4

Comparison of Methodologies for Analysis of Longitudinal Data Using MATLAB

João Eduardo da Silva Pereira, Janete Pereira Amador and Angela Pellegrin Ansuj
Federal University of Santa Maria
Brazil

1. Introduction

In several areas of scientific knowledge there is a need for studying the behavior of one or more variables using data generated by repeated measurements of the same unit of observations along time or spatial region. Due to this, many experiments are constructed in which various treatments are applied on the same plot at different times, or only one treatment is applied to an experimental unit and it is made a measurement of a characteristic or a set of features in more than one occasion [Khattree & Naik, 2000]. Castro and Riboldi [Castro & Riboldi, 2005] define data collected under these kinds of experimental setups as repeated measures. More specifically, he asserts that "repeated measures is understood as the data generated by repeatedly observing a number of investigation units under different conditions of evaluation, assuming that the units of investigation are a random sample of a population of interest". In order to analyze repeated measures data it is necessary to take a care about not independency between observations. This is so because it is expected a high degree of correlation between data collected on the same observation unit over time, and there is usually more variability in the measurements between the subjects than within a given subject. A very common type of repeated measures is longitudinal data, i.e., repeated measures where the observations within units of investigation were not or can not have been randomly assigned to different conditions of evaluation, usually time or position in space.

There are basically two paths to be taken in the analysis of longitudinal data; univariate analysis, which requires as a precondition a rigid structure of covariances, or multivariate analysis, which, despite being more flexible, is less efficient in detecting significant differences than the univariate methodology.

In Advances in Longitudinal Data Analysis [Fitzmaurice et al., 2009], Fitzmaurice comments that despite the advances made in statistical methodology in the last 30 years there has been a lag between recent developments and their widespread application to substantive problems, and adds that part of the problem why the advances have been somewhat slow to move into the mainstream is due to their limited implementation in widely available standard computer software.

In this context this work proposes to develop a single and easy computational implementation to solve a great number of practical problems of analysis of longitudinal

data, through the decomposition of the sum of squares error of the polynomial models of regression.

In light of the above, not independents the computational support MatLab looks likes an ideal tool for the implementation and dissemination of this kind of statistical analysis methods, and linear models, first because its matrix structure fits perfectly well for linear models which facilitates the construction of models for univariate and multivariate analysis, and second because being a large diffusion tool of, it allows for that the models to be implemented, modified and reused in several uses in different situations by several users who have access to a MatLab community on the internet. This avoids the need for the acquisition of expensive software with black box structure.

2. Review

As far as the analysis of experiments using longitudinal data is concerned the methods traditionally used are: univariate analyis or Univariate Profile Model whereby longitudinal data is considered as if it were observations done in subdivisions of the slots, usually requiring that the variance of the response be constant in the occasions of evaluation and that the covariance between responses in different occasions be equal; multivariate analysis or Multivariate Profile Model whereby it is admitted that these variances and covariances be distinct. Despite its apparent versatility, as far as the dimension of the matrix of variances and covariances, the multivariate model becomes less attractive, because its results are hard to interpret, and its estimates are not consistent. The univariate profile model gives consistent estimates and should be used every time when its presuppositions are met. Otherwise, the multivariate profile model is a viable alternative [Castro & Riboldi, 2005; Johnson & Wichern, 1998].

Using the univariate analysis in split-plot designs, regarding time as a sub-plot may cause problems because, as it is known, this design presupposes that the covariance matrix meets the condition of sphericity which does not always happen. What is found in the literature is that repeated measures in one same experimental unit along time are in general correlated, and that these correlations are greater for closer times [Malheiros, 1999].

Xavier [Xavier, 2000] asserts that a sufficient condition for the F test of the analysis of variance of the sub-plots for the time factor and the interaction time*treatments, be valid, is that the covariance matrix has a so called composite symmetry shape. The composite symmetry occurs when the variance and covariance matrix may be expressed as:

$$\Sigma = \begin{bmatrix} (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) \end{bmatrix} \quad (1)$$

where:

σ^2 : is the variance of the sub-plot (within-subjects);

σ_1^2 : is the variance of the plot (among-subjects).

The composite symmetry condition implies that the random variable be equally correlated and has equal variances considering the different occasions. A more general condition of the

$\sum$ is described by Huynh and Feldt [Huynh & Feldt, 1970]. This condition, called HUYNH-FELDT (H-F) or sphericity condition (circularity), specifies that the elements of the $\sum$ matrix be expressed for one $\lambda > 0$, as:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \frac{(\sigma_1^2+\sigma_2^2)}{2}-\lambda & \frac{(\sigma_1^2+\sigma_3^2)}{2}-\lambda & \frac{(\sigma_1^2+\sigma_4^2)}{2}-\lambda \\ \frac{(\sigma_1^2+\sigma_2^2)}{2}-\lambda & \sigma_2^2 & \frac{(\sigma_2^2+\sigma_3^2)}{2}-\lambda & \frac{(\sigma_2^2+\sigma_4^2)}{2}-\lambda \\ \frac{(\sigma_3^2+\sigma_1^2)}{2}-\lambda & \frac{(\sigma_3^2+\sigma_2^2)}{2}-\lambda & \sigma_3^2 & \frac{(\sigma_3^2+\sigma_4^2)}{2}-\lambda \\ \frac{(\sigma_4^2+\sigma_1^2)}{2}-\lambda & \frac{(\sigma_4^2+\sigma_2^2)}{2}-\lambda & \frac{(\sigma_4^2+\sigma_3^2)}{2}-\lambda & \sigma_4^2 \end{bmatrix} \quad (2)$$

where λ is the difference between the means of the variances and the means of the covariances.

The H-F condition is necessary and sufficient for the F test in the usual analysis of variance in split-plot in time to be valid. This condition is equivalent to specifying that the variances of the difference between pairs of errors are equal, and if the variances are all equal then the condition is equivalent to compound symmetry [Xavier, 2000].

To check the condition of circularity Mauchly [Mauchly, 1940] presents the test of sphericity. This test uses H-F condition for the covariance matrix of (t-1) normalized orthogonal contrasts for repeated measures not correlated with equal variances. Vonesh and Chinchilli [Vonesh & Chinchilli, 1997] state that the sphericity test is not very powerful for small samples and is not robust when there is violation of the normality assumption.

According to Box; Greenhouse & Geisser; and Huynh & Feldt [Box, 1954; Greenhouse & Geisser, 1959; Huynh & Feldt, 1976], although the matrix $\sum$ may not satisfy the condition of sphericity, the central F distribution may be used, in an approximate form, if a correction in the degrees of freedom associated with the causes of variation involving the time factor is made. The degrees of freedom correction in these sources of variation is done by multiplying the original degrees by a factor ε. When $\sum$ is uniform, the value of ε =1.

According to Freitas [Freitas, 2007] the correction of the number of degrees of freedom should be made only in statistics that involve comparisons within subjects (time factor and interaction time*treatments). The statistics involving comparisons between subjects do not need corrections in the degrees of freedom because there is always an exact central F distribution.

When the pattern of the $\sum$ matrix is not satisfied, not even close, the multivariate techniques are used since this type of solution is applicable to any $\sum$ matrix. The only requirement of the multivariate procedure is that the $\sum$ matrix should be common to all treatments.

Due to the essentially multivariate nature of the response vectors, in studies involving longitudinal data, the multivariate analysis technique also known as multivariate profile analysis is a natural alternative to the problem at hand [Wald, 2000]. The multivariate profile analysis is well discussed in the literature by authors such as [Lima, 1996; Morrison, 1990; Singer, 1986].

The multivariate profile analysis is one of the statistics technique used to analyze observations derived from experiments that use longitudinal data. This technique bases itself both in the number of experimental units and the sample size [Castro, 1997].
Unlike the univariate profile analysis model, the multivariate profile analysis model does not require that the variance of the repeated measures or that the correlation between pairs of repeated measures remain constant along time. Nevertheless, both models require that the variances and the correlations be homogeneous in each moment in time [Vieira, 2006].
The routine techniques for analysis of variance impose the condition of independence of observations. However, this restriction generally does not apply to longitudinal data where the observations in the same individual are usually correlated. In such case, the adequate manner for treating the observations would be the multivariate form [Vonesh & Chinchilli, 1997].
Cole & Grizzle [Cole & Grizzle, 1966] use the multivariate analysis of variance according to the Smith et al. [Smith et al., 1962] formulation and comment on its versatility in the construction of specific hypothesis testing that may be obtained as particular cases of the general linear multivariate hypothesis test procedure. They assert that such hypothesis may be tested by three alternative criterions, all of which dependent on characteristic roots of matrix functions due to the hypothesis and of the matrix due to the error: criterion of the maximum characteristic root, criterion of the product of the roots (criterion of the verosimilarity ratio) and criterion of the sum of the roots. The authors illustrate the application of the multivariate analysis of variance and demonstrate that the information requested from these experiments may be formulated in terms of the following null hypotheses:

i. there are no principal effects of "measured conditions" (occasions);
ii. there are no effects of treatments;
iii. there is no interaction of treatment and occasions.

The multivariate analysis of variance is a powerful instrument to analyze longitudinal data but if the uniformity hypothesis of the variance and covariance matrix is not rejected the univariate analysis should be employed. Nonetheless, if the variance and covariance matrix of repeated measures has the serial correlation structure one should use an analysis method that takes into account the structure of this matrix in order that one might have an increment in the testing power. In this way the multivariate analysis of variance becomes the most convenient one if not the only appropriate one among the available procedures [Cole & Grizzle, 1966; Smith et al., 1962].
Lima [Lima, 1996] asserts that the multivariate profile analysis possesses as its main advantage the fact that is allows for the adoption of a very general model to represent the structure of covariances admitting that the variances of responses in each time and the covariances of responses between distinct times be different.
In studying longitudinal data investigation methods, Greenhouse & Geisser [Greenhouse & Geisser, 1959] observed that the ratios between the mean squares obtained in the analysis of variance for the mixed univariate model will only have exact distribution of probability F if the observations in time be normally distributed with equal variances and be mutually independent or equally correlated. Because these presuppositions are strict, the authors prefer considering the observations in time as a vector of samples of a normal multivariate distribution with an arbitrary variance and covariance matrix. Being so, the multivariate perspective presented by Morrison [Morrison, 1990] allows for the adoption of a general model to represent the covariance structure of the observations. In this case, the covariance

matrix is known as being non structured where all variances and covariances might be different and, as pointed out by Andreoni [Andreoni, 1989], it is only applicable when:

- there be no theoretical or empirical basis to establish any pattern for this matrix;
- there be no need to extrapolate the model beyond the occasions of the considered observations.

The quantity of parameters associated with the non structured matrix that need to be estimated is proportional to the number of conditions of evaluation. In situations where the number is large, when the number of experimental units is small in relation to the number of evaluation events or when there is the presence of many incomplete observations the efficiency of the estimators might be affected. In some cases it may be impossible to estimate the parameters of this covariance matrix [Wald, 2000].

Meredith & Stehman [Meredith & Stehman, 1991] state that the disadvantage of the multivariate analysis is the lack of power to estimate the parameters of the covariance matrix in case when t (number of measurement events or times) is large and n is small.

Stuker [Stuker, 1986] comments on the restriction of the multivariate analysis of covariance in which the number of experimental units minus the number of treatments should be greater than the number of observations taken in each experimental unit otherwise the required matrix due to error for these tests is singular.

Timm [Timm, 1980] claims that the restrictions to the application of the multivariate profile analysis occur due to the need for complete individual response profiles and to the low power of these hypothesis tests due to excessive parametering. On the other hand, except for these restrictions, the majority of the cases in longitudinal data studies, the analysis procedure of multivariate analysis of variance is the most convenient if not the only appropriate one among the available techniques.

3. Materials and methods

3.1 Data

In order to conduct the study it was created a data matrix with the following structure: $\underline{Y} = y_{ijk}$, where y_{ijk} is the observation j belonged the period i of the treatment k. To simulate growth curves composed of two treatments, seven observations over time and five repetitions, each observation of Y matrix was defined as $y_{ijk} = f_i + \varepsilon_{ijr}$ with a fixed part f_i , with i = 1,2, and where $f_1 = 46 + 88X - 57X^2$ and $f_2 = 42 + 88X + 53X^2$ and a variable portion ε_{ijk} randomly generated with normal distribution with zero mean and variance proportional to E(f_i) in which the variation coefficient remains constant in 0,05, under these conditions is imposed on the model f_1 a linear growth higher than compared to the f_2 model and both with the same regression model.

3.1.1 Data base structure

To analyze the longitudinal data, the data base was structured in the following way; the first column refers to the independent variable $X = [x_i]$ or to periods with $i = 1...p$, the second column the response variable $Y = [y_{ijk}]$, in which y_{ijk}, refers to the observation referring to the repetition j of the period i of treatment k, with $j = 1...r$., and the third column refers to the control variable $F = [f_k]$ or treatments, with $j = 1...t$.

$$\text{File.txt}=\begin{bmatrix} 1 & y_{111} & 1 \\ . & y_{121} & . \\ . & . & . \\ . & . & . \\ p & y_{prt} & t \end{bmatrix} \tag{3}$$

The following Matlab commands upload and dimension the file in addition to determining the index of the column of each variable.

```
M=load('-ascii', 'file.txt');
[n,c]=size(M);
a=input('column of the independent variable X   =');
b=input('column of the dependent variable Y   =');
aa=input('initial column of the control variable curve =');
nc=input('number of curves to be compared =');
npc=input('number of points per curve =');
```

3.2 Data analysis

Once the data base is correctly structured the first step is to adjust the best polynomial model that explains the variation of Y in function of the X periods. Towards this, the parameters of the polynomial of adjustment will be estimated by the matrix expression below.

$$\hat{Y} = BX \tag{4}$$

$$\beta = [bo \quad b_1 \quad .... \quad b_g] \tag{5}$$

in which g is a degree of the polynomial

$$X = [1 \quad x_i \quad x_i^2 . \quad . \quad . \quad x_i^g] \tag{6}$$

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{7}$$

To determine what is the best degree of the polynomial for the data under analysis it was used a scatterplot. The following commands prepare the data for visualization.

```
x=M(:,a:(b-1));
Y=M(:,b:(aa-1));
Trat=M(:,aa);
M=[x Y Trat];
[tmp,idx]=sort(M(:,aa));
M=M(idx,:),
set(plot(x,Y,'o'))
```

From the scatterplot, choose the degree of polynomial to be adjusted.

```
g=input('choose the degree of polynomial Degree   =');
```

The following procedures were used to estimate $\hat{\beta}$

```
[n,r]=size(M);
X=ones(n,npc);
```

```
y1=ones(n,1);
for i=2:npc
X(:,i)=M(:,a).*y1;
y1=X(:,i);
end
X=X(:,1:(d+1));
BT=(inv((X'*X)))*(X'*(M(:,b:aa-1)));
```

To test the hypothesis: $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$, the F test is employed.

$$F = \frac{QMr}{QM\varepsilon} \tag{8}$$

that have the Snedecor F distribution with (g-1) and (n-g) degrees of freedom.

$$QMr = \frac{1}{(g-1)}\beta(X'Y) - n\bar{Y}^2 \tag{9}$$

$$QM\varepsilon = \frac{1}{(n-g)}[(Y'Y) - \beta(X'Y)] = \frac{1}{(n-g)}\ (Y - \hat{Y})'(Y - \hat{Y}) \tag{10}$$

And to measure the degree of explanation of the variability of Y according to the polynomial model it is used the coefficient of determination.

$$R^2 = \frac{SQr}{SQT} \tag{11}$$

$$SQr = \beta(X'Y) - n\bar{Y}^2 \tag{12}$$

$$SQT = (Y'Y) - n\bar{Y}^2 \tag{13}$$

And to measure the degree of explanation of the Y variability in function of the polynomial model it is employed the determination coefficient.

$$R^2 = \frac{SQr}{SQT} \tag{14}$$

in which

$$SQr = \beta(X'Y) - n\bar{Y}^2 \tag{15}$$

and

$$SQT = (Y'Y) - n\bar{Y}^2 \tag{16}$$

After the adjustment of the polynomial model for the data set, the next step is to adjust the same model for each of the k treatments separately, so

$$\hat{Y}_k = (X'X)^{-1}(X'Y_k) \tag{17}$$

Where

$$\hat{b}_k = (X'X)^{-1}(X'Yk) \tag{18}$$

The test for comparing the curves is based on the decomposition of $SQ\varepsilon$ in one part explained by the variation between the curves and the other by the variation within the curves.

$$SQ\varepsilon = \sum_{k=1}^{t} (\hat{Y} - \hat{Y}k)'(\hat{Y} - \hat{Y}k) + \sum_{k=1}^{t} (\hat{Y}k - \dot{Y}k)'(\hat{Y}k - Yk) \tag{19}$$

in which $\sum_{K=1}^{t} (\hat{Y} - \hat{Y}k)'(\hat{Y} - \hat{Y}k)$ is the variation explained by the treatments, and $\sum_{K=1}^{t} (\hat{Y}k - \dot{Y}k)'(\hat{Y}k - Yk)$ is the variation within each treatment.

$$\text{And the } F = \frac{\sum_{k=1}^{t} ((\hat{Y} - \hat{Y}k)'(\hat{Y} - Yk))(n-t)(p-1)}{\sum_{K=1}^{t} ((\hat{Y} - Yk)'(\hat{Y} - Yk))(p-1)(t-1)} \tag{21}$$

has a Snedecor F distribution with $(p-1)(t-1)$ and $(n-t)(p-1)$ degrees of freedom.

$$\text{And the reason } F = \frac{\sum_{k=1}^{t} ((\hat{Y} - \hat{Y}k)'(\hat{Y} - Yk))(n-t)(p-1)}{\sum_{K=1}^{t} ((\hat{Y} - Yk)'(\hat{Y} - Yk))(p-1)(t-1)} \tag{22}$$

Has a Snedecor F distribution with $(p-1)(t-1)$ and $(n-t)(p-1)$ degrees of freedom.

The following commands calculate the regression parameters for the individual curves.

```
[c,r]=size(M1);
Yobs(:,i)=M1(:,b)
X=ones(c,npc);
y1=ones(c,1);
for j=2:npc
X(:,j)=M1(:,a).*y1;
y1=X(:,j);
end
Y=M1(:,b);
X=X(:,1:(d+1));
B(:,i)=(inv((X'*X)))*(X'*Y);
Y1est(:,i)=X*BT;
end
```

The following commands print the graph with the curves estimated.

```
Yest=X*B;
y=[Y1est Yest];
x=X(:,b);
plot(x,y)
```

The following commands execute analysis of variance.

```
[n,c]=size(M)
 SQmodelo=sum(sum((Y1est-Yest).*(Y1est-Yest)))
SQerro=sum(sum((Yest-Yobs).*(Yest-Yobs)))
SQtotal=sum(sum((Y1est-Yobs).*(Y1est-Yobs)))
glmodelo=(npc-1)*(nc-1)
gltotal=(n-nc)
glerro=gltotal-glmode
R=(SQmodelo/SQtotal)
F=(SQmodelo/glmodelo)/(SQerro/glerro)
p=fpdf(F,glmodelo,glerro)
```

The following commands format the ANOVA Table printout.

```
Table=zeros(3,5);
Table(:,1)=[ RSS SSE TSS]';
Table(:,2)=[df1 df2 df3]';
Table(:,3)=[ RSS/df1 SSE/df2 Inf ]';
Table(:,4)=[ F Inf Inf ]';
Table(:,5)=[ p Inf Inf ]';
colheads = ['Source       ';'        SS  ';'          df ';...
         '       MS    ';'         F  ';'     Prob>F  '];
  atab = num2cell(Table);
for i=1:size(atab,1)
  for j=1:size(atab,2)
    if (isinf(atab{i,j}))
      atab{i,j} = [];
    end
  end
end
if (nargout > 1)
  anovatab = atab
end
```

The following commands prepare the file for the multivariate analysis.

```
M=[X Yobs];
nt=M(:,2);
    x=nt(1);
    n=1;
    idx=1;
    for i=2:length(nt)
       if nt(i)==x(idx)
          n(idx)=n(idx)+1;
       else
```

```
        idx=idx+1;
        x(idx)=nt(i);
        n(idx)=1;
    end
n=cumsum(n);
      B=ones(d+1,nc);
      for i=1:length(n)
        idx=find(M(:,aa)==M(n(i),aa));
M2=M(idx,:);
      end
```

For the multivariate analysis it was employed a new structure of the data file.

File2.txt=[x_{jk} y_{1jk} . . . y_{pjk}]

File2.txt=[X Y]

In which, the first column has the values for the j repetitions for each of the k treatments, each one of the following i columns contains the values of Y for the j repetitions of the k treatments. As seen in the structure below.

$$\text{File2.txt} = \begin{bmatrix} x_{11} & y_{111} & \cdots & y_{711} \\ x_{21} & y_{121} & \cdots & y_{721} \\ x_{31} & y_{131} & \cdots & y_{731} \\ x_{41} & y_{141} & \cdots & y_{741} \\ x_{51} & y_{151} & \cdots & y_{751} \\ x_{12} & y_{112} & \cdots & y_{712} \\ x_{22} & y_{122} & \cdots & y_{722} \\ x_{32} & y_{132} & \cdots & y_{732} \\ x_{42} & y_{142} & \cdots & y_{742} \\ x_{52} & y_{152} & \cdots & y_{752} \end{bmatrix} \quad (23)$$

The following commands change the structure of the file.

```
        M=Mtemp;
    [n,c]=size(M);
 nt=M(:,1);
z=nt(1);
n=1;
idx=1;
for i=2:length(nt);
   if nt(i)==z(idx)
      n(idx)=n(idx)+1;
   else
      idx=idx+1;
      z(idx)=nt(i);
      n(idx)=1;
   end
```

```
end
 n=cumsum(n);
 for i=1:npc
     idx=find(M(:,a)==M(n(i),a));
     M1=M(idx,:);
     trat=M1(:,aa);
     M1=M1(:,b);
     Ymult(:,i)=M1;
 end
 Y=[trat Ymult];
```

For the multivariate data analysis the employed procedure was proposed by Johnson and Wishern [Johnson and Wishern, 1998] in which the standardized variable employed for the comparison of the curves is

$$\Lambda = \frac{|W|}{|W_p|} \tag{24}$$

$$\text{in which } \mathrm{W} = (\mathrm{X}'\beta - \mathrm{X}'\beta T)' * (\mathrm{X}'\beta - \mathrm{X}'\beta T) \tag{25}$$

$$\beta = \begin{bmatrix} b_{10} & b_{20} \\ b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix} = \left[b_{dk} \right] \tag{26}$$

where $\left[b_{dk} \right]$ is the polynomial coefficient of d order of the k treatment.

$$\beta = (X'S^{-1}X)^{-1}(X'S^{-1}\bar{Y}) \tag{27}$$

$$S = \frac{1}{(rt)}(Y - \bar{Y})'(Y - \bar{Y}) \tag{28}$$

$$\mathrm{W_p} = (\mathrm{Y} - \mathrm{X}'\beta\mathrm{T})' * (\mathrm{Y} - \mathrm{X}'\beta\mathrm{T}) \tag{29}$$

$$\beta T = (X'S_p^{-1}X)^{-1}(X'S_p^{-1}\bar{Y}) \tag{30}$$

$$S_p = \frac{1}{(n-t)} \sum_{K=1}^{t} (n_k - 1)S_k^2 \tag{31}$$

Where $S_k^2 = matrix.of.\mathrm{cov}\, ariance.of.the.treatment.k$

In order to test if there is a difference between curves the standardized variable is employed

$$\chi^2 = -\left(N - \frac{1}{2}(p - d - t) \right) \ln \Lambda \tag{32}$$

has a chi square distribution with (p-q-1) degrees of freedom.
The following commands run the multivariate analysis.

```
    x=x';
    [n,c]=size(Y);
    M=Y
    temp=M(:,b:c);
Y1obs=temp';
V=cov(temp);
S=inv(V);
temp=(sum(temp))./n;
Y=temp;
[n,c]=size(V);
    X=ones(n,c);
    y1=ones(n,1);
    for i=2:c
       X(:,i)=x.*y1;
       y1=X(:,i);
    end
    d=input('choose the polynomial degree =');
    X=X(:,1:(d+1));
    BT=(inv(X'*S*X))*(X'*S*Y');
    Y1est=X*BT;
    plot(x,Y1est)
    [n,c]=size(Y1obs);
    temp=ones(n,c);
    for i=1:c
       temp(:,i)=Y1est;
    end
    Y1est=temp;
    Temp=Y1obs-Y1est;
    W=Temp*Temp';
```

The following commands run the analysis of individual curves.

```
    nt=M(:,a);
   z=nt(a);
  n=1;
   idx=1;
   for i=2:length(nt)
      if nt(i)==z(idx)
         n(idx)=n(idx)+1;
      else
         idx=idx+1;
         z(idx)=nt(i);
         n(idx)=1;
      end
   end
  k=n;
  n=cumsum(n);
```

```
B=zeros(d+1,length(n));
v=zeros(length(W));
sp=zeros(length(S));
for i=1:length(n)
      idx=find(M(:,a)==M(n(i),a));
      M1=M(idx,:);
      [r,c]=size(M1);
      temp=M1(:,b:c);
      Yobs=temp';
      V=cov(temp);
      V=(k(i)-1)*V;
      temp=(sum(temp))./r;
      Y(i,:)=temp;
      temp=(V+v)/(n(i)-2);
      v=temp;
      S=inv(v);
      B=(inv(X'*S*X))*(X'*S*Y');
end
Temp=zeros(npc);
 for i=1:length(n)
      idx=find(M(:,1)==M(n(i),1));
      M1=M(idx,:);
      [r,c]=size(M1);
      temp=M1(:,b:c);
      Yobs=temp';
      temp=zeros(npc,r);
      Yest=X*B(:,z(i));
      for j=1:5
          temp(:,j)=Yest;
       end
       temp=((Yobs-temp)*(Yobs-temp)');
       Temp=temp+Temp;
       Wp=Temp;
 end
       Wilks=((det(Wp))/(det(W)))
      Qsquare=-(N-0.5*(npc-nc-2+d))*log(Wilks)
      df=(npc-nc-1)*d
      chi2pdf(Qsquare,df)
```

4. Results

The following parameters must be furnished when running the program:
independent variable column X =1
dependent variable column Y =2
initial column of the control variable curve =3
number of curves to be compared =2
number of points per curve =7

The following graph is generated in order to choose the degree of the polynomial to be adjusted.

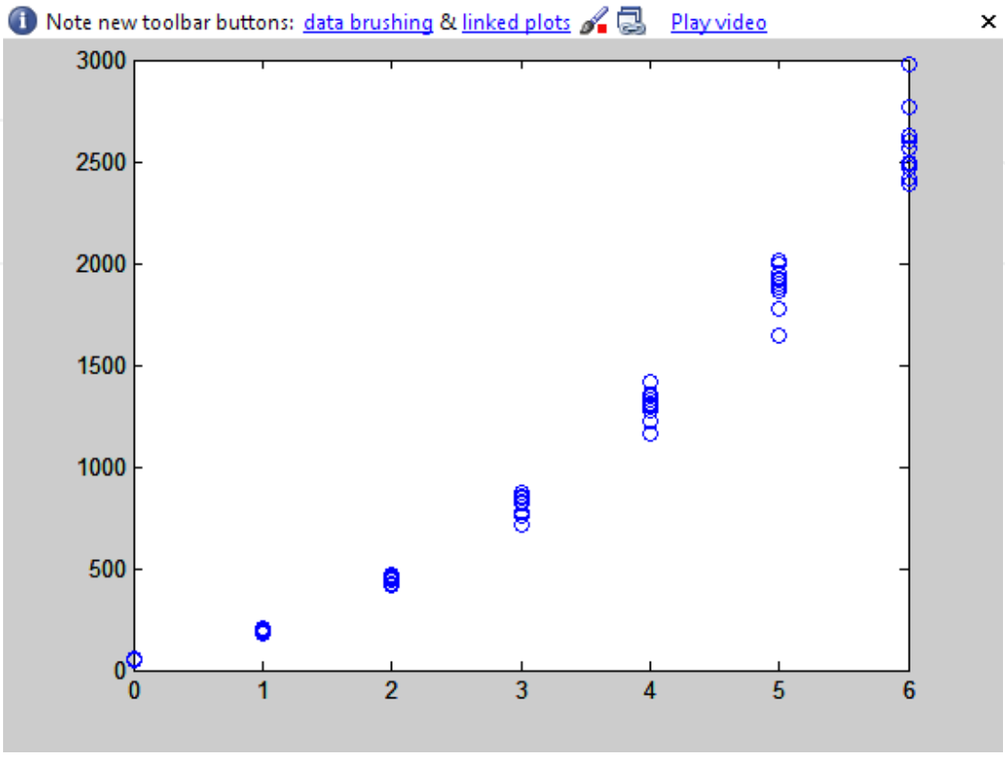

Fig. 1. Scatterplot of the data.

A second degree polynomial was chosen to model the data according to the scatterplot above.
The estimated coefficients for the second order polynomial were:

$$\hat{Y}_T = 48.464 + 81965X + 56.806X^2 \,]$$

with (P<0000,1)

$$R^2 = 0.3405$$

The analysis of variance of the complete polynomial model is presented in table 1.

Causes of variation	DF	SS	SQ	F	P
Polynomial	2	5.2799E+007	2.6384E+007	3.8235E+003	1.2054E-072
Error	68	4.6924E+005	6.9006E+003		
Total	69	5.3238E+007			

Table 1. ANOVA for polynomial model

The output of the program has the following format
anovatab1 =

[5.2769e+007] [2] [2.6384e+007] [3.8235e+003] [1.2054e-072]
[4.6924e+005] [68] [6.9006e+003] [] []
[5.3238e+007] [69] [] [] []

After the choice of the polynomial model and its test of significance, the same model was applied on each one of the treatments separately, the results are as follows:

$$\hat{Y}_1 = 50.31 + 89.10X + 57.10X^2$$

$$\hat{Y}_2 = 46.90 + 76.82X + 56.55X^2$$

$$R^2 = 0.3405$$

The analysis of variance for decomposition of error was employed to test the difference between the curves and is presented in table 2.

Causes of variation	DF	SS	SQ	F	P
Polynomial	6	1.5976E+005	2.6626E+004	5.4202	2.52652E-004
Error	63	3.0948E+005	4.912E+003		
Total	69	4.6924E+005			

Table 2. Analysis of variance to compare the curves.

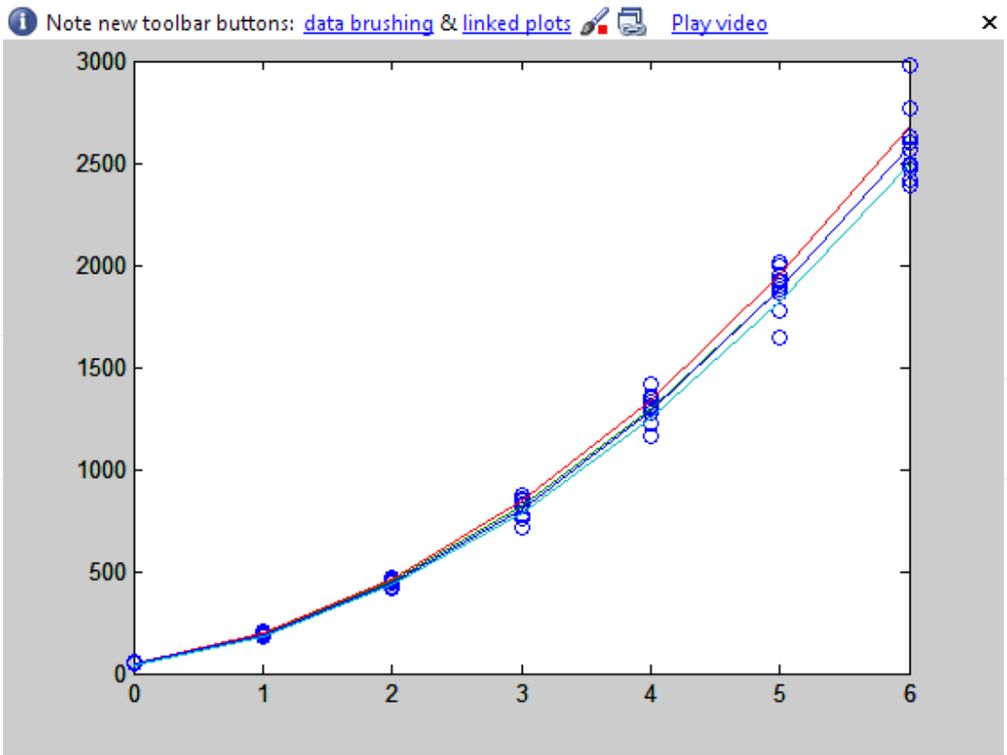

Fig. 2. Central line resulting from the estimated polynomial for the entire data set and the external lines, one for each treatment.

The output of the program has the following format:

```
B =   50.3069   46.9023
      89.0928   76.8230
      57.0744   56.5484
```

After the choice of the polynomial model and its test of significance, the same model was applied on each one of the treatments separately, the results are as follows:

```
anovatab =
   [1.5976e+005]   [ 6]   [2.6626e+004]   [5.4202]   [2.5262e-004]
   [3.0948e+005]   [63]   [4.9124e+003]       []         []
   [4.6924e+005]   [69]         []            []         []
```

The graph below presents a central line resulting from the estimated polynomial for the whole data set and the external lines are one for each treatment.

For the multivariate test it was calculated the standardized variable

$$\chi^2 = -\left(N - \frac{1}{2}(p - d - t)\right)\ln\Lambda = 192.0591 \text{ with } (P<0{,}001).$$

The program outputs were as follows.

```
Wilks =    0.0581
Chi square =  192.0591
df =     8
ans =  1.4552e-037
```

5. Conclusion

Given its matrix structure, Matlab presented itself as an efficient tool for linear models. The programs and the methodology presented were efficient to the comparing of polynomial growth curves. The modular sequence in which the programs were developed allows the user to implement new routines as well as new methodology proposals for the solution of the proposed problem. The solutions presented for the problem of comparison of polynomial growth curves may be used in part or in conjunction for the solution of other linear models problems.

6. References

Andrade, D. F. & Singer, J. M. (1986). Análise de dados longitudinais. *Proceedings of VII Simpósio Nacional de Probabilidade e Estatística*, pp. 19-26, Campinas, SP, Brazil, 1986.

Andreoni, S. (1989). Modelos de efeitos aleatórios para a análise de dados longitudinais não balanceados em relação ao tempo. Dissertation (MS. in Statistics) Institute of Mathematics, São Paulo University, São Paulo, Brazil, 1989.

Box, G. E. P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *Annals of Mathematical Statistics*, Vol.25, No. 2, 1954, DOI: 10.1214, pp. 290-302.

Castro, S. M. J. (1997). A metodologia de análise de dados longitudinais. Thesis (BS. in Statistics), Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Castro, S. M. J. & Riboldi, J. (2005). A construção do modelo em dados longitudinais: escolha dos efeitos fixos e aleatórios, modelagem das estruturas de covariância. *Proceedings of the Annual Meeting of the Brazilian Region of the International Biometrics Society, and of the Simpósio de Estatística Aplicada a Experimentação Agronômica*, pp. 157-158, Londrina, Paraná, Brazil, 2005.

Cole, J. W. L. & Grizzle, J. E. (1966). Applications of Multivariate Analysis of Variance to Repeated Measurements Experiments. *Biometrics*, Vol. 22, 1966, ISSN 0006341x, pp. 810 – 828.

Fitzmaurice, G. et al. Ed(s).). (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC Taylor & Francis Group, ISBN 978-1-58488-658-7, Boca Raton, Florida.

Freitas, G. E. (2007). Análise de dados longitudinais em experimentos com cana-de-açúcar. Dissertation (MS. in Agronomics), Escola Superior de Agricultura "Luis de Queiroz", São Paulo University, Piracicaba, Brazil, 2007.

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, Vol.24, No. 2, June 1959, DOI: 10.1007, pp.95-112.

Huynh, H. & Feldt, L.S. (1970) Condition under which mean square ratios in repeated measurements designs have exact F-distributions. *J. Am. Stat. Assoc.*, Vol.72, 1970, ISSN 0162-1459, pp.320-340.

Huynh, H. & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J. Educational and Behavioral Statistics*, Vol.1, No.1, March 1976, DOI: 10.3102/10769986001001069, pp.69-82.

Johnson, R. A. & Wichern, D. W. (1998). *Applied multivariate statistical analysis*, 4 ed. Prentice Hall, ISBN 0-13-834194-x, Upper Saddle River, New Jersey.

Khattree, R. & Naik, D. N. (2000). Multivariate data reduction and discrimination with SAS software. SAS Institute Inc., ISBN 1-58025-696-1, Cary, North Carolina, USA.

Lima, C. G. (1996). Análise de dados longitudinais provenientes de experimento em blocos casualizados. Dissertation (PhD in Agronomics), Escola Superior de Agricultura "Luiz de Queiroz", São Paulo University, Piracicaba, Brazil, 1996.

Malheiros, E.B. (1999). Precisão da análise de dados longitudinais, com diferentes estruturas para a matriz de variâncias e covariância, quando se utiliza o esquema em parcelas subdivididas. *Revista de Matemática e Estatística UNESP*, Vol.17, 1999, ISSN 0102-0811, pp.263-273.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *An. Math. Stat.*, Vol.11, No. 2, June 1940,ISSN 00034851, pp.204-209.

Meredith, M.P.; Stehman, S.V. (1991). Repeated measures experiments in forestry: focus on analysis of response curves. *Canadian Journal of Forest Research*, Vol.21, 1991, ISSN 0045-5067, pp.957-965.

Smith, H. R. et al. (1962). Multivariate Analysis of Variance. (MANOVA). *Biometrics*, Vol.2, 1962, ISSN 0006-341X, pp. 61 – 67.

Morrison, D. F. (1990). *Multivariate Statistical Methods*. 3 ed, McGraw-Hill, ISBN 0-07-043187-6, Singapore.

Stuker, H. (1986). Análise multivariada para dados onde a característica observada é subdividida em K classes. Dissertation (MS in Agronomics), Escola Superior de Agricultura "Luiz de Queiroz", São Paulo University, Piracicaba, 1986.

Timm, N.H. (1980). Multivariate analysis of variance of repeated measurements. In: *Handbook of Statistics Analysis of Variance*, Vol. 1, P. R. Krishnaiah (Ed), pp.41-87, North-Holland, ISBN:0444853359, New York.

Vieira, F. T. P. A. (2006). Uma Abordagem Multivariada em Experimento Silvipastoril com Leucaena leucocephala (Lam.) de Wit. no Agreste de Pernambuco. Dissertation (MS. in Biometrics), Rural Federal University of Pernambuco, Pernambuco, Brazil, 2006.

Vonesh, F. E. & Chinchilli, V.M. (1997). *Linear and nonlinear models for the analysis of repeated measurements.* Marcel Dekker, ISBN 0-8247-8248-8, New York.

Wald, V. B. (2000). A metodologia de modelos mistos não lineares aplicados à análise de dados longitudinais em plantas forrageiras. Dissertation (MS in Zootechnique), Federal University of Rio Grande do Sul, Brazil, 2000.

Xavier, L. H. (2000). Modelos univariado e multivariado para análise de medidas repetidas e verificação da acurácia do modelo univariado por meio de simulação. Dissertation (MS in Statistics and Agronomics experimentation), Escola Superior de Agricultura "Luiz de Queiroz", São Paulo University, Piracicaba, Brazil, 2000.

MATLAB - A Ubiquitous Tool for the Practical Engineer

Edited by Prof. Clara Ionescu

ISBN 978-953-307-907-3
Hard cover, 564 pages
Publisher InTech
Published online 13, October, 2011
Published in print edition October, 2011

A well-known statement says that the PID controller is the â€œbread and butterâ€ of the control engineer. This is indeed true, from a scientific standpoint. However, nowadays, in the era of computer science, when the paper and pencil have been replaced by the keyboard and the display of computers, one may equally say that MATLAB is the â€œbreadâ€ in the above statement. MATLAB has became a de facto tool for the modern system engineer. This book is written for both engineering students, as well as for practicing engineers. The wide range of applications in which MATLAB is the working framework, shows that it is a powerful, comprehensive and easy-to-use environment for performing technical computations. The book includes various excellent applications in which MATLAB is employed: from pure algebraic computations to data acquisition in real-life experiments, from control strategies to image processing algorithms, from graphical user interface design for educational purposes to Simulink embedded systems.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

João Eduardo da Silva Pereira, Janete Pereira Amador and Angela Pellegrin Ansuj (2011). Comparison of Methodologies for Analysis of Longitudinal Data Using MATLAB, MATLAB - A Ubiquitous Tool for the Practical Engineer, Prof. Clara Ionescu (Ed.), ISBN: 978-953-307-907-3, InTech, Available from: http://www.intechopen.com/books/matlab-a-ubiquitous-tool-for-the-practical-engineer/comparison-of-methodologies-for-analysis-of-longitudinal-data-using-matlab

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen