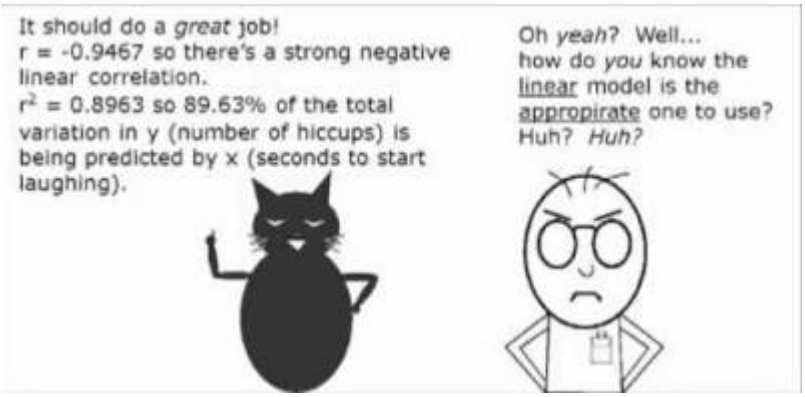
Modelos Lineares



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

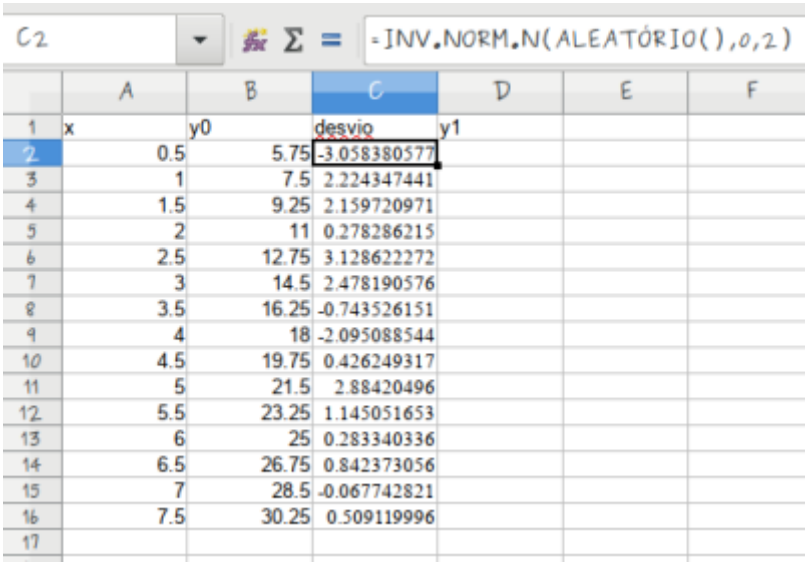
	A	B	C	D
1	x	y0	desvio	y1
2	0.5			
3	1			
4	1.5			
5	2			
6	2.5			
7	3			
8	3.5			
9	4			
10	4.5			
11	5			
12	5.5			
13	6			
14	6.5			
15	7			
16	7.5			
17				
18				

- 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;

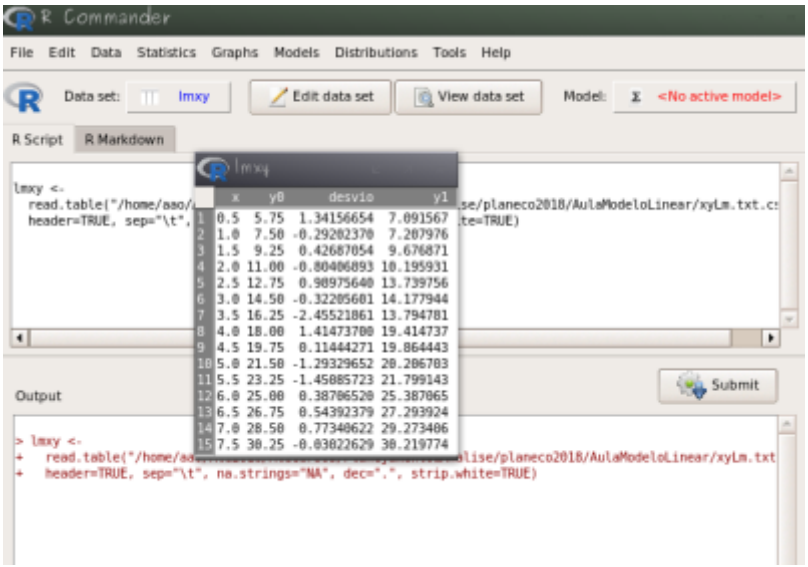
- A variável **y1** na coluna ***D*** é a soma do valor da coluna ***B*** com o valor da coluna ***C*** (y0+ desvio). Para fazer isso, coloque na célula D2 a função ***=soma(B2:C2)***, depois copie para as outras células da coluna
- salve a planilha como texto separado por vírgulas e use o nome “xy.csv”



	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.

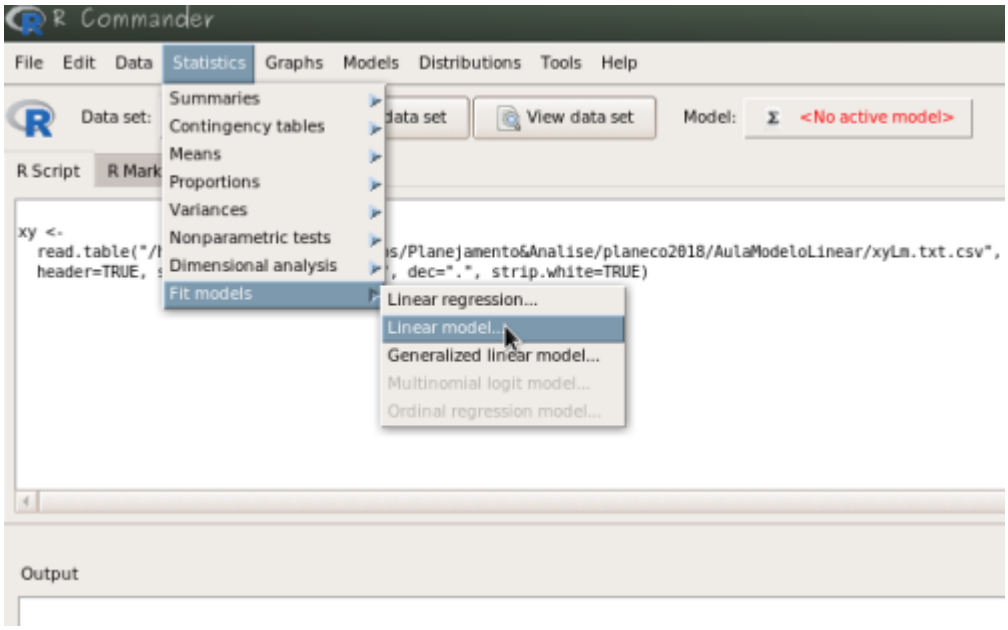
- importe os dados da planilha para o Rcommander (lembrando de selecionar como separador a vírgula) e use o nome **xy** ;
- garanta que os dados foram lidos corretamente, clicando em *View data set*



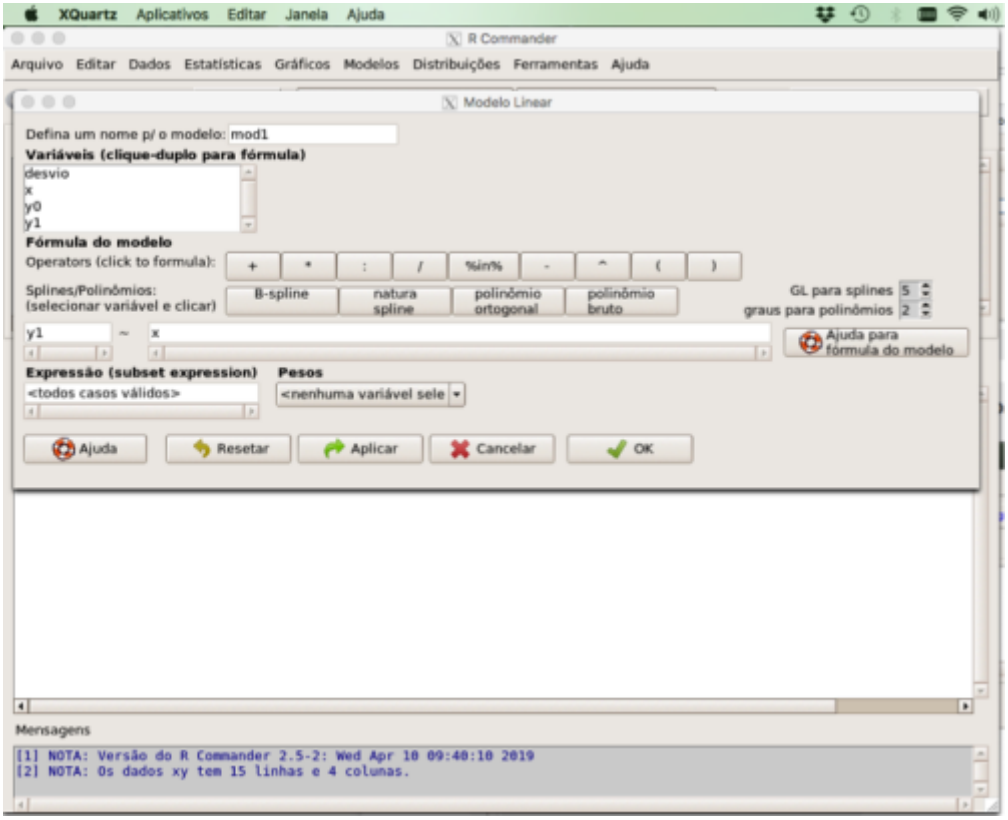
Modelo Linear Simples

Criando o modelo no Rcmdr

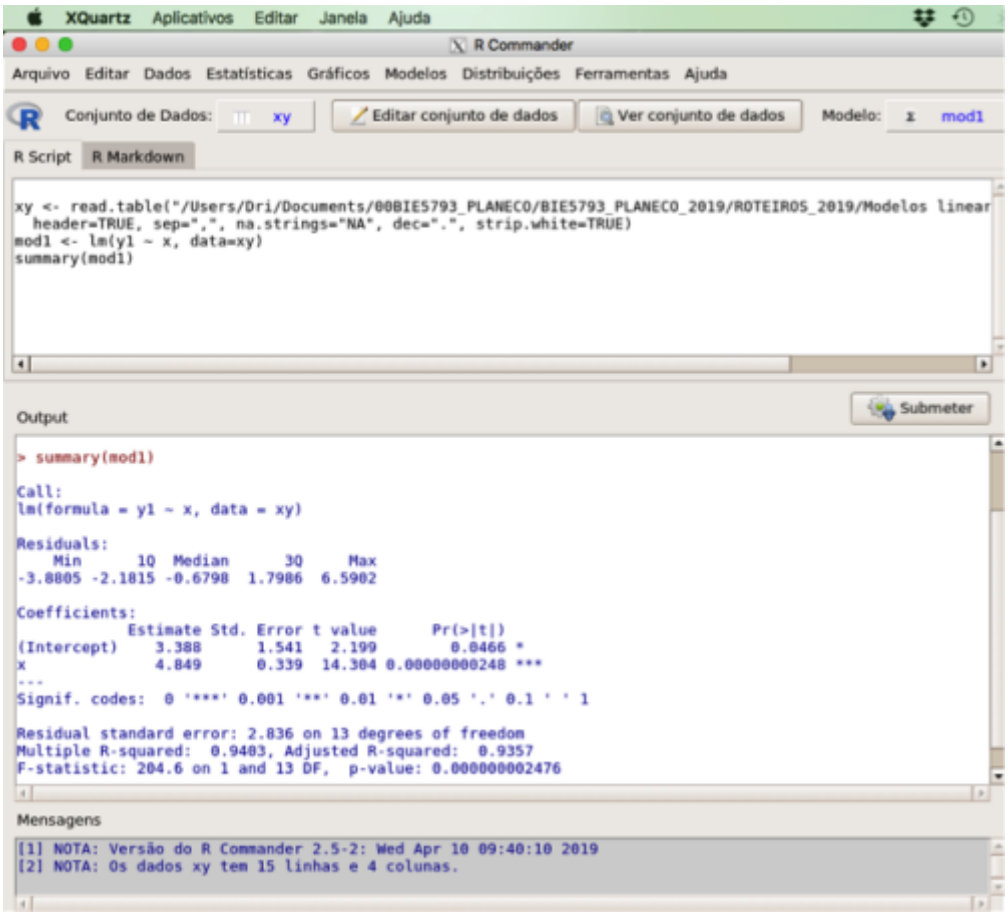
Abra o menu **Statistics** > **Fit Models** > **Linear Models...**



- 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**



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, por isso utilizamos os dados de outros anos. **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.

Incertezas

Para entendermos melhor o que afeta nossas estimativas e também o resíduo do modelo (ou erro), vamos fazer uma pequena modificação nos nossos dados simulados, aumentando (MUITO!) a variabilidade do nosso sistema. Para isso precisamos apenas mudar o parâmetro dos dados simulados associados à sua variância (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 ; 2)

para:

INV.NORM.N(ALEATÓRIO(); 0 ; 4)

- refaça todos os cálculos

Resultado do Modelo II

Guarde os resultados base do modelo na planilha modelo linear simples II

Salve o arquivo com os dado simulados pois iremos utilizá-lo no próximo roteiro.

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/xKbJrBhEQgvzQ6cG6. Em caso de mais de uma submissão, a última, antes do final do prazo, será considerada.

Exercício Modelo Linear Simples

Responda o formulário abaixo.

Para enviar as respostas é necessário estar logado no wiki.

Utilize:

- usuário: *alunos*
- senha: *planeco2020*

<form> action pagemod cursos:planeco:respostas:lm01 table_adder thanks “Respostas enviadas”

fieldset “Seus dados” textbox “Nome” @ textbox “Email” @ select “Nível” “mestrado|doutorado|nenhum” @ select “Programa” “Ecologia IBUSP| IBUSP| USP| Outra Universidade”

fieldset “Quais parâmetros definem a população?” multiselect “Selecione:” “Intercepto|P_valor|Resíduos|Inclinação|Amostra|Desvio_padrão|R_quadrado|Graus_de_liberdade”

fieldset “Anote os valores médios de todos os alunos da planilha Modelo Linear simples I e II”

number “Intercept I” number “Intercept II” number “Slope I” number “Slope II” number “Erro Padrão I” number “Erro Padrão II” number “R squared I” number “R squared II”

fieldset “Anote quantas vezes o p-valor foi maior que 0.05:”

number “modelo I” number “modelo II”

fieldset “Explique o que aconteceria aos valores médios das estimativas se acrescentássemos mais 1000 alunos na turma?” textarea “Resposta 1:” !

fieldset “Descreva quais as diferenças observadas nos resultados médios do modelo I e II” textarea “Resposta 2:” !

fieldset “Qual(is) valores(s) apresentado(s) no modelo indica(m) ” textbox “variabilidade do sistema:” textbox “incerteza nas estimativas:”

fieldset “Qual a interpretação do p-valor e do r-squared nos modelos lineares?” textarea “p-valor:” ! textarea “r-squared:” !

submit

</form>

Tabela de Anova de uma Regressão

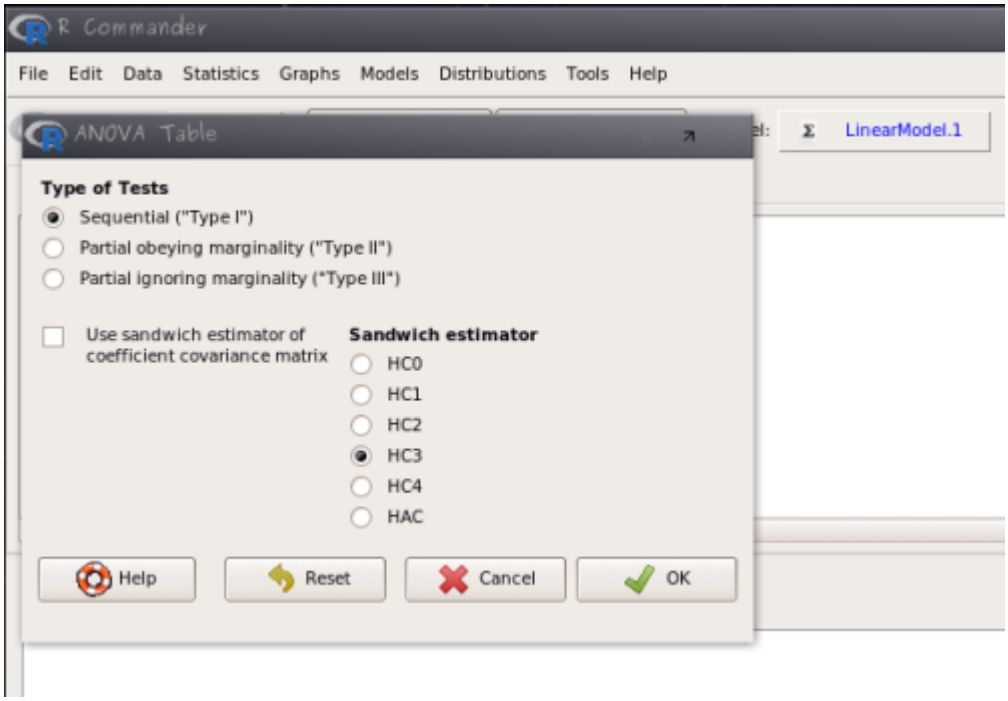
Os modelos lineares podem ser analisados através do método de partição de variância que aprendemos no roteiro de Princípios da Estatística Frequentista. Caso não tenha sedimentado bem o conceito, retorne ao roteiro e reveja a videaula, isso será importante para acompanhar o restante deste roteiro. Assim como na análise de variância clássica, podemos particionar a variação total existente nos dados de uma variável preditora contínua nas porções explicadas e não explicadas pelo modelo linear. Assista ao vídeo abaixo para entender como se dá o particionamento da variação no caso de um modelo linear simples e como essa partição é análoga ao que acontece em uma análise de variância.

Video

Agora, vamos utilizar os dados do modelo II que simulamos (dp = 4) no tutorial do tópico simulando_dados do roteiro anterior e criar novamente o modelo linear no Rcmdr[2].

Tabela de Anova do LM

- confira se o modelo está ativo. 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

O modelo gerado e seu resultado, apresentado na tabela de partição de variância, nada mais é do que uma comparação com o modelo nulo ou mínimo [3]. A tabela de Anova de um modelo isolado é

equivalente a comparar o modelo em questão com o modelo nulo. Para verificarmos isso, vamos comparar o resultado da tabela de anova do modelo com uma tabela de anova que compara o modelo com o modelo mínimo, sem preditoras. O entendimento desse conceito será fundamental para entendermos a comparação de modelo por partição de variância, interpretando a tabela de ANOVA na comparação de dois modelos. Assista ao video abaixo com a explicação sobre este conceito.

Video

Criando o modelo mínimo (nulo) no Rcmdr

- 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

Nesse ponto, é desejável que tenha entendido que a partição da variância de um modelo é correspondente a compará-lo com o modelo nulo, ou seja, quanta variância o modelo é capaz de explicar em relação ao modelo nulo. Esse modelo nulo, representa o modelo mais simples com a variação total dos dados e é representado por apenas um parâmetro, a média da variável resposta.

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 das lagartas;
- 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** [4)]
- **INTERCEPÇÃO** [5)]

- 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

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

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

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

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

$\hat{y_i}$ = 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)**[6] para o calculo 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 do cálculo dos desvios quadráticos médios dos resíduos (n - 2).

- calcule o $r2$ (coeficiente de determinação) da regressão [7];

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

- 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;

Diagnóstico do Modelo Linear

O diagnóstico do modelo linear é feito baseado nas premissas associadas ao modelo e para verificar a influência de cada observação na estimativa dos parâmetros do modelo. Os nossos dados precisam estar acoplados às premissas do modelo linear e não é desejável que o modelo seja definido apenas por uma ou por poucas observações influentes. As principais premissas dos modelos lineares são:

- a relação entre a variável preditora e a resposta é linear;
- a variabilidade tem estrutura de uma variável aleatória normal;
- a variabilidade na resposta é constante ao longo de toda a amplitude da preditora;

Além disso, avaliamos, para cada observação, sua alavancagem (leverage), definida pelo quanto a observação se afasta da média dos dados, e a sua influência (distância de Cook), definida como o quanto os parâmetros estimados são alterados ao se retirar esta observação dos dados.

Faça ou refaça o tutorial Regressão Linear para entender ou sedimentar o diagnóstico dos modelos lineares.

Variáveis Indicadoras (Dummies)

Uma das razões para a unificação do testes clássicos em modelos lineares foi a transformação das variáveis categóricas em variáveis indicadoras, também chamadas de dummies. As variáveis indicadoras são definidas pelas categorias da variável aleatória, indicando `1` quando a observação pertence ao nível e `0` quando não pertence. Para cada nível precisamos de uma indicadora, com exceção do nível que é considerado basal, indicado pelo `0` em todas as variáveis indicadoras dos outros níveis. Portanto, precisamos de:

$$n_{levels} - 1$$

variáveis indicadoras para cada variável categórica em nosso modelo. Dessa forma, para uma variável preditora categórica com `4 níveis` teremos `3 variáveis indicadoras` no modelo e se tivermos duas variáveis categóricas preditoras, cada uma com `3 níveis`, teremos `4 variáveis indicadoras`, duas para cada. Com a transformação para variáveis indicadoras, o modelo linear pode tratar as variáveis categóricas como variáveis numéricas binárias e assim, podemos inserir variáveis numéricas e categóricas como preditoras indistintamente no modelo linear. Entretanto, entender que as categorias foram transformadas em indicadoras é essencial para a interpretação destas variáveis nos outputs do modelo. Veja a explicação mais detalhada na videoaula abaixo:

Video

- 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) [8)]
- 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 construir o modelo:

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

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

* Ajuste o modelo normal de ANOVA seguindo os mesmo passos anteriores, apenas mudando a

fórmula do modelo para:

```
colhe~solo
```

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

PARA ENTREGAR ANTES DO INÍCIO DA PRÓXIMA AULA

- Entre em uma conta google e preencha o formulário abaixo.
- Caso não tenha conta ou não consiga preencher pelo link do formulário, encaminhe as repostas e documentos aos professores (**planecousp@gmail.com**), indicando como "Assunto": **Modelos Lineares Simples II**.

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))*.
2)
caso não lembre, volte ao roteiro e refaça a construção do modelo com os dados gerados com dp = 4
3)
quando não há nenhuma variável preditora
4)
SLOPE no LibreOffice
5)
INTERCEPT no LibreOffice
6)
F.DIST no LibreOffice
7)
desvios quadráticos da regressão dividido pelo soma dos desvios quadrático total
8)
"1;0;0", "0;1;0" e "0;0;1" representando cada uma uma variável. Note que um nível (alagado) não foi representado como dummy, esse será representado pelo intercepto do modelo