

ViSta

The Visual Statistics System

Missing Data Analysis

¹Pedro Valero & ²Forrest Young

¹DEPARTAMENTO DE METODOLOGÍA-UNIVERSITAT DE VALENCIA

²THE L.L. THURSTONE PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA

CB 3270 DAVIE HALL, CHAPEL HILL N.C., USA 27599-3270

VISUAL STATISTICS PROJECT
WWW.VISUALSTATS.ORG
REPORT NUMBER 2000-4
MARCH 2000

Missing Data Imputation

Pedro Valero and Forrest Young

This chapter presents ViSta-MissDat, the module for Missing Data Imputation in ViSta. This procedure is capable of estimating parameters of data with missing values, as well as the missing values themselves. The visualization for ViSta-MissDat includes a parallel boxplot comparing the correlations between variables using Listwise, Pairwise or Maximum Likelihood methods; a scatterplot showing the bivariate relationship between pairs of variables that incorporates the estimations for the missing values; a normal probability plot of a measure for multivariate outliers; and a parallel boxplot for the data including the estimations for missing values. The report includes univariate statistics for the available cases, Listwise cases and for the data after estimation of missing values. ViSta-MissDat also allows for the creation of imputed data that can be submitted for further analysis under certain conditions.

1 Introduction

Missing values are values that the researcher had planned to collect but that for reasons not under his/her control, they were not observed. For example, some people may answer some of the questions in a survey but refuse to answer others. In other occasions, some samples can be destroyed for mistakes in the manipulation process or they are simply not available due to ethical, organizational or practical issues. All these causes will produce a matrix of data that contains “holes” in some cells as in table 1.

Missing values introduce complexities to data analysis that should be dealt with before applying standard statistical techniques. These complexities are of two kinds:

1. **Computational problems.** Most of the standard analysis techniques like regression analysis or principal components assume as starting point a multivariate data matrix that is reduced to a vector of means $\bar{y}=(\bar{y}_1, \dots, \bar{y}_k)$ and a covariances matrix $S=(s_{jk})$. If some of the values in the data matrix are missing, computing the means and the covariances becomes a somewhat convoluted problem. This problem has been managed traditionally by quick methods like casewise deletion of data or formulae that use only the available cases. These methods can result mistaken in many cases so better alternatives have been developed in the last years. In particular, the EM algorithm (Little and Rubin, 1987) provides maximum likelihood estimates for the means and the covariances and is in general more recommendable than any of the quick methods.
2. **Bias due to the missingness pattern.** The pattern of the missing values can be an important problem, bigger even than the amount of incomplete data. Of course, the greater the amount of incomplete data, the greater the loss of information, and hence the lower the statistical precision. But, if the points are scattered randomly along the data matrix, almost any method dealing with missing values will produce similar results. However, when the missing data depend on the values of the variables included in the analysis, or, even worse, on unknown variables not even measured, the statistical results can be unacceptable.

Id	Y1	Y2	Y3	Y4	Y5
1	1	5	3	8	9
2	5	X	7	X	8
3	4	5	6	7	9
4	5	2	2	7	5
5	9	X	7	6	5
6	4	5	6	8	7
7	6	5	X	3	X
8	6	7	8	9	3
9	4	X	5	6	X
10	8	5	6	7	8

TABLE 1: A example table with missing values

2 Example

We will discuss an example based on demographic data of countries in the year 1995. The variables `calories`, `litmale` and `litfem` have many missing-values. These data are appropriate for missing data analysis. The variables are:

```

logpop      LOG OF POPULATION x 1000
log-densit  LOG OF INHABITANTS PER SQUARE KM
litmale     PROPORTION OF LITERATE MALE (transformed using
folded-power, p=0.34)
litfem      PROPORTION OF LITERATE FEMALE (transformed
using folded-power, p=0.33)
log-aids    LOG OF CASES OF AIDS PER 100.000 INHABITANTS
explfem     LIFE EXPECTANCY FEMALE (YEARS)
explmale    LIFE EXPECTANCY OF MALE (YEARS)
mortchil    INFANT DEATH RATE (CHILDREN DYING PER ONE THOU-
SAND BIRTHS)
calories    DAILY CALORIES INTAKE
loggdp      LOG OFGROSS NATIONAL PRODUCT
Birtdeat    LOG OF BIRTHS DIVIDED BY DEATHS
fertili     LOG OF AVERAGE NUMBER OF CHILDREN PER WOMAN

```

Notice that many of these variables have been transformed previously to analysis. Exploration of the original dataset revealed many variables with non-asymmetric distributions. However, logarithms of these variables had a more satisfactory shape except in the case of the two variables that originally were proportions. The visual folded power transformation in ViSta was used for modifying these variables (see section about transformations in this guide) and the values that resulted in symmetric histograms were chosen. The methods for imputation of missing values discussed in this chapter assume a normal multivariate distribution so these sort of manipulations can be necessary when data do not seem to satisfy this assumption.

The Visualize Data command in the menu Data can be used to explore a dataset with missing values. Notice that non-numerical variables are not considered in the visualization. This dataset did not include any categorical variable but the dummy variable transformation can provide a convenient way of transforming these types of variables as if they were numerical. This is explained in the section about transformations. Therefore, the visualization shows the data after the aforementioned transformations have been carried out. We did a first visualization in order to determine the variables to be transformed, carried them out and then re-ran the visualization to obtain the image in figure 1. We decided to ignore the asymmetry of variables `explfem`, `explmale` and `mortchil`.

This visualization includes four plots and a list of observations. Both the histogram and the scatterplot matrix are designed to examine the univariate and bivariate distribution of a dataset with missing values and they work very much like as standard plots. Hence, looking at the figure 1 we can observe in the histogram that the variable `logbirtdeat` has approximately a sym-

Missing Data Imputation

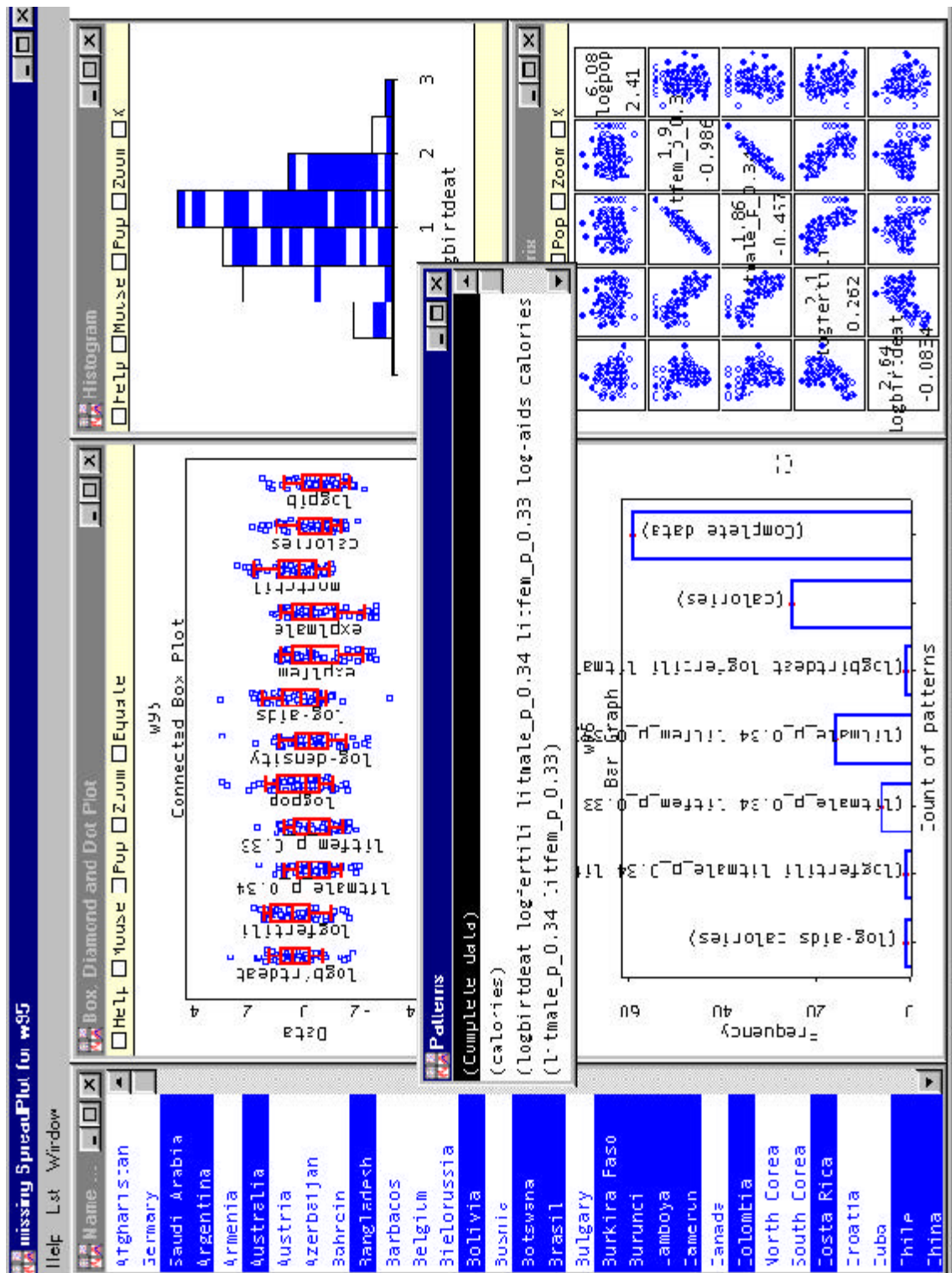


Figure 1: Visualization for numerical missing data

metric distribution and that the five variables in the scatterplot matrix do not appear as being non-linearly related. Also, heteroskedastic relationships do not show up in this plot.

Notice that the scatterplot matrix has an X button in the plot overlay. This button leads to a dialog box that allows for changes in the variables shown in this plot. This is necessary because the scatterplot matrix will only display five variables simultaneously so it may be necessary to exchange variables in and out so as to examine the whole data-set.

The histogram and the scatterplot matrix for data-sets that have missing values present peculiarities that will be commented. These peculiarities stem from the variability in the number of cases that are included into them. Hence, the histogram contains the number of cases available for the variable currently represented and the scatterplot-matrix only contains the number of cases for the variables currently represented. This can be a source of confusion because even though the histogram, the scatterplot matrix and the observations list are linked, highlighting or changing a given point does not automatically results in the corresponding effect in the other plots, as this particular observation can be missing in them. For example, Taiwan is missing in the variable logbirtdeat so selecting it in the observations list will not result in any point selected in the histogram for that variable.

The other windows in figure 1 correspond to information that is more specifically linked to the missing data problem. In particular, they show information about the different patterns of missing values in the data. A pattern of missing values is a list of the variables not observed in a case. The dialog box in figure 1 labeled Patterns shows all the missing patterns in our data. For example, the second line reads (Calories). This means that there are at least one and possibly several cases that have missing values for only the variable Calories. A missing pattern can have many variables, like the third, which includes six variables. A variable can be in more than one missing pattern, like the variable calories, which is in the second and the third patterns. Finally, there will often be some cases without missing values. This pattern is labeled Complete data and is the first in the list of patterns in Figure 1.

The barchart of figure 1 informs about the number of cases for each of the patterns. Looking at this figure we can say that the pattern with more cases corresponds to complete data, followed by the pattern with the variable Calories missing. We can observe that the pattern where the variables Litmale and Litfem are missing is also important, and that those two variables are quite related in this aspect. Another pattern of importance corresponds to the variables Calories, Litmale and Litfem which are missing simultaneously.

It is convenient to discuss at this point the actions that are available to the researcher in this situation and the possible consequences of them. These are the following:

1. Omit the variables with many missing values. As variables with many missing values are a challenge for standard statistical techniques and introduce new complexities that the researcher may not want to face, it is possible to avoid the problem just omitting the offending variables. This option, however, can not be acceptable when the question of scientific importance involves the variables with missing values. If, for example, the researcher wants to study the proportion of literate individuals in each country as a

function of other economical and demographical variables, omitting these variables is not possible. On the other hand, even though this were possible in some research scenarios, making an effort towards understanding the missingness patterns may still be worth the effort.

2. Omit the cases with missing values. This method receives the name of complete or listwise data methods. ViSta provides the following way of computing a listwise dataset from a dataset with missing values. Choose create data from the Data menu. This will provide an option of keeping or removing the cases with missing values. The option of removing will result in a dataset complete, without missing values. This method is not appropriate in many cases due to two reasons: 1) cases with missing values may be related with observed variables and deleting it throws away this information, resulting in biased results, and 2) even if the cases are not related to observed variables, deleting them means losing statistical precision due to the reduction in sample size.
3. Obtain the summary statistics after controlling for missing values. This option is available in ViSta in the menu Data->Impute missing data. This command uses the Estimation-Maximization (EM) algorithm to compute maximum likelihood estimates of the summary statistics of the data. Furthermore, this algorithm derives estimates of the imputed values that can be used for exploratory purposes. Other methods based on this algorithm provide a way to use the imputed values for hypothesis testing under the uncertainty underlying the situation where some information is missing.

Omitting the cases with missing values with the data about countries in the world results in a dataset with only 59 observations while the original had 109. Many statistical packages remove data in this way internally, before computing the requested statistical analysis. Consequently, the analysis will be based on less cases than originally available, where the proportion of deleted cases depends on the proportion of missing values by variable and how these values are spread along the data matrix. This will produce lower statistical precision and wider intervals of confidence than originally expected. Also, when several models are tested, which include different variables, direct comparison between them can be difficult because they will be based on sometimes quite dissimilar sets of observations. However, this is not the worse case scenario that can arise, but a description of the so-called missing data mechanisms is necessary before describing it.

A missing data mechanism is a description of the hypothetical mechanism that has produced the missing values in the dataset. Little and Rubin (1987) propose the following classification:

1. Missing Completely at Random (MCAR). This case the missing values are a random subsample of the data. This case implies that the probability of response is independent of the values of the variables. Therefore, summary statistics like means or variances should be the same for the data observed and the non-observed. Likewise, the matrix of variances-covariances for the whole data-matrix should be equal for the complete and the incomplete data.
2. Missing at Random (MAR). This condition is less restrictive and it assumes that the probability of response on a variable Y with missing values is related to other variables X, but not with itself. Despite its name, MAR does not mean that missing values

are a simple random subsample of all values. MAR is less restrictive than MCAR because it requires only that the missing values behave like a random sample of all values *within subclassess defined by observed data* (Schaffer, 1997). In this case, incomplete data has different summary statistics than complete data but it is assumed that the observed data possesses enough information to recover the lost information, helping to estimate the correct summary statistics. MAR and MCAR are said to be *ignorable* missing data mechanisms.

3. Non Missing at Random (NMAR). Sometimes, the researcher will have additional information on the data that leads him to be suspicious of data being MAR or MCAR. In this case, it is supposed that there is a non-measured variable that could be related to the missing values. This mechanism is called a *non-ignorable* missing data mechanism. This situation is particularly problematic because there is no test that you can use to evaluate whether the missing data mechanism is random or not. Most of the literature concerning missing data is based on the assumptions of MAR or MCAR. However, it is important to mention that ignorability can be regarded as relative. Sometimes, a missing data mechanism may not be known by the researcher, but there are variables that can explain the missingness to a great extent. Including these variables in the analysis will make the assumption of MAR much more plausible (Graham et al., 1994; Schaffer, 1997).

Once the different mechanisms for missing data that a researcher might hypothesize have been discussed, it should be clear to the reader that the option of omitting cases with missing values will only be acceptable if the missing data mechanism is MCAR. If data are MAR but not

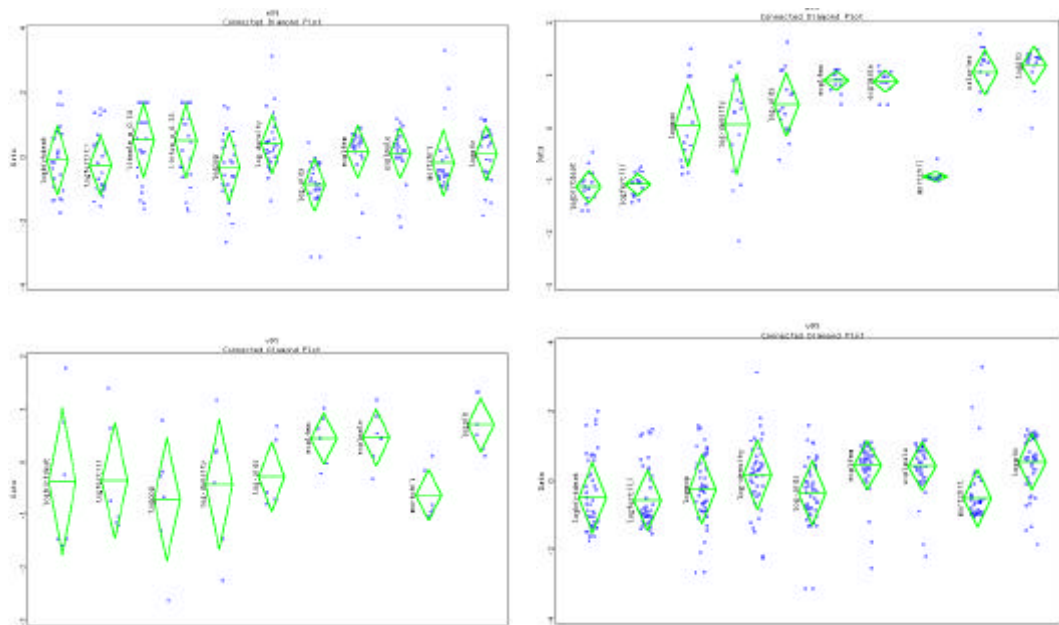


Figure 2: Boxplots of patterns of missing data in the example about countries. They are from left to right and up to down, a) pattern for the variable Calories missing, b) pattern for the variables Litfem and Litmale missing, c) pattern for the variables Calories, Litfem and Litmale missing simultaneously, d) cases in the three previous patterns.

MCAR, this option will remove the information that would allow estimating the right summary statistics, producing biased results.

The boxplots in figure 1 are designed to explore the plausibility of the MCAR assumption. If this assumption is valid, means and standard deviations of a variable should be approximately equal under the different patterns of missing data. Selecting a pattern in the Patterns window, plots the data corresponding to the observed variables for this pattern (i.e. the other variables in the dataset that are *not* in the list of variables for the pattern). Variables have been normalized previously to plotting so they all have mean equal to zero and standard deviation equal to one. Therefore, missing data patterns that stand against the MCAR assumption will reveal because the means and standard deviations of one or several boxplots will deviate from zero and one respectively. The figure 2 shows the diamond plots for several patterns. Figure 2-a shows the pattern for the variable Calories missing. Figure 2-b shows the pattern for variables Litfem and Litmale missing. Figure 2-c shows the data for the pattern with Calories, Litfem and Litmale missing. Finally, Figure 2-d shows the data for the previous three patterns together. This is obtained by clicking with the control key in the Patterns window and select-

calories									
Descriptives	Mean	Median	Std	Skew	Kurt	Min	Max	Missing	N
logbirtdeat									
MISSING	0.90	0.88	0.67	0.29	-0.93	-0.08	2.23	1.00	33.00
OBSERVED	1.01	1.06	0.60	0.06	-0.57	-0.08	2.64	0.00	75.00
logfertili									
MISSING	1.00	0.76	0.50	0.68	-0.88	0.34	1.93	2.00	32.00
OBSERVED	1.19	1.16	0.54	0.01	-1.38	0.26	2.10	0.00	75.00
litmale									
MISSING	1.16	1.47	0.72	-0.67	-0.92	-0.22	1.86	8.00	26.00
OBSERVED	0.62	0.67	0.52	-0.16	-0.46	-0.46	1.86	16.00	59.00
litfem									
MISSING	0.99	1.37	0.94	-0.65	-0.94	-0.82	1.90	8.00	26.00
OBSERVED	0.31	0.37	0.66	-0.01	-0.73	-0.99	1.90	16.00	59.00
logpop									
MISSING	3.88	3.91	0.67	-0.30	0.05	2.41	5.17	0.00	34.00
OBSERVED	4.22	4.03	0.62	0.68	0.35	3.11	6.08	0.00	75.00
log-density									
MISSING	4.54	4.45	1.43	0.18	1.78	0.92	8.61	0.00	34.00
OBSERVED	3.91	3.89	1.40	0.01	0.66	0.83	8.40	0.00	75.00
log-aids									
MISSING	14.02	14.10	4.83	-1.29	3.27	0.00	22.42	3.00	31.00
OBSERVED	20.08	19.77	5.47	-0.18	1.90	0.00	36.43	0.00	75.00
explfem									
MISSING	73.12	75.00	7.79	-2.43	6.44	44.00	81.00	0.00	34.00
OBSERVED	68.81	72.00	11.41	-0.77	-0.54	43.00	82.00	0.00	75.00
explmale									
MISSING	67.21	68.00	6.66	-1.79	4.33	45.00	76.00	0.00	34.00
OBSERVED	63.88	66.00	10.11	-0.83	-0.36	41.00	76.00	0.00	75.00
mortchil									
MISSING	31.93	21.60	34.93	2.64	7.49	4.00	168.00	0.00	34.00
OBSERVED	47.02	39.30	38.73	0.62	-0.82	4.40	137.00	0.00	75.00
loggdp									
MISSING	3.58	3.72	0.48	-0.96	0.57	2.31	4.25	0.00	34.00
OBSERVED	3.35	3.33	0.66	0.02	-1.23	2.09	4.37	0.00	75.00

Figure 3: Descriptive statistics comparing the values of several descriptive statistics for the missing and observed values of each variable with missing values.

ing several patterns simultaneously. The other patterns of missing data were ignored because of their smaller size.

Taking all the plots together it may be said that the missingness in variables *Litfem* and *Litmale* is clearly associated with other variables observed in the dataset. Figure 2-b shows that when this variables are missing, values in observed variables deviate considerably of having mean zero and standard deviation 1. A similar effect, but not so extreme is seen for variable *Calories*. In conclusion, we may affirm that the MCAR assumption does not hold for this dataset. Therefore, omitting observations with missing values (i.e. listwise deletion of cases) will produce biased data and should be avoided.

ViSta also offers a numerical summary of a dataset with missing values. This summary includes statistics comparing available data for each of the variables against the data left after listwise deletion. If data are MCAR, means, standard deviations and the rest of descriptive statistics should be the same for both conditions. The report also offers the listwise correlations and the pairwise correlations. Gross deviations among them would be also indicative of non-MCAR, but maybe MAR, data. Figure 3 shows a third summary of data with missing values offered by ViSta. This summary compares the usual descriptive statistics of the variables in the dataset for the observed and missing parts of each variable with missing values. The variable *Calories* again results quite illustrative.

The EM algorithm (Little and Rubin, 1987; Schafer, 1997; McLachlan and Krishnan, 1997) provides a method to deal with data that are MAR and is consequently the best option available. In fact, even in the case that data are MCAR, this algorithm would still be better than listwise deletion in presence of missing values.

2.1 Visualization.

Figure 4 shows the visualization for the missing data model. This visualization includes several plots that help to compare the result of the EM algorithm with other methods.

There are six windows in the visualization for missing data.

The window labeled **Variables-paired** lists the variables in the analyzed dataset *by pairs*. There is a pair of variables selected in this list that corresponds to the variables *litfem* and *logbirtdeat*. Clicking an item in this list selects a point in the window titled **Correlations**. The boxplots in this window correspond to the correlations between pairs of variables computed using three methods. The boxplot on the right are listwise correlations. The one placed in the middle are pairwise correlations. The boxplot on the left are the maximum likelihood correlations computed using the EM algorithm.

The scatterplot behind the Variables-paired window plots the pair of variables *litfem* and *logbirtdeat*. This plot includes several enhancements related with the missing values. Points in RED means that the points have been estimated using the Maximum Likelihood estimations of the matrix of parameters. Points in GREEN are points that are available but they would not

Missing Data Imputation

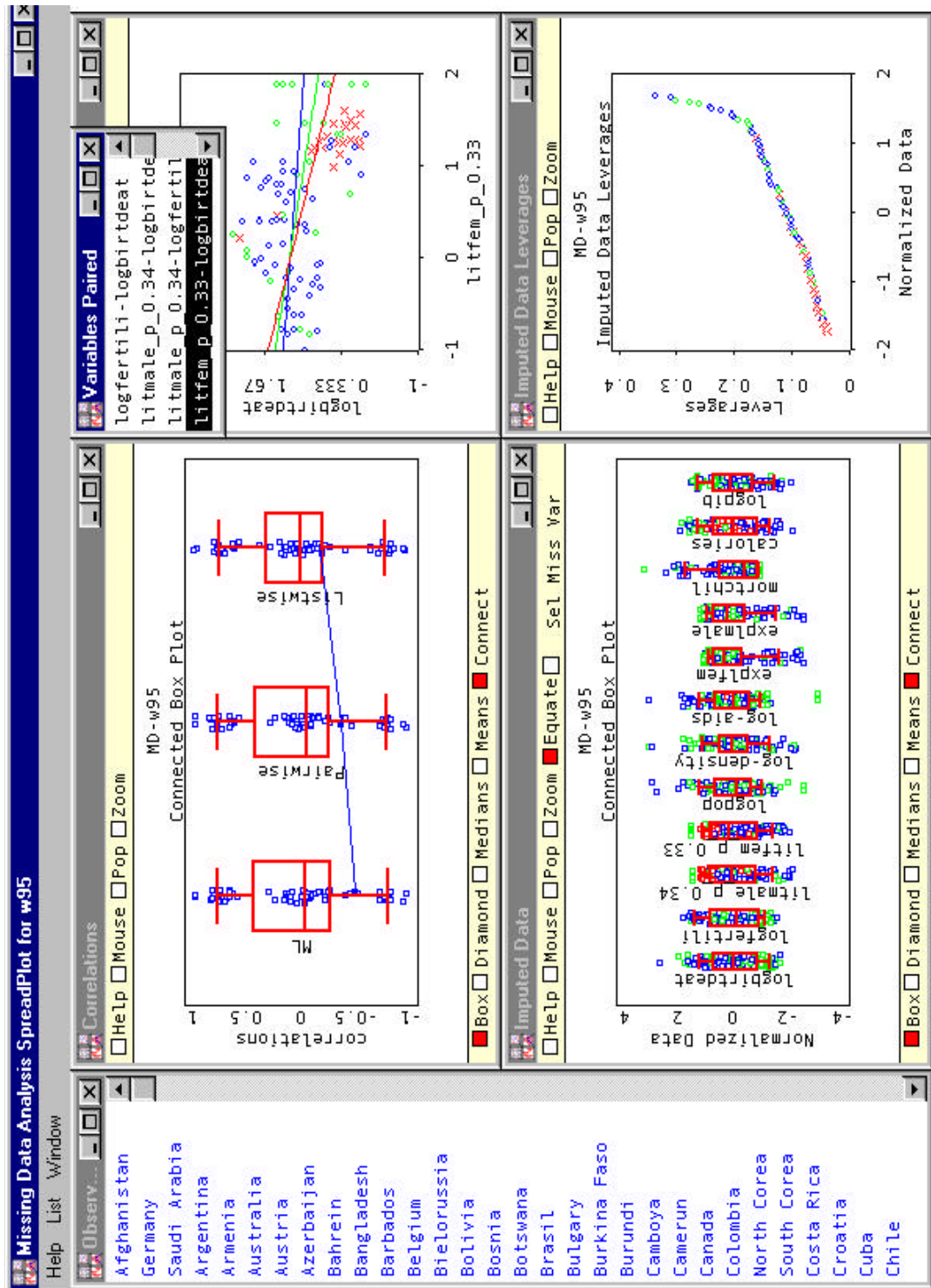


Figure 4: Spreadplot for missing data analysis

Missing Data Imputation

if listwise method is used. Points in BLUE are listwise or complete data. Points in RED can have three different symbols. An X indicates that the point is missing in the X axis; a CROSS corresponds to a point not observed in the Y axis; finally, a DIAMOND refers to a point that is missing in both axis.

This plot also has three lines on it. They are Least Squares regression lines and the color has a meaning similar to the color of the points. The black line corresponds to the black points and is, therefore, a listwise regression. The green line includes black points and green points and is a pairwise line. Finally, the red line fits all the points in the plot, including imputed data.

The boxplot titled **Imputed Data** shows the points for the variables in the model. The points are connected and have the same coding scheme as the scatterplot for pairs of variables.

The last plot on the spreadplot is called **Imputed Data Leverages** and shows the leverages for data after imputation. The leverages is a diagnostic tool that is explained in the chapter about regression. The leverage value is a measure of how an observation deviates from the centroid or mean of all the variable. Leverages are between 0 and 1 and bigger values indicate that the observation is farther from this centroid. This means that this point is an outlier. Observations that are outliers and have missing values might be examined individually because they might unduly influence the analysis.

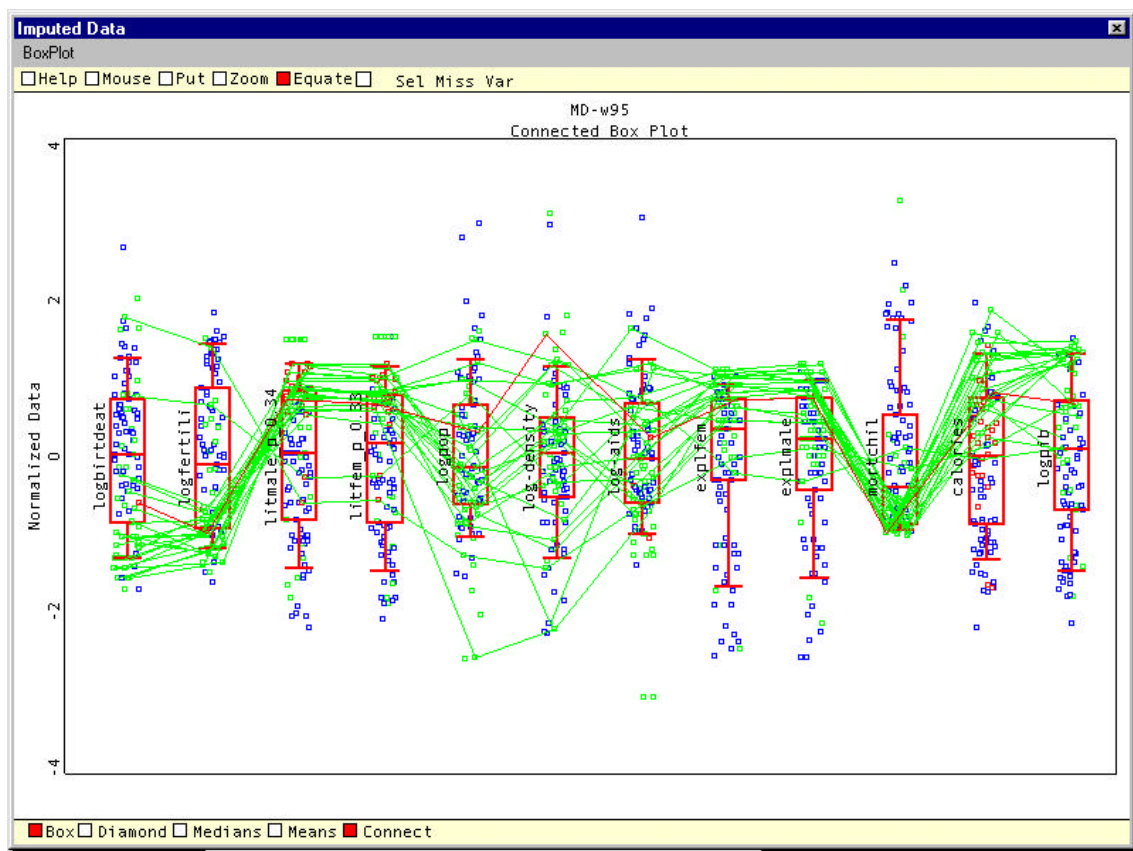


Figure 5: Parallel boxplot of imputed missing values

This spreadplot helps to understand the consequences of using different methods for dealing with missing values in a dataset. The boxplot for **Correlations** helps to find those that are different depending on the method used to compute it. For example, the correlation for `litfem` and `logbirtdeat` is quite different depending on the method used to compute it. The scatterplot for this pair of variables shows that there are many observations for the variable `litfem` and that the imputed values for these observations concentrate on one side of the plot. This results in regression lines clearly separated. This is an evidence of non completely random missing data mechanism. In this case, EM estimations of parameters should be used instead of listwise estimations because there is evidence of bias.

The parallel boxplot for imputed data is shown in the figure 5. The `Sel Miss Var` button has been used to select the points with missing values in the variable `litmale`. This button helps to select the points that are missing in a given variable. We can see that imputed data for the variable `litmale` are generally high. Hence, all these countries seem to have a high proportion of educated males. This trend extends multivariately as it corresponds to countries with high values in `litfem`, `explfem`, `explmale`, `calories` and `logpib` as well as low values in `logbirtdeat`, `logfertili` and `mortchil`. Looking at the labels for these observations it can be seen that they generally correspond to developed countries.

2.2 Report

Figure 6 and 7 show the report for missing data provided by ViSta. The first window is

```
DESCRIPTIVES FOR MISSING AND IMPUTED DATA

Number of cases                109.00
Number of values in the data matrix 1308.00
Number of values missing       88.00
Percentage of values missing    6.73

Little's MCAR means test
Chisq    df    Prob
114.75   53.00   0.00
If the probability is low we reject
that data are Missing
Completely At Random (MCAR)
```

Descriptives	Mean	Median	Std	Skew	Kurt	Min	Max	Missing	N
<code>logbirtdeat</code>									
Uniwise	0.97	0.98	0.62	0.12	-0.74	-0.08	2.64	1.00	108.00
Listwise	1.20	1.26	0.52	-0.21	0.61	-0.08	2.64	50.00	59.00
EM	0.97	0.98	0.62	0.13	-0.73	-0.08	2.64	0.00	0.00
Imp-rand	0.97	0.98	0.62	0.12	-0.72	-0.08	2.64	0.00	0.00
<code>logfertili</code>									
Uniwise	1.13	1.12	0.53	0.19	-1.36	0.26	2.10	2.00	107.00
Listwise	1.36	1.39	0.48	-0.45	-0.75	0.26	2.10	50.00	59.00
EM	1.12	1.06	0.53	0.23	-1.35	0.26	2.10	0.00	0.00
Imp-rand	1.13	1.06	0.53	0.22	-1.34	0.26	2.10	0.00	0.00
<code>litmale</code>									
Uniwise	0.79	0.85	0.63	0.04	-0.79	-0.46	1.86	24.00	85.00
Listwise	0.62	0.67	0.52	-0.16	-0.46	-0.46	1.86	50.00	59.00
EM	0.92	0.95	0.64	-0.35	-0.84	-0.46	1.86	0.00	0.00
Imp-rand	0.91	0.95	0.62	-0.33	-0.79	-0.46	1.86	0.00	0.00

Figure 6: Descriptives for missing data

Missing Data Imputation

focused on standard univariate statistics for the data with missing values. This window shows the number of cases in the rectangular data matrix, the number of values (obtained multiplying the number of rows by the number of columns) and the actual number of missing values in the whole data matrix. Dividing the last two numbers and multiplying by 100 it results in the percentage of missing values with respect to the possible values in the data matrix. In this case, this number is 6.73%. Then, the Little's MCAR test of means is printed. This test is based in the EM algorithm and is based on the information on patterns of missing data. It stems from the following reasoning: means of a variable with missing values in the different patterns where it is observed should be homogeneous if data are MCAR. Little's test evaluates the homogeneity of means for all variables, across data patterns, simultaneously (Kim and Bentler, 1999; Little, 1988). In particular, it checks the means of the variables with missing values, weighted by number of cases, against the means computed from the variables using the EM algorithm. The result of the test follows the Chi Square distribution with degrees of freedom equal to the number of means available across data patterns minus the number of variables. The probability of the result is also printed.

Values of probability lower than 0.05 or 0.01 allow the null hypothesis of MCAR data to be rejected.

We repeated the imputation of missing data process a number of times in order to test the effect of excluding some variables of the dataset. The results are in table 2. Neither excluding Calories nor Litfem and Litmale will result in MCAR data but the second is closer to it. Excluding the three variables results in MCAR data because the proportion of missing values in the rest of the data matrix is very low.

COMPARISONS BETWEEN CORRELATIONS USING DIFFERENT METHODS							
Ntotal	109.00						
NListwise	59.00						
Variables	EM	Pair	List	N PW	N V1	N V2	CorrMissing
logfertili-logbirtdeat	0.66	0.66	0.39	107.00	107.00	108.00	0.70
litmale-logbirtdeat	-0.52	-0.40	-0.20	85.00	85.00	108.00	0.18
litmale-logfertili	-0.81	-0.79	-0.81	85.00	85.00	107.00	0.26
litfem-logbirtdeat	-0.48	-0.36	-0.15	85.00	85.00	108.00	0.18
litfem-logfertili	-0.80	-0.79	-0.82	85.00	85.00	107.00	0.26
litfem-litmale	0.98	0.98	0.97	85.00	85.00	85.00	1.00
logpop-logbirtdeat	-0.08	-0.08	-0.19	108.00	109.00	108.00	NIL
logpop-logfertili	-0.07	-0.07	-0.25	107.00	109.00	107.00	NIL
logpop-litmale	0.01	0.03	0.16	85.00	109.00	85.00	NIL
logpop-litfem	-0.03	-0.01	0.11	85.00	109.00	85.00	NIL

Figure 7: Correlations for the imputation of missing data

The rest of the output shows descriptive statistics for the data in different conditions. These are:

- Unwise, data available univariately,
- Listwise, data after removing rows with missing values,

Missing Data Imputation

DESCRIPTIVES FOR MISSING DATA

Number of cases	109.00								
Number of values in the data matrix	1308.00								
Number of values missing	88.00								
Percentage of values missing	6.73								
Descriptives	Mean	Median	Std	Skew	Kurt	Min	Max	Missing	N
logbirtdeat									
Uniwise	0.97	0.98	0.62	0.12	-0.74	-0.08	2.64	1.00	108.00
Listwise	1.20	1.26	0.52	-0.21	0.61	-0.08	2.64	50.00	59.00
logfertili									
Uniwise	1.13	1.12	0.53	0.19	-1.36	0.26	2.10	2.00	107.00
Listwise	1.36	1.39	0.48	-0.45	-0.75	0.26	2.10	50.00	59.00
litmale_p_0.34									
Uniwise	0.79	0.85	0.63	0.04	-0.79	-0.46	1.86	24.00	85.00
Listwise	0.62	0.67	0.52	-0.16	-0.46	-0.46	1.86	50.00	59.00
litfem_p_0.33									
Uniwise	0.52	0.44	0.81	0.12	-0.90	-0.99	1.90	24.00	85.00
Listwise	0.31	0.37	0.66	-0.01	-0.73	-0.99	1.90	50.00	59.00
logpop									
Uniwise	4.11	4.02	0.65	0.25	0.59	2.41	6.08	0.00	109.00
Listwise	4.24	4.12	0.65	0.66	0.35	3.11	6.08	50.00	59.00
log-density									
Uniwise	4.11	4.16	1.44	0.07	0.89	0.83	8.61	0.00	109.00
Listwise	3.82	3.83	1.41	0.23	1.14	0.83	8.40	50.00	59.00
log-aids									
Uniwise	18.30	18.15	5.95	-0.25	1.58	0.00	36.43	3.00	106.00
Listwise	19.82	19.29	5.88	-0.11	1.62	0.00	36.43	50.00	59.00
explfem									
Uniwise	70.16	74.00	10.57	-1.11	0.21	43.00	82.00	0.00	109.00
Listwise	65.83	68.00	11.08	-0.53	-0.82	43.00	81.00	50.00	59.00
explmale									
Uniwise	64.92	67.00	9.27	-1.08	0.34	41.00	76.00	0.00	109.00
Listwise	61.34	63.00	9.93	-0.58	-0.66	41.00	76.00	50.00	59.00

Figure 8: Descriptives for missing data

- EM, data after imputing the missing values using regression estimation with the coefficients computed using the EM algorithm.

Imp-rand, data after imputing the missing values using the previous method and then adding to the estimations a variable normally distributed with mean zero and standard deviation equal to the prediction error of a regression model computed using the rest of variables apart from the predicted. This correction to imputed data will be discussed in the section about creating data.

In general, differences between results in the different conditions should be checked.

Other output is referred to correlations using different methods. This output permits comparing the EM method with Listwise and Pairwise methods, as well as the data available pairwise and in each variable. Finally, the correlations between indicator variables for missing and observed data for the pairs of variables. High values of these correlation reveals that the two variables involved in the computation are simultaneously missing.

2.3 Creating data

The last step after using the missing data analysis methods in ViSta is creating new data that

Missing Data Imputation

Model	Chi Square	Df	Prob.
All 12 variables	114.75	53	< 0.0001
Excluding Calories	62.99	32	<0.001
Excluding Litfem and Litmale	46.96	32	0.025
Excluding the previous three variables	11.31	21	0.95

TABLE 2: Comparison of the effect of excluding some variables of the dataset on the Little's test.

incorporates the results of the previous analysis. This is a very controversial issue for some people because it involves something similar to "fabrication" of data, and suggests some kind of cheating by the researcher. Listwise deletion, on the other hand, may appear more honest because it discards uncertain information. However, as it should have been made clear before, listwise deletion can be quite bad and almost any reasonable method of imputation will generally improve it.

As pointed out to the reader, the solution obtained using the EM algorithm can be regarded as the best from the point of view of estimating the parameters (means and variances-covariances) of data with missing values. Therefore, analysis involving only these parameters should be apparently the best way to go after this estimation. However, these parameters are often not adequate for practical purposes and the researcher will want to have estimations of

PAIRWISE AND LISTWISE CORRELATIONS

Ntotal	109.00						
NListwise	59.00						
Variables	Pair	List	N PW	N V1	N V2	CorrMiss	
logfertili-logbirtdeat	0.66	0.39	107.00	107.00	108.00	0.70	
litmale_p_0.34-logbirtdeat	-0.40	-0.20	85.00	85.00	108.00	0.18	
litmale_p_0.34-logfertili	-0.79	-0.81	85.00	85.00	107.00	0.26	
litfem_p_0.33-logbirtdeat	-0.36	-0.15	85.00	85.00	108.00	0.18	
litfem_p_0.33-logfertili	-0.79	-0.82	85.00	85.00	107.00	0.26	
litfem_p_0.33-litmale_p_0.34	0.98	0.97	85.00	85.00	85.00	1.00	
logpop-logbirtdeat	-0.08	-0.19	108.00	109.00	108.00	NIL	
logpop-logfertili	-0.07	-0.25	107.00	109.00	107.00	NIL	
logpop-litmale_p_0.34	0.03	0.16	85.00	109.00	85.00	NIL	
logpop-litfem_p_0.33	-0.01	0.11	85.00	109.00	85.00	NIL	
log-density-logbirtdeat	-0.13	-0.10	108.00	109.00	108.00	NIL	
log-density-logfertili	-0.25	-0.17	107.00	109.00	107.00	NIL	
log-density-litmale_p_0.34	0.09	0.05	85.00	109.00	85.00	NIL	
log-density-litfem_p_0.33	0.07	0.03	85.00	109.00	85.00	NIL	
log-density-logpop	0.14	0.28	109.00	109.00	109.00	NIL	
log-aids-logbirtdeat	-0.22	-0.40	106.00	106.00	108.00	0.57	
log-aids-logfertili	-0.08	-0.14	106.00	106.00	107.00	0.81	
log-aids-litmale_p_0.34	-0.10	0.16	84.00	106.00	85.00	0.18	
log-aids-litfem_p_0.33	-0.08	0.20	84.00	106.00	85.00	0.18	
log-aids-logpop	0.24	0.19	106.00	106.00	109.00	NIL	
log-aids-log-density	-0.08	-0.08	106.00	106.00	109.00	NIL	
explfem-logbirtdeat	-0.25	0.14	108.00	109.00	108.00	NIL	
explfem-logfertili	-0.82	-0.77	107.00	109.00	107.00	NIL	
explfem-litmale_p_0.34	0.74	0.77	85.00	109.00	85.00	NIL	

Figure 9: Pairwise and Listwise correlations

the missing values. A reason for this is that many statistical packages are not designed to accept only the parameters as input. For example, the only analysis that ViSta can compute on matrixes of square matrixes like a matrix of correlations or covariances is multidimensional scaling analysis. For this reason, ViSta outputs matrixes of data with the imputed data.

The figure 10 shows the dialog box for creating data in ViSta. We can clasify the results in two groups: Square matrixes that refer to matrixes of parameters, and rectangular matrixes that refer to raw (after manipulation) data.

4. Square Matrixes

ViSta MissDat outputs the following square matrixes.

Pairwise correlations: Outputs a matrix with the pairwise correlations.

Listwise correlations: Outputs a matrix with the listwise correlations.

Maximum Likelihood Correlations: Outputs a matrix with the Maximum Likelihood Correlations computed using the EM algorithm.

Maximum Likelihood Covariances: Outputs a matrix with the Maximum Likelihood Correlations computed using the EM algorithm.

Missingness Matrix: This matrix is an indicator matrix with ones for observed values and zeros for missing values. This matrix can be used for analysis involving patterns of missingness.

5. Rectangular Matrixes

The rectangular matrixes that ViSta MissDat outputs can be classified in matrixes of available data, single imputation, and multiple imputation. The methods in the dialog box are sorted so those on the top can be regarded as less correct than those placed down.

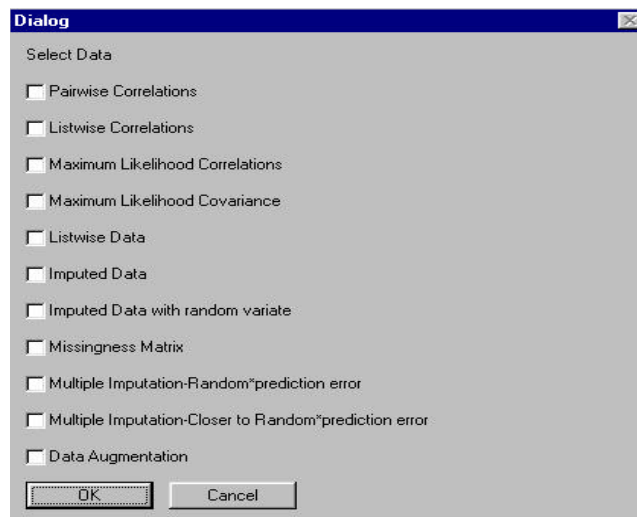


Figure 10: Dialog box for types of data that can be created with missing data imputation in ViSta

Missing Data Imputation

Available data

Listwise data: This is the same matrix that can be obtained in ViSta using the command **Create Data** and then choosing the option Remove Rows with missing data. If you have arrived to here you will probably not prefer this matrix as opposed to some of the following options.

Single Imputation

Single Imputation just fills the missing data and creates a new data set. This method is more efficient than Listwise deletion because it uses more of the original data matrix. There are two methods of this kind available in ViSta.

Imputed Data: This is the matrix with missing values filled with regression predictions obtained using the EM estimation of the missing values. This implies that each variable with missing values occupies the place of predicted in a multiple regression with the rest of variables occupy the place of predictors. These imputations can be useful as approximations but they deflate the variances of the variables and inflate their covariances because they underestimate the error component. In general, they are only appropriate when the number of missing values is not high.

Imputed Data with random variate. This matrix is the same as the previous but adding to each filled value other value coming from a variable normally distributed with mean zero and standard deviation equal to the residual variance of a regression model computed using the rest of variables apart from the predicted. This correction reintroduces variability into the data so that deflation of variances and inflation of covariances is corrected partially.

Single Imputation is not a perfect solution because it fails to produce an estimation of the right standard errors of parameters. Therefore, using a standard statistical analysis with the imputed data may be wrong because the printed standard errors assume that the data matrix was completely observed and they do not reflect missing data uncertainty. This leads to an overstatement of the size of N's, intervals of confidence too narrow and Type I errors too high (Schafer, 1999). The problem becomes worse when the rate of missing information and the number of parameters increases. Rate of missing information is discussed in the section about convergence. A solution that improves single imputation is multiple imputation (Rubin, 1987; Schafer, 1999).

2.4 Convergence

The EM algorithm is quite simple and stable. However, it is not absolutely guaranteed that abnormalities will not arise. Typically, these abnormalities happen with small samples, high rates of missingness and models with too many parameters relative to the amount of information in the observed part of the data matrix (Schafer, 1999). ViSta provides some options that allows exploring these problems. These are the possibility of choosing a starting point for the algorithm, the printing of the loglikelihood of the result at each step and a spreadplot designed to monitor the convergence.

1. Starting point

The Impute missing Data command has three choices as starting points for the algorithm.

Null covariances is a matrix with variances of the variables in the diagonal and zeros in the rest of the matrix. Listwise covariances are the covariances of the data after removing rows with missing data. Pairwise covariances are the covariances of the variables computed vari-

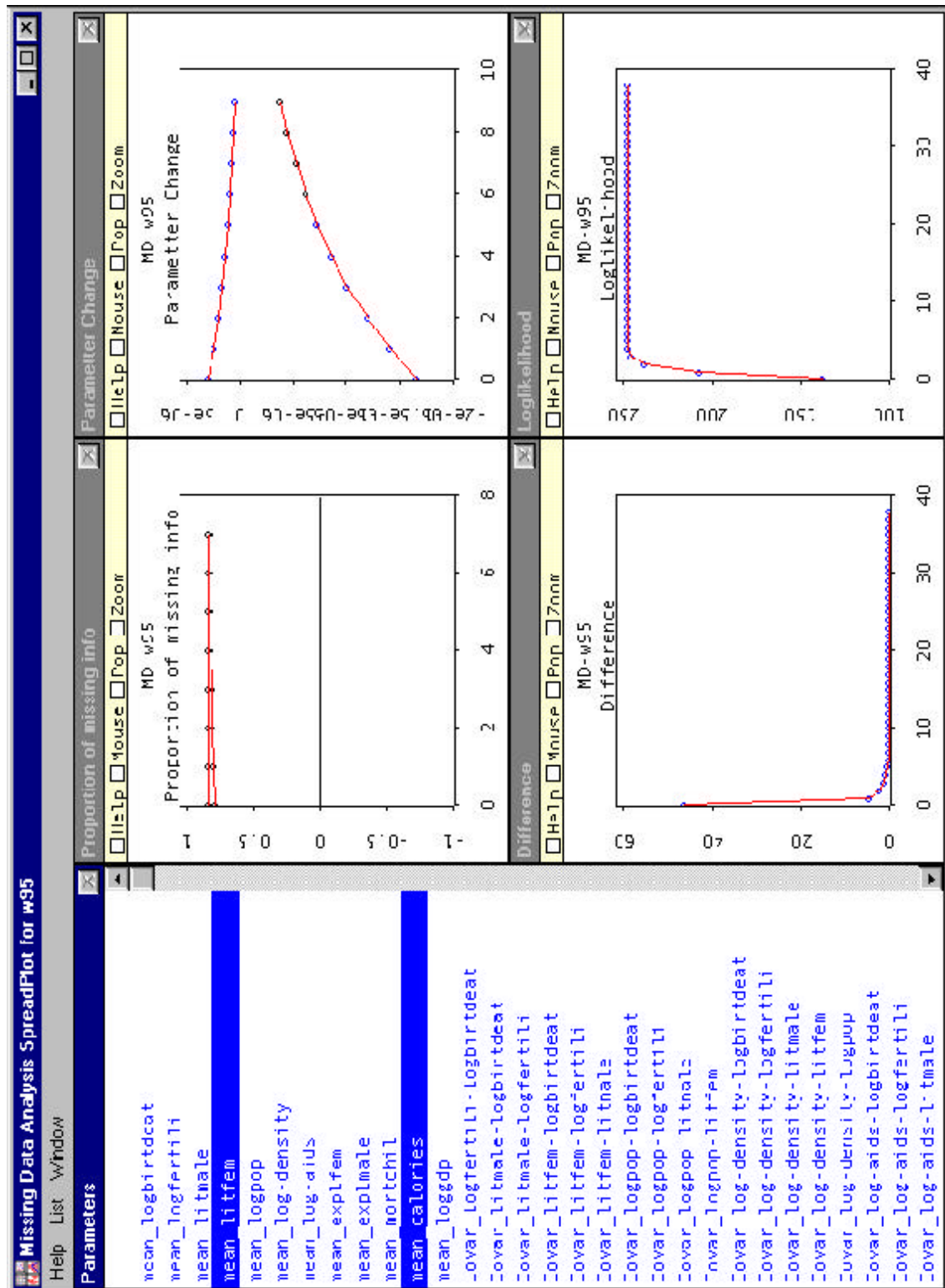


Figure 11: Spreadplot for convergence

able by variable, using the available data in each case. In all the cases, the initial means are obtained from available data. The first method provides a solution that is always workable as initial estimation. The second may produce problems when a variable becomes a constant after deletion and the analysis will simply not proceed. The third solution will not be adequate sometimes in the first iteration because the matrix may not be positive definite. It is usually wise to run the analysis from these different starting points to check if all the solutions converge to the same result. Depending on the starting point chosen the number of iterations needed to converge may change but there is not clear guidelines indicating which is the most adequate for each case.

2. Printing of loglikelihood

ViSta prints the loglikelihood of the observed data at each step of the process until convergence. This loglikelihood is a measure of the likelihood of the data given the parameters found at this step. The EM algorithm is guaranteed not to decrease this likelihood. However, ViSta does not utilize the change in loglikelihood as criteria for stopping the process but uses the change in the matrix of parameters instead.

Problems where the matrix of parameters keep changing without much increase in the loglikelihood indicate that the maximum fraction of missing information (see below) is close to one. In this case, additional iterations are probably of little interest. A reasonable strategy in this case can be excluding the variables with high proportion of missingness.

3. Spreadplot for convergence

ViSta includes a spreadplot specially designed to explore some aspects related with convergence. This spreadplot for the data about countries is shown in figure 11. This spreadplot includes four plots and a list of names of the parameters in the data.

The window titled **Parameters** has a list of all the parameters in the data. Means of the variables are listed first, then covariances and finally variances. Selecting one or several parameters in this window results in changes in the windows titled **Proportion of missing info** and **Parameter Change**.

The window titled **Proportion of missing info** shows the elementwise rate of convergence for the parameter. The rate of convergence is computed dividing the difference between a matrix of parameters in the step t and the step $t+1$ by the difference between the matrix of parameters in the step t and the step $t-1$. Rates of convergence are between zero and one and are a useful estimation of the largest fraction of missing information, where zero means no missing information and one means much missing information. The fraction of missing information will be approximately the proportion of missing values respect to the total of data for univariate problems but for multivariate data it is a more complex function (Schafer, 1997). Not every parameter will have the same rate of convergence. Those without missing information will have rates equal to zero but, typically, several parameters will have rates of convergence close to a maximum. This maximum can be observed in figure xx. Means of the variables *litfem* and *calories* have been selected and the rates of convergence shown are about .80. It should be warned that due to numerical inaccuracy some of the values will occasionally exceed one or

will wander without a precise trend. These plots should be ignored. Also, these rates are not reliable except when computed from the last few iterations as far as sign of numerical inaccuracy does not arise. ViSta also shows the computation for the last 10 iterations, which results in only the 8 values shown in this plot.

The window titled **Parameter Change** shows the elementwise differences between parameters for the last 10 iterations. This plot may help to diagnose numerical problems in the previous plot.

The last two plots are different of the previously discussed because they refer to global changes instead of elementwise. The plot titled **Difference** is a plot of the values of the sum of differences in absolute value in the successive iterations of the EM algorithm. The plot titled **Loglikelihood** is a plot of the values of the loglikelihood at the iterations of the EM algorithm. These two plots will help to diagnose if the algorithm has converged effectively or if it happens that it runs along several steps without practical consequences.

3 USING MISSING DATA ANALYSIS

The Impute missing data command works on multivariate numerical data. Categorical or ordinal data will not be put into the analysis and the create data command will only output the numerical variables in the analysis. Also, as the imputation technique is supported on multivariate regression analysis, all the assumptions of this technique apply here. Therefore, transformations of numerical variables may be necessary in order to satisfy these assumptions.

1. Analysis options

Missing data imputation includes three options as shown in figure 12.

1. Difference between parameters: This is value that ViSta uses as criteria to stop the EM algorithm. This is the sum, in absolute values, of the sum of the difference between the matrix of parameters in the iteration t and the iteration $t+1$. ViSta normalizes the vari-

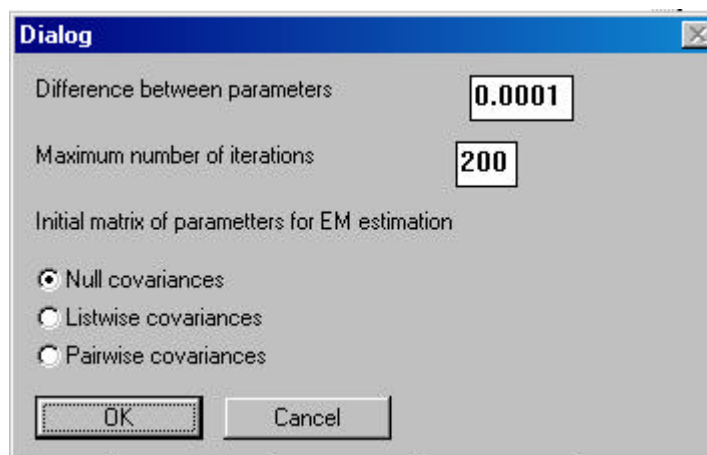


Figure 12: Choosing an initial matrix of parameters for EM estimation

ables to mean zero and standard deviation one the original data before starting the process of computation. This avoids that variables with bigger variances control this sum. Changing this number to a bigger number results in less iterations.

2. Maximum number of iterations: This value limits the number of iterations that the EM algorithm will run. It is usually wise to set this number to a small value (say 2 or 3) when the problem to be estimated is relatively complex and each iteration is anticipated to take too long. This will give an indication of the computer time that will be necessary for the problem
3. Initial matrix of parameters. This is the initial matrix used as starting point for the analysis. It has been discussed in the section on convergence.

The process for imputing data can also be started by typing (impute-missing-data) in the Listener window. This analyzes the current data and produces a missing data model object in the screen. There are several keywords that can be used additionally when typing in the Listener window. These keywords and the default values are as follows:

:dialog followed by t or nil (the default). Controls if the dialog box for options pops up. It can be useful to carry out analysis automatically without user interaction.

:title followed by a title enclosed by quotes. The default is Missing Data Analysis.

:difference followed by the difference between parameters that controls the convergence. Default is 0.001

:iterations followed by the maximum number of iterations computed. Default is 200.

:initial matrix followed by a number. 0 is a null or identity matrix. 1 is a listwise matrix. 2 is a pairwise matrix.

4 Algorithm

The EM algorithm can be regarded as a general strategy that can be applied to a variety of problems involving incomplete data. The EM algorithm formalizes an old ad hoc idea for handling missing values: (1) replace missing values by estimated values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameters are correct, (4) reestimate parameters, and so forth, until convergence (Little and Rubin, 1987). The algorithm implemented here corresponds to Incomplete Multivariate Normal Samples.

The E step of this algorithm at the iteration t consists in calculating:

$$\sum_{i=1}^n y_{ij}^t$$

$$\sum_{i=1}^n \frac{y_{ij}^t}{\sum_{j=1}^p y_{ij}^t} \quad \text{and} \quad \sum_{i=1}^n \frac{y_{ij}^t y_{ik}^t}{\sum_{j=1}^p y_{ij}^t}$$

Where

$$y'_{ij} = \begin{cases} y_{ij}^t & \text{if } y_{ij} \text{ is observed} \\ E(y_{ij} \mid y_{obs,i}, \mathbf{q}^t) & \text{if } y_{ij} \text{ is missing} \end{cases}$$

and

$$c'_{jki} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ are observed} \\ \text{cov}(y_{ij}, y_{ik} \mid y_{obs,i}, \mathbf{q}^t) & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases}$$

Computing the value expected for the missing values would be a formidable task without the SWEEP matrix operation. This operation provides a simple and convenient way to compute multiple regression coefficients. Lisp-Stat provides a version of this operation.

The M step of the algorithm consists in recomputing the parameters for the step $t+1$. These parameters are obtained from the completed data. Two algorithms can be used in this case. The one used here follows Schafer (1997) and stores the sums and the sums of squares by patterns of missing data. Additional details can be consulted there.

5 References

- Graham, J. W., Hofer, S. M. And Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In Collins, L. M. and Seitz, L. A., *Advances in Data Analysis for Prevention Intervention Research*, 13-63. National Institute on Drug Abuse.
- Kim, K. H., and Bentler, P. M. (1999). *Tests of homogeneity of means and covariance*. Report. University of California. Los Angeles.
- Little, R. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198-1202.
- Little, R. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley and Sons.
- McLachlan, G. J., and Krishnan, T. (1997). *The EM algorithm and Extensions*. New York: Wiley and Sons.
- Rubin, D. B. (1987). *Multiple Imputation for Non Response in Surveys*. New York: Wiley and Sons.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.
- Schafer, J. L. (1999). Fabricate your data well. *Slides for a presentation in the Joint Statistical Meetings*, Baltimore.