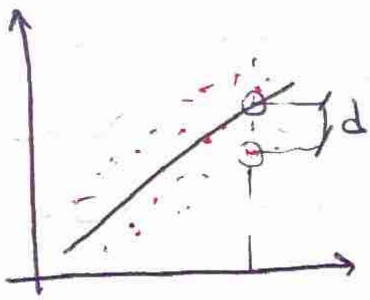
ETAPA 2 - PROJETO 3

LUCAS FONTENLA & GUILHERME SCHOVERI

(A) Partindo de uma dispersão, como demonstrado abaixo, podemos calcular o erro como sendo uma distância entre um ponto y "original" (dos dados) e um $\hat{y}$, que é um ponto em uma reta. O Erro ($\varepsilon$) é ao quadrado pois elimina o sinal negativo



$$d \rightarrow \underline{d^2 = \varepsilon} = (y_i - \hat{y}_i)^2$$

Assim, podemos escrever a equação da reta $\hat{y}$:

$$\hat{y} = \beta_0 + \beta_1 x_i$$

Desse modo,

$$\varepsilon = [y_i - (\beta_0 + \beta_1 x_i)]^2$$

A partir do Erro ($\varepsilon$) podemos calcular o Erro total de $x_1$ a $x_n$, como a soma de todos os erros $\varepsilon_i$.

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \rightarrow (1)$$

Desse modo vamos desenvolver a soma em (1)

$$[y_i - (\beta_0 + \beta_1 x_i)]^2 = y_i^2 - 2y_i\beta_0 - 2x_i y_i \beta_1 + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2$$

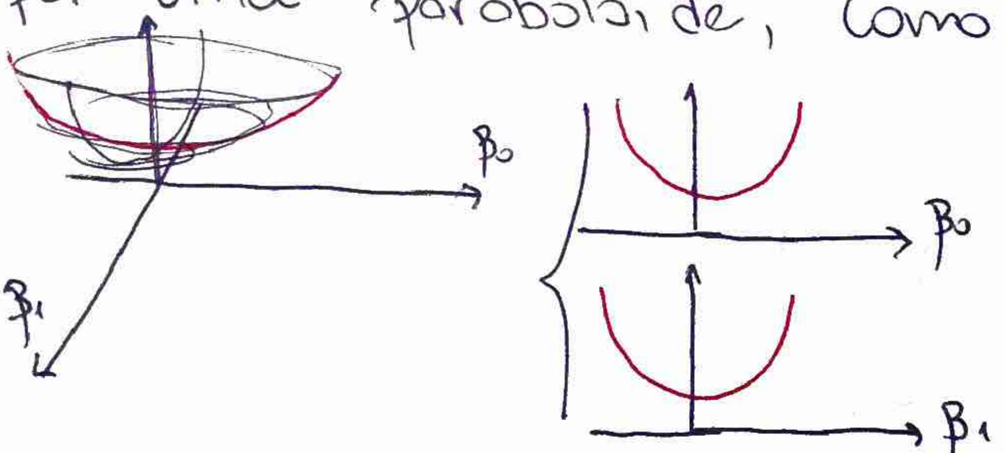
Como a porção anterior que foi "aberta" está em uma somatório, podemos escrever:

$$\frac{\sum y_i^2}{n} = \overline{y^2} \;\Big|\; \frac{\sum x_i^2}{n} = \overline{x^2} \;\Big|\; \frac{\sum y_i}{n} = \overline{y} \;\Big|\; \frac{\sum x_i}{n} = \overline{x} \;\Big|\; \frac{\sum x_i y_i}{n} = \overline{xy}$$

Ou seja, podemos reescrever tudo como:

$$\varepsilon^2 = n.\overline{y^2} - 2\beta_1 n \overline{xy} - 2\beta_0 \overline{y} + \beta_1^2 n \overline{x^2} + 2\beta_1\beta_0\overline{x} + n\beta_0^2$$

A equação acima é em função de duas variáveis $\beta_0$ e $\beta_1$, e ambos estão em sua forma quadrática, a partir disso, podemos inferir que o gráfico será dado por uma paraboloide, como abaixo.




Ambos, $\beta_0$ e $\beta_1$ são parábolas e para isso queremos os pontos no eixo nos quais a outra variável não varia, ou seja a derivada parcial em $\beta_0$ e $\beta_1$ tem de ser 0 (zero).

$$\text{I} \quad \frac{\partial \varepsilon}{\partial \beta_0} = 0 \qquad e \qquad \text{II} \quad \frac{\partial \varepsilon}{\partial \beta_1}$$

Assim,

$$\text{I} \quad \frac{\partial \varepsilon}{\partial \beta_0} = -2n\overline{xy} + 2\beta_1 n\overline{x^2} + 2\beta_0 n\overline{x} = 0 \quad /2n$$

$$= \boxed{-\overline{xy} + \beta_1 \overline{x^2} + \beta_0 \overline{x} = 0}$$

$$\text{II} \quad \frac{\partial \varepsilon}{\partial \beta_1} = -2n\overline{y} + 2\beta_1 n\overline{x} + 2n\beta_0 = 0 \quad /2n$$

$$= \boxed{-\overline{y} + \beta_1 \overline{x} + \beta_0 = 0}$$

A partir de tudo podemos reescrever:

$$\overline{y} - \beta_1\overline{x} = \frac{\overline{xy} - \beta_1\overline{x^2}}{\overline{x}} \Longrightarrow \beta_1\left(\frac{\overline{x^2}}{\overline{x}} - \overline{x}\right) = \frac{\overline{xy}}{\overline{x}} - \overline{y}$$

$$\beta_1 = \frac{\overline{xy} - \overline{y}\,\overline{x}}{\overline{x^2} - (\overline{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

Desse modo, provamos que, o COEFICIENTE ANGULAR DA RETA É DADO PELA RAZÃO COVARIÂNCIA DE X E Y E A VARIÂNCIA DE X.

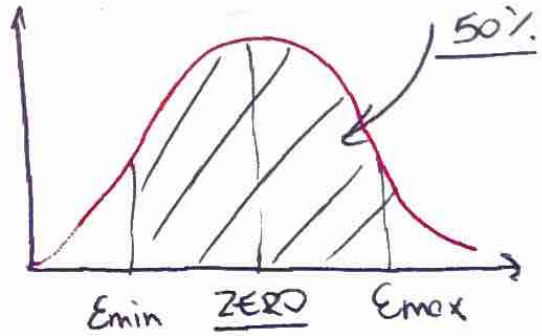
(B) As suposições:

ERRO: O Erro total segue uma distribuição normal com média 0 (zero) e o desvio-padrão $\sigma$, ou seja:

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

Podemos demonstrar seguindo a seguinte pergunta. Qual a probabilidade de termos um erro menor que 50% do valor.

Construindo a normal:




Desse modo, a média está em zero, pois é o ponto ajustado na reta que buscamos ter.

VARIÂNCIA: $Var(\varepsilon_i) = \sigma^2 = E(\varepsilon_i^2) - E^2(\varepsilon_i)$

FENÔMENO DA HOMOCELASTICIDADE: A variância é constante ao longo da reta.

(C) Podemos criar duas hipóteses (alternativa e a nula) simples.

$H_0: \beta_1 = 0 \longrightarrow$ (não há relação entre x e y)

$H_1: \beta_1 \neq 0 \longrightarrow$ (há relação entre x e y)

Se $H_0$ for rejeitado, demonstramos que há alguma relação entre X e Y, caso ~~seja~~ ~~todo~~ não rejeitada, dizemos que não há relação entre x e y.

(D) É possível realizar regressão com quantas variáveis se desejar, mas para cada variável adicionada, mais complicada é a equação.

Por exemplo se, fossemos realizar a regressão de x e w, a equação seria a seguinte:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 w$$

As suposições permaneceriam as mesmas do item b, mas adicionaríamos mais um eixo e consequentemente, avaliaríamos duas curvas normais.

Para o teste de hipóteses, seria igual ao c, mas como no caso b, duplicaríamos o número de hipóteses.

Desse modo, ficaríamos com algo:

<u>x</u>

$H_0 : \beta_1 = 0$
(não há relação entre x e y)

$H_1 : \beta_1 \neq 0$
(há relação entre x e y)

<u>w</u>

$H_0 : \beta_2 = 0$
(não há relação entre w e y)

$H_1 : \beta_2 \neq 0$
(há relação entre w e y)