# Capítulo 11 — Introdução à Regressão

## Modelo de Regressão Linear Simples

Um modelo de regressão deste tipo traduz a relação entre uma variável quantitativa independente, $X$, e uma variável quantitativa dependente, $Y$, nos termos seguintes:

$$Y_n = \alpha + \beta \cdot (X_n - \bar{X}) + E_n$$

(modelo alternativo: $Y_n = \alpha' + \beta \cdot X_n + E_n$, com $\alpha' = \alpha - \beta \cdot \bar{X}$)

em que:

$n$: índice denotando as observações do par de variáveis $X$ e $Y$ ($n = 1, 2, \ldots, N$)
$\bar{X}$: média aritmética das observações $X_n$
$E_n$: erro aleatório associado ao valor observado $Y_n$

### Estimativas de $\alpha$ e $\beta$

$$\hat{\alpha} = a = \frac{1}{N} \cdot \sum_n y_n = \bar{y} \qquad \hat{\beta} = b = \frac{S_{XY}}{S_{XX}}$$

com,

$$S_{XY} = \sum_n [(X_n - \bar{X}) \cdot (Y_n - \bar{Y})] \qquad S_{XX} = \sum_n (X_n - \bar{X})^2 \qquad S_{YY} = \sum_n (Y_n - \bar{Y})^2$$

### Intervalos de Confiança a $(1-\gamma) \cdot 100\%$ para os parâmetros $\alpha'$, $\alpha$ e $\beta$

| | |
|---|---|
| Parâmetro $\alpha$ | $A \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{\frac{1}{N}}$ |
| Parâmetro $\alpha'$ (ordenada na origem do eixo dos $XX$) | $(A - \bar{X} \cdot B) \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{S_{XX}}}$ |
| Parâmetro $\beta$ | $B \pm t_{N-2}(\gamma/2) \cdot S \cdot \sqrt{\frac{1}{S_{XX}}}$ |

Notas: $A$ e $B$ são os estimadores de $\alpha$ e $\beta$, respectivamente;

$S^2$ é um estimador de $\sigma^2$ e é dado por $S^2 = \frac{1}{N-2} \cdot \sum_n [Y_n - A - B \cdot (X_n - \bar{X})]^2$

$\sigma^2$ pode também ser estimado a partir do $DQMR$ da tabela ANOVA (ver detalhes mais abaixo): $DQMR = \frac{S_{YY} - B \cdot S_{XY}}{N-2}$

| **Testes de Hipóteses aos parâmetros $\alpha'$, $\alpha$ e $\beta$** | |
|---|---|
| **Hipóteses** | **Estatística de Teste** |
| $H_0: \alpha = \alpha_0$<br>$H_1: \alpha \neq \alpha_0, \alpha > \alpha_0$ ou $\alpha < \alpha_0$ | $ET = \dfrac{A - \alpha_0}{S/\sqrt{N}}$<br>Quando $H_0$ é verdadeira, $ET$ segue uma distribuição $t_{N-2}$ |
| **Hipóteses** | **Estatística de Teste** |
| $H_0: \alpha' = \alpha'_0$<br>$H_1: \alpha' \neq \alpha'_0, \alpha' > \alpha'_0$ ou $\alpha' < \alpha'_0$ | $ET = \dfrac{(A - \bar{X} \cdot B) - \alpha'_0}{S \cdot \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{S_{XX}}}}$<br>Quando $H_0$ é verdadeira, $ET$ segue uma distribuição $t_{N-2}$ |
| **Hipóteses** | **Estatística de Teste** |
| $H_0: \beta = \beta_0$<br>$H_1: \beta \neq \beta_0, \beta > \beta_0$ ou $\beta < \beta_0$ | $ET = \dfrac{B - \beta_0}{S/\sqrt{S_{XX}}}$<br>Quando $H_0$ é verdadeira, $ET$ segue uma distribuição $t_{N-2}$ |

| **Teste de Hipótese ao parâmetro $\beta$ baseado na tabela ANOVA (ver tabela ANOVA no final do formulário)** | |
|---|---|
| **Hipóteses** | **Estatística de Teste** |
| $H_0: \beta = \beta_0$<br>$H_1: \beta \neq \beta_0$ | $ET = \dfrac{DQMDR}{DQMR}$<br>Quando $H_0$ é verdadeira, $ET$ segue uma distribuição $F_{1,N-2}$ |

### Previsões com base no modelo de regressão linear simples

O valor de uma variável dependente, $Y$, pode ser obtida a partir do novo valor da variável $X$ dependente. Para isto recorre-se à seguinte expressão para a melhor previsão de $Y$:

$$\hat{Y} = \hat{\mu}_Y = A + B \cdot (X - \bar{X})$$

A variância do erro de previsão ($\delta$) é dada por:

$$Var(\delta) = \left[1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{S_{XX}}\right] \cdot \sigma^2$$

**Intervalos de Confiança a $(1-\gamma)\cdot 100\%$ para a previsão**

| | |
|---|---|
| Para valores de $Y$ | $\hat{Y} \pm t_{N-2}(\gamma/2)\cdot S\cdot\sqrt{1+\frac{1}{N}+\frac{(X-\bar{X})^2}{S_{XX}}}$ |
| Para valores de $\mu_Y$ | $\hat{Y} \pm t_{N-2}(\gamma/2)\cdot S\cdot\sqrt{\frac{1}{N}+\frac{(X-\bar{X})^2}{S_{XX}}}$ |

**Correlação entre Variáveis**

A análise da correlação entre variáveis é uma técnica muito semelhante à regressão no que se refere aos objectivos. No entanto, esta é menos potente, não indicando a forma como as variáveis estão relacionadas.

| | |
|---|---|
| Coeficiente de Correlação Amostral | $R_{XY}=\frac{S_{XY}}{\sqrt{S_{XX}}\cdot\sqrt{S_{YY}}}$ |
| Coeficiente de Correlação Populacional | $\rho_{XY}=\frac{E[(X-\mu_X)\cdot(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]\cdot E[(Y-\mu_Y)^2]}}$ |
| Coeficiente de Determinação | $R_{XY}^2=\frac{B^2\cdot S_{XX}}{S_{YY}}$ |

**Regressão Não Linear**

Os modelos não lineares que traduzem relações entre a variável dependente e uma só variável independente podem ser convertidos em modelos lineares por aplicação de transformações às variáveis dependente ou independente. Seguem-se exemplos de modelos onde a linearização é facilmente obtida:

| Modelo Original | Novas Variáveis | Modelo Linearizado |
|---|---|---|
| $Y_n=\alpha'+\frac{\beta}{X_n}+E_n$ | $U_n=\frac{1}{X_n}$ | $Y_n=\alpha'+\beta\cdot U_n+E_n$ |
| Modelo Exponencial<br>$Y_n=e^{\alpha'+\beta\cdot X_n+E_n}$ | $Z_n=\ln Y_n$ | $Z_n=\alpha'+\beta\cdot X_n+E_n$ |
| Modelo "Curva em S" (com $\alpha'>0$ e $\beta<0$)<br>$Y_n=e^{\alpha'+\beta/X_n+E_n}$ | $U_n=\frac{1}{X_n}$ $Z_n=\ln Y_n$ | $Z_n=\alpha'+\beta\cdot U_n+E_n$ |

**Tabela ANOVA para Teste de Hipótese ao parâmetro $\beta$ baseado na tabela ANOVA**

| **Fontes de Variação** | **Variações** *(Somas de desvios quadráticos)* | **Graus de Liberdade** *(Nº de termos independentes)* | **Desvios Quadráticos Médios (DQMs)** | **Valores Esperados** |
|---|---|---|---|---|
| **Devida à Regressão (DR)** | *(Variação "explicada" pela regressão)*<br>$\boldsymbol{VDR=B^2\cdot S_{XX}=B\cdot S_{XY}}$ | $\boldsymbol{GL_1=1}$ | $\boldsymbol{DQMDR=VDR}$ | $\boldsymbol{E[DQMDR]=\sigma^2+\beta^2\cdot S_{XX}}$ |
| **Residual (R)** | *(Variação residual "não explicada")*<br>$\boldsymbol{VR=S_{YY}-B\cdot S_{XY}}$ | $\boldsymbol{GL_2=N-2}$ | $\boldsymbol{DQMR=\frac{VR}{GL_2}}$ | $\boldsymbol{E[DQMR]=\sigma^2}$ |
| **Total (T)** | *(Variação total)*<br>$\boldsymbol{VT=S_{YY}}$ | $\boldsymbol{GL=GL_1+GL_2=N-1}$ | | |

Formulário adaptado de:
Estatística
Rui Campos Guimarães, José A. Sarsfield Cabral
Verlag Dashöfer