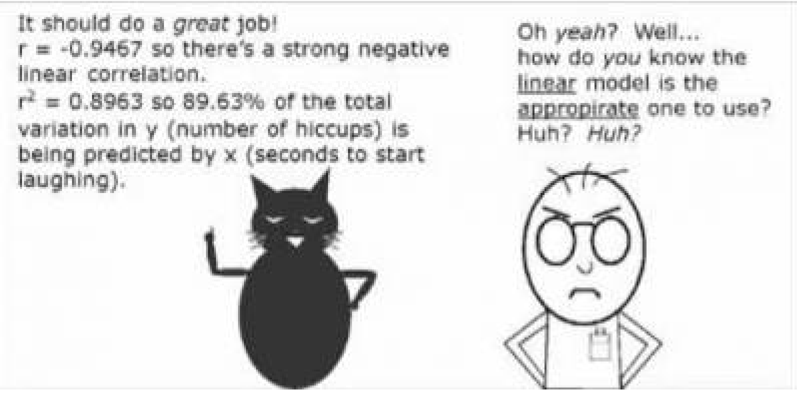
Modelos Lineares I



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.

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

- Link do canal do vídeo no youtube

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:

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

Note que estamos usando uma notação diferente da aula de regressão linear, mas a expressão é a mesma:

α = A

β = B

Ou seja, os parâmetros da população ao qual não temos acesso. O componente aleatório é expresso por uma variável probabilística Gaussiana da seguinte forma:

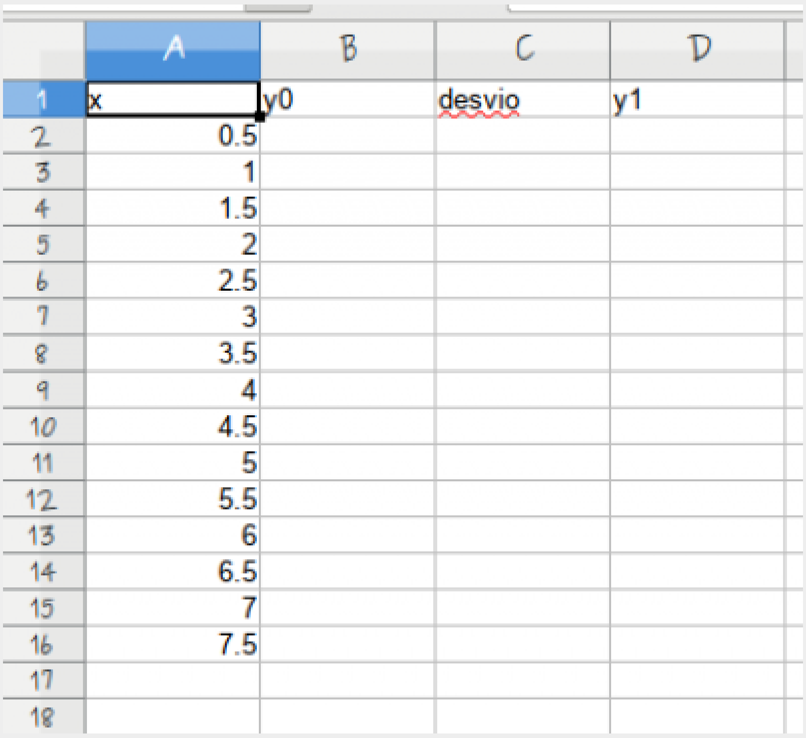
$$ \epsilon = N(0, \sigma) $$

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 ; 7)*** [1]. **Essa fórmula vai retornar valores aleatórios tomados de uma distribuição normal com média 0 e desvio padrão 7**;
- 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 **+**.

C2 =INV.NORM.N(ALEATÓRIO(), 0, 7)

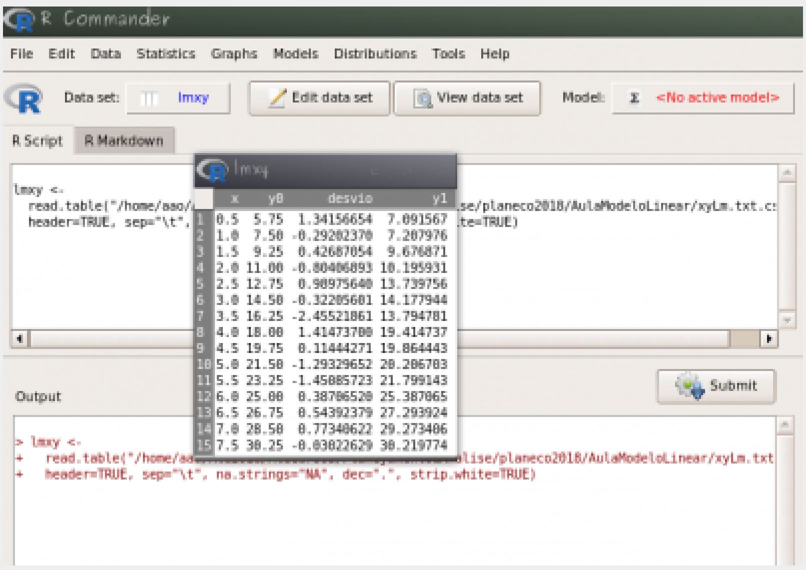
	A	B	C	D
1	x	y0	desvio	
2	0.5	5.75	-5.38956884	
3	1	7.5	-8.59748141	
4	1.5	9.25	-9.50622887	
5	2	11	-4.60569083	
6	2.5	12.75	2.807467015	
7	3	14.5	6.0259677	
8	3.5	16.25	3.53594984	
9	4	18	-0.22545112	
10	4.5	19.75	-8.6177537	
11	5	21.5	-5.64474034	
12	5.5	23.25	-1.00914875	
13	6	25	7.048986761	
14	6.5	26.75	1.930846798	
15	7	28.5	-22.8184108	
16	7.5	30.25	-6.57969081	
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 7.

- 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)*** ou ***=B2+C2*** , depois copie para as outras células da coluna;
- salve a planilha como texto separado por vírgulas e use o nome "xy.csv"

Note que a cada vez que faz algum cálculo na planilha os valores dos desvios são atualizados, ou seja, novas amostras são feitas da pela função ***INV.NORM.N*** os valores de desvios atualizados. Para evitar esse comportamento podemos selecionar os valores desta coluna e usar `Editar > Colar especial` e usar a opção de colar apenas os valores numericos, com isso a formula some e os valores não são mais atualizados a todo momento.

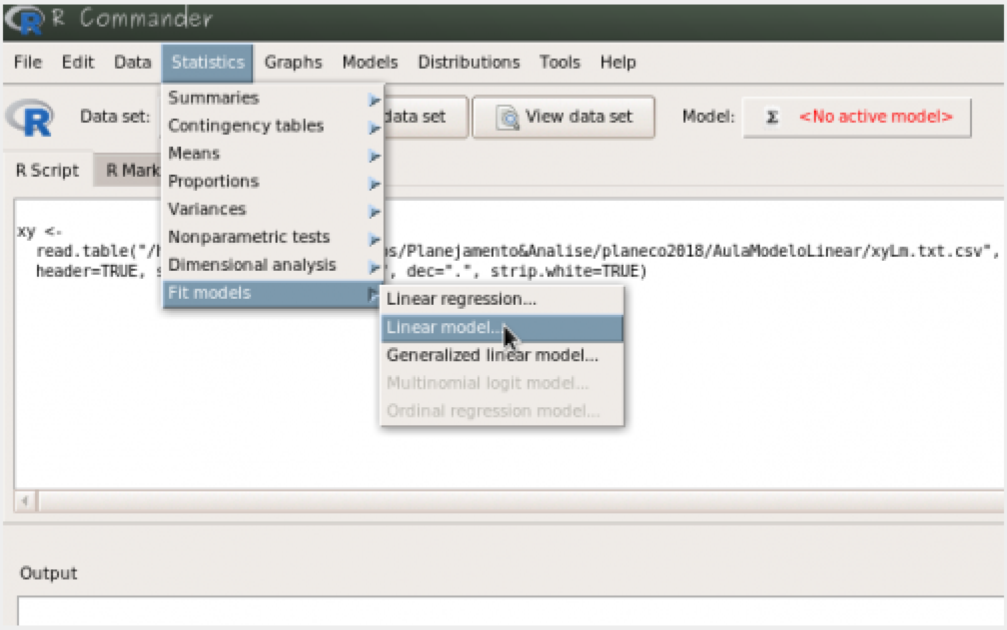
- importe os dados da planilha para o Rcommander (lembrando de selecionar como separador a vírgula) e use o nome **xy** ;
- garanta que os dados foram lidos corretamente, clicando em *View data set*



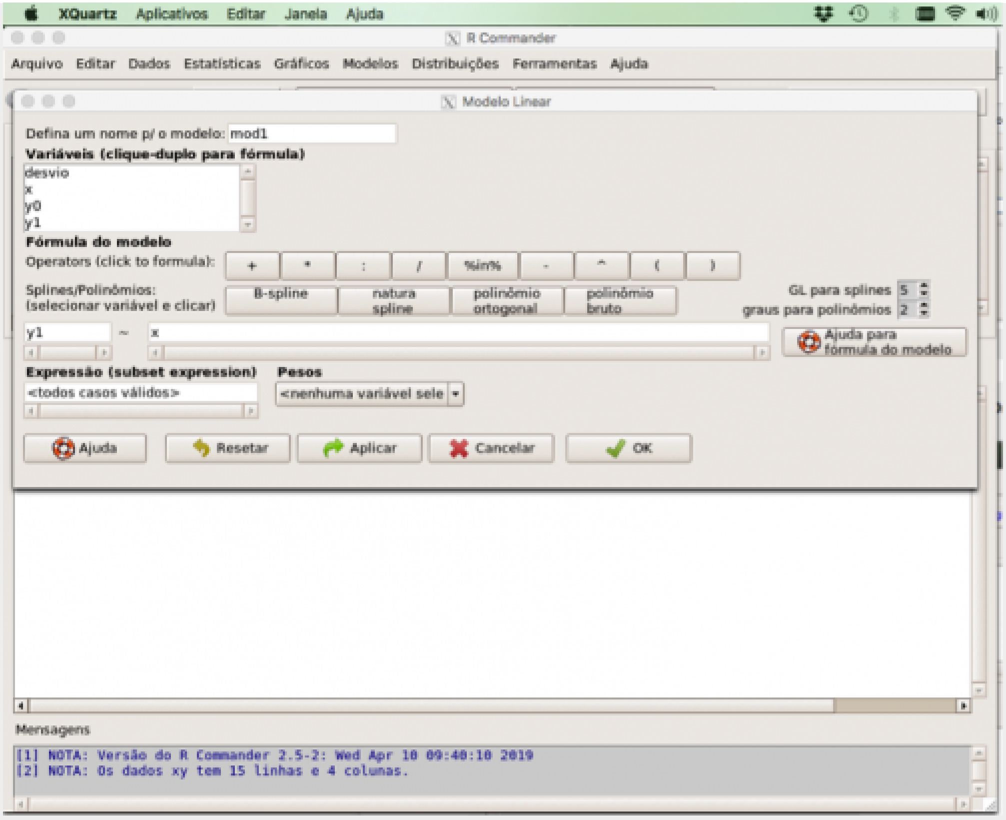
Modelos Lineares Simples

Criando o modelo no Rcmdr

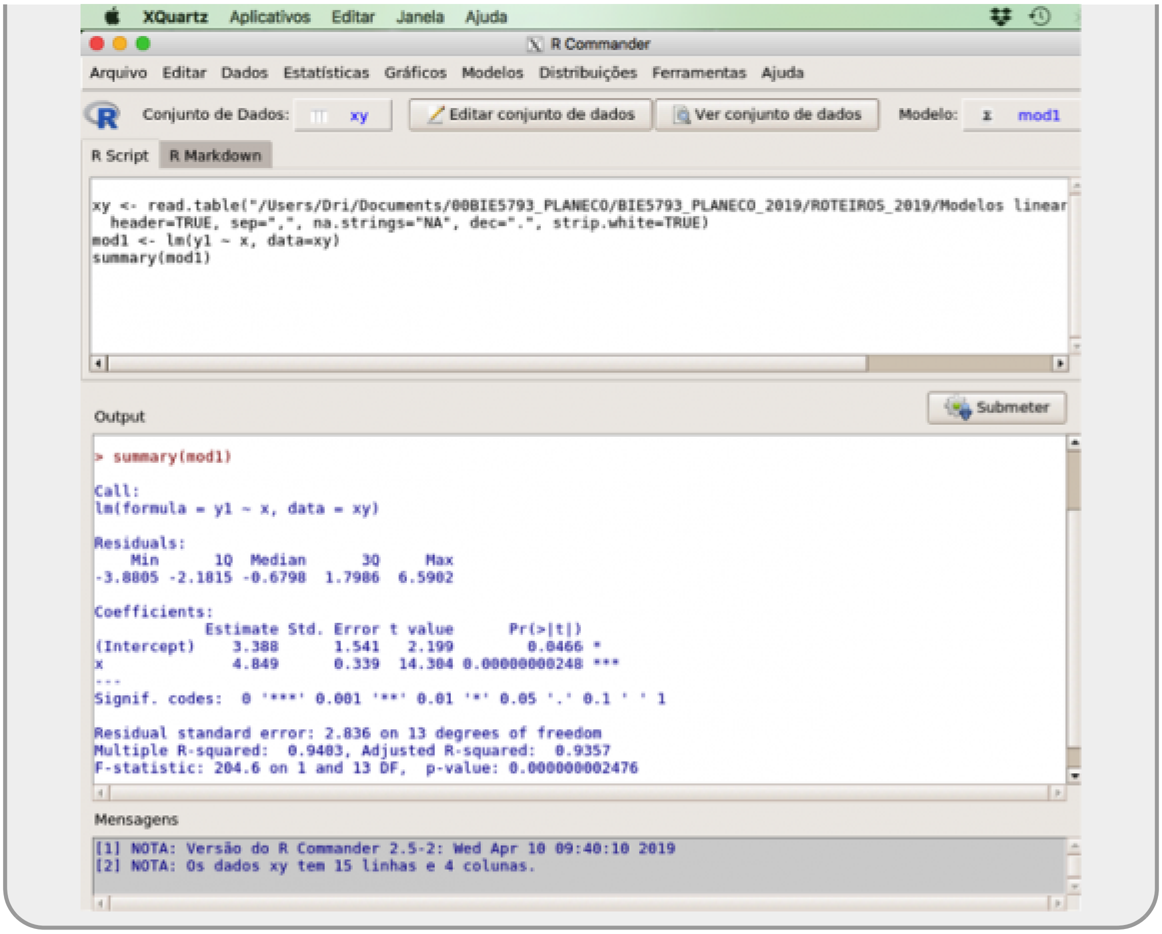
Abra o menu **Statistics** > **Fit Models** > **Linear Models…**



- 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**[2)]



Resultados do Modelo I

Anote os valores do resultado da análise na planilha modelo linear I

ATENÇÃO A PLANILHA GOOGLE PODE ESTAR FORMATADA PARA DECIMAL COM ,. CONFIRA AO FAZER A TRANSPOSIÇÃO DE VALORES

Múltiplos Experimentos

A base da estatística frequentista é que uma amostra e seus resultados são apenas uma realização dentre os possíveis resultados provenientes de uma população real, a qual não temos acesso. Utilizando os resultados de outros alunos na tabela modelo linear I, vamos investigar alguns conceitos importantes.

1. Baixe a planilha modelo linear I no seu computador, depois de incluir o seu dado. Não se preocupe em esperar todos os colegas completarem a planilha, repetimos algumas

vezes a simulação de dados para que possam usar, mesmo que nenhum outro aluno tenha feito ainda. **Não calcule nenhum valor diretamente na planilha do Google**

2. Calcule a média e o desvio padrão dos parâmetros dessa planilha
3. Conte o número de vezes que o p-valor foi maior do que 0.05.
4. Responda as perguntas indicadas no questionário no final dessa atividade.

Variabilidades e Incertezas

Para entendermos melhor uma das fontes de variabilidade que afeta nossas estimativas e também o resíduo do modelo, vamos fazer uma pequena modificação nos nossos dados simulados, aumentando (MUITO!) a variabilidade do nosso sistema. Para isso precisamos apenas mudar o parâmetro da nossa população associados à sua variabilidade (no caso, o parâmetro `desvio padrão`). Desta forma, a nossa população estatística incorpora maior variabilidade. Isso, por consequência, afeta nossas estimativas. Vamos investigar como:

- simule um novo conjunto de dados usando os mesmo passos anteriores, mudando apenas o comando:

INV.NORM.N(ALEATÓRIO(); 0 ; 7)

para:

INV.NORM.N(ALEATÓRIO(); 0 ; 14)

- Salve o arquivo com os dado simulados pois iremos utilizá-lo no próximo roteiro;
- suba os dados para o Rcommander;
- construa o modelo no Rcommander;
- salve os resultados do modelo.

Resultado do Modelo II

- anote os resultados base do modelo na planilha modelo linear simples II
- depois de anotar seus resultados baixe a planilha no seu computador;
- faça os cáculos de médias e desvios padrão para todas os parâmetros desta planilha;
- compare esses valores com os da resultado do modelo.

Tamanho Amostral

Uma outra fonte de imprecisão no nosso modelo tem relação com a próprio desenho experimental e está associada ao tamanho da nossa amostra. Essa fonte de imprecisão, apesar de estar acoplada à variabilidade da sistema, pode ser minimizada com o aumento do esforço amostral. Vamos simular uma amostra maior para o caso acima onde o desvio padrão da população é 7, modificando a sequência de valores de x na amplitude de 0,5 a 7,5 para intervalos de 0,14, totalizando 51 observações na nossa amostra.

Para agilizar a construção desta sequência podemos criar um valor de referência para as observações de 0 a 50 e operar esse valor de referência.

- na célula ***A2*** inicie em `0` e crie uma sequencia de inteiros até 50 (célula ***A51***);
- na célula ***B2*** coloque a fórmula `=0.5+(1.4*A2)` e copie a fórmula para todas a coluna até a célula ***B51***;
- a partir deste ponto é só seguir os passos da simulação anterior;
- garanta que calculou os desvios com `INV.NORM.N(ALEATÓRIO(); 0 ; 14)`, como no exemplo anterior;
- salve os dados simulados em um arquivo para uso posterior;
- crie o modelo no Rcommander;
- salve o resultado do modelo;
- anote os resultados do modelo gerado na planilha modelo linear III ;
- salve a planilha no seu computador;
- calcule a média e o desvio padrão para todos os parâmetros;
- compare esses valores com os resutados do modelo da sua simulação de dados.

PARA ENTREGAR ANTES DO INÍCIO DA PRÓXIMA AULA

Preencha as perguntas no formulário abaixo até antes da próxima aula ou a data estipulada pela equipe da disciplina. Caso tenha algum problema, faça pelo link https://forms.gle/LuRFrjnTEmrNCccJ8. Em caso de mais de uma submissão, a última, antes do final do prazo, será considerada.

[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; 7)*, no calc do LibreOffice use *=NORMINV(RAND(),0,7))*.

[2]
a imagem do resumo do modelo aqui é meramente ilustrativa, não se basei nela como referência