UNIVERSIDADE FEDERAL DE LAVRAS
DEPARTAMENTO DE CIÊNCIAS EXATAS

# Uso de Recursos Computacionais

**Daniel Furtado Ferreira**

LAVRAS
Minas Gerais - Brasil
21 de março de 2007

# Sumário

# Lista de Tabelas

# Lista de Figuras

# Capítulo 1

# Introdução ao sistema SAS

O sistema SAS é um dos melhores software existentes na atualidade. Atualmente somente o programa R tem competido com o SAS®. O sistema SAS é um software que propicia grandes vantagens no tratamento de bancos de dados, nas análises estatísticas e na geração de relatórios das mais variadas formas. Para utilizarmos o SAS precisamos conhecer como é sua estrutura e como se dá o seu funcionamento. O ambiente de interação com o usuário do SAS possui três janelas, que por sua vez possuem funções específicas, a saber:

1. Janela de programas: nesta janela digitamos os programas, que são seqüências de passos e comandos para utilizarmos o sistema SAS de acordo com a finalidade que almejamos. Temos que utilizar determinados comandos específicos para chamar rotinas prontas do SAS ou podemos utilizar programas desenvolvidos para um ambiente de programação interativo, o IML.

2. Janela de erros: esta janela é conhecida como janela log e deve ser utilizada para consultarmos a ocorrência de erros de sintaxe em nossos comandos ou programas. O SAS marca os erros com letras em cor vermelha e aponta a linha do programa onde este erro ocorreu.

3. Janela de saída ou output: nesta janela obtemos os resultados não gráficos da análise recém executada. O seu conteúdo pode ser salvo em diferentes formatos ou impressos diretamente.

Todo o conteúdo das janelas pode ser salvo, marcado e eliminado utilizando os recursos do Windows e da barra de ferramentas. Não daremos maiores detalhes destes procedimentos por julgá-los muito simples. Devemos ter o cuidado único de que esses comandos são específicos para a janela que estiver ativa e não para o conteúdo de todas elas.

O SAS infelizmente não é um programa com muita interatividade, a menos que o módulo ASSIST esteja presente. Um outro recurso extremamente útil ao se utilizar o SAS é o sistema de auxílio (help on line), que permite a consulta, através de uma navegação não linear, dos principais comandos e bibliotecas do programa. Existem manuais on line em HTML e que podem ser consultados pela internet e ainda manuais em PDF que podem ser baixados e utilizados gratuitamente. Nestas notas veremos apenas os principais procedimentos do sistema SAS para realizarmos análises estatísticas. Enfatizaremos os principais recursos relacionados as análises de estatística básica, regressão e estatística experimental. Estes recursos são os mais variados e flexíveis e são abordados de maneira simples, sendo que daremos ênfase nas interpretações estatísticas dos fundamentos dos métodos e da inferência. Utilizaremos apenas exemplos acadêmicos simples, que muitas vezes foram simulados ou são dados fictícios.

## 1.1 Entrada de dados

O SAS possui inúmeros recursos de importação dos mais diferentes banco de dados e planilhas. Utilizaremos o recurso mais comum de simplesmente "colarmos" os dados de outro programa na janela de programa e salvarmos o arquivo resultante como texto (ASCII). Este formato é mais robusto, livre de vírus, além de os arquivos resultantes ocuparem menos memória. Quando possuímos valores perdidos no nosso arquivo ou banco de dados, podemos substituir a célula do arquivo que foi perdida por um ponto. Este é o default do programa SAS, podendo ser mudado de acordo com a preferência do usuário.

O arquivo SAS pode ser lido de inúmeras maneiras diferentes, porém utilizaremos as formas mais simples. Temos que pensar que cada variável deve ocupar uma coluna do arquivo e cada observação ou unidade amostral

uma linha. Esta é a estrutura utilizada pela maioria dos programas de análise estatística. Internamente, ao criarmos o banco de dados e executarmos o programa, temos que dar um nome, o qual o programa SAS utilizará para criar no seus diretórios de trabalho SASWORK ou SASUSER o conjunto de dados que estamos utilizando. Este conjunto de dados é *SAS Data Set*. Antes dos dados devemos criar três linhas de comando indicando o nome deste conjunto de dados, as variáveis e um comando de iniciação da leitura dos dados.

Cada linha de comando do SAS tem algumas palavras reservadas de comandos e termina com um <;>. Apesar de termos inúmeros comandos diferentes para entrarmos com o *SAS Data Set*, utilizaremos quase sempre a seguinte estrutura:

```
/*exemplo de um arquivo de dados com peso em kg de coelhos híbridos Norfolk abatidos
aos 90 dias de idade. Tudo que está aqui dentro é um comentário do programa.*/
data coelhos;
input peso;
cards;
2.50
2.58
2.60
2.62
2.65
2.66
2.58
2.70
2.55
2.57
2.70
2.62
2.59
2.54
2.53
2.20
;
proc print;
var peso;
run;
```

Podemos explicar os comandos usados neste simples programa da seguinte forma:

1. <data coelhos;>: este comando indica o nome do SAS Data Set. A palavra data é um comando do SAS para indicar o nome do conjunto de dados e coelhos foi o nome que escolhemos para este exemplo específico. Podemos observar que terminamos sempre com um ; a linha de comando. Assim, apesar de não ter vantagem alguma, poderíamos colocar data em uma linha, coelhos na outra e o ponto e vírgula na terceira. Fisicamente teríamos três linhas, mas uma só linha de comando.

2. <input peso;>: este comando vem com a palavra input para designar as variáveis que o nosso conjunto de dados possui. Como temos somente o peso dos coelhos híbridos Norfolk abatidos aos 90 dias em kg, somente esta variável apareceu após o comando input. Se houvesse mais variáveis, estas deveriam ser separadas por pelo menos um espaço em branco, antes do ponto e vírgula.

3. <cards;>: este comando indica que os dados virão na seqüência.

4. <proc print;>: este é um dos procedimentos, *procedure*, do SAS. Os procedimentos aparecem depois da palavra proc, utilizada como indicativo de procedimento e seguida do nome do procedimento, no caso, print. Este procedimento é utilizado para gerar relatórios de impressão na janela output.

5. <run;>: comando utilizado após cada procedimento para indicar ao SAS para executá-lo.

Depois de digitados estes comandos e colocados na janela de programas do SAS devemos submetê-lo ao compilador do programa. Para isso utilizamos o comando *submit*, que possui o atalho por meio da tecla $F8$ ou do ícone (run) correspondente na janela de programas.

Podemos utilizar na linha de comando do *input* os seguintes caracteres @@. Isto nos permite digitar o arquivo na seqüência de variáveis do arquivo,

mas não necessariamente obedecendo a estrutura de colunas. Para este exemplo teríamos:

```
/*exemplo de um arquivo de dados com peso em kg de coelhos híbridos Norfolk abatidos
aos 90 dias de idade. Tudo que está aqui dentro é um comentário do programa.*/
data coelhos;
input peso @@;
cards;
2.50 2.58 2.60 2.62 2.65
2.66 2.58 2.70 2.55 2.57
2.70 2.62 2.59
2.54 2.53 2.20
;
proc print;
    var peso;
run;
```

Um segundo exemplo com mais de uma variável é apresentado na seqüência com dados de dez árvores de *Araucaria angustifolia*. A primeira variável $Y$ é o volume em $m^3/acre$, a segunda variável $X_1$ é a área basal das árvores, a terceira variável $X_2$ é esta mesma área basal, mas tomada com referência a área basal de outra espécie (*Pinus taeda*) e a quarta variável $X_3$ é a altura das árvores em pés. Observamos que a utilização do @@ possibilita a leitura dos dados em uma estrutura de uma aparente desorganização. No entanto, podemos observar que existe uma seqüência dos valores obedecendo a seqüência das variáveis do input $Y$, $X_1$, $X_2$ e $X_3$.

```
/*exemplo de um arquivo de dados com dados de 10 árvores de araucária, com 4 variáveis.
*/
data arvores;
input Y X1 X2 X3 @@;
cards;
65 41 79 35 78 71 48 53
82 90 80 64 86 80 81 59
87 93 61 66 90 90 70 64
```

```
93 87 96 62 96 95 84 67
104 100 78 70
113 101 96 71
;
proc print;
    var Y X1 X3;
run;
```

---

Uma importante situação que acontece em exemplos reais é a ocorrência de variáveis qualitativas. Estas variáveis são identificadas por nomes alfanuméricos e o SAS permite sua presença. Assim, se um conjunto de dados possui 3 variáveis, sendo por exemplo blocos, tratamentos e produção e a variável tratamento possui seus níveis qualitativos (nomes), então devemos formar o conjunto de dados normalmente e no input após as variáveis cujos níveis são alfanuméricos, devemos colocar um $, conforme o exemplo a seguir. Isto indicará que aquelas variáveis possuem níveis que são nomes e não números.

---

```
/*exemplo de um arquivo com dados experimentais fictícios, onde os níveis dos tratamentos são alfanuméricos.*/
data exper;
input bl trat $ prod;
cards;
1 A 12.23
1 B 10.31
1 C 11.90
2 A 14.56
2 B 10.17
2 C 13.45
3 A 16.11
3 B 19.12
3 C 14.73
4 A 12.78
4 B 10.67
4 C 11.34
;
proc print data=exper;
run;
```

---

## 1.2 Transformações de variáveis

Para obtermos novas variáveis no SAS a partir de um grupo de variáveis já existentes, não precisamos criá-las fisicamente no SAS data set que temos. Podemos fazer isso utilizando alguns comandos em determinados lugares de nosso programa utilizando as funções dos SAS. O arquivo interno do SAS terá as variáveis criadas ou transformadas. Vamos descrever duas formas básicas de fazermos isso. A primeira delas utilizamos simples comandos de transformação de variáveis situados entre a linha de comando do input e do cards. Podemos utilizar uma série de operadores, sejam eles lógicos ou não. Alguns exemplos destes operadores são: $+$: soma; $-$: subtração; log: logaritmo neperiano; $\log 2$: logaritmo na base 2; $\log 10$: logaritmo na base 10; $*$: multiplicação; /: divisão; e $**$: potenciação do tipo $X^Y$, que no SAS é obtido por $X**Y$. O comando ^ não é reconhecido pelo SAS para potenciação. Operadores lógicos como $>$, GE ($\geq$), $<$, LE ($\leq$) ou $=$ podem ser usados também. Estruturas condicionais *if* $\cdots$ *then; else* são permitidas, entre outras.

Apresentamos um exemplo na seqüência um exemplo utilizando algumas destas transformações de variáveis para ilustrarmos os procedimentos.

---

```
/*exemplo de um arquivo de dados com peso em kg de coelhos híbridos Norfolk abatidos
aos 90 dias de idade.*/
data coelhos;
input peso @@;
sqrtp=peso**0.5;
pln=log(peso);
if peso<2.55 then classe=1;
  else classe=2;
cards;
2.50 2.58 2.60 2.62 2.65
2.66 2.58 2.70 2.55 2.57
2.70 2.62 2.59
2.54 2.53 2.20
;
proc print;
```

```
    var peso sqrtp pln classe;
run;
```

---

A segunda alternativa nos possibilita realizarmos transformações sobre variáveis de um *SAS Data Set* em um lugar qualquer do programa após a definição do data set original. Usamos o comando *Data* para denominarmos um novo ou o mesmo conjunto de dados e o comando *Set* para selecionar o conjunto de dados existente para realizarmos as programações que almejarmos. Apresentamos o seguinte exemplo utilizando o data set coelhos, onde não alteramos o seu nome. Veja que teremos o mesmo efeito do exemplo anterior.

---

```
/*exemplo de um arquivo de dados com peso em kg de coelhos híbridos Norfolk abatidos
aos 90 dias de idade.*/
data coelhos;
input peso @@;
cards;
2.50 2.58 2.60 2.62 2.65
2.66 2.58 2.70 2.55 2.57
2.70 2.62 2.59
2.54 2.53 2.20
;
data coelhos; set coelhos;
sqrtp=peso**0.5;
pln=log(peso);
if peso<2.55 then classe=1;
  else classe=2;
run;quit;
proc print;
    var peso sqrtp pln classe;
run;
```

---

## 1.3 Ordenamento de dados

Podemos utilizar o *proc sort* do SAS para ordenarmos conjuntos de dados especificando as variáveis que almejamos utilizar como chaves do processo de ordenação dos valores do conjunto de dados. Podemos ordenar em ordem crescente ou decrescente. Por default o SAS ordena em ordem crescente cada variável chave. Se quisermos uma ordem decrescente, devemos utilizar o comando *descending*. Ilustramos o uso do *proc sort* em um exemplo, em que uma sala de aula foi dividida em dois grupos de acordo com os lugares que os alunos sentavam. Os da bancada da direita foram denominados de grupo 1 e os da esquerda de grupo 2. Foram mensurados os pesos e altura destes alunos. Usamos o *proc sort* para ordenar por grupos em ordem crescente e por peso em ordem decrescente dentro de cada grupo.

```
/*exemplo de ordenação utilizando o proc sort.*/
data sala;
input grupo peso alt;
cards;
2 72 1.80
1 48.5 1.58
2 88 1.80
1 86 1.83
2 62 1.72
1 79 1.69
2 95 1.93
1 53 1.60
;
proc sort data=sala;
   by grupo descending peso;
run;
proc print data=sala;
run;
```

## 1.4 Procedimentos para análise estatística

Vamos utilizar neste material basicamente alguns procedimentos SAS para realizarmos análise estatística. Estes procedimentos no SAS são referenciados por *proc* que é a abreviatura de *procedure*. Vamos neste material apresentar a lógica de tais procedimentos, suas sintaxes e principalmente vamos enfatizar os métodos estatísticos que estão envolvidos neste procedimento. Vamos procurar também mostrar o *proc IML*. O programa SAS/IML fornece ao usuário uma poderosa e flexível linguagem de programação (*Interactive Matrix Language*) em um ambiente dinâmico e interativo. O objeto fundamental da linguagem é uma matriz de dados. A programação é dinâmica por causa do dimensionamento das matrizes e da alocação de memória serem feitos de forma automática.

Vamos utilizar alguns procedimentos do SAS para efetuarmos análises de estatística básica, quais sejam, *proc univariate*, *proc summary* e *proc ttest*. Para realizarmos análises de regressão linear utilizaremos o *proc reg* e para regressão não-linear o *proc nlin*. Para análises de modelos lineares vamos utilizar o *proc anova*, *proc glm* e o *proc mixed* para modelos lineares mistos. Estimaremos componentes de variâncias com o *proc varcomp*. Poderemos eventualmente utilizar algum outro procedimento específico para realizarmos algumas análises multivariadas.

O SAS é um programa que consideramos praticamente completo. Vamos neste material abordar situações específicas da estatística para fazermos uma introdução ao sistema SAS. Não temos de forma alguma a pretensão de que este seja um material de consulta imprescindível, mas que sirva de um roteiro básico para aqueles que desejam ter uma noção inicial de como efetuar análises estatísticas pelo SAS.

# Capítulo 2

# Estatística básica no SAS

O SAS possui muitos recursos para realizarmos análises estatísticas descritivas de uma amostra de tamanho $n$. Neste capítulo vamos abordar as principais estatísticas descritivas utilizando o proc univariate e o proc summary. Vamos ilustrar a obtenção de estimativas pontuais de vários parâmetros, histogramas e estimadores de Kernel. Vamos realizar inferência sobre média de uma população e de dados emparelhados, tanto testes de hipóteses como estimação intervalar e vamos inferir sobre a distribuição de probabilidade dos dados amostrais. Para dados de duas amostras independentes vamos utilizar o proc ttest para inferirmos sobre a média e sobre a variância das populações amostradas. Para alguns parâmetros vamos utilizar o IML para construirmos intervalos de confiança utilizando os recursos do SAS e a teoria de inferência. Vamos utilizar diferentes recursos dentro do contexto da estatística básica.

## 2.1 Estatísticas descritivas

Vamos utilizar basicamente o *proc univariate* e *summary* para obtermos as estatísticas descritivas de uma população. Vamos supor que temos uma população com parâmetros desconhecidos. Vamos considerar inicialmente que essa população possui uma determinada distribuição de probabilidade e que este modelo probabilístico é o normal, dado por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.1}$$

em que os parâmetros $\mu$ e $\sigma^2$ são a média e a variância respectivamente.

Este modelo é simétrico em relação à média e o parâmetro usado para medir a simetria é o coeficiente de assimetria que pode ter dois estimadores, o estimador beta e o estimador gama. No SAS o estimador gama de simetria é obtido e o seu valor de referência na distribuição normal é o valor 0. Este estimador (Ferreira, 2005[3]) é dado por:

$$g_1 = \frac{m_3\sqrt{n(n-1)}}{(n-2)m_2^{3/2}}, \tag{2.2}$$

em que $m_r = \sum_{i=1}^{n}(X_i - \bar{X}_.)^r/n$ é o estimador de centrado de momento de ordem $r$, sendo $r \geq 2$.

O coeficiente de curtose populacional da distribuição normal tem como referência o valor zero, se for considerado o estimador gama ou o valor 3 se for considerado o estimador beta. O coeficiente de curtose mede o grau de achatamento da curva. Como o SAS estima somente o parâmetro gama, temos o seguinte estimador do coeficiente de curtose:

$$g_2 = \frac{(n-1)\left[(n+1)m_4 - 3(n-1)m_2^2\right]}{(n-2)(n-3)m_2^2}. \tag{2.3}$$

Assim uma distribuição com coeficiente de assimetria igual a zero é considerada simétrica; se o coeficiente de assimetria for maior que zero, esta distribuição será assimétrica à direita e se for menor que zero, assimétrica à esquerda. Da mesma forma uma distribuição com coeficiente de curtose igual a 0 será considerada mesocúrtica; se o coeficiente de curtose for negativo, será considerada platicúrtica e se for maior que zero, será considerada leptocúrtica.

Caracterizada a distribuição, o interesse se volta para a caracterização da locação e da dispersão da população. A média amostral é dada por:

$$\bar{X}_. = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.4}$$

A variância amostral é dada por:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]. \tag{2.5}$$

O SAS estima ainda várias outras estatísticas descritivas, como o desvio padrão $S$, o coeficiente de variação $CV$, o erro padrão da média $S_{\bar{X}}$, a mediana $m_d$, a soma de quadrados corrigida e não corrigida, alguns percentis entre outras estimativas. Podemos utilizar o *proc univariate* para esta finalidade. Este procedimento ainda apresenta a vantagem de propiciar a estimação do histograma, bem como de permitir um ajuste da distribuição normal a este histograma. Permite que outras distribuições sejam plotadas e que seus parâmetros sejam estimados. Estas distribuições são: beta, exponencial, gama, Weibull e lognormal. Permite ainda que um estimador de Kernel de densidade seja estimado e plotado no mesmo gráfico. Calcula ainda gráficos de probabilidade e os qqplots para as mesmas distribuições utilizadas no comando *histogram*. Na seqüência apresentamos os principais comandos do *proc univariate*, descrevendo suas principais opções.

Vamos ilustrar a utilização do *proc univariate* com um conjunto de dados de feijão, onde foram avaliadas as produtividades em g/planta de 20 plantas da geração $F_2$. Neste programa optamos por apresentar no mesmo histograma o estimador kernel com suas três opções (normal, quadratic e triangular) e com o tamanho do parâmetro de suavização de cada igual a 1. A opção $c = 1 \quad 1 \quad 1$ é que definiu este valor para cada método. Escolhemos a opção normal para ajustar o polígono da normal ao histograma e também traçamos os gráficos da probabilidade e dos quantis utilizando os comandos *qqplot* e *probplot*.

---

/*Exemplo de um arquivo de dados com $n = 20$ plantas $F_2$ de feijão com o peso de cada uma em g/plantas.*/
data feijao;
input prod @@;

```
cards;
   1.38    3.65    3.78    3.87
   4.14    4.54    5.64    5.67
   6.23    6.79    8.21    9.79
  12.13   12.56   13.19   15.60
  17.12   19.68   21.26   24.57
;
proc univariate data=feijao;
    var prod;
    histogram prod/ normal kernel(c=1 1 1 k=normal quadratic triangular);
    probplot prod/normal;
    qqplot prod/normal;
run;
```

Ao observamos os resultados, podemos verificar que embora as evidências descritivas não sejam muito fortes, não parece haver uma boa concordância da distribuição dos dados amostrais com a distribuição normal. Testes formais precisam ser feitos para que haja uma confirmação ou não destas evidências descritivas. Um outro comentário simples que gostaríamos de fazer neste instante diz respeito à forma que devemos sumariar os resultados descritivos de posição e dispersão em um trabalho científico. Em geral, se a distribuição é simétrica utilizamos a média como medida de posição. Associada a esta medida devemos apresentar uma medida de dispersão. Podemos escolher o desvio padrão ou o erro padrão, conforme o objetivo do trabalho. Se queremos retratar a variabilidade dos dados populacionais em relação a média desta população, devemos utilizar o desvio padrão como uma estimativa desta medida. O coeficiente de variação também pode ser utilizado se pretendemos apresentar esta variabilidade em uma escala relativa e não absoluta. Se por outro lado desejamos caracterizar a precisão com que a média populacional foi estimada, ou seja, a precisão da estimativa obtida, deveremos reportar o erro padrão da média.

A forma como estas medidas devem ser apresentadas também é alvo de muita polêmica no meio científico. Muitas críticas surgem quando apresentamos em uma tabela ou no texto, os resultados por $\bar{X}_{\cdot} \pm S$ ou por $\bar{X}_{\cdot} \pm S_{\bar{X}_{\cdot}}$. O uso do $\pm$ é muito criticado, pois gera ambigüidade dos resultados e das interpretações. Isto porque pode dar idéia de que o resultado se trata de

um intervalo de confiança, o que não é verdade. Assim, é preferível que os resultados sejam apresentados por $\bar{X}_{\cdot}(S)$ ou por $\bar{X}_{\cdot}\left(S_{\bar{X}_{\cdot}}\right)$. Em ambos os casos deve ficar claro para o leitor que se trata da estimativa da média seguida, entre parênteses, pelo desvio padrão ou pelo erro padrão. Não temos restrições ao uso particular de um destes estimadores: coeficiente de variação, desvio padrão ou erro padrão. Isto porque podemos calcular a partir de um deles os demais. Então se torna preponderante a apresentação do tamanho da amostra $n$ utilizado no experimento ou no levantamento amostral (Ferreira, 2005[3]).

Podemos utilizar ainda o *proc summary* para obtermos algumas estatísticas descritivas. Este procedimento é interessante por realizar estimação por intervalo de médias de populações normais. Assim, podemos complementar a informação do *proc univariate* que realiza testes de hipóteses paramétricos e não-paramétricos sobre a média. Utilizamos os dados de feijão anteriormente apresentados para mostrar uma aplicação do *proc summary* e de sua sintaxe básica. Por *default* este procedimento não produz *output*. Devemos utilizar a opção *print* para obtermos o resultado na janela de saída. As estatísticas descritivas que almejamos devem ser solicitadas ao procedimento. A lista de opções é: alpha, clm, range, css, skewness (skew), cv, stddev (std), kurtosis (kurt), stderr, lclm, sum, max, sumwgt, mean, uclm, min, n, uss, var, nmiss. As opções de quantis são: median (p50), q3 (p75), p1, p90, p5, p95, p10, p99, q1 (p25) e qrange. A opção qrange é a amplitude interquartílica: $p75 - p25$. O exemplo final com algumas das opções é:

---

/*Exemplo de um arquivo de dados com $n = 20$ plantas $F_2$ de feijão com o peso de cada uma em g/plantas.*/
data feijao;
input prod @@;
cards;
1.38 3.65 3.78 3.87
4.14 4.54 5.64 5.67
6.23 6.79 8.21 9.79
12.13 12.56 13.19 15.60
17.12 19.68 21.26 24.57

```
;
proc summary data=feijao print range css skew cv std kurt stderr sum max mean min n
uss var nmiss p5 p95 qrange;
    var prod;
run;
```

## 2.2 Estimação de Parâmetros

Vamos apresentar vários procedimentos para estimação dos principais parâmetros de uma população. Nesta seção vamos considerar a estimação de média, proporção, variância, desvio padrão, coeficiente de variação e diferenças de médias.

### 2.2.1 Estimação de Médias, Desvio Padrão e Variâncias

Vamos apresentar o procedimento SAS para estimação intervalar de médias de uma população normal. Para isso vamos utilizar novamente o proc summary. Neste caso utilizamos a opção clm (*confidence limits for the mean*) e a opção alpha para determinarmos o valor do coeficiente de confiança que é dado por $1-\alpha$. Por *default* o SAS utiliza $\alpha = 0,05$. O intervalo de confiança para a média de uma normal é dado por:

$$IC_{1-\alpha}(\mu) : \bar{X}_{\cdot} \pm t_{\alpha/2;\nu}\frac{S}{\sqrt{n}}, \tag{2.6}$$

em que $t_{\alpha/2;\nu}$ é o quantil superior $100\alpha/2\%$ da distribuição $t$ de Student com $\nu = n-1$ graus de liberdade.

O programa SAS para realizarmos a estimação por intervalo para a média de uma população normal, considerando os dados de feijão como exemplo, está apresentado na seqüência. Vamos a partir deste instante fazer algumas simplificações nos programas, apresentando somente a parte contendo os comandos de interesse e omitindo a parte de entrada de dados. Só apresentaremos a parte de entrada de dados quando se tratar de conjuntos de valores que ainda não foram descritos anteriormente. O programa simplificado é:

```
/*Exemplo da utilização dos dados de feijão para obtermos o intervalo de 95% para a
média.*/
proc summary data=feijao print alpha=0.05 mean stderr n std clm;
    var prod;
run;
```

Também podemos utilizar o *proc univariate* para realizarmos intervalo de confiança para média, desvio padrão e variância de uma população normal utilizando a opção *cibasic*. O intervalo de confiança para a variância de uma população normal é dado por:

$$IC_{1-\alpha}(\sigma^2) : \left[ \frac{(n-1)S^2}{\chi^2_{\alpha/2;\nu}}; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;\nu}} \right], \tag{2.7}$$

em que $\chi^2_{\alpha/2;\nu}$ e $\chi^2_{1-\alpha/2;\nu}$ são os quantis superiores $100\alpha/2\%$ e $100(1-\alpha/2)\%$ da distribuição qui-quadrado com $\nu = n-1$ graus de liberdade, respectivamente.

O intervalo de confiança para o desvio padrão populacional ($\sigma$) é obtido calculando a raiz quadrada dos limites do intervalo de confiança para variância. O programa SAS para obtenção destes intervalos, utilizando os dados do feijão, é dado por:

```
/*Exemplo da utilização dos dados de feijão para obtermos o intervalo de 95% para a
média, desvio padrão e variância.*/
proc univariate data=feijao alpha=0.05 cibasic;
    var prod;
run;
```

### 2.2.2 Estimação de Proporções

Para estimarmos por intervalo proporções binomiais podemos utilizar a aproximação normal em grandes amostras e o intervalo de confiança exato.

Estes métodos serão implementados no *proc iml* para ilustrarmos a sua utilização e a de algumas funções do SAS para obtenção de quantis dos modelos probabilísticos necessários em cada caso. Dada uma amostra de tamanho $n$ de eventos Bernoulli independentes e com probabilidade de sucesso constante $p$, em que exatamente $y$ sucessos foram observados, o intervalo de confiança normal aproximado para $p$ é dado por:

$$IC_{1-\alpha}(p) : \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{2.8}$$

em que $\hat{p} = y/n$ é estimador pontual de $p$ e $z_{\alpha/2}$ é o quantil superior $\alpha/2$ da distribuição normal padrão.

O intervalo de confiança exato para as proporções binomiais deve ser utilizado principalmente se $n$ for pequeno e se $p$ se afastar muito de $1/2$. Este intervalo é baseado na relação da binomial com a beta incompleta e portanto com a distribuição $F$. O intervalo de confiança exato para as proporções binomiais é dado por:

$$IC_{1-\alpha}(p) : \left[ \frac{1}{1 + \dfrac{(n-y+1)F_{\alpha/2;2(n-y+1),2y}}{y}}; \frac{1}{1 + \dfrac{n-y}{(y+1)F_{\alpha/2;2(y+1),2(n-y)}}} \right], \tag{2.9}$$

em que $F_{\alpha/2;\nu_1,\nu_2}$ é o quantil superior $100\alpha/2\%$ da distribuição $F$ com $\nu_1$ e $\nu_2$ graus de liberdade.

Implementamos um programa no *proc iml* utilizando os recursos da linguagem SAS, onde o usuário deve trocar os valores de $y$ e de $n$ apresentados no programa, conforme forem os resultados de sua pesquisa. O valor de $\alpha$ também deve ser alterado se tivermos interesse em outro coeficiente de confiança do que aquele utilizado no programa.

---

/*Utilização do Proc IML para a obtenção de intervalos exato e aproximado para o parâmetro binomial $p$ em uma amostra de tamanho $n$, com coeficiente de confiança de $(1-\alpha)100\%$, onde foram observados $y$ sucessos.*/

```
proc iml;
   /*Intervalo de confiança exato*/;
```

```
y=2;n=10;p=y/n;alpha=0.05;
if y=0 then F1=0;
else F1=Finv(1-alpha/2,2*(n-y+1),2*y);
if y=n then F2=0;
else F2=Finv(1-alpha/2,2*(y+1),2*(n-y));
if y=0 then LIE=0;
else LIE=1/(1+(n-y+1)*F1/y);
if y=n then LSE=1;
else LSE=1/(1+(n-y)/(F2*(y+1)));
print "IC exato para p: " LIE LSE " alpha: " alpha " phat: " p;
/*Intervalo de confiança normal aproximado*/;
z=probit(1-alpha/2);
LIap=p-z*(p*(1-p)/n)**0.5;
LSap=p+z*(p*(1-p)/n)**0.5;
print "IC aproximado para p: " LIap LSap " alpha: " alpha;
quit;
```

### 2.2.3 Estimação de Coeficientes de Variação

Para estimar o intervalo de confiança do coeficiente de variação populacional de uma normal, seja $\hat{\kappa} = S/\bar{X}_{\cdot}$ o estimador do coeficiente de variação. O intervalo aproximado proposto por Vangel (1996)[15] é dado por:

$$IC_{1-\alpha}(\kappa): \begin{cases} LI = \dfrac{\hat{\kappa}}{\sqrt{\left(\dfrac{\chi^2_{\alpha/2}+2}{\nu+1}-1\right)\hat{\kappa}^2+\dfrac{\chi^2_{\alpha/2}}{\nu}}} \\ \\ LS = \dfrac{\hat{\kappa}}{\sqrt{\left(\dfrac{\chi^2_{1-\alpha/2}+2}{\nu+1}-1\right)\hat{\kappa}^2+\dfrac{\chi^2_{1-\alpha/2}}{\nu}}}, \end{cases} \tag{2.10}$$

em que $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$ são os quantis superiores $100\alpha/2\%$ e $100(1-\alpha/2)\%$ da distribuição de qui-quadrado com $\nu = n-1$ graus de liberdade.

Novamente utilizamos o *proc iml* para obter o intervalo de confiança para o coeficiente de variação, dadas as estimativas da variância e da média e o tamanho da amostra. O programa resultante é dado por:

```
/*Utilização do Proc IML para a obtenção de intervalos de confiança para o coeficiente
de variação em uma amostra de tamanho n, com coeficiente de confiança de (1-alpha)100%,
sendo dado a média e variância amostral.*/
proc iml;
    /*Intervalo de confiança para o CV*/
    xbar=194.8333;S2=26.2947;n=6;alpha=0.05;
    khat=S2**0.5/xbar;
    qui1=cinv(1-alpha/2,n-1);
    qui2=cinv(alpha/2,n-1);
    LICV=khat/(((qui1+2)/n-1)*khat**2+qui1/(n-1))**0.5;
    LSCV=khat/(((qui2+2)/n-1)*khat**2+qui2/(n-1))**0.5;
    print "IC para o CV: " LICV LSCV " alpha: " alpha " khat: " khat;
quit;
```

### 2.2.4 Diferença de Duas Médias Independentes

Esta é uma situação de muito interesse para os pesquisadores, pois é muito comum obter amostras independentes de duas populações. O objetivo é obter o intervalo de confiança para a diferença das médias $\mu_1 - \mu_2$ das duas populações. Algumas suposições são feitas para a utilização dos procedimentos estatísticos adequados. Inicialmente pressupomos que ambas as populações possuem distribuição normal com médias $\mu_1$ e $\mu_2$ e variâncias $\sigma_1^2$ e $\sigma_2^2$, respectivamente. Ao obtermos as amostras aleatórias de tamanhos $n_1$ e $n_2$ das populações 1 e 2, respectivamente, devemos supor independência entre as observações das diferentes amostras e também das observações dentro das duas amostras. Finalmente, supomos que as variâncias das duas populações são homogêneas, ou seja, que $\sigma_1^2 = \sigma_2^2$.

Sejam $\bar{X}_1$ e $\bar{X}_2$ os estimadores das médias da populações 1 e 2 e $S_1^2$ e $S_2^2$ os estimadores das variâncias populacionais obtidos em amostras de tamanho $n_1$ e $n_2$, respectivamente, então duas situações distintas podem ser consideradas. A primeira quando $\sigma_1^2 = \sigma_2^2$ e a segunda quando $\sigma_1^2 \neq \sigma_2^2$. Estas duas situações estão destacadas na seqüência.

a. Se $\sigma_1^2 = \sigma_2^2$: O intervalo de confiança quando as variâncias são homogêneas é dado por:

$$IC_{1-\alpha}(\mu_1 - \mu_2) : \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2;\nu}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \tag{2.11}$$

em que $t_{\alpha/2;\nu}$ é o quantil superior $\alpha/2$ da distribuição $t$ de Student com $\nu = n_1 + n_2 - 2$ graus de liberdade e $S_p^2$ é a variância combinada (*pooled*) dada por:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{2.12}$$

b. Se $\sigma_1^2 \neq \sigma_2^2$: Neste caso a distribuição $t$ de Student não é mais exata para obtermos o intervalo de confiança. No entanto, esta distribuição é utilizada de forma aproximada, ajustando somente os graus de liberdade. Este ajuste aos graus de liberdade é atribuído a Satterthwaite (1946)[11]. O intervalo de confiança aproximado é dado por:

$$IC_{1-\alpha}(\mu_1 - \mu_2) : \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2;\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}. \tag{2.13}$$

Neste caso os graus de liberdade $\nu$ para a obtenção do quantil superior da distribuição $t$ de Student é ajustado (Satterthwaite, 1946) por:

$$\nu \cong \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}. \tag{2.14}$$

O procedimento mais apropriado para estimar duas médias populacionais por intervalo requer que tenhamos o conhecimento sobre a homogeneidade ou não das variâncias das duas populações. Como se tratam de parâmetros desconhecidos podemos inferir apenas a este respeito. Para isso podemos utilizar o teste F. Um artifício que utilizamos é considerar a variância maior no numerador da expressão, multiplicando o valor encontrado por 2. Assim, para testarmos a hipótese $H_0 : \sigma_1^2 = \sigma_2^2$ calculamos:

$$F_c = \frac{S^2_{Maior}}{S^2_{Menor}} \tag{2.15}$$

e o valor-p é determinado por $2 \times P(F > F_c)$. Se valor-p for menor ou igual ao valor nominal $\alpha$, rejeitamos $H_0$. O programa SAS resultante deste procedimento é dado por:

---

```
/*Utilização do Proc IML para a obtenção de intervalos de confiança para o diferença
de duas médias, testando antes a igualdade de variâncias, utilizando uma confiança de
(1 − α)100%.*/
proc iml;
   /*Dados amostrais H pop. 1 e M= pop 2*/;
   h={72,88,62,95};m={48.5,86,79,53};
   n1=nrow(h);n2=nrow(m);alpha=0.05;
   xb1=sum(h)/n1;xb2=sum(m)/n2;
   s21=(t(h)*h-sum(h)**2/n1)/(n1-1);
   s22=(t(m)*m-sum(m)**2/n2)/(n2-1);
   /*teste de hipótese*/
   smaior=max(s21,s22);
   smenor=min(s21,s22);
   if s21>s22 then v1=n1-1;
   else v1=n2-1;
   if s21>s22 then v2=n2-1;
   else v2=n1-1;
   Fc=smaior/smenor;
   pval=2*(1-probF(fc,v1,v2));
   print “FC ” fc “ alpha: ” alpha “ pval: ” pval;
   if pval>alpha then
   do;
      sp=((n1-1)*s21+(n2-1)*s22)/(n1+n2-2);
      t=tinv(1-alpha/2, n1+n2-2);
      LIE=xb1-xb2-t*(sp*(1/n1+1/n2))**0.5;
      LSE=xb1-xb2+t*(sp*(1/n1+1/n2))**0.5;
      print “LI ” LIE “LS ” LSE;
   end;
   else do;
      v=(s21/n1+s22/n2)**2/((s21/n1)**2/(n1-1)+(s22/n2)**2/(n2-1));
      t=tinv(1-alpha/2, v);
      LIA=xb1-xb2-t*(s21/n1+s22/n2)**0.5;
```

```
        LSA=xb1-xb2+t*(s21/n1+s22/n2)**0.5;
        print "LI " LIA "LS " LSA;
    end;
quit;
```

### 2.2.5 Estimação da Diferenças de Duas Médias Em Dados Emparelhados

Em muitas ocasiões experimentais nos deparamos com a necessidade de inferir sobre o efeito de algum medicamento, fertilizante, fungicida entre outros tratamentos. Realizamos experimentos onde temos o maior grau de controle local possível, ou seja, mensuramos os indivíduos ou as unidades experimentais antes da aplicação do tratamento e após a sua aplicação. Neste experimento temos a mesma unidade experimental servindo de controle local. Isto torna este experimento mais eficiente que o experimento em que as amostras são tomadas de forma independente na população tratada e não tratada. Uma alternativa a este delineamento experimental é possível de ser obtida se utilizarmos duas parcelas experimentais locadas e submetidas sob as mesmas condições e sorteamos uma para receber o tratamento e a outra para não recebê-lo.

Se $X_i$ e $Y_i$ são as respostas mensuradas antes e após a aplicação do tratamento, respectivamente, na $i-$ésima unidade amostral, para $i = 1, 2, \cdots, n$, então podemos gerar a variável aleatória $d_i = Y_i - X_i$. A estimação pontual do valor esperado desta variável aleatória $E(d_i) = \delta = \mu_Y - \mu_X$ pode ser feita por:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}. \tag{2.16}$$

O estimador da variância populacional das diferenças é dado por:

$$S_d^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right]. \tag{2.17}$$

Assim, o intervalo de confiança pode ser obtido por:

$$IC_{1-\alpha}(\delta) : \bar{d} \pm t_{\alpha/2;\nu=n-1}\frac{s_d}{\sqrt{n}}. \tag{2.18}$$

O artifício que usaremos para obter o intervalo de confiança almejado consiste em considerar com um conjunto de dados, para o qual especificamos em cada parcela a variável $X$ e a variável $Y$ (antes e após). Em seguida utilizando o processo de transformação de variáveis descritos na seção 1.2 devemos gerar $D = Y - X$. Finalmente utilizamos o *proc summary* ou o *proc univariate* para obtermos o intervalo de confiança para a média. No programa seguinte descrevemos este processo com a utilização do *proc summary*. Este exemplo refere-se a produção de leite média diária em kg de todos os animais de uma fazenda em uma amostra de 6 fazendas da região de Marechal Cândido Rondom antes $X$ e após $Y$ um plano governamental. A questão era responder se o plano foi eficiente e se sim, qual foi o aumento na produção média diária de leite dos animais em kg. Tomamos apenas uma parte dos dados $n = 6$ para ilustrar de forma didatica esta situação. O programa SAS é:

---

```
/*Utilização do Proc Summary para a obtenção de intervalos de confiança para o dife-
rença de duas médias emparelhadas, utilizando uma confiança de (1 − α)100%.*/
data leite;
input X Y;
d=Y-X;
cards;
12.00 12.56
11.58 13.98
11.67 14.23
12.32 14.56
```

```
11.23 13.71
11.25 16.78
;
proc summary data=leite print alpha=0.05 n mean std stderr clm;
    var d;
run;quit;
```

## 2.3 Testes de Hipóteses

Neste seção trataremos dos testes de hipóteses sobre os principais parâmetros de uma ou duas populações. Antes de apresentarmos os métodos e recursos computacionais para realizarmos os testes de hipóteses, devemos atentar para o fato de que existe uma relação estreita entre os procedimentos de estimação e decisão.

Se já temos um intervalo de confiança construído, podemos testar uma hipótese bilateral apenas verificando se este intervalo contém o valor hipotético. Caso o valor hipotético pertença ao intervalo de confiança não temos evidências significativas para rejeitar a hipótese nula. Por outro lado, se o valor hipotético não pertence ao intervalo de confiança, podemos concluir a favor da hipótese alternativa, rejeitando a hipótese nula. Assim, vamos apresentar somente os procedimentos para testarmos médias de uma população e de duas, sejam elas independentes ou emparelhadas. Testes sobre variâncias, desvios padrões ou coeficientes de variação poderão ser realizados com o uso dos intervalos de confiança apresentados anteriormente.

### 2.3.1 Teste Sobre Médias

Para testarmos hipóteses sobre médias normais devemos utilizar o teste $t$ de Student. Assim, para testarmos a hipótese nula $H_0 : \mu = \mu_0$ utilizamos os seguintes procedimentos. Inicialmente calculamos a estatística do teste por

$$t_c = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}. \tag{2.19}$$

Se a hipótese alternativa for do tipo bilateral $H_1 : \mu \neq \mu_0$, calculamos o valor-p por $P(t > |t_c|)$; se a hipótese alternativa for unilateral do tipo $H_1 : \mu > \mu_0$, calculamos o valor-p por $P(t > t_c)$; e se a hipótese alternativa for unilateral do tipo $H_1 : \mu < \mu_0$, calculamos o valor-p por $P(t < t_c)$. Finalmente, confrontamos o valor-p com o valor nominal do nível de significância $\alpha$. Se o valor-p for inferior ou igual a $\alpha$, devemos rejeitar a hipótese nula neste nível de significância; caso contrário, não devemos rejeitar $H_0$.

Se a distribuição dos dados não for normal podemos utilizar dois testes não-paramétricos: o teste do sinal e o teste dos postos com sinais de Wilcoxon. Vamos descrever o teste do sinal com detalhes e realizar apenas uma breve descrição do teste de Wilcoxon.

Para aplicarmos o teste do sinal, inicialmente calculamos o número de sinais positivos e negativos para a diferença de cada observação amostral com o valor hipotético. Se $X_i - \mu_0$ representa esta diferença, então podemos definir $n_+$ como o número de observações para as quais $X_i > \mu_0$ (sinais positivos) e $n_-$ com o número de observações para as quais $X_i < \mu_0$ (sinais negativos). Devemos desprezar todas as observações para as quais $X_i = \mu_0$. Assim, o número de observações efetivas amostrais é $n_e = n_+ + n_-$. Ao realizarmos este teste estamos supondo que se a hipótese nula for verdadeira, o número de sinais positivos deve ser igual ao número de sinais negativos. Aplicamos, então, um teste binomial para $p = 1/2$, em que $p$ é a proporção de sinais positivos ou negativos. Assim, a estatística do teste sinal é dada por:

$$M_c = \frac{n_+ - n_-}{2}. \tag{2.20}$$

O valor-p é calculado utilizando a distribuição binomial em um teste bilateral por:

$$valor - p = P(M > |M_c|) = \left(\frac{1}{2}\right)^{(n_e-1)} \sum_{j=0}^{\min(n_+,n_-)} \binom{n_e}{j}. \tag{2.21}$$

O valor-p é confrontado com o valor de $\alpha$ e tomamos a decisão de rejeitar ou não a hipótese nula utilizando procedimentos semelhantes ao que apresentamos anteriormente para o teste $t$.

A estatística do teste do sinal com postos de Wilcoxon é obtida calculando-se todos os desvios das observações em relação ao valor hipotético e tomando-se os postos dos valores destas diferenças em módulo $d_i = |X_i - \mu_0|$. Se algum valor amostral for igual a zero, devemos eliminá-lo da amostra, como fazemos no teste do sinal. Se houver empates, tomamos a média dos postos que seriam atribuídos a estas observações empatadas. Retornamos os sinais de $X_i - \mu_0$ aos postos das diferenças e somamos os valores positivos. Esta soma é representada por $W^+$ e é a estatística do teste. Os valores-p podem ser obtidos utilizando-se uma aproximação normal ou a distribuição nula da estatística $W^+$, derivada pela atribuição de sinais positivos ou negativos a cada posto amostral em todas as combinações possíveis. O teste de Wilcoxon é, em geral, mais poderoso do que o teste do sinal. Nenhum detalhe adicional será apresentado neste material.

Podemos utilizar o *proc univariate* para testarmos hipóteses sobre a média de uma população. O *proc univariate* utiliza as três opções apresentadas nesta seção para realizarmos o teste de hipótese. Devemos optar pelo teste mais apropriado conforme for o caso. Esta escolha deve ser pautada no atendimento ou não das pressuposições básicas de cada teste. Um procedimento SAS é apresentado na seqüência para testarmos a hipótese da igualdade da média do peso dos coelhos híbridos Norfolk abatidos aos 90 dias a $2,50$ kg, ou seja, para testarmos $H_0 : \mu = 2,50$. Se várias variáveis são apresentadas no comando *var*, devemos utilizar a opção $mu0 = 0.5\ 2.5\ \cdots$, indicando que o valor sob $H_0$ para a primeira variável é $0,5$, para a segunda é $2,5$ e assim sucessivamente até completar o número de variáveis do comando *var*. O programa resultante é:

---

/*Utilização do *Proc Univariate* para testarmos a hipótese sobre a média de uma população normal e não normais (testes não-paramétricos). Utilizamos o exemplo dos coelhos Norfolk para ilustrar os testes.*/
data coelhos;
input peso @@;
cards;
2.50 2.58 2.60 2.62 2.65
2.66 2.58 2.70 2.55 2.57
2.70 2.62 2.59

```
2.54 2.53 2.20
;
proc univariate data=coelhos mu0=2.5 alpha=0.05;
    var peso;
run;quit;
```

### 2.3.2 Teste Sobre Médias de Duas Populações Emparelhadas

Quando temos dados emparelhados, antes e após a aplicação de um tratamento podemos estar interessados em testes de hipóteses sobre o efeito deste tratamento. Podemos utilizar o mesmo procedimento descrito anteriormente para média e assim testar hipóteses sobre o efeito do tratamento. A hipótese nula de interesse é dada por $H_0 : \delta = \delta_0$. Podemos utilizar o teste $t$ de Student se as variáveis $(X_i, Y_i)$ tiverem distribuição normal bivariada ou, em caso contrário, os testes não-paramétricos do sinal e do sinal com postos de Wilcoxon.

Seja $d_i = Y_i - X_i$ a diferença entre a observação da i-ésima unidade amostral após $Y_i$ e antes $X_i$ da aplicação do tratamento, sendo $i = 1, 2, \cdots n$. Sejam $\bar{d}$ e $S_d^2$ a média e a variância amostral destas $n$ observações, então a estatística do teste da hipótese $H_0 : \delta = \delta_0$ supondo normalidade bivariada é dado por:

$$t_c = \frac{\bar{d} - \delta_0}{\frac{S_d}{\sqrt{n}}}, \tag{2.22}$$

que segue a distribuição $t$ de Student com $\nu = n - 1$ graus de liberdade sob a hipótese nula.

O teste do sinal é obtido contando-se o número de vezes que $d_i > \delta_0$ e desprezando-se os casos em que $d_i = \delta_0$. As expressões 2.20 e 2.21 são usadas para testar a hipótese de interesse. O teste do sinal com postos de Wilcoxon também é obtido da mesma forma considerando tanto o posto da diferença $d_i - \delta_0$ considerada em módulo, quanto o sinal da diferença. Como se trata apenas de uma aplicação do mesmo procedimento adaptado para

esta situação, não faremos nenhum comentário adicional, por julgarmos suficiente o que já abordamos.

A seguir detalharemos o programa SAS para aplicar o teste de avaliação da eficiência de um plano governamental no aumento da média dos índices zootécnicos da região de Marechal Cândido Rondom. A produção média diária de seis fazendas foi avaliadas antes ($X$) e após ($Y$) o plano governamental. Inicialmente criamos uma variável com a diferença e então utilizamos o *proc univariate* da mesma forma que fizemos nos testes de hipóteses sobre a média de uma população. Neste exemplo, a hipótese nula consiste na afirmativa que o plano não foi eficiente, ou seja, $H_0 : \delta = \delta_0 = 0$. Assim, ao utilizarmos o *proc univariate* devemos especificar a hipótese com a opção *mu0=0* ou simplesmente não especificar nada, pois o valor 0 é o default deste procedimento. O programa resultante é dado por:

---

```
/*Utilização do Proc univariate para a testarmos a hipótese de não haver efeito do plano
governamental panela cheia na melhoria da produtividade leiteira das fazendas da cidade
de Marechal Cândido Rondom no Paraná.*/
data leite;
input X Y;
d=Y-X;
cards;
12.00 12.56
11.58 13.98
11.67 14.23
12.32 14.56
11.23 13.71
11.25 16.78
;
proc univariate data=leite mu0=0;
    var d;
run;quit;
```

---

Podemos utilizar um procedimento especializado do SAS para aplicar o teste de hipótese sobre a diferença de duas médias emparelhadas. Este procedimento é o *proc ttest*. Uma vantagem deste procedimento é podermos

obter, além do teste de hipótese, o intervalo de confiança para a diferença de médias e para o desvio padrão da diferença. Utilizamos a opção $H0 = \delta_0$ para especificarmos o valor nulo da hipótese. O programa ilustrativo desta situação é dado por:

---

```
/*Utilização do Proc ttest para testarmos a hipótese de não haver efeito do plano governamental panela cheia na melhoria da produtividade leiteira das fazendas da cidade de Marechal Cândido Rondom no Paraná.*/
data leite;
input X Y;
cards;
12.00 12.56
11.58 13.98
11.67 14.23
12.32 14.56
11.23 13.71
11.25 16.78
;
proc ttest data=leite h0=0;
    paired y*x;
run;quit;
```

---

### 2.3.3 Teste Sobre Médias de Duas Populações Independentes

Finalmente podemos testar a hipótese da igualdade de duas médias populacionais independentes. Para este caso o SAS possui um procedimento especializado, o *proc ttest*. Conforme já apresentamos na seção de estimação por intervalo, devemos inicialmente aplicar o teste de igualdade de variâncias e de acordo com os resultados obtidos, escolhemos entre o teste $t$ de Student exato ou aproximado. O teste exato ocorre quando as variâncias são consideradas homogêneas; o teste é aproximado quando as variâncias são heterogêneas. Devemos neste último caso utilizar o ajuste de graus de liberdade pelo procedimento de Satterthwaite (1946)[11] ou o procedimento

de Cochran e Cox que aproxima o nível de probabilidade da estatística $t$ de Student aproximada.

Vamos apresentar na seqüência o *proc ttest* com o objetivo de ilustra sua utilização. Para isso, um exemplo em dois grupos de alunos foram avaliados com relação ao peso em kg e a altura em m. Os grupos referem-se aos alunos que sentam na bancada da direita (grupo 1) e da esquerda (grupo 2) do laboratório de informática. A primeira turma desta disciplina foi amostrada para esta finalidade. Esperamos a princípio que não haja diferenças significativas entre os dois grupos, uma vez que a distribuição é completamente aleatória nas duas bancadas da sala de aula.

Devemos fazer um conjunto de dados criando uma variável para identificarmos os grupos. Esta variável tem que ter sempre dois níveis para podermos utilizar o *proc ttest*. Sejam $\bar{X}_1$ e $\bar{X}_2$ as médias das amostras aleatórias de tamanhos $n_1$ e $n_2$, respectivamente, retiradas das populações 1 e 2. Sejam $S_1^2$ e $S_2^2$ as variâncias amostrais relativas às populações 1 e 2. Pressupomos que as amostras sejam aleatórias e independentes e que a distribuição das duas populações seja normal.

Inicialmente devemos testar a hipótese sobre a igualdade das variâncias $H_0 : \sigma_1^2 = \sigma_2^2$. Assim, de acordo com este teste devemos aplicar o teste de igualdade da diferença das médias populacionais a um valor de interesse, ou seja, $H_0 : \mu_1 - \mu_2 = \delta_0$ utilizando os seguintes procedimentos:

a) Se $\sigma_1^2 = \sigma_2^2$:

   Neste caso, o teste de igualdade da diferença das médias populacionais a um valor de interesse é exato e a estatística do teste, dada por

$$t_c = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \tag{2.23}$$

   segue a distribuição $t$ de Student com $\nu = n_1 + n_2 - 2$ graus de liberdade. O significado de $S_p^2$ foi apresentado na equação 2.12.

b) Se $\sigma_1^2 \neq \sigma_2^2$:

   Neste caso, a estatística do teste não segue de forma exata a distribuição $t$ de Student. Então, ajustamos os graus de liberdade pelo procedimento

de Satterthwaite (1946)[11] ou ajustamos as probabilidades pelo procedimento de Cochran e Cox. A estatística do teste dada por

$$t_c = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2.24}$$

segue aproximadamente a distribuição $t$ de Student com $\nu$ graus de liberdade obtidos com o uso da expressão 2.14.

Para utilizarmos o *proc ttest* devemos especificar o valor $\delta_0$. Isto é feito utilizando a opção $H0 = \delta_0$. A opção *Cochran* também foi utilizada. Devemos, no entanto, alertar o leitor que, via de regra, os dois procedimentos utilizados para ajustar os graus de liberdade ou as probabilidades, fornecem resultados similares dos testes. Além disso, a decisão tomada, em geral, é a mesma. O programa SAS utilizando o exemplo dos grupos de alunos é dado por:

```
/*exemplo do uso do proc ttest para duas amostras independentes.*/
data sala;
input grupo peso alt;
cards;
1 48.5 1.58
1 53.0 1.60
1 86.0 1.83
1 79.0 1.69
2 62.0 1.72
2 95.0 1.93
2 88.0 1.80
2 72.0 1.80
;
proc ttest data=sala cochran h0=0;
    class grupo;
    var peso alt;
run;
```

Devemos especificar no comando *class* a variável com dois níveis que são usados para identificar as populações. Devemos também determinar quais

variáveis vamos analisar com o comando *var* e o valor hipotético. Infelizmente o SAS não permite especificar um valor diferente para cada variável com o comando *H0*. Se quisermos testar um valor diferente para cada variável, devemos fazer vários comandos repetidos, como no programa anterior, especificando um valor hipotético diferente para cada variável. Por default o *proc ttest* utiliza o valor zero se nada for especificado. Obtivemos para ambas variáveis resultados não significativos para os testes da igualdade variâncias e de médias dos dois grupos, como era esperado.

O *proc ttest* nos permite calcular o intervalo de confiança para a média de cada população e para a diferença de médias. Também fornece o intervalo de confiança para as variâncias. No entanto, o intervalo de confiança da diferença de duas médias deste procedimento do SAS ignora completamente o teste de igualdade de variâncias e estima a diferença de duas médias por intervalo utilizando o procedimento de quando as variâncias são homogêneas. Assim, se o teste de homogeneidade de variâncias for rejeitado, o intervalo de confiança fornecido é via de regra muito impreciso e deve ser desconsiderado. Recomendamos o uso do programa utilizando o *proc iml* que fornecemos anteriormente.

### 2.3.4 Teste de Normalidade

O SAS nos permite realizar teste de normalidade para os dados amostrais coletados em $n$ unidades. Anteriormente já apresentamos alguns destes testes quando utilizamos o comando *histogram prod/normal;* no *proc univariate*. Os testes aplicados no SAS são Kolmogorov-Smirnov, Cramer-von Mises e Anderson-Darling. Também é possível chamar o teste de normalidade sem solicitar o histograma e a estimação dos parâmetros da normal. Podemos utilizar a seguinte linha de comando: *proc univariate data=feijao normal;*. Assim, teremos os mesmos testes de normalidade, incorporando, porém, o poderoso teste de Shapiro-Wilk.

O SAS fornece o valor da estatística de cada teste e o valor-p associado. Se este valor-p for menor do que o valor nominal de significância $\alpha$ previamente adotado, então devemos rejeitar a hipótese nula de normalidade; caso contrário, não haverá evidências significativas neste nível para rejeitar

a hipótese de normalidade.

Devemos enfatizar que o teste de normalidade aplicado no contexto de uma amostra aleatória simples onde não há controle local e efeitos de diferentes tratamentos atuando é totalmente justificável, pois estamos diante de um modelo linear simples do tipo:

$$Y_i = \mu + \epsilon_i,$$

em que $Y_i$ é a observação amostral da i-ésima unidade amostral, $\mu$ a média geral e $\epsilon_i$ o erro associado a i-ésima unidade amostral.

Nos modelos lineares a suposição de normalidade é feita sobre os resíduos e não sobre a variável dependente. Neste modelo linear simples, ao erro de todas as observações é acrescido uma única constante e esta constante somente faz uma translação dos valores de $Y$, não alterando a sua distribuição. Assim, testar a normalidade de $Y$ ou de $\epsilon$ são procedimentos equivalentes. O que muitos pesquisadores fazem muitas vezes dentro do contexto da experimentação é testar a hipótese de normalidade da variável resposta para verificar se esta pressuposição foi atendida, para validar as inferências realizadas. Isto muitas vezes é incorreto, pois se pressupõe resíduos e não variáveis respostas normais. Então, sob um modelo mais complexo, onde existe controle local, efeito de bloco ($\beta_j$) e\ou efeitos de tratamentos ($\tau_i$), a variável resposta $Y$ terá uma distribuição que é na verdade uma mistura de distribuições normais com diferentes médias. Observe que para o modelo linear

$$Y_{ij} = \mu + \beta_j + \tau_i + \epsilon_{ij},$$

a variável $Y_{ij}$ tem a seguinte média: $E(Y_{ij}) = \mu+\beta_j+\tau_i$. Assim, se variarmos a unidade experimental $(i, j)$, teremos diferentes valores médios para $Y_{ij}$. Como supomos independência e homocedasticidade de variâncias, a mistura de distribuições terá diferentes distribuições normais com diferentes médias, mas com a mesma variância. Então, em uma amostra de tamanho $n$, não podemos testar a hipótese de normalidade utilizando os valores de $Y$, mas devemos estimar o erro cuja média é zero e a variância é constante para realizarmos tal teste.

# Capítulo 3

# Regressão Linear

Os modelos de regressão linear desempenham um grande papel nas mais diferentes áreas do conhecimento. Os pesquisadores buscam sempre modelar seus dados por um modelo e então passam a compreender melhor o fenômeno sob estudo. Os modelos lineares são apenas uma das classes utilizadas pelos pesquisadores na compreensão dos problemas de suas pesquisas. A classificação de um modelo como linear é muitas vezes confundida com o tipo de curva matemática que aquele modelo descreve e, ainda, é mal compreendida. Assim, iniciaremos nossa discussão com a classificação de dois modelos como linear ou não-linear. O primeiro modelo é dado por $Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i$, em que $Y_i$ e $X_i^2$ são as variáveis resposta e regressoras, respectivamente; $\beta_0$ e $\beta_1$ são os seus parâmetros; e $\epsilon_i$ é o resíduo ou erro. O segundo modelo é $Y_i = \beta_0 X_i^{\beta_1} + \epsilon_i$. Ambos os modelos descrevem curvas que não são uma reta simples. Esta é uma das causas de confusões na classificação de um modelo como linear. Nestes exemplos, o primeiro modelo é linear e o segundo é não-linear.

Para esclarecermos e definirmos um modelo como linear, devemos apresentar inicialmente um conceito filosófico. Dizemos que um modelo é linear ou não-linear nos parâmetros e com isso não estamos interessado no tipo de curva que a função representa. Formalmente, podemos dizer que um modelo é linear se as derivadas parciais da variável dependente em relação a cada parâmetro não forem funções dos próprios parâmetros. Assim, as derivadas parciais do primeiro modelo são: $\partial Y_i/\partial\beta_0 = 1$ e $\partial Y_i/\partial\beta_1 = X_i^2$.

Como nenhuma das derivadas parciais dependem dos próprios parâmetros, então este modelo é linear. No segundo caso, as derivadas parciais são: $\partial Y_i/\partial\beta_0 = X_i^{\beta_1}$ e $\partial Y_i/\partial\beta_1 = \beta_0 X_i^{\beta_1}\ln(X_i)$. O segundo modelo é não-linear nos parâmetros, pois as duas derivadas parciais são funções dos próprios parâmetros. Bastaria uma de estas derivadas ser função dos parâmetros para classificarmos o modelo como não-linear.

Dois procedimentos, entre outros, podem ser utilizados para analisarmos os modelos lineares e não lineares. Utilizaremos o *proc reg* para os modelos lineares e o *proc nlin* para modelos não-lineares. Neste capítulo estudaremos apenas os modelos lineares nos parâmetros. O *proc reg* é, entre os possíveis procedimentos de regressão do SAS, aquele que tem um amplo propósito, enquanto os demais possuem objetivos mais específicos. Este procedimento permite entre outras as seguintes análises:

- Especificação de múltiplos modelos
- Métodos de seleção de modelos
- Diagnósticos de regressão
- Obtenção de valores preditos
- Diagnose de multicolinearidade
- Gráficos de resíduos

## 3.1 Método dos Quadrados Mínimos

O *proc reg* foi idealizado para ajustar modelos lineares e fornecer várias ferramentas de diagnóstico da qualidade de ajuste. Seja o modelo linear de regressão com $m+1$ parâmetros definido por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + \epsilon_i \tag{3.1}$$

em que $Y_i$ é a i-ésima observação da variável resposta; $X_{hi}$ é i-ésima observação da $h$-ésima variável; $\beta_h$ são os parâmetros do modelo; $\epsilon_i$ é o resíduo de regressão associado a $i$-ésima unidade amostral; $h = 0, 1, 2, \cdots, m$ e

$i = 1, 2, \cdots, n$; $X_{0i}$ é constante com todos os valores iguais a 1; $m$ representa o número de variáveis e $n$ o tamanho da amostra.

O método dos quadrados mínimos é baseado na idéia de minimizar a soma de quadrados dos resíduos dos modelos lineares. Assim, se $Q = \sum_i^n \epsilon_i^2$ é a soma de quadrados de resíduos, o seu valor mínimo deve ser encontrado para obtermos uma solução de quadrados mínimos. Matricialmente temos o modelo 3.1 expresso da seguinte forma:

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \tag{3.2}$$

em que $\underset{\sim}{Y}$ é o vetor de observações de dimensões $n \times 1$; $X$ é a matriz do modelo de dimensões $n \times (m+1)$ das derivadas parciais de $Y_i$ em relação aos parâmetros; $\underset{\sim}{\beta}$ é o vetor de parâmetros $[(m+1) \times 1]$; e $\underset{\sim}{\epsilon}$ é o vetor de resíduos $(n \times 1)$.

Os resíduos podem ser isolados por $\underset{\sim}{\epsilon} = \underset{\sim}{Y} - X\underset{\sim}{\beta}$ e a soma de quadrados do resíduos matricialmente é expressa por:

$$Q = \underset{\sim}{\epsilon}'\underset{\sim}{\epsilon} = \left(\underset{\sim}{Y} - X\underset{\sim}{\beta}\right)' \left(\underset{\sim}{Y} - X\underset{\sim}{\beta}\right)$$

$$Q = \underset{\sim}{\epsilon}'\underset{\sim}{\epsilon} = \left(\underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta}\right)$$

Obtemos as derivadas de $Q$ com relação a $\beta$ e encontramos:

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}} = -2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{\beta}$$

Igualamos a zero e obtemos as conhecidas equações normais (EN) na seqüência. Assim, temos:

$$-2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{\hat{\beta}} = 0$$

$$X'X\underset{\sim}{\hat{\beta}} = X'\underset{\sim}{Y} \tag{3.3}$$

em que $\underset{\sim}{\hat{\beta}}$ é o estimador de mínimos quadrados do parâmetro $\beta$.

A matriz de derivadas parciais ou de modelo $X$, em geral, possui posto coluna completo nos modelos de regressão. Assim, a matriz $X'X$ possui inversa única e a solução do sistema é:

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y} \tag{3.4}$$

O valor esperado de $\underset{\sim}{Y}$ é $E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$. Podemos obter os valores estimados substituindo $\underset{\sim}{\beta}$ por $\underset{\sim}{\hat{\beta}}$. Assim, os valores preditos são dados por:

$$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}} \tag{3.5}$$

É importante obtermos as somas de quadrados do modelo e do resíduo, para aplicar uma análise de variância e realizarmos inferência a respeito do modelo ajustado. Nenhuma pressuposição foi feita até o momento sobre a distribuição dos resíduos, mas se temos a intenção de realizar inferências é necessário pressupormos normalidade e ainda distribuição idêntica e independente de todos os componentes do vetor de resíduos. Podemos estimar $Q$ substituindo $\underset{\sim}{\beta}$ por $\underset{\sim}{\hat{\beta}}$. Obtemos após algumas simplificações:

$$\hat{Q} = \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\hat{\beta}}'X'\underset{\sim}{Y}$$

Assim, podemos interpretar esta expressão da seguinte forma:

$$\text{SQRes} = \text{SQTotal não corrigida} - \text{SQModelo}$$

Assim, a soma de quadrados de modelo é dada por:

$$\text{SQModelo} = \underset{\sim}{\hat{\beta}}'X'\underset{\sim}{Y} \tag{3.6}$$

Os graus de liberdade associado ao modelo é igual ao posto coluna da matriz $X$. Se esta matriz tem posto coluna completo $m+1$, concluímos que a soma de quadrados do modelo está associada a $m+1$ graus de liberdade e a soma de quadrados do resíduo a $n-m-1$ graus de liberdade. O que fazemos é definir sub-modelos a partir do modelo completo com $m+1$

parâmetros. Desta forma podemos definir dois tipos básicos de soma de quadrados: a seqüencial (tipo I) e a parcial (tipo II). Na seqüencial tomamos o modelo completo e o reduzimos eliminando a variável $m$. Obtemos a soma de quadrado do modelo completo, que representamos por $R(\beta_0, \beta_1, \cdots, \beta_m)$, e a do modelo reduzido, representada por $R(\beta_0, \beta_1, \cdots, \beta_{m-1})$. A notação $R$ indica uma redução particular do modelo que estamos abordando. Se tomarmos a diferença da soma de quadrados dos dois modelos teremos $R(\beta_m/\beta_0, \beta_1, \cdots, \beta_{m-1}) = R(\beta_0, \cdots, \beta_m)$-$R(\beta_0, \cdots, \beta_{m-1})$. Se do modelo com $m-1$ variáveis eliminarmos a última e repetirmos este procedimento, teremos a soma de quadrado da $(m-1)$-ésima variável ajustada para todas as outras que a precedem. Se fizermos isso repetidas vezes até reduzirmos o modelo ao termo constante apenas, teremos as somas de quadrados de cada variável ajustada para todas as outras que a precedem, ignorando as variáveis que a sucedem. Esta é a soma de quadrados tipo I ou seqüencial.

Para obtermos as somas de quadrados parciais ou do tipo II, devemos a partir do modelo completo formar um novo modelo eliminando uma das variáveis. A soma de quadrados do modelo reduzido é comparada com a soma de quadrado do modelo completo e a sua diferença é a soma de quadrados do tipo II. Assim, teremos o ajuste de cada variável para todas as outras do modelo. Podemos perceber que as somas de quadrados tipo I e tipo II da $m$-ésima variável são iguais. Via de regra as somas de quadrados tipo I e tipo II não serão iguais para as demais variáveis, a menos de ortogonalidade. Podemos resumir o dois tipos de somas de quadrados conforme esquema apresentado na Tabela 3.1.

Tabela 3.1: Tipos de somas de quadrados de um modelo de regressão contendo $m$ variáveis.

| FV | SQ Tipo I | SQ Tipo II |
|---|---|---|
| $X_1$ | $R(\beta_1/\beta_0)$ | $R(\beta_1/\beta_0, \beta_2, \cdots, \beta_m)$ |
| $X_2$ | $R(\beta_2/\beta_0, \beta_1)$ | $R(\beta_2/\beta_0, \beta_1, \cdots, \beta_m)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $X_m$ | $R(\beta_m/\beta_0, \beta_1, \cdots, \beta_{m-1})$ | $R(\beta_m/\beta_0, \beta_1, \cdots, \beta_{m-1})$ |

Uma forma alternativa bastante útil para podermos obter as somas de

quadrados tipo II é baseada no método da inversa de parte da inversa de Searle (1971, 1987)[12, 13]. Por este método podemos obter as somas de quadrados tipo II de uma forma mais direta do que por redução de modelos. Vamos apresentar o método no contexto de regressão linear na seqüência. Seja a matriz $(X'X)^{-1}$ definida por:

$$(X'X)^{-1} = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0m} \\ x_{10} & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m0} & x_{m1} & \cdots & x_{mm} \end{bmatrix} \tag{3.7}$$

Assim, para obtermos a soma de quadrados do tipo II para a variável $X_h$ podemos simplesmente calcular:

$$R(\beta_h/\beta_0, \cdots, \beta_{h-1}, \beta_{h+1}, \cdots, \beta_m) = \frac{\hat{\beta}_h^2}{x_{hh}} \tag{3.8}$$

## 3.2 Um Exemplo de Regressão Pelo Proc IML

Vamos mostrar um exemplo de um ajuste de um modelo de regressão utilizando o *proc iml*. O objetivo é mostrar todos os cálculos utilizando as fórmulas anteriormente apresentadas por meio de um programa matricial. Seja para isso um exemplo em que a variável $X$ representa o número de horas de exposição solar de uma planta e a variável resposta $Y$ o crescimento da planta. Os dados deste exemplo estão apresentados na Tabela 3.2.

Vamos ajustar um modelo linear quadrático do tipo:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{3.9}$$

em que $\beta_0$, $\beta_1$ e $\beta_2$ são os parâmetros que desejamos estimar.

Para este modelo vamos estimar os parâmetros e obter as somas de quadrados dos tipos I e II utilizando o *proc iml*. A matriz $X$ do modelo é dada por:

Tabela 3.2: Crescimento de uma planta $Y$ após ser submetida a um tempo $X$ de exposição solar em horas.

| $X$ | $Y$ |
|---|---|
| 0,1 | 0,88 |
| 0,2 | 0,90 |
| 0,3 | 0,99 |
| 0,5 | 1,12 |
| 0,8 | 1,40 |
| 1,0 | 1,62 |
| 1,5 | 2,20 |
| 2,0 | 3,10 |

$$X = \begin{bmatrix} 1 & 0,1 & 0,01 \\ 1 & 0,2 & 0,04 \\ 1 & 0,3 & 0,09 \\ 1 & 0,5 & 0,25 \\ 1 & 0,8 & 0,64 \\ 1 & 1,0 & 1,00 \\ 1 & 1,5 & 2,25 \\ 1 & 2,0 & 4,00 \end{bmatrix}$$

O vetor de parâmetros é dado por:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

O vetor de observações é dado por:

$$\underset{\sim}{Y} = \begin{bmatrix} 0,88 \\ 0,90 \\ 0,99 \\ 1,12 \\ 1,40 \\ 1,62 \\ 2,20 \\ 3,10 \end{bmatrix}$$

Desta forma podemos formular o programa IML para ajustar este modelo e obter as somas de quadrados e testes de hipóteses relativo aos parâmetros. Vamos apenas ilustrar uma parte de todos os cálculos, pois felizmente podemos utilizar o proc reg do SAS que nos fornece todas as estimativas e testes de hipóteses que desejarmos, com comando mais simples. O nosso objetivo é possibilitar ao leitor obter um maior conhecimento de todo o processo de regressão linear. O programa resultante desta análise é:

---

```
/*Exemplo de programa IML para realizar regressão linear.*/
proc iml;
   x={ 1 0.1 0.01,
       1 0.2 0.04,
       1 0.3 0.09,
       1 0.5 0.25,
       1 0.8 0.64,
       1 1.0 1.00,
       1 1.5 2.25,
       1 2.0 4.00};
   y={ 0.88,
       0.90,
       0.99,
       1.12,
       1.40,
       1.62,
       2.20,
       3.10};
   /*modelo completo y = b0 + b1x + b2x^2*/
```

```
    n=nrow(y);
    xlx=t(x)*x;
    xly=t(x)*y;
    print xlx xly;
    ixlx=inv(xlx);
    print ixlx;
    betam1=ixlx*xly;
    print betam1;
    /*somas de quadrados*/
    glm1=3;
    sqb0b1b2=t(betam1)*xly;
    sqtotal=t(y)*y;
    sqresm1=sqtotal-sqb0b1b2;
    glrm1=n-glm1;
    print sqb0b1b2 sqtotal sqresm1;
    /*Soma de quadrados do tipo II*/
    sqb1=betam1[2]**2/(ixlx[2,2]);
    sqb2=betam1[3]**2/(ixlx[3,3]);
    print sqb1 sqb2;
    /*teste t H0 bi=0*/
    b0=betam1[1];
    tcb0=(b0-0)/(ixlx[1,1]*sqresm1/glrm1)**0.5;
    prtcb0=2*(1-probt(abs(tcb0),glrm1));
    print b0 tcb0 prtcb0;
    b1=betam1[2];
    tcb1=(b1-0)/(ixlx[2,2]*sqresm1/glrm1)**0.5;
    prtcb1=2*(1-probt(abs(tcb1),glrm1));
    print b1 tcb1 prtcb1;
    b2=betam1[3];
    tcb2=(b2-0)/(ixlx[3,3]*sqresm1/glrm1)**0.5;
    prtcb2=2*(1-probt(abs(tcb2),glrm1));
    print b2 tcb2 prtcb2;
quit;
```

---

Os principais resultados obtidos neste procedimento são apresentados na seqüência. Iniciamos pelas matrizes $X'X$ e $X'\underset{\sim}{Y}$, dadas por:

$$X'X = \begin{bmatrix} 8 & 6,4 & 8,28 \\ 6,4 & 8,28 & 13,048 \\ 8,28 & 13,048 & 22,5444 \end{bmatrix}$$

e

$$X'\underset{\sim}{Y} = \begin{bmatrix} 12,21 \\ 13,365 \\ 20,2799 \end{bmatrix}$$

A matriz inversa $(X'X)^{-1}$ é dada por:

$$(X'X)^{-1} = \begin{bmatrix} 0,7096 & -1,5667 & 0,6461 \\ -1,5667 & 4,8322 & -2,2213 \\ 0,6461 & -2,2213 & 1,0927 \end{bmatrix}$$

Finalmente, o vetor $\underset{\sim}{\beta}$ é estimado por:

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 0,8289504 \\ 0,4048794 \\ 0,3607692 \end{bmatrix}$$

Portanto, o modelo de regressão ajustado é $\hat{Y}_i = 0,8289504 + 0,4048794\ X_i + 0,3607692 X_i^2$. O gráfico desta função quadrática está apresentado na Figura (3.1)

As somas de quadrados para modelo $(\beta_0, \beta_1, \beta_2)$, total não corrigido e resíduo foram iguais a $22,84906$, $22,8533$ e $0,0042399$, respectivamente. O $R^2$, proporção da variação total corrigida explicada pelo modelo de regressão, é dado por: $R^2 = 1 - \text{sqresíduo/sqtotal corrigida} = 99,90\%$. Um excelente ajuste foi encontrado, mas é necessário que se faça a análise de resíduo para termos uma confirmação disto, o que não será feito neste instante. A soma de quadrado total corrigida foi obtida por SQtotal nc = sqtotal c $- G^2/n$, em que $G = \sum_{i=1}^{n} Y_i = 12,21$.

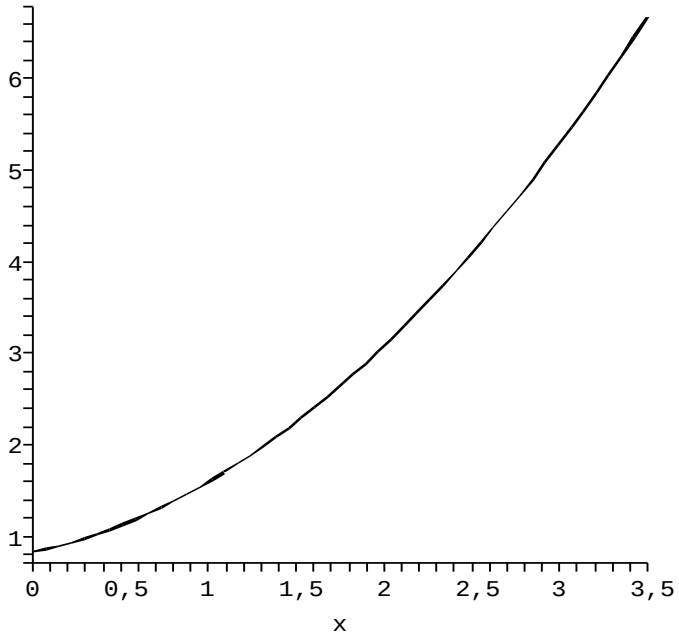


Figura 3.1: Equação quadrática resultante do ajuste de quadrados mínimos do exemplo tratado.

No passo seguinte obtivemos as somas de quadrados do tipo II para $X$ e $X^2$ por $0,4048794^2/4,8322 = 0,03392$ e $0,3607692^2/1,0927 = 0,1191$, respectivamente. Podemos efetuar um teste $F$ para a hipótese $H_0 : \beta_i = 0$ se desejarmos, dividindo o quadrado médio do tipo II de cada variável pelo quadrado médio do erro e calcularmos o valor-p utilizando a distribuição $F$ de Snedecor. O quadrado médio do tipo II para cada parâmetro é igual a soma de quadrados, pois está associado a 1 grau de liberdade. Finalmente podemos utilizar o teste $t$ de Student para obtermos um teste de hipótese equivalente ao realizado pelo teste $F$, baseado em somas de quadrados parciais ou somas de quadrados do tipo II. Este teste está descrito formalmente nas equações (3.13) a (3.16). Os resultados destes testes de hipótese bilateral estão apresentados na Tabela 3.3.

Podemos fazer muitas outras análises no *proc iml*. Isso não será necessário, pois o SAS possui alguns procedimentos apropriados para lidarmos com ajustes de modelos lineares. Entre estes procedimentos destacamos o

Tabela 3.3: Testes de hipótese do tipo $H_0 : \beta_i = 0$, com $i = 0, 1, 2$ utilizando a distribuição $t$ de Student com $\nu = 5$ graus de liberdade.

| Parâmetro | Estimativa | $t_c$ | $Pr(t > \|t_c\|)$ |
|---|---|---|---|
| $\beta_0$ | 0,82895 | 33,793 | $4,267 \times 10^{-7}$ |
| $\beta_1$ | 0,40488 | 6,325 | $0,0014562$ |
| $\beta_2$ | 0,36077 | 11,852 | $0,0000753$ |

*proc reg*, para o qual, anteriormente, já apontamos suas principais características, ou seja, as análises com que é capaz de lidar. Como o IML é um procedimento poderoso, mas que requer conhecimentos especiais de estatística e de álgebra matricial, não abordaremos mais o *proc iml*, neste capítulo. Faremos todas as análises de modelos lineares de regressão utilizando o *proc reg*.

## 3.3 O *Proc Reg*

Vamos apresentar o proc reg para realizarmos o ajuste do modelo anterior e em seguida apresentaremos um exemplo de regressão múltipla, onde aparentemente ocorre um resultado paradoxal na inferência realizada. Utilizamos este exemplo para elucidar aspectos de testes de hipóteses que são muitas vezes ignorados. Inicialmente vamos apresentar os comandos necessários para ajustarmos o modelo (3.9). O *proc reg* não permite a criação de variáveis no próprio modelo como faz um outro procedimento do SAS chamado *glm*. Neste caso, devemos criar o arquivo de dados e após o input criar a variável $X2 = X^2$. Assim, criamos nosso arquivo com as variáveis necessárias e o programa simplificado para o ajuste é dado por:

```
/*Exemplo do proc reg para realizar regressão linear.*/
data rlq;
input x y;
x2=x**2;
cards;
0.1 0.88
```

```
0.2 0.90
0.3 0.99
0.5 1.12
0.8 1.40
1.0 1.62
1.5 2.20
2.0 3.10
;
proc reg data=rlq;
    model y=x x2/ss1 ss2;
run;quit;
```

---

A linha de comando do *proc reg* dada por *<model y=x x2/ss1 ss2;>*, nos permite fazer o ajuste do modelo (3.9). As opções *ss1* e *ss2* solicitam o cálculo das somas de quadrados dos tipos I e II. Não necessitamos especificar nada mais, pois por default o SAS apresenta as estimativas dos parâmetros do modelo com seus erros padrões e testes de hipóteses associados, a análise de variância, o $R^2$, média geral e algumas outras estimativas de parâmetros específicos. O teste $F$ da análise de variância está relacionado a seguinte hipótese:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_m = 0 \\ H_1 : \beta_i \neq 0 \quad \text{Para algum } i = 1, 2, \cdots, m \end{cases} \tag{3.10}$$

Neste exemplo observamos que o $F$ observado foi igual a $2484,4$ e o -valor associado $Pr(F > F_c) < 0,0001$. Assim a hipótese nula global de que nenhuma variável explica significativamente a variação na variável resposta $Y_i$ foi rejeitada. O SAS realiza o teste $t$ para as hipóteses do tipo $H_0 : \beta_i = 0$, $i = 1, 2, \cdots, m$. Neste exemplo os valores da estatística $t$ e as respectivas significâncias estão apresentadas na Tabela 3.3. Concluímos que ambas as variáveis tem efeito significativamente diferente de zero na variação de $Y$. O teste $t$ de Student é equivalente ao teste $F$ parcial. Embora este teste tenha sido aplicado por ser padrão no SAS, é conveniente utilizar para este exemplo um teste seqüencial. Isto porque esta análise refere-se ao ajuste de um modelo polinomial e usualmente nestes casos utilizamos

testes que envolvem somas de quadrados tipo I. Este tipo de procedimento é comumente encontrado nos livros de estatística experimental.

Vamos apresentar um segundo exemplo, como dissemos anteriormente, para elucidarmos alguns pontos interessantes da análise de regressão linear. Nosso exemplo, refere-se a uma amostra de $n = 10$ árvores, na qual foram mensurados o volume ($Y$), em $m^3.acre^{-1}$, sendo que 1 *acre* é igual a 4.064 $m^2$, a área basal ($X_1$) em $dm^2$, a área basal tomada em % em relação à área de outra espécie ($X_2$) e a altura em pés ($X_3$) (1 pé = 30, 48 cm). Na Tabela 3.4 temos os dados amostrados na população de *Araucaria angustifolia*.

Tabela 3.4: Dados de uma amostra de $n = 10$ árvores de araucária (*Araucaria angustifolia*) mensuradas em relação ao volume $Y$, área basal $X_1$, área basal relativa $X_2$ e altura em pés $X_3$.

| $Y$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 65 | 41 | 79 | 35 |
| 78 | 71 | 48 | 53 |
| 82 | 90 | 80 | 64 |
| 86 | 80 | 81 | 59 |
| 87 | 93 | 61 | 66 |
| 90 | 90 | 70 | 64 |
| 93 | 87 | 96 | 62 |
| 96 | 95 | 84 | 67 |
| 104 | 100 | 78 | 70 |
| 113 | 101 | 96 | 71 |

Vamos inicialmente ajustar um modelo linear simples para cada variável utilizando o modelo linear dado por:

$$Y_i = \beta_0 + \beta_1 X_{hi} + \epsilon_i, \quad \text{Para} \quad h = 1, 2 \text{ ou } 3, \quad i = 1, 2, \cdots, n \qquad (3.11)$$

O programa para realizarmos estes ajustes, para cada uma das variáveis regressoras, mas de forma simultânea simultânea, é dado por:

```
/*Exemplo do proc reg para realizar regressão linear.*/
data arvores;
input y x1 x2 x3;
datalines;
   65    41   79   35
   78    71   48   53
   82    90   80   64
   86    80   81   59
   87    93   61   66
   90    90   70   64
   93    87   96   62
   96    95   84   67
  104   100   78   70
  113   101   96   71
;
proc reg data=arvores;
    model y=x1;
    model y=x2;
    model y=x3;
run;quit;
```

Na Tabela 3.5 apresentamos os resultados mais importantes destes ajustes, que iremos mencionar futuramente. Selecionamos o $F$ calculado e sua significância e o $R^2$ do modelo.

Tabela 3.5: Resultados mais importantes do ajuste dos modelos lineares simples para os dados dos volumes das $n = 10$ árvores de araucária *Araucaria angustifolia*.

| Modelo | $F_c$ | $Pr(F > F_c)$ | $R^2$ |
|---|---|---|---|
| 1: $E(Y_i) = \beta_0 + \beta_1 X_{1i}$ | 24,17 | 0,0012 | 0,7513 |
| 2: $E(Y_i) = \beta_0 + \beta_1 X_{2i}$ | 2,43 | 0,1579 | 0,2328 |
| 3: $E(Y_i) = \beta_0 + \beta_1 X_{3i}$ | 24,73 | 0,0011 | 0,7556 |

Observamos que o modelo 2 não se ajustou aos dados, embora isso fosse esperado, uma vez que a variável $X_2$ é resultante de uma medida relativa entre uma variável mensurada diretamente na espécie e outra medida em outra espécie. Portanto, o resultado é perfeitamente justificável, pois a

covariação existente entre $X_2$ e $Y$ pode ser atribuída meramente à fatores de acaso. As demais variáveis apresentam explicações significativas ($P < 0,05$) da variação que ocorre na variável resposta, com $R^2$ igual a $75,13\%$ para $X_1$ e $75,56\%$ para $X_3$. Agora vamos ajustar o modelo linear múltiplo dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \tag{3.12}$$

O programa SAS, que faz uso do *proc reg* para ajustar o modelo 3.12, é dado por:

```
/*Exemplo do proc reg para realizar regressão linear múltipla.*/
data arvores;
input y x1 x2 x3;
datalines;
   65    41   79   35
   78    71   48   53
   82    90   80   64
   86    80   81   59
   87    93   61   66
   90    90   70   64
   93    87   96   62
   96    95   84   67
  104   100   78   70
  113   101   96   71
;
proc reg data=arvores;
    model y=x1 x2 x3;
run;quit;
```

Os principais resultados obtidos do ajuste do modelo 3.12 são apresentados e discutidos na seqüência. A princípio, vamos apresentar (Tabela 3.6) o resumo da análise de variância.

Podemos concluir que pelo menos uma variável explica significativamente a variação que ocorre na variável resposta $Y$, ou seja, a hipótese nula (3.10) deve ser rejeitada se for considerado o nível nominal de 5%.

Tabela 3.6: Resumo da análise de variância do ajuste de regressão múltipla aos dados do volume das árvores de araucária.

| FV | GL | QM | $F_c$ | $Pr(F > F_c)$ |
|---|---|---|---|---|
| Regressão | 3 | $455,85296$ | $10,65$ | $0,0081$ |
| Erro | 6 | $42,80685$ | | |
| Total Corrigido | 9 | | | |

Na Tabela 3.7 apresentamos os testes $t$ de Student para a hipótese nula $H_0 : \beta_h = 0$, em que $h = 1, 2, 3$. Devemos neste instante apresentar a expressão geral para realizarmos os testes de hipóteses sobre componentes do vetor de parâmetros. A variância do estimador do vetor de parâmetros é dada por:

$$V\left(\underset{\sim}{\hat{\beta}}\right) = (X'X)^{-1}\sigma^2 \tag{3.13}$$

O estimador desta variância é obtido substituindo a variância paramétrica pelo estimador da variância ($S^2 = QME$). Assim, temos o estimador da variância do estimador dos parâmetros dada por:

$$\hat{V}\left(\underset{\sim}{\hat{\beta}}\right) = (X'X)^{-1}S^2 \tag{3.14}$$

Desta forma, o erro padrão de $\hat{\beta}_i$ é dado por:

$$S_{(\hat{\beta}_i)} = \sqrt{x_{ii}S^2} \tag{3.15}$$

em que $x_{ii}$ é o elemento correspondente a i-ésima diagonal da matriz inversa $(X'X)^{-1}$.

Logo, o teste $t$ de Student para a hipótese $H_0 : \beta_i = \delta_0$, em que $\delta_0$ é uma constante real de interesse pode ser aplicado, pois sob $H_0$ a distribuição da estatística do teste dada por

$$t_c = \frac{\hat{\beta}_i - \delta_0}{S_{(\hat{\beta}_i)}} \tag{3.16}$$

é $t$ de Student com $\nu = n - m - 1$ graus de liberdade.

O SAS testa a hipótese nula, assumindo que a constante $\delta_0$ é igual a zero. Os resultados para este caso estão apresentados na Tabela 3.7.

Tabela 3.7: Estimativas dos parâmetros e teste $t$ de Student para a nulidade das estimativas.

| Parâmetros | Estimativas | $S_{(\hat{\beta}_i)}$ | $t_c$ | $Pr(t > \lvert t_c \rvert)$ |
|---|---|---|---|---|
| $\beta_0$ | -33,82268 | 75,35853 | -0,45 | 0,6693 |
| $\beta_1$ | -2,22672 | 4,02805 | -0,55 | 0,6004 |
| $\beta_2$ | 0,26976 | 0,15332 | 1,76 | 0,1290 |
| $\beta_3$ | 4,76590 | 6,78649 | 0,70 | 0,5088 |

Quando observamos os resultados dos testes de hipóteses na Tabela 3.7, verificamos que nenhuma variável explicou significativamente a variação da variável resposta $Y$. Este resultado é aparentemente contraditório ao resultado do teste da hipótese global do modelo de regressão, hipótese esta que foi significativamente rejeitada. Este suposto paradoxo na verdade é um problema de interpretação do que está sendo realmente testado pelos testes $t$ individuais. O que ocorre é que o teste $t$ é equivalente ao teste $F$, obtido a partir das somas de quadrados parciais ou do tipo II. Assim, o que o $t$ realmente testa é a contribuição de uma variável, eliminando a explicação das demais variáveis no modelo. Então, se a explicação da variável para a variação de $Y$ for expressiva, após ser eliminada a redundância da informação com as outras variáveis do modelo, a estatística do teste tenderá a pertencer a região crítica. Essa redundância é dependente da estrutura de correlação existente entre a variável que está sendo testada e as demais variáveis do modelo.

O que acontece neste exemplo é que temos uma forte estrutura de correlação entre as três variáveis do modelo e, portanto, na presença das outras, a variável que está sendo testada não contribui com uma explicação significativa da variação total. Podemos perceber que duas das variáveis que apresentaram resultados não significativos para o teste $t$, são individualmente importantes para a variação do volume, pois apresentaram significâncias menores que 5% nos testes individuais. Portanto, não tem nada de parado-

xal nos resultados encontrados. O que temos são variáveis correlacionadas que não necessitariam estar todas no modelo e parte delas nem precisaria ser mensurada, onerando menos os experimentos de campo.

Um outro parâmetro que é estimado pelo *proc reg* é o $R^2$, o qual mede a proporção da variação do total dos dados que é explicada pelo modelo de regressão. Um outro importante parâmetro é o coeficiente de determinação ajustado ($R^2_{Aj.}$). Este ajuste, feito para o número de parâmetros no modelo, fornece uma medida mais adequada para comparar modelos com diferentes quantidades de parâmetros. O $R^2$ ajustado é dado por:

$$R^2_{Aj.} = 1 - \frac{n-i}{n-p}\left(1 - R^2\right) \tag{3.17}$$

em que $n$ é o tamanho da amostra, $p$ é o número de parâmetros (incluindo o intercepto) e $i$ é igual a 1, se o modelo inclui o intercepto ou 0, se o modelo não inclui $\beta_0$.

Duas opções interessantes para calcularmos as somas de quadrados tipos I e II são dadas por SS1 e SS2. Estas opções devem aparecer após o modelo. Para isso, ao terminarmos de especificar o modelo, colocamos uma barra / e em seguida as opções SS1 e SS2. O programa simplificado ilustrando o uso das opções SS1 e SS2 é dado por:

```
/*Exemplo do proc reg para realizar regressão linear múltipla utilizando SS1 e SS2.*/
proc reg data=arvores;
   model y=x1 x2 x3/ss1 ss2;
run;quit;
```

Juntamente com as estimativas dos parâmetros podemos observar as somas de quadrados tipo I e II resultantes das opções de modelo utilizadas. Outros comandos que são importantes no *proc reg* são: p, clm e cli. Estas opções nos possibilitam predizer os valores de $Y_i$, estimar por intervalo de confiança o valor médio da resposta (clm) ou intervalo de confiança para uma predição estocástica ou predição futura (cli). Para apresentarmos estes conceitos, sejam $Y_i$ a observação da variável resposta na i-ésima unidade

amostral e o vetor $\underset{\sim}{z}_i = [1 \quad X_{1i} \quad X_{2i} \quad \cdots \quad X_{mi}]'$ o vetor de variáveis regressoras, incluindo a indicadora do intercepto, então o valor predito $\hat{Y}_i$ é dado por:

$$\hat{Y}_i = \underset{\sim}{z}'\underset{\sim}{\hat{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_1 X_{mi} \tag{3.18}$$

Este vetor $\underset{\sim}{z}_i$ não necessita necessariamente ser observado entre o conjunto de observações. O estimador do erro padrão desta predição para o intervalo da média ($clm$) é dado por:

$$S(\hat{Y}_i) = \sqrt{\underset{\sim}{z}'(X'X)^{-1}\underset{\sim}{z}S^2} \tag{3.19}$$

O intervalo de confiança $clm$ é dado por:

$$\hat{Y}_i \pm t_{\alpha/2,\nu} S(\hat{Y}_i) \tag{3.20}$$

Se diferenciarmos a predição futura da predição média simplesmente utilizando a notação $\tilde{Y}_i$, mas mantivermos a mesma combinação linear determinada pelo vetor $\underset{\sim}{z}$, teremos o intervalo de confiança $cli$ dado por:

$$\tilde{Y}_i \pm t_{\alpha/2,\nu} S(\tilde{Y}_i) \tag{3.21}$$

Este intervalo distingue-se do anterior somente pelo estimador do erro padrão do valor da predição futura, o qual envolve uma variância residual a mais em relação ao erro padrão da predição do valor médio. Este estimador do erro padrão da predição futura é dado por:

$$S(\tilde{Y}_i) = \sqrt{\left[1 + \underset{\sim}{z}'(X'X)^{-1}\underset{\sim}{z}\right] S^2} \tag{3.22}$$

O programa SAS simplificado para ilustrarmos o uso destas opções está apresentado na seqüência. Podemos especificar o valor de $\alpha$ com a opção *alpha=0.05*. Claro que se o valor de 5% for mantido, que é o padrão, esta opção não precisa ser utilizada.

```
/*Exemplo do proc reg para realizar regressão linear múltipla utilizando p clm e cli.*/
proc reg data=arvores;
   model y=x1 x2 x3/alpha=0.05 p clm cli;
run;quit;
```

Podemos utilizar ainda algumas outras opções do modelo de regressão. Particularmente interessante são os coeficientes de determinações semi-parciais dos tipos I e II. Os comandos para obtermos estas correlações semi-parciais quadráticas são *scorr1* e *scorr2*. Os coeficientes de determinação semi-parciais são estimados por:

$$R^2_{sp1} = \frac{R(\beta_h/\beta_0, \cdots, \beta_{h-1})}{\text{SQtotal corrigida}} \tag{3.23}$$

e

$$R^2_{sp2} = \frac{R(\beta_h/\beta_0, \cdots, \beta_{h-1}, \beta_{h+1}, \cdots, \beta_m)}{\text{SQtotal corrigida}} \tag{3.24}$$

em que $R^2_{sp1}$ e $R^2_{sp2}$ são os coeficientes de determinação semi-parciais dos tipos I e II, respectivamente, para a h-ésima variável.

Também são úteis os coeficientes de determinação parciais dos tipos I e II. As opções que devemos utilizar são, respectivamente, *pcorr1* e *pcorr2*. Os estimadores correspondentes são dados por:

$$R^2_{p1} = \frac{R(\beta_h/\beta_0, \cdots, \beta_{h-1})}{R(\beta_h/\beta_0, \cdots, \beta_{h-1}) + SQE^*} \tag{3.25}$$

em que $SQE^*$ é a soma de quadrados do erro resultante do ajuste de um modelo contendo as variáveis $X_1$, $X_2$, $\cdots$, $X_h$ e

$$R^2_{p2} = \frac{R(\beta_h/\beta_0, \cdots, \beta_{h-1}, \beta_{h+1}, \cdots, \beta_m)}{R(\beta_h/\beta_0, \cdots, \beta_{h-1}, \beta_{h+1}, \cdots, \beta_m) + \text{SQE}} \tag{3.26}$$

em que $SQE$ é a soma de quadrados do erro resultante do ajuste do modelo completo.

```
/*Exemplo do proc reg para realizar regressão linear múltipla e ilustrar a obtenção dos
coeficientes de determinação parciais e semi-parciais.*/
proc reg data=arvores;
    model y=x1 x2 x3/ss1 ss2 scorr1 scorr2 pcorr1 pcorr2;
run;quit;
```

## 3.4 Seleção de Modelos

A seleção de modelos é bastante interessante na pesquisa científica, pois muitas vezes temos variáveis correlacionadas que não contribuem para a variação da variável resposta de forma significativa, na presença das outras. Dizemos que existe uma redundância da informação. Assim, procedimentos para selecionarmos modelos de regressão linear são importantes no sentido de evitarmos a inclusão em um modelo de variáveis que são correlacionadas com outras variáveis candidatas. Evitamos com isso mensurações desnecessárias e onerosas. O SAS nos permite utilizar diferentes métodos de seleção de modelos, quais sejam, *forward*, *backward*, *stepwise*, *maxr*, *minr*, *rsquare*, *adjrsq*, *cp* ou *none* (usar o modelo completo). Cada um destes métodos tem uma característica especial. Enfocaremos nesta seção apenas os três primeiros: *forward*, *backward* e *stepwise*.

Vamos apresentar algumas características de cada um destes três métodos escolhidos. Vamos iniciar pelo *forward*. Neste método as $m$ variáveis regressoras são submetidas a um ajuste individual (modelo linear simples). Cada modelo deste é ajustado e entre aqueles modelos em que as variáveis regressoras apresentaram teste $F$ parcial significativo para a hipótese $H_0 : \beta_h = 0$, fixado o valor de $\alpha$, devemos escolher aquela variável que apresentou maior valor desta estatística ou equivalentemente, aquela que apresentou maior $R^2$ parcial. A variável escolhida é fixada no modelo e todas as outras são introduzidas um a uma neste modelo, formando $m-1$ modelos de duas variáveis. Estes modelos são formados pela variável escolhida no passo 1 com a outra escolhida entre as variáveis candidatas a entrar neste modelo. Novamente entre aquelas variáveis que apresentaram $F$ parcial significativo

na presença da variável selecionada no primeiro passo, escolhemos aquela de maior $F$ parcial ou $R^2$ parcial. Se nenhuma variável apresentou significância para entrar, encerramos o processo e ficamos com um modelo com a variável que entrou no primeiro passo. Se uma das candidatas foi escolhida, formamos um modelo com esta variável e aquela escolhida no passo 1. As variáveis candidatas são testadas uma por vez na presença destas duas variáveis e todo o processo é repetido. Devemos parar quando nenhuma das candidatas atingiu o nível de significância estabelecido a priori para entrar no modelo ou quando não temos mais variáveis candidatas para entrar.

O procedimento *stepwise* é muito parecido com o *forward*, exceto pelo fato de que em cada passo, após a entrada de uma das variáveis candidatas, devemos testar as variáveis que estavam no modelo. Se uma ou mais delas apresentarem $F$ parcial não significativo, aquela que tiver menor valor de $F$ parcial deve sair do modelo. Esta saída é de apenas uma variável por vez, até não ter mais variáveis no modelo que apresentem $F$ parcial não significativos. As variáveis que saíram do modelo, não são mais candidatas a entrar. As variáveis remanescentes, candidatas a entrar no modelo, são colocadas um por vez no modelo final e o processo continua com entradas e saídas até não termos mais candidatas para entrarem ou as candidatas não atingirem o nível mínimo de significância para entrarem no modelo e as variáveis do modelo forem todas significativas.

O procedimento de *backward* testa todas as variáveis candidatas simultaneamente. Entre aquelas que apresentarem $F$ parciais não significativos, a que tiver menor valor observado deve sair do modelo. Se todas as variáveis no modelo apresentarem $F$ parciais significativos, em um nível pré-estabelecido $\alpha$ de significância para a permanência no modelo, então encerramos o processo. Neste caso o modelo resultante será o completo. Se por outro lado, for eliminada um variável, o procedimento é repetido para as $m-1$ variáveis remanescentes. Paramos o processo se todas as variáveis de um passo apresentarem $F$ parcial significativo ou se modelo resultar em um modelo nulo, somente com o intercepto.

Devemos especificar para o SAS o nível de significância de permanência ou de entrada das variáveis do modelo. No *forward* devemos especificar somente o nível de significância de entrada, no *backward*, o nível de signi-

ficância de permanência e no *stepwise*, os dois níveis de significância, de permanência e de entrada. Os comandos que devemos usar são *slstay* para nível de significância de permanência e *slentry* para entrada.

O comando que utilizamos para indicarmos que utilizaremos um método de seleção de modelos é o *selection=method*. O programa SAS para realizarmos a escolha de modelos de regressão, para os dados das árvores, é dado por:

```
/*Exemplo do proc reg para realizar seleção de modelos de regressão linear múltipla.*/
proc reg data=arvores;
    model y=x1 x2 x3/selection=backward slstay=0.05;
    model y=x1 x2 x3/selection=forward slentry=0.05;
    model y=x1 x2 x3/selection=stepwise slentry=0.05 slstay=0.05;
run;quit;
```

Nos três métodos obtivemos o mesmo modelo ajustado, da variável resposta $Y$ em função da variável $X_3$. Algumas vezes os procedimentos podem resultar em conclusões conflitantes quanto ao modelo e o pesquisador deve escolher o que melhor lhe convier. Esta escolha, entre outras coisas, pode ser embasada na análise de resíduos e na qualidade da predição da variável aleatória $Y$.

## 3.5 Diagnóstico em Regressão Linear

Seja o modelo de regressão linear dado por

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

em que $\underset{\sim}{Y}$ é o vetor de observações de dimensões $n \times 1$; $X$ é a matriz do modelo de dimensões $n \times (m+1)$ das derivadas parciais de $Y_i$ em relação aos parâmetros; $\underset{\sim}{\beta}$ é o vetor de parâmetros $[(m+1) \times 1]$; e $\underset{\sim}{\epsilon}$ é o vetor de resíduos $(n \times 1)$ não observáveis e com $E\left(\underset{\sim}{\epsilon}\right) = \underset{\sim}{0}$ e $V\left(\underset{\sim}{\epsilon}\right) = I\sigma^2$.

Na metodologia clássica de modelos lineares, onde se encontram os modelos de regressão linear, pressupomos que exista uma linearidade nos parâmetros do preditor e aditividade dos erros e, ainda, que os erros são independentes, têm média zero, variância constante e que sua distribuição seja normal, ou seja, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Além disso outras condições são importantes, como por exemplo, supomos que algumas poucas observações não devam ter influência demasiada sobre as estimativas dos parâmetros do modelo e de suas variâncias. Assim, diagnósticos numéricos são funções dos dados cujos valores permitem detectar respostas que são anormalmente grandes ou pequenas (outliers ou valores discrepantes) ou que estão afastadas do grupo majoritário dos dados, influenciando em demasia o ajustamento. Assim, temos interesse particular nas análises denominadas de influência, onde utilizamos um conjunto de técnicas destinadas a detecção de pontos influentes e/ou discrepantes que podem afetar o ajustamento.

Muitas causas podem ser atribuídas a alguns problemas normalmente encontrados na análise de regressão. Algumas destas possibilidades são, entre outras, devidas à medidas erradas ou erro no registro da realização da variável resposta, ou ainda, erros de transcrição; observações tomadas em condições distintas das demais; modelo mal especificado; e distribuição não normal dos resíduos, apesar de o modelo e a escala estarem corretos.

A forma utilizada normalmente para verificar a influência de uma observação é retirá-la do modelo e verificar como as estimativas dos parâmetros, predições e variâncias são afetadas. Assim, se retirarmos a i-ésima observação e reestimarmos as quantidades mais importantes do modelo, poderemos avaliar a influência da observação retirada na estimação destes parâmetros de interesse. Podemos, no entanto, evitar que todos os cálculos sejam refeitos, utilizando algumas relações e propriedades apresentadas por Velleman e Welsch, (1981)[16]. Vários métodos de avaliar a influência de observações no ajuste de um modelo de regressão linear são apresentados por Chatterjee e Hadi (1986)[2].

### 3.5.1 Análise de resíduos

O preditor dos resíduos é dado por:

$$\underset{\sim}{e} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}} \tag{3.27}$$

Podemos reescrever o erro como uma combinação linear de $\underset{\sim}{Y}$ por:

$$\underset{\sim}{e} = \underset{\sim}{Y} - X(X'X)^{-1}X'\underset{\sim}{Y} = [I - X(X'X)^{-1}X']\underset{\sim}{Y}$$

A matriz $X(X'X)^{-1}X'$ é denominada projetor e representada por $P$, pois projeta o vetor de observações $\underset{\sim}{Y}$, $n$-dimensional, no sub-espaço $(m+1)$-dimensional. Aplicando esta matriz ao vetor de observações, obtemos o vetor de valores preditos $\underset{\sim}{\hat{Y}}$, ou seja, $\underset{\sim}{\hat{Y}} = P\underset{\sim}{Y}$. Na análise de regressão linear simples, a matriz $P$ é denominada de matriz *Hat* e representada por $H$. Vamos representar a i-ésima observação pelo vetor composto por $[Y_i \quad \underset{\sim}{z_i}']'$, sendo que $\underset{\sim}{z_i} = [1 \quad X_{1i} \quad X_{2i} \quad \cdots \quad X_{mi}]'$ é o vetor dos elementos da i-ésima linha da matriz $X$ do modelo. O elemento da diagonal correspondente na matriz $H$ é denominado simplesmente por $h_i$. Assim,

$$\underset{\sim}{e} = (I - H)\underset{\sim}{Y} \tag{3.28}$$

é o preditor do vetor de erros, que é equivalente a equação (3.27).

A esperança de $\underset{\sim}{e}$ é dada por:

$$\begin{aligned} E\left(\underset{\sim}{e}\right) =& E\left[(I - H)\underset{\sim}{Y}\right] = (I - H)E\left(\underset{\sim}{Y}\right) \\ =& [I - X(X'X)^{-1}X']X\underset{\sim}{\beta} = X\underset{\sim}{\beta} - X(X'X)^{-1}X'X\underset{\sim}{\beta} \\ =& X\underset{\sim}{\beta} - X\underset{\sim}{\beta} = \underset{\sim}{0} \end{aligned}$$

Assim, a covariância do vetor de resíduos preditos é:

$$\begin{aligned} V\left(\underset{\sim}{e}\right) =& (I - H)V\left(\underset{\sim}{Y}\right)(I - H) = (I - H)I\sigma^2(I - H)' \\ =& (I - H)(I - H')\sigma^2 = (I - H) - (I - H)H' \\ =& (I - H - H' + HH')\sigma^2 = (I - H - H + H)\sigma^2 \\ =& (I - H)\sigma^2 \end{aligned}$$

Para a i-ésima observação temos que a variância $V(e_i)$ é dada por:

$$V(e_i) = (1 - h_i)\sigma^2 \tag{3.29}$$

em que $e_i$ é o i-ésimo elemento do vetor de resíduos preditos, ou seja, é o erro predito para a i-ésima observação. Neste contexto é denominado de resíduo ordinário.

O problema básico destes resíduos é que eles não são comparáveis entre si, por possuírem variâncias distintas. Devemos buscar alguma forma de padronização para termos a mesma dispersão em todos os $n$ resíduos preditos. Temos basicamente três formas de padronizações que podemos efetuar e que discutiremos na seqüência. Podemos ter os resíduos padronizados, resíduos estudentizados internamente e resíduos estudentizados externamente, também conhecidos por resíduos de *jackknife* (Chatterjee e Hadi, 1986[2]). Em todos os casos vamos substituir a variância $\sigma^2$ pelo seu estimador $S^2 = QME$.

A primeira opção, não computada pelo SAS, é obtida pela divisão dos resíduos ordinários pelo desvio padrão $S = \sqrt{QME}$. Este artifício reduz a variabilidade a uma faixa específica, mas não elimina o problema de variâncias distintas. Este resíduo padronizado é dado por:

$$z_i = \frac{e_i}{S} \tag{3.30}$$

Pela razão anteriormente apontada, os resíduos estudentizados foram propostos na literatura especializada. Os resíduos estudentizados internamente são obtidos por meio da razão entre o resíduo ordinário e o seu estimador do erro padrão específico, ou seja, por

$$r_i = \frac{e_i}{\sqrt{(1 - h_i)S^2}} \tag{3.31}$$

Este tipo de resíduo é mais interessante que o anterior, devido ao fato de considerar a variância individual de cada resíduo ordinário. Entretanto, se a i-ésima observação for um *outlier* pode ocorrer que a estimativa da variância estará afetada por este valor.

A última proposta de padronização foi feita para contornar este problema e tem ainda algumas propriedades mais interessantes do que as demais formas de padronização. Esta última padronização resulta nos resíduos estudentizados externamente, também denominados de resíduos de *jackknife*. A idéia é eliminar a i-ésima observação e obtermos uma estimador da variância, digamos, $S^2_{(i)}$. O subscrito $i$ apresentado entre parênteses foi utilizado para indicar que se trata de um estimador aplicado a todos as $n-1$ observações resultante da eliminação da i-ésima observação da amostra completa. Felizmente, não precisamos reajustar o modelo eliminando a i-ésima observação para obtermos uma estimativa desta variância (Chatterjee e Hadi, 1986[2]). Um estimador obtido a partir da análise original (Beckman e Trussell, 1974[1]) é dado por:

$$S^2_{(i)} = \frac{(n-m-1)S^2}{n-m-2} - \frac{e_i^2}{(n-m-2)(1-h_i)} \tag{3.32}$$

O resíduo estudentizado externamente é definido por:

$$t_i = \frac{e_i}{\sqrt{(1-h_i)S^2_{(i)}}} \tag{3.33}$$

Este resíduo é denominado por $RSTUDENT$ na literatura especializada de regressão. Observações que apresentarem este tipo de resíduo superior em módulo a 2, devem receber atenção especial. Existe uma preferência por este tipo de resíduo na literatura e as razões para isso podem ser apontadas (Chatterjee e Hadi, 1986[2]) por:

- Os resíduos estudentizados externamente $t_i$ sob a hipótese de normalidade seguem a distribuição $t$ de Student com $\nu = n-m-2$ graus de liberdade, enquanto $r_i^2/(n-m-1)$ segue a distribuição beta;
- podemos mostrar facilmente que:

  $$t_i = r_i\sqrt{\frac{n-m-2}{n-m-1-r_i^2}}$$

  de onde se observa que $t_i$ é uma transformação monotônica de $r_i$ e que $t_i \to \infty$ à medida que $r_i \to (n-m-1)$. Assim, $t_i$ reflete um resíduo fora de faixa de forma mais acentuada do que faz $r_i$; e

- o estimador $S^2_{(i)}$ é robusto à grandes e grosseiros erros da i-ésima observação, ou seja, se esta observação for discrepante.

É importante ressaltarmos que a detecção de valores discrepantes não deve implicar em descarte automático de observações. É possível, por exemplo, que o valor discrepante se deva a erro de transcrição, situação em que esse valor pode ser facilmente corrigido ou então pode ser um indicativo de modelo inadequado, possibilitando que modelos melhores sejam adotados e ajustados.

### 3.5.2 Influência no Espaço das Variáveis Preditoras

Além dos resíduos podemos verificar a influência das observações em uma série de quantidades importantes da análise de regressão. Uma interessante medida de diagnóstico é o próprio elemento $h_i$ da matriz de projeção $H$. Esta estatística é denominada de influência (*leverage*). O critério utilizado é baseado em algumas propriedades (Velleman e Welsch, 1981[16]) de $h_i$, dadas por: $0 \leq h_i \leq 1$ e $\sum_{i=1}^{n} h_i = (m+1)$. Assim, o valor médio da influência é $(m+1)/n$. Como $h_i = \partial \hat{Y}_i / \partial Y_i$, uma estimativa igual a zero é indicativo de que não há influência no ajuste do modelo e uma estimativa igual a 1, é indicativo que um grau de liberdade foi efetivamente atribuído ao ajuste daquela observação. O problema é determinar quais observações amostrais têm alta influência no ajuste e, portanto, receber atenção especial. Se $m > 14$ e $(n-m) > 31$ podemos utilizar o critério de que a i-ésima observação merece atenção se $h_i > 2(m+1)/n$. Se estas condições envolvendo $m$ e $n$ não forem verificadas, podemos utilizar $h_i > 3(m+1)/n$ como um melhor critério.

Devemos chamar a atenção de que a influência medida pelo $h_i$ refere-se ao papel das variáveis regressoras (fatores). Assim, medimos a influência, com $h_i$, no espaço dos fatores e, com a análise de resíduos, no espaço da variável resposta. Assim, a influência pode ocorrer no espaço dos fatores, no espaço das respostas ou em ambos os casos.

### 3.5.3 Influência no Vetor de Estimativas dos Parâmetros

A idéia de medir a influência da i-ésima observação na estimativa do vetor de parâmetros pode ser desenvolvida a partir da eliminação desta observação. Após esta eliminação, estimamos novamente os parâmetros do modelo e aplicamos uma medida de distância entre as estimativas. Esta distância pode ser dada pela diferença entre as estimativas obtidas com e sem a eliminação da i-ésima observação. Em geral é isso que fazemos, tomando-se o cuidado apenas de padronizar os resultados. Seja $\hat{\beta_{ij}}$, o estimador do j-ésimo parâmetro após eliminarmos a i-ésima observação, para $i = 1, 2, \cdots, n$ e $j = 0, 1, \cdots, m$. A estatística que utilizaremos para isso é conhecida por $DFBETA_{ij}$, em que DF são as iniciais de *Deviation of Fit*. Por meio dela podemos determinar a influência de cada observação na estimativa de cada parâmetro do modelo. Esta estatística é dada por:

$$DFBETA_{ij} = \frac{\hat{\beta}_j - \hat{\beta_{ij}}}{\hat{V}\left(\hat{\beta}_j\right)} \tag{3.34}$$

A dificuldade é obter as estimativas do vetor de parâmetros para cada um dos $n$ casos, em que um das variáveis é eliminada. Felizmente, não precisamos estimar $n$ vezes o vetor de parâmetros para calcularmos os $DFBETAS$. Existe uma relação interessante (Chatterjee e Hadi, 1986[2]) para a diferença entre os vetor de estimativas com e sem a i-ésima observação que é dada por:

$$\underset{\sim}{\hat{\beta}} - \underset{\sim}{\hat{\beta_{(i)}}} = \frac{1}{1 - h_i}(X'X)^{-1}\underset{\sim}{Z_i}e_i \tag{3.35}$$

em que $\underset{\sim}{\hat{\beta_{(i)}}}$ é o estimador do vetor de parâmetros após a eliminação da i-ésima observação.

Também sabemos que o vetor de estimadores dos parâmetros é dado por:

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y} = C\underset{\sim}{Y} \tag{3.36}$$

Assim, o $DFBETA$ não padronizado é dado por:

$$DFBETA_{ij} = c_{ji}\frac{e_i}{1-h_i} \tag{3.37}$$

em que $c_{ji}$ é o elemento da j-ésima linha e i-ésima coluna da matriz $C = (X'X)^{-1}X'$.

Se a expressão (3.37) for dividida pelo erro padrão do vetor de parâmetros $\hat{V}\left(\hat{\beta}_j\right)$, obteremos uma expressão equivalente (3.34). A expressão resultante é utilizada para obtermos os $DFBETAS$, sendo dada por:

$$DFBETA_{ij} = \frac{c_{ji}t_i}{\sqrt{(1-h_i)\underset{\sim}{C_j}'\underset{\sim}{C_j}}} \tag{3.38}$$

em que $\underset{\sim}{C_j}$ é vetor obtido a partir da j-ésima linha da matriz $C$.

Estas estatísticas são muito dependentes do número de observações, sendo que tanto menor será o efeito da observação sobre os valores de $DFBETAS$, quanto maior for o número de observações. Para estabelecer um valor limite para essa estatística, podemos tomar como base o valor limite para os resíduos, que é igual a 2. Assim, teremos que observações cujos $|DFBETA_{ij}| > 2/\sqrt{n}$ devem ter atenção especial, pois o vetor de estimativas pode ter sofrido alterações significativas.

### 3.5.4 Influência no Vetor de Valores Preditos

O impacto da i-ésima observação no i-ésimo valor predito pode ser medido pela padronização da mudança no valor predito na presença e ausência desta observação. A estatística utilizada para fazer tal mensuração é denominada de $DFFITS$ e é dada por:

$$DFFITS_i = \frac{\left|\underset{\sim}{Y_i} - \underset{\sim}{\hat{Y}_{i(i)}}\right|}{\sqrt{(1-h_i)S^2_{(i)}}} = |t_i|\sqrt{\frac{h_i}{1-h_i}} \tag{3.39}$$

Podemos verificar que quanto maior a influência da i-ésima observação, mais $h_i$ se aproxima de 1 e, conseqüentemente, maior será o coeficiente $|t_i|$.

Como vimos anteriormente $h_i/(1-h_i)$ está relacionada a uma medida da distância entre as linhas de $X$. Assim, a grandeza do valor de $DFFITS$ pode ser atribuída à discrepância do valor da resposta, do conjunto de valores das variáveis preditoras ou de ambos. Um ponto geral para a determinação de observações influentes é considerado o valor 2. Um ponto de corte ajustado para determinar a influência é $2\sqrt{(m+1)/n}$.

A distância de Cook é outra estatística utilizada para medir a influência de uma observação na predição dos valores da variável resposta $Y$. Esta estatística pode ser vista como a distância Euclidiana entre os valores preditos com e sem a i-ésima observação. O estimador da distância de Cook é dado por:

$$D_i = \frac{1}{(m+1)} \frac{h_i}{(1-h_i)} r_i^2 \tag{3.40}$$

Apesar de que a distância de Cook não deva ser usada como teste de significância, sugere-se o uso dos quantis da distribuição $F$ central com $m+1$ e $n-m-1$ graus de liberdade para servir de referência para o valor $D_i$. Outros autores sugerem que se $D_i > 1$, a i-ésima observação deve ser considerada influente.

A distância de Cook utiliza $r_i^2$, sendo que implicitamente está utilizando $S^2$ para padronizar a variância. Existe uma sugestão de que esta estatística possa ter melhores propriedades se for utilizado o estimador $S^2_{(i)}$ no lugar de $S^2$. Assim, a distância modificada de Cook utiliza esta substituição e faz um ajuste para o número de observações e toma ainda a raiz quadrada da distância transformada. A distância modificada de Cook é dada por:

$$D_i^* = |t_i| \sqrt{\frac{h_i(n-m-1)}{(1-h_i)(m+1)}} = DFFITS \sqrt{\frac{n-m-1}{m+1}} \tag{3.41}$$

Com essa modificação, temos que: a nova estatística enfatiza mais os pontos extremos; o gráfico de probabilidade normal pode ser utilizado para checagem; nos casos perfeitamente balanceados $[h_i = (m+1)/n]$ para qualquer i, a distância modificada tem comportamento idêntico ao $DFFITS$; a distância modificada com sinal pode ser plotada contra variáveis exploratórias do modelo.

Dado o limite máximo estabelecido para $DFFITS$, um valor da distância modificada de Cook maior que 2 pode ser considerado um indicativo de observação influente.

### 3.5.5 Influência na Matriz de Covariâncias

Uma medida da influência da i-ésima observação na $V\left(\underset{\sim}{\hat{\beta}}\right)$ é obtida comparando a razão de variâncias generalizadas (determinantes) da estimativa da covariância com e sem a i-ésima observação. Esta estatística é dada por:

$$COVRATIO_i = \frac{\det\left[S^2_{(i)}\left(X'_{(i)}X_{(i)}\right)^{-1}\right]}{\det\left[S^2\left(X'X\right)^{-1}\right]}$$

$$= \frac{\left(\dfrac{n-m-1-r_i^2}{n-m-2}\right)^{m+1}}{(1-h_i)} \tag{3.42}$$

em que $X_{(i)}$ é a matriz do modelo obtida após a eliminação da i-ésima observação amostral.

Um valor não muito preciso para determinar pontos influentes é dado por $|COVRATIO_i - 1| > 3(m+1)/n$.

### 3.5.6 Comandos SAS

Felizmente todas estes métodos de diagnóstico em regressão linear podem ser obtidas utilizando duas opções simples do comandos model: $r$ e *influence*. Apresentamos na seqüência um exemplo do programa SAS utilizado para obter o diagnóstico de regressão para o exemplo do volume de madeira das árvores.

---

```
/*Exemplo do proc reg para realizar análise de diagnose em modelos de regressão linear
múltipla.*/
proc reg data=arvores;
```

```
    model y=x1 x2 x3/r influence;
run;quit;
```

## 3.6 Exercícios

1. Utilize os dados do exemplo da amostra de $n = 10$ árvores e ajuste o seguinte modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} + \beta_5 \frac{1}{X_{3i}} + \epsilon_i$$

2. Existe alguma variável redundante? Se houver utilize os métodos de seleção de modelos apresentados neste capítulo e determine qual é o melhor modelo.

3. Os métodos de seleção de modelo chegaram a um mesmo modelo?

4. Para o modelo final utilizar as opções apresentadas e verificar a qualidade da predição, fazer o gráfico dos valores preditos e do intervalos de confiança (clm e cli) e plotar os resíduos em relação aos valores preditos na abscissa.

5. Utilize variáveis candidatas diferentes das apresentadas no exercício (1) e aplique os métodos de seleção de modelos. Você chegou a um modelo melhor do que o anteriormente obtido? Justifique devidamente suas conclusões.

6. Utilizando os dados da amostra de $n = 10$ árvores ajuste o modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} + \beta_5 \frac{1}{X_{3i}} + \epsilon_i$$

   Faça a análise de diagnose e verifique se existe alguma observação influente. Justifique devidamente suas conclusões.

# Capítulo 4

# Regressão Não-Linear

Outro assunto extremamente importante para os pesquisadores em geral é o ajuste de regressões não-lineares em suas pesquisas aplicadas. Temos o objetivo de apresentar neste capítulo as principais idéias sobre os processos de estimação de parâmetros de modelos não-lineares e os comandos do *proc nlin* para realizar esta tarefa. O que devemos considerar é que os modelos não-lineares nos parâmetros têm uma maior plasticidade e portanto são considerados mais apropriados para modelarem fenômenos biológicos.

Neste capítulo vamos discutir um pouco sobre métodos de estimação de parâmetros de modelos não-lineares e sobre a sintaxe do proc nlin. Vamos apresentar programas de modelos de *Response Plateau* linear e não-linear. Ambos são não-lineares nos parâmetros, mas descrevem curvas lineares e quadráticas, respectivamente, além do *plateau* no ponto de junção dos segmentos, que é uma linha reta paralela à abscissa.

Os procedimentos de estimação não-linear são em geral iterativos. O processo deve iniciar para um valor específico inicial de seus parâmetros e a soma de quadrado do resíduo é avaliada. Então uma nova estimativa dos parâmetros é obtida, buscando-se minimizar a soma de quadrados do resíduo. Este processo é repetido até que este mínimo seja alcançado. Vários algoritmos e métodos existem para realizar este processo de estimação. Não faremos uma descrição detalhada destes métodos, que aceleram a convergência e são eficientes para estimarmos os parâmetros que conduzem ao mínimo global para a soma de quadrados de resíduos, por causa de as di-

ficuldades teóricas do assunto ultrapassarem o limite estipulado para este material.

## 4.1 Introdução aos Modelos Não-Lineares

Um modelo é considerado não-linear nos parâmetros e esta classificação não é influenciada pela função matemática descrita (hipérbole, parábola, etc.). Como já dissemos no capítulo 3, se as derivadas parciais forem funções dos próprios parâmetros, teremos um modelo não-linear. Podemos ter múltiplos parâmetros neste modelo ou apenas um e da mesma forma, podemos ter apenas uma variável regressora ou mais de uma. Assim, $Y = \alpha\beta^Z$ é um modelo não-linear com dois parâmetros $\alpha$ e $\beta$ e $Y = \alpha + \beta Z^2$ é um modelo linear, independentemente de a função descrever uma parábola, pois este modelo é linear nos parâmetros $\alpha$ e $\beta$.

Os detalhes computacionais envolvidos nos procedimentos não-lineares são muito complexos. Vamos simplificar o máximo que pudermos, sem no entanto deixarmos de ter o rigor necessário. Seja o modelo não-linear $F$ definido de forma geral para o vetor de parâmetros $\underset{\sim}{\beta} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_m]'$ e para o vetor de variáveis regressoras da j-ésima unidade amostral $\underset{\sim}{Z}'_j = [Z_{1j} \quad Z_{2j} \quad \cdots \quad Z_{pj}]$ por

$$Y_j = F_j\left(\underset{\sim}{\beta}, \underset{\sim}{Z}_j\right) + \epsilon_j. \tag{4.1}$$

Podemos expressar este modelo em notação matricial por:

$$\underset{\sim}{Y} = \underset{\sim}{F}\left(\underset{\sim}{\beta}\right) + \underset{\sim}{\epsilon}. \tag{4.2}$$

em que podemos expressar o vetor do modelo $\underset{\sim}{F}\left(\underset{\sim}{\beta}\right)$, simplesmente por $\underset{\sim}{F}$.

Para ficar claro a notação que estamos utilizando, consideremos o modelo $Y_j = \alpha\theta^{Z_j} + \epsilon_j$. Neste caso temos um vetor de parâmetros dado por $\underset{\sim}{\beta}' = [\alpha \quad \theta]$ e uma única variável regressora $Z$. O vetor do modelo é dado por:

$$\underset{\sim}{F} = \begin{bmatrix} \alpha\theta^{Z_1} \\ \alpha\theta^{Z_2} \\ \vdots \\ \alpha\theta^{Z_n} \end{bmatrix}$$

O vetor de observações é dado por:

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Finalmente, o vetor de resíduos é dado por:

$$\underset{\sim}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

O modelo pode ser escrito por:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \alpha\theta^{Z_1} \\ \alpha\theta^{Z_2} \\ \vdots \\ \alpha\theta^{Z_n} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Um dos métodos utilizados baseia-se na minimização da soma de quadrados dos resíduos $L\left(\underset{\sim}{\beta}\right) = \underset{\sim}{\epsilon}'\underset{\sim}{\epsilon}$. Substituindo $\underset{\sim}{\epsilon} = \underset{\sim}{Y} - \underset{\sim}{F}$ e derivando com respeito a $\underset{\sim}{\beta}$, obtivemos:

$$L\left(\underset{\sim}{\beta}\right) = \underset{\sim}{\epsilon}'\underset{\sim}{\epsilon} = \left(\underset{\sim}{Y} - \underset{\sim}{F}\right)'\left(\underset{\sim}{Y} - \underset{\sim}{F}\right) = \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{Y}'\underset{\sim}{F} + \underset{\sim}{F}'\underset{\sim}{F}$$

$$\frac{\partial L}{\partial \underset{\sim}{\beta}} = \frac{-\partial 2\underset{\sim}{Y}'\underset{\sim}{F}}{\partial \underset{\sim}{\beta}} + \frac{\partial \underset{\sim}{F}'\underset{\sim}{F}}{\partial \underset{\sim}{\beta}}$$

Mas,

$$\frac{-\partial 2\underset{\sim}{Y}'\underset{\sim}{F}}{\partial \underset{\sim}{\beta}} = \frac{-\partial 2\underset{\sim}{Y}'\underset{\sim}{F}}{\partial \underset{\sim}{F}} \times \frac{\partial \underset{\sim}{F}}{\partial \underset{\sim}{\beta}} = -2\underset{\sim}{Y}'X$$

em que $X = \partial \underset{\sim}{F}/\partial \underset{\sim}{\beta}$ é a matriz de derivadas parciais, em que cada coluna é formada pela derivada da função linear em relação aos parâmetros.

Também podemos simplificar $\partial \underset{\sim}{F}'\underset{\sim}{F}/\partial \underset{\sim}{\beta}$ por:

$$\frac{\partial \underset{\sim}{F}'\underset{\sim}{F}}{\partial \underset{\sim}{\beta}} = \frac{\partial \underset{\sim}{F}'\underset{\sim}{F}}{\partial \underset{\sim}{F}} \times \frac{\partial \underset{\sim}{F}}{\partial \underset{\sim}{\beta}} = 2\underset{\sim}{F}'X$$

Logo,

$$\frac{\partial L}{\partial \underset{\sim}{\beta}} = -2\underset{\sim}{Y}'X + 2\underset{\sim}{F}'X$$

Igualando a zero a primeira derivada, temos as equações normais para os modelos não-lineares:

$$X'\underset{\sim}{F} = X'\underset{\sim}{Y} \tag{4.3}$$

Como $\underset{\sim}{F}$ e $X$ são funções de $\underset{\sim}{\beta}$, então uma forma fechada para a solução, em geral, não existe. Então devemos utilizar um processo iterativo. Para isso precisamos de um valor inicial para o vetor de parâmetros, que deve ser melhorado continuamente até que a soma de quadrados de resíduos $\underset{\sim}{\epsilon}'\underset{\sim}{\epsilon}$ seja minimizada.

Se considerarmos o modelo $Y_j = \alpha\theta^{Z_j} + \epsilon_j$, que utilizamos anteriormente para ilustrar alguns aspectos do modelo, podemos construir a matriz $X$ das derivadas parciais facilmente. Sejam as derivadas parciais $\partial Y_j/\partial\alpha = \theta^{Z_j}$ e $\partial Y_j/\partial\theta = Z_j\alpha\theta^{(Z_j-1)}$

$$X = \begin{bmatrix} \theta^{Z_1} & Z_1\alpha\theta^{(Z_1-1)} \\ \theta^{Z_2} & Z_2\alpha\theta^{(Z_2-1)} \\ \vdots & \vdots \\ \theta^{Z_n} & Z_n\alpha\theta^{(Z_n-1)} \end{bmatrix}$$

As equações normais para este exemplo são:

$$\begin{bmatrix} \theta^{Z_1} & \dots & \theta^{Z_n} \\ Z_1\alpha\theta^{(Z_1-1)} & \dots & Z_n\alpha\theta^{(Z_n-1)} \end{bmatrix} \begin{bmatrix} \alpha\theta^{Z_1} \\ \alpha\theta^{Z_2} \\ \vdots \\ \alpha\theta^{Z_n} \end{bmatrix} =$$

$$= \begin{bmatrix} \theta^{Z_1} & \dots & \theta^{Z_n} \\ Z_1\alpha\theta^{(Z_1-1)} & \dots & Z_n\alpha\theta^{(Z_n-1)} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Devemos iniciar o processo iterativo para um determinado valor inicial $\underset{\sim}{\beta_0}$. Para o valor corrente (k-ésimo passo do processo iterativo) do vetor de parâmetros, devemos calcular a matriz $X$ e estimar o vetor de resíduos por $\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{F}\left(\underset{\sim}{\beta_k}\right)$. No ponto inicial ($k = 0$), avaliamos $X$ e o vetor de resíduos, considerando o valor arbitrário do vetor de parâmetros especificado. Neste caso, se $SQE\left(\underset{\sim}{\beta_k}\right) = \underset{\sim}{e}'\underset{\sim}{e}$ for a soma de quadrados dos resíduos avaliada na k-ésima iteração, então $X$ e $\underset{\sim}{Y}$ são usados para calcular um vetor $\underset{\sim}{\Delta}$ de tal forma que

$$SQE\left(\underset{\sim}{\beta_k} + \lambda\underset{\sim}{\Delta}\right) < SQE\left(\underset{\sim}{\beta_k}\right)$$

para uma constante $\lambda$ qualquer.

Existem quatro métodos implementados no SAS. Estes quatro métodos diferem na forma como $\underset{\sim}{\Delta}$ é calculado para propiciar as trocas no vetor de parâmetros. De uma forma geral os critérios básicos são:

$$\begin{cases} \text{Gradiente:} & \underset{\sim}{\Delta} = X'\underset{\sim}{e} \\ \text{Gauss-Newton:} & \underset{\sim}{\Delta} = (X'X)^{-}X'\underset{\sim}{e} \\ \text{Newton:} & \underset{\sim}{\Delta} = G^{-}X'\underset{\sim}{e} \\ \text{Marquardt:} & \underset{\sim}{\Delta} = [X'X + \delta\text{diag}(X'X)]^{-}X'\underset{\sim}{e} \end{cases} \tag{4.4}$$

em que $(X'X)^-$ é uma inversa generalizada. Pode ser uma inversa reflexiva ($g_2$), mas o ideal é que seja uma inversa de Moore-Penrose ($g_4$).

Os métodos Gauss-Newton e Marquardt realizam a regressão dos resíduos em relação as primeiras derivadas do modelo não-linear em relação aos parâmetros, até que haja a convergência. O método de Newton faz a regressão destes resíduos em relação a uma função das segundas derivadas do modelo não-linear com relação aos parâmetros ($G^-$).

### 4.1.1 Método do Gradiente

Este método é baseado no gradiente ou grau de variação de $\underset{\sim}{\epsilon}'\underset{\sim}{\epsilon}$. Seja $\underset{\sim}{\beta_k}$ a estimativa do vetor de parâmetros na k-ésima iteração do processo. Assim, este gradiente é definido por:

$$\frac{1}{2}\frac{\partial L\left(\underset{\sim}{\beta_k}\right)}{\partial \underset{\sim}{\beta_k}} = -\underset{\sim}{X'Y} + \underset{\sim}{X'F} = -\underset{\sim}{X'e}$$

pois $X$ e $\underset{\sim}{F}$ são avaliados no ponto $\underset{\sim}{\beta_k}$.

A quantidade $-\underset{\sim}{X'e}$ é o gradiente para o qual $\underset{\sim}{\epsilon}'\underset{\sim}{\epsilon}$ cresce. Sendo assim, $\underset{\sim}{\Delta} = \underset{\sim}{X'e}$ é o grau de variação para o método de gradiente. Para utilizarmos o método do gradiente devemos inicialmente estipular um valor arbitrário para o vetor de parâmetros, digamos $\underset{\sim}{\beta_0}$. Calculamos $\underset{\sim}{e}$ e $\underset{\sim}{\Delta}$. Assim, podemos obter o valor do parâmetro no (k+1)-ésimo passo, tomando a estimativa do k-ésimo passo anterior por:

$$\underset{\sim}{\beta_{k+1}} = \underset{\sim}{\beta_k} + \lambda\underset{\sim}{\Delta} \tag{4.5}$$

em que o escalar $\lambda$ é escolhido no k-ésimo passo para que

$$SQE\left(\underset{\sim}{\beta_k} + \lambda\underset{\sim}{\Delta}\right) < SQE\left(\underset{\sim}{\beta_k}\right). \tag{4.6}$$

O método do gradiente possui convergência muito lenta e, em geral, não é utilizado para estimar parâmetros dos modelos não-lineares. Quando, no entanto, as estimativas iniciais são pobres, este método se torna particularmente útil.

### 4.1.2 Método de *Newton*

O método de *Newton* utiliza a segunda derivada do erro em relação aos parâmetros e obtém o vetor $\underset{\sim}{\Delta}$ por:

$$\underset{\sim}{\Delta} = G^{-} X' \underset{\sim}{e} \tag{4.7}$$

em que

$$G = (X'X) + \sum_{j=1}^{n} H_j \left( \underset{\sim}{\beta_k} \right) e_j \tag{4.8}$$

sendo que a matriz $H_j$, de dimensão $r \times r$, avaliada para o vetor de parâmetros $\underset{\sim}{\beta_k}$ no k-ésimo passo para a j-ésima observação amostral, é a matriz Hessiana do vetor de erros $\underset{\sim}{\epsilon}$. O elemento $(\ell, k)$ desta matriz, $[H_j]_{\ell k}$, é dado por:

$$[H_j]_{\ell k} = \left[ \frac{\partial^2 \epsilon_j}{\partial \beta_\ell \partial \beta_k} \right]_{\ell k} \tag{4.9}$$

Estimado o vetor $\underset{\sim}{\Delta}$, devemos aplicar as equações (4.5) e (4.6) para obtermos uma nova equação e recalcularmos o vetor de parâmetros.

Para o exemplo anterior, considerando o modelo $Y_j = \alpha\theta^{Z_j} + \epsilon_j$, a matriz de segundas derivadas para a j-ésima observação é:

$$H_j = \begin{bmatrix} 0 & -Z_j\theta^{(Z_j-1)} \\ -Z_j\theta^{(Z_j-1)} & -Z_j(Z_j-1)\alpha\theta^{(Z_j-2)} \end{bmatrix}$$

### 4.1.3 Método de *Gauss-Newton*

O método de *Gauss-Newton* usa a expansão em série de *Taylor* do vetor de funções

$$\underset{\sim}{F} \left( \underset{\sim}{\beta} \right) = \underset{\sim}{F} \left( \underset{\sim}{\beta_0} \right) + X \left( \underset{\sim}{\beta} - \underset{\sim}{\beta_0} \right) + \cdots$$

em que a matriz de primeiras derivadas $X$ é avaliada no ponto $\underset{\sim}{\beta_0}$.

Se substituirmos os dois termos desta expansão nas equações normais obtemos

$$\begin{aligned} X'\underset{\sim}{F}\left(\underset{\sim}{\beta}\right) &= X'\underset{\sim}{Y} \\ X'\left[\underset{\sim}{F}\left(\underset{\sim}{\beta_0}\right) + X\left(\underset{\sim}{\beta} - \underset{\sim}{\beta_0}\right)\right] &= X'Y \\ X'X\left(\underset{\sim}{\beta} - \underset{\sim}{\beta_0}\right) &= X'Y - X'\underset{\sim}{F}\left(\underset{\sim}{\beta_0}\right) \\ X'X\underset{\sim}{\Delta} &= X'\underset{\sim}{e} \end{aligned}$$

e portanto,

$$\underset{\sim}{\Delta} = (X'X)^{-}X'\underset{\sim}{e} \tag{4.10}$$

Estimado o valor de $\underset{\sim}{\Delta}$ para o vetor $\underset{\sim}{\beta_0}$, aplicam-se as equações (4.5) e (4.6) para se obter o vetor de estimativas do passo 1. O processo é repetido um determinado número de vezes até que o vetor de estimativas não se altere mais dentro de uma precisão pré-estipulada.

### 4.1.4 Método de *Marquardt*

O método de *Marquardt* mantém um compromisso entre o método de *Gauss-Newton* e o método do gradiente. A fórmula de atualização do vetor de parâmetros é dada por:

$$\underset{\sim}{\Delta} = \left[(X'X) + \delta \mathrm{diag}(X'X)\right]^{-} X'\underset{\sim}{e} \tag{4.11}$$

Se $\delta \to 0$, há uma aproximação ao método de *Gauss-Newton* e se $\delta \to \infty$, há uma aproximação ao método do gradiente. Por padrão o *proc nlin* começa com valor de $\delta = 10^{-7}$. Se $SQE\left(\underset{\sim}{\beta_0} + \underset{\sim}{\Delta}\right) < SQE\left(\underset{\sim}{\beta_0}\right)$, então $\delta = \delta/10$ na próxima iteração; se por outro lado ocorrer o contrário, ou seja, se $SQE\left(\underset{\sim}{\beta_0} + \underset{\sim}{\Delta}\right) > SQE\left(\underset{\sim}{\beta_0}\right)$, então $\delta = 10\delta$. Assim, se a soma de quadrados do resíduo decresce a cada iteração, estaremos utilizando essencialmente o método de *Gauss-Newton*; se ocorrer o contrário o valor de $\delta$ é aumentado em cada iteração, sendo que passaremos a utilizar o método de gradiente.

### 4.1.5 Tamanho do passo da iteração

Devemos estipular o tamanho do passo que daremos em cada iteração. Assim, se $SQE\left(\underset{\sim}{\beta_k} + \lambda\underset{\sim}{\Delta}\right) > SQE\left(\underset{\sim}{\beta_k}\right)$, começando com $\lambda = 1$, devemos reduzir o valor pela metade em cada passo $SQE\left(\underset{\sim}{\beta_k} + 0,5\underset{\sim}{\Delta}\right)$, $SQE\left(\underset{\sim}{\beta_k} + 0,25\underset{\sim}{\Delta}\right)$, e assim por diante até que um quadrado médio do resíduo menor seja encontrado. Podemos muitas vezes encontrar dificuldades em obter avanços na redução da soma de quadrados dos resíduos. Quando isso acontece, o SAS interrompe o processo e comunica ao usuário da não ocorrência de ganhos na redução do $SQE$ com no passo atual da iteração. As possíveis causas podem ser: derivadas mal especificadas e valores iniciais inadequados.

## 4.2 O *Proc Nlin*

O *proc nlin* é o procedimento SAS apropriado para ajustarmos modelos não-lineares. Este procedimento possui além dos métodos descritos anteriormente uma quinta opção, o método de DUD. Este método é livre de derivadas, ou seja, não utiliza a matriz Jacobiana $X$. Assim, o usuário não precisa especificar as derivadas parciais. Isso não é uma grande vantagem, pois nas novas versões, o SAS faz o cálculo numérico das derivadas parciais necessárias, se elas não forem especificadas.

Vamos ilustrar nesta seção os comandos básicos para ajustarmos um modelo de regressão não-linear utilizando o *proc nlin*. Vamos especificar a forma de entrar com o modelo e com as derivadas parciais e, também, como escolher os métodos de estimação a ser utilizado. Antes de fazermos isso, devemos fazer algumas considerações a respeito de como atribuir valores iniciais para os parâmetros. Podemos utilizar, entre outras possibilidades, estimativas publicadas na literatura especializada, que utilizam modelos e conjuntos de dados similares aos de nossa pesquisa. Se o modelo pode ser linearizado, ignorando o fato de ter resíduos aditivos, podemos aplicar uma transformação para linearizá-lo e então, ajustar, o modelo linear resultante. As estimativas de quadrados mínimos, devidamente transformadas

para a escala original, quando for o caso, são utilizadas como valores iniciais. Algumas vezes, antes da linearização, podemos efetuar algum tipo de reparametrização e proceder da mesma forma. Os processos iterativos possuem convergência bem mais rápida, quando os valores iniciais estão mais próximos das estimativas de mínimos quadrados.

Para apresentarmos os comandos básicos do *proc nlin*, vamos utilizar os dados da Tabela 3.2 e o seguinte modelo não-linear nos parâmetros:

$$y_i = \alpha\beta^{x_i} + \epsilon_i \tag{4.12}$$

Neste caso temos $n = 8$ árvores e as seguintes derivadas parciais em relação aos parâmetros $\alpha$ e $\beta$: $\partial y_i/\partial\alpha = \beta^{x_i}$ e $\partial y_i/\partial\beta = x_i\alpha\beta^{(x_i-1)}$. Como estas derivadas parciais são funções dos parâmetros $\alpha$ e $\beta$, temos um modelo não-linear nos parâmetros caracterizado. Vamos atribuir valores iniciais arbitrários iguais a $0,5$ e $1,8$ para $\alpha$ e $\beta$, respectivamente. Poderíamos ter linearizado este modelo facilmente aplicando a função logaritmo, ignorando é claro o fato de o erro ser aditivo. Este seria um artifício para obtermos valores iniciais mais acurados. O modelo linearizado é dado por $\ln(y_i) = \ln(\alpha) + \beta\ln(x_i) + \epsilon_i^*$, que poderia ser rescrito por $z_i = A + \beta w_i + \epsilon_i^*$. Neste caso a estimativa do parâmetro $A$ do modelo linear dever ser transformada para a escala original por $\hat{\alpha} = \exp(\hat{A})$. A estimativa de $\beta$ não precisa ser modificada, pois o parâmetro $\beta$ não foi alterado pela transformação efetuada. Isto é deixado a cargo do leitor na forma de exercício. O programa SAS resultante é:

---

```
Data regnlm1;
input X Y;
Cards;
0.1 0.88
0.2 0.90
0.3 0.99
0.5 1.12
0.8 1.40
1.0 1.62
1.5 2.20
```

```
2.0 3.10
;
Proc nlin Method=Gauss;
    Parms a=0.5 b=1.8;
    Model y=a*(b**x);
    Der.a=b**x;
    Der.b=a*x*(b**(x-1));
run;quit;
```

Neste programa $a$ e $b$ representam os parâmetros $\alpha$ e $\beta$, respectivamente; os comandos *<der.a=b**x;>* e *<der.b=a*x*(b**(x-1));>* indicam as derivadas parciais da variável resposta em relação aos parâmetros $\alpha$ e $\beta$, respectivamente; o modelo é especificado com o comando *<model y=a*(b**x);>*.

O SAS utilizou 4 iterações e apresentou uma mensagem que o ajuste do modelo atingiu convergência. O modelo ajustado foi $\hat{y}_i = 0,8117 \times 1,9542^{x_i}$. Ambos os parâmetros foram significativamente diferentes de zero, pois os intervalos assintóticos de 95% de confiança não abrangeram o valor 0. O intervalo assintótico de 95% confiança para o parâmetro $\alpha$ foi $[0,7903; 0,8330]$ e para o parâmetro $\beta$, $[1,9206; 1,9877]$. O $R^2$ do modelo pode ser estimado por $R^2 = 1 - SQRes/SQTotal$. Para este exemplo, o $R^2 = 1 - 0,00276/4,2178 = 0,9993$, indicando que $99,93\%$ da variação do crescimento das plantas foi explicado pelo modelo de regressão.

Vamos ilustrar o *proc nlin* com o ajuste de mais um modelo aos dados da Tabela 3.2 dado por:

$$y_i = \alpha x_i^{\beta} + \epsilon_i \tag{4.13}$$

As derivadas parciais em relação a cada parâmetro são dadas pelas funções $\partial y_i/\partial\alpha = x_i^{\beta}$ e $\partial y_i/\partial\beta = \alpha x_i^{\beta} \ln(x_i)$. O programa correspondente a este exemplo é dado por:

```
Data regnlm2;
input X Y;
Cards;
```

```
0.1 0.88
0.2 0.90
0.3 0.99
0.5 1.12
0.8 1.40
1.0 1.62
1.5 2.20
2.0 3.10
;
Proc nlin Method=Gauss maxiter=500;
    Parms a=0.5 b=1.8;
    Model y=a*(x**b);
    Der.a=x**b;
    Der.b=a*x**b*log(x);
run;quit;
```

Especificamos um número máximo de iterações igual a 500. O padrão do SAS, se nada for especificado, é 100. Neste caso ocorreu a convergência com apenas 8 iterações. Este comando (*maxiter=nit*) se torna útil apenas quando o valor inicial é precário, requerendo um número grande de iterações, principalmente se houver correlações elevadas entre os estimadores dos parâmetros. Neste exemplo, o modelo ajustado foi $\hat{y}_i = 1,8548x_i^{0,575}$, sendo que este ajuste foi um pouco inferior ao ajuste do modelo anterior. Isto pode ser constatado observando o valor do coeficiente de determinação $R^2 = 89,61\%$ deste modelo e comparando com o valor anteriormente obtido. Os dois modelos ajustados estão apresentados na Figura 4.1. Devemos procurar sempre, além de um bom ajuste, modelos que possam ter uma relação com o fenômeno que estamos estudando. Apesar dos bons ajustes alcançados, podemos para este exemplo escolher, do ponto de vista biológico, melhores modelos não-lineares.

## 4.3 Modelos Segmentados

Dentre os modelos segmentados existe o modelo de "*response plateau*" que é muito utilizado na pesquisa em diversas áreas. Esse modelo possui dois segmentos, sendo que o primeiro descreve uma curva crescente ou de-

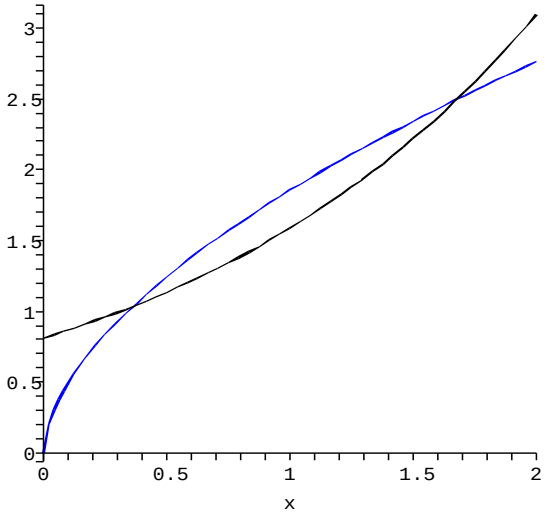


Figura 4.1: Modelos não lineares ajustados - modelo $\hat{y}_i = 1,8548x_i^{0,575}$ iniciando pela origem e modelo $\hat{y}_i = 0,8117 \times 1,9542^{x_i}$ iniciando pelo ponto $0,8117$.

crescente até uma determinada altura da ordenada ($P$) que é o platô. A partir desse ponto o valor $Y$ assume um valor constante $P$. O ponto correspondente ao valor $P$ na abscissa é o ponto $X_0$, que também é um parâmetro a ser estimado. Vários modelos podem ser utilizados para modelar o comportamento da curva entre a origem e o ponto onde se encontra o platô. Nesta seção apresentamos o exemplo do manual do SAS (*proc nlin*) com um modelo quadrático anterior ao platô. Na Figura 4.2 é apresentado um exemplo de um modelo de *response plateau*, destacando-se os pontos $X_0$ e $P$.

Para ilustrarmos o ajuste de um modelo bi-segmentado desta natureza é considerado o exemplo apresentado no manual do SAS, relativo ao *proc nlin*. Seja para isso o seguinte modelo quadrático de response platô:

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \text{ se } X_i < X_0 \\ P \text{ se } X_i \geq X_0 \end{cases} \tag{4.14}$$

Para valores de $X < X_0$, os de $Y$ são explicados por um modelo quadrático (parábola) e para valores de $X \geq X_0$, a equação explicativa é constante

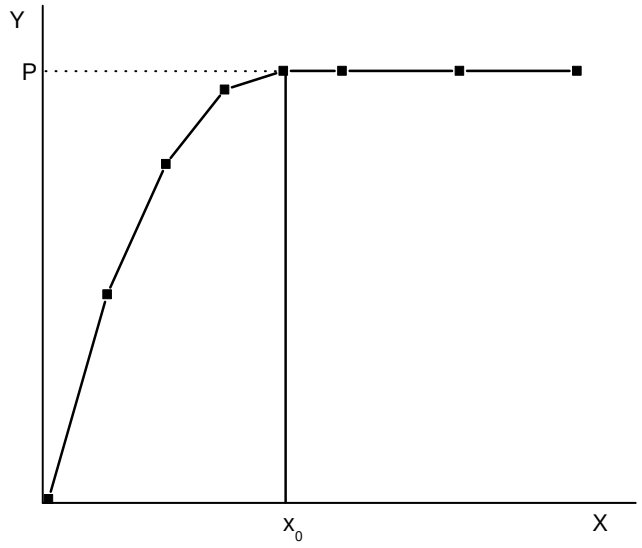


Figura 4.2: Modelo segmentado considerando um plateau no ponto $X = X_0$ com valor de $Y = P$ e um modelo crescente para $X < X_0$.

e paralela a abscissa. O ponto $X_0$ é considerado desconhecido e deve ser estimado juntamente com os demais parâmetros do modelo. Este ponto representa a junção do segmento quadrático com o segmento de platô. As curvas devem ser contínuas (os dois segmentos devem se encontrar em $X_0$) e suavizada, ou seja, as primeiras derivadas com relação a $X$ nos dois segmentos devem ser a mesma no ponto $X_0$. Essas condições implicam em algumas conseqüências descritas a seguir.

A primeira derivada de $Y$ em relação a $X$ no modelo quadrático é dada por:

$$\frac{dY_i}{dX_i} = \beta_1 + 2\beta_2 X_i$$

Se igualarmos esta deriva a zero, resolvermos a equação resultante em $X$ e substituirmos o valor de $X$ por $X_0$, ponto em que a curva deve ser contínua e suavizada, obtemos:

$$X_0 = \frac{-\beta_1}{2\beta_2}$$

Substituindo esse valor na equação (4.14) obtemos o máximo, que corresponde ao platô almejado. Assim, este platô é dado por:

$$Y = P = \beta_0 + \beta_1 X_0 + \beta_2 X_0^2 = \beta_0 - \frac{\beta_1^2}{2\beta_2} + \frac{\beta_1^2 \beta_2}{4\beta_2^2} = \beta_0 - \frac{\beta_1^2}{4\beta_2}$$

Neste caso temos apenas três parâmetros efetivos, pois tanto $X_0$, quanto $P$ são determinados a partir de $\beta_0$, $\beta_1$ e $\beta_2$. Este é um modelo não linear nos parâmetros, pois as derivadas parciais de $Y$ são funções dos parâmetros em alguns casos, justificando o uso do *proc nlin*. O programa final é apresentado na seqüência. Podemos destacar que ele é dividido em duas partes: a primeira com a parte quadrática polinomial e a segunda, com a parte do platô. Em cada ciclo do processo iterativo imprimimos nos resultados, juntamente com os demais parâmetros, as estimativas de $X_0$ e de $P$. Utilizamos o *proc plot* para produzir um gráfico de baixa qualidade dos valores ajustados. Neste modelo, $a$ representa $\beta_0$, $b$ representa $\beta_1$ e $c$ representa $\beta_2$.

```
/* Ajuste do modelo segmentado usando o NLIN */
/* y= a + b*x + c*x*x e y=P se x>x0 */
/* restrição de continuidade: P= a +b*x0+c*x0*x0 */
/* restrição de suavização: 0=b+2*c*x0, então, x0=-b/(2*c) */
title "Modelo quadrático com platô";
data reg;
input x y @@;
cards;
  1 0.46 2 0.47 3 0.57 4 0.61 5 0.62 6 0.68 7 0.69
  8 0.78 9 0.70 10 0.74 11 0.77 12 0.78 13 0.74 13 0.80
 15 0.80 16 0.78
;
proc nlin data=reg;
    parms a=0.45 b=0.05 c=-0.0025;
    file print;
    x0=-0.5*b/c; /*estimação do ponto comum */
    db=-0.5/c; /* derivada de xo em relação a b */
```

```
    dc=0.5*b/c**2; /* derivada de xo em relação a c */
    if x<x0 then /* parte quadrática do modelo */
    do;
       model y=a+b*x+c*x**2;
       der.a=1;
       der.b=x;
       der.c=x**2;
    end;
    else /* parte do modelo relativo ao platô de resposta*/
    do;
       model y=a+b*x0+c*x0**2;
       der.a=1;
       der.b=x0+b*db+2*c*x0*db;
       der.c=b*dc+x0*x0+2*c*x0*dc;
    end;
    if _obs_=1 then
    do;
       plateau=a+b*x0+c*x0**2;
       put x0= plateau=;
    end;
    output out=reg1 predicted=yp;
run;quit;
proc plot data=reg1;
    plot y*x yp*x="*"/overlay vpos=35;
run;quit;
```

O modelo ajustado foi $\hat{Y}_i = 0,3921 + 0,0605X_i - 0,00237X_i^2$ se $X_i < 12,7477$ e $\hat{Y}_i = 0,7775$, caso contrário. As estimativas de $\beta_0$ e $\beta_1$ foram significativamente ($P < 0,05$) superiores a zero e a de $\beta_2$, significativamente inferior a zero. Estes resultados foram obtidos analisando os intervalos de confiança assintóticos. O $R^2$ do modelo foi igual a $1 - 0,0101/0,1869 = 0,9460$.

Outro modelo que aparece freqüentemente na literatura é o *linear response plateau* ou LRP. Este modelo possui um segmento de reta antes do ponto de junção ($X_0$) com o platô e é dado por:

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \epsilon_i \text{ se } X_i \leq X_0 \\ P + \epsilon_i \text{ se } X_i > X_0 \end{cases} \tag{4.15}$$

É comum utilizarmos uma variável binária (*Dummy*) para representarmos o modelo. Neste caso utilizaremos a variável $Z_i$, que receberá o valor 1 se $X_i \leq X_0$, ou 0 se $X_i > X_0$. Este modelo poderá ser reescrito por $Y_i = (\beta_0 + \beta_1 X_i)\, Z_i + P(1 - Z_i)$. Para termos continuidade em $X_0$, devemos igualar $\beta_0 + \beta_1 X_0 = P$, ou seja, $X_0 = (P - \beta_0)/\beta_1$.

Neste caso temos um modelo com três parâmetros ($\beta_0$, $\beta_1$ e $P$). Diferentemente do modelo anterior, $P$ não pôde ser expresso em função dos demais parâmetros. Apesar de as variáveis parciais não dependerem dos parâmetros, este é um modelo não-linear uma vez que a matriz Jacobiana depende de $X_0$ para ser construída, sendo que $X_0$ é função de $\beta_0$, $\beta_1$ e de $P$. Assim, as derivadas parciais, dadas por $\partial Y_i/\partial\beta_0 = Z_i$, $\partial Y_i/\partial\beta_1 = X_i Z_i$ e $\partial Y_i/\partial P = 1 - Z_i$, dependem dos parâmetros por meio de $X_0$. A cada passo do processo iterativo, o parâmetro $X_0$ é estimado e a matriz do modelo é composta, pois os $Z_i$'s ficam completamente definidos.

Utilizamos os recursos do *proc nlin* para estimar os parâmetros deste modelo segmentado do tipo LRP. O resultado final está apresentado na seqüência para um conjunto simulado de dados. Neste conjunto de dados os parâmetros são $\beta_0 = 2$, $\beta_1 = 2$ e $P = 10$.

---

```
/* Ajuste do modelo segmentado usando o NLIN */
/* y= a + b*x se x<x0 e y=P se x>=x0 */
/* restrição de continuidade: P= a +b*x0 */
title "Modelo Linear com platô";
data LRP;
input x y;
cards;
1.0     4.10
2.0     5.90
2.5     7.10
3.0     7.80
4.0     9.90
```

```
5.0    10.10
6.0    10.20
7.0     9.80
8.0     9.78
;
proc nlin data=LRP;
    parms a=1 b=2 p=2.0;
    X0=(p-a)/b;
    if x<=x0 then /* Parte não-plateau do modelo */
    do;
       model y=a+b*x;
       der.a=1;
       der.b=X;
    end;
    else /* Parte plateau do modelo */
    do;
       model y=p;
       der.a=0;
       der.b=0;
       der.p=1;
    end;
    if _obs_=1 then /*Para imprimir a saída se for a 1^a observação*/
    do;
       put x0=;
    end;
    output out=saida predicted=yp Residual=Res parms=a b p ess=sqe;
run;quit;
```

---

O modelo ajustado foi $\hat{Y}_i = 2,135 + 1,93X_i$ se $X_i \leq 4,06$ e $\hat{Y}_i = 9,97$ se $X_i > 4,06$. O coeficiente de determinação do modelo foi igual a $R^2 = 99,53\%$. Todos os valores paramétricos estão dentro do intervalo de confiança assintótico construído.

Apresentamos na seqüência um outro exemplo, também simulado, em que temos os parâmetros iguais a $\beta_0 = 5$, $\beta_1 = 2,4$, $P = 29$ e $\sigma^2 = 1$.

---

```
/* Ajuste do modelo segmentado usando o NLIN */
/* y= a + b*x se x<x0 e y=P se x>=x0 */
/* restrição de continuidade: P= a +b*x0 */
```

```
title "Modelo Linear com platô";
data LRP;
input x y;
cards;
    1 8.6264841
    2 8.9408731
    3 11.909886
    4 13.936262
    5 17.945067
    6 18.732450
    7 21.847226
    8 23.769043
    9 27.671300
   10 28.441954
   11 27.811677
   12 30.827451
   13 28.817408
   14 30.665168
   15 28.813364
   16 29.127870
   17 28.218656
   18 28.309338
   19 28.651342
   20 29.230743
;
proc nlin data=LRP;
    parms a=1 b=2 p=2.0;
    X0=(p-a)/b;
    if x<=x0 then /* Parte não-plateau do modelo */
    do;
       model y=a+b*x;
       der.a=1;
       der.b=X;
    end;
    else /* Parte plateau do modelo */
    do;
       model y=p;
       der.a=0;
       der.b=0;
       der.p=1;
    end;
    if _obs_=1 then /*Para imprimir a saída se for a 1ª observação*/
```

```
    do;
       put x0=;
    end;
    output out=saida predicted=yp Residual=Res parms=a b p ess=sqe;
run;quit;
```

O modelo ajustado para este exemplo foi $\hat{Y}_i = 5,0731 + 2,3834X_i$ se $X_i \leq 10,06$ e $\hat{Y}_i = 29,05$ se $X_i > 10,06$. O coeficiente de determinação do modelo foi igual a $R^2 = 98,64\%$. Também neste caso, todos os valores paramétricos estão dentro do intervalo de confiança assintótico construído.

## 4.4 Exercícios

1. Utilize os dados da Tabela 3.2 e o *proc nlin* do SAS para ajustar o seguinte modelo:
$$Y_i = \frac{\alpha}{\beta_0 + \beta_i X_i} + \epsilon_i$$

2. Este modelo se ajustou melhor do que aqueles da seção 4.2? Justifique sua resposta.

3. Tente ajustar um modelo LRP aos dados da Tabela 3.2. Qual foi o modelo encontrado? Este modelo é um modelo LRP? Justifique sua resposta. Plote os dados e verifique se existe uma dispersão dos pontos que justifique a representação por meio de um modelo LRP.

4. Utilize os resíduos gerados no exemplo apresentado em aula do ajuste do modelo LRP e realize a análise gráfica dos resíduos.

5. Busque em sua área de atuação dados que poderiam se enquadrar dentro do modelo segmentado quadrático. Descreva as situações e os possíveis benefícios de ajustar um modelo deste tipo. Se os dados estiverem disponíveis, utilize o programa apresentado em aula para ajustar o modelo de platô de resposta quadrático.

# Capítulo 5

# Análise de Variância para Dados Balanceados

Para realizarmos inferências sobre a hipótese de igualdade entre várias médias dos níveis de algum fator de interesse, utilizamos o teste $F$ da análise de variância (Anava). Esta hipótese pode ser formalizada por:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_\ell = \mu \\ H_1 : \text{pelo menos uma média difere das demais} \end{cases} \tag{5.1}$$

em que $\ell$ é o número de níveis deste fator de interesse e $\mu_i$ é a média do i-ésimo nível, $i = 1, 2, \cdots, \ell$.

Um valor de $F$ observado superior a um valor crítico da distribuição $F$ para um nível $\alpha$ de significância indica que devemos rejeitar a hipótese nula $H_0$; caso contrário, não existirão evidências significativas para rejeitar a hipótese nula. Podemos ter mais de um fator. Neste caso teremos uma hipótese nula para cada fator separadamente. Além disso, estes fatores podem interagir. Se houver algum tipo de interação entre eles, um teste $F$ específico para a hipótese de haver interação irá apresentar efeito significativo da estatística. Também podemos ter efeitos hierarquizados, onde os níveis de um fator $A$, por exemplo, dentro de um determinado nível de outro fator, digamos $B$, são diferentes dos níveis de $A$ em outro nível de $B$. Isto ocorre, por exemplo, quando temos diferentes procedências de eucalipto e dentro de cada procedência, temos diferentes progênies.

Neste capítulo estaremos interessados nestes diferentes modelos estatísticos, contendo um ou mais fatores, cujos efeitos podem ser cruzados ou hierarquizados, porém em uma estrutura experimental balanceada. Entenderemos por estrutura balanceada, aquele conjunto de dados cujo número de observações em cada combinação dos níveis dos fatores é o mesmo. Cada nível de um fator, ou cada nível resultante da combinação dos níveis de dois ou mais fatores, é denominado de casela. Se houver diferenças neste número de observações por casela, teremos dados não balanceados. O procedimento do SAS apropriado para lidar com estas estruturas é o *proc anova*. Se a estrutura é não-balanceada devemos utilizar o *proc glm*.

## 5.1 O Proc Anova

O *proc anova* é o procedimento apropriado para realizarmos análises de variância envolvendo dados balanceados. Podemos utilizar muitas opções específicas entre os comandos deste procedimento. Vamos apresentar na seqüência alguns dos comandos básicos e específicos para ilustrar a sintaxe do *proc anova*.

---

```
proc anova data=conjdados options;
    class variables;
    model dependents=effects / options;
    means effects / options;
    test H=effects E=effect;
    manova H= effects E=effect / options;
    by variables;
run; quit;
```

---

São comandos obrigatórios *<class variables;>* e *<model dependents = effects /options;>*. No primeiro caso, especificamos as variáveis classificatórias após o comando *class*, separadas por espaços em branco. Estas variáveis classificatórias são os fatores da análise. Não devemos especificar as interações entre estes fatores e nem os efeitos aninhados, mas somente os efeitos

principais. Obviamente devemos usar os mesmos nomes especificados no comando *input*. No comando *model* devemos colocar do lado esquerdo da igualdade, as variáveis respostas e do lado direito, as fontes de variação do modelo adotado (*effects*). Ainda podemos especificar algumas opções associadas ao modelo. Estas opções aparecem após a barra (/). Duas opções estão disponíveis no *proc anova*: *nouni* e *intercept*. A opção *nouni* suprime as análises univariadas da saída do programa. Em geral é utilizada de forma associada com o comando *manova*, para realizarmos análises de variância multivariadas. A opção *intercept* ou simplesmente *int* é utilizada quando pretendemos testar hipóteses relativas ao intercepto como um efeito do modelo.

Os demais comandos são opcionais, ou seja, devemos utilizá-los conforme nosso interesse particular em algum tipo de análise. O comando *<means effects /options;>* é utilizado para estimarmos as médias de um determinado fator na análise de variância, podendo ser inclusive um efeito de interação ou hierárquico. Podemos utilizar vários comandos *means*, desde que eles apareçam após o comando *model*. As opções deste comando permitem que façamos testes de comparações múltiplas. Entre as opções podemos destacar: *alpha=p* para determinar o valor da significância $p$ (0,05 é o padrão), *cldiff* para obter os intervalos de confiança de um determinado teste em relação a todas as diferenças entre médias, *clm* para obter os intervalos de confiança dos níveis dos fatores para um determinado teste, *E=effect* para determinar o efeito que irá ser utilizado como erro nos testes de comparações múltiplas, *Bon* para o teste de Bonferroni, *Duncan* para o teste de Duncan, *Dunnett("Controle")* para realizar o teste de Dunnett de um tratamento com o controle especificado entre aspas e entre parênteses após a opção. As opções *GABRIEL*, *LSD* ou *T*, *Scheffe*, *SNK*, *Tukey*, *Waller* são utilizadas para solicitar os testes de Gabriel, $t$ de Student, Scheffé, Student-Newman-Keuls, Tukey e Waller-Duncan, respectivamente. A opção *nosort* é utilizada para solicitar que as médias não sejam ordenadas; a opção *lines*, para listar as médias ordenadas com o indicativo das médias consecutivas não significativamente diferentes por uma linha.

Finalmente, a opção *HovTest=teste* possibilita que seja aplicado o teste de homogeneidade de variâncias para os grupos de tratamentos, no modelo

inteiramente casualizado. Se outros modelos forem especificados, a opção é ignorada. Os testes escolhidos podem ser: *Bartlett, Levene(type=abs| square), BF, OBrien*. O teste *BF* é o de Brown e Forsythe, que é uma variação do teste de Levene que utiliza desvios da mediana; o teste OBrien é também uma variação do teste Levene atribuída a O'Brien. Ferreira (2005)[3] descreve com detalhes estes testes.

O comando *<test H=effects E=effect;>* é bastante útil em modelos com mais de um erro ou em modelos mistos, para realizarmos testes de hipóteses de alguns efeitos da análise de variância (opção H=effects) com um erro particular de interesse (opção E=effect). Os riscos de utilização inadequada são relegados aos usuários. O comando *<manova H= effects E=effect / options;>* possibilita a realização de testes de hipóteses multivariados para os fatores especificados em *H=effects*, utilizando como erro o efeito especificado em *E=effect*. As opções que podemos utilizar são *canonical, printe e printh* entre outras. A opção *printe* é particularmente interessante por proporcionar a estimação das correlações parciais entre as variáveis dependentes, dadas as variáveis independentes (fatores). Finalmente o comando *<by variables;>* permite a obtenção das análises de variâncias para cada grupo das variáveis especificadas após o comando *by*. Esta opção exige que as variáveis, utilizadas no comando *by*, estejam em ordem crescente. Caso isso não seja verdade, é necessário utilizar o *proc sort* antes de chamar o *proc anova*.

Vamos ilustrar algumas formas que podemos utilizar para especificar o modelo de análise de variância. Suponhamos que $A$, $B$ e $C$ sejam fatores de interesse e $Y$ a variável resposta. Podemos especificar diferentes modelos utilizando os seguintes comandos:

a) Exemplos de modelos com efeitos simples: *<model Y=A;>* ou *<model Y=A B;>* ou *<model Y=A B C;>*.

b) Exemplos de efeitos cruzados: *model Y=A B A*B;* ou simplesmente *<model Y=A | B;>*. Neste último caso a | é uma notação geral para a estrutura de efeitos. No exemplo particular significa que o modelo ajustado é função dos efeitos principais e da interação, ou seja, é igual ao primeiro modelo deste item.

c) Exemplos de efeitos hierárquicos: *<model Y=B A(B);>*, indicando que temos um modelo com o fator principal $B$ e com o fator $A$ hierarquizado, dentro dos níveis de $B$. Isto significa que os níveis de $A$ não são os mesmos quando consideramos dois diferentes níveis de $B$. Um outro exemplo onde temos os níveis de $A$ dentro da combinação dos níveis de $B$ e $C$ é dado por: *<model Y=B C A(B C);>*. A sintaxe para este caso no *proc glm* seria: *<model Y=B C A(B*C);>*. Assim, os dois procedimentos diferem pela utilização ou não do asterisco, nos fatores que estão dentro dos parênteses.

d) Exemplos de modelos com efeitos cruzados e hierárquicos: *<model Y=A B(A) C(A) B*C(A);>*

## 5.2 Delineamento Inteiramente Casualizado

Os delineamentos inteiramente casualizados, com um fator, serão utilizados para ilustrarmos inicialmente os comandos básicos do *proc anova*. Para isso, utilizaremos os dados apresentados por Gomes (2000)[5], onde os efeitos no ganho de peso de animais em kg de 4 rações foram comparados. Os dados estão apresentados na Tabela 5.1.

Tabela 5.1: Ganho de peso (gp), em kg, de animais que foram submetidos a uma dieta com determinadas rações. Um delineamento inteiramente casualizado com cinco repetições (animais) e 4 rações foi utilizado (Gomes, 2000)[5].

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 35 | 40 | 39 | 27 |
| 19 | 35 | 27 | 12 |
| 31 | 46 | 20 | 13 |
| 15 | 41 | 29 | 28 |
| 30 | 33 | 45 | 30 |

O modelo de análise de variância adotado é dado por:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{5.2}$$

em que $Y_{ij}$ é o ganho de peso observado no j-ésimo animal para a i-ésima ração, $\mu$ é a constante geral, $\tau_i$ é o efeito da i-ésima ração e $\epsilon_{ij}$ é o efeito do erro experimental suposto normal e independentemente distribuído com média 0 e variância comum $\sigma^2$.

O programa SAS para obtenção da análise de variância do modelo 5.2 é dado por:

```
/* Exemplo da utilização do Proc Anova*/
data dic;
input racoes gp;
cards;
 1 35
 1 19
 1 31
 1 15
 1 30
 2 40
 2 35
 2 46
 2 41
 2 33
 3 39
 3 27
 3 20
 3 29
 3 45
 4 27
 4 12
 4 13
 4 28
 4 30
  ;
proc anova;
    class racoes;
    model gp=racoes;
```

```
    means racoes / snk alpha=0.05 lines;
run; quit; /* fim do programa */
```

Os principais resultados do SAS estão apresentados na seqüência. Neste programa, modelamos o ganho de peso em função do fator rações. Não precisamos especificar nem o erro do modelo e nem a constante geral. Solicitamos as médias de tratamentos e a aplicação do teste SNK para realização das comparações múltiplas. Os resultados da análise de variância estão apresentados nas Tabelas 5.2 e 5.3.

Tabela 5.2: Análise de variância para o delineamento inteiramente casualizado com um fator (rações) com quatro níveis e cinco repetições.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Modelo | 3 | 823,7500 | 274,5833 | 3,99 | 0,0267 |
| Erro | 16 | 1100,0000 | 68,7500 | | |
| total corrigido | 19 | 1923,7500 | | | |
| $R^2$ | 0,4282 | | | | |
| CV | 27,8708 | | | | |
| Média | 29,7500 | | | | |

Tabela 5.3: Análise da variação contendo as fontes de variação do modelo para o delineamento inteiramente casualizado das rações.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Rações | 3 | 823,7500 | 274,5833 | 3,99 | 0,0267 |

O resultado do teste $F$ da análise de variância indica que devemos rejeitar a hipótese nula de igualdade de efeitos das rações. Assim, pelo menos uma delas difere das demais. Devemos utilizar um teste de comparações múltiplas para identificar estas diferenças. Neste exemplo foi utilizado o teste SNK para identificar quais rações diferiram entre si. Na Tabela 5.4 apresentamos o resultado do teste SNK e as respectivas diferenças mínimas significativas (dms). As médias que possuem a mesma letra não são

consideradas significativamente diferentes pelo teste SNK no nível nominal de significância de 5%. Neste caso, as rações 2, 3 e 1 não são estatisticamente diferentes em média, como ocorre também com as rações 3, 1 e 4. No entanto, as rações 2 e 4 são significativamente diferentes ($P < 0,05$).

Tabela 5.4: Teste de SNK e médias para a fonte de variação rações juntamente com as diferenças mínimas significativas dms.

| Grupo | Média | $r_i$ | Rações |
|---|---|---|---|
| A | 39,000 | 5 | 2 |
| A B | 32,000 | 5 | 3 |
| A B | 26,000 | 5 | 1 |
| B | 22,000 | 5 | 4 |

$dms_4$=11,116861, $dms_3$=13,53137, $dms_2$=15,003329.

Um aspecto importante deste teste é apresentado juntamente com os resultados. Esta característica refere-se ao fato de que este teste controla o erro tipo I por experimento sob $H_0$ completa, mas não sob a hipótese nula parcial.

Podemos realizar inferências de interesse sobre parâmetros decorrentes de uma combinação linear das médias por meio dos testes hipóteses e construindo intervalos de confiança. A realização de inferências sobre combinações lineares (usualmente contrastes) de médias, em geral, é o passo seguinte à rejeição da hipótese global da equação (5.1), às vezes denominada hipótese nula completa.

Como o teste $F$, que testa a hipótese global, não informa quais são as médias que diferem entre si, passamos, então, a realizar uma seqüência de testes de hipóteses sobre um conjunto de combinações lineares de médias utilizando os mesmos dados observados. A estes testes estão associados erros de decisão. Se a hipótese nula global for verdadeira e se uma destas hipóteses for rejeitada, estaremos cometendo o erro tipo I. O controle do erro tipo I, no caso de comparações múltiplas, envolve alguns conceitos diferentes. Se por outro lado não rejeitamos uma hipótese que deveria ser rejeitada, estaremos cometendo o erro tipo II. Acontece, também, que as taxas de erro dos tipos I e II, decorrentes da aplicação de um único teste,

têm comportamentos diferentes daquelas associadas à aplicação de uma seqüência de testes.

Um grande número de estratégias existem para garantir uma taxa de erro global $\alpha$ para todas as comparações. Procedimentos de inferência que asseguram uma probabilidade conjunta $1 - \alpha$ contra o erro do tipo I são denominados procedimentos de *inferência simultânea* ou conjunta e procedimentos que asseguram proteção apenas para a comparação que está sendo realizada são denominados procedimentos de *inferência individual*. Nos procedimentos de inferência individual não é feito nenhum ajuste na probabilidade por causa da multiplicidade dos testes.

Algumas definições conduzem a uma taxa de erro que são dependentes da nulidade da hipótese global. Outras conduzem a uma taxa de erro dependente do número de inferências erradas em relação ao número total de inferências feitas. Assim, O'Neill e Wetherill (1971)[9] definem duas maneiras básicas para calcularmos a taxa de erro do tipo I. Uma delas diz respeito à probabilidade de a família de testes conter pelo menos uma inferência errada e a outra, ao número esperado de inferências erradas na família.

De acordo O'Neill e Wetherill (1971)[9] as possibilidades para as taxas de erro observadas são:

i. Taxa de erro por comparação (*comparisonwise error rate*):

$$\frac{\text{Número de inferências erradas}}{\text{Número total de inferências}}$$

ii. Taxa de erro por experimento (*experimentwise error rate*):

$$\frac{\text{Número de experimentos com pelo menos uma inferência errada}}{\text{Número total de experimentos}}$$

Os vários procedimentos de comparações múltiplas possuem diferentes controle do erro tipo I por experimento. O teste Tukey por exemplo, controla a taxa de erro por experimento sob $H_0$ nula e parcial, mas na medida em que o número de níveis do fator aumenta, o teste se torna mais conservador. Assim, este teste possui elevadas taxas de erro tipo II, ou seja, baixo poder quando temos muitos níveis do fator. O teste Duncan e $t$ de Student

são muito liberais e apresentam elevadas taxas de erro tipo I por experimento, com baixas taxas de erro tipo II ou com elevado poder. Por causa de não haver controle do erro tipo I por experimento os elevados poderes não são vantajosos. O teste SNK, como já afirmamos, controla o erro tipo I sob a hipótese de nulidade completa, mas não sob a nulidade parcial. O teste $t$ com proteção de Bonferroni é na maioria das vezes mais conservador do que o teste de Tukey, da mesma forma que ocorre com teste Scheffé quando utilizado no contexto de comparações múltiplas.

Uma importante pressuposição na análise de variância é a homogeneidade de variâncias. Podemos testar hipóteses de igualdade de variâncias facilmente no SAS. Como já mencionamos em outra oportunidade, devemos utilizar a opção *hovtest* do comando *means*. A hipótese de interesse neste caso é dada por:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2 \\ H_1 : \text{pelo menos uma variância difere das demais} \end{cases} \tag{5.3}$$

em que $k$ é o número de níveis do fator de interesse e $\sigma_i^2$ é a variância do i-ésimo nível, $i = 1, 2, \cdots, k$.

Existem vários testes para esta hipótese na literatura. O SAS apresenta a implementação para alguns deles. Vamos descrever estes testes de forma bastante simplificada. Maiores detalhes podem ser vistos em Ferreira (2005)[3]. O teste de Bartlett é um teste de razão de verossimilhanças. Para apresentarmos a estatística deste teste, devemos considerar que $S_i^2$ é o estimador da variância do i-ésimo nível do fator estudado em $n_i$ repetições; $S_p^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2/(n - k)$ é o estimador da variância comum das $k$ populações (ou dos $k$ níveis do fator); e $n = \sum_{i=1}^{k} n_i$ é total de parcelas experimentais. Assim, a estatística

$$\chi_c^2 = \frac{(n-k)\ln(S_p^2) - \sum_{i=1}^{k}\left[(n_i - 1)\ln(S_i^2)\right]}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{n-k}\right]} \tag{5.4}$$

sob $H_0$ possui distribuição assintoticamente de qui-quadrado com $\nu = k-1$ graus de liberdade. Assim, se o valor calculado da estatística superar o quantil superior $100\alpha\%$ ($\chi^2_{\nu;\alpha}$) da distribuição de qui-quadrado com $\nu$ graus de liberdade, a hipótese nula (5.3) deve ser rejeitada.

Os demais testes que veremos na seqüência são os de Levene e Brown e Forsythe (Ferreira (2005)[3]). Estes testes são baseados em uma análise de variância, onde os valores originais da variável resposta são substituídos por outra variável $Z_{ij}$. O teste $F$ é aplicado e a sua estatística é obtida entre a razão da variação entre grupos e dentro de grupos. A diferença básica entre os procedimentos é determinada pela forma como os valores desta nova variável são obtidos. Para o teste de Levene, duas opções existem. A primeira é baseada nos desvios da i-ésima média, tomados em módulo. Assim, os valores para a variável $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$ são obtidos e o teste $F$ é aplicado. Para a segunda opção, devemos obter os valores da variável $Z_{ij} = (Y_{ij} - \bar{Y}_{i.})^2$, a qual refere-se aos desvios da média do i-ésimo nível do fator tomados ao quadrado. Para realizarmos o teste de Brown e Forsythe devemos obter esta variável por: $Z_{ij} = |Y_{ij} - \tilde{Y}_i|$, sendo $\tilde{Y}_i$ a mediana do i-ésimo nível do fator.

Obtidos os valores desta variável para as $n$ observações amostrais, devemos utilizar a estatística do teste:

$$F_c = \frac{(n-k)\sum_{i=1}^{k} n_i \left(\bar{Z}_{i.} - \bar{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_{i.}\right)^2} \tag{5.5}$$

em que:

$$\bar{Z}_{i.} = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i} \qquad \text{e} \qquad \bar{Z}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}}{n}$$

para testarmos a hipótese nula (5.3), utilizando a distribuição $F$ com $\nu_1 = k-1$ e $\nu_2 = n-k$ graus de liberdade. Devemos rejeitar a hipótese nula se $F_c$ de (5.5) for superior ao quantil superior $100\alpha\%$ ($F_{\alpha,\nu_1,\nu_2}$) da distribuição $F$.

Todos estes testes podem ser obtidos com a opção *hovtest=teste* do comando means. Onde no lugar de *teste*, podemos utilizar *levene(type = square)*, *levene(type=abs)*, *BF*, *Bartlett* e o teste não apresentado *OBrien*. O programa SAS na seqüência ilustra a aplicação do teste de Levene com desvios absolutos da média. Obtivemos um valor-p para a estatística $F_c$ de $19,5\%$ e tomamos a decisão de não rejeitar a hipótese de homogeneidade de variâncias.

```
/* Exemplo da utilização do Proc Anova para realizar testes de homogeneidade de variâncias*/
data dic;
input racoes gp @@;
cards;
 1 35   1 19   1 31
 1 15   1 30   2 40
 2 35   2 46   2 41
 2 33   3 39   3 27
 3 20   3 29   3 45
 4 27   4 12   4 13
 4 28   4 30
  ;
proc anova;
    class racoes;
    model gp=racoes;
    means racoes / hovtest=levene(type=abs);
run; quit; /* fim do programa */
```

## 5.3 Estrutura Cruzada de Tratamentos

Em muitas situações experimentais temos delineamentos mais complexos que o inteiramente casualizado, ou mesmo para este delineamento, podemos ter mais de um fator em estruturas mais intrincadas. Entre estes delineamentos mais complexos, encontram-se os blocos casualizados, os quadrados latinos e os látices. Além da estrutura experimental ser mais

complexa, a estrutura de tratamentos também pode não ser a de um simples fator. Uma estrutura muito comum é a cruzada, onde os fatores são combinados fatorialmente. Como a modelagem no SAS é bastante simples, independentemente das estruturas experimental e de tratamentos, vamos ilustrar o seu uso com um caso onde temos um delineamento em blocos casualizados com dois fatores quantitativos (adubo mineral e torta de filtro). Foram utilizados os níveis 0 e 20 kg/ha de adubo mineral e 10% e 20% de torta de filtro. Cada combinação fatorial dos tratamentos foi repetida 4 vezes e a produtividade das plantas foi mensurada. O programa SAS para a análise de variância deste modelo está apresentado na seqüência. O modelo estatístico da análise de variação é dado por:

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \tau_k + \delta_{jk} + \epsilon_{ijk} \tag{5.6}$$

em que $\mu$ é a constante geral do modelo, $\beta_i$ é o efeito do i-ésimo bloco, $\alpha_j$ é o efeito do j-ésimo adubo mineral, $\tau_k$ é o efeito da k-ésima torta de filtro, $\delta_{jk}$ é o efeito da interação entre a j-ésima dose do adubo mineral e a k-ésima dose da torta de filtro e $\epsilon_{ijk}$ é o erro experimental suposto normal e independentemente distribuído com média 0 e variância $\sigma^2$.

---

```
/* Exemplo da utilização do Proc Anova para uma estrutura fatorial em um DBC*/
data Fat;
input A T bloco prod;
cards;
 0 10 1 18.0
20 10 1 20.6
 0 20 1 19.6
20 20 1 19.2
 0 10 2  8.6
20 10 2 21.0
 0 20 2 15.0
20 20 2 19.6
 0 10 3  9.4
20 10 3 18.6
 0 20 3 14.6
20 20 3 18.4
```

```
 0 10 4 11.4
20 10 4 20.6
 0 20 4 15.8
20 20 4 20.2
;
proc anova data=fat;
    class A T bloco;
    model prod = bloco A T A*T;
run; quit;
```

---

O resultado da análise de variação foi reapresentado na Tabela 5.5 em uma forma que encontramos mais comumente nos livros textos.

Tabela 5.5: Análise da variação para o modelo fatorial (2 fatores) em um delineamento de blocos casualizados.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Bloco | 3 | 37,83 | 12,6100 | 3,01 | 0,09 |
| A | 1 | 131,10 | 131,1000 | 31,30 | 0,00 |
| T | 1 | 12,60 | 12,6000 | 3,01 | 0,12 |
| A*T | 1 | 27,55 | 27,5500 | 6,58 | 0,03 |
| Erro | 9 | 37,70 | 4,1889 | | |
| Total | 15 | 246,80 | | | |

Podemos observar efeitos significativos ($P < 0,05$) para adubo mineral e interação. Poderíamos pensar inicialmente em desdobrar a interação adubo mineral e torta de filtro $A * T$, estudando o efeito do adubo mineral em cada nível de torta. Uma abordagem um pouco mais interessante consiste em utilizar um modelo de regressão contendo efeitos de ambos os fatores simultaneamente. Este tipo de modelo é conhecido como superfície de resposta. Vamos utilizar um modelo com três parâmetros, sem considerar o intercepto. O modelo de análise de variância para as fontes de variação adubo mineral, torta de filtro e interação adubo mineral e torta de filtro ($A * T$) possui 3 graus de liberdade associados. O modelo escolhido deveria conter apenas 2 parâmetros, para que o grau de liberdade remanescente

fosse utilizado para testar a falta de ajuste do modelo. Neste exemplo não poderemos aplicar tal teste, por termos esgotados os três graus de liberdade disponíveis. O $R^2$ será igual à unidade, mostrando que podemos obrigar a superfície a passar exatamente sobre os pontos observados. Utilizaremos esta superfície apenas para ilustrar como recalcular determinadas quantidades como $R^2$, erros padrões e testes F e $t$ para as hipóteses de interesse. O modelo que ajustaremos é dado por:

$$\bar{Y}_{.jk} = \beta_0 + \beta_1 A_j + \beta_2 T_k + \beta_3 AT_{jk} + \bar{\epsilon}_{jk} \tag{5.7}$$

em que $\bar{Y}_{.jk}$ é a resposta média para os níveis j e k do adubo mineral e da torta de filtro, $\beta_\ell$ são os parâmetros da regressão, $A_j$ é o nível j do adubo mineral, $T_k$ é o k-ésimo nível da torta de filtro, $AT_{jk}$ é o produto dos níveis j e k do adubo mineral e da torta de filtro e $\bar{\epsilon}_{jk}$ é o erro médio associado com variância $\sigma^2/r$, sendo $r = 4$.

Para ajustar o modelo da equação (5.7) foi utilizado o *proc reg* com todas as observações experimentais. Poderíamos ter utilizado somente as médias da interação para realizarmos este ajuste. Neste caso as somas de quadrados deveriam ser recalculadas para a escala original e optamos por não fazê-lo e utilizarmos todos os dados. Assim, criamos a variável AT dada pelo produto dos níveis de A pelos de T. O programa resultante é dado por:

---

```
/* Exemplo da utilização do Proc Anova para uma estrutura fatorial em um DBC*/
data Fat;
input A T bloco prod;
AT=A*T;
cards;
 0 10 1 18.0
20 10 1 20.6
 0 20 1 19.6
20 20 1 19.2
 0 10 2  8.6
20 10 2 21.0
 0 20 2 15.0
20 20 2 19.6
```

```
 0 10 3  9.4
20 10 3 18.6
 0 20 3 14.6
20 20 3 18.4
 0 10 4 11.4
20 10 4 20.6
 0 20 4 15.8
20 20 4 20.2
;
proc reg data=fat;
    model prod= A T AT/ss1;
Run;Quit;
```

Como fizemos as análises utilizando os dados originais, a soma de quadrados de modelo de regressão (171, 2675), apresentada na Tabela 5.6, representa a soma das somas de quadrados de $A$, $T$ e $A * T$ (131,10, 12,60 e 27,55) obtidas na análise de variância (Tabela 5.5). A soma de quadrados do resíduo (75, 53) desta análise contempla a soma de quadrados do erro puro (37, 70) e a soma de quadrados de blocos (37, 83). Também conteria a soma de quadrados do desvio do modelo ajustado, se não tivéssemos utilizado um modelo completo. Como neste exemplo esgotamos os graus de liberdade do modelo, não houve desvios. Devemos sempre isolar todos estes componentes *manualmente*, pois o SAS não tem uma opção que nos possibilita ajustar o modelo dentro do contexto da análise de variância. Devemos utilizar o *proc reg* e os resultados obtidos devem ser corrigidos posteriormente pelo usuário.

Tabela 5.6: Análise da variação para o modelo de regressão para o exemplo fatorial da adubação com 2 fatores.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Modelo | 3 | 171,27 | 57,0900 | 9,070 | 0,002 |
| Erro | 12 | 75,53 | 6,2942 | | |
| Total | 15 | 246,80 | | | |

Não precisamos ajustar nenhum coeficiente de regressão, mas devemos ajustar os erros padrões e os testes associados, o $R^2$ do modelo e outros

testes e estimativas. O $R^2 = 0,6940$ utilizou a soma de quadrados de totais corrigido como denominador, mas deveria utilizar a soma de quadrados de tratamentos $SQA + SQT + SQAT = 171,27$. Assim, o real valor do coeficiente de determinação é $R^2 = 1$. As estimativas dos parâmetros do modelo e os seus erros padrão estão apresentados na Tabela 5.7. Estes resultados referem-se as estimativas originais do programa SAS, as quais devemos ajustar.

Tabela 5.7: Estimativas dos parâmetros do modelo com seus erros padrões e teste da hipótese para $\beta_i = 0$ fornecidas originalmente pelo SAS.

| Parâmetro | GL | Estimativas | Erro padrão | $t_c$ para $H_0 : \beta_i = 0$ | $Pr > \|t\|$ |
|---|---|---|---|---|---|
| $\beta_0$ | 1 | 7,4500 | 2,8049 | 2,66 | 0,021 |
| $\beta_1$ | 1 | 0,6800 | 0,1983 | 3,43 | 0,005 |
| $\beta_2$ | 1 | 0,4400 | 0,1774 | 2,48 | 0,029 |
| $\beta_3$ | 1 | -0,0263 | 0,0125 | -2,09 | 0,058 |

O erro padrão de uma determinada estimativa é obtido pela expressão (3.15), ou seja, por $\sqrt{x_{ii}S^2}$, em que $S^2$ é o estimador da variância residual e $x_{ii}$ a diagonal de $(X'X)^{-1}$. Como $S^2$ utilizada foi a variância contendo outros efeitos do modelo, como o efeito de blocos, de outros fatores do modelo, do desvio de regressão e do erro puro, então devemos obter o quadrado do erro padrão, multiplicar pela estimativa da variância do erro do modelo de regressão do *proc reg* e assim obter $x_{ii}$. O novo erro padrão é estimado multiplicando $x_{ii}$ pelo $QME$ da análise de variância (Tabela 5.5) e extraindo a raiz quadrada. Para ilustrarmos, vamos considerar o erro padrão da estimativa de $\beta_0$. Este erro padrão foi igual a $2,8049$. Devemos elevá-lo ao quadrado e dividi-lo por $6,2942$, obtendo $2,8049^2/6,2942 = 1,25$. Este valor deve ser multiplicado pelo quadrado médio do erro puro $(4,1889)$ e em seguida extrair sua raiz quadrada. O valor obtido é $\sqrt{1,25 \times 4,1889} = 2,2883$. Repetindo este processo para todos os demais parâmetros, encontramos os resultados apresentados na Tabela 5.8, após recalcular os valores-p da última coluna. Concluímos que todos os efeitos foram significativamente importantes na presença dos demais, o

que não havia acontecido para $A*T$ ou $\beta_3$, quando consideramos a análise original do *proc reg*.

Tabela 5.8: Estimativas dos parâmetros do modelo com seus erros padrões e teste da hipótese para $\beta_i = 0$ devidamente corrigidas.

| Parâmetro | GL | Estimativas | Erro padrão | $t_c$ para $H_0 : \beta_i = 0$ | $Pr > \|t\|$ |
|---|---|---|---|---|---|
| $\beta_0$ | 1 | 7,4500 | 2,2882 | 3,26 | 0,010 |
| $\beta_1$ | 1 | 0,6800 | 0,1618 | 4,20 | 0,002 |
| $\beta_2$ | 1 | 0,4400 | 0,1447 | 3,04 | 0,014 |
| $\beta_3$ | 1 | -0,0263 | 0,0102 | -2,58 | 0,030 |

A análise de variância para o modelo de regressão devidamente corrigida foi apresentada na Tabela 5.9. Não temos neste caso graus de liberdade para o desvio de regressão, que nos possibilitaria aplicar o conhecido teste da falta de ajuste, um dos mais importantes testes na análise de regressão. O ideal é ajustarmos modelos que não esgotem os graus de liberdade de tratamentos, permitindo que haja pelo menos um grau de liberdade para realizarmos o teste da falta de ajuste.

Tabela 5.9: Análise da variação devidamente corrigida para o modelo de regressão do exemplo fatorial da adubação com 2 fatores.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Modelo | 3 | 171,27 | 57,0900 | 13,62 | 0,001 |
| Desvios | 0 | - | - | - | - |
| Erro | 9 | 37,70 | 4,1889 | | |
| Tratamento | 3 | 171,27 | | | |

Muitos pesquisadores não se atentam para estas correções da análise de regressão quando submetida ao *proc reg*, sendo os dados oriundos de uma análise de variância. Assim, muitas inferências podem estar comprometidas e até mesmo incorretas.

O modelo ajustado é dado por:

$$\hat{\hat{Y}}_{.jk} = 7,45 + 0,68A_j + 0,44T_k - 0,0263AT_{jk}$$

Na Figura 5.1 apresentamos a superfície de resposta ajustada para os valores médios dos níveis dos fatores $A$ e $T$ em relação a produção. Observamos que as respostas máximas foram obtidas quando se utilizou a dose 20 kg/ha de adubo mineral com a dose mínima de torta de filtro (10%).

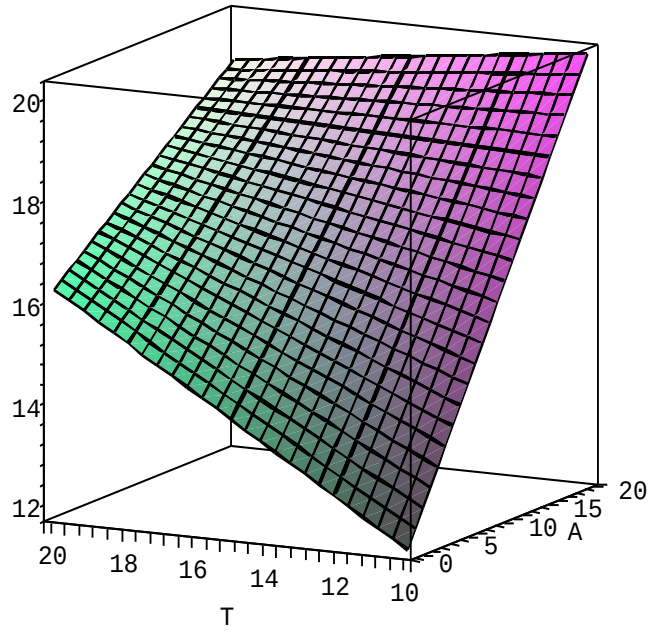


Figura 5.1: Modelo ajustado de superfície de resposta para os dados de produção em função da adubação mineral ($A$) e da adubação orgânica com torta de filtro ($T$).

Podemos observar que haverá uma queda acentuada da produtividade se não for utilizado adubo químico. Nesta mesma condição se passarmos do nível de 10% de torta para 20%, observamos um incremento na produtividade. No entanto, se estamos utilizando a dose de 20 kg/ha de adubo químico, este aumento de 10% para 20% na torta de filtro provoca uma redução da produtividade média. Assim, devemos recomendar as doses de 20

kg/ha de adubo mineral e 10% de torta de filtro para obtermos a máxima resposta.

## 5.4 Modelos Lineares Com Mais de Um Erro

Em algumas situações reais nos deparamos com modelos que contém mais de um erro experimental. Isso acontece em delineamentos experimentais como o de parcelas subdivididas, sub-subdivididas ou em faixas. Um outro caso que ocorre normalmente é o de parcela subdividida no tempo. Neste caso o delineamento em geral é simples, como o inteiramente casualizado ou o de blocos casualizados e cada parcela ou unidade experimental é avaliada ao longo do tempo. Se pudermos supor que existe uma variância constante entre as observações ao longo do tempo e que a estrutura de correlação entre diferentes tempos é a mesma, então podemos fazer uma abordagem biométrica bastante simples, tratando este modelo com um modelo de parcelas subdividas no tempo. Assim, mais de um erro irá aparecer no modelo e este caso pode ser encaixado dentro desta seção. Esta estrutura de correlação é denominada de simetria composta.

Vamos ilustrar este tipo de modelo, contendo mais de um erro, com um exemplo de parcela subdividida no tempo. Um adubo mineral foi utilizado como fator principal, onde desejávamos comparar seus três níveis 0, 10 e 20 kg/ha. Estas três dosagens foram submetidas a um delineamento em blocos completos casualizados com 2 repetições. O interesse era o crescimento das plantas ao longo do tempo. Assim, foram avaliadas as alturas das plantas durante 3 meses consecutivos. O modelo estatístico para este experimento é dado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ij} + \gamma_k + \epsilon_{jk} + \delta_{ik} + \epsilon_{ijk} \tag{5.8}$$

em que $Y_{ijk}$ é a observação da altura das plantas em metros, $\mu$ é a constante geral do modelo, $\alpha_i$ é o efeito do i-ésimo nível da adubação química, $\beta_j$ é o efeito do j-ésimo bloco, $\epsilon_{ij}$ é o efeito do erro experimental entre a i-ésima dose e o j-ésimo bloco, $\gamma_k$ é o efeito do k-ésimo mês, $\epsilon_{jk}$ é efeito do erro experimental do j-ésimo bloco com o k-ésimo mês, $\delta_{ik}$ é o efeito da interação

entre a i-ésima dose de adubo químico com o k-ésimo mês e $\epsilon_{ijk}$ é o erro experimental entre a i-ésima dose, j-ésimo bloco e k-ésimo mês.

O programa SAS contendo os dados experimentais e a sintaxe para especificar os erros do modelo e determinar os testes corretos é apresentado na seqüência. Como os erros intermediários do modelo não são prontamente reconhecidos pelo SAS, estes devem ser indicados para que possamos realizar os testes de hipóteses corretamente. Se esta indicação dos erros intermediários não for feita, os resultados dos testes de hipóteses serão incorretos.

---

```
/* Programa para realizar análise de variância de um modelo contendo múltiplos erros.
O modelo escolhido foi o de parcela subdividida no tempo.*/
data sub;
input bloco trat mes alt;
cards;
1   0 1 1.00
1 10 1 1.05
1 20 1 1.08
2   0 1 1.02
2 10 1 1.06
2 20 1 1.09
1   0 2 1.10
1 10 2 1.12
1 20 2 1.14
2   0 2 1.08
2 10 2 1.15
2 20 2 1.18
1   0 3 1.14
1 10 3 1.20
1 20 3 1.22
2   0 3 1.15
2 10 3 1.21
2 20 3 1.23
;
proc anova data=sub;
    class bloco trat mes;
    model alt = bloco trat bloco*trat mes bloco*mes mes*trat;
    test h=bloco trat e=bloco*trat;
    test h=mes e=bloco*mes;
```

```
    means mes/ Tukey e=bloco*mes;
run; quit;
```

---

Se os níveis dos tratamentos fossem qualitativos, o que não é o caso deste exemplo, o comando "*<means trat / tukey e=bloco*trat;>*", poderia ser utilizado. Com este comando, são requisitados o cálculo das médias de tratamento e a aplicação do teste de Tukey usando como erro o efeito de *bloco*trat*. Se for utilizado apenas o comando "*<means trat / tukey;>*", o *proc anova* irá aplicar o teste de Tukey com o erro inadequado, ou seja, com o erro geral do modelo. Os testes de hipóteses sobre os efeitos dos fatores são aplicados corretamente se for especificado o comando *test*, indicando ao SAS qual deve ser o procedimento adequado. Neste comando as hipóteses a serem testadas são determinadas no comando *h=efeito* e o erro apropriado para testá-las, no comando *e=efeito*. Os resultados incorretos do SAS, que utiliza o erro do modelo para testar estas hipóteses, devem ser ignorados. A opção *test* não é checada pelo *proc anova* e é de inteira responsabilidade do usuário a correta aplicação do teste $F$. Os resultados da análise de variância devidamente reorganizada está apresentada na Tabela 5.10.

Tabela 5.10: Análise da variação devidamente apresentada para o modelo de parcela subdividida no tempo.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Bloco | 1 | 0,00080000 | 0,00080000 | 6,86 | 0,1201 |
| Trat | (2) | (0,01750000) | 0,00875000 | 75,00 | 0,0132 |
| RL | 1 | 0,01687000 | 0,01687000 | 144,60 | 0,0068 |
| Desvio | 1 | 0,00062500 | 0,00062500 | 5,35 | 0,1468 |
| Erro a | 2 | 0,00023333 | 0,00011667 | | |
| Mês | 2 | 0,06043333 | 0,03021667 | 1.813,00 | 0,0006 |
| Erro b | 2 | 0,00003333 | 0,00001667 | | |
| Trat*Mês | 4 | 0,00016667 | 0,00004167 | 0,20 | 0,9259 |
| Erro | 4 | 0,00083333 | 0,00020833 | | |
| Total | 17 | 0,08000000 | | | |

Ajustamos um modelo linear simples da variável resposta altura em função da adubação química utilizando o *proc reg* e obtivemos o seguinte modelo: $\hat{\hat{Y}}_{i..} = 1,08583 + 0,00375A_i$, em que $A_i$ é o i-ésimo nível do adubo químico. O coeficiente de determinação deve ser reestimado por $R^2 = 0,01687/0,0175 = 0,964$. A análise de variância do modelo de regressão, apresentando o teste de falta de ajuste foi incorporado na Tabela 5.10. Neste caso, obtivemos um teste de falta de ajuste não significativo, um $R^2$ alto e o modelo de regressão com teste $F$ significativo, ou seja, obtivemos resultados considerados ideais.

Consideramos ainda que os níveis de mês sejam qualitativos e não quantitativos e aplicamos o teste Tukey. Todas as médias diferiram entre si pelo teste de Tukey. Deve-se observar que foi utilizado o erro apropriado para realizarmos o teste de comparações múltiplas de Tukey. As maiores médias para a altura em relação ao mês, como era esperado, estavam associadas ao 3, seguidas pelo 2 e finalmente pelo 1.

## 5.5 Modelos lineares multivariados

Na pesquisa agropecuária e de outras áreas é comum as situações em que várias variáveis são mensuradas simultaneamente. Os fenômenos estudados respondem aos tratamentos não apenas com relação a uma variável, mas sim em relação ao conjunto total de variáveis associadas aquele fenômeno. Nestes casos, duas aproximações podem ser feitas: a primeira utilizando uma análise para cada variável separadamente, produzindo uma grande quantidade de informações, além de não levar em consideração a estrutura de covariância entre as variáveis; a segunda utilizando a análise multivariada, que considera esta estrutura de covariância entre as variáveis sob estudo.

Para ilustrar como são realizados os ajustes dos modelos e obtidas as somas de quadrados e de produtos, vamos utilizar um modelo linear multivariado com $m$ parâmetros associados a cada uma das $p$ variáveis respostas. Diferentemente dos casos univariados, onde são calculadas apenas somas de quadrados, nos modelos lineares multivariados são obtidas somas de produtos entre as variáveis. Isto deve ser feito para cada fonte de variação (ou efeito) do modelo. As somas de quadrados e produtos são apresentadas em

uma matriz $p \times p$ e os testes de hipóteses envolvem estatísticas que são relacionadas com razões de determinantes ou de funções dos autovalores das matrizes de somas de quadrados e produtos associadas à hipótese e ao erro.

Os modelos lineares multivariados podem ser escritos matricialmente por:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.9}$$

em que $\boldsymbol{Y}$ é matriz das variáveis respostas com $n$ linhas (observações) e $p$ colunas (variáveis), $\boldsymbol{X}$ é a matriz de modelo com $n$ linhas e $m$ colunas (parâmetros do modelo), $\boldsymbol{\beta}$ é a matriz de parâmetros com $m$ linhas e $p$ colunas e $\boldsymbol{\epsilon}$ é a matriz de erros $n \times p$ supostos normal multivariados e independentemente distribuídos com média $\underset{\sim}{0}$ e covariância comum $\boldsymbol{\Sigma}$.

A solução de mínimos quadrados é obtida por:

$$\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^g\boldsymbol{X}'\boldsymbol{Y} \tag{5.10}$$

A matriz de somas de quadrados e produtos do modelo determinado por 5.9 é dada por:

$$H = R(\boldsymbol{\beta}) = \boldsymbol{\beta}^{*\prime}\boldsymbol{X}'\boldsymbol{Y} \tag{5.11}$$

A matriz de soma de quadrados e produtos do resíduo $\boldsymbol{E}$ é obtida por $\boldsymbol{E} = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{\beta}^{*\prime}\boldsymbol{X}'\boldsymbol{Y}$. Mediante reduções de modelos hierárquicos, aplicamos as expressões 5.10 e 5.11 para estimarmos as matrizes de somas de quadrados e produtos dos efeitos de um modelo ajustados para os efeitos de outros, da mesma forma como é feito para regressão e para modelos univariados. A diferença neste caso é o resultado matricial obtido. Não daremos nenhum outro resultado adicional neste material, devido às dificuldades teóricas deste assunto.

Vamos ilustrar a utilização do *proc anova* para realizarmos uma análise de variância multivariada, com os respectivos testes de hipóteses. O exemplo que vamos utilizar refere-se a três métodos de ensino diferentes aplicados a uma determinada série do ensino básico. As notas de duas disciplinas em

cada método de ensino foram anotadas em amostras de diferentes tamanhos. O programa SAS com os três métodos de ensino ($A$, $B$ e $C$) juntamente com os comandos da opção Manova são apresentados na seqüência.

---

```
/* Programa ilustrativo da Manova */
data multi;
input met $ n1 n2;
cards;
A 69 75
A 69 70
A 71 73
A 78 82
A 79 81
A 73 75
B 69 70
B 68 74
B 75 80
B 78 85
B 68 68
B 63 68
B 72 74
B 63 66
B 71 76
B 72 78
B 71 73
B 70 73
B 56 59
B 77 83
C 72 79
C 64 65
C 74 74
C 72 75
C 82 84
C 69 68
C 76 76
C 68 65
C 78 79
C 70 71
C 60 61
;
```

```
proc anova;
    class met;
    model n1 n2 = met;
    manova h = met / printe printh;
run;quit;
```

---

Os principais resultados desta análise foram sumariados na seqüência. Inicialmente foram obtidas as análises de variâncias para cada uma das notas das matérias. Os resultados para a variável 1 estão apresentados na Tabela 5.11. Observamos que não foram detectadas diferenças significativas entre os métodos.

Tabela 5.11: Análise da variação para nota da disciplina 1 para testar a hipótese de igualdade dos efeitos dos métodos de ensino.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Métodos | 2 | 60,6051 | 30,3025 | 0,91 | 0,4143 |
| Erro | 28 | 932,8788 | 33,3171 | | |
| Tratamento | 30 | 993,4839 | | | |

Os resultados para a variável 2 estão apresentados na Tabela 5.12. Da mesma forma que ocorreu para a variável 1, observamos que não foram detectadas diferenças significativas entre os métodos.

Tabela 5.12: Análise da variação para nota da disciplina 2 para testar a hipótese de igualdade dos efeitos dos métodos de ensino.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Métodos | 2 | 49,7359 | 24,8679 | 0,56 | 0,5776 |
| Erro | 28 | 1243,9416 | 44,4265 | | |
| Tratamento | 30 | 1293,6774 | | | |

Os comandos *printe* e *printh* geram saídas com as matrizes de somas de quadrados e produtos do resíduo e de métodos. Além disso, o primeiro comando permite que se obtenha as estimativas das correlações parciais

entre as variáveis ajustadas paras as fontes de variação do modelo. As matrizes de soma de quadrados e produtos são:

$$E = \begin{bmatrix} 932,8788 & 1018,6818 \\ 1018,6818 & 1243,9416 \end{bmatrix} \quad \text{e} \quad H = \begin{bmatrix} 60,6051 & 31,5117 \\ 31,5117 & 49,7359 \end{bmatrix}$$

A matriz de correlações parciais acompanhada das probabilidade para os testes de hipóteses $H_0 : \rho = 0$ é dada por:

$$R = \begin{bmatrix} 1,0000 & 0,94564 \\ & < 0,0001 \\ 0,945640 & 1,0000 \\ < 0,0001 & \end{bmatrix}$$

Concluímos que as duas variáveis são altamente correlacionadas, eliminando-se o efeito dos métodos. Os testes de hipóteses multivariados sobre a igualdade do vetor de médias são feitos basicamente por 4 critérios distintos. O critério de Wilks é um deles e é um teste via razão de verossimilhanças. Muitos pesquisadores preferem tomar a decisão de rejeitar a hipótese nula quando pelo menos 3 dos 4 critérios apresentarem estimativas significativas das estatísticas dos testes. Outros preferem utilizar o critério de Wilks para tomar esta decisão. Para testarmos a hipótese nula, qualquer que seja a opção escolhida, os valores destas estatísticas são convertidos para F, que é a distribuição utilizada para aproximar as exatas. Em alguns casos dependendo do número de tratamentos e de variáveis a estatística F resultante possui distribuição F exata. Na versão 9, o SAS já apresenta uma opção para solicitar que os testes exatos sejam computados. Os resultados do teste de hipótese de igualdade dos vetores de médias dos três métodos foram apresentados na Tabela 5.13. Todos os critérios apresentaram valores correspondentes de $F$ significativos.

Uma outra observação que pode ser feita neste exemplo, refere-se ao fato de os níveis de significância multivariados terem sido muito menores que os univariados, indicando os casos clássicos em que os testes univariados

Tabela 5.13: Testes de hipóteses multivariados para a igualdade dos efeitos dos métodos de ensino.

| Estatística | Estimativa | F | GL num. | GL den. | $Pr > F$ |
|---|---|---|---|---|---|
| Wilks' Lambda | 0,67310116 | 2,95 | 4 | 54 | 0,0279 |
| Pillai's Trace | 0,33798387 | 2,85 | 4 | 56 | 0,0322 |
| Hotelling-Lawley Trace | 0,46919220 | 3,13 | 4 | 31,389 | 0,0281 |
| Roy's Greatest Root | 0,43098027 | 6,03 | 2 | 28 | 0,0066 |

falham em detectar alguma diferença entre os tratamentos, mas os multivariados não. Este fato provavelmente pode ser em parte explicado pela alta correlação parcial entre as variáveis respostas.

## 5.6 Exercícios

1. Utilizar dados balanceados resultantes de pesquisas desenvolvidas em sua área e realizar análises de variâncias utilizando o *proc anova*. Aplicar os testes de médias, se os níveis forem qualitativos, ou ajustar modelos de superfície de resposta ou de regressão, se os níveis dos fatores forem quantitativos.

2. Em sua opinião, qual foi a vantagem de se utilizar uma modelagem multivariada para o exemplo deste capítulo que comparava três métodos de ensino em relação a análise de variância univariada. Você utilizaria análises multivariadas de variância em sua área profissional?

# Capítulo 6

# Análise de Variância para Dados Não-Balanceados

Muitas vezes precisamos realizar inferência sobre a igualdade de médias de um determinado fator. Se o conjunto de dados for não-balanceado, apresentando perdas de parcelas ou até mesmo de caselas devemos utilizar a análise de variância para isso. A análise de variância neste caso deve ser realizada por meio de métodos matriciais para lidarmos com o não-balanceamento dos dados. A partição da variação entre as observações em partes associadas a certos fatores, que são definidos pelo esquema de classificação dos dados experimentais, pode ser realizada de diferentes formas. Assim, diferentes hipóteses podem ser testadas a partir de um mesmo conjunto de dados.

O *proc anova* é apropriado para conjuntos de dados que sejam balanceados. O *proc glm* nos permite analisar conjuntos de dados não-balanceados, incluindo casos extremos de desconexão. Neste capítulo aplicaremos o *proc glm* a conjuntos de dados não-balanceados. Estudaremos três dos quatro tipos de somas de quadrados que podem ser estimados por este procedimento. No caso de delineamentos balanceados, estas somas de quadrados, são todas iguais, não havendo diferenças nas hipóteses que são testadas, exceto se para a soma de quadrados tipo I for utilizada uma ordem em que um efeito de interação aparece antes dos efeitos principais ou de interações de menor ordem destes efeitos principais que compõem esta interação.

A soma de quadrados tipo I refere-se à soma de quadrados seqüencial. Esta soma de quadrado é obtida com a redução no modelo de um fator por vez, na ordem inversa à de entrada dos fatores no modelo. Para ilustrarmos, vamos considerar um modelo com dois fatores $(\alpha, \beta)$ e interação $(\delta)$ dado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk} \tag{6.1}$$

em que $Y_{ijk}$ é o valor observado da variável resposta, $\mu$ é a constante geral, $\alpha_i$ é o efeito do i-ésimo nível do fator $\alpha$, $\beta_j$ é o efeito do j-ésimo nível do fator $\beta$, $\delta_{ij}$ é o efeito da interação entre o i-ésimo nível do fator $\alpha$ com o j-ésimo nível do fator $\beta$ e $\epsilon_{ijk}$ é o efeito do erro experimental suposto normal e independentemente distribuído com média 0 e variância comum $\sigma^2$.

A soma de quadrados tipo I, II e III para os efeitos do modelo da equação (6.1) está apresentada na Tabela 6.1.

Tabela 6.1: Tipos de somas de quadrados de um modelo de análise de variância contendo dois fatores $\alpha$ e $\beta$ e interação $\delta$.

| FV | SQ Tipo I | SQ Tipo II | SQ Tipo III |
|---|---|---|---|
| $\alpha$ | $R(\alpha/\mu)$ | $R(\alpha/\mu, \beta)$ | $R(\alpha^*/\mu^*, \beta^*, \delta^*)$ |
| $\beta$ | $R(\beta/\mu, \alpha)$ | $R(\beta/\mu, \alpha)$ | $R(\beta^*/\mu^*, \alpha^*, \delta^*)$ |
| $\delta$ | $R(\delta/\mu, \alpha, \beta)$ | $R(\delta/\mu, \alpha, \beta)$ | $R(\delta^*/\mu^*, \alpha^*, \beta^*)$ |

$^*$ indica parâmetros obtidos sob o uso de restrição paramétrica.

A soma de quadrado tipo II para um dado fator é obtida ajustando esta fonte de variação para todas as outras que não contenha o efeito em questão. Assim, a soma de quadrados para $\alpha$, não pode ser ajustada para a fonte de variação $\delta$, uma vez que esta última contém o efeito de $\alpha$, por ser a interação deste fator com $\beta$. A soma de quadrados tipo III, ou parcial, refere-se ao ajuste de cada fator para todos os demais efeitos do modelo sob restrição paramétrica do tipo soma de efeitos igual a zero.

As somas de quadrados do tipo I são dependentes da ordem de entrada dos fatores no modelo. As somas de quadrados do tipo II e III não dependem desta ordem de entrada. Como dissemos, elas são iguais quando os dados

são balanceados, tomando-se o cuidado de entrar com uma ordem dos efeitos no modelo, em que os fatores principais vêm antes das interações de que participam.

O *proc glm* é um dos procedimentos do SAS utilizados para lidar com estes casos não-balanceados. As sintaxes deste procedimento e do *proc anova* são praticamente idênticas. As principais diferenças são, entre outras, a possibilidade de estimar efeitos e testar contrastes, de realizar análise de covariância e de estimar componentes de variância.

Vamos utilizar alguns dos conjuntos de dados anteriores, provocando artificialmente algum tipo de não balanceamento em algumas ocasiões e em outras utilizando os dados balanceados, para ilustrarmos as principais peculiaridades do *proc glm*.

## 6.1 Delineamento Inteiramente Casualizado

No modelo inteiramente casualizado com um fator (equação 5.2), vamos considerar o mesmo conjunto de dados apresentados na Tabela 5.1, para ilustrarmos o uso de contrastes no *proc glm*. A variável resposta é o ganho de peso dos animais submetidos a quatro rações diferentes. Um delineamento inteiramente casualizado com 5 repetições foi utilizado. Vamos imaginar que houvesse uma estrutura dos níveis dos tratamentos, estabelecida por diferentes firmas produtoras das rações e diferentes fontes de proteínas. Assim, a ração 1 é proveniente da firma A e as rações 2, 3 e 4 da firma B. A ração 2 possui fonte de proteína animal e as rações 3 e 4 têm proteína de origem vegetal. As rações 3 e 4 diferem quanto ao nível de energia que possuem.

Devido aos tratamentos serem estruturados é natural que façamos contrastes sugeridos por esta estrutura. Um conjunto de contrastes ortogonais que poderíamos desejar testar seria: 1 vs 2, 3, e 4, contrastando firma A contra firma B, 2 vs 3 e 4, contrastando proteína animal contra proteína vegetal e finalmente 3 vs 4, contrastando os níveis de energia. Como temos 3 graus de liberdade e 3 contrastes ortogonais, então, teríamos feito uma decomposição ortogonal das somas de quadrados de tratamento. Para estimarmos os efeitos dos contrastes, aplicamos o comando *estimate* e para testarmos o

contraste, o comando *contrast*. O programa resultante, para estimarmos e testarmos os efeitos dos contrastes, é apresentado na seqüência.

---

```
/* Exemplo da utilização do Proc GLM para testarmos contrastes em um DIC balance-
ado*/
data dic;
input racoes gp @@;
cards;
 1 35   1 19   1 31   1 15
 1 30   2 40   2 35   2 46
 2 41   2 33   3 39   3 27
 3 20   3 29   3 45   4 27
 4 12   4 13   4 28   4 30
  ;
proc glm;
    class racoes;
    model gp=racoes;
    means racoes / tukey alpha = 0.05 lines;
    lsmeans racoes / pdiff adjust = tukey;
    lsmeans racoes / pdiff = control("1") adjust = dunnett;
    contrast "1 vs 2, 3 e 4" racoes 3 -1 -1 -1;
    contrast "2 vs 3 e 4" racoes 0 2 -1 -1;
    contrast "3 vs 4" racoes 0 0 1 -1;
    estimate "1 vs 2, 3 e 4" racoes 3 -1 -1 -1/divisor=3;
    estimate "2 vs 3 e 4" racoes 0 2 -1 -1/divisor=2;
    estimate "3 vs 4" racoes 0 0 1 -1;
run; quit; /* fim do programa */
```

---

Utilizamos os comandos *means* e *lsmeans*, neste exemplo, simplesmente para ilustrarmos as sintaxes, pois como os tratamentos são qualitativos estruturados, devemos utilizar contrastes para otimizarmos as comparações realizadas. Ilustramos o uso de um teste de comparações múltiplas sobre médias não ajustadas e ajustadas e o teste de Dunnett bilateral, utilizando a ração 1 como controle. O objetivo foi de apresentar a sintaxe dos comandos para podermos obter médias ajustadas e para aplicarmos os testes de comparações múltiplas e de Dunnett. Todos estes resultados devem ser

ignorados neste exemplo e somente os resultados dos contrastes e das estimativas devem ser considerados. Somente o contraste entre os tipos de origem das proteínas na formulação das rações da firma B foi significativo ($P < 0,0177$). Como a estimativa é positiva, podemos afirmar que em média teremos um ganho superior em 12 kg/animal/período, se utilizarmos ração com proteína animal em vez de proteína de origem vegetal. Não solicitamos somas de quadrados de nenhum tipo, mas o padrão do *glm* é apresentar tanto a soma de quadrados do tipo I, quanto à do tipo III. Nos modelos lineares para os quais temos apenas um efeito, além do intercepto e do erro, não faz sentido diferenciar as somas de quadrados, pois todas elas são idênticas. Neste caso, a soma de quadrados do tipo I para rações foi de $823,75$, sendo o mesmo resultado obtido para as somas de quadrados dos tipos II e III.

Uma outra vantagem do *proc glm* é obter predições para os valores da variável resposta, que neste caso, são as médias de caselas. Adicionalmente os valores residuais são preditos. Para isso basta substituir o comando *<model gp=racoes;>* por *<model gp=racoes/p;>*. Este comando, além destas estimativas e predições, fornece a estatística de Durbin-Watson, para realizarmos testes de autocorrelação. Outra estimativa, que utilizamos com freqüência na análise de dados não-balanceados, é a da média ajustada. Em vez de utilizarmos o comando *<means racoes / tukey alpha=0.05 lines;>* podemos utilizar o comando *<lsmeans racoes / pdiff adjust=tukey;>*. Neste caso, o SAS calculará os valores-p das comparações entre as lsmeans utilizando o procedimento ajustado de Tukey. Para comparação com o controle fazemos *pdiff = control('trat')* com o comando *adjust = opção*. A opção que devemos utilizar é a do teste de Dunnett, determinada por *dunnett*. Apesar de o natural ser a escolha do comando *adjust=dunnett*, podemos escolher outras formas de ajustes como Bon, Sidak, Scheffe, entre outras. É claro que para um delineamento inteiramente casualizado com um fator balanceado ou não-balanceado não existem diferenças entre as médias ajustadas e não-ajustadas. Mas, entre os testes utilizando as médias ajustadas e as médias não ajustadas existem diferenças nos casos não balanceados. Devemos optar por utilizar as médias ajustadas solicitando o teste apropriado.

## 6.2 Estrutura Cruzada de Tratamentos

Para ilustramos a análise de modelos mais complexos, onde temos conjuntos de dados não-balanceados, vamos retornar ao exemplo apresentado na seção 5.3, simulando algumas perdas de parcelas. Com este exemplo, vamos mostrar as dificuldades existentes para realizar uma análise de dados não-balanceados e as diferenças entre os três tipos de somas de quadrados que estamos considerando. Posteriormente consideraremos, ainda, uma análise de covariância. Os dados apresentados na seção 5.3 com algumas perdas de unidades experimentais simuladas e o modelo da equação (5.6) foram utilizados. Temos um delineamento em blocos casualizados com 4 repetições e 2 fatores (adubo mineral e torta de filtro) com 2 níveis cada.

O programa ilustrando a análise de variância e os principais resultados alcançados estão apresentados na seqüência. Vamos destacar o uso da opção *slice* do comando *lsmeans* neste programa, a qual possibilita que seja realizado o desdobramento de interações entre efeitos do modelo.

---

```
/* Exemplo da utilização do proc GLM para uma estrutura fatorial de tratamentos em
um DBC e não-balanceada*/
data Fat;
input A T bloco prod;
cards;
 0 10 1 18.0
20 10 1 20.6
 0 20 1 19.6
20 20 1 19.2
 0 10 2  8.6
 0 20 2 15.0
20 20 2 19.6
 0 10 3  9.4
20 10 3 18.6
 0 20 3 14.6
20 20 3 18.4
 0 10 4 11.4
 0 20 4 15.8
20 20 4 20.2
;
```

```
proc glm data=fat;
    class A T bloco;
    model prod = bloco A T A*T/ss1 ss2 ss3;
    means A T/Tukey;
    lsmeans A T/pdiff adjust=Tukey;
    lsmeans A*T/slice=A slice=T;
run; quit;
```

---

Inicialmente, observamos que uma análise de variação contendo as fontes de variação de modelo e de resíduos foi obtida. Estes resultados estão apresentados na Tabela 6.2. Na Tabela 6.3 apresentamos os três tipos de somas de quadrados solicitadas (I, II e III). Podemos observar um efeito significativo de $A$ e de $T$ para os três tipos de somas de quadrados, exceto para o efeito da torta de filtro com a soma de quadrado do tipo III. Em todos os casos (I, II e III) tivemos um efeito não significativo da interação, sendo as somas de quadrados tipo I, II e III para este efeito iguais.

Tabela 6.2: Análise da variação para o modelo fatorial (2 fatores) em um delineamento de blocos casualizados, destacando-se as fontes de variação de modelo e erro.

| FV | G.L. | SQ | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Modelo | 6 | 180,89 | 30,15 | 6,75 | 0,0120 |
| Erro | 7 | 31,29 | 4,47 | | |
| Total | 13 | 212,17 | | | |

CV = 12,92% $\bar{Y}_{...} = 16,36$

Houve uma diferença muito grande entre algumas das somas de quadrados, sendo que no efeito da adubação mineral, isto foi mais pronunciado. Era esperado, por exemplo, que as somas de quadrados do tipo I e do tipo II para efeito da torta de filtro fossem iguais, considerando a ordem que os fatores entraram no modelo. Dessa forma, podemos observar a importância de saber exatamente o que testamos, para interpretar adequadamente as saídas do *proc glm*. Detalhes técnicos a respeito das hipóteses associadas a estas somas de quadrados podem ser obtidos em publicações especializadas.

Tabela 6.3: Resumo da análise da variação para o modelo fatorial (2 fatores) em um delineamento de blocos casualizados, destacando as somas de quadrados tipo I, II e III e as significâncias correspondentes.

| FV | G.L. | SQ I | SQ II | SQ III |
|---|---|---|---|---|
| Bloco | 3 | 53,1543$ns$ | 42,7233$ns$ | 42,7233$ns$ |
| A | 1 | 88,7520** | 66,9780** | 77,0133** |
| T | 1 | 27,3780* | 27,3780* | 17,7633$ns$ |
| A*T | 1 | 11,6033$ns$ | 11,6033$ns$ | 11,6033$ns$ |

*, ** e $ns$: significativo a 5, 1% e não significativo, respectivamente.

Se observarmos as saídas do SAS, podemos verificar que existem diferenças entre as médias ajustadas e não-ajustadas, destacando-se a importância de utilizar o comando adequado para o caso balanceado. Neste exemplo observamos que tanto para torta de filtro, como para a adubação mineral, obtivemos diferenças significativas para as médias. No entanto, quando utilizamos o teste com correção de Tukey sobre as médias ajustadas, somente detectamos diferenças significativas para adubo mineral, mas não para torta de filtro.

Finalmente o comando *slice* nos possibilita obter a análise do desdobramento da interação $A * T$. Solicitamos os dois tipos de desdobramento: o de $A$ dentro dos níveis de $T$ e o de $T$ fixados os níveis de $A$. Nenhum destes dois casos serão apresentados, pois a interação foi não significativa. Assim, recomendamos utilizar a maior dose de adubo mineral (teste marginal significativo) e a menor porcentagem de torta de filtro (teste marginal não significativo).

Reiteramos que as somas de quadrados do tipo I são afetadas pela ordem dos efeitos na especificação do modelo. Podemos ver claramente que se alterarmos esta ordem, teremos diferentes somas de quadrados do tipo I, mas as mesmas somas de quadrados dos tipos II e III obtidas anteriormente. O caso mais crítico desta alteração ocorre quando colocamos o efeito da interação dos fatores antes dos efeitos principais. Como o espaço paramétrico da interação contém os espaços paramétricos dos efeitos principais, teremos resultados nulos para os graus de liberdade e somas de quadrados

associados. O leitor é conclamado a verificar este resultado para o modelo em questão.

Alguns outros aspectos interessantes da análise merecem destaques. Como todos os procedimentos são realizados por meio de álgebra matricial e vetorial, podemos solicitar a matriz inversa, a matriz $X'X$, valores preditos, solução mínimos de quadrados, entre outras opções. Para isso bastaria substituir o comando <model prod = bloco A T A*T/ss1 ss2 ss3;> por <model prod = bloco A T A*T/ss1 ss2 ss3 p solution XPX I;>.

Outra grande vantagem do *proc glm* é a possibilidade de realizarmos análises de regressão. Um fator omitido do comando *class* será considerado variável regressora e não variável classificatória. Assim, temos a possibilidade de realizar análises de covariância. A análise de covariância ocorre quando temos variáveis classificatórias (fatores qualitativos) e variáveis regressoras (fatores quantitativos) no mesmo modelo. Em geral estas covariáveis devem ser mensuradas em todas as unidades experimentais e não devem ser influenciadas pelo tratamento. Por exemplo, se estamos testando diferentes cultivares, utilizar o estande final como covariável, pode não ser uma boa estratégia. Isso porque pode existir um efeito de cultivares no estande final, ou seja, o efeito de estande é influenciado pelo efeito de cultivares. Assim, uma análise como essa vai produzir um ajuste do efeito de cultivar pelo efeito de estande. Como os dois efeitos podem estar relacionados, como acabamos de discutir, teremos o efeito de cultivar ajustado, de forma indireta, para o próprio efeito de cultivar. Assim, devemos utilizar covariáveis que não sejam influenciadas pelos tratamentos. Neste caso, poderíamos, por exemplo, ter tomado medidas da fertilidade do solo em cada parcela experimental, antes de as cultivares terem sido semeadas. Estas variáveis de fertilidade poderiam ser utilizadas como covariáveis.

Neste exemplo fatorial foi simulada a avaliação de uma covariável em cada parcela, para podermos ilustrar uma análise de covariância. Assim, em cada parcela experimental foi avaliado o teor de nitrogênio. Uma amostra de cada unidade foi coletada e os níveis de nitrogênio do solo foram mensurados, antes da implantação dos tratamentos, correspondentes ao adubo mineral e a torta de filtro. Um aspecto da análise de covariância que empiricamente podemos mencionar, refere-se ao fato de que ao utilizarmos

uma covariável e ajustarmos o efeito de tratamentos para essa covariável, estaríamos fazendo algo semelhante a ter um experimento cujas condições iniciais seriam homogêneas para os níveis desta covariável. Assim, é como se indiretamente estivéssemos utilizando um controle local.

No exemplo que se segue apresentamos a análise de covariância utilizando como covariável os níveis de nitrogênio nas unidades experimentais mensurados anteriormente a implantação do experimento. A especificação de uma covariável no modelo é feita de maneira bastante simples. Para isso omitimos no comando *class* a covariável, mas a introduzimos no comando *model*. O *proc glm* irá reconhecer a variável omitida como uma variável regressora e o comando *lsmeans* irá ajustar as médias dos fatores para a covariável ou covariáveis presentes no modelo. O programa SAS, ilustrativo deste caso, é dado por:

```
/* Exemplo da utilização do proc GLM para uma estrutura fatorial dos tratamentos com
covariável em um DBC não-balanceado*/
data Fat;
input A T bloco prod N;
cards;
 0 10 1 18.0 3
20 10 1 20.6 4
 0 20 1 19.6 5
 0 10 2  8.6 3
 0 20 2 15.0 4
20 20 2 19.6 4
 0 10 3  9.4 6
20 10 3 18.6 5
 0 20 3 14.6 2
20 20 3 18.4 7
 0 10 4 11.4 4
 0 20 4 15.8 3
20 20 4 20.2 3
;
proc glm data=fat;
    class A T bloco;
    model prod = bloco A T A*T N/solution ss1 ss2 ss3;
    means A T/Tukey;
    lsmeans A T/pdiff adjust=Tukey;
```

```
    lsmeans A*T/slice=A slice=T;
run; quit;
```

---

Se realizarmos uma análise de variância com e sem a covariável podemos observar que os resultados para este exemplo apresentam uma ligeira diferença nas somas de quadrados dos dois modelos. É claro que a soma de quadrados do tipo I não foi afetada, pois a covariável apareceu após todos os demais efeitos do modelo. A opção *solution* permitiu que fosse apresentada a solução de mínimos quadrados. A covariável foi único efeito do modelo cuja estimativa era não viesada. As demais conclusões são similares às já apresentadas anteriormente para este modelo de análise de variação.

## 6.3 Modelos Com Mais de Um Erro

Para analisarmos experimentos mais complexos, contendo mais de um erro e em estruturas não balanceadas, devemos definir quais tipos de somas de quadrados desejamos utilizar, tanto para o tratamento quanto para o resíduo. Além disso, temos que especificar quais são os testadores das fontes de variação do modelo e também qual tipo de soma de quadrados deve ser utilizada para realizar o teste de interesse. Vamos ilustrar este tipo de análise considerando modelos que contenham mais de um erro, a partir do mesmo exemplo de parcela subdividida no tempo, apresentado na seção 5.4. Vamos provocar artificialmente um desbalanceamento no conjunto original de dados para ilustrarmos a análise almejada. Um adubo mineral foi utilizado como fator principal, onde desejávamos comparar seus três níveis 0, 10 e 20 kg/ha. Estas três dosagens foram submetidas a um delineamento em blocos completos casualizados com 2 repetições. O interesse focava o crescimento das plantas ao longo do tempo. Assim, foram avaliadas as alturas das plantas durante 3 meses consecutivos. O modelo estatístico para este experimento é dado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ij} + \gamma_k + \epsilon_{jk} + \delta_{ik} + \epsilon_{ijk} \tag{6.2}$$

em que $Y_{ijk}$ é a observação da altura das plantas em metros, $\mu$ é a constante geral do modelo, $\alpha_i$ é o efeito do i-ésimo nível da adubação química, $\beta_j$ é o efeito do j-ésimo bloco, $\epsilon_{ij}$ é o efeito do erro experimental entre a i-ésima dose e o j-ésimo bloco, $\gamma_k$ é o efeito do k-ésimo mês, $\epsilon_{jk}$ é efeito do erro experimental do j-ésimo bloco com o k-ésimo mês, $\delta_{ik}$ é o efeito da interação entre a i-ésima dose de adubo químico com o k-ésimo mês e $\epsilon_{ijk}$ é o erro experimental entre a i-ésima dose, j-ésimo bloco e k-ésimo mês.

O programa SAS contendo os dados experimentais modificados artificialmente para se tornarem não balanceado e a sintaxe para especificar os erros do modelo e determinar os testes corretos com o tipo de soma de quadrados pretendida é apresentado na seqüência. O comando *test* deve ser utilizado e em suas opções devemos nos preocupar em indicar o tipo de soma de quadrados que utilizaremos. O programa resultante é dado por:

```
/* Programa para realizar análise de variância de um modelo contendo múltiplos erros.
O modelo escolhido foi o de parcela subdividida no tempo com dados não-balanceados.*/
data sub;
input bloco trat mes alt;
cards;
1   0 1 1.00
1 10 1 1.05
1 20 1 1.08
2 10 1 1.06
2 20 1 1.09
1   0 2 1.10
1 10 2 1.12
1 20 2 1.14
2   0 2 1.08
2 10 2 1.15
2 20 2 1.18
1   0 3 1.14
1 10 3 1.20
1 20 3 1.22
2 10 3 1.21
2 20 3 1.23
;
proc glm data=sub;
    class bloco trat mes;
```

```
    model alt = bloco trat bloco*trat mes bloco*mes mes*trat/ss1 ss2 ss3;
    test h=bloco trat e=bloco*trat / htype = 3 etype = 3;
    test h=mes e=bloco*mes /htype = 3 etype = 3;
    lsmeans trat/e=bloco*trat etype = 3 stderr;
    lsmeans mes/e=bloco*mes etype = 3 pdiff stderr adjust=Tukey;
    lsmeans trat*mes/ etype = 3 stderr slice = trat slice = mes;
run; quit;
```

---

Nesta análise podemos destacar que os testes são inicialmente realizados utilizando o erro do modelo (erro C) como testador. Somente com o uso do comando *test* é que este problema foi corrigido. Assim, o teste para bloco e para tratamento foi realizado com o erro A (bloco*trat) e o efeito de mês foi testado com erro B (bloco*mes). No comando *<test h=bloco trat e=bloco*trat / htype = 3 etype = 3;>* especificamos que iríamos utilizar as somas de quadrados do tipo III para tratamento e bloco e também para o resíduo. Comando similar é utilizado para o teste do efeito relativo a mês.

Os comandos solicitando as médias ajustadas de tratamento e de mês são acrescidos das opções para que sejam estipulados o erro e o tipo de somas de quadrados que serão utilizados. Também possibilitam obtermos os erros padrões dos efeitos e no caso de efeitos qualitativos, permitem realizarmos testes de comparações múltiplas com ajuste das probabilidade pelo método de Tukey-Kramer. No caso de efeitos de interação, permitem que sejam realizados desdobramentos com o comando *slice*. O problema do comando *<lsmeans trat*mes/ etype = 3 stderr slice = trat slice = mes;>* é não possibilitar que em alguns desdobramentos pudéssemos utilizar variâncias complexas, como é o caso destes dois tipos de desdobramento realizados. O SAS não permite que especifiquemos erros que são combinações de quadrados médios distintos. Então, apesar de as somas de quadrados estarem corretamente calculadas, os testes de hipóteses desta opção devem ser refeitos manualmente. Um outro problema é a impossibilidade de aplicar um teste de médias para algum desdobramento que tenha apresentado teste de hipótese significativo, utilizando o próprio programa.

## 6.4 Componentes de Variância

Podemos utilizar o *proc glm* para obtermos componentes de variância. Componentes de variância surgem quando alguns dos fatores que estamos estudando são aleatórios. Estes fatores são considerados aleatórios quando temos interesse na população de origem. Os níveis destes fatores são amostras aleatórias destas populações. Assim, temos interesse na média geral daquele efeito e principalmente na variância. Em geral, não temos nenhum interesse particular de comparar os níveis de fator aleatório.

A idéia de um dos métodos para estimarmos os componentes da variância dos efeitos aleatórios do modelo consiste em igualarmos as estimativas dos quadrados médios às suas esperanças $E(QM)$ e resolvermos as equações resultantes. Este método é conhecido como método dos momentos. O *proc glm* permite que obtenhamos as esperanças dos quadrados médios por meio do comando *random*. Um modelo pode ser classificado como fixo, quando todos os seus efeitos, excetuando a média geral e o erro, são fixos. Se todos os efeitos forem aleatórios, temos um modelo aleatório. Se por outro lado, tivermos efeitos fixos e efeitos aleatórios, teremos um modelo misto.

Quando temos efeitos aleatórios no modelo, os testes de hipóteses em muitas situações podem não ser feitos utilizando o quadrado médio do resíduo na obtenção da estatística. A decisão de qual deve ser o denominador da estatística do teste $F$, depende das esperanças dos quadrados médios. Nem sempre a especificação deste denominador é trivial, pois pode haver a necessidade de composição de quadrados médios. A opção *test* do comando *random* permite que testes $F$ adequados sejam feitos nos modelos mistos ou aleatórios. Este comando (*random*) é essencialmente útil quando temos dados não balanceados.

Vamos ilustrar o uso do *proc glm* com um delineamento em blocos casualizados com 2 repetições. Uma amostra aleatória de 5 cultivares foi obtida pelo pesquisador e constituiu o fator de interesse da análise. Adicionalmente, este experimento foi implantado em 2 locais. Assim, este é um exemplo em que aplicaremos uma análise conjunta. Ocorreu, no experimento do local 1, uma perda de parcela. A repetição 1 da cultivar 5 foi perdida.

O interesse reside no componente de variância para cultivar, que foi considerada de efeito aleatório. O efeito de bloco, em geral, é considerado como aleatório na literatura. Pelo fato de o efeito de cultivar ter sido considerado aleatório e o de local fixo, a interação é considerada aleatória. Os comandos SAS, necessários para estimarmos os componentes de variância dos efeitos aleatórios, são dados por:

```
/* Programa para realizar análise de variância conjunta de um modelo misto.*/
data rand;
input cult bl local prod;
cards;
1 1 1 8.4
1 2 1 8.6
2 1 1 5.7
2 2 1 5.8
3 1 1 4.5
3 2 1 6.7
4 1 1 5.9
4 2 1 7.8
5 2 1 8.9
1 1 2 6.2
1 2 2 7.6
2 1 2 8.3
2 2 2 9.5
3 1 2 3.5
3 2 2 4.9
4 1 2 7.4
4 2 2 8.8
5 1 2 8.9
5 2 2 9.0
;
proc glm data=rand;
    class cult bl local;
    model prod = bl(local) cult local cult*local / e3 ss3;
    random bl(local) cult cult*local / test;
run; quit;
```

Merecem destaques alguns comandos e especificações de modelo utilizados. O comando *<model prod = bl(local) cult local cult*local / e3 ss3;>* possui o efeito de bloco hierarquizado em local. Não podemos especificar apenas o efeito de bloco, pois estaríamos ignorando o fato de que os blocos dos diferentes locais não são os mesmos. Assim, o bloco 1 do local 1 é diferente do bloco 1 do local 2. As opções *e3* e *ss3* indicam que as esperanças dos quadrados médios, utilizando somas de quadrados do tipo III, devem ser utilizadas. No comando *<random bl(local) cult cult*local / test;>*, que aparece após o comando *model*, indicamos ao *proc glm* quais são os efeitos aleatórios do modelo. Neste exemplo foram os efeitos de bloco dentro de local, de cultivar e da interação cultivar × local.

Inicialmente o SAS apresenta o resultado da análise de variância do tipo III, cujo resumo apresentamos na Tabela 6.4. Se o modelo possui efeitos aleatórios, os testes de significância (teste $F$) apresentados nesta análise provavelmente podem estar incorretos. Neste exemplo, como apenas o efeito de local é considerado fixo, sendo todos os demais aleatórios, a maioria dos testes $F$ está incorreta. O correto é utilizar as esperanças dos quadrados médios para especificar os testes de hipóteses adequados e também para estimar os componentes de variância.

Tabela 6.4: Análise da variação para o modelo de análise conjunta (2 locais) em um delineamento de blocos casualizados.

| FV | G.L. | SQ III | QM | $F$ | $Pr > F$ |
|---|---|---|---|---|---|
| Modelo | (11) | (52,9816) | 4,8165 | 13,65 | 0,0011 |
| bl(local) | 2 | 5,4450 | 2,7225 | 7,72 | 0,0170 |
| cult | 4 | 27,4770 | 6,8693 | 19,47 | 0,0007 |
| local | 1 | 0,7111 | 0,7111 | 2,02 | 0,1987 |
| cult*local | 4 | 15,5483 | 3,8871 | 11,02 | 0,0038 |
| Erro | 7 | 2,4700 | 0,3529 | | |
| Total | 18 | 55,4516 | | | |

CV = 8,27% $\quad \bar{Y}_{...} = 7,1789$

Um segundo resultado apresentado pelo SAS, associado a análise de variação, refere-se as esperanças dos quadrados médios. Estes resultados

estão sumariados na Tabela 6.5. Uma análise das esperanças dos quadrados médios mostra que o testador para bloco(local) e para a interação cultivar × local é o erro experimental. O testador para cultivar é a interação cultivar × local e o testador para local tem de ser obtido por uma combinação de quadrados médios. A opção *test* do comando *random* nos permite obter as estatísticas destes testes automaticamente.

Tabela 6.5: Esperança dos quadrados médios e resumo da análise da variação para o modelo de análise conjunta (2 locais) em um delineamento de blocos casualizados.

| FV | G.L. | QM | E(QM) |
|---|---|---|---|
| bl(local) | 2 | 2,7225 | $\sigma^2 + 4,5\sigma^2_{b(L)}$ |
| cult | 4 | 6,8693 | $\sigma^2 + 1,8333\sigma^2_{CL} + 3,6667\sigma^2_C$ |
| local | 1 | 0,7111 | $\sigma^2 + 1,7778\sigma^2_{CL} + 4,4444\sigma^2_{b(L)} + Q_L$ |
| cult*local | 4 | 3,8871 | $\sigma^2 + 1,8333\sigma^2_{CL}$ |
| Erro | 7 | 0,3529 | $\sigma^2$ |

$Q_L$ é a forma quadrática associada a local

A estimativa do componente de variância de cultivar pode ser obtida por: $\hat{\sigma}_C = (QMCult - QMCult \times Local)/3,6667 = 0,8133$. Os demais componentes de variância podem ser obtidos de maneira similar. Muitas vezes temos dificuldades em determinar qual é o quadrado médio que devemos subtrair do quadrado médio correspondente ao fator aleatório para o qual desejamos estimar o componente. Para a interação, isso foi obtido de uma maneira bastante simples por $\hat{\sigma}_{CL} = (QMCult \times Local - QMErro)/1,8333 = 1,9278$. Quando precisamos combinar quadrados médios, o melhor indicativo para determinarmos esta combinação é fornecida pelo comando *test*. Por exemplo, se desejássemos testar a hipótese de que o efeito quadrático $Q_L$ devido a local, que é fixo, seja nulo, poderíamos utilizar a seguinte combinação de quadrados médios como denominador da expressão da estatística do teste $F$:

$$0,9877\text{QMbl(local)} + 0,9697\text{QMcult} \times \text{local} - 0,9574\text{QMErro},$$

cujos graus de liberdade associados seriam obtidos pelo processo de Sat-

terthwaite (1946)[11].

Utilizando os testes adequados apenas os efeitos de bloco(local) e da interação cultivar × local foram significantes, indicando que os componentes de variância associados são diferentes de zero. Para cultivar não foi detectada significância estatística, sendo considerado nulo o componente de variância associado. Outras tipos de somas de quadrados podem ser utilizadas para estimarmos componentes de variância e para realizarmos os testes $F$. Para selecionarmos, por exemplo, as somas de quadrados do tipo II, bastaria trocar o comando *<model prod = bl(local) cult local cult*local / e3 ss3;>* por *<model prod = bl(local) cult local cult*local / e2 ss2;>*. Quando aplicamos esta mudança, os resultados dos testes são praticamente idênticos aos obtidos com as somas de quadrados do tipo III.

O SAS possui outros procedimentos para estimarmos componentes de variância. Podemos destacar o *proc mixed* e o proc *proc varcomp*. Estes procedimentos são muitas vezes mais adequados para estimarmos componentes de variância, além de oferecerem mais alternativas de métodos. Discutiremos o varcomp posteriormente neste material. Os modelos mistos são uma generalização dos modelos lineares utilizados no *proc glm*.

## 6.5 Exercícios

1. Utilizar dados não balanceados resultantes de pesquisas desenvolvidas em sua área e realizar análises de variâncias utilizando o *proc glm*. Aplicar os testes de médias, se os níveis forem qualitativos, ou ajustar modelos de superfície de resposta ou de regressão, se os níveis dos fatores forem quantitativos.

2. Dar sua opinião sobre o fato de muitos autores ainda recomendarem estimação de parcelas, em conjuntos de dados onde foram perdidas uma ou mais delas. Como você lidaria com conjuntos de dados não balanceados? Estimaria os valores perdidos?

# Capítulo 7

# Componentes de Variância

O *varcomp* foi designado para lidar com modelos lineares que possuam efeitos aleatórios. Efeitos aleatórios são fatores cujos níveis são amostras aleatórias de uma população de possíveis infinitos níveis. O *proc varcomp* estima a contribuição de cada fator aleatório para a variância da variável resposta. Vários métodos existem para a estimação dos componentes de variância. O *proc varcomp* possui implementado os métodos *type 1* (baseado no cômputo da soma de quadrados do tipo I para cada efeito do modelo), *MIVQUE0*, máxima verossimilhança (ML) e máxima verossimilhança restrita (REML).

Componentes de variância são, por definição, positivos. No entanto, estimativas negativas podem ocorrer. Algumas razões potenciais para que estimativas negativas de componentes de variância ocorram podem ser destacadas por:

- Variabilidade muito grande dos dados, produzindo estimativas negativas, apesar do valor verdadeiro do componente ser positivo;
- Presença de *outliers* nos dados experimentais;
- Especificação incorreta do modelo estatístico.

Alguns métodos específicos para lidarmos com cada uma destas situações existem. No caso de *outliers*, análises exploratórias de dados podem ser aplicadas facilmente para identificação e eliminação destas observações

discrepantes. A especificação incorreta do modelo está diretamente sob o controle do pesquisador que ao identificar o problema pode prontamente corrigí-lo.

## 7.1 Métodos de Estimação de Componentes de Variância

O método denominado por *Type 1* é um método dos momentos. As esperanças dos quadrados médios são determinadas e igualadas aos quadrados médios de uma análise de variância seqüencial (somas de quadrados do tipo I). O método *Mivque0* é baseado no método de Hartley, Rao e LaMotte (1978)[7], o qual produz estimativas que são invariantes em relação aos efeitos fixos do modelo e são localmente os melhores estimadores quadráticos não viciados. Possui estimação semelhante a do método *Type 1*, exceto pelo fato de que os efeitos aleatórios são ajustados somente para os efeitos fixos.

Os estimadores de Máxima Verossimilhança (ML) para os componentes de variância usam a transformação W, desenvolvida por Hemmerle e Hartley (1973)[8] e Goodnigth e Hemmerle (1978)[6] e o algoritmo de Newton-Raphson, aplicado iterativamente até que o logaritmo da função de verossimilhança seja maximizado. O método da máxima verossimilhança restrita (REML) é semelhante ao ML, só que há uma separação da função de verossimilhança em duas partes. A primeira com os efeitos fixos e a segunda com os aleatórios (Patterson e Thompson, 1971[10]).

## 7.2 O Proc Varcomp

Para apresentarmos os comandos do *proc varcomp*, ilustrando a forma de especificar tanto os métodos, quanto os efeitos fixos, vamos utilizar o delineamento em blocos casualizados com 2 repetições, apresentado no capítulo 6. Uma amostra aleatória de 5 cultivares foi obtida. Adicionalmente, este experimento foi conduzido em 2 locais. Ocorreu, no local 1, a perda da parcela correspondente à repetição 1 da cultivar 5. Todos os efeitos do modelo foram considerados aleatórios, exceto a média geral (por razões óbvias) e o efeito de local. O programa SAS resultante é dado por:

```
/* Programa para estimar componentes de variância em um modelo misto.*/
data rand;
input cult bl local prod;
cards;
1 1 1 8.4
1 2 1 8.6
2 1 1 5.7
2 2 1 5.8
3 1 1 4.5
3 2 1 6.7
4 1 1 5.9
4 2 1 7.8
5 2 1 8.9
1 1 2 6.2
1 2 2 7.6
2 1 2 8.3
2 2 2 9.5
3 1 2 3.5
3 2 2 4.9
4 1 2 7.4
4 2 2 8.8
5 1 2 8.9
5 2 2 9.0
;
proc varcomp data=rand maxiter=500 method=type1;
    class cult bl local;
    model prod = local bl(local) cult cult*local /fixed = 1;
run; quit;
```

Na linha de comando *<proc varcomp data = rand maxiter = 500 method = type1;>* declaramos o número máximo de iterações para o processo iterativo, por meio da opção *maxiter=500*, e o método que desejamos utilizar, com a opção *method=type1*. Neste caso, limitamos em no máximo 500 iterações e utilizamos o método *type 1*. Podemos alterar o método, substituindo *type1* por *mivque0*, *ML* ou *RML*. Diferentemente do *proc glm*, onde com o comando *random* especificamos os efeitos aleatórios, no *proc varcomp* devemos mencionar o número de efeitos fixos do modelo. Assim, com

o comando *<model prod = local bl(local) cult cult*local /fixed = 1;>*, informamos ao programa que temos um efeito fixo (*fixed=1*) e que o efeito de local é este efeito fixo. O programa ao ser informado do número de efeitos fixos, começa a reconhecê-los a partir da igualdade (primeiro efeito do modelo) entre a parte dependente e independente do modelo. Devemos, portanto, posicionar os efeitos fixos antes dos efeitos aleatórios no modelo especificado, quando utilizamos o *proc varcomp*.

O SAS apresenta entre os seus resultados a análise de variância e as esperanças dos quadrados médios para o método *Type 1*. Para os demais métodos, alguns outros resultados particulares são apresentados. Em todos os casos temos as estimativas dos componentes de variância dos efeitos aleatórios. Alteramos a opção *method = type1*, considerando as demais possibilidades, para estimarmos os componentes de variância utilizando todos os métodos (*mivque0*, *ml* ou *reml*) e apresentamos os resultados na Tabela 7.1.

Tabela 7.1: Estimativas dos componentes de variância para o modelo de análise conjunta (2 locais) em um delineamento de blocos casualizados utilizando os 4 métodos de estimação do *proc varcomp*.

| | | Método | | | |
|---|---|---|---|---|---|
| FV | G.L. | Type 1 | Mivque0 | ML* | REML* |
| bl(local) | 2 | 0,69760 | 0,71978 | 0,38173(0,37) | 0,54146(0,62) |
| cult | 4 | 0,83428 | 0,89047 | 0,78798(1,18) | 0,96363(1,55) |
| cult*local | 4 | 1,92776 | 2,03984 | 1,51873(1,10) | 1,79084(1,39) |
| Erro | 7 | 0,35286 | 0,19096 | 0,35252(0,20) | 0,34854(0,17) |

* Erro padrão das estimativas entre parênteses.

O SAS apresenta a matriz de covariância dos estimadores dos componentes de variância dos efeitos aleatórios do modelo para os métodos da máxima verossimilhança e da máxima verossimilhança restrita. A raiz quadrada dos elementos da diagonal são os erros padrões das estimativas destes componentes de variâncias, que foram apresentados na Tabela 7.1. Em geral, os erros padrões das estimativas associadas ao método da máxima verossimilhança restrita foram maiores do que os do método da máxima

verossimilhança.

Um segundo exemplo, para ilustrar a estimação de componentes de variância negativos, é apresentado na seqüência. Para isso um delineamento em blocos casualizados com 5 cultivares e 2 repetições foi considerado. Duas repetições dentro de cada bloco foram obtidas. Uma das repetições dentro do bloco 1, para a cultivar 5, foi perdida. O modelo foi considerado aleatório e dado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ij} + \delta_{k(ij)} \tag{7.1}$$

em que $Y_{ijk}$ é o valor observado da variável resposta, $\mu$ é a constante geral, $\alpha_i$ é o efeito aleatório do i-ésimo nível das cultivares, $\beta_j$ é o efeito aleatório do j-ésimo nível dos blocos, $\epsilon_{ij}$ é o efeito aleatório do erro experimental suposto normal e independentemente distribuído com média 0 e variância comum $\sigma_e^2$ e $\delta_{kij}$ é o efeito do erro amostral aleatório suposto normal e independentemente distribuído com média 0 e variância comum $\sigma^2$.

O programa SAS para estimarmos os componentes de variância é dado por:

```
/* Programa para estimar componentes de variância em um modelo aleatório.*/
data vc2;
input cult bl rep prod;
cards;
1 1 1 8.4
1 2 1 7.6
2 1 1 5.7
2 2 1 5.8
3 1 1 4.5
3 2 1 6.7
4 1 1 8.9
4 2 1 7.8
5 2 1 8.9
1 1 2 6.2
1 2 2 7.6
2 1 2 8.3
2 2 2 2.5
```

```
3 1 2 3.5
3 2 2 4.9
4 1 2 7.4
4 2 2 8.8
5 1 2 8.9
5 2 2 9.0
;
proc varcomp data=vc2 maxiter=500 method=type1;
    class cult bl;
    model prod = cult bl bl*cult;
run; quit;
```

O erro amostral dado pelo efeito de repetição dentro de cada combinação de cultivar × bloco foi obtido por diferença e o erro experimental é dado pela interação bloco × cultivar. Alterando a opção *<method=type1>* para os demais métodos, obtivemos as estimativas dos componentes de variância apresentados na Tabela 7.2.

Tabela 7.2: Estimativas dos componentes de variância para o modelo de blocos casualizados com repetição dentro de cada bloco em um ensaio de cultivares, utilizando os 4 métodos de estimação do *proc varcomp*.

| | | Método | | | |
|---|---|---|---|---|---|
| FV | G.L. | Type 1 | Mivque0 | ML* | REML* |
| cult | 4 | 2,11787 | 1,96139 | 1,70757(1,54) | 2,30153(2,12) |
| bl | 1 | -0,30145 | -0,34551 | 0,00000(0,00) | 0,00000(0,00) |
| Erro | 4 | 0,63854 | 0,80142 | 0,40027(0,85) | 0,39980(0,85) |
| Erro amostral | 9 | 1,66611 | 1,66676 | 1,62392(0,75) | 1,62262(0,75) |

* Erro padrão das estimativas entre parênteses.

Grandes diferenças podem ser observadas nas estimativas dos componentes de variância. Uma delas são as estimativas negativas dos componentes de variância nos métodos *Type 1* e *Mivque0*. É uma prática comum tratar as estimativas negativas como se elas fossem nulas. Nos métodos *ML* e *REML* este procedimento já é feito automaticamente durante o processo de estimação e componentes de variância negativos são evitados.

## 7.3 Exercícios

1. Exemplificar situações em sua área em que componentes de variância poderiam ser estimados.

2. Podemos utilizar intervalos de confiança normais para componentes de variância se considerarmos a propriedade de normalidade assintótica dos estimadores de máxima verossimilhança. Assim, construir intervalos de confiança normais para os componentes de variância de cultivares $\sigma^2_C$ nos dois exemplos, utilizando a seguinte expressão:

$$IC_{1-\alpha}(\sigma^2_C) : \hat{\sigma}^2_C \pm Z_{\alpha/2} EP(\hat{\sigma}^2_C)$$

em que $Z_{\alpha/2}$ é o quantil superior $100\alpha/2\%$ da distribuição normal padrão e $EP(\hat{\sigma}^2_C)$ é o erro padrão do estimador do componente de variância de cultivar.

# Capítulo 8

# Pressuposições da Análise de Variância

A validade da análise de variância depende que algumas condições pressupostas sejam atendidas. Quando um estatístico formula um modelo e estima seus parâmetros e propõe algum método de estimação ou teste, há a necessidade de que algumas condições sejam ratificadas. A validade desta inferência depende de algumas restrições impostas aos efeitos deste modelo, como por exemplo, a suposição de normalidade dos erros. Se o pesquisador obtiver um conjunto de dados amostrais, em que essas condições não foram obedecidas, então a validade das inferências realizadas é no mínimo questionável. Especificamente no caso dos modelos lineares, fazemos suposições de distribuição normal dos erros, aditividade dos efeitos do modelo e homogeneidade das variâncias dos erros associados aos níveis de um determinado efeito ou fator. Estas pressuposições muitas vezes não são checadas, o que pode comprometer a validade dos resultados dos testes e da estimação realizados. Desta forma, o pesquisador pode eventualmente tomar decisões errôneas.

Uma das razões de se ignorar a checagem das pressuposições para validade da análise de variância é a dificuldade de se encontrar recursos computacionais para realizar esta tarefa. A maioria dos *softwares* não checa estas pressuposições, ou não possui rotinas para realização destes testes.

O programa SAS, pela sua flexibilidade e facilidade de programação,

permite que muitos métodos, existentes para esta finalidade, sejam implementados. No entanto, os testes existentes na literatura, para checarmos se as pressuposições foram atendidas, são específicos para alguns modelos, o que dificulta a sua aplicação em casos mais gerais. Um outro fator limitante diz respeito ao fato de que estes procedimentos ficariam limitados a pesquisadores que tivessem uma maior familiaridade com a linguagem SAS. Desta forma, a busca de procedimentos mais gerais e mais fáceis de utilizar, facilitaria a verificação das pressuposições feitas aos efeitos do modelo. Para isso, Gill (1978)[4] apresenta alguns métodos mais abrangentes, que são tratados nas próximas seções. Vamos apresentar os testes para verificar a normalidade dos resíduos e a aditividade dos efeitos do modelo.

## 8.1 Normalidade dos Resíduos

A pressuposição de normalidade, exigida na análise de variância, é na maioria das vezes mal interpretada e checada de forma incorreta. A exigência que se faz, a respeito da distribuição normal, é para a distribuição dos resíduos de um determinado modelo linear e não para os dados observados nas unidades experimentais. Muitos pesquisadores desavisados, ou por desconhecimento, realizam o teste de normalidade nos dados experimentais observados, o que é uma prática incorreta. Este procedimento só seria válido se estivéssemos avaliando uma amostra aleatória de uma única população, cujos dados pudessem ser explicados pelo modelo linear simples dado por $Y_i = \mu + \epsilon_i$. Em modelos onde temos um ou mais fatores, os valores da variável $Y_i$ são explicados por diferentes constantes ao longo da amostra aleatória de tamanho $n$. Assim, por exemplo, para o modelo $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ temos diferentes constantes $\mu + \tau_i$, que são funções do i-ésimo nível do efeito $\tau_i$. Então a distribuição da variável $Y$ é na verdade uma mistura de normais com diferentes médias. Quanto maior a complexidade do modelo, mais complexa fica esta mistura de distribuições normais.

Como a suposição de normalidade que fazemos é para o erro deste modelo, que é uma variável aleatória não observável, temos de estimá-lo e então aplicar os testes de normalidade. Podemos utilizar os recursos do SAS para realizar esta tarefa. O SAS permite que estimemos e salvemos os erros dos

modelos em um SAS *data set* em cada procedimento. Se utilizarmos o teste de normalidade de Shapiro-Wilk do *proc univariate*, poderemos avaliar se a pressuposição de normalidade foi atendida. Vamos utilizar um exemplo de um experimento realizado em blocos casualizados com 4 repetições e 3 tratamentos de um único fator. O modelo estatístico é dado por:

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \tag{8.1}$$

em que $Y_{ij}$ é o valor observado da variável resposta produção, $\mu$ é a constante geral, $\tau_i$ é o efeito do i-ésimo nível dos tratamentos, $\beta_j$ é o efeito do j-ésimo nível dos blocos e $\epsilon_{ij}$ é o efeito do erro experimental suposto normal e independentemente distribuído com média 0 e variância comum $\sigma^2$.

Os valores preditos da variável resposta são dados por $\hat{Y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j$, que de forma matricial podem ser obtidos por $\underset{\sim}{\hat{Y}} = X\underset{\sim}{\tilde{\beta}}$, em que $\underset{\sim}{\hat{Y}}$ é o vetor de observações, $X$ é matriz do modelo e $\underset{\sim}{\tilde{\beta}}$ é o vetor de soluções de mínimos quadrados. Assim, os resíduos são estimados por $\hat{\epsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}$ ou simultaneamente por $\underset{\sim}{\hat{\epsilon}} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}}$. Após estimarmos os resíduos do modelo, aplicamos o teste de Shapiro-Wilk utilizando o *proc univariate*. O programa SAS para realizarmos o teste de normalidade dos resíduos do exemplo que estamos considerando é dado por:

```
/* Programa para testar a pressuposição de erros normais em um modelo linear em blocos
casualizados.*/
data press1;
input bl trat prod;
cards;
1 1 12.34
1 2 13.45
1 3 14.56
2 1 12.34
2 2 16.78
2 3 17.89
3 1 10.32
3 2 15.67
3 3 16.01
```

```
4 1 13.45
4 2 16.78
4 3 17.89
;
proc glm data=press1;
    class bl trat;
    model prod = bl trat;
    output out=norm P=pred R=res;
run;quit;
proc univariate data=norm normal;
    var res;
run;quit;
```

---

Realizamos a análise de variância para estimarmos os resíduos, utilizando o *proc glm* para isso. Armazenamos os resíduos e os valores preditos em um SAS *data set* utilizando o comando *<output out=norm P=pred R=res;>*. Definimos que a variável correspondente aos valores preditos seria denominada de *pred* e a dos resíduos de *res*. Utilizamos o *proc univariate* na seqüência para aplicar o teste de normalidade a variável *res* do SAS *data set norm*. O resultado que nos interessa é o do teste de Shapiro-Wilk. O valor observado da estatística foi $W = 0,946844$ e o valor-p associado foi igual a $0,5914$. Assim, não devemos rejeitar a hipótese nula de normalidade dos resíduos, se considerarmos um nível nominal de significância de $\alpha = 0,05$.

## 8.2 Aditividade

Em um modelo linear, assumimos que os efeitos são aditivos e não multiplicativos (Tukey, 1949[14]). O método de Tukey decompõe a soma de quadrado do erro em duas partes. Uma delas com apenas 1 grau de liberdade e a outra com os graus de liberdade remanescentes. Um teste $F$ é aplicado e denominado de teste da não-aditividade de Tukey. Este teste da não-aditividade de Tukey pode ser generalizado para possibilitar sua aplicação em diversos modelos lineares. Esta generalização consiste em obtermos os valores preditos e em seguida introduzirmos o seu quadrado como covariável no modelo de análise de variância. Esta análise se prestará

unicamente para testarmos a hipótese de aditividade dos efeitos. Se houver efeito significativo da covariável, deveremos rejeitar a hipótese nula de efeitos aditivos.

Utilizando o exemplo da seção 8.1 e definindo os valores preditos por $\hat{Y}_{ij}$, devemos ajustar o seguinte modelo linear:

$$Y_{ij} = \mu + \tau_i + \beta_j + \lambda \hat{Y}_{ij}^2 + \epsilon_{ij} \tag{8.2}$$

em que $\lambda$ é o coeficiente de regressão associado à covariável determinada pelos valores preditos ao quadrado; os demais efeitos têm os mesmos significados do modelo 8.1.

A hipótese de interesse $H_0 : \lambda = 0$ é equivalente à hipótese nula de que o modelo é aditivo. Devemos realizar uma análise de covariância e realizar o teste de interesse sobre o efeito da covariável, que como já dissemos, é equivalente ao teste de aditividade dos efeitos. Infelizmente este procedimento não pode ser utilizado em experimentos inteiramente casualizados com um fator, por razões óbvias, ou com dois fatores e interação, pois haverá um confundimento da interação com o efeito da covariável. O programa SAS utilizado para aplicarmos este teste aos dados do exemplo da seção 8.1 é dado por:

---

```
/* Programa para testar a pressuposição de efeitos aditivos em um modelo linear em
blocos casualizados.*/
data press2;
input bl trat prod;
cards;
1 1 12.34
1 2 13.45
1 3 14.56
2 1 12.34
2 2 16.78
2 3 17.89
3 1 10.32
3 2 15.67
3 3 16.01
4 1 13.45
```

```
4 2 16.78
4 3 17.89
;
proc glm data=press2;
    class bl trat;
    model prod = bl trat;
    output out=norm P=pred R=res;
run;quit;
data norm; set norm;
    pred2=pred*pred;
run;quit;
proc glm data=norm;
    class bl trat;
    model prod= bl trat pred2;
run;quit;
```

Observamos um valor da estatística $F$ para o teste de $F_c = 1,02$ com $\nu_1 = 1$ e $\nu_2 = 5$ graus de liberdade. O valor-p associado foi de $0,3581$, portanto não devemos rejeitar a hipótese nula, indicando que não existem evidências significativas (5%) para afirmarmos que haja não-aditividade dos efeitos do modelo. Para o caso de rejeitarmos a hipótese nula, Tukey (1949)[14] recomenda algum tipo de transformação dados para corrigir o problema. A justificativa para tentar eliminar o problema é baseada no fato de que o teste $F$ na presença da não-aditividade é considerado bastante conservador.

## 8.3 Homogeneidade de Variâncias

A suposição de que os erros $\epsilon_{ij}$ de um modelo têm distribuição normal e variância comum, indica que as variâncias dos diferentes níveis dos fatores presentes no modelo devem ser homogêneas. Para o modelo inteiramente casualizado com um fator, apresentamos o teste de homogeneidade de variâncias na seção 5.2 de acordo com os procedimentos descritos por Ferreira (2005)[3]. O *proc anova* do SAS, no caso de um fator único no modelo, nos possibilita testar a homogeneidade de variâncias entre os níveis do fator.

Em casos mais gerais Gill (1978)[4] recomenda utilizar como covariável

os valores preditos do resíduo ao quadrado. Por não termos avaliado este procedimento e não conhecermos na literatura nenhum indicativo científico de sua validade, optamos por não apresentar maiores detalhes deste método.

## 8.4 Exercícios

1. Aplicar testes de normalidade para alguns modelos de regressão apresentados no capitulo 3.

2. Em sua opinião qual dos três pressupostos causaria mais impacto sobre a validade das inferências?

# Referências Bibliográficas


[1] BECKMAN, R. J.; TRUSSELL, H. J. The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association*, 69:179–201, 1974. 62

[2] CHATTERJEE, S.; HADI, A. S. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986. 59, 61, 62, 64

[3] FERREIRA, D. F. *Estatística básica*. Editora UFLA, Lavras, 2005. 676p. 12, 15, 92, 98, 99, 148

[4] GILL, J. W. *Design and analysis of experiments in the animal and medical sciences.*, volume 2. Iowa State University, Ames, 1978. 301p. 144, 148

[5] GOMES, F. P. *Curso de estatística experimental*. Esalq/Usp, Piracicaba, 14 edition, 2000. 476p. vii, 93

[6] GOODNIGTH, J. H.; HEMMERLE, W. J. A simplified algorithm for the W-transformation in variance component estimation. *Technometrics*, 21:265–268, 1978. 136

[7] HARTLEY, H. O.; RAO, J. N. K.; LaMOTTE, L. A simple synthesis-based method of variance component estimation. *Biometrics*, 34:233–244, 1978. 136

[8] HEMMERLY, W. J.; HARTLEY, H. O. Computing maximum likelihood estimates for mixed AOV model using the W-transformation. *Technometrics*, 15:819–831, 1973. 136

[9] O'NEILL, R.; WETHERILL, G. B. The present state of multiple comparison methods. *Journal of the Royal Statistical Society*, 33(2):218–250, 1971. 97

[10] PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971. 136

[11] SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946. 21, 30, 32, 134

[12] SEARLE, S. R. *Linear models*. John Wiley, New York, 1971. 532p. 40

[13] SEARLE, S. R. *Linear models for unbalanced models*. John Wiley, New York, 1987. 536p. 40

[14] TUKEY, J. W. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949. 146, 148

[15] VANGEL, M. G. Confidence intervals for a normal coefficient of variation. *The American Statistician*, 15(1):21–26, 1996. 19

[16] VELLEMAN, P. F.; WELSCH, R. E. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981. 59, 63

# Índice Remissivo