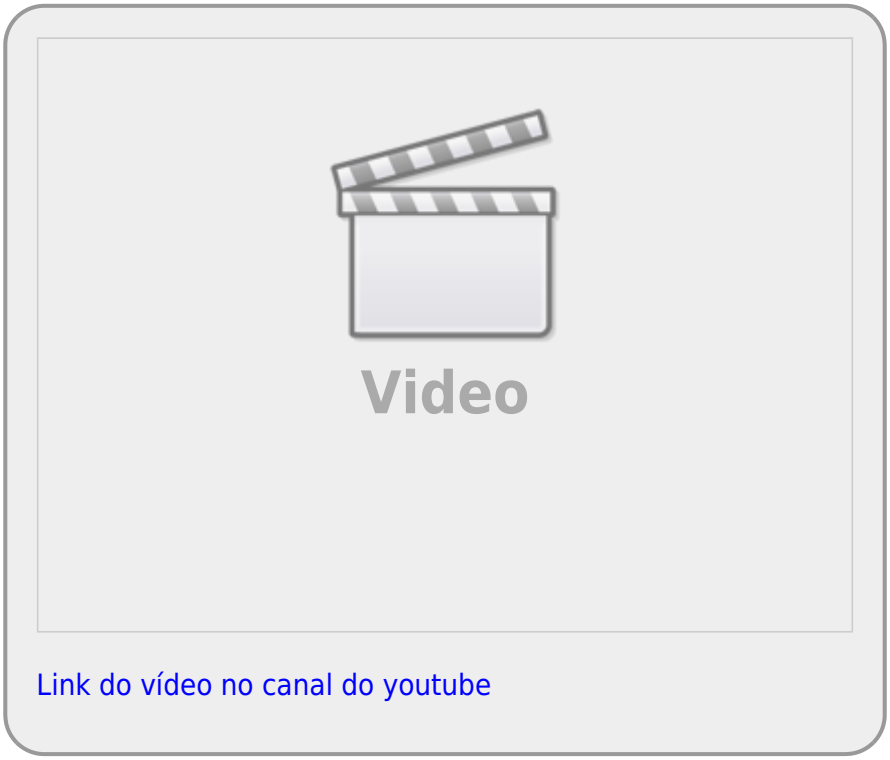
Modelos Lineares Simples II

Os modelos lineares são a base para o entendimento de todos os modelos mais complexos que iremos abordar durante este curso. Caso ainda não tenha feito o roteiro Modelos Lineares Simples I, retorne a ele.

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 onde a preditora é uma variável categórica, podemos particionar a variação total existente nos dados nas porções explicadas e não explicadas por uma **variável contínua preditora**. Esse particionamento da variação no caso de um modelo linear simples é análogo ao que acontece em uma análise de variância tradicional, com a diferença que essa última só pode ser aplicada para variáveis preditoras categóricas.



Link do vídeo no canal do youtube

A nossa próxima atividade usa os dados de crescimento de lagartas submetidas a dietas de folhas com diferentes concentrações de taninos presente no livro The R Book (Crawley, 2012). 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.

Desvios Quadráticos

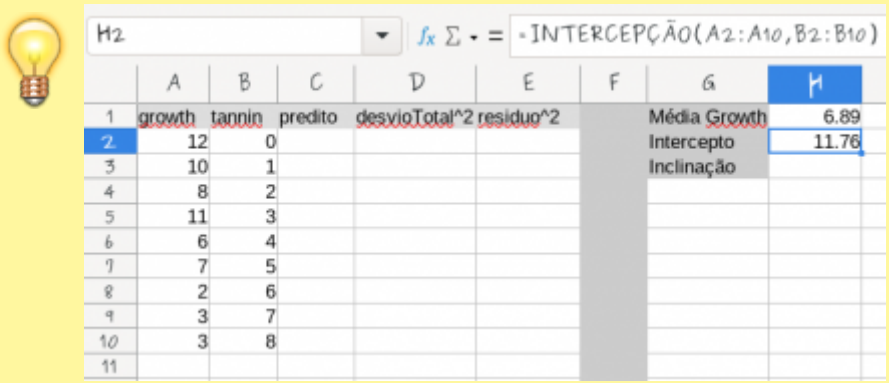
- baixe o arquivo

  regression.txt

  ;
- abra o arquivo no Excel, selecionando a separação de campo como tabulação;
- calcule a média de crescimento das lagartas;
- 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** [1]: veja documentação da função aqui.
- **INTERCEPÇÃO** [2]: Veja a documetação da função aqui



- em uma coluna chamada **desvio total** calcule o desvio total de cada

observação (o crescimento observado menos a média do crescimento);
- nomei uma coluna **desvios quadráticos totais** e eleve ao quadrado os valores da coluna criada anteriormente;
- some esses valores para obter a soma dos desvios quadráticos total nomeado como **Variação Total**
- calcule o valor predito pelo modelo em uma coluna chamada **predito**;

Predito pelo modelo

A predição do modelo é calculada pela equação da reta:

$$ \hat{y_i} = a + b * x_i $$

a = intercepto

b = inclinação

x_i = valor de x da observação `i`

$\hat{y_i}$ = valor predito para a observação `i`

- em uma coluna chamada **residuo** calcule a diferença entre cada observação e o respectivo valor predito pelo modelo;
- crie uma outra coluna (**residuo^2**) com os valores de resíduos quadrático do modelo para cada observação (observado menos o predito pelo modelo ao quadrado);
- some os desvios quadráticos dos resíduos para calcular a soma dos desvios quadráticos do modelo e nomeie esse valor como **Variação Resido^2**;
- faça a diferença entre a soma dos desvios quadráticos total pela soma dos desvios quadráticos dos resíduos para calcular a `Variação Explicada` pelo modelo;

Tabela de Anova de um Modelo Linear

A partir da partição da variação dos desvios quadráticos explicado pela preditora (`tannin`) e não explicado (`residuos`) podemos montar uma tabela de anova da mesma forma que fizemos no tutorial Testes Clássicos: ANOVA

Tabela de Anova Dieta de Lagarta

A tabela de anova tem as seguintes colunas e linhas:

- colunas: `soma quadrática`, `graus de liberdade`, `média quadrática`, `F` e `p-valor`
- linhas: `Modelo`, `Resíduo`, `Total`

- 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)`[3] 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 do cálculo dos desvios quadráticos médios dos resíduos (n - 2) que é o número de observações menos dois graus relativos ao cálculo do intercepto e da inclinação.

- calcule o $r2$ (coeficiente de determinação) da regressão [4];
- salve a planilha completa para envio no formulário.

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

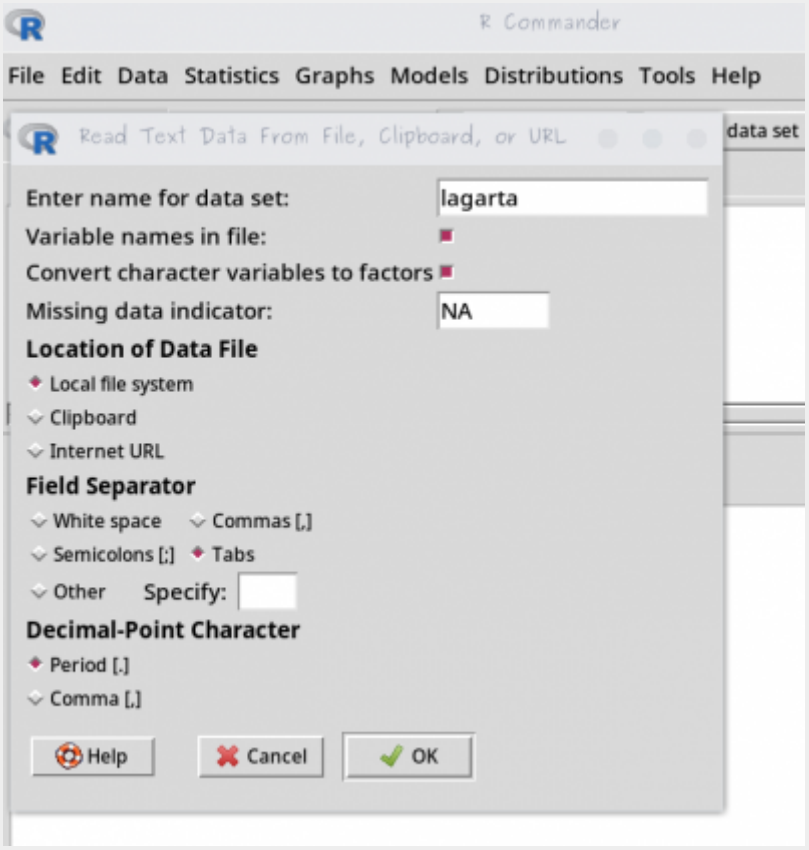
Modelo Linear: tabela de anova no R

Vamos agora fazer a `tabela de Anova` no R

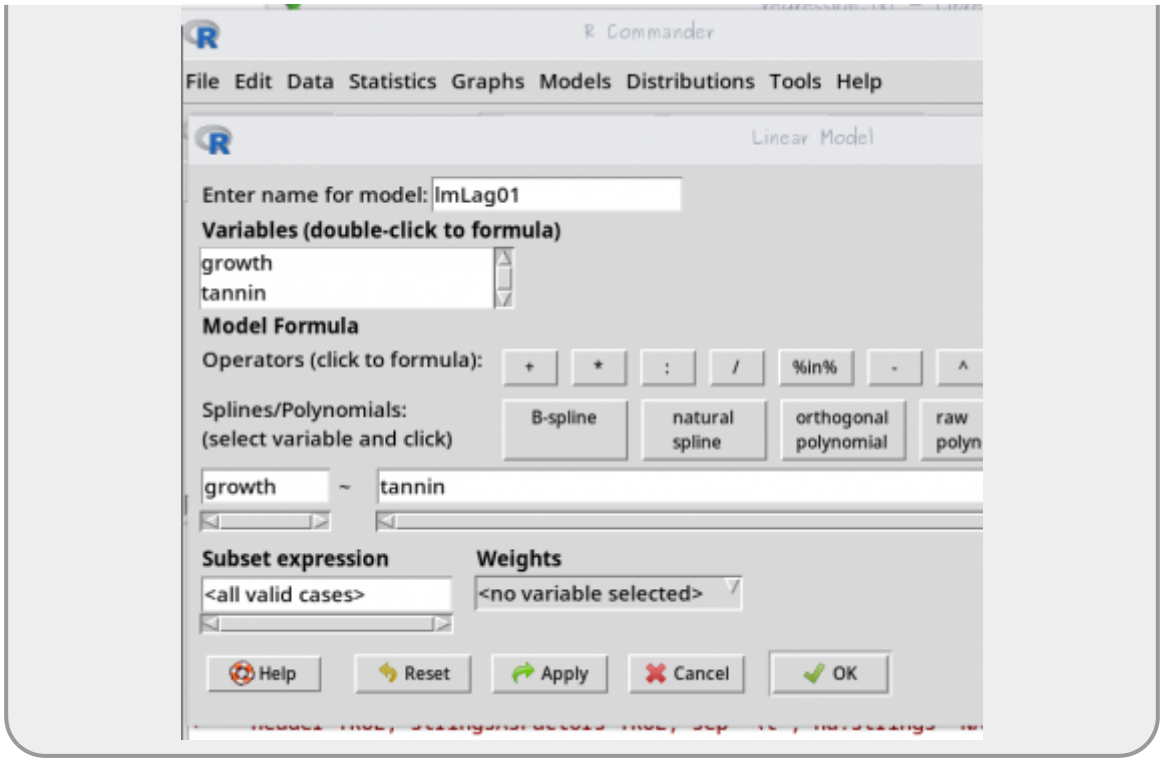
- leia os dados

  lagarta.txt

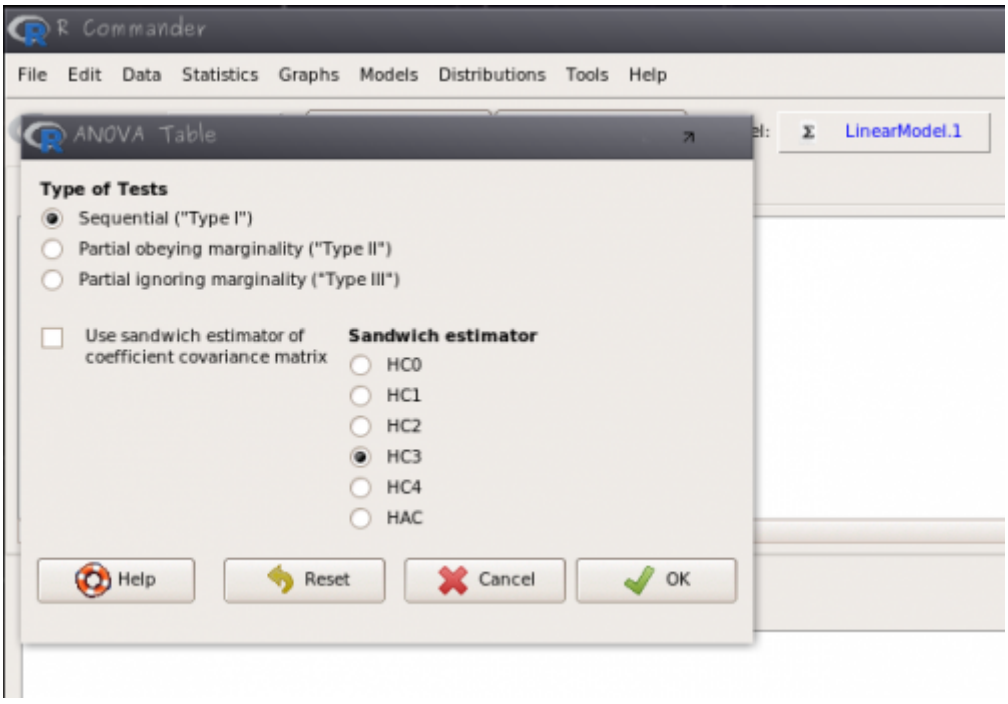
  no Rcommander, não esqueça de selecionar `Tabs` como separador de campo[5];



- monte um novo modelo linear, chamado `lmLag01`, pelo menu (`Statistics` > `Fit Models` > `Linear Models`), selecione:
  - `growth` como variável resposta;
  - `tannin` como variável preditora;



- interprete o resultado desse modelo
- faça a tabela de ANOVA do modelo gerado (`Models` > `Hipothesis test` > `Anova table`);
- durante o curso iremos usar a `tabela de ANOVA tipo I` onde a partição de variância é sequencial na ordem que os fatores são incluídos no modelo[6)];
- marque a opção: **Sequential (“Type I”)**;



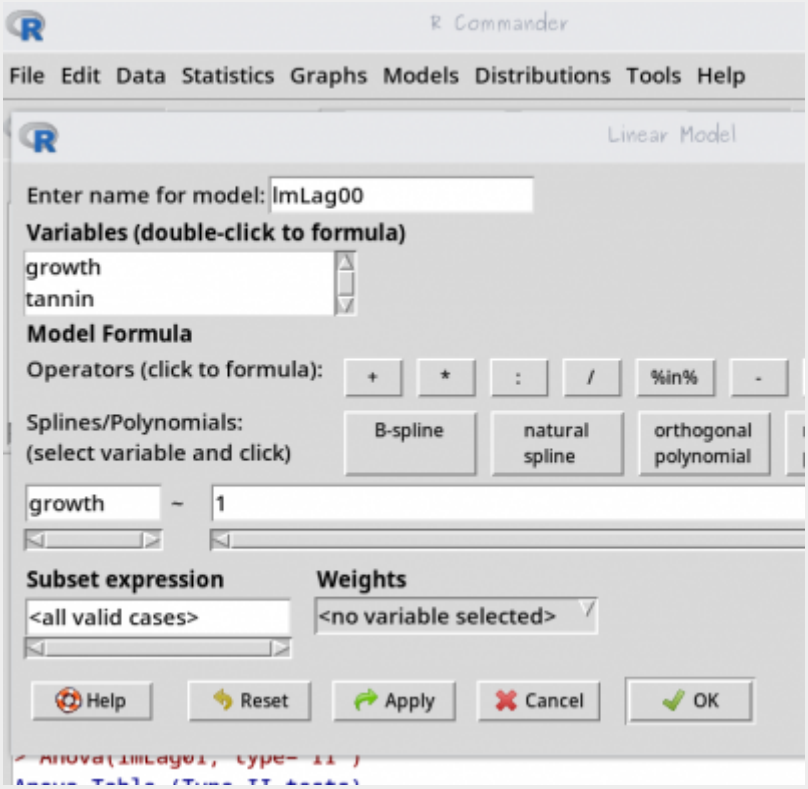
- compare os valores calculados na planilha eletrônica com a tabela de

ANOVA do modelo linear do Rcmdr, reconheça a partição da variação em ambos.

Modelo Mínimo

Com esses mesmos dados podemos construir o modelo denominado **mínimo** ou **nulo**. No experimento de crescimento da lagarta, a hipótese nula é que `tannin` não tem efeito em `growth`. Podemos construir o modelo que representa esse cenário, criando o modelo em que `growth` não tem preditoras.

- garanta que o os dados `lagarta` estão ativos no Rcmdr;
- monte um novo modelo linear, chamado `lmLag00`, pelo menu (`Statistics` > `Fit Models` > `Linear Models`), selecione:
  - `growth` como variável resposta;
  - inclua `1`,numeral um, como variável preditora[7];



- monte a tabela de anova do modelo `lmLag00` no menu: `Models` > `Hipothesis tests` > `ANOVA table`

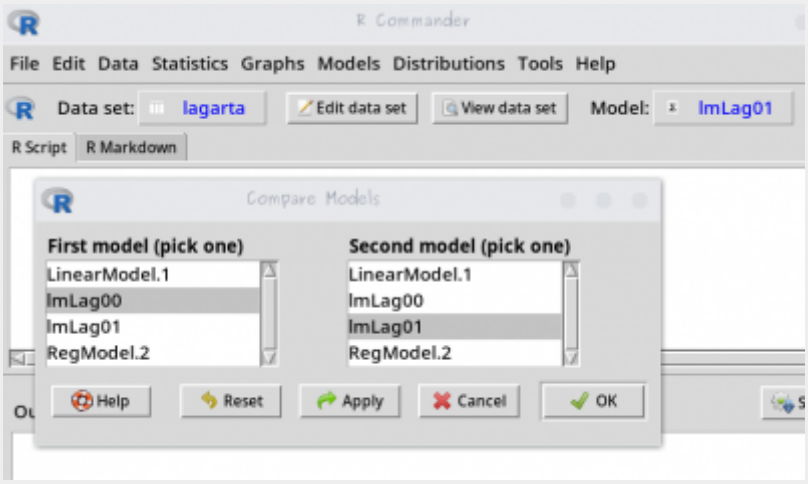
Não há muito a ser interpretado nos resultados do modelo mínimo, mas reconheça os valores que são estimados no resultado do modelo em `Coefficients Estimate`. Note que neste modelo não há inclinação, pois não existe preditora. Na tabela de ANOVA verifique o valor do `Sum Sq Residuals` e reconheça onde ele se encontra na tabela de ANOVA montada na planilha eletrônica.

Comparando Modelos

O procedimento de partição da variação e calculo da razão entre variâncias pode ser generalizado e utilizada como critério para comparação de modelos aninhados. Modelo são considerados aninhados quando o mais complexo engloba todos as variáveis do mais simples, e por consequência, o modelo mais simples não pode explicar mais variação do que o mais complexo. O modelo `lmLag00` é aninhado ao modelo `lmLag01` e por isso podemos fazer a comparação entre eles pelo critério de partição da variação como segue.

Comparando modelo com o mínimo (nulo) no Rcmdr

- confira se na caixa `Model:` existem os modelos `lmLag00` e `lmLag01`;
- utilize o menu `Models` > `Hypothesis Test` > `Compare two models`;
- na caixa que se abre selecione `lmLag00` e `lmLag01` para comparação;



- compare os valores dessa tabela de comparação entre modelos com a tabela de ANOVA do modelo `lmLag01`;
- reconheça os valores das partições de variação em ambos os casos.

Na comparação de modelos a razão de variância é relacionada ao quanto o modelo mais complexo explica da variação dos dados em relação ao modelo mais simples. De uma certa forma, a `tabela de ANOVA` no R sempre apresenta a partição da variância da comparação de dois modelos aninhados. A `tabela de ANOVA` de um modelo isolado é equivalente a comparar o modelo em questão com o modelo mínimo (nulo) correspondente. O entendimento desses conceitos é fundamental para utilizarmos a partição de variação como crítério para a tomada de decisão sobre qual modelo melhor explica nossos dados.

Link do vídeo no canal do youtube

Nesse ponto, é desejável que tenha entendido que a partição da variância de um modelo é correspondente a compará-lo com o modelo mínimo (nulo), ou seja, quanta variância o modelo é capaz de explicar em relação ao modelo sem nenhuma preditora. Este modelo mínimo, representado por apenas um parâmetro, a média da variável resposta, apresenta toda a variação dos dados contida nos seus resíduos.

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.

Caso ainda tenha dúvidas sobre o diagnóstico dos modelos revisite o tutorial Regressão Linear para sedimentar o diagnóstico dos modelos lineares.

PARA ENTREGAR ANTES DO INÍCIO DA PRÓXIMA AULA

- Preencha o formulário neste link. Caso não consiga, encaminhe as repostas e documentos aos professores (**planecousp@gmail.com**), indicando como "Assunto": **Modelos Lineares Simples II**.

Variável Indicadoras (Dummies)

No inicio deste tutorial dissemos que os modelos lineares unificaram muitos dos testes clássicos da estatística frequentista. Uma dos elementos importantes para essa unificação foi a transformação das variáveis preditoras categóricas em **variáveis indicadoras**, também chamadas de **dummies**. O procedimento consiste basicamente em criar novas variável para representar as categoria da variável preditora. Para cada categoria há uma indicadora contendo `1` quando a observação pertence ao nível referente 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 outras variáveis indicadoras relativas aos outros níveis da variável categórica. 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 variável.

Link do vídeo no canal do youtube

No nosso exemplo de anova a variável preditora `solo` tinha os níveis: `arenoso`, `argiloso` e `húmico`.Neste caso, cada nível de `solo` seria representada pelas indicadoras da seguinte forma:

	variável indicadoras:	
nível:	**indica arenoso**	**indica húmico**
arenoso	0	0
argiloso	1	0
húmico	0	1

O resultado deste modelo irá apresentar um intercepto e dois coeficientes, um associado ao nível `argiloso`, outro ao nível `humico`. O nível `arenoso`, não contemplado com uma variável indicadora [8] é estimado no intercepto. Essa estimativa do intercepto, no caso do exemplo apresentado na aula de anova, representa a produção média nesse tipo de solo. Os outros coeficientes apresentados pelo modelo representam o quanto os solos `argiloso` ou `humico` são em média diferentes do `arenoso`. Vamos criar um modelo e interpretar os coeficientes em um conjunto de dados que tem a variável `solo` agora com quatro níveis.

- baixe o arquivo
  colheita.csv
  ;
- abra no excel;
- note que a variável `solo` tem agora 4 níveis: `arenoso`, `argiloso`, `humico` e `alagado`;
- calcule a média de produtividade para cada tipo de `solo`;
- Importe o arquivo original `colheita.csv` para o Rcommander;
- Ajuste um modelo denominado de `lmSolo` no menu `Estatística` > `Ajuste de Modelos` > `Modelo Linear`. O modelo deve ser definido como na fórmula abaixo:

```
colhe~solo
```

- compare os coeficientes estimados pelo modelo com os valores de produtividade média para cada tipo de solo.

Para entender o procedimento das variáveis `indicadoras` vamos construir explicitamente nossas variáveis indicadoras.

- abra o arquivo
  colheita.csv
  no excel;
- crie 3 novas colunas nomeadas de: `arenoso`, `argiloso`, `humico`.
- para cada observação (linha) represente o nível do solo com o valor `1` na respectiva indicadora e `0` nas outras. Note que um nível não precisa de indicadora pois será representado pela indicação de `0` em todas as indicadoras, no nosso caso o nível `alagado` [9];
- salve a planilha no formato `.csv`;
- importe essa planilha com as variáveis indicadoras para o Rcommander;
- ajuste um modelo denominado de `lmSoloIndica` com as variáveis indicadoras no menu `Estatística` > `Ajuste de Modelos` > `Modelo Linear`. O modelo deve ser definido como na fórmula abaixo:

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

- Avalie o modelo com variáveis indicadoras no menu `Modelos` > `Resumir modelo` [10] e clique em OK;
- Para olhar a tabela de partição de variância, vá ao menu `Modelos` > `Testes de hipóteses` > `Tabela de ANOVA`
- Compare os dois modelos `lmSolo` e `lmSoloIndica`

A transformação de variáveis resposta categóricas para variáveis indicadoras permite que o modelo linear possa tratar indistintamente variáveis categóricas e contínuas. Essa unificação simplifica muito a construção de modelos e sua operacionalização, entretanto, entender que as categorias foram transformadas em indicadoras é essencial para entender e interpretar o resultado apresentado pelos modelos lineares.

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 por esse link. Em caso de mais de uma submissão, a última, antes do final do prazo, será considerada.

1) SLOPE no LibreOffice

2) INTERCEPT no LibreOffice

3) F.DIST no LibreOffice

4) desvios quadráticos da regressão dividido pelo soma dos desvios quadrático total

5) confira que os dados foram lidos corretamente

6) Quando se tem mais de uma preditora é possível calcular a partição da variação em diferentes sequências, por isso existem tipos diferentes de tabelas de ANOVA

7) esta é a forma de dizer ao R que nosso modelo não tem preditoras

8) representado por `00` nas outras indicadoras

9) os valores `0;0;0` `1;0;0`, `0;1;0` e `0;0;1` em cada indicadora representam respectivamente: `alagado`,`arenoso`, `argiloso` e `humico`

10) `Models` > `Summarize model`