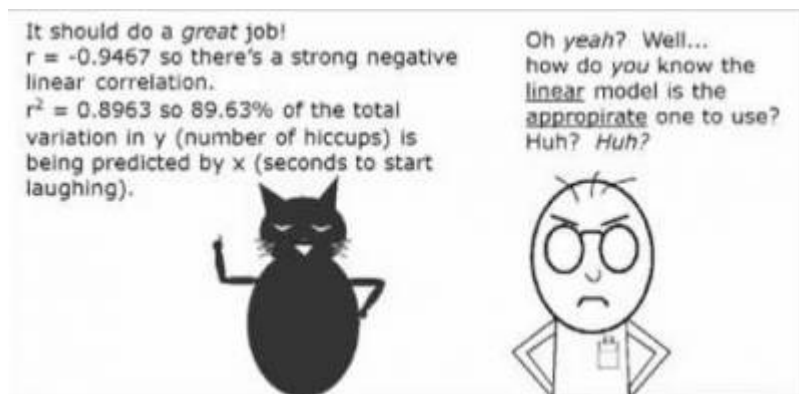


# Modelos Lineares



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 as variáveis são contínuas. Os modelos lineares não tem essa limitação, podemos usar variáveis contínuas ou categóricas indistintamente.

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 níveis)	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	$\text{logit}(\beta_1) = 1$

No caso, os modelos lineares dão conta de todos os testes nessa tabela 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 trabalhando. Isso continua sendo imprescindível para tomar boas decisões ao longo do processo de análise e interpretação destes 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. Iremos inicialmente criar nossos dados. 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	<u>desvio</u>	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 fórmula 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 células **C2** com a fórmula = **INV.NORM.N(ALEATÓRIO()); 0 ; 4)** (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; 4)**). Essa fórmula vai retornar valores aleatórios tomados de uma distribuição normal com média 0 e desvio padrão 4;
- 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** ( $y_0 + \text{desvio}$ ). 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			$\Sigma$	=	=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 4.

- importe os dados da planilha para o Rcommander e use o nome **xy** ;
- garanta que os dados foram lidos corretamente, clicando em **View data set**

The screenshot shows the R Commander interface. The 'Data set' dropdown is set to 'lmxy'. The 'View data set' button is highlighted. A preview window for 'lmxy' is open, displaying the following data:

	x	y0	desvio	y1
1	0.5	5.75	1.34156654	7.091567
2	1.0	7.50	-0.29202370	7.207976
3	1.5	9.25	0.42687054	9.676871
4	2.0	11.00	-0.80406893	10.195931
5	2.5	12.75	0.90975640	13.739756
6	3.0	14.50	-0.32205601	14.177944
7	3.5	16.25	-2.45521861	13.794781
8	4.0	18.00	1.41473700	19.414737
9	4.5	19.75	0.11444271	19.864443
10	5.0	21.50	-1.29329652	20.206703
11	5.5	23.25	-1.45085723	21.799143
12	6.0	25.00	0.38706520	25.387065
13	6.5	26.75	0.54392379	27.293924
14	7.0	28.50	0.77340622	29.273406
15	7.5	30.25	-0.03022629	30.219774

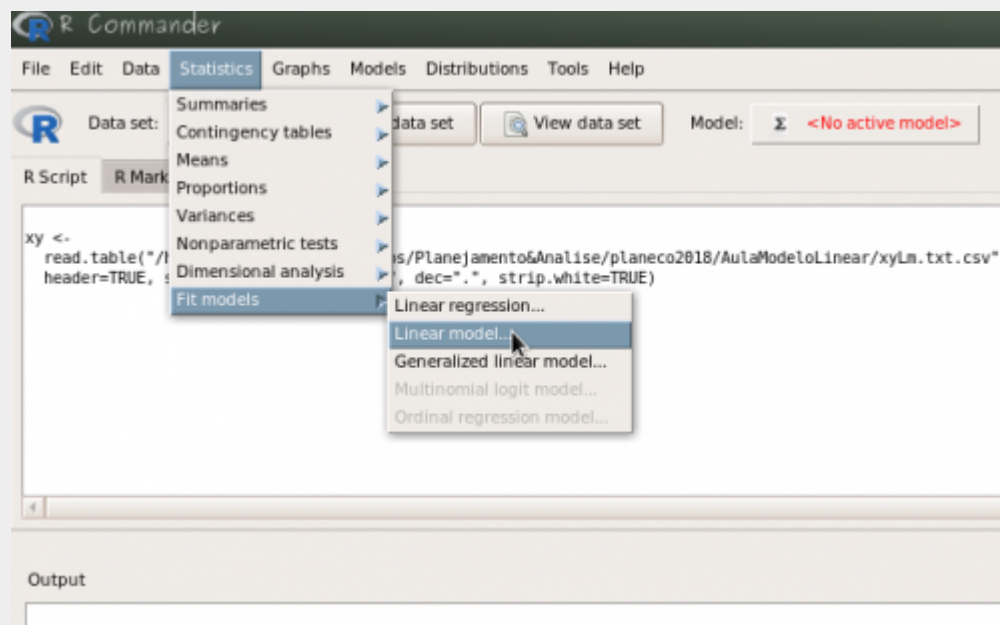
The R Script window shows the following code:

```
lmxy <- read.table("/home/aso/.../planeco2018/AulaModeloLinear/xyLm.txt", header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
```

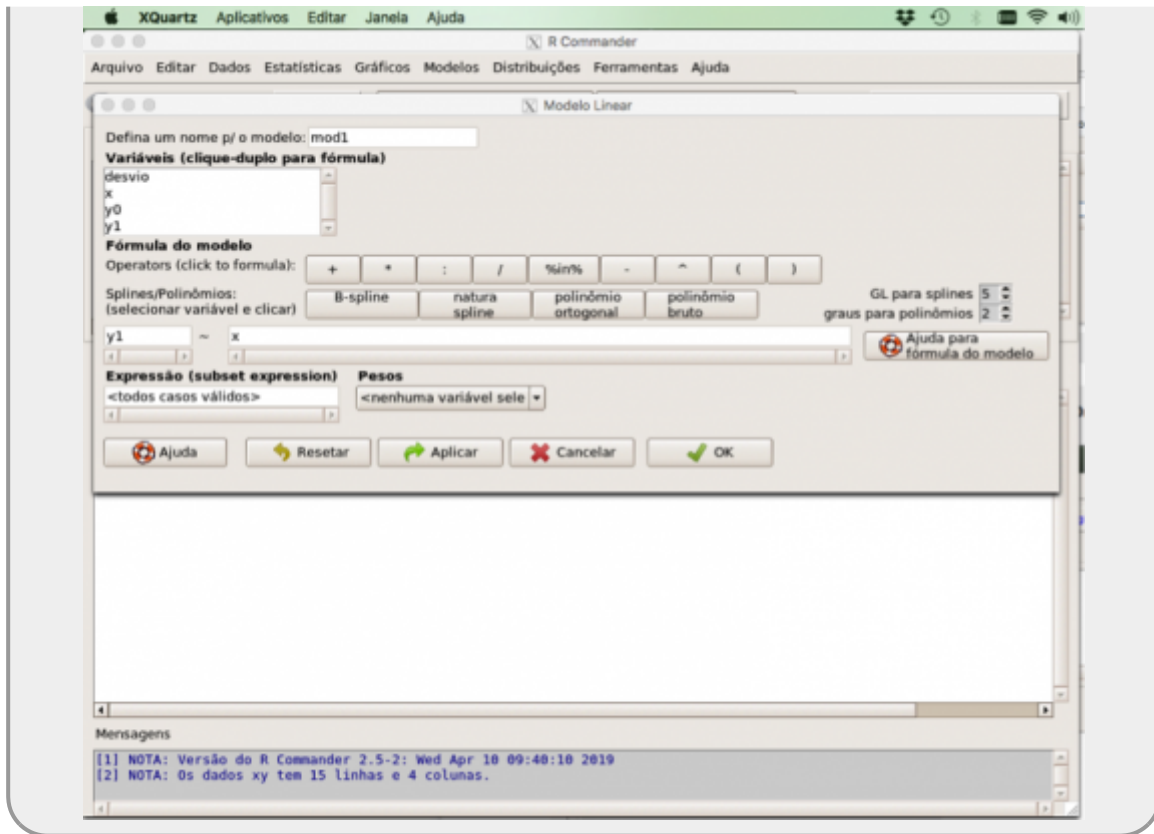
# Modelo Linear Simples

## Criando o modelo no Rcmdr

Abra o menu **Statistics > Fit Models > Linear Models...**



- Defina o nome desse modelo como **mod1**
- A fórmula do modelo tem duas caixas. Na caixa da esquerda (antes do símbolo ~) você deve colocar a variável resposta, que nesse caso é a nossa variável **y1**.
- Na caixa da direita (após o ~) coloque a variável preditora, que nesse caso é a variável **x**



- interprete o resultado do ajuste. Onde está o valor da inclinação da reta ajustada?
- copie o resultado do **summary** do modelo que aparece na janela **Output**

The screenshot shows the R Commander window with the following content:

**R Script**

```
xy <- read.table("/Users/Dri/Documents/00BIE5793_PLANECO/BIE5793_PLANECO_2019/ROTEIROS_2019/Modelos linear
header=TRUE, sep=".", na.strings="NA", dec=".", strip.white=TRUE)
mod1 <- lm(y1 ~ x, data=xy)
summary(mod1)
```

**Output**

```
> summary(mod1)

Call:
lm(formula = y1 ~ x, data = xy)

Residuals:
    Min       10   Median       30      Max
-3.8805 -2.1815 -0.6798  1.7986  6.5902


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.388      1.541   2.199   0.0466 *
x             4.849      0.339  14.304 0.0000000248 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.836 on 13 degrees of freedom
Multiple R-squared:  0.9483, Adjusted R-squared:  0.9357
F-statistic: 204.6 on 1 and 13 DF, p-value: 0.00000002476
```

**Mensagens**

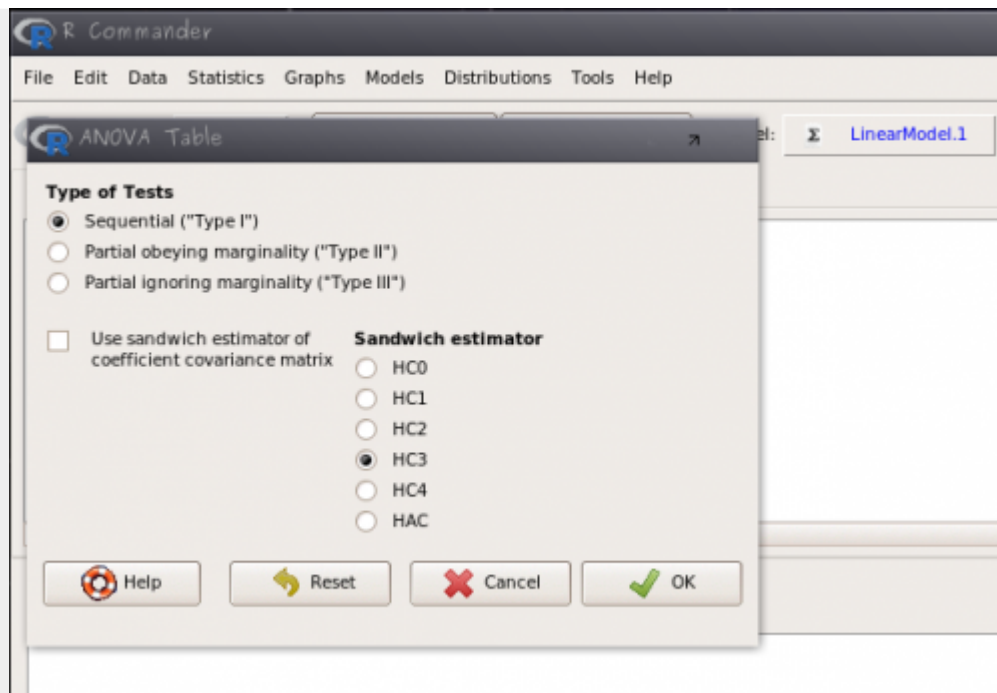
```
[1] NOTA: Versão do R Commander 2.5-2: Wed Apr 10 09:40:10 2019
[2] NOTA: Os dados xy tem 15 linhas e 4 colunas.
```

## Resultados do Modelo Linear Simples

 Anote os valores do resultado da análise na planilha [resultados do modelo linear](#)

## Tabela de Anova do LM

- confira se o modelo ativo é o que montou anteriormente (**mod1**). Isso deve ser checado na caixa “Model” que fica no canto superior direito da tela do Rcommander.
- vá ao menu **Models > Hypothesis Test > ANOVA table...**
- marque a opção: **Sequential (“Type I”)**



- copie o resultado da tabela de ANOVA
- interprete o resultado da tabela

## Comparando com Modelo Nulo

- monte um novo modelo, chamado “mod0” ( **Statistics > Fit Models > Linear Models** )
- como variável resposta use y1
- no lugar da preditora coloque o valor 1
- interprete o resultado desse modelo
- compare o **mod0** com o **mod1** ( **Models > Hypothesis Test > Compare two models...** )
- compare esse resultado com a tabela de ANOVA do modelo **mod1**

## Resíduo do Modelo

- simule um novo conjunto de dados usando os mesmo passos anteriores, mudando apenas o comando `INV.NORM.N(ALEATÓRIO(); 0 ; 4)` para `INV.NORM.N(ALEATÓRIO(); 0 ; 2)`
- refaça o modelo com os novos dados e compare o resultado com o modelo anterior



## Múltiplos experimentos

A base da estatística frequentista é que uma amostra e seus resultados são uma realização dos possíveis resultados provenientes de uma população ideal, na qual não temos acesso. Utilizando os resultados de outros alunos da tabela [resultados do modelo linear](#) e o que aprendeu no exercício, responda:

- 1. Qual o parâmetro da população referentes a relação entre  $x$  e  $y$ ?
- 2. Porque as estimativas variam entre os modelos ajustados se todos seguiram o mesmo roteiro?
- 3. Quais os valores médios estimados de alfa (intercepto), beta (inclinação) e desvio padrão das estimativas de todos os alunos? O valor médio é mais ou menos parecido com o parâmetro populacional? Explique.
- 4. Qual a fonte principal das variações encontradas nas estimativas dos parâmetros e seu paralelo em um experimento real?

## Tabela de Anova de uma Regressão



O nosso próximo exercício usa os dados de crescimento de lagartas submetidas a dietas de folhas com diferentes concentrações de taninos. São apenas duas variáveis, **growth**, o crescimento da lagarta, e **tannins**, a concentração de taninos. O objetivo é verificar se há relação entre o crescimento da lagarta e a concentração de taninos da dieta.

## Desvio Quadrático Total

- baixe o arquivo `regression.txt`;
- abra o arquivo no Excel;
- calcule a média de crescimento dos dados;
- calcule o valor de desvio total dos dados (o crescimento observado menos a média do crescimento);
- calcule o desvio quadrático total;

## Estimação dos Parâmetros e Resíduos

- calcule o intercepto e a inclinação do modelo linear no próprio excel, usando as funções descritas no quadro abaixo;

Para o cálculo dos parâmetros da reta use as funções do Excel:



- **INCLINAÇÃO**
- **INTERCEPÇÃO**

- a partir da inclinação e do intercepto estimado, calcule o valor predito pelo modelo em uma coluna chamada **predMod**
- crie uma outra coluna (**resdMod**) com os valores de resíduos do modelo para cada observação (observado menos o predito pelo modelo);
- calcule o desvio quadrático do resíduo para cada observação;
- some os desvios quadráticos dos resíduos;


## Tabela de Anova de um Modelo Linear

- monte uma tabela de ANOVA com as somas quadráticas como no [tutorial de anova](#);

### Equações

#### Somas Quadráticas

$$SS_{TOTAL} = \sum_{i=1}^n (y_i - \bar{y})^2$$


$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SS_{TOTAL} = SS_{regr} + SS_{res}$$

$$\bar{y} = \text{média da variável resposta}$$

$$\hat{y}_i = \text{valor estimado pelo modelo para } x_i$$

- Calcule o p-valor associado à estatística F do modelo

Utilize no excel o valor **1- DIST.F(F, df1, df2, VERDADEIRO)** para o cálculo do p-valor sendo F o valor da estatística F calculada, df1 o grau de liberdade da regressão (normalmente 1) e df2 o valor de graus de liberdade da desvios quadráticos médios dos resíduos.

- calcule o  $r^2$  (coeficiente de determinação) da regressão <sup>1)</sup>;

$$R^2 = \frac{SS_{\text{regr}}}{SS_{\text{TOTAL}}}$$

- entre os dados no Rcmdr e faça um modelo linear do crescimento em função da concentração de taninos;
- faça o teste de hipótese por ANOVA do modelo gerado;
- compare o resultado obtido na planilha com a ANOVA do modelo linear do Rcmdr;

## Variáveis Dummies

- baixe o arquivo

colheita.csv

- abra no excel
- note que a variável solo tem agora 4 níveis: arenoso, argiloso, húmico e alagado
- transforme a variável solo em dummy (3 novas colunas: arenoso, argiloso, húmico) <sup>2)</sup>
- salve o arquivo com texto e abra no Rcmdr
- crie o modelo linear desses dados com as novas variáveis como preditoras

- Importe os dados para o Rcommander
- Ajuste um modelo com as variáveis dummy no menu **Estatística > Ajuste de Modelos > Modelo Linear**. Use a fórmula abaixo para construí o modelo.

```
colhe ~ arenoso + argiloso + humico
```

- Ajuste o modelo normal de ANOVA seguindo os mesmos passos anteriores mudando a fórmula do modelo para:

```
colhe~solo
```

- Avalie o modelo ANOVA indo no menu **Modelos > Resumir modelo** e clique em OK.
- Para olhar a tabela de ANOVA, vá no menu no menu **Modelos > Testes de hipóteses > Tabela de ANOVA**

- compare os coeficientes dos dois modelos (veja os resultados na janela **Outputs**)

1)

desvios quadráticos da regressão dividido pelo soma dos desvios quadrático total

2)

“000”, “100”, “010” e “001” representando cada uma uma variável, note que um nível não foi representado como dummy, esse será representado pelo intercepto do modelo

From:

<http://labtrop.ib.usp.br/> - **Laboratório de Ecologia de Florestas Tropicais**

Permanent link:

[http://labtrop.ib.usp.br/doku.php?id=cursos:planeco2019:roteiro:08-lm\\_base](http://labtrop.ib.usp.br/doku.php?id=cursos:planeco2019:roteiro:08-lm_base)



Last update: **2019/12/11 12:31**