Estatística II

Licenciatura em Gestão

2.º Ano/1.º Semestre

2023/2024

Aulas Teóricas N.ºs 18 e 19 (Semana 10)

Docente: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

Conteúdos Programáticos

Aulas Teóricas (Semanas 1 a 5)	Aulas Teóricas (Semanas 5 a 7)	Aulas Teóricas (Semanas 7 a 9)	Aulas Teóricas (Semanas 10 a 13)
• **Capítulo 1**: Estimação	• **Capítulo 2**: Testes de Hipóteses	• **Capítulo 3**: Modelo de Regressão Linear	• **Capítulo 4**: Complementos ao Modelo de Regressão Linear

Material didático: Exercícios do Livro Murteira et al (2015), Formulário e Tabelas Estatísticas

Bibliografia: B. Murteira, C. Silva Ribeiro, J. Andrade e Silva, C. Pimenta e F. Pimenta; *Introdução à Estatística*, 2ª ed., Escolar Editora, 2015.

7ª semana (31/10 a 02/11)

T12 - Teste de hipóteses

Testes em universos normais com amostras emparelhadas. Exemplo. Teste de hipóteses para grandes amostras. Aplicação ao universo de Bernoulli (média e diferença de médias). Exemplos.

T13 - Modelo de regressão linear

Introdução; modelo linear e linearizável; exemplos; Hipóteses básicas; estimação dos coeficientes da regressão pelos Mínimos Quadrados. Exemplo.

8ª semana (07/11 e 09/11)

T14 - Modelo de Regressão Linea (MRL)r

Interpretação dos parâmetros da regressão; exemplos; Resíduos MQ e regressão ajustada; Propriedades dos estimadores MQ dos coeficientes da regressão; Estimador não enviesado da variância da variável residual; Exemplo.

T15 - Modelo de regressão Linear

Coeficiente de determinação e sua interpretação. Hipótese adicional (H6) e inferência estatística sobre o modelo; Inferência sobre um parâmetro beta. Exemplos

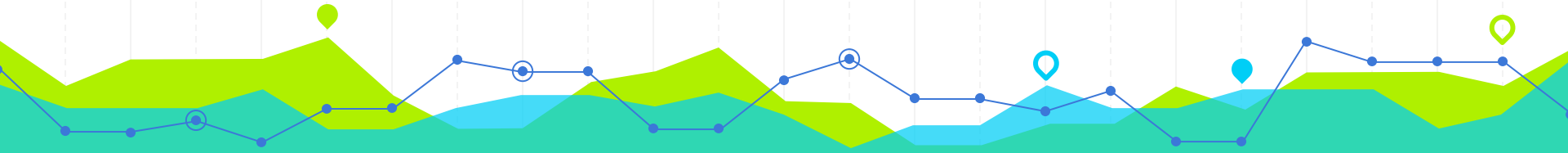
9ª semana (14/11 e 16/11)

T16 - Modelo de Regressão Linear

Mais exemplos de inferência sobre um parâmetro beta; Inferência sobre uma combinação linear de betas; exemplos.

T17 - Modelo de Regressão Linear

Teste de nulidade conjunta de vários coeficientes; exemplo; Teste F à significância global da regressão; Teste de um conjunto de restrições lineares; exemplo.



Modelo de Regressão Linear Simples

1

Coeficientes da reta de regressão

Revisão

Modelo de Regressão Linear

Modelos de Regressão

São modelos utilizados para compreender a relação entre

- uma variável resposta Y

  e

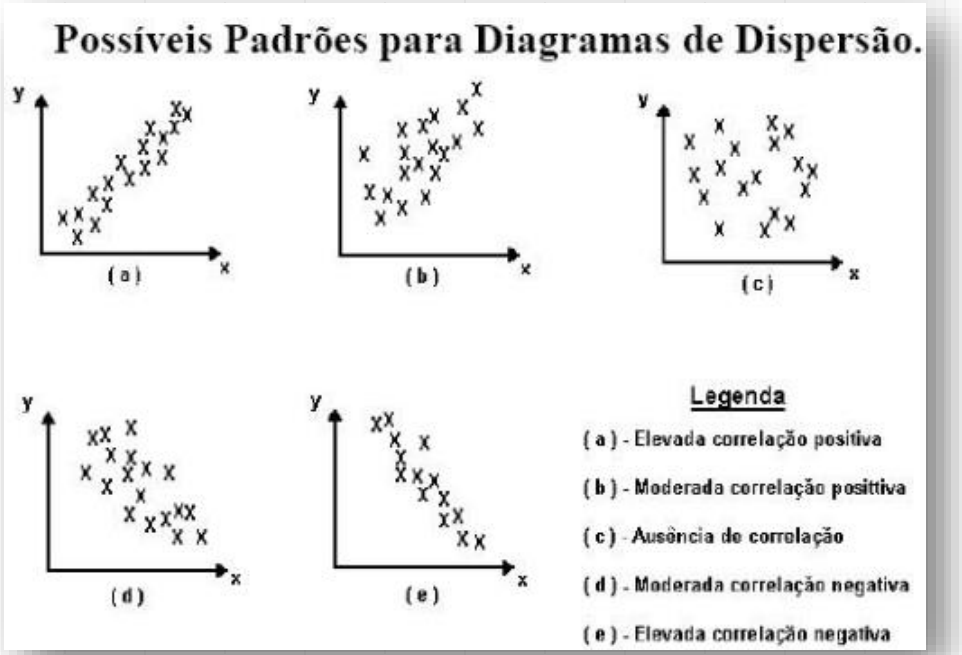
- uma ou mais variáveis explicativas (X ou $X_1, X_2, \dots$).

Um problema de análise da associação entre **duas variáveis** quantitativas começa pela recolha de uma amostra de pares de pontos:

$$\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\},$$

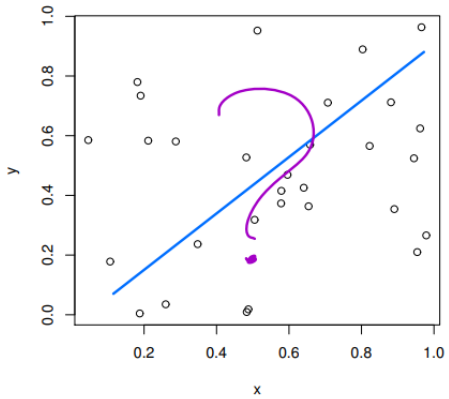
após o qual se deve proceder à representação gráfica desses pontos, usando os diagramas de dispersão.

Diagramas de Dispersão

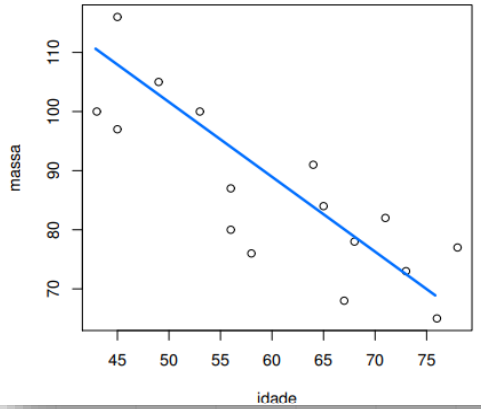


Diagramas de Dispersão

Quando duas variáveis não estão relacionadas, este gráfico representará uma mancha de pontos aleatória:



Quando existe uma relação entre duas variáveis, o diagrama de dispersão mostrará uma mancha de pontos não aleatória e poder-se-á verificar se a relação entre x e y poderá ser modelada por uma reta ou sugerir alguma outra relação.



Slides Professora Conceição Amado

Modelo de Regressão Linear Simples

Modelo de Regressão Linear Simples (RLS)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

onde

Y_i: variável aleatória (variável dependente ou resposta)

x_i: valor fixo (medição sem erro ou com erro desprezável) (variável independente ou de controlo)

β_0: parâmetro (ordenada na origem) –> a estimar

β_1: parâmetro regressor (declive, coeficiente angular) –> a estimar

ε_i: erro aleatório associado à i-ésima prova, verificando:

- $E[\varepsilon_i] = 0$ (valor esperado nulo);
- $Var[\varepsilon_i] = \sigma^2$ (variância constante); –> a estimar
- $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ (i \neq j)$ (não correlacionados);

Y é a v.a. dependente, resposta, explicada

X é a v.a. independente, de controlo, explicativa, regressora

Modelo de Regressão Linear Simples

Equação da Regressão Linear

Para se estimar o valor esperado, usa-se de uma equação, que determina a relação entre ambas as variáveis.

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

, onde:

y_i: Variável explicada (dependente); representa o que o modelo tentará prever

α: É uma constante, que representa a interceptação da reta com o eixo vertical;

β: Representa a inclinação (coeficiente angular) em relação à variável explicativa;

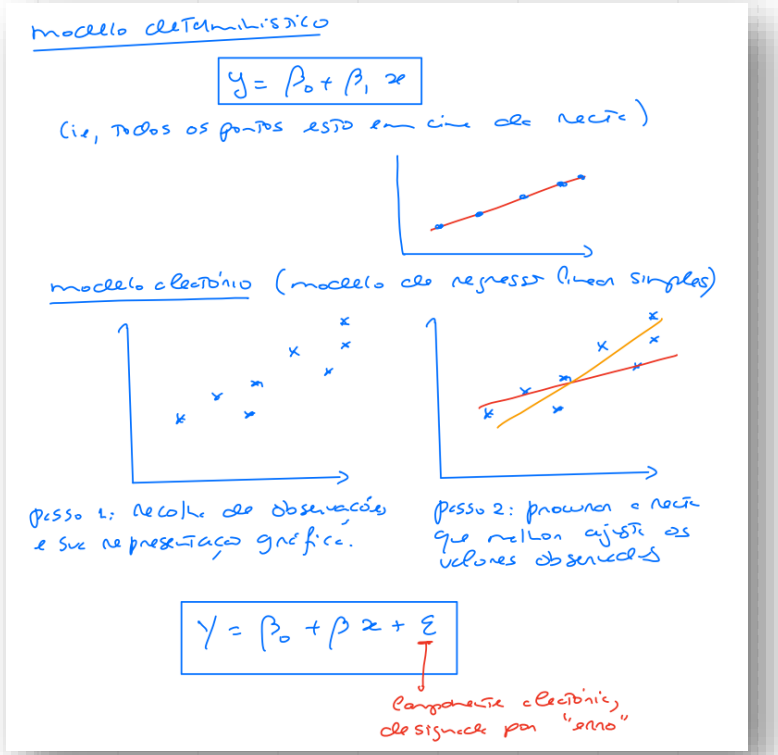
X_i: Variável explicativa (independente);

A v.a. X, denominada variável regressora, explicativa ou independente, é considerada uma variável controlada pelo investigador e medida com erro desprezível

ε_i: Representa todos os factores residuais mais os possíveis erros de medição. O seu comportamento é aleatório, devido à natureza dos factores que encerra. Para que essa fórmula possa ser aplicada, os erros devem satisfazer determinadas hipóteses, que são: terem distribuição normal, com a mesma variância σ^2 , independentes e independentes da variável explicativa X, ou seja, i.i.d. (independentes e identicamente distribuídas).

Regressão linear – Wikipédia, a enciclopédia livre (wikipedia.org)

Modelos de Regressão Linear Simples

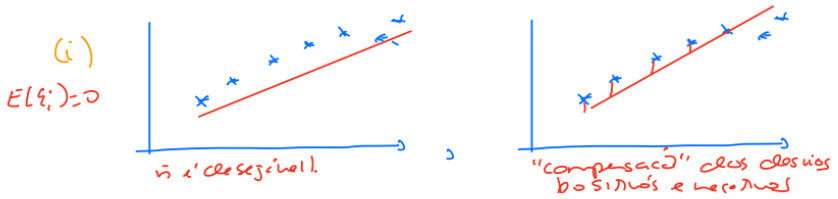


Slides Professora Conceição Amado

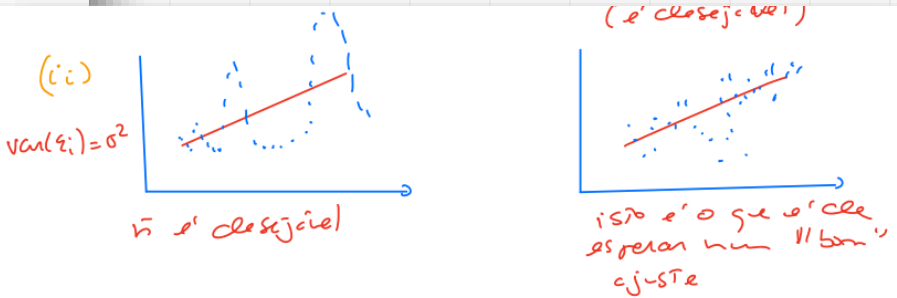
Modelo de Regressão Linear Simples

ε_i: erro aleatório associado à i-ésima prova, verificando:

- $E[\varepsilon_i] = 0$ (valor esperado nulo); (i)
- $Var[\varepsilon_i] = \sigma^2$ (variância constante); –> a estimar (ii)
- $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ (i \neq j)$ (não correlacionados); (iii)



Slides Professora Conceição Amado



(iii)
$cov(\varepsilon_i, \varepsilon_j) = 0$

veremos mais tarde, e não de forma gráfica.

Modelo de Regressão Linear Simples

Em consequência do modelo:

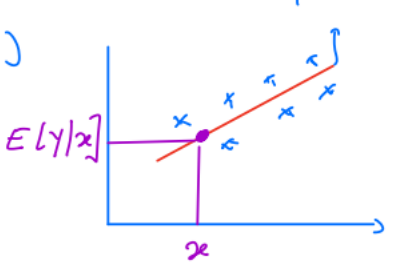
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

e das hipóteses sobre os ε_i's, decorre-se:

$$E[Y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i]$$

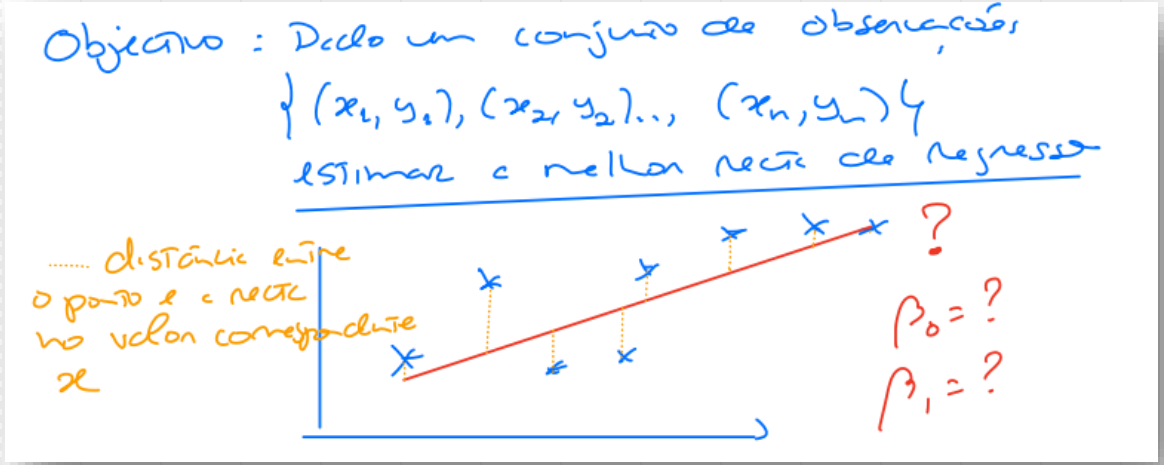
$$= \beta_0 + \beta_1 x_i + E(\varepsilon_i)$$

$$= \beta_0 + \beta_1 x_i$$



Slides Professora Conceição Amado

Modelos de Regressão Linear Simples



Slides Professora Conceição Amado

Modelos de Regressão Linear Simples

Observações:

1. Como $E[\varepsilon_i] = 0$ então

$$E[Y|x_i] = \beta_0 + \beta_1 x_i \quad \text{(recta de regressão)}$$

Ou seja, para cada valor x_i o ponto sobre a recta tem ordenada $E[Y|x_i] = \beta_0 + \beta_1 x_i$.

2. Interpretação dos parâmetros β_0 e β_1:
   - $\beta_0 = E(Y|x = 0) \quad \rightarrow \quad$ ordenada na origem;
   - $\beta_1 = E(Y|x = x^* + 1) - E(Y|x = x^*), \forall_{x^*} \quad \rightarrow \quad$ declive;

Modelos de Regressão Linear Simples (MRLS)

O modelo RLS diz-se:

1. simples, pelo facto de apenas existir uma variável explicativa, (x), na relação.
   (O modelo $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ é um exemplo de um modelo de regressão linear múltiplo, já que envolve duas variáveis explicativas x_1 e x_2.)
2. linear, referindo à linearidade nos parâmetros $(\beta_0, \beta_1 \dots)$.
   ou seja, modelos que em vez de x e Y apareçam funções destas variáveis também são considerados modelos de regressão linear, por exemplo, os modelos

$$Y = \beta_0 + \beta_1 x^2 + \varepsilon$$

ou

$$\log Y = \beta_0 + \beta_1 \log x + \varepsilon$$

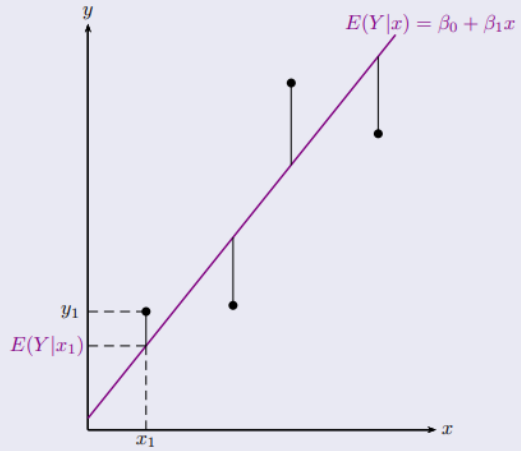
são modelos de regressão linear.

Slides Professora Conceição Amado

Método dos Mínimos Quadrados: MRLS

Estimação de β_0 e β_1

Como as observações estão afetadas de erros não é possível saber qual o valor exato dos coeficientes β_0 e β_1, mas é possível estimá-los...usando, por exemplo, o método dos mínimos quadrados.



Para cada valor de x_i o ponto sobre a recta tem ordenada $E(Y|x_i) = \beta_0 + \beta_1 x_i$.

Seja

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Pretende-se então determinar β_0 e β_1 que minimizam $Q(\beta_0, \beta_1)$.

Método dos Mínimos Quadrados: MRLS

Estimação de β_0 e β_1

Amostra, $\{(x_i, y_i), i = 1, \ldots, n\}$, pretende-se $\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min Q(\beta_0, \beta_1)$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \end{cases} \Leftrightarrow \begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

trata-se de um sistema de duas equações lineares a duas incógnitas que tem como solução as

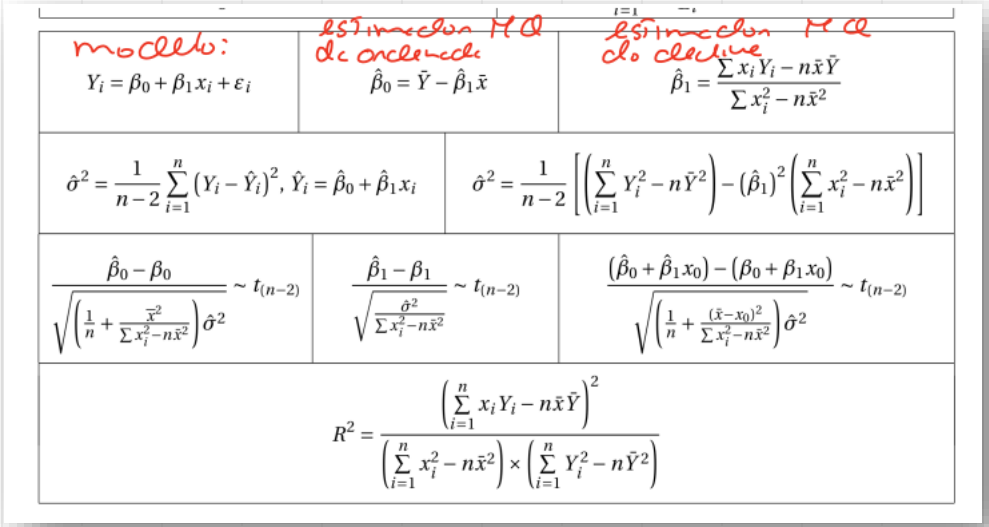
estimativas dos mínimos quadrados:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

com $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Slides Professora Conceição Amado

Método dos Mínimos Quadrados: MRLS



modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

estimador MQ do ordenado

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

estimador MQ do declive

$$\hat{\beta}_1 = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,\ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right]$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$$

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \times \left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}$$

Slides Professora Conceição Amado

Exemplo

Uma nutricionista está a investigar a relação entre o índice de colesterol total (x, em miligramas por decilitro) e o índice de massa corporal (Y, em quilogramas por altura ao quadrado). Numa amostra de 10 utentes de um centro de saúde, obtiveram-se os seguintes resultados:

$$\sum_{i=1}^{10} x_i = 2459, \quad \sum_{i=1}^{10} x_i^2 = 620155, \quad \sum_{i=1}^{10} y_i = 263.63, \quad \sum_{i=1}^{10} y_i^2 = 7091.388, \quad \sum_{i=1}^{10} x_i y_i = 65530.71,$$

onde $[\min_{i=1,\ldots,10}(x_i), \max_{i=1,\ldots,10}(x_i)] = [200, 310]$. Admita que x e Y estão relacionadas de acordo com o modelo de regressão linear simples, $Y = \beta_0 + \beta_1 x + \epsilon$. Obtenha a reta de mínimos quadrados com base nos dados fornecidos; além disso, calcule o coeficiente de determinação e interprete o seu valor.

Exemplo: Estimação dos Coeficientes da Reta de Regressão

Estimativas de MQ dos parâmetros desconhecidos β_0 e β_1

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{10} x_i y_i - 10 * \bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - 10 * \bar{x}^2} \\ &= \frac{65530.71 - 10 * 245.9 * 26.36}{620155 - 10 * 245.9^2} \\ &= \frac{704.093}{15486.9} \\ &= 0.0454.\end{aligned}$$

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 * \bar{x} \\ &= 26.36 - 0.0454 * 245.9 \\ &= 15.196.\end{aligned}$$

obter $\hat{\beta}_1$ e $\hat{\beta}_0$, ver formulário

ie:

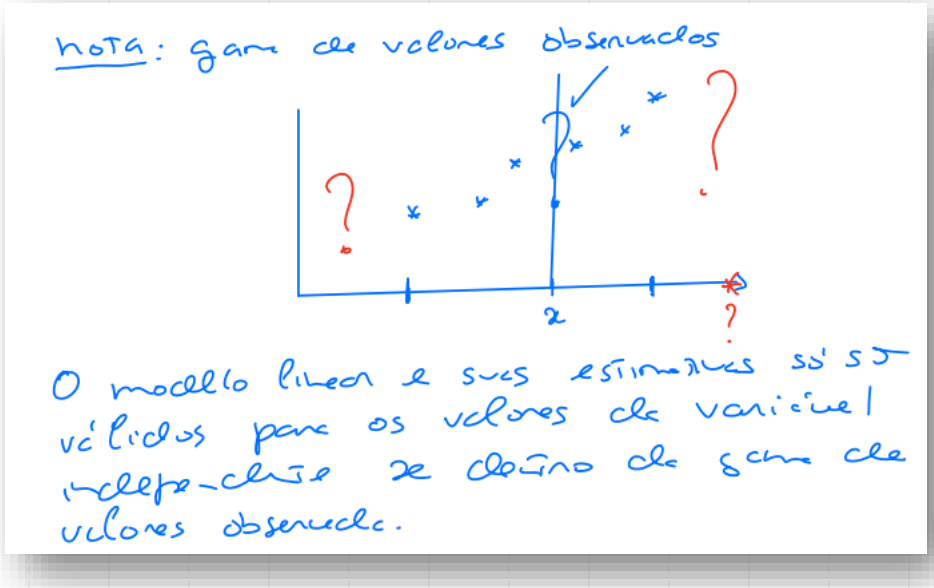
$Y_i = 15.196 + 0.0454 x_i + \varepsilon_i$ errado (muito!)

$E(Y_i) = 15.196 + 0.0454 x_i$ errado

$x_i \in [200, 310]$

$\hat{E}[Y_i] = 15.196 + 0.0454 x_i$ certo

Exemplo: Estimação dos Coeficientes da Reta de Regressão



Slides Professora Conceição Amado

Método dos Mínimos Quadrados: MRLS

Observações:

1. Pode mostrar-se, pelo determinante da matriz Hessiana, que se existir solução ela corresponde sempre a um mínimo.

2. Existe solução se e só se $\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$, ou seja, se na amostra existirem pelo menos dois valores distintos de x.

3. $$\hat{y} = \widehat{E[Y|x]} = \hat{\beta}_0 + \hat{\beta}_1 x$$

   é a estimativa de mínimos quadrados da reta de regressão. A estimação pontual de $E(Y|x)$ deve restringir-se ao domínio dos valores observados na amostra da variável explicativa x.

4. Às diferenças

   $$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

   chamam-se **resíduos**. A análise dos resíduos permite avaliar se o modelo assumido é adequado.

Slides Professora Conceição Amado

Método dos Mínimos Quadrados: MRLS

Os correspondentes estimadores dos mínimos quadrados de β_0 e β_1 são, respetivamente:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x};$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2};$$

onde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{e} \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

$$\hat{Y}_i = \widehat{E[Y|x_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

é o estimador de mínimos quadrados da reta de regressão.

Slides Professora Conceição Amado

Estimador da Variância

Recordar que no modelo apresentado aparece mais um parâmetro: $Var[\varepsilon_i] = \sigma^2$. Essa estimação faz-se usando os resíduos ($e_i = y_i - \hat{y}_i$).

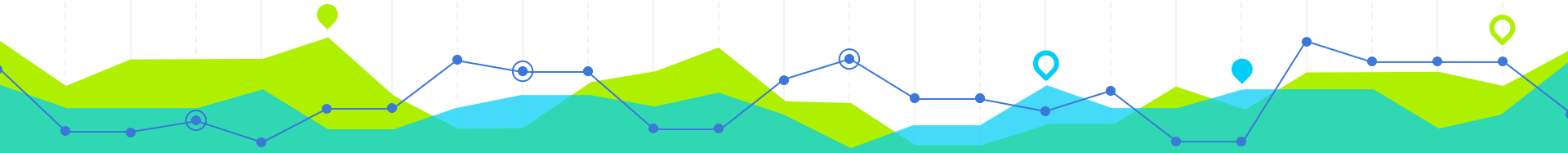
Estimação de $\sigma^2 = Var[\varepsilon_i]$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \quad (2)$$

$$= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right] \quad (3)$$

onde $\hat{Y}_i = \widehat{E[Y|x_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Nota:
- A $\hat{\sigma}$ chama-se erro padrão dos resíduos;
- $\hat{\sigma}^2$ é também um estimador da variância de Y_i.



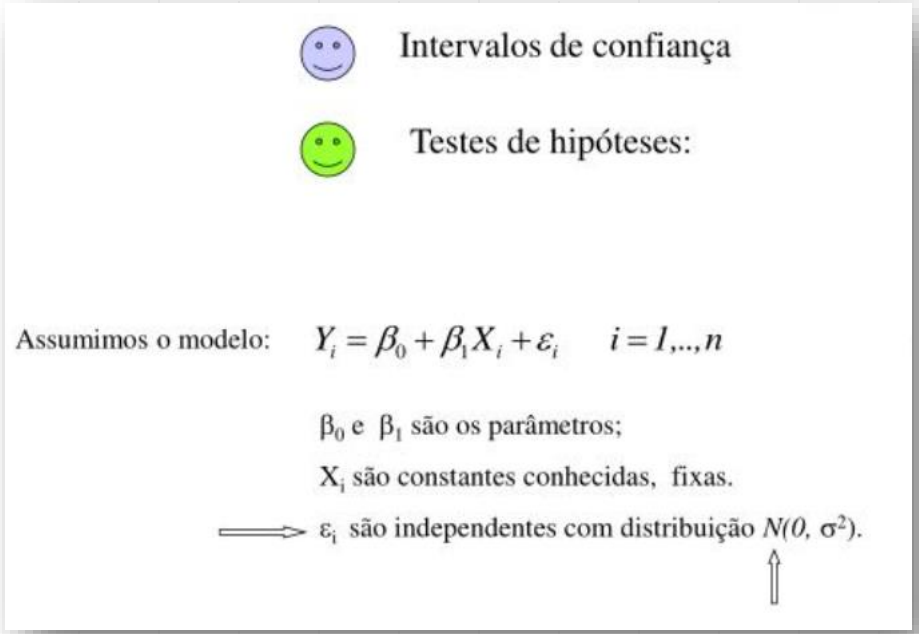
Inferência na Regressão Linear Simples

Intervalos de Confiança e Testes de Hipóteses

2

Revisão

Inferência na Análise de Regressão



PPT - Regressão Linear Simples PowerPoint Presentation, free download - ID:3374089 (slideserve.com)

Inferência sobre os Coeficientes da Reta RLS

Para se fazerem inferências em RLS (i.e, realizar testes de hipóteses e calcular intervalos de confiança) é necessário admitir que:

os erros tenham distribuição normal

As suposições do modelo de RLS são então:

- $E[\varepsilon_i] = 0$,
- $Var[\varepsilon_i] = \sigma^2$ e
- e $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ \ (i \neq j)$

e com a nova hipótese de trabalho tem-se agora que

$$\varepsilon_i \underset{i.i.d.}{\sim} N(0, \sigma^2)$$

(i), (ii), (iii) + normalidade

.

Slides Professora Conceição Amado

Inferência sobre os Coeficientes da Reta RLS

Resultados importantes

Mostra-se que:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)},$$

e

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}.$$

Estes resultados permitem realizar testes de hipóteses e calcular intervalos de confiança relativos a β_0 e β_1, respectivamente.

Inferência sobre os Coeficientes da Reta RLS

Exemplos: Intervalos de confiança e Testes de hipóteses para:

β_0, β_1, $E[Y|x]$

v. fulcrais:

ordenada na origem

declive

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)} \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)} \qquad \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

valor esperado de Y dado um valor particular x_0

I.C. p/ β_0

T.H. p/ β_0

I.C. p/ β_1

T.H. p/ β_1

(significância estatística)

I.C. p/ $E[Y|x_0]$

T.H. p/ $E[Y|x_0]$

Slides Professora Conceição Amado

Pergunta 10 2 valores

O interesse crescente na utilização da Internet para fins comerciais tem levado muitas companhias a vender os seus produtos através deste meio. Um estatístico levou a cabo um estudo para determinar até que ponto o grau de escolaridade e o uso da Internet estão ligados entre si. Para o efeito considerou uma amostra selecionada aleatoriamente de 20 adultos, para os quais registou o número de anos de escolaridade (x, com valores observados entre 8 e 14 anos) e o número de horas despendidas na Internet, Y, na semana anterior ao decorrer do questionário. Obtiveram-se os seguintes resultados:

$$\sum_{i=1}^{20} x_i = 228,\ \sum_{i=1}^{20} x_i^2 = 2696,\ \sum_{i=1}^{20} y_i = 157,\ \sum_{i=1}^{20} y_i^2 = 1671,\ \sum_{i=1}^{20} x_i y_i = 1852.$$

Admita a validade do modelo de regressão linear simples, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1,\ldots,20$. Para indivíduos com 14 anos de escolaridade, obtenha um intervalo de confiança a 90% para o número esperado de horas semanais despendidas na Internet.

Pede-se: I.C. $_{0.90}(E[Y|x=14])$

em geral, I.C: (exp-7)

1. escolher a v. fulcral ✓
2. determinar os quantis ✓
3. inverter as desigualdades
4. concretizar, p/ os valores amostrais

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

Exemplo: Intervalos de Confiança dos Coeficientes da Reta de Regressão

- **Hipóteses de trabalho**

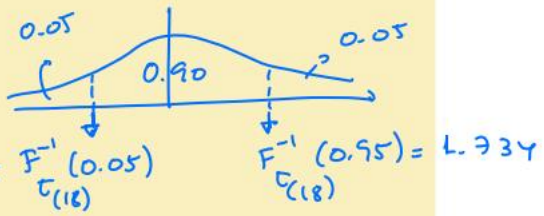
  No modelo de RLS, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, consideraremos $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$, $i = 1, \ldots, n$.

  Pretende-se um intervalo de confiança a 90% para $E[Y|x = 14]$.
  Observar que $x = 14 \in [\min(x_i), \max(x_i)] = [8, 14]$

- **Variável aleatória fulcral**

$$T = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(18)},$$

  com $x_0 = 14$.

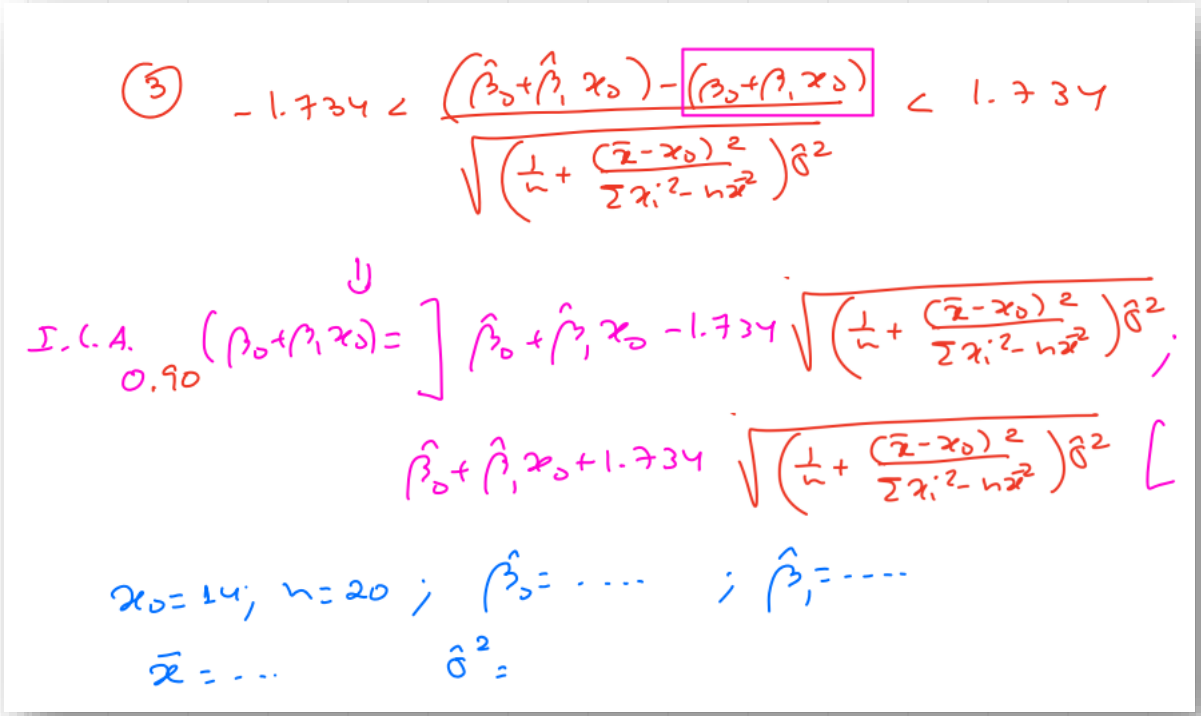


- **Intervalo aleatório** Como $1 - \alpha = 0.9$ então $\alpha = 0.10$ e $a = F^{-1}_{t(18)}\left(1 - \frac{\alpha}{2}\right) = F^{-1}_{t(18)}(1 - 0.05) = F^{-1}_{t(18)}(0.95) = 1.734$

  Como a distribuição da t-Student é simétrica vem:

Slides Professora Conceição Amado

Exemplo: Intervalos de Confiança dos Coeficientes da Reta de Regressão



Slides Professora Conceição Amado

Exemplo: Intervalos de Confiança dos Coeficientes da Reta de Regressão

- **Concretização:** Precisamos de calcular as estimativas de β_0 e β_1, que serão dadas por:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{1852 - 20 \times 228/20 \times 157/20}{2696 - 20 \times (228/20)^2} \\
&\simeq 0.6426; \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \times \bar{x} \\
&\simeq 157/20 - 0.6425 \times 228/20 \\
&= 0.5244,
\end{aligned}$$

e também a estimativa de σ^2

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - (\hat{\beta}_1)^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right] \\
&= \frac{1}{20-2}\left[438.55 - 0.6425^2 \times 96.8\right] \\
&= 22.1432.
\end{aligned}$$

O intervalo de confiança, a 90% de confiança é dado por:

$$\begin{aligned}
IC_{90\%}(\beta_0 + \beta_1 \times 14) &= \left((0.5244 + 0.6426 \times 14) \pm 1.73\sqrt{\left(\frac{1}{20} + \frac{(228/20 - 14)^2}{96.8}\right)22.1432}\right) \\
&= (6.6961, 12.3454).
\end{aligned}$$

Slides Professora Conceição Amado

Pergunta 10 2 valores

O interesse crescente na utilização da Internet para fins comerciais tem levado muitas companhias a vender os seus produtos através deste meio. Um estatístico levou a cabo um estudo para determinar até que ponto o grau de escolaridade e o uso da Internet estão ligados entre si. Para o efeito considerou uma amostra selecionada aleatoriamente de 20 adultos, para os quais registou o número de anos de escolaridade (x, com valores observados entre 8 e 14 anos) e o número de horas despendidas na Internet, Y, na semana anterior ao decorrer do questionário. Obtiveram-se os seguintes resultados:

x = nº de anos de escolaridade (8 – 14)

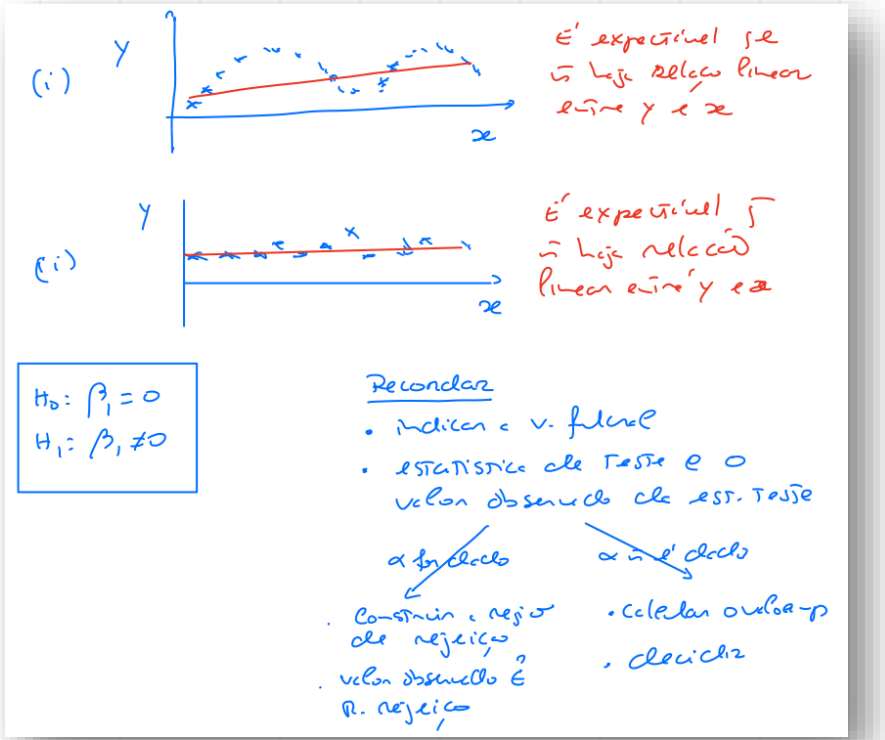
$$\sum_{i=1}^{20} x_i = 228, \quad \sum_{i=1}^{20} x_i^2 = 2696, \quad \sum_{i=1}^{20} y_i = 157, \quad \sum_{i=1}^{20} y_i^2 = 1671, \quad \sum_{i=1}^{20} x_i y_i = 1852.$$

Admita a validade do modelo de regressão linear simples, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, 20$. Para indivíduos com 14 anos de escolaridade, obtenha um intervalo de confiança a 90% para o número esperado de horas semanais despendidas na Internet.

- $I.C._{0.90}(Y \mid x_0 = 14)$ ✓

- Testar a significância do modelo de regressão (ou se há relação linear entre Y e x, estatisticamente significativa)

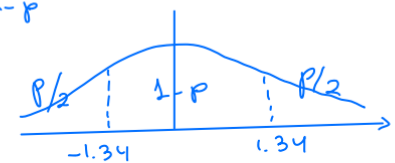
Exemplo: Testes de Hipóteses dos Coeficientes da Reta de Regressão



Slides Professora Conceição Amado

Exemplo: Testes de Hipóteses dos Coeficientes da Reta de Regressão

- variável fulcral: $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$
- EST. TESTE $T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \overset{H_0}{\sim} t_{(n-2)}$
- VALOR OBS: $t_0 = \frac{0.6426}{\sqrt{\frac{22.1432}{2696 - 20 \times (\frac{228}{20})^2}}} = 1.34$
- valor-p



$1 - p_2 = F_{t_{(18)}}(1.34) \Leftrightarrow p = 0.197$ (0.901541)

- decisão: Rejeitar H_0 p/ $\forall \alpha \geq 19.7\%$, ie, H_0 não deve ser rejeitada p/ os valores usuais de significância

Exemplo: Estimativas da Reta de Regressão

- Estimativa do nº médio de horas dispendidas na net para 2 pessoas cuja escolaridade difere de 3 anos?

Se x e $x+3 \in [8,14]$: ok, podemos resolver!

$$E[y|x] = \beta_0 + \beta_1 x$$

$$E[y|x+3] = \beta_0 + \beta_1 (x+3) =$$

$$= \beta_0 + \beta_1 x + \boxed{3\beta_1}$$

Resposta: diferença estimada é $3\hat{\beta}_1$

Slides Professora Conceição Amado

Teste de Hipóteses para a Ordenada na Origem

Teste de hipóteses relativo a β_0

1. **Definir H_0 e H_1:** $H_0 : \beta_0 = \beta_{0,0}$ *vs* $H_1 : \beta_0 >, <, \neq \beta_{0,0}$
2. **Estatística de teste:**

$$T_{H_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \underset{Sob\ H_0}{\sim} t_{(n-2)}$$

3. **Região Crítica (ao nível de significância α):**
   1. $H_1 : \beta_0 \neq \beta_{0,0}$, $RC_\alpha : |T_{H_0}| > c$, com $c = F^{-1}_{t_{(n-2)}}(1 - \alpha/2)$
   2. $H_1 : \beta_0 > \beta_{0,0}$, $RC_\alpha : T_{H_0} > c$, com $c = F^{-1}_{t_{(n-2)}}(1 - \alpha)$
   3. $H_1 : \beta_0 < \beta_{0,0}$, $RC_\alpha : T_{H_0} < c$, com $c = F^{-1}_{t_{(n-2)}}(\alpha)$

   Ou cálculo do valor-p.

Decisão usual.

Intervalo de Confiança para a Ordenada na Origem

Intervalo de confiança a $(1-\alpha)\times 100\%$ para β_0

Usando a seguinte variável aleatória fulcral:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

obtém-se, após dedução, o seguinte intervalo para β_0 a $(1-\alpha)\times 100\%$ de confiança

$$I.C._{(1-\alpha)\times 100\%}(\beta_0) = \hat{\beta}_0 \pm F^{-1}_{t_{n-2}}(1-\alpha/2)\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}$$

Teste de Hipóteses para o Declive

Teste de hipóteses relativo a β_1

1. **Definir H_0 e H_1:** $H_0 : \beta_1 = \beta_{1,0}$ *vs* $H_1 : \beta_1 >, <, \neq \beta_{1,0}$
2. **Estatística de teste:**

$$T_{H_0} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \underset{Sob\ H_0}{\sim} t_{(n-2)}.$$

3. Região crítica (ao nível de significância α):
   1. $H_1 : \beta_1 \neq \beta_{1,0}$, $RC_\alpha : |T_{H_0}| > c$, com $c = F^{-1}_{t_{(n-2)}}(1 - \alpha/2)$
   2. $H_1 : \beta_1 > \beta_{1,0}$, $RC_\alpha : T_{H_0} > c$, com $c = F^{-1}_{t_{(n-2)}}(1 - \alpha)$
   3. $H_1 : \beta_1 < \beta_{1,0}$, $RC_\alpha : T_{H_0} < c$, com $c = F^{-1}_{t_{(n-2)}}(\alpha)$.

   Ou cálculo do valor-p.

Decisão usual.

Teste de Hipóteses para o Declive

Observações:

1. Um teste importante em RLS é

$$H_0 : \beta_1 = 0 \ \ vs \ \ H_1 : \beta_1 \neq 0$$

pois não rejeitar H_0 significa que há evidência para a não existência de uma associação linear entre x e y, ou seja não há associação ou a associação não é linear. A este teste costuma designar-se por:

teste à significância da regressão

.

2. Como já foi referido no Capítulo 8 também se pode realizar um teste de hipóteses usando a relação com intervalos de confiança (desde que a v.a. fulcral e a estatística de teste sejam da mesma forma!).

Intervalo de Confiança para o Declive

Intervalo de confiança a $(1-\alpha)\times 100\%$ para β_1

Usando a seguinte variável aleatória fulcral:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$$

obtém-se, após dedução, o seguinte intervalo para β_1 a $(1-\alpha)\times 100\%$ de confiança:

$$I.C._{(1-\alpha)\times 100\%}(\beta_1) = \hat{\beta}_1 \pm F^{-1}_{t_{n-2}}(1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

Slides Professora Conceição Amado

Teste de Hipóteses para o Valor Médio de Y dado um dado x

Em algumas situações é necessário efetuar inferências sobre o valor médio de Y para um dado valor fixo de x, digamos, x_0.

Inferência para $E[Y|x_0] = \beta_0 + \beta_1 x_0$

Um estimador pontual de $E[Y|x_0] = \beta_0 + \beta_1 x_0$ é:

$$\widehat{E[Y|x_0]} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

E pode mostrar-se que:

$$\frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}.$$

Testes de hipóteses e **intervalos de confiança** para $E[Y|x_0]$ são então baseados nesta variável, e o procedimento é o habitual.

Slides Professora Conceição Amado

Qualidade do Modelo de Regressão Linear Simples

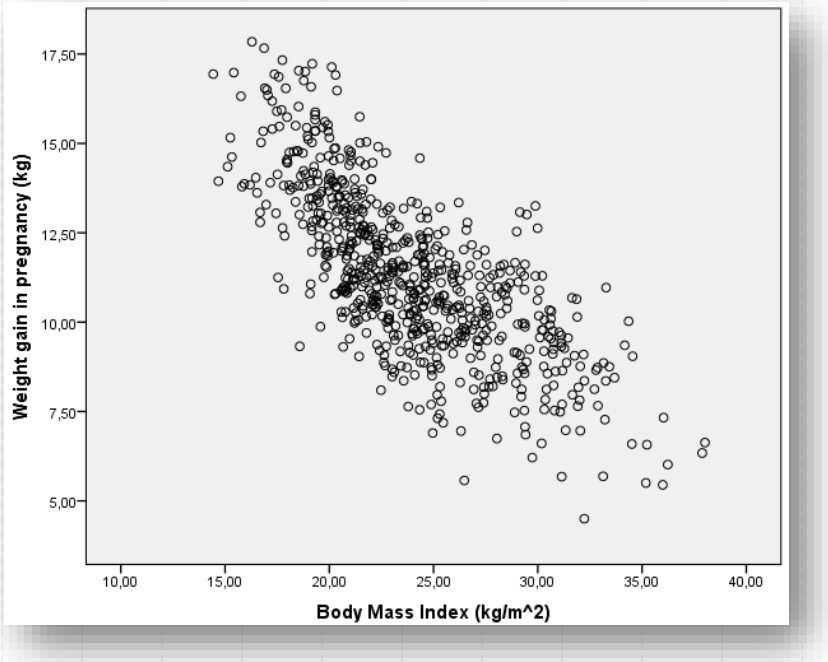
3

Coeficiente de determinação

Revisão

Diagrama de Dispersão

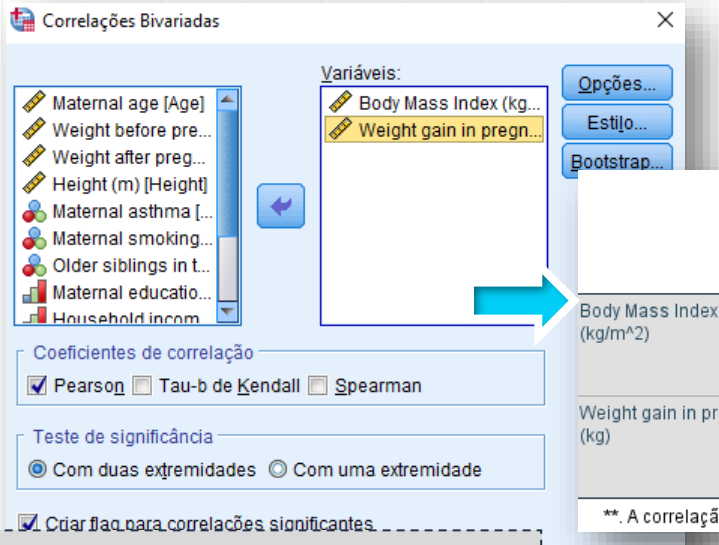
BMI vs. WGain



Coeficiente de Correlação de Pearson e Coeficiente de Determinação

BMI vs. WGain

SPSS: Analyze ⇨ Correlate ⇨ Bivariate [✓ Pearson …]



		Body Mass Index (kg/m^2)	Weight gain in pregnancy (kg)
Body Mass Index (kg/m^2)	Correlação de Pearson	1	-,738**
	Sig. (bilateral)		,000
	N	699	699
Weight gain in pregnancy (kg)	Correlação de Pearson	-,738**	1
	Sig. (bilateral)	,000	
	N	699	699

**. A correlação é significativa no nível 0,01 (bilateral).

Coeficiente de Correlação de Pearson:
r = -0,738 ⇒ Correlação linear negativa moderada, pois -0,8 < r ≤ -0,5.

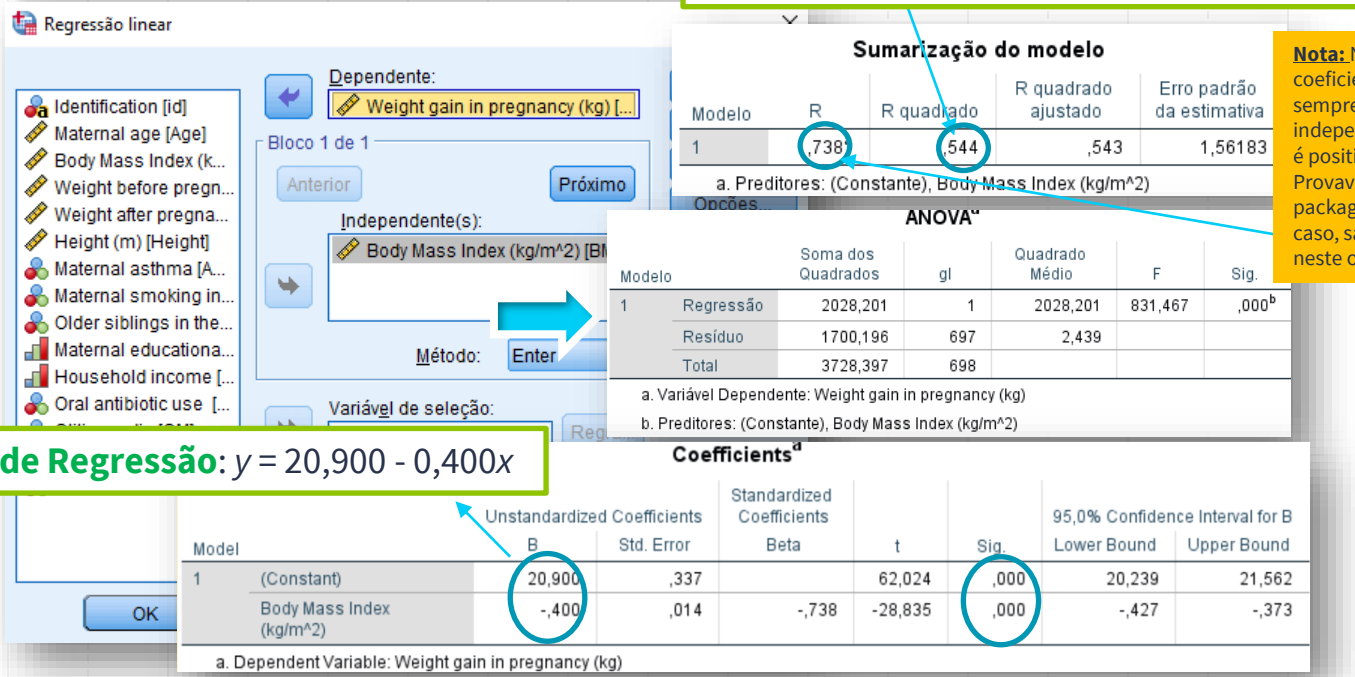
54,4% da variabilidade no WGain (variável dependente) pode ser explicada pelo BMI (variável independente) no modelo.

Coeficiente de Determinação: $r^2 = 0,544$

Regressão Linear

BMI vs. WGain

SPSS: Analyze → Regression → Linear



Coeficiente de Determinação: $r^2 = 0{,}544$

Nota: Neste output, o valor do coeficiente de correlação surge sempre positivo independentemente se de fato é positivo ou negativo. Provavelmente é um lapso do package. Por exemplo, neste caso, sabe-se que r=-0,738, mas neste output surge r=0,738.

Reta de Regressão: $y = 20{,}900 - 0{,}400x$

Regressão Linear

BMI vs. WGain

X - BMI
- v.a. independente, explicativa
- Quantitativa contínua (escala métrica)

Y - WGain
- v.a. dependente
- Quantitativa contínua (escala métrica)

Considere-se a reta de regressão: $y = \alpha + \beta x$, onde α = ordenada na origem e β = declive

Hipóteses

a) H_0: $\alpha = 0$ versus H_1: $\alpha \neq 0$

b) H_0: $\beta = 0$ versus H_1: $\beta \neq 0$

Testes de Hipóteses para os Coeficientes da Reta de Regressão

BMI vs. WGain

Valores observados das estatísticas de teste (VOE)

a) $t_0 = 62{,}024$ e valor-p $< 0{,}0005 < 0{,}05 \Rightarrow$ Rejeita-se H_0 para $\alpha = 0{,}05$
b) $t_0 = -28{,}835$ e valor-p $< 0{,}0005 < 0{,}05 \Rightarrow$ Rejeita-se H_0 para $\alpha = 0{,}05$

Conclusão

- Faz sentido ajustar um modelo de regressão linear simples.
- O valor de beta ou coeficiente estandardizado/padronizado corresponde ao coeficiente de correlação linear de Pearson: r = -0,738 ~ -0,74.
- Existe evidência estatística para afirmar que existe uma associação (correlação) linear negativa moderada entre X e Y (ou seja, parece existir uma relação de dependência), pois $-0{,}8 < r \leq -0{,}5$.
- O coeficiente de determinação é igual a $r^2 = 0{,}544$. Logo a qualidade do ajustamento do modelo aos dados é moderada.

Nota: O sinal de beta (declive da reta) indica se a relação linear é positiva ou negativa (direta ou inversa).

Coeficientes[a]

Modelo		Coeficientes não padronizados B	Coeficientes não padronizados Erro Padrão	Coeficientes padronizados Beta	t	Sig.
1	(Constante)	20,900	,337		62,024	,000
	Body Mass Index (kg/m^2)	-,400	,014	-,738	-28,835	,000

a. Variável Dependente: Weight gain in pregnancy (kg)

Equação da Reta de Regressão e Previsões

BMI vs. WGain

Estimativas dos parâmetros α e β

a) a = 20,900

b) b = -0,400 (declive negativo)

Reta de regressão ajustada

y = 20,900-0,400x

Previsões

Se x = 30 então y* = 8,9kg

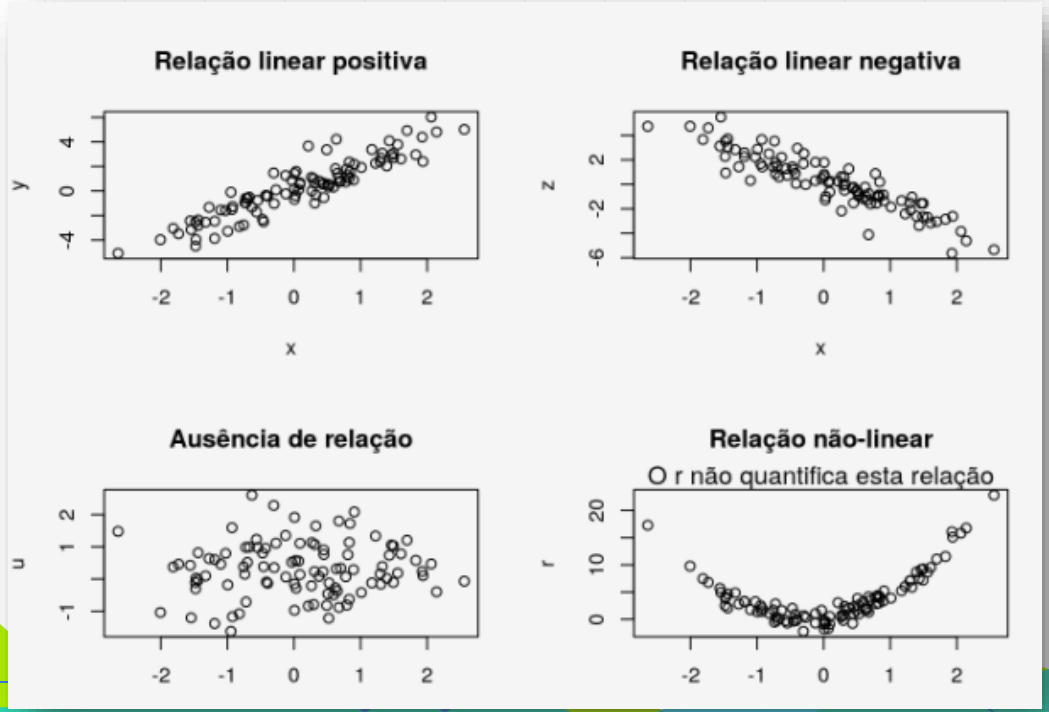
Estatísticas Descritivas

	N	Mínimo	Máximo	Média	Desvio Padrão
Body Mass Index (kg/m^2)	699	14,43	38,02	23,9178	4,26199
N válido (listwise)	699				

Nota: Não é possível obter-se uma previsão do ganho de peso durante a gravidez para as mulheres que tenham um BMI inferior a 14,43 e superior a 38,02, pois esses valores não pertence ao intervalo de valores de x (14,43; 38,02) utilizados na estimação dos coeficientes da reta de regressão.

Associação/Correlação Linear?

https://lec.pro.br//2-nao-categorizado/150-cor-pearson

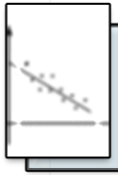


Coeficiente de Correlação de Pearson

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

r = -1

- Correlação linear negativa perfeita



-1 < r ≤ -0,8

- Correlação linear negativa forte

-0,8 < r ≤ -0,5

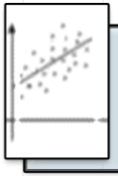
- Correlação linear negativa moderada

-0,5 < r < 0

- Correlação linear negativa fraca

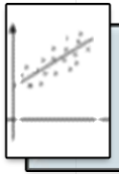
r = 0

- Não existe correlação linear



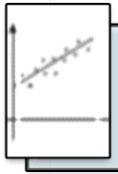
0 < r < 0,5

- Correlação linear positiva fraca



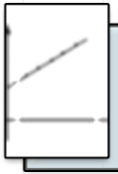
0,5 ≤ r < 0,8

- Correlação linear positiva moderada



0,8 ≤ r < 1

- Correlação linear positiva forte



r = 1

- Correlação linear positiva perfeita

Coeficiente de Determinação

O **coeficiente de determinação** ou simplesmente R^2. É uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra.

https://pt.slideshare.net/rodrigomuribec

Diagrama de dispersão

+

Coeficiente de Correlação de Pearson (r = 0,889) e Coeficiente de Determinação

($r^2 = 0,790$)

→ Existe uma relação linear positiva forte entre as duas variáveis.

79,0% da variabilidade do peso (variável dependente) pode ser explicada pela altura (variável independente) no modelo.

Faixa de variação de r: $0 \leq r^2 \leq 1$

Interpretação:

a) **Próximo de 1**: Forte poder de explicação do modelo (reta de regressão);

b) **Próximo de 0**: Fraco poder de explicação do modelo (reta de regressão);

c) **Exempo:** $r^2 = 0,92$. Então, 92% da variação de y pode ser explicada pera relação entre x e y.

https://slideplayer.com.br/slide/48204

Coeficiente de Determinação

Decomposição das somas dos quadrados

O coeficiente de determinação é uma medida descritiva indicadora da qualidade do ajustamento da recta estimada. Mostra-se que:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}_{SST}=\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{SSE}+\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}_{SSR}$$

onde

- SST corresponde à soma de quadrados total
- SSE corresponde à soma dos quadrados dos resíduos
- SSR é a soma dos quadrados da regressão

Coeficiente de Determinação

Coeficiente de determinação: R^2

O coeficiente de determinação é, então, dado pela expressão:

$$R^2 = \frac{SSR}{SST} = \frac{\text{variação explicada pelo modelo}}{\text{variação total}} =$$

$$= 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} =$$

$$= \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \right)}.$$

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \times \left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \right)} \quad \in [0, 1]$$

Slides Professora Conceição Amado

Coeficiente de Determinação

Interpretação: $R^2 \times 100\%$ representa a % de variabilidade do Y explicada pelo modelo de regressão linear simples.

R^2 "pequeno" → modelo é pouco adequado
[provavelmente ao testar se $\beta_1 = 0$, vamos aceitar)

R^2 "grande" → modelo é adequado
[provavelmente ao testar se $\beta_1 = 0$, vamos rejeitar]

Slides Professora Conceição Amado

Coeficiente de Determinação

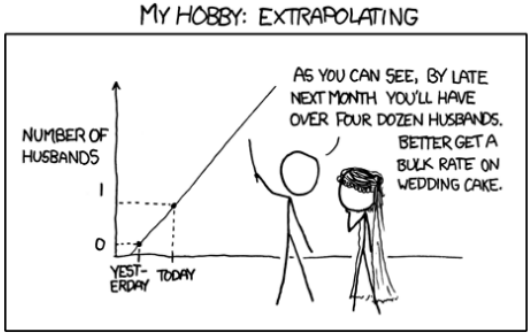
O coeficiente de determinação e o ajuste do modelo

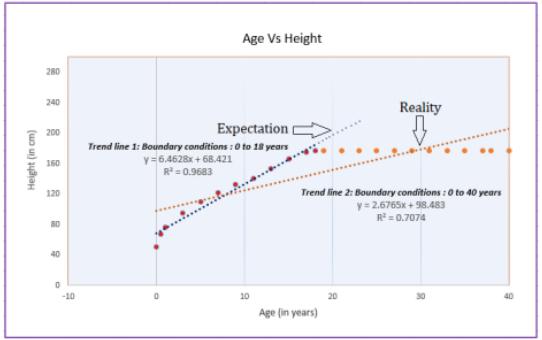
- $R^2 \times 100\%$ representa a percentagem de variabilidade total de Y que é explicada pelo modelo de regressão.
- $\hat{y}_i = y_i, \forall_i, \Leftrightarrow R^2 = 1$. Todos os pontos estão sobre a recta, o modelo de regressão explica toda a variabilidade observada em y. **Modelo óptimo!**
- $\hat{\beta}_1 = 0 \Leftrightarrow \hat{\beta}_0 = \bar{y} \Leftrightarrow \hat{y}_i = \bar{y}, \forall_i, \Leftrightarrow R^2 = 0$. A recta é horizontal, o modelo de regressão não explica nada da variabilidade observada em y. **Modelo péssimo!**
- Em geral, $0 < R^2 < 1$.
- Observar que a expressão do coeficiente de **determinação** em RLS corresponde ao **quadrado** do coeficiente de **correlação** linear (r_{xy}) que vimos no Cap.1.

Coeficiente de Determinação

Extrapolação: pode ser uma má ideia!

Extrapolação – uso do modelo de regressão para valores fora do domínio da variável explicativa.





Slides Professora Conceição Amado

Para explorar a relação entre a massa muscular e a idade (no género feminino) um nutricionista seleccionou aleatoriamente 16 mulheres com idades compreendidas entre os 40 e os 79 anos. Os resultados observados encontram-se na tabela seguinte (x representa a idade e y é um índice de massa muscular):

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_i	71	64	43	67	56	73	68	56	76	65	45	58	45	53	49	78
y_i	82	91	100	68	87	73	78	80	65	84	116	76	97	100	105	77

$$\sum_{i=1}^{16} x_i = 967,\ \sum_{i=1}^{16} y_i = 1379,\ \sum_{i=1}^{16} x_i^2 = 60409,\ \sum_{i=1}^{16} y_i^2 = 121887,$$

$$\sum_{i=1}^{16} x_i y_i = 81331.$$

Admita que o modelo de regressão linear simples ($Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) é adequado.

a) Calcule:

   i) uma estimativa pontual da diferença entre as massas musculares médias de mulheres cujas idades diferem de um ano;
   ii) uma estimativa pontual da massa muscular média para as mulheres de 60 anos;
   iii) o valor do resíduo para a 8^aobservação;
   iv) uma estimativa pontual de $Var(\varepsilon_i) = \sigma^2$;
   v) o coeficiente de determinação e interprete o valor obtido.

b) O nutricionista pensa que (na gama de idades considerada) a massa muscular é significativamente influenciada pela idade. Acha que as observações confirmam esta hipótese? Use um nível de significância de 5% e indique as hipóteses de trabalho de que necessita para efectuar o teste.

c) Calcule o intervalo de confiança a 99% para o valor esperado da massa muscular para uma mulher de 45 anos. Acha legítimo usar o mesmo procedimento tratando-se de uma mulher com 20 anos em vez de 45?

Modelo de RLS

Resolução:
Modelo de RLS:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

onde
Y_i: variável aleatória que representa o índice da massa muscular associada à i-ésima mulher;
x_i: variável explicativa que representa a idade da i-ésima mulher;
β_0, β_1: parâmetros;
ε_i: erro aleatório associado à i-ésima mulher, verificando:

(i) $E[\varepsilon_i] = 0$;

(ii) $Var[\varepsilon_i] = \sigma^2$;

(iii) $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ \ (i \neq j)$;

$i = 1, 2, \ldots, 16$.

Exercício a) i)

Cálculos auxiliares:

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 81331 - 16 \times \frac{967}{16} \times \frac{1379}{16} = -2012.3125$$

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 60409 - (967^2)/16 = 1965.9375$$

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 121887 - (1379^2)/16 = 3034.4375$$

a) Calcule:

i) uma estimativa pontual da diferença entre as massas musculares médias de mulheres cujas idades diferem de um ano;

Como
$E[Y|(x_j+1)] - E[Y|(x_j)] = \beta_0 + \beta_1(x_j+1) - \beta_0 - \beta_1 x_j = \beta_1$,
então uma estimativa dessa diferença será

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{-2012.3125}{1965.9375} \approx -1.0236$$

Exercícios a) ii) e iii)

ii) uma estimativa pontual da massa muscular média para as mulheres de 60 anos;

A estimativa pretendida é
$\hat{E}[Y|x = 60] = \hat{\beta}_0 + \hat{\beta}_1 \times 60 = 148.05 - 1.0236 \times 60 \approx 86.63$, pois
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1379/16 + 1.0236 \times (967/16) \approx 148.05.$

iii) o valor do resíduo para a 8ª observação;

Representando $\hat{y}_i = \hat{E}[Y|x_i]$ sabe-se que o oitavo resíduo é dado por $e_8 = y_8 - \hat{y}_8 = 80 - 90.728 \approx -10.7284$, já que
$\hat{y}_8 = 148.05 - 1.0236 \times 56 = 90.728.$

Exercícios a) iv) e v)

iv) uma estimativa pontual de $Var(\varepsilon_i) = \sigma^2$;

Uma estimativa pontual de $Var(\varepsilon_i) = \sigma^2$, $\forall i,\ i = 1, 2, \ldots, 16$ é

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right] =$$

$$= \frac{1}{14}(3034.4375) - (-1.0236)^2(1965.9375) \approx 69.62$$

v) o coeficiente de determinação e interprete o valor obtido.

O valor para o coeficiente de determinação é

$$r^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = \frac{(-2012.3125)^2}{(1965.9375)(3034.4375)} \approx 0.679.$$

Este valor indica que cerca de 67% da variabilidade deste índice de massa muscular é explicada pela idade (da mulher).

Exercício b)

b) O nutricionista pensa que (na gama de idades considerada) a massa muscular é significativamente influenciada pela idade. Acha que as observações confirmam esta hipótese? Use um nível de significância de 5% e indique as hipóteses de trabalho de que necessita para efetuar o teste.

O que o nutricionista pretende testar pode ser traduzido no seguinte teste de hipóteses:

Hipóteses: $H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$

Para se efetuar este teste de hipóteses é necessário considerar a seguinte hipótese de trabalho:

$$\varepsilon_i \underset{i.i.d.}{\sim} N\left(0, \sigma^2\right), \quad \forall_i, \quad i = 1, 2, \ldots, 16.$$

Estatística de teste: $T_{H_0} = \dfrac{\hat{\beta}_1}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \underset{Sob\, H_0}{\sim} t_{(14)}.$

Exercício b)

Valor observado da estatística de teste:

$$t_{obs} = \frac{-1.0236}{\sqrt{\frac{69.62}{1965.9375}}} \approx -5.44$$

Região Crítica para T_{H_0} ao nível de significância $\alpha = 0.05$:

$RC_{0.05} = (-\infty, -c) \cup (c, +\infty)$, onde $c : P(T_{H_0} \in RC | H_0) = 0.05$. Então o valor de c será dado por
$c = F^{-1}_{t(14)}\left(1 - \frac{\alpha}{2}\right) = F^{-1}_{t(14)}(1 - 0.025) = F^{-1}_{t(14)}(0.975) = 2.145$,
logo

$$RC_{0.05} = (-\infty, -2.145) \cup (2.145, +\infty).$$

Decisão:
Como o valor observado da estatística de teste, $t_{obs} = -5.44 \in RC_{0.05}$, devemos rejeitar H_0 ao nível de significância de 5%, ou seja, parece haver evidência de que a idade da mulher influencia a sua massa muscular.

Exercício c)

c) Calcule o intervalo de confiança a 99% para o valor esperado da massa muscular para uma mulher de 45 anos. Acha legítimo usar o mesmo procedimento tratando-se de uma mulher com 20 anos em vez de 45?

Pretende-se um intervalo de confiança a 99% para $E[Y|x = 45]$. Observar que $x = 45 \in (\min(x_i), \max(x_i))$.

Variável aleatória fulcral:

$$T = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(14)},$$

com $x_0 = 45$.

Exercício c)

Como $1-\alpha = 0.99$ então $\alpha = 0.01$ e
$a = F^{-1}_{t(14)}\left(1-\frac{\alpha}{2}\right) = F^{-1}_{t(14)}(1-0.005) = F^{-1}_{t(14)}(0.995) = 2.977.$

Como a distribuição da t-Student é simétrica vem:

$$P\left(-2.977 \leq \frac{\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right)-\left(\beta_0+\beta_1 x_0\right)}{\sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \leq 2.977\right) = 0.99$$

$$\equiv P\left(\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) - 2.977\sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2} \leq \left(\beta_0+\beta_1 x_0\right) \leq \right.$$

$$\left. \leq \left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) + 2.977\sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}\right) = 0.99$$

Exercício c)

Obtendo-se então o intervalo aleatório:

$$IAC_{99\%}\left(\beta_0+\beta_1 x_0\right)=\left(\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) \pm 2.977 \sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum_{i=1}^{n} x_i^2-n\bar{x}^2}\right)\hat{\sigma}^2}\right)$$

Concretização: O intervalo de confiança, a 99% de confiança é dado por:

$$IC_{99\%}\left(\beta_0+\beta_1\times 45\right)=$$

$$\left((148.05-1.0236\times 45) \pm 2.977\sqrt{\left(\frac{1}{16}+\frac{\left(\frac{967}{16}-45\right)^2}{1965.9375}\right)69.62}\right)$$

$$=(91.34, 112.63)$$

Para $x=20$ anos não é correto utilizar o mesmo procedimento pois este valor não pertence ao $(\min(x_i), \max(x_i))$. Qualquer extrapolação para valores que não fizeram parte do ajuste do modelo será um abuso.

Obrigada!

Questões?