



Regressao Simples

Parte II: Anova, Estimaco Intervalar e
Predico

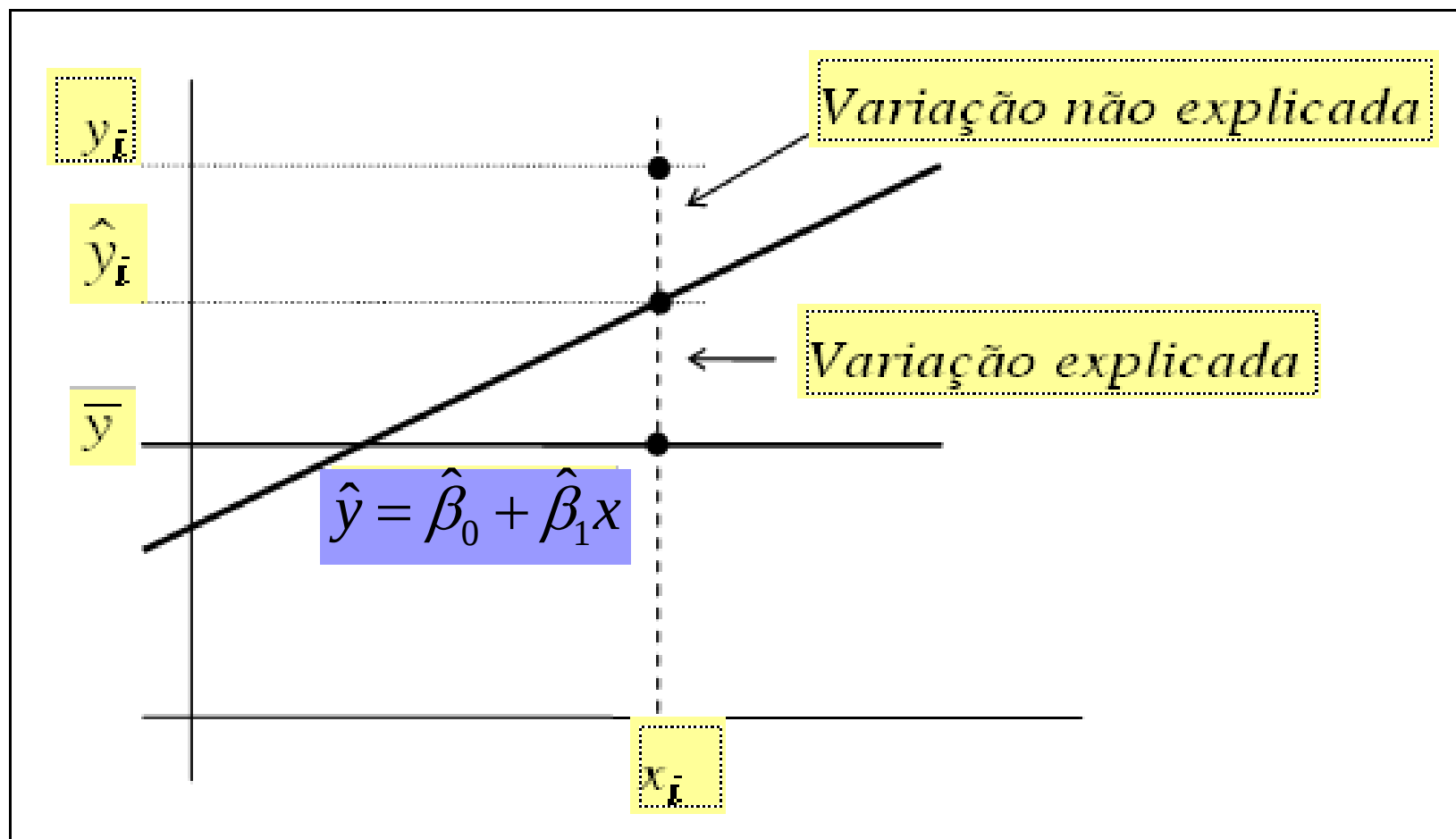
Análise de Variância

- Nem todos os valores das amostras estão contidos na reta de regressão, e quanto mais afastados estiverem pior, a reta representará a relação entre as amostras.
- A reta obtida pelo método dos quadrados mínimos é um resumo útil da tendência entre as variáveis, pois não explica perfeitamente os dados.
- Quão útil é a reta de regressão obtida pelo procedimento apresentado?
- Testar a significância do modelo por analisando a variabilidade.
- Anova: É baseada no particionamento da variabilidade total na variável resposta

Análise de Variância

- Vamos supor que escolhemos como modelo de regressão a reta de regressão horizontal $\hat{y} = \bar{y}$ a equação que representa a média da variável dependente y .
- Nesse caso, o coeficiente b da reta de regressão é igual a zero e, conseqüentemente, o coeficiente de correlação também é nulo.
- Embora a reta da média pouco explique, ela é um ponto importante de partida para medir variações. Analisando a reta de regressão com os coeficientes a e b , pode-se ver que a maioria dos valores das amostras está dispersa ao redor da reta, como mostra a figura do slide seguinte para um par de valores fora da reta.

Análise de Variância



Análise de Variância

- *Variação total* é o resultado da soma dos quadrados dos desvios dos valores y com relação à média:

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

- *Variação explicada* é o resultado da soma dos quadrados dos desvios dos valores estimados com relação à média:

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

- *Variação não-explicada* é o resultado da soma dos quadrados dos desvios de y com relação aos valores projetados:

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Análise de Variância

- Variação total é o resultado da soma dos quadrados dos desvios dos valores y com relação à média:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y})^2 + \sum_{i=1}^n (\hat{y} - \bar{y})^2$$

$$SQT = SSM + SSR$$

- SST tem $n-1$ graus de liberdade; um grau de liberdade é perdido devido a restrição de que a soma dos desvios em torno da média é zero. De outra forma: um grau de liberdade é perdido porque a média da amostra é usada para estimar a média populacional.
- SSR tem $n-2$ graus de liberdade. Dois graus de liberdade são perdidos pois dois parâmetros são estimados para obter $\hat{Y}$ (chapéu).
- SSM tem 1 grau de liberdade pois os graus de liberdade são aditivos. Isto é $n-1 = 1 + (n-2)$

Análise de Variância

	Soma de quadrados		Graus de liberdade	Média de Soma de Quadrados	F_0
Regressão	SSM	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	$MS_M = \frac{\left[\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \right]}{1}$	$\frac{MS_M}{MS_R}$
Residual	SSR	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	n-2	$MS_R = \frac{\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]}{n-2}$	
Total	SST	$\sum_{i=1}^n (y_i - \bar{y})^2$	n-1		

Análise de Variância

- SSR/σ^2 tem distribuição qui-quadrado com $n-2$ graus de liberdade.
- SSM/σ^2 tem distribuição qui-quadrado com 1 grau de liberdade.
- SSR e SSM são independentes.
- Por definição, F_0 segue uma distribuição
- F-Snedecor com 1 e $n-2$ graus de liberdade.
- Se F_0 é grande então coeficiente β_1 é diferente de zero.
- Hipoteses:
 - $H_0: \beta_1 = 0$
 - $H_1: \beta_1 \neq 0$
- Rejeita H_0 , se $F_0 > F\text{-Snedecor com } 1 \text{ e } n-2 \text{ graus de liberdade.}$

- Obs: Se não rejeita H_0 isso não pode significar que x e y não estão relacionados. Isto pode significar que o range dos valores de x não é adequado ou que a variabilidade do processo de medição foi insuficiente para expressar a relação.

Estimação Intervalar: β_1

- Se os erros são normalmente distribuídos, então

$$\hat{\beta}_1 \rightarrow ND(\beta_1, \sigma^2 / S_{XX})$$

- e $Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{XX}}} \rightarrow N(0,1)$ com variância conhecida e o intervalo de confiança é .

$$\hat{\beta}_1 - z_{\alpha/2} \sqrt{MS_R / S_{XX}} \leq \beta_1 \leq \hat{\beta}_1 + z_{\alpha/2} \sqrt{MS_R / S_{XX}}$$

- Se a variância é desconhecida $t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_R / S_{XX}}} \rightarrow t_{n-2}$

$$\hat{\beta}_1 - t_{n-2, \alpha/2} \sqrt{MS_R / S_{XX}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2, \alpha/2} \sqrt{MS_R / S_{XX}}$$

Estimação Intervalar: β_0

- Se os erros são normalmente distribuídos com variância conhecida, então

$$\hat{\beta}_0 \rightarrow ND(\beta_0, \sigma^2 (1/n + \bar{x}^2 / S_{XX}))$$

- Tipicamente a variância não é conhecida

$$\hat{\beta}_0 \rightarrow ND(\beta_0, MS_R (1/n + \bar{x}^2 / S_{XX}))$$

- e o intervalo de confiança é .

$$\hat{\beta}_0 - t_{n-2, \alpha/2} \sqrt{MS_R (1/n + \bar{x}^2 / S_{XX})} \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2, \alpha/2} \sqrt{MS_R (1/n + \bar{x}^2 / S_{XX})}$$

Estimação Intervalar: σ^2

- Se os erros não normalmente distribuídos

$$(n-2) \frac{MS_R}{\sigma^2} \rightarrow \chi_{n-2}^2$$

- Então

$$\frac{(n-2)MS_R}{\chi_{\alpha/2, n-2}^2} \leq \sigma^2 \leq \frac{(n-2)MS_R}{\chi_{1-\alpha/2, n-2}^2}$$

Estimação intervalar: Média da variável resposta

- Sabemos que $E(y / x_0) = \beta_0 + \beta_1 x_0$ e deseja-se IC para a média da variável resposta.
- Um estimador não enviesado é encontrado como $\hat{E}(y / x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- O estimador acima é uma variável aleatória normalmente distribuída pois ele é uma combinação linear das observações y_i . E a variância é

$$Var(\hat{E}(y / x_0)) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right]$$

- IC

$$\left(\hat{E}(y / x_0) - t_{\alpha/2, n-2} \sqrt{MS_R \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right]}, \hat{E}(y / x_0) + t_{\alpha/2, n-2} \sqrt{MS_R \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right]} \right)$$

Predição

- Se x_0 é uma observação de interesse, então a predição para a média de y

$$\hat{E}(y / x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Predição intervalar

- Considere obtendo uma estimativa intervalar para uma observação futura de y .
- Essa estimativa intervalar para o parâmetro $E(y/x)$ não é apropriado.

$$\hat{E}(y / x_0) = \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Seja $\phi = y_0 - \hat{y}_0$ uma variável aleatória normalmente distribuída com média zero e variância

$$Var(\phi) = Var(y_0 - \hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$y_0 \in \left[\hat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{MS_R \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \right]$$

Predição intervalar

- A diferença entre estimar uma resposta média, $E(y_0/x)$ e fazer a predição de uma nova observação, y_0 , é que no primeiro caso estimamos a *média* da distribuição de Y . Agora, vamos predizer uma resposta *individual* da distribuição de Y .
- A predição intervalar de nova observação é sempre maior que o IC para a resposta média

Considerações

- Autores de livros recomendam que nunca devemos usar valores de x que extrapolem do range dos valores originais. O erro ou o modelo pode prejudicar a predição.
- Através da equação do IC para a média da variável resposta, pode-se verificar que para valores remotos de x o desvio produz uma estimativa mais variável da média de y dado x : $\hat{X} - \bar{X}$
- Engenheiros e economistas fazem extrapolação com um ou mais períodos no futuro. Contudo deve se tomar cuidado com o impacto disso nos resultados.