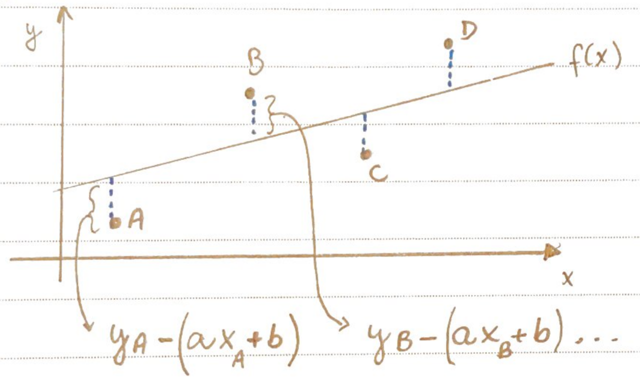
REGRESSÃO LINEAR

• FREDERICO VILELA CURTI E RAPHAEL COSTA

I Podemos encontrar a função que melhor descreve uma distribuição de uma variável aleatória, para uma reta, generalizando-a na forma

$$f(x) = ax + b$$

Os coeficientes a e b podem ser obtidos com as seguintes etapas:



I. Encontramos o valor do erro (diferença entre o valor de y e f(x) em cada ponto

Como alguns erros serão negativos, elevamos todos os erros ao quadrado para contornar isso, obtendo que o erro total acumulado assume:

$$\rightarrow \quad E^2 = (y_A - (ax_A + b))^2 + (y_B - (ax_B + b))^2 + \ldots (y_n - (ax_n + b))^2$$

* Simplificando a expressão, através dos produtos notáveis:

$$E^2 = y_A^2 - 2 \cdot y_A \cdot (ax_A + b) + (ax_A + b)^2 + y_B^2 - 2y_B \cdot (ax_B + b) + (ax_B + b)^2 + \ldots$$
$$+ y_n^2 - 2 \cdot y_n \cdot (ax_n + b) + (ax_n + b)^2$$

* Repetindo o processo e aplicando propriedades distributivas

$$= y_A^2 - 2y_A a x_A - 2y_A b + a^2 x_A^2 + 2ax_A b + b^2 + \ldots$$
$$+ y_n^2 - 2y_n \cdot ax_n - 2y_n b + a^2 x_n^2 + 2 \cdot ax_n b + b^2$$

II. Colocamos nossos termos de interesse em evidência:

$$E^2 = \Big[(y_A^2 + y_B^2 + \dots y_n^2) - 2a.(x_A y_A + x_B y_B + \dots + x_n y_n) - 2b(y_A + y_B + \dots + y_n) + a^2.(x_A^2 + x_B^2 + \dots + x_n^2) + 2.ab.(x_A + x_B + \dots + x_n) + n.b^2 \Big]$$

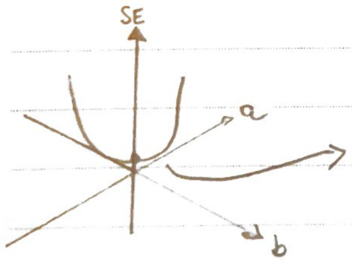
→ Os termos de x e y da expressão podem ser simplificados para suas médias por exemplo:

$$* \quad \frac{y_A^2 + y_B^2 + \dots + y_n^2}{n} = \overline{y^2} \cdot n$$

III. Obtemos a fórmula:

$$\rightarrow \quad (n.\overline{y^2}) - 2an\overline{xy} - 2bn.\overline{y} + a^2.n\overline{x} + 2abn\overline{x} + nb^2$$

c/ essa expressão, temos todos os valores possíveis para m e b. Com as derivadas parciais, podemos encontrar o menor valor destes coeficientes, minimizando o erro



$$E^2 = SE \text{ (Squared Error)}$$

$$\frac{\partial SE}{\partial a} = \frac{\partial SE}{\partial b} = 0$$

IV Derivação parcial

$$\frac{\partial SE}{\partial a} = -2n\overline{xy} + 2an\overline{x^2} + 2bn\overline{x} = 0 \quad (\div 2n)$$

$$\rightarrow -\overline{xy} + a\overline{x^2} + b\overline{x} = 0 \rightarrow a\overline{x^2} + b\overline{x} = \overline{xy} \quad (\div \overline{x}) \longrightarrow$$

$$\frac{\partial SE}{\partial b} = -2n\overline{y} + 2an\overline{x} + 2nb = 0 \quad (\div 2n)$$

$$\rightarrow -\overline{y} + a\overline{x} + b = 0 \rightarrow a\overline{x} + b = \overline{y} \longrightarrow$$

$$y = a\bar{x} + b \quad \rightarrow \left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right) \quad \rightarrow (\bar{x}, \bar{y})$$

V. Simplificando para isolar a em função de x e y

$$a \cdot \left(\bar{x} - \frac{\overline{x^2}}{\bar{x}}\right) = \bar{y} - \frac{\overline{xy}}{\bar{x}} \rightarrow a = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} \left(\frac{\bar{x}}{\bar{x}}\right)$$

$$\boxed{a = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}, \quad b = \bar{y} - a\bar{x}}$$

Para não limitarmos nossos coeficientes a e b, vamos chamá-los de β_1 e β_0, respectivamente. Logo, a função da reta terá a forma

$$\hat{y} = \beta_1 \cdot x + \beta_0$$

Ou seja:

$$\beta_0 = \bar{y} - a\bar{x} \quad e \quad \beta_1 = \frac{\bar{x}\,\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$$

Para efeitos de visualização ao multiplicarmos β_1 por -1, teremos:

$$\beta_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}$$, que equivale a:

$$\beta_1 = \frac{Cov(x, y)}{Var(x)}$$

b) As suposições feitas sobre os erros em termos de distribuição, valor esperado e variância são: os resíduos seguem uma distribuição normal; os erros são independentes entre si, ou seja, a covariância entres os erros é 0; o modelo é linear nos parâmetros; e a homocedasticidade, que consiste em considerar que os erros além de seguirem uma distribuição normal, seguem a mesma normal ao longo de toda reta.; As suposições são importantes para que seja possível efetuar o teste de Hipóteses com os erros da equação, para que o modelo torne-se ainda mais confiável.

c) O teste de hipótese efetuado na regressão simples consiste em:

Hipótese nula: $\beta 1 = 0$

Hipótese alternativa: $\beta 1 \neq 0$

Este teste consiste em: ao rejeitarmos a hipótese nula, afirmamos que x e y possuem uma relação, ou seja, faz sentido fazermos uma regressão linear com as variáveis. Ao não rejeitarmos a hipótese nula, o contrário ocorre, comprovamos que não existe relação entre x e y.

d) Sim. Quanto ao modelo conforme descrito na equação, teríamos a aparição de mais um $\beta 2.X2$; Quanto as suposições do modelo, teríamos que acrescentar um eixo no gráfico da regressão, pois teríamos y dependendo de 2 variáveis (X1 e X2); Quanto ao teste de hipótese, teríamos que efetuar o teste de hipótese duas vezes, sendo uma com o beta de y e X1 e outra com o beta de y e X2.