# COMPUTING AND VISUALIZING LOG-LINEAR ANALYSIS INTERACTIVELY

Pedro Valero-Mora

Forrest W. Young

## INTRODUCTION

Log-linear models provide a method for analyzing associations between two or more categorical variables and they have become widely accepted as a tool for researchers during the last two decades. There is therefore a range of excellent textbooks that give overviews of categorical data analysis by means of log-linear analysis (Agresti, 1990; Ato and López 1996; Andersen 1990, 1996; Christensen 1990) and almost every major statistical package includes capabilities for computing them. As well as this, it is not rare at this moment to find universities offering intermediate level courses for undergraduates, including not only those focused on statistical science but also on social sciences, business, biology and medicine.

However, log-linear analysis is regarded in many cases as an arduous subject for those undergraduates and researchers who are not interested in statistical science per se because it involves concepts and strategies that apparently differ greatly from those usually studied by them. Computer software oriented towards log-linear analysis has not always contributed to facilitating the task of learning and using these kind of models. Friendly (2000), for example, signals that the usual computer programs for log-linear analysis present the results in terms of tables of parameter estimates. This makes it difficult to understand the nature of the associations, particularly for large tables. Friendly emphasizes the use of graphics and plots for helping to understand more easily the results of log-linear models. He outlines that, while the use of plots has become very common to the researcher dealing with numerical data, they are scarcely applied to categorical data by comparison.

Another difficulty with computer programs for log-linear analysis and novice students is that very often it is necessary to learn a moderately complex command language to take advantage of them. This complexity stems from the need for indicating the terms included in models with several variables and interactions. On the other hand, direct manipulation interfaces (i.e. point and click) have demonstrated to be often easier for students (when implemented appropriately) than command languages. Hence, there should be an interest in exploring ways of implementing these kinds of interfaces for log-linear analysis.

Finally, while there has been a progressive increase in the use of interactive, dynamic and linked plots to many types of statistical analysis (Cook & Weisberg 1999; Tierney 1990; Velleman 1995) throughout the last two decades, to our knowledge there have not been many specific applications of them to log-linear analysis. On the other hand, as we have been applying a specific way of using interactive, dynamic and linked plots, to different types of statistical models, we are interested in extending our experiences also to log-linear analysis.

The purpose of this paper is to describe a simple program for computing log-linear analysis based on a direct manipulation interface that emphasizes the use of plots for guiding the analysis and evaluating the results obtained. The program described here works as a plugin for ViSta (Young 1997) and receives the name of LoginViSta (for Log-linear analysis in ViSTa). ViSta is a statistical package based on Lisp-Stat. Lisp-Stat is a statistical programming environment developed by Luke Tierney (1990) that features an object-oriented approach for statistical computing and one that allows for interactive and dynamic graphs. In particular, Tierney (1990) extended the Lisp language to support vectorized arithmetic, basic statistical calculations, a window system and tools for building graphics, with special emphasis on dynamic graphics. ViSta incorporates the object-oriented approach as part of its internal and external functioning. In particular, it extends Lisp-Stat with additional graphical, statistical model and data objects; it provides objects for mapping the process of data analyses and has objects that guide novices through their early attempts to carry out analyses. All these characteristics shape a system that has been shown to be appropriate for students and teachers of statistics as well as for researchers and developers in computational and graphical statistics.

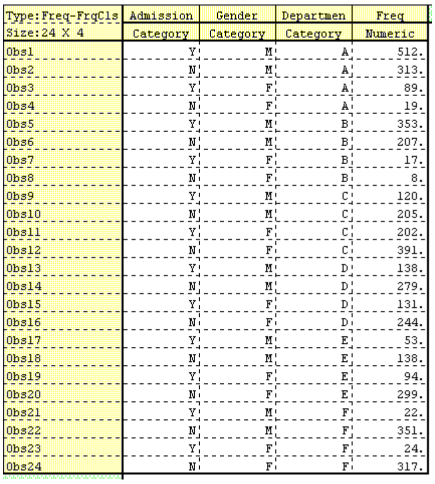
Spreadplots, one of the innovative aspects of ViSta (Young, Faldowsky & McFarlane 1993), are groups of several plots that simultaneously provide alternative views of data or statistical model objects. The plots in a spreadplot are linked, they are dynamic, and they work coordinately as a set. For example, the plots in a spreadplot show up and are removed from the screen at the same time as response to a single action carried out by the user. Spreadplots also automatically lay out the plots so there is no overlapping between them. Finally, spreadplots provide a standard programming method for implementing innovative linking features.

In summary, the advantages of LoginViSta with respect to other programs for log-linear analysis are as follows:

1) It provides an easy to use graphical (point and click) method for specification of log-linear models, whether they are hierarchical or not.
2) Plots are automatically recomputed so the user can visualize the consequences of a model without additional actions.
3) The history of the models evaluated by the user is kept automatically. This also makes it easier to carry out specific comparisons between models.

The plan of this paper is the following: we will illustrate how to use LoginViSta for computing a sequence of log-linear analysis. Then we will describe some additional features not covered in the example. A short section with some final comments about the software will close this paper.

## Working with Login-ViSta



| Type:Freq-FrqCls | Admission | Gender | Departmen | Freq |
|---|---|---|---|---|
| Size:24 X 4 | Category | Category | Category | Numeric |
| Obs1 | Y | M | A | 512. |
| Obs2 | N | M | A | 313. |
| Obs3 | Y | F | A | 89. |
| Obs4 | N | F | A | 19. |
| Obs5 | Y | M | B | 353. |
| Obs6 | N | M | B | 207. |
| Obs7 | Y | F | B | 17. |
| Obs8 | N | F | B | 8. |
| Obs9 | Y | M | C | 120. |
| Obs10 | N | M | C | 205. |
| Obs11 | Y | F | C | 202. |
| Obs12 | N | F | C | 391. |
| Obs13 | Y | M | D | 138. |
| Obs14 | N | M | D | 279. |
| Obs15 | Y | F | D | 131. |
| Obs16 | N | F | D | 244. |
| Obs17 | Y | M | E | 53. |
| Obs18 | N | M | E | 138. |
| Obs19 | Y | F | E | 94. |
| Obs20 | N | F | E | 299. |
| Obs21 | Y | M | F | 22. |
| Obs22 | N | M | F | 351. |
| Obs23 | Y | F | F | 24. |
| Obs24 | N | F | F | 317. |

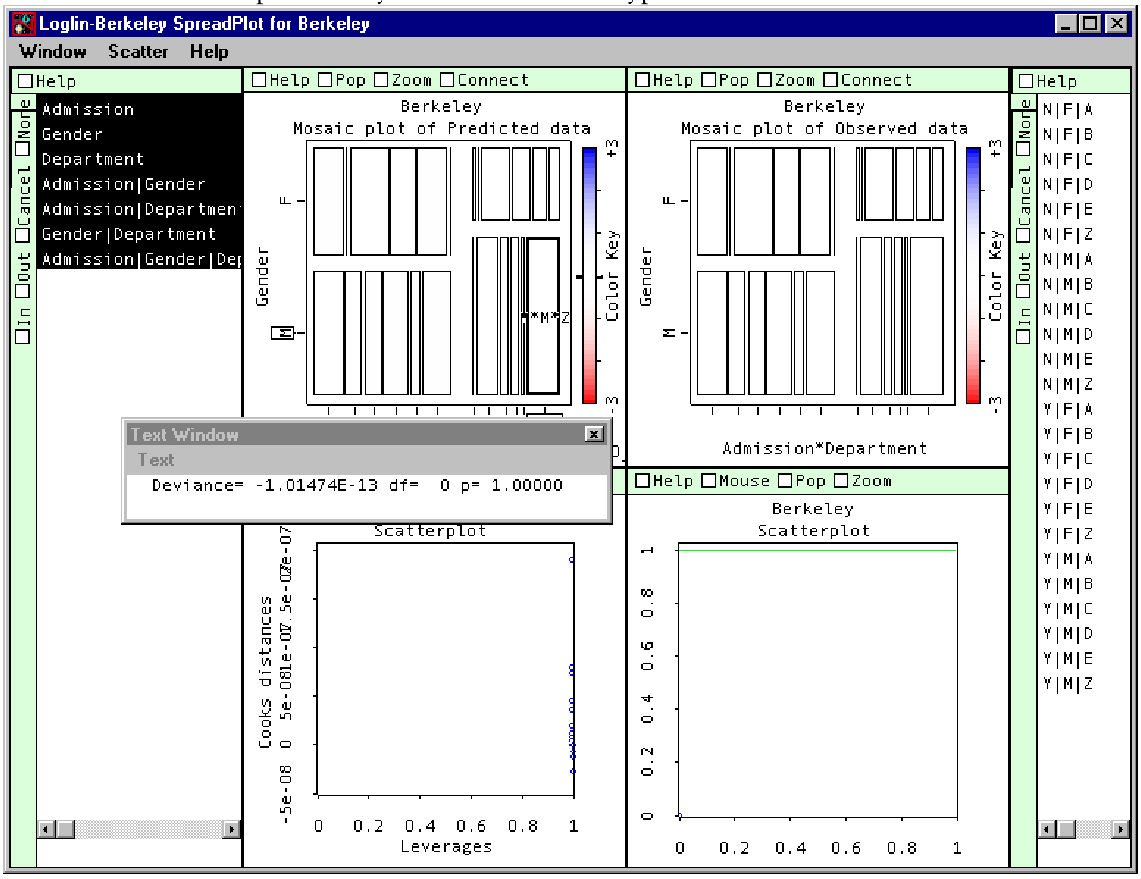
Table 1: Berkeley data.

The Berkeley admission data were described and analyzed by Bickel et al (1975). The data in table 1 reflects the six largest departments in 1971 as listed by Friendly (2000). LoginViSta uses dummy coding for creating the indicator variables required to compute the model. In particular, the default procedure is that the categories of each variable are sorted alphabetically and then the first category of them is excluded in order to avoid collinearity. The user can change this default interactively by means of a dialog box described later.

Figure 1 shows the spreadplot for log-linear analysis in ViSta. The spreadplot of Login-ViSTa includes two list windows, four plots and a floating text output window. The user can compute log-linear models, visualize them and carry out model comparisons using this spreadplot by selecting the terms to be included in the fitted model in the window placed on the left of the spreadplot. This window lists all the possible terms for the data analyzed, including every possible interaction. Therefore, in the Berkeley dataset, the window has one term for each of the three variables in the dataset (Admission, Gender and Department), three terms for the two-way interactions (Admission|Gender, Department|Gender and Admission|Department) and one additional term for the three-way interaction (Admission|Gender|Department).

The plots include two mosaic plots, a scatterplot and a connected points plot. The mosaic plot on the right shows rectangles proportional to the *observed* values (Friendly, 1994) with residuals from the current model coded as colors (red-negative; blue-positive). The mosaic plot on the left shows rectangles proportional to the *predicted* values (Thaus & Lauer, 1999) with residuals coded the same as the other mosaic plot.

The scatterplot portrays adjusted Cook distances versus leverages. The connected points plot shows the $\chi^2$ statistic of each computed model divided by the degrees of freedom of the model and can be used to trace the different models tested along a session of analysis. Finally, the floating text window shows the deviance of the model, the degrees of freedom and the probability associated to the hypothesis of null deviance.



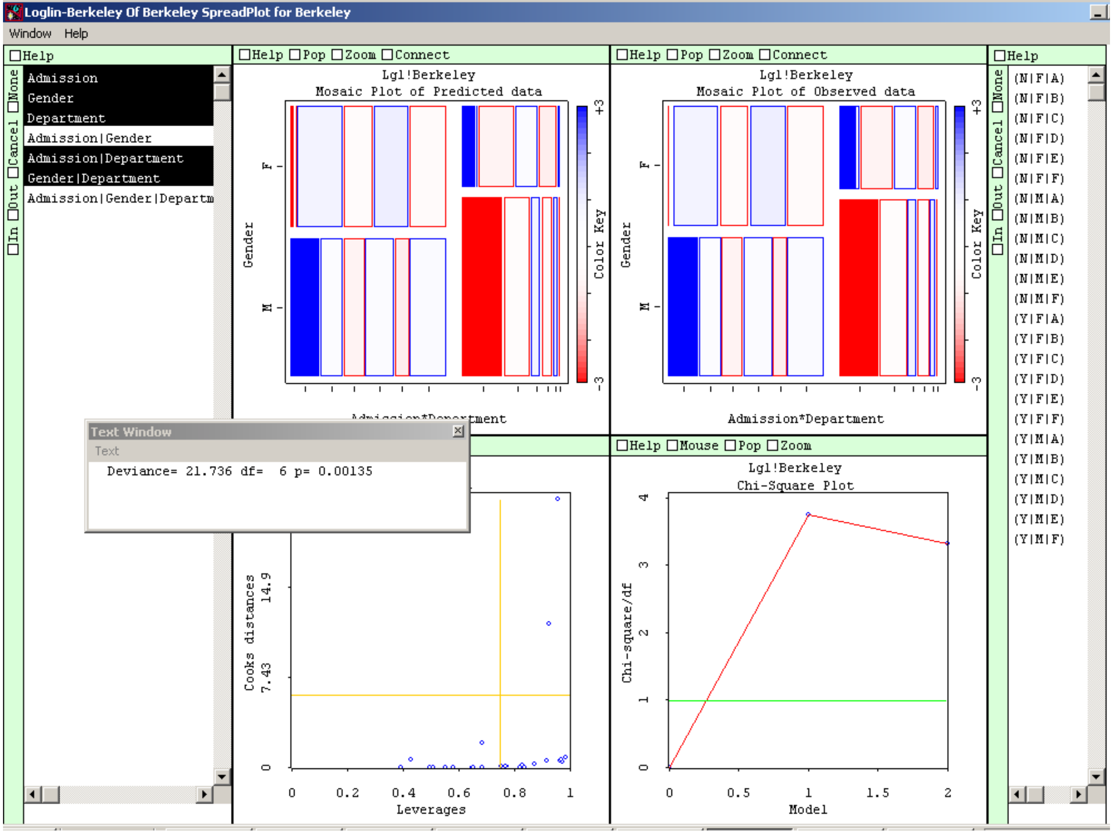
```
Deviance= -1.01474E-13 df=  0 p= 1.00000
```

Figure 1: Initial spreadplot for the Berkeley data.

The initial model computed by LoginViSta is the so-called saturated model. This model includes all the possible terms and results in a perfect fit as well as providing a useful starting point for the analysis. The user can remove or add terms to the list in the left window by pressing the Ctrl key while clicking or dragging on them with the mouse. So, for example, pressing the Ctrl key and clicking on the three-way interaction term in the figure 1 and on the interaction between Admission|Gender terms will result in the spreadplot shown in the figure 2. It can be observed that the four plots and the floating text window have all changed automatically so that the user has immediate feedback of the consequences of this action.

The floating text window in figure 2 shows that the fit of the model currently portrayed is not adequate so we reject the hypothesis of deviance equal to zero. However, the value of the deviance provides a measure of the global fit but does not provide information about the source of the differences. At this moment, the user could check the table of parameter estimates as shown by many statistical packages. A table of this sort can be obtained in LoginViSta using the menu item named Report in the Window menu but we will use the plots instead.

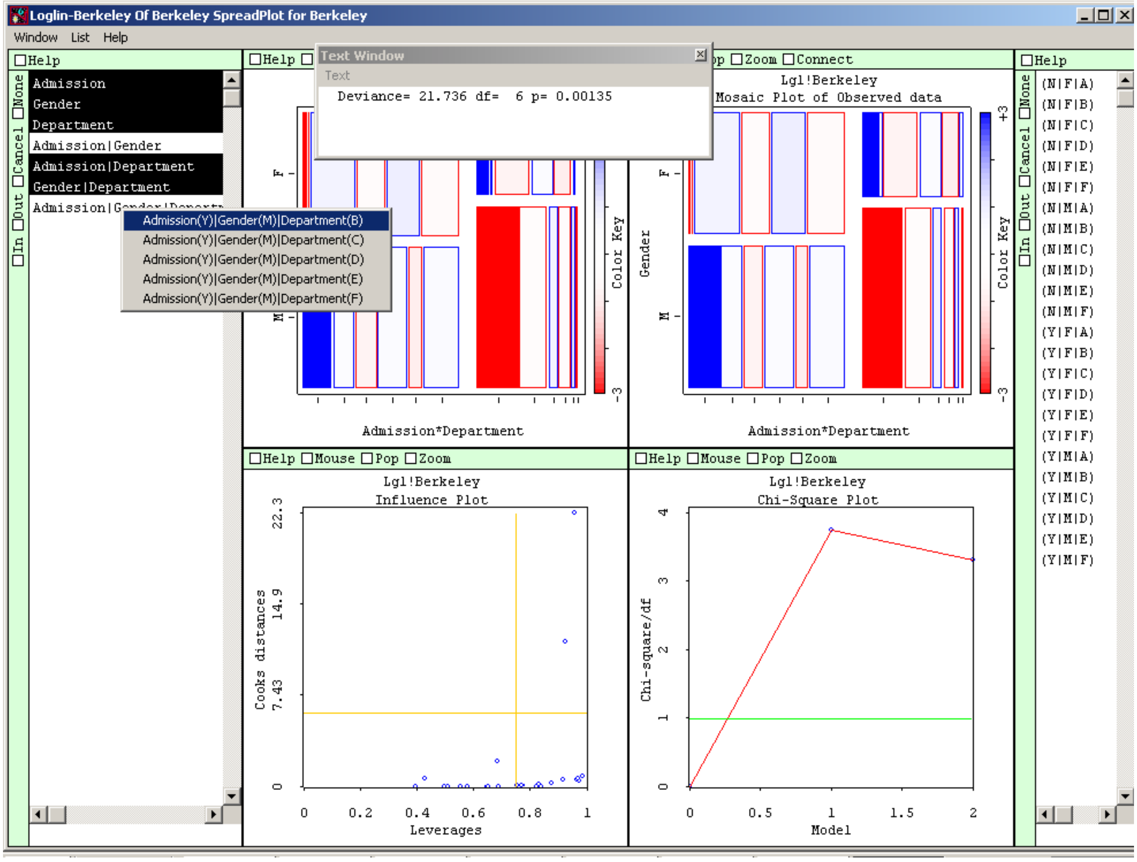
Comparison of mosaic plots will help to diagnose the source of the differences between model and data. Four cells, corresponding to the department A, stand out from the rest. Close examination of this department shows that it seems to admit more females than males given the model. This department appears special, so it looks justified to specify a model that fits a parameter expressly for it. We can indicate this type of parameters right-clicking on non included terms in the terms window as shown in Figure 3.



```
Deviance= 21.736 df=  6 p= 0.00135
```

Figure 2: Spreadplot for log-linear model

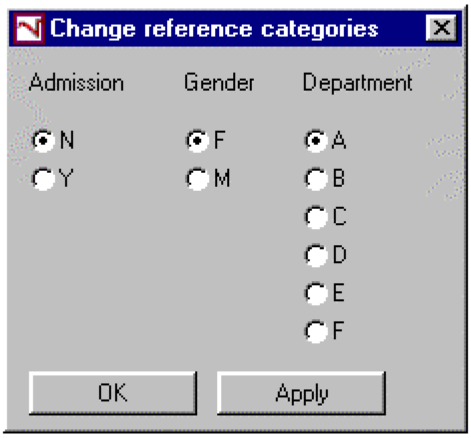
Figure 3 shows a menu for the individual parameters that would be included in the model if the three way interactions were chosen. Selecting any of them will add the parameter to the list of terms so the user can use it for including specific parameters in the model. However, this menu does not show any parameter that includes department A. This happens because the program is using department A as the reference category for the variable and all the terms that mention this department are automatically excluded.



```
Deviance= 21.736 df=  6 p= 0.00135
```

Figure 3: Spreadplot showing the menu for selecting individual indicators

At this point the user can select the menu item **Reference Categories...** and obtain the dialog box shown in Figure 4. This dialog box let the user choose reference category for the variables. Pressing the OK button will result in a new model that uses as reference categories those chosen in this dialog box.



Change reference categories

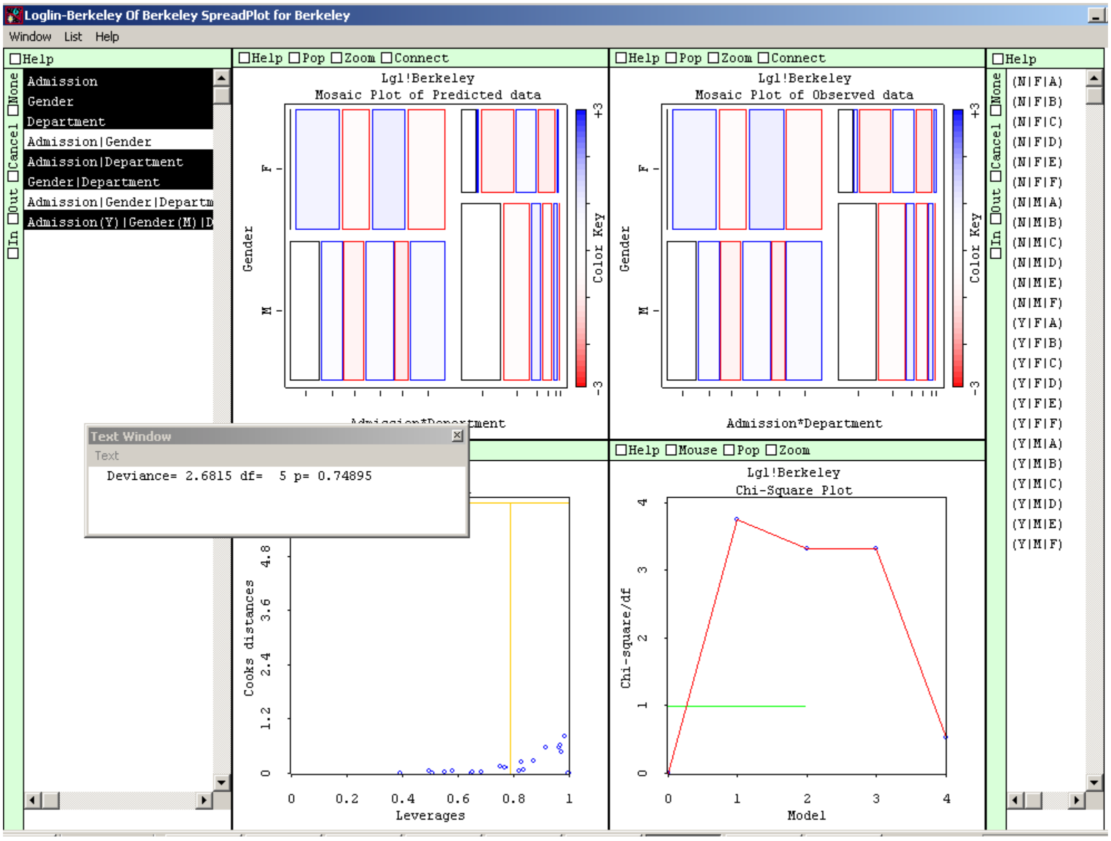
Admission: ● N ○ Y

Gender: ● F ○ M

Department: ● A ○ B ○ C ○ D ○ E ○ F

Figure 4: Dialog box for selecting the reference categories

Figure 5 shows a model that includes the parameter for the department A (last menu item in the terms window) but does not include the two-way interaction Admission|Gender. The model fits quite well as can be observed in the deviance window and in the plots of residuals, which show values close to zero for all the cells. Taking this spreadplot together with that shown in figure 2 suggests that the association between Gender and Admission is not important in the data except for the three-way interaction for Department A.



```
Deviance= 2.6815 df=  5 p= 0.74895
```

Figure 5: Model including an individual indicator for department A

## OTHER FEATURES IN LOGINVISTA

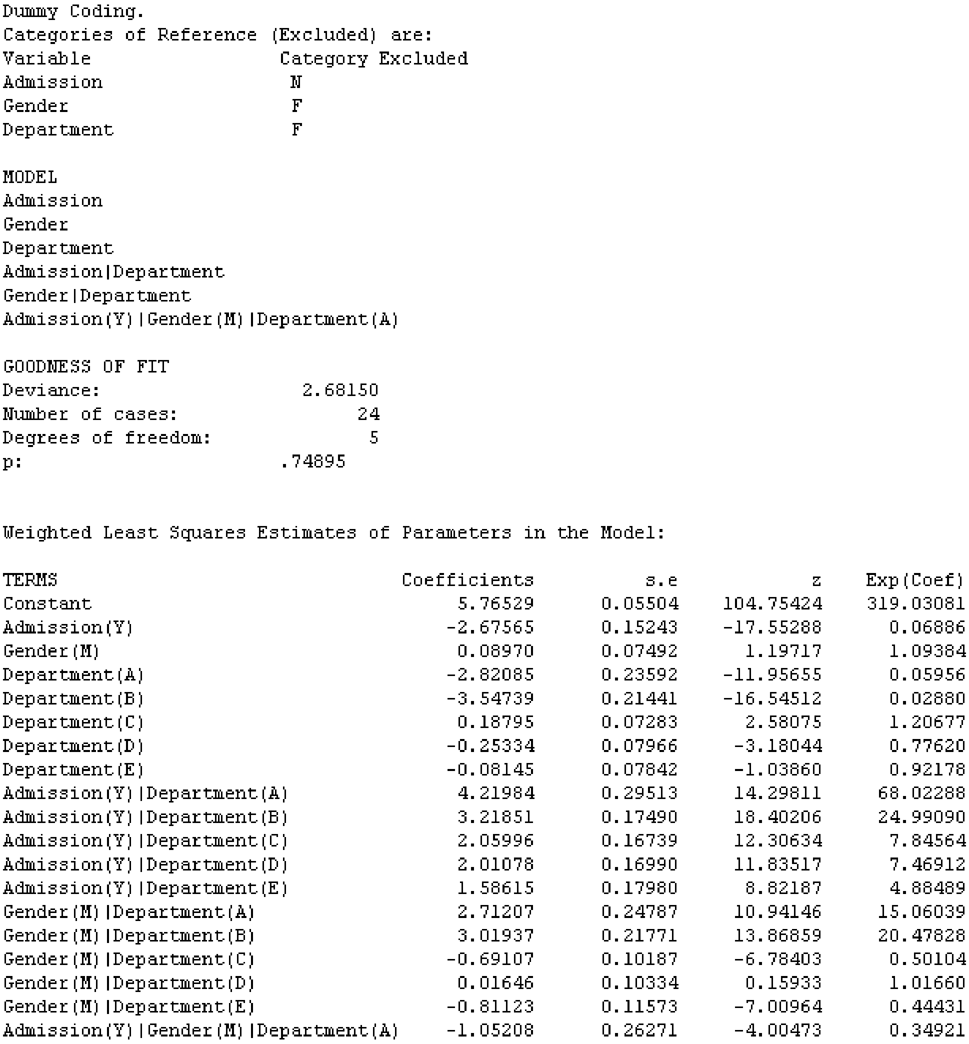
The previous analysis has demonstrated the basics about using LoginViSta for fitting log-linear models. However, there are features that have not been fully discussed yet. This section will focus on specific parts of the spreadplot in order to introduce some of the additional features. We will describe the plots from left to right, top to bottom. We will only describe characteristics that are specific to LoginViSta, leaving till the references section other material related to log-linear modeling. The parts of the spreadplot are:

1) The list of terms: The list of terms allows for selection of the terms in the model. All possible terms are listed in this window. The user selects or deselects those required for a model by using Ctrl-mouse_click or Ctrl-mouse_drag. This list has two working modes. When the list is in the hierarchical mode, including an interaction term involves automatically adding all the terms hierarchically below the (added) one. On the other hand, when the list is in the non-hierarchical mode, each term is added or removed individually. Note that LoginViSta uses the same computational method for the fitting of hierarchical or non-hierarchical models (Tierney, 1991).
2) The Leverages versus the Cook distances plot: Friendly (2000) discusses scatterplots of residuals versus leverages for diagnosing log-linear models, adding the Cook distances as bubbles associated to each point. The plots here showed use linking to displaying this information in two separate plots.
3) The Mosaic plots: The version of mosaic plots here shown is part of ViSta. It uses colors to display adjusted $\chi^2$ residuals and admits up to four variables. As cells with very low frequency are difficult to identify, LoginViSta permits

plotting the square root of the values. LoginViSta plots a scatterplot of the raw frequencies versus the residuals when the data analyzed includes more than four variables.

4) The plot for fitted models: This plot keeps track of the $\chi^2$ divided by the degrees of freedom of the different models fitted during a session. Models below the horizontal line would pass the criteria of $\chi^2/df < 1$. Selecting a point in this plot changes the spreadplot to show the values for the model selected, including the list of terms. This makes it easy to return to past successful models when the currently tested one is inadequate. Also, this plot allows for testing differences between nested models. When the user selects two points in the plot, the software will check if the models are nested and will then print the test if they are. Also, the mosaic plots will show the observed v. residual plots for both models.

LoginViSta also prints tables of Estimates of Parameters, standard errors and associated t-values under request by the user by means of the **Report** menu item in the **Window** menu of the spreadplot. An example is shown in figure 6 for the model of the figure 5.



```
Dummy Coding.
Categories of Reference (Excluded) are:
Variable                Category Excluded
Admission                N
Gender                   F
Department               F

MODEL
Admission
Gender
Department
Admission|Department
Gender|Department
Admission(Y)|Gender(M)|Department(A)

GOODNESS OF FIT
Deviance:                  2.68150
Number of cases:                24
Degrees of freedom:              5
p:                       .74895


Weighted Least Squares Estimates of Parameters in the Model:

TERMS                                Coefficients         s.e           z    Exp(Coef)
Constant                                  5.76529     0.05504   104.75424    319.03081
Admission(Y)                             -2.67565     0.15243   -17.55288      0.06886
Gender(M)                                 0.08970     0.07492     1.19717      1.09384
Department(A)                            -2.82085     0.23592   -11.95655      0.05956
Department(B)                            -3.54739     0.21441   -16.54512      0.02880
Department(C)                             0.18795     0.07283     2.58075      1.20677
Department(D)                            -0.25334     0.07966    -3.18044      0.77620
Department(E)                            -0.08145     0.07842    -1.03860      0.92178
Admission(Y)|Department(A)                4.21984     0.29513    14.29811     68.02288
Admission(Y)|Department(B)                3.21851     0.17490    18.40206     24.99090
Admission(Y)|Department(C)                2.05996     0.16739    12.30634      7.84564
Admission(Y)|Department(D)                2.01078     0.16990    11.83517      7.46912
Admission(Y)|Department(E)                1.58615     0.17980     8.82187      4.88489
Gender(M)|Department(A)                   2.71207     0.24787    10.94146     15.06039
Gender(M)|Department(B)                   3.01937     0.21771    13.86859     20.47828
Gender(M)|Department(C)                  -0.69107     0.10187    -6.78403      0.50104
Gender(M)|Department(D)                   0.01646     0.10334     0.15933      1.01660
Gender(M)|Department(E)                  -0.81123     0.11573    -7.00964      0.44431
Admission(Y)|Gender(M)|Department(A)     -1.05208     0.26271    -4.00473      0.34921
```

Figure 6: A table of estimates of parameters from LoginViSta

## FINAL COMMENTS

We consider that LoginViSta should be a useful program for students and teachers interested in log-linear analysis and visualization. The point-and-click interface, the instant feed-back of the actions, the relative lean outputs and the examples included with it make for a program that is, in our opinion, very enjoyable and fun. Of course, nothing prevents the user from applying this program to more advanced applications. It is also possible to expand the capabilities of the program using the tools provided for ViSta and Lisp-Stat.

LoginViSta has been made possible thanks to free contributions from the Internet open-software community, mainly those related to the language Lisp-Stat. So, the list of people that has made LoginViSta possible starts with Luke Tierney, that developed the Lisp-Stat language, based on the Xlisp interpreter programmed by David Betz. Then, Forrest W. Young developed ViSta, which runs almost entirely based on standard Lisp-Stat and provides the general environment where LoginViSta resides. Again, Luke Tierney developed the basic engine for computing generalized linear models that LoginViSta uses (Tierney, 1991). Sandy Weisberg contributed to code for generalized linear models with diagnostics. Jan de Leeuw implemented a syntax for fitting log-linear models that is internally used for LoginViSta to create the appropriate design matrix. Ernest Kwan developed the basic algorithm for mosaic plots that ViSta uses. Finally, Michel Friendly, Manual Ato and María Rodrigo contributed with valuable comments on previous versions of this software.

## REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley-Interscience.

Andersen, E.B. (1990). *The statistical analysis of Categorical Data*. New York: Springer-Verlag.

Andersen, E.B. (1996). *Introduction to the Statistical Analysis of Categorical Data*. New York: Springer-Verlag.

Ato, M. and López, J.J. (1996). *Análisis estadísticos para datos categóricos*. Madrid: Síntesis.

Christensen, R. (1990). *Log-Linear Models*. New York: Springer-Verlag.

Cleveland, W.S. and McGill, M.E. (1988). *Dynamic Graphics for Statistics*. Belmont: Wadsworth.

Cook, R.D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *The American Statistician, 49*, 190-200.

Friendly, M. (2000). *Visualizing Categorical Data.* SAS Press.

Rindskopf, D. (1990). Nonstandard Log-Linear Models. *Psychological Bulletin, 108*(1), 150-162.

Stine, R. and Fox, J. (Eds.). (1997). *Statistical Computing Environments for Social Research*. Thousand Oaks: Sage.

Theus, M. and Lauer, S.R.W. (1999). Visualizing Loglinear Models. *Journal of Computational and Graphical Statistics, 8*(3), 396-412.

Tierney, L. (1990). *Lisp-Stat. An Object Oriented Environment for Statistical Computing and Dynamic Graphics*. New York: Wiley.

Tierney, L. (1991). *Generalized Linear Models in Lisp-Stat* (Technical Report Nº 557). School of Statistics, University of Minnesota.

Velleman, P.F. (1995). *DataDesk Handbook.* Ithaca NY: Data Description Inc.

Young, F.W. and Bann, C.M. (1997). ViSta: A Visual Statistics System. En R. Stine and J. Fox (Eds.). *Statistical Computing Environments for Social Research* (pp. 207-235). Thousand Oaks: Sage.

Young, F.W., Faldowsky, R.A. and McFarlane, M.M. (1993). Multivariate Statistical Visualization. En C.R. Rao (Ed.). *Handbook of Statistics* (pp. 959-998). Amsterdam: North Holland.