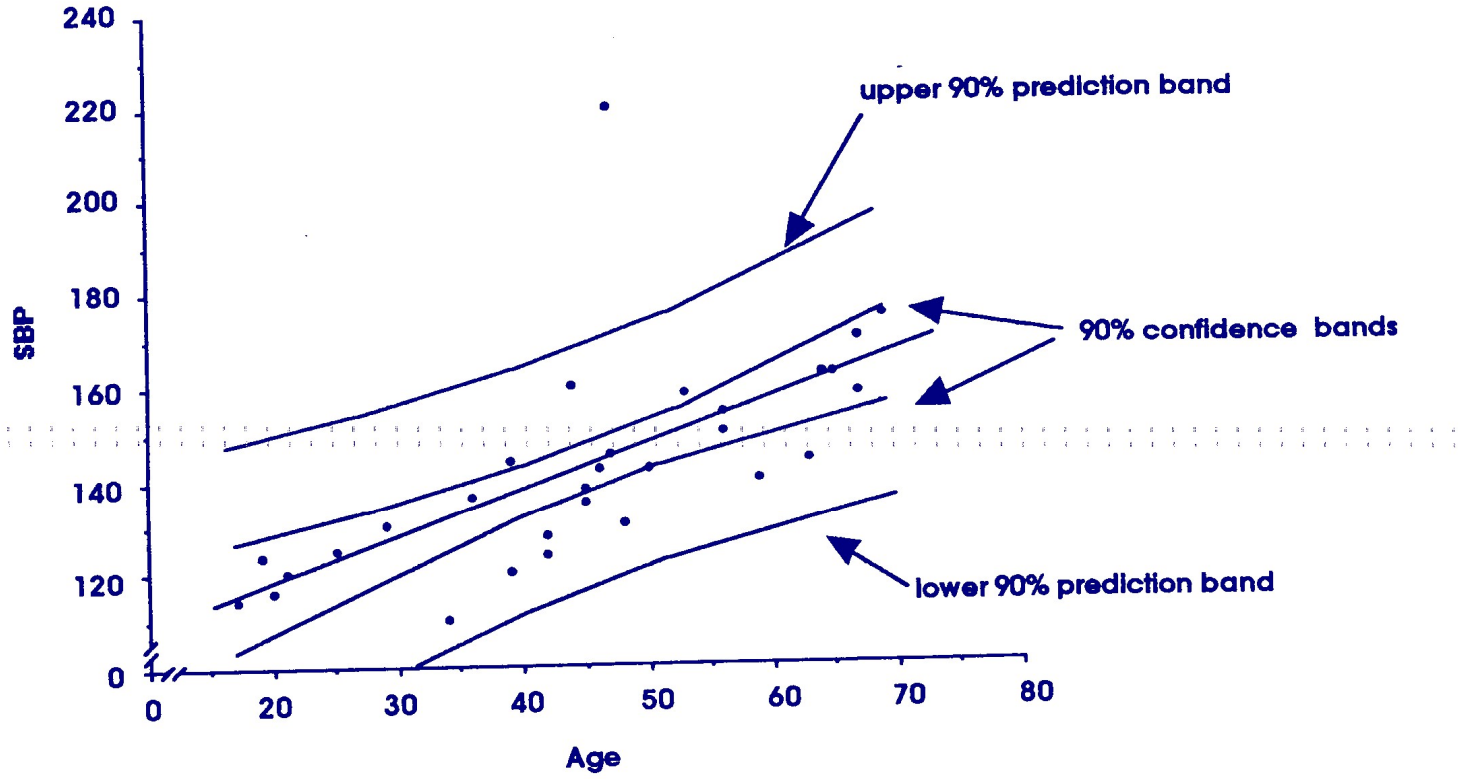
$$\hat{Y}_h \mp W S_{\hat{Y}_h}$$

$$W^2 = 2F_{(1-\alpha;2,n-2)}$$

- ✓ Note que a fórmula para as bandas de confiança é parecida com a do intervalo de estimação da reposta média μ_h para um dado valor X_h, exceto que a distribuição t foi substituida pela F.
- ✓ Com isto, as bandas terão amplitude maior do que o intervalo de estimação, o que faz sentido, já que devem compreender toda a reta de regressão e não apenas um único ponto, como no caso do intervalo de estimação.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

Bandas de confiança e de predição



GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O coeficiente de correlação e a análise de regressão

O coeficiente de correlação pode ser definido como

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

onde σ_{XY} é a covariância entre *X* e *Y*, definida como

$$\sigma_{XY} = E[X - E(X)][Y - E(Y)]$$

Anteriormente, aprendemos sobre *r* que, na verdade, é o estimador de ρ, obtido a partir da amostra.

$$\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \mathrm{r} = \frac{cov(X,Y)}{S_X S_Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O coeficiente de correlação e a análise de regressão

Não é difícil mostrar que

$$\rho = \frac{\sigma_X}{\sigma_Y}\beta_1 \quad \Rightarrow \quad \beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$$

Note que o sinal de ρ é o mesmo de β_1.

O mesmo vale para o estimador de ρ:

$$\hat{\rho} = r = \frac{S_X}{S_Y}b_1 \quad \Rightarrow \quad b_1 = r\frac{S_Y}{S_X}$$

O sinal de r é o mesmo de b_1.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O coeficiente de correlação e a análise de regressão

Teste de hipóteses para ρ

H_0: $\rho = 0$

H_1: $\rho \neq 0$

Isto equivale a testar as hipóteses

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

uma vez que $\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$

A estatística do teste é:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

✓ Note que, como H_0: $\rho = 0$ pode ser escrito inteiramente em termos de r e de n, pode-se realizar o teste de hipótese mesmo sem o ajuste de uma reta de regressão.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP

O coeficiente de correlação e a análise de regressão

Intervalo de confiança para ρ

Uma vez que distribuição de *r* é complicada quando $\rho \neq 0$, o intervalo de confiança é obtido por meio de uma transformação com aproximação pela Normal

$$z' = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$$

e o intervalo de confiança será dado por

$$IC(z', 1-\alpha) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \mp z_{(1-\frac{\alpha}{2})} \frac{1}{\sqrt{n-3}}$$

Uma vez obtido o intervalo para *z'*, é necessário transformar de volta para obter o intervalo para ρ.

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP - USP