Resumo P3 | Akira 2017.1

1) Testes não paramétricos:

→ Teste de Aderência: → Queremos estimar o tipo de distribuição que a população segue a partir dos dados de uma amostra (dados com "boa" ou "má" aderência ao modelo")

A hipótese testada refere-se a forma de distribuição da população.

H_0: os dados observados provém de uma distribuição "x" com parâmetro "γ"

H_1: os dados observados NÃO provém de uma distribuição "x" com parâmetro "γ"

• Teste de Aderência pelo χ^2

$$\boxed{\chi^2_{calc} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n}$$

→ média dos O_i calculado a partir da distribuição que estamos analisando com o dado parâmetro ($E_i = n \cdot P_i(x=x)$) poisso, binomial, etc.

→ observações realizadas = frequências

Obs: $E_i \geq 5$, caso contrário somar os valores de $E_i \leq 5$ com a classe anterior em que $E_i \geq 5$. (fazer o mesmo com as observações)

↳ $n = \sum_{i=1}^{k} E_i = \sum_{i=1}^{k} O_i$

$$\boxed{g.l. = \gamma = k - 1 - m}$$

k → número de parcelas após a junção

m → número de parâmetros estimados a partir dos dados da amostra (para o calculos dos p_i).

Se $\chi^2_{calc} > \chi^2_{\gamma,\alpha}$ → rejeito H_0!

• Teste de Aderencia pelo "Papel de Probabilidade Normal"

→ para estimar μ tomamos o ponto de 50%

→ para estimar σ tomamos -1σ e $+1\sigma$

No eixo horizontal devemos colocar as frequências (observações) e no eixo vertical devemos colocar as frequências relativas acumuladas dada por: $\boxed{x' = \frac{50(2i-1)}{n}}$ → $i = 1, \ldots, n$

↳ Se obtivermos uma reta, os dados tem aderencia pela normal!

• Teste de Aderência pelo Método de Kolmogrov - Smirnov

→ ordenar os dados da amostra!

Trata-se de um Teste mais poderoso que o χ^2 para testar a aderência. A variável de teste é a maior diferença observada entre a função de distribuição acumulada do modelo e da amostra.

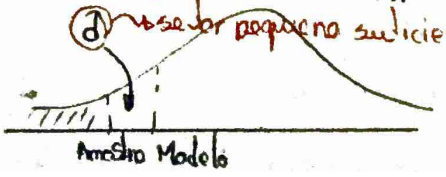
↳ f.d.a do modelo: $F(x) = P(x \geq x)$ → utilizamos a normal!

↳ f.d.a da amostra: $G(x)_{esquerda} = \frac{(i-1)}{n}$, $G(x)_{direita} = \frac{i}{n}$ ↳ $Z_i = \frac{x_i - \mu}{\sigma}$

$$\boxed{d_{calc} = \max\{|F(x) - G(x)|\}}$$

Obs: se $n > 50$: $d_{crit} = \frac{1{,}36}{\sqrt{n}}$ ($\alpha = 5\%$) e $d_{crit} = \frac{1{,}63}{\sqrt{n}}$ ($\alpha = 1\%$)

Se $d_{calc} > d_{crit}$ → rejeito H_0!



→ se for pequeno suficiente → "aceito" H_0

→ Teste de Independência:

$$\begin{cases} H_0: \text{as variáveis são independentes} \\ H_1: \text{as variáveis não são independentes} \end{cases}$$

$$\boxed{\chi^2_{\nu} = \chi^2_{calc} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}} \; ; \; \boxed{E_{ij} = \frac{f_i \cdot f_j}{n}} \; ; \; \boxed{\nu = (r-1)(s-1)}$$

O_{ij} → frequência observada na intersecção da linha i com a coluna j

f_i → total da linha

f_j → total da coluna

r → nº de linhas

s → número de colunas

II) ANOVA

→ Análise de Variancia (ANOVA) → comparar as médias das populações ⇒ identifica diferenças entre médias populacionais devido a diversas causas (classificações) atuando simultaneamente sobre os elementos da população.

Hipóteses
- Populações homocedásticas → variâncias iguais
- Populações com distribuições iguais normais

1) Uma classificação – amostras de mesmo tamanho

$$\begin{cases} H_0: \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1: \text{pelo menos uma das médias é diferente das demais} \end{cases}$$

k = número de amostras
n = tamanho da amostra
$\boxed{N = n \cdot k}$ = tamanho total

Notação:

$$\begin{cases} T_i = \sum_{i=1}^{n} x_i \longrightarrow T = \sum_{j=1}^{k} T_j \\ Q_i = \sum_{i=1}^{n} x_i^2 \longrightarrow Q = \sum_{j=1}^{k} Q_j \end{cases}$$

→ Posso subtrair o menor termo para facilitar os calculos!
$\sigma^2(cte + x) = \sigma^2(x)$

Fonte de Variação	Soma de Quadrados	G.L.	Quadrado Medio	F_{calc}	F_{α}
Entre Amostras	$SQE = \sum_{i=1}^{k} \frac{T_i^2}{n} - \frac{T^2}{nk}$	$k-1$	$S_E^2 = \frac{SQE}{k-1}$	$F = \frac{S_E^2}{S_R^2}$	$F_{(k-1),(n-1),\alpha}$
Residual	$SQR = Q - \sum_{i=1}^{k} \frac{T_i^2}{n}$	$k(n-1)$	$S_R^2 = \frac{SQR}{k(n-1)}$		
Total	$SQT = Q - \frac{T^2}{nk}$	$nk-1$			

→ numerador
→ denominador

Equação Fundamental: $\boxed{SQT = SQR + SQE}$

Probabilidades que não sabemos aplicar / separar (entre elementos de cada amostra ⇒ dentro da amostra)

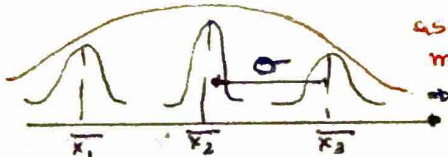
→ Sendo H_0 verdadeiro podemos estimar σ^2 como de 3 maneiras possíveis: S_T^2, S_E^2, S_R^2

$$\begin{cases} H_0: \sigma_E^2 = \sigma_R^2 \\ H_1: \sigma_E^2 > \sigma_R^2 \end{cases}$$

Para amostras de mesmo tamanho: $\boxed{S_R^2 = \frac{S_1^2 + S_2^2 + \ldots + S_k^2}{k}}$

Se não há variações dentro das amostras e entre as amostras podemos estimar que as amostras provem da mesma população.

Se $F_{calc} > F_{crit}$ ⇒ rejeito H_0



⇒ H_1 é verdadeiro

⇒ H_0 é verdadeiro

2) Uma classificação – amostras de tamanhos diferentes

Fonte de Variação	Soma de Quadrados	G.L.	Quadrado Médio	Fcalc	Fα
Entre Amostras	$SQE = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{\sum_{i=1}^{k} n_i}$	$k-1$	$S_E^2 = \frac{SQE}{k-1}$	$F_{calc} = \frac{S_E^2}{S_R^2}$	$F_{k-1;\Sigma n_i - k,\alpha}$
Residual	$SQR = Q - \sum_{i=1}^{k} \frac{T_i^2}{n_i}$	$\sum_{i=1}^{k} n_i - k$	$S_R^2 = \frac{SQR}{\Sigma n_i - k}$		
Total	$SQT = Q - \frac{T^2}{\sum_{i=1}^{k} n_i}$	$\sum_{i=1}^{k} n_i - 1$			

$$\boxed{SQT = S_x^2 \cdot G.L}$$

$$\boxed{S_R^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_k-1)S_k^2}{n_1 + n_2 + \ldots + n_k - k}}$$

Quais médias devem ser consideradas diferentes de quais outras?

Método de Tukey: para k linhas e n colunas, as médias μ_i e μ_m serão consideradas distintas se:

$$\boxed{|\bar{x}_i - \bar{x}_m| > \underbrace{q_{k,\nu,\alpha}}_{\text{amplitude studentizada}} \sqrt{\frac{S_R^2}{n}}}$$

$\nu = k(n-1)$

para cada amostra

Método de Scheffé: as médias μ_i e μ_m serão consideradas distintas entre si, se:

$$\boxed{|\bar{x}_i - \bar{x}_m| > \sqrt{S_R^2 \cdot \frac{2(k-1)}{n} F_{k-1,k(n-1),\alpha}}}$$

Duas Classificações: os elementos observados são ~~observados~~ classificados segundo dois critérios. Admitindo que haja nk observações, do primeiro critério temos n elementos e k amostras, e do segundo critério temos k elementos e duas amostras.

H_{0_1}: $\mu_{1,linha} = \mu_{2,linha} = \dots = \mu_{k,linha}$

H_{0_2}: $\mu_{1,coluna} = \mu_{2,coluna} = \dots = \mu_{n,coluna}$

H_1: pelo menos uma das médias é diferente das demais → linha interfere

H_2: pelo menos uma das médias é diferente das demais → coluna interfere

<u>Modelo fixo</u>: os efeitos resultantes das classificações segundo linhas e colunas são ambos fixo, ou seja representam a totalidade das condições existentes.

3) Duas Classificações sem repetição: → Modelo fixo ⇒ (hipótese) inexistência de interação entre linhas e colunas.

↳ podemos estimar σ^2 através de S_T^2, S_L^2, S_C^2 e S_R^2.

Fonte de Variação	Soma de Quadrados	G.L.	Quadrado Médio	Fcalc	Fα
Entre Linhas	$SQL = \frac{\sum_{i=1}^{k} T_i^2}{n} - \frac{T^2}{nk}$	$k-1$	$S_L^2 = \frac{SQL}{k-1}$	$F_L = \frac{S_L^2}{S_R^2}$	$F_{k-1,(k-1)(n-1),\alpha}$
Entre Colunas	$SQC = \frac{\sum_{j=1}^{n} T_j^2}{k} - \frac{T^2}{nk}$	$n-1$	$S_C^2 = \frac{SQC}{n-1}$	$F_C = \frac{S_C^2}{S_R^2}$	$F_{n-1,(k-1)(n-1),\alpha}$
Residual	$SQR = SQT - SQC - SQL$	$(k-1)(n-1)$	$S_R^2 = \frac{SQR}{(k-1)(n-1)}$		
Total	$SQT = Q - \frac{T^2}{nk}$	$nk-1$			

k = linhas

n = colunas

4) Duas Classificações com repetição:

se houver interação! (testa o efeito fixo)

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	Fcalc	Fα
Entre Linhas	$SQL = \frac{\sum_{i=1}^{k} T_i^2}{nr} - \frac{T^2}{nkr}$	$k-1$	$S_L^2 = \frac{SQL}{k-1}$	$F_L = \frac{S_L^2}{S_I^2}$	
Entre Colunas	$SQC = \frac{\sum_{j=1}^{n} T_j^2}{kr} - \frac{T^2}{nkr}$	$n-1$	$S_C^2 = \frac{SQC}{n-1}$	$F_C = \frac{S_C^2}{S_I^2}$	
Interação	$SQI = SQT_r - SQL - SQC$	$(k-1)(n-1)$	$S_I^2 = \frac{SQI}{(k-1)(n-1)}$	$F_I = \frac{S_I^2}{S_R^2}$	$F_{(k-1)(n-1),nk(r-1),\alpha}$
Entre Tratamentos	$SQT_r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} T_{ij}^2}{r} - \frac{T^2}{nkr}$	$nk-1$	$S_{Tr}^2 = \frac{SQT_r}{nk-1}$	$F_{T_r} = \frac{S_{T_r}^2}{S_R^2}$	$F_{nk-1,nk(r-1),\alpha}$
Residual	$SQR = SQT - SQT_r$	$nk(r-1)$	$S_R^2 = \frac{SQR}{nk(r-1)}$		
Total	$SQT = Q - \frac{T^2}{nkr}$	$nkr-1$			

"Tratamento" é cada combinação entre linha e coluna.

↳ pode haver uma variação entre os nk tratamentos aos quais cada elemento é submetido

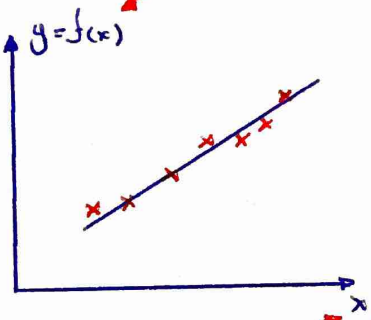
Portanto, também podemos estimar σ^2 por $S_{T_r}^2$.

"Interação" está ligado ao modo que os critérios se relacionam

↳ Devemos verificar se há ou não interação entre os critérios (linhas e colunas). Se não houver, o resíduo "absorve a interação" (somar os SQ e os G.L.).

Depois recalcular S_R^2

III) Correlação e Regressão



Variável dependente

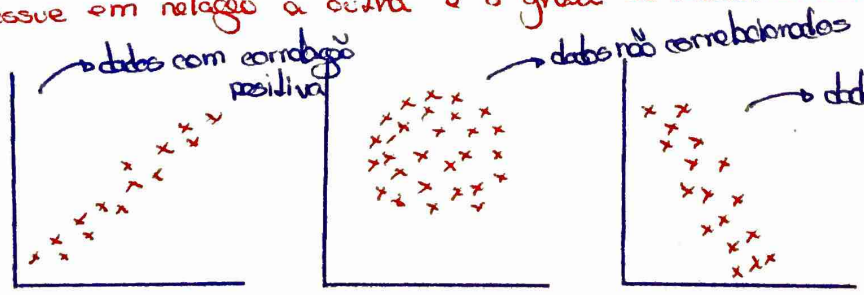
→ Regressão Linear: trata-se da reta que minimiza a distância dos pontos em y.

Variável independente (temos total controle sobre ela)

Aplicar o conceito de estratificação (dados provenientes de conjuntos homogêneos)

→ Correlação Linear:

↳ nos mostra como as variáveis tendem a se relacionar fornecendo a magnitude do conhecimento que uma possue em relação a outra e o grau de relacionamento



$-1 < r < 0$ $\quad$ $r \cong 0$ $\quad$ $0 < r < 1$

Obs:

$S_{xy} = \sum(x_i - \bar{x})y_i = \sum(y_i - \bar{y})x_i$

Coeficiente de Person:

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}}$$

só serve p/ reta!

$-1 \leq r \leq +1$

se $r = 1$ ou $r = -1$ temos que as variáveis se relacionam por uma reta

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i \cdot \sum y_i)}{n} \sim \cancel{\sum(x_i - \bar{x})}$$

pode ser ⊕ ou ⊖

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum(x_i - \bar{x})^2$$

⊕ sempre

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum(y_i - \bar{y})^2$$

⊕ sempre

Coeficiente R^2:

$$\boxed{R^2 = \frac{SQT - SQR}{SQT} = r^2}$$

SQE

→ é mais geral

→ equação que relaciona as variáveis de interesse.

Teste para o coeficiente de correlação:

Coef. de Correlação populacional: $H_0: \rho = 0$; $H_1: \rho \neq 0$

$t_{calc} = r\sqrt{\frac{n-2}{1-r^2}}$

$t_{crit} = t_{n-2, \alpha/2}$

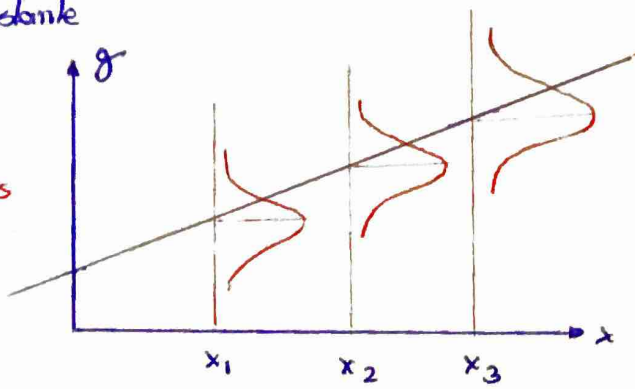
→ Regressão → após verificada a correlação entre os dados, faz-se necessário encontrar uma função que exprima

esse nota [illegible]

temos o total controle sobre elas

Hipóteses:
- $x_1, \dots, x_n$ são admitidos sem erros
- y é admitido sem erro. Este erro é $\varepsilon \sim N(0; \sigma_\varepsilon^2)$ → média 0 → variância residual
- σ_ε^2 = constante

Regressão Linear Simples: → Duas variáveis



→ média de y

Modelo Teórico:

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

média de y para cada x

Reta Teórica: $y_i = \alpha + \beta x$ → reta verdadeira

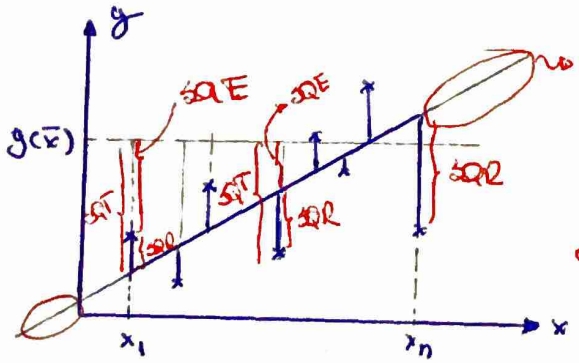
Reta Estimativa: $\hat{y} = a + bx$

coef. linear → a; coef. angular → b

Para achar a e b utilizaremos o método dos mínimos quadrados (MMQ)

Método dos Mínimos Quadrados:

→ a e b que mais se ajustam ao modelo são aqueles que tornam mínima a diferença entre cada ponto e a reta estimada (distância mínima)



→ não sabemos o que ocorre

$SQT_y = \sum (y_i - \bar{y})^2$

$SQR_y^{min} = \min \sum d_i^2 =$

$= \min \sum (y_i - \hat{y}_i) =$

$= \min \sum (y_i - a - bx_i)^2$

o MMQ vai reduzir ao máximo a variação residual (SQR)

$$\boxed{b = \frac{S_{xy}}{S_{xx}}} \; ; \; \boxed{a = \bar{y} - b\bar{x}}$$

$\hat{y} = a + bx$

Valor Médio!

Obs:
- Sempre há uma reta que representa um conjunto de dados
- Mas nem sempre esta reta é o melhor modelo

Variancia Residual:

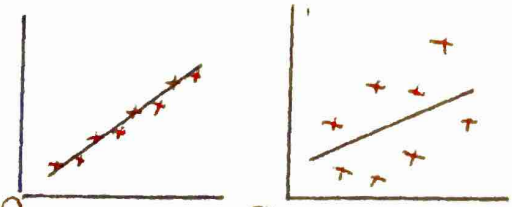
$$\boxed{S_\varepsilon^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}}$$

→ estimação de a e b

→ Variância em torno da reta dos mínimos quadrados.

$SQR + \sum (\hat{y} - \bar{y})^2 = S_{yy} = SQT$

$\sum (y_i - \hat{y}_i)^2$



Variancia Residual pequena → S_ε^2 grande

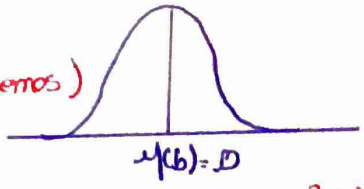
a e b são variáveis aleatórias → a,b ~ Normal (μ; σ)

Variável b:

$\mu(b) = \beta$ ← Estamos supondo que o modelo linear é satisfatório

$\sigma^2(b) = \dfrac{\sigma_R^2}{S_{xx}}$ → Residual (não controlamos)

↳ Maior a amostra, menor será o $\sigma^2(b)$



$\hat{S}^2(b) = \dfrac{\hat{S}_R^2}{S_{xx}}$

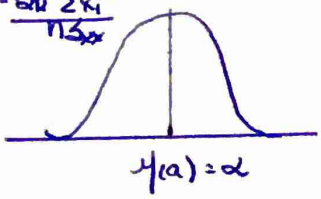
Variável a:

$\mu(a) = \alpha$

$\sigma^2(a) = \dfrac{\sigma_R^2 \sum x_i^2}{n S_{xx}}$

↳ Maior a amostra, menor será o $\sigma^2(a)$

$\hat{S}^2(a) = \dfrac{\hat{S}_R^2 \sum x_i^2}{n S_{xx}}$



} Intervalos de confiança

↳ a, b, $\hat{y}$ → α, β, y

média → previsões

Intervalo de Confiança (para $y_i = \alpha + \beta x_i$)

↳ onde a reta teórica estará dado x_i

$\mu(y_i) = \alpha + \beta x_i$

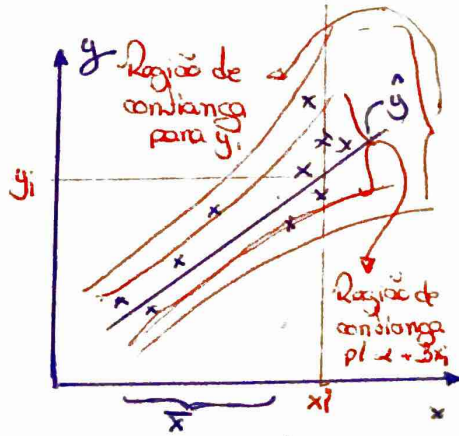
Estimativa: $a + b x_i$ ∴ $Var(a + b x_i) = \sigma_R^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$

→ ponto qualquer!

→ quanto mais afastado da média dos "dados controlados" maior será a variância da reta estimada!

quanto mais longe x_i maior o IC!

IC: $\underbrace{\hat{y}_i \pm t_{n-2;\alpha/2} \sqrt{\hat{S}_R^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}}_{\text{Região de confiança para } \alpha + \beta x_i \text{ (reta verdadeira)}}$



Intervalo de Confiança (para uma previsão de y_i dado x_i) → onde y_i estará dado x_i

Estimativa: $a + b x_i$ ∴ $Var(a + b x_i + \varepsilon_i) = \sigma_R^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$

IC: $\hat{y}_i \pm t_{n-2;\alpha/2} \sqrt{\hat{S}_R^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}$

→ Teste de Hipótese e Análise de Variância aplicados à regressão.

Quando desejamos verificar se determinada reta de regressão tem inclinação β_0 podemos fazer o seguinte teste:

ou inclinação maior/menor que β_0

$\begin{cases} H_0: \beta = \beta_0 \\ H_1: \beta > \beta_0 \end{cases}$

$\boxed{t_{calc} = \dfrac{b - \beta_0}{\sqrt{\hat{S}^2(b)}} = \dfrac{b - \beta_0}{\hat{S}_R / \sqrt{S_{xx}}}}$; $\boxed{t_{crit} = t_{n-2, \alpha \text{ ou } \alpha/2}}$

↳ teste de correlação (se $\beta = 0$)

Caso particular (teste de correlação):

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

H_0: → Se H_0 for verdadeiro então y não depende de x, pois, $y = \alpha + \cancel{\beta x}^{\,0} + \varepsilon_i$ ∴ não há correlação

H_1: → Modelo tem validade (há correlação)

Quando queremos saber se a reta passa pela origem : $\begin{cases} H_0: \alpha = 0 \\ H_1: \alpha \neq 0 \end{cases}$ → passa pela origem

$$\boxed{t_{calc} = \frac{a-\alpha_0}{s(a)}}, \text{ onde } \boxed{s^2(a) = \frac{s_R^2 \sum x_i^2}{n s_{xx}}} \qquad \boxed{t_{crit} = t_{n-2,\alpha/2}}$$

ANOVA para regressão → semelhante a fazer o teste $\beta_0 = 0$.

Verificar se o modelo de regressão é satisfatório

Lembrar que : $\boxed{\underbrace{\sum (y_i - \bar{y})^2}_{(n-1)} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{(1)} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{(n-2)}}$

SQT (variação em torno da média) proporcionará

SQR (proporcionará a variação em torno da reta de regressão)

SQE (proporcionará a variância explicada pela regressão)

Intuição: se a variação em torno da média (s_y^2) for muito maior que a variação em torno da reta de regressão (s_R^2) ⇒ a regressão é válida.

Fórmulas :

$$\boxed{SQT = \sum (y_i - \bar{y})^2 = s_{yy}}$$

$$\boxed{SQE = \sum (\hat{y} - \bar{y})^2 = b^2 s_{xx} = b s_{xy}}$$

$$\boxed{SQR = \sum (\hat{y} - y_i)^2 = SQT - SQE = s_{yy} - b^2 s_{xx}}$$

Fonte Variação	Soma de Quadrados	G.L.	Quadrado Médio	Fcalc	Fcrit
Explicada pela Regressão	SQE	1	$\frac{b^2 s_{xx}}{1}$	$F_{calc} = \frac{b^2 s_{xx}}{s_R^2}$	$F_{1,n-2,\alpha}$
Residual	SQR	n-2	s_R^2		
Total	SQT	n-1			

→ Fcalc > Fcrit : há evidências para validar a regressão.

Se $\sigma_E >> \sigma_R$ a variação explicada pela regressão não provém da variância residual. Logo a uma outra razão para sua existência. Está razão é a dependência de x na forma linear.

→ **Regressão Polinomial**

Tratando-se de um polinômio de segundo grau : $y = \alpha + \beta x + \gamma x^2$

Estimativa : $\hat{y}_p = a + bx + cx^2$

Para achar a,b,c basta resolver : $\hat{y}_p = a + b(x-\bar{x}) + c(x-\bar{x})^2$

$$\begin{cases} \sum y_i = na + c\sum (x_i - \bar{x})^2 \\ \sum (x_i - \bar{x}) y_i = b \sum (x_i - \bar{x})^2 \\ \sum (x_i - \bar{x})^2 y_i = a \sum (x_i - \bar{x})^2 + c \sum (x_i - \bar{x})^4 \end{cases}$$

s_{xx} → Pode ser calculado igual na reta

s_{xy}

Tratando-se de um polinômio de grau maior que 2 :

$$\begin{cases} \sum y_i = na + b\sum x_i + c\sum x_i^2 + d\sum x_i^3 + \dots \\ \sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 + d\sum x_i^4 + \dots \\ \sum x_i y_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 + d\sum x_i^5 + \dots \\ \vdots \end{cases}$$

→ Regressão múltipla

↳ y agora depende de mais de um x

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

calcular como S_{xx}

$$\begin{cases} S_{1y} = b_1 S_{11} + b_2 S_{12} + \dots + b_k S_{1k} \\ S_{2y} = b_1 S_{21} + b_2 S_{22} + \dots + b_k S_{2k} \\ \vdots \\ S_{ky} = b_1 S_{k1} + b_2 S_{k2} + \dots + b_k S_{kk} \end{cases}$$

calcular como S_{xy}

$$S_{kj} = \sum_{j=1}^{n} (x_{kj} - \bar{x}_k)(y_j - \bar{y}) = \sum x_{kj} y_j - \frac{\sum x_{kj} \sum y_j}{n}$$

~~$S_{kk} = \sum x_{ij}$~~