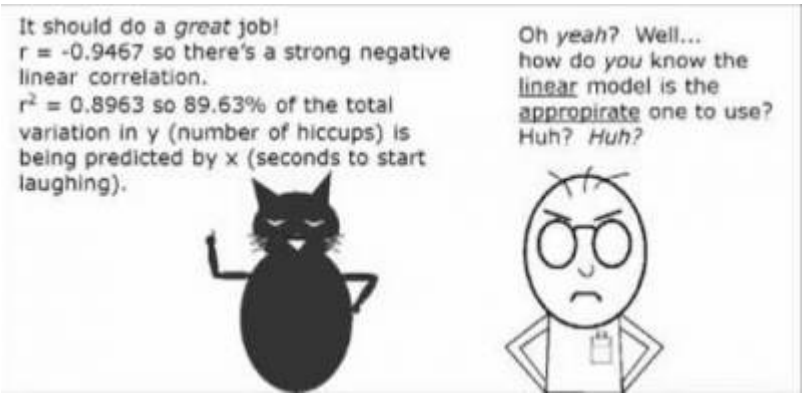
# Modelos Lineares




Os modelos lineares são uma generalização dos testes de hipótese clássicos mais simples. Uma regressão linear, por exemplo, só pode ser aplicada para dados em que tanto a variável preditora quanto a resposta são contínuas, enquanto uma análise de variância é utilizada quando a variável preditora é categórica. Os modelos lineares não têm essa limitação, podemos usar variáveis contínuas ou categóricas indistintamente.

Video

**ERRATA**: por volta de 16'28“ digo que o valor da inclinação na população é 3,5 quando o correto é 2,5

No nosso quadro de testes clássicos frequentistas, definimos os testes, baseados na natureza das variáveis respostas e preditoras.

| Tipo de Variável | | Estatística Clássica | |
|---|---|---|---|
| Resposta | Preditora | Teste | Hipótese |
| Categórica | Categórica | Qui-quadrado | independência |
| Contínua | Categórica (2 niveis) | Teste t | $\mu_1=\mu_2$ |
| Contínua | Categórica | Anova | $\mu_1=\mu_2=\mu_n$ |
| Contínua | 1 Contínua | Regressão | $\beta_1=0$ |
| Contínua | >1 Contínua | Reg. múltipla | $\beta_1=0;\beta_n=0$ |
| Contínua | Cont + Categ | Ancova | $\beta_1=\beta_2;\alpha_1=\alpha 2$ |
| Proporção | Contínua | Reg. Logística | $logit(\beta_1)=1$ |

Os modelos lineares dão conta de todos os testes apresentados na tabela acima que tenham a **variável resposta contínua**. Portanto, já não há mais necessidade de decorar os nomes: *teste-t*, *Anova*, *Anova Fatorial*, *Regressão Simples*, *Regressão Múltipla*, *Ancova* entre muitos outros nomes de testes que foram incorporados nos modelos lineares. Isso não livra o bom usuário de estatística de entender a natureza das variáveis que está utilizando. Isso continua sendo imprescindível para tomar boas decisões ao longo do processo de análise e interpretação dos dados.

## Simulando Dados

Vamos começar com um exemplo simples de regressão, mas de forma diferente da usual. Vamos usar a engenharia reversa para entender bem o que os modelos estatísticos estão nos dizendo e como interpretar os resultados produzidos. Para isso vamos inicialmente gerar dados fictícios. Esses dados terão dois componentes: uma estrutura determinística e outra aleatória. A primeira está relacionada ao processo de interesse e relaciona a variável resposta à preditora. No caso, essa estrutura é linear e tem a seguinte forma:

$$ y = {\alpha} + {\beta} x$$

O componente aleatório é expresso por uma variável probabilística Gaussiana da seguinte forma:

$$ \epsilon = N(0, \sigma) $$

Portanto, nossos dados serão uma amostra de uma população com a seguinte estrutura:

$$ y = {\alpha} + {\beta} x + \epsilon$$

Parece complicado, mas é razoavelmente simples gerar dados aleatórios em nosso computador baseado nessa estrutura. Para isso, abra uma planilha eletrônica e siga os passos descritos abaixo:

- nomeie a coluna ***A*** como **x** na célula A1;
- preencha as células A2:A16 com uma sequência de valores de 0.5 a 7.5,

em intervalos de 0.5

| | A | B | C | D |
|---|---|---|---|---|
| 1 | x | y0 | desvio | y1 |
| 2 | 0.5 | | | |
| 3 | 1 | | | |
| 4 | 1.5 | | | |
| 5 | 2 | | | |
| 6 | 2.5 | | | |
| 7 | 3 | | | |
| 8 | 3.5 | | | |
| 9 | 4 | | | |
| 10 | 4.5 | | | |
| 11 | 5 | | | |
| 12 | 5.5 | | | |
| 13 | 6 | | | |
| 14 | 6.5 | | | |
| 15 | 7 | | | |
| 16 | 7.5 | | | |
| 17 | | | | |
| 18 | | | | |

- nomeie a coluna ***B*** como **y0** na célula B1;
- preencha a célula B2 com a fórmula ***= 4 + 3.5 * A2***
- copie a formula para as células B3:B16, clicando e arrastando o mouse quando aparecer no canto inferior esquerdo da célula B2 o sinal de **+**.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | x | y0 | desvio | y1 |
| 2 | 0.5 | = 4 + 3.5*A2 | | |
| 3 | 1 | | | |
| 4 | 1.5 | | | |
| 5 | 2 | | | |
| 6 | 2.5 | | | |
| 7 | 3 | | | |
| 8 | 3.5 | | | |
| 9 | 4 | | | |
| 10 | 4.5 | | | |
| 11 | 5 | | | |
| 12 | 5.5 | | | |
| 13 | 6 | | | |
| 14 | 6.5 | | | |
| 15 | 7 | | | |
| 16 | 7.5 | | | |
| 17 | | | | |
| 18 | | | | |

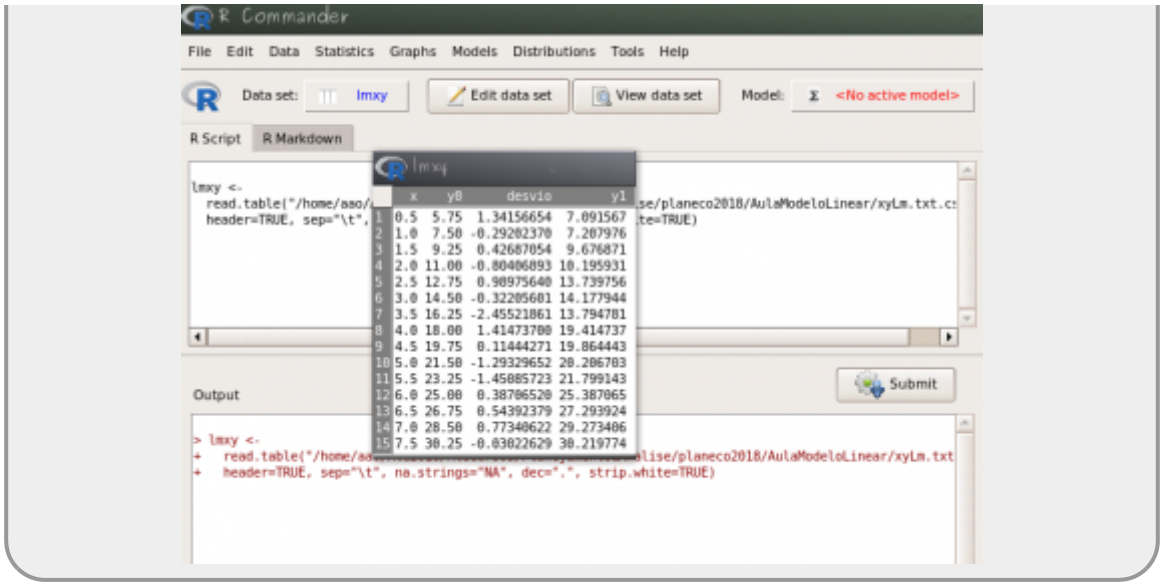
- nomeie a coluna ***C*** como **desvio** na célula C1;
- preencha a célula **C2** com a fórmula **= *INV.NORM.N(ALEATÓRIO(); 0 ; 2)*** [1]. **Essa fórmula vai retornar valores aleatórios tomados de uma distribuição normal com média 0 e desvio padrão 2**;
- copie a formula para as células C3:C16, clicando e arrastando o mouse quando aparecer no canto inferior esquerdo da célula ***B2*** o sinal de **+**.
- nomeie a coluna ***D*** como **y1** na célula D1;
- A variável **y1** na coluna ***D*** é a soma do valor da coluna ***B*** com o valor da coluna ***C*** (y0+ desvio). Para fazer isso, coloque na célula D2 a função ***=soma(B2:C2)***, depois copie para as outras células da coluna
- salve a planilha como texto separado por vírgulas e use o nome “xy.csv”

C2 | Σ = | =INV.NORM.N(ALEATÓRIO(),0,2)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | x | y0 | desvio | y1 | | |
| 2 | 0.5 | 5.75 | -3.058380577 | | | |
| 3 | 1 | 7.5 | 2.224347441 | | | |
| 4 | 1.5 | 9.25 | 2.159720971 | | | |
| 5 | 2 | 11 | 0.278286215 | | | |
| 6 | 2.5 | 12.75 | 3.128622272 | | | |
| 7 | 3 | 14.5 | 2.478190576 | | | |
| 8 | 3.5 | 16.25 | -0.743526151 | | | |
| 9 | 4 | 18 | -2.095088544 | | | |
| 10 | 4.5 | 19.75 | 0.426249317 | | | |
| 11 | 5 | 21.5 | 2.88420496 | | | |
| 12 | 5.5 | 23.25 | 1.145051653 | | | |
| 13 | 6 | 25 | 0.283340336 | | | |
| 14 | 6.5 | 26.75 | 0.842373056 | | | |
| 15 | 7 | 28.5 | -0.067742821 | | | |
| 16 | 7.5 | 30.25 | 0.509119996 | | | |
| 17 | | | | | | |

A função INV.NORM.N() tem três parâmetros, (1) probabilidade, (2) média e (3) desvios padrão. Ao definir o terceiro parâmetro, estamos amostrando valores de uma distribuição normal com desvio padrão igual a 2.

- importe os dados da planilha para o Rcommander (lembrando de selecionar como separador a vírgula) e use o nome **xy** ;
- garanta que os dados foram lidos corretamente, clicando em *View data set*




```
lmdummy <- lm(colhe ~ dummy1 + dummy2 + dummy3 , data = colheitaDummy)
## avalie o modelo
summary(lmdummy)
anova(lmdummy)
```

- ajuste o modelo normal de anova

```
lmAnova <- lm(colhe~solo, data=colheita)

## avalie o modelo
summary(lmAnova)
anova(lmAnova)
```

- compare os coeficientes dos dois modelos

1)

Em versões mais antigas do Excel, essa função tinha o nome de *INV.NORM* e para computadores em inglês use a função no seguinte formato: *=NORM.INV(RAND(); 0; 2)*, no calc do LibreOffice use *=NORMINV(RAND(),0,2))*.