# Uma introdução ao modelo de regressão linear simples

Gustavo Valencio Tofolo

Instituto de Matemática e Estatística, Universidade de São Paulo

2025

# Sumário

# Curvas de Regressão

# Distribuições Condicionais Contínuas

Revisitarei algumas definições e interpretações geométricas mais simples, a começar pela **distribuição condicional contínua**.

## Definição

Sejam $X$ e $Y$ variáveis contínuas. A densidade condicional de $X$ dado que $Y = y$ é definida por:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \quad f_Y(y) > 0, \tag{1}$$

onde $f_Y(y)$ é a densidade marginal de $Y$.

Obs.: $f_{X|Y}(x|y)$ define uma densidade de probabilidade, para $y$ fixado.

# Distribuições Condicionais Contínuas: Exemplo

Tome a densidade condicional:

$$f_{X|Y}(x|y) = \frac{x+y}{4(2+y)}, \quad 0 \leq x \leq 4, \quad 0 \leq y \leq 4.$$

Graficamente, temos

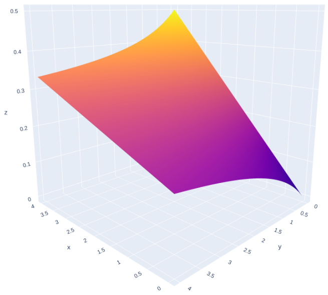


Figura 1: Representação gráfica da densidade condicional.

# Distribuições Condicionais Contínuas: Interpretação

Fixado $y = y_0$, temos que a intersecção deste plano com a $f_{X|Y}(x|y)$ define $f_{X|Y}(x|y_0)$. Se $y = 2$, $f_{X|Y}(x|2) = \frac{x+2}{16}$. Veja:

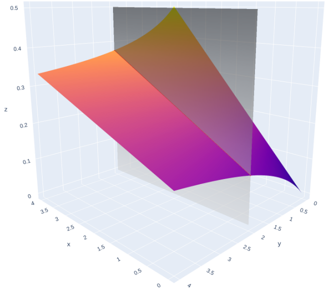



Figura 2: Intersecção entre $f_{X|Y}(x|y)$ e $y = y_0$.

Como isso define uma f.d.p., é razoável estudarmos $E[X|y]$.

# Curvas de Regressão: Intuição

Podemos traçar planos em $y = y_0,\ y_0 \in \{0, 0.25, 0.50, \ldots, 4\}$ e calcular o valor médio das distribuições condicionais resultantes das interseções. Se a diferença entre valores consecutivos de $y$ tende a zero, ou seja, $y_{i+1} - y_i \to 0$, teremos a curva de $E[X|y]$. Veja:

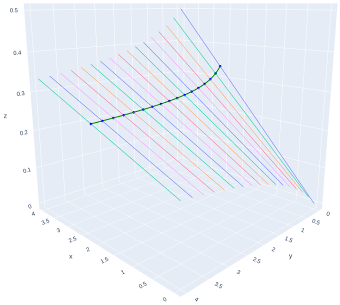



Figura 3: Intersecções, valores médios de $X|y$ e interpolação para $E[X|y]$.

Após apresentarmos a intuição, temos a seguinte definição.

## Definição

A esperança condicional de $X$, dado que $Y = y$, é definida como:

$$E[X|y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx \tag{2}$$

Note que $E[X|y]$ é uma função de $Y$, ou seja, $E[X|y] = s(y)$, sendo denominada **curva de regressão** de $X$ sobre $y$. O mesmo desenvolvimento pode ser feito para $Y|x$.

# Modelo com Efeitos Fixos

# Motivação

O estatístico busca criar modelos que revelem estruturas do fenômeno observado, reduzindo a incerteza associada. Uma abordagem comum é assumir que cada observação segue a relação:

$$\text{observação} = \text{previsível} + \text{aleatório} \tag{3}$$

A primeira componente reflete o conhecimento do pesquisador e é expressa por uma função matemática com parâmetros desconhecidos. A segunda representa o erro aleatório, que, devido a suposições, também obedece a alguma função com parâmetros desconhecidos. Cabe ao estatístico estimar esses parâmetros com base em amostras.

# Modelo com Efeitos Fixos

Dada uma população $P$ e uma variável de interesse $Y$, podemos dividi-la em $I$ subpopulações $P_1, P_2, ..., P_I$, onde $i = 1, 2, ..., I$ é um nível de fator.

Suponha que $E[Y] = \mu$ e que $E[Y|P_i] = \mu_i$, $i = 1, ..., I$. O objetivo é estimar $\mu_i$, $i = 1, ..., I$ e testar hipóteses como:

$$H_0 : \mu_1 = \mu_2 = \ ... \ = \mu_I = \mu \tag{4}$$

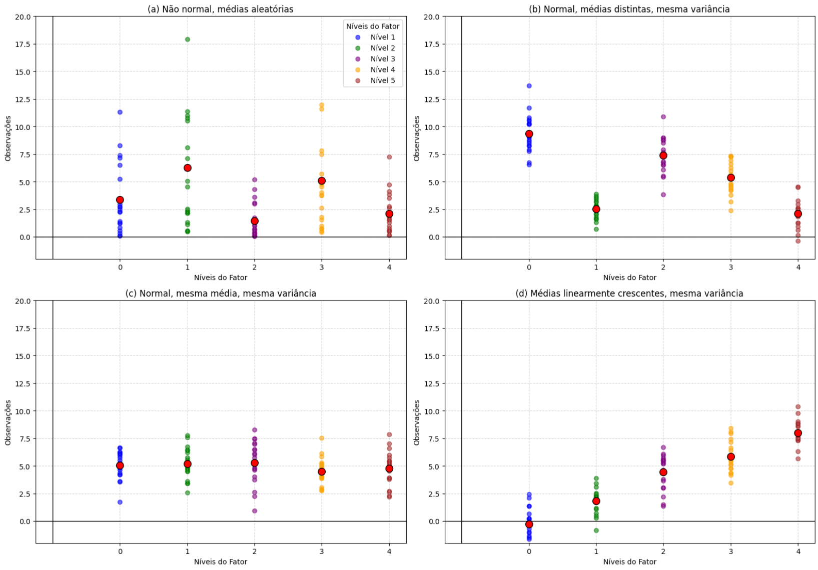
$$H_1 : \mu_i \neq \mu_j, \text{ para algum par } (i, j) \tag{5}$$

Um modelo conveniente para descrever essa situação é:

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, ..., I, \quad j = i, ..., n_i \tag{6}$$

onde $n_i$ é o tamanho da amostra desta subpopulação. Supomos também que $e_{ij} \sim N(0, \sigma_e^2)$.

# Modelos com Efeitos Fixos: Gráficos



# Modelos com Efeitos Fixos: Comentários

A Figura 4(a) apresenta um comportamento mais amplo, com distribuições distintas. Nas demais figuras, assume-se a hipótese mais comum de normalidade com variância constante. A Figura 4(b) ilustra a hipótese alternativa (5), enquanto a Figura 4(c) representa a hipótese nula (4). Por fim, a última figura mostra um caso em que as médias se comportam de forma linear.

Os pontos vermelhos representam as médias das subpopulações. Como estamos propondo um modelo para essas médias, a notação apropriada para o modelo (6), chamado de **modelo com efeitos fixos**, é a seguinte:

$$y_i = \mu_i + e_i, \quad i = 1, ..., I \tag{7}$$

# Regressão Linear Simples

# Motivação

Anteriormente, vimos que $E[Y|x] = \mu(x)$. Da Figura 4(d), observamos que as médias de $Y$ aumentam conforme os níveis de fator aumentam. Se, ao invés de níveis, tivermos valores contínuos, seria razoável propormos:

$$E[Y|x] = \mu(x) = \alpha + \beta x \tag{8}$$

tal que

$$y_i = E[Y|x_i] + e_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, ..., n \tag{9}$$

Devemos encontrar os valores mais prováveis para $\alpha$, $\beta$ e $\sigma_e^2$, seguindo algum critério, dada uma amostra observada.

# Estimação dos Parâmetros: Suposições

Suposições:

- Os valores de $X$ são conhecidos. Isso é viável: $E[xY] = xE[Y]$.
- Os erros se distribuem ao redor da média com média zero. Se não, estaríamos consistentemente subestimando ou superestimando $Y$.

$$E[e_i|x] = 0 \tag{10}$$

- A variância dos erros é constante para todo $x$.

$$Var[e_i|x] = \sigma_e^2 \tag{11}$$

- O erro associado a uma observação não influencia o erro de outra observação.

# Estimação dos Parâmetros: Mínimos Quadrados

Colhidos $n$ pares $(x_i, y_i),\ i = 1, 2, \ldots, n$ que satisfazem (9), temos

$$e_i = y_i - (\alpha + \beta x_i), \quad i = 1, 2, \ldots, n \tag{12}$$

Disso, obtemos a quantidade de informação perdida pelo modelo:

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 \tag{13}$$

A solução de mínimos quadrados é a que torna essa soma mínima. Derivando em relação aos coeficientes e igualando a zero, temos:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{14}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{15}$$

# Estimação dos Parâmetros: Exemplo

De uma amostra aleatória, obtemos a seguinte reta estimada:

$$\hat{y}_i = 3.32 + 1.43x_i$$

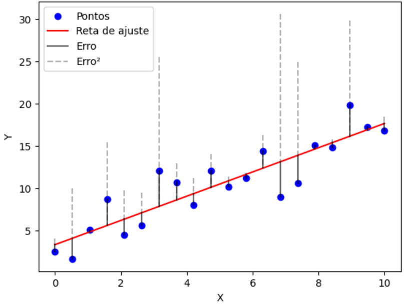



Figura 5: Nuvem de pontos, reta ajustada e mensuração dos erros.

```python
import pandas as pd
import statsmodels.formula.api as smf

model = smf.ols("Y ~ X", data=df).fit()
print(model.params)
```

Figura 6: Código exemplo para estimação dos parâmetros.

# Avaliação do Modelo: estimador de $\sigma_e^2$

Iremos avaliar a adoção do modelo linear simples comparando-o com o proposto em (7), no cenário da Figura 4(c). As somas dos resíduos quadráticos dos modelos são:

$$SQTot = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{16}$$

$$SQRes = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{17}$$

Em cada cenário, temos os seguintes estimadores não-viciados para $\sigma_e^2$:

$$S^2 = \frac{SQTot}{n-1} \tag{18}$$

$$S_e^2 = \frac{SQRes}{n-2} \tag{19}$$

# Avaliação do Modelo: Decomposição de SQ

Observando a Figura 7, temos que $y_i - \bar{y} = e_i + (\hat{y}_i - \bar{y})$. Podemos elevar estes membros ao quadrado e obter:

$$SQTot = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + SQRes, \quad \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SQReg \tag{20}$$

É possível derivar que:

$$SQReg = \hat{\beta}^2 \sum_{i=1}^{n} (\hat{x}_i - \bar{x})^2 \tag{21}$$

Note que quanto maior $\hat{\beta}$, mais nos afastamos do modelo (7).

# Avaliação do Modelo: Figura

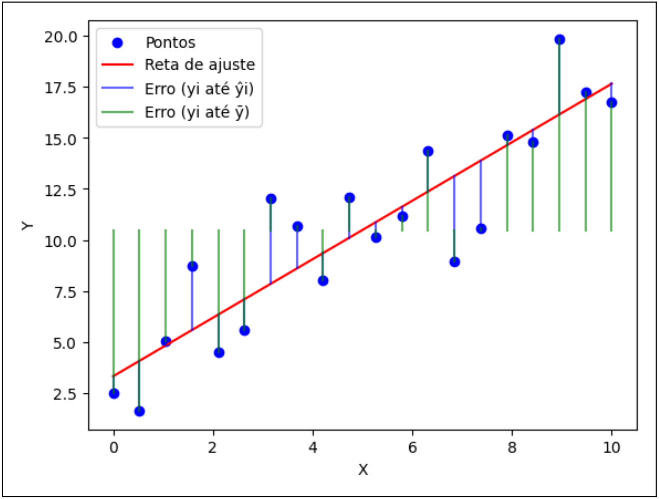



Figura 7: Decomposição das variações do modelo.

# Avaliação do Modelo: Interpretação

Podemos nos aprofundar um pouco mais na interpretação das métricas antes apresentadas:

- SQResíduos: mede as variações que ainda existem nos dados após o ajuste do modelo. Ou seja, o que **não é** explicado pelo modelo.
- SQRegressão: mede as variações que existem **por conta do modelo**. Ou seja, como a variável explicativa interfere no valor esperado da variável resposta (o que é explicada pelo modelo).
- SQTotal: é a soma total das variações observadas nos dados em relação à média. É também a soma dos quadrados do modelo simplificado.

Podemos utilizar $SQReg = SQTot - SQRes$ e $R^2 = \frac{SQReq}{SQTot}$ para realizar comparativos entre as propostas.

# Avaliação do Modelo: Tabela ANOVA

Podemos resumir essas informações numa tabela como a que segue:

| F.V. | g.l. | SQ | QM | F |
|---|---|---|---|---|
| Regressão | 1 | SQReg | SQReg $=$ QMReg | QMReg$/S_2^2$ |
| Resíduo | $n-2$ | SQRes | SQRes$/(n-2) = S_e^2$ | |
| Total | $n-1$ | SQTot | SQTot$/(n-1) = S^2$ | |

Consta a fonte de variação, a soma dos quadrados e os graus de liberdade associados ao estimador. Também, a estatística F, que virá em breve.

```
# Gerando a tabela ANOVA
anova_table = sm.stats.anova_lm(model, typ=1)
print(anova_table)
```

```
            df  sum_sq     mean_sq         F    PR(>F)
X          1.0   810.0  810.000000  25.89698  0.000077
Residual  18.0   563.0   31.277778       NaN       NaN
```

Figura 8: Código exemplo para tabela ANOVA.

# Propriedades dos Estimadores: Média e Variância

Para o estimador $\hat{\beta}$:

$$E[\hat{\beta}] = \beta \tag{22}$$

$$\text{Var}(\hat{\beta}) = \frac{\sigma_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{23}$$

Para o estimador $\hat{\alpha}$:

$$E[\hat{\alpha}] = \alpha \tag{24}$$

$$\text{Var}(\hat{\alpha}) = \sigma_e^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{25}$$

# Propriedade dos Estimadores: Distribuições Amostrais

Temos que $y_i \sim N(\alpha + \beta x_i;\ \sigma_e^2)$ e $e_i \sim N(0, \sigma_e^2)$. Como $\hat{\alpha}$ e $\hat{\beta}$ são combinações lineares destas (independentes), então:

$$\hat{\alpha} \sim N\left(\alpha; \frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right) \tag{26}$$

$$\hat{\beta} \sim N\left(\beta; \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}\right) \tag{27}$$

Podemos normalizar variáveis aleatórias:

$$\frac{\hat{\beta} - \beta}{\sigma_e} \sqrt{\sum (x_i - \bar{x})^2} \sim N(0, 1) \tag{28}$$

$$\frac{\hat{\alpha} - \alpha}{\sigma_e} \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}} \sim N(0, 1) \tag{29}$$

# Propriedades dos Estimadores: Intervalo de Confiança

Se $W = \frac{Z}{\sqrt{Y/v}}$, $Z \sim N(0,1)$ e $Y \sim \chi^2_{(v)}$, então $W \sim t_{(v)}$. Disso, derivamos

$$t(\hat{\beta}) = \frac{\hat{\beta} - \beta}{S_e} \sqrt{\sum (x_i - \bar{x})^2} \sim t_{(n-2)} \tag{30}$$

$$t(\hat{\alpha}) = \frac{\hat{\alpha} - \alpha}{S_e} \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}} \sim t_{(n-2)} \tag{31}$$

Sabe-se que existe $\gamma$ tal que $P(-t_{\gamma(n-2)} < t(\hat{\alpha}) < t_{\gamma(n-2)}) = \gamma$. Expandindo (similarmente para $\hat{\beta}$), obtemos

$$\text{IC}(\alpha; \gamma) = \hat{\alpha} \pm t_{\gamma(n-2)} S_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}} \tag{32}$$

$$\text{IC}(\beta; \gamma) = \hat{\beta} \pm t_{\gamma(n-2)} S_e \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}} \tag{33}$$

# Propriedades dos Estimadores: IC em Python

Podemos obter o intervalo de confiança pros coeficientes da reta via:

```
intervalos = model.conf_int(alpha=0.05)

# Nomeando as colunas para facilitar a leitura
intervalos.columns = ["Limite Inferior", "Limite Superior"]

# Adicionando os nomes dos coeficientes
intervalos.index = ["Intercepto", "Coeficiente de X"]

print(intervalos)
```

```
                  Limite Inferior  Limite Superior
Intercepto              69.047780        91.952220
Coeficiente de X         0.528441         1.271559
```

Figura 9: Código exemplo para construir IC.

# Propriedades dos Estimadores: Sumário em Python

Podemos obter um resumo do modelo via:

```
model = smf.ols("Y ~ X", data=df).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.590
Model:                            OLS   Adj. R-squared:                  0.567
Method:                 Least Squares   F-statistic:                     25.90
Date:                Tue, 11 Mar 2025   Prob (F-statistic):           7.66e-05
Time:                        10:09:04   Log-Likelihood:                -61.754
No. Observations:                  20   AIC:                             127.5
Df Residuals:                      18   BIC:                             129.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     80.5000      5.451     14.768      0.000      69.048      91.952
X              0.9000      0.177      5.089      0.000       0.528       1.272
==============================================================================
Omnibus:                        1.840   Durbin-Watson:                   2.565
Prob(Omnibus):                  0.399   Jarque-Bera (JB):                1.395
Skew:                           0.459   Prob(JB):                        0.498
Kurtosis:                       2.088   Cond. No.                         134.
==============================================================================
```

Figura 10: Código exemplo para sumarizar modelo.

Hipóteses nulas comuns:

$$H_0 : \alpha = \alpha_0 \tag{34}$$

$$H_0 : \beta = \beta_0 \tag{35}$$

Em $P\left(t(\hat{\beta}) \in \mathrm{RC} | \beta = 0\right)$ e $P\left(t(\hat{\alpha}) \in \mathrm{RC} | \alpha = 0\right)$, são estatísticas do teste:

$$t(\hat{\alpha}) = \frac{\hat{\alpha}}{S_e}\sqrt{\frac{n\sum(x_i - \bar{x})^2}{\sum x_i^2}} \tag{36}$$

$$t(\hat{\beta}) = \frac{\hat{\beta}}{S_e}\sqrt{\sum(x_i - \bar{x})^2} \tag{37}$$

E da tabela ANOVA, é possível usar a $F$ para testar (35). Segue:

$$[t(\hat{\beta})]^2 = \frac{\hat{\beta}^2 \sum(x_i - \bar{x})^2}{S_e^2} = \frac{SQReg}{S_e^2} \sim F_{(1,n-2)} \tag{38}$$

# Análise de Resíduos

Útil para investigar se as suposições feitas são satisfeitas. Estenderemos os níveis de fator da Figura 4 para 30 (com 10 obs.). São pares interessantes:

- $(x_i, \hat{e}_i)$, onde $\hat{e}_i = y_i - \hat{y}_i$
- $(x_i, \hat{z}_i)$, onde $\hat{z}_i = \frac{\hat{e}_i}{S_e}$
- $(x_i, \hat{r}_i)$, onde $\hat{r}_i = \frac{\hat{e}_i}{S_e\sqrt{1-v_{ii}}}$

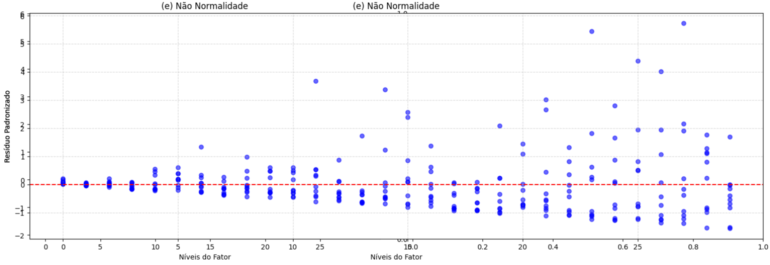



Figura 11: Diagnóstico não-normal.

# Análise de Resíduos: Outros Cenários

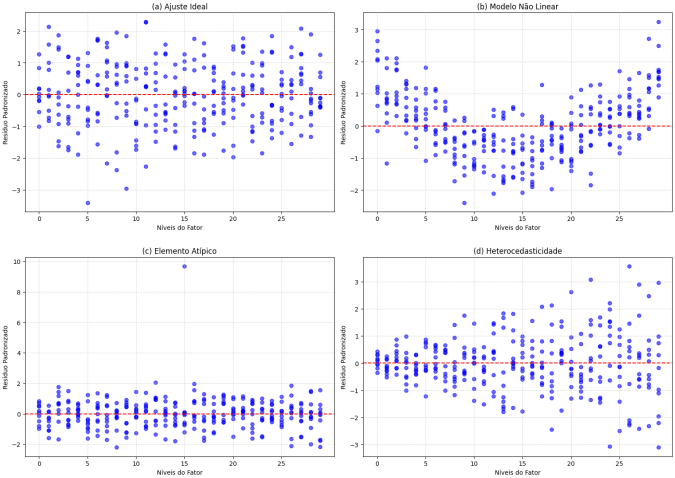



Figura 12: Demais diagnósticos.

# Bibliografia e Contato

# Bibliografia


Bussab, W. O., & Morettin, P. A. *Estatística Básica*. 6ª edição. Saraiva, 2009.

Statsmodels. *Statsmodels Documentation*. Disponível em `https://www.statsmodels.org/stable/index.html`. Acesso em: 11 mar. 2025.

# Contato

Para mais informações, entre em contato:

- E-mail: valenciotofolo@usp.br
- Site: https://linktr.ee/gustatofolo

O material apresentado será disponibilizado em um repositório no GitHub, contendo os códigos utilizados para gerar as imagens e tabelas numeradas conforme mostrado neste trabalho.