

Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2004 - 16th Annual Conference Proceedings

APPLICATIONS OF STATISTICAL DATA MINING METHODS

George Fernandez

Follow this and additional works at: <https://newprairiepress.org/agstatconference>



Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Fernandez, George (2004). "APPLICATIONS OF STATISTICAL DATA MINING METHODS," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1148>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

APPLICATIONS OF STATISTICAL DATA MINING METHODS

George Fernandez
College of Agriculture, Biotechnology, and Natural Resources
University of Nevada – Reno
Reno NV 89434

Abstract

Data mining is a collection of analytical techniques to uncover new trends and patterns in large databases. These data mining techniques stress visualization to thoroughly study the structure of data and to check the validity of statistical model fit to the data and lead to knowledge discovery. Data mining is an interdisciplinary research area spanning several disciplines such as database management, machine learning, statistical computing, and expert systems. Although data mining is a relatively new term, the technology is not. Data mining allows users to analyze data from many different dimensions or angles, explore and categorize it, and summarize the relationships identified. **Large investments in technology and data collection are currently being made in the area of precision agriculture, remote sensing, and in bioinformatics. Experiments conducted in these disciplines are generating mountains of data at a rapid rate. Analyzing such massive data combined with the biological and environmental information would not be possible without automated and efficient data mining techniques.** Effective statistical and graphical data mining tools can enable agricultural researchers to perform quicker and more cost-effective experiments. **Commonly used statistical and graphical data mining techniques in data exploration and visualization, model selection, model development, checking for violations of statistical assumptions, and model validation** are presented here.

Keywords: Data exploration, supervised learning, unsupervised learning, model validation

1. Introduction

Data Mining is the process of extracting knowledge hidden from large volumes of raw data using analytical techniques. These data mining techniques stress visualization to thoroughly study the structure of data and to check the validity of statistical model fit to the data and lead to proactive decision making. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst. The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various business, scientific, and governmental organizations around the world is daunting. Large investments in technology and data collection are currently being made in the area of precision agriculture, remote sensing, and in bioinformatics. Experiments conducted in these disciplines are generating large amount of data at a rapid rate. Analyzing such massive data combined with the biological and environmental information would not be possible without automated and

efficient data mining techniques. Effective statistical and graphical data mining tools can enable agricultural researchers to perform quicker and more cost-effective experiments.

The first step toward building a productive data mining program is, of course, to gather data! Most institutions already perform these data gathering tasks to some extent -- the key here is to locate the data critical to your research, refine it and prepare it for the data mining process. The data mining solution is considered a process rather than a set of analytical tools. The acronym SEMMA (SAS institute, 2000) sample, explore, modify, model, assess -- refers to a methodology proposed by the SAS software that clarifies this process. Beginning with a taking statistically representative sample of your data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. The steps in the SEMMA process include:

Sample: your data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly.

Explore: your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify: your data by creating, selecting, and transforming the variables to focus the model selection process.

Model: your data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess: your data by evaluating the usefulness and reliability of the findings from the data mining process.

By assessing the results gained from each stage of the SEMMA process, you can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. Effective statistical and graphical data mining tools can enable agricultural researchers to perform quicker and more cost-effective experiments. Commonly used statistical-graphical data mining techniques in data exploration and visualization, model selection, model development, checking for violations of statistical assumptions, and model validation are presented here.

2. Data exploration and visualization

Simple scatter plots are very useful in exploring the relationship between a response and a predictor variable in simple linear regression. However, these simple scatter plots are not effective in revealing the complex relationships or detecting the trend and data problems in multiple regression models. The use and interpretation of multiple regressions depends on the estimates of individual regression coefficient. Influential outliers can bias parameter estimates and make the resulting analysis less useful. However, identifying influential outliers are not always easy in simple scatter plots. Failure to include significant quadratic or interaction terms or omitting other important predictor variables in multiple linear regression models results in model specification errors. However, identifying significant model terms in multiple linear regressions are not always easy in simple scatter plots. When the predictors are nearly perfectly related, the regression coefficients tend to be unstable and the inferences based on the regression model can be misleading and erroneous. This condition is known as multicollinearity (Mason et.

al, 1975). Severe multicollinearity in OLS regression model results in large variances and covariances for β_i and these coefficients are usually too large in absolute values with wrong signs. Interpretation of the partial regression coefficient is difficult. Multicollinearity in multiple linear regression can be detected by examining variance inflation factors (VIF) and condition indices (Neter et. al. 1989). However, identifying multicollinearity is not realistic by examining simple scatter plots.

Partial plots are considered better substitutes for scatter plots in multiple linear regression. These partial plots illustrate the partial effects or the effects of a given predictor variable after adjusting for all other predictor variables in the regression model. Two kinds of partial plots, partial regression and partial residual or added variable plot are documented in the literature (Belsley et.al 1980; Cook and Weisberg 1982).

2.1 Partial regression plots

A multiple regression model with 3 (X1-X3) predictor variables and a response variable Y is defined as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \quad (1)$$

The partial regression plot for X_1 is derived as follows:

1) Fit the following two regressions:

$$Y_i = \theta_0 + \theta_2 X_2 + \theta_3 X_3 + \epsilon_{y|x_2,x_3} \quad (2)$$

$$X_{1i} = \gamma_0 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon_{x_1|x_2,x_3} \quad (3)$$

2) Fit the following simple linear regression using the residuals of models 2 and 3.

$$\epsilon_{y|x_2,x_3} = 0 + \beta_1 \epsilon_{x_1|x_2,x_3} + \epsilon_i$$

The partial regression plot for the X_1 variable shows two sets of residuals, those from regressing the response variable (Y) and X_1 on other predictor variables. The associated simple regression has the slope of β_1 , zero intercept and the same residuals (ϵ) as the multiple linear regression. This plot is considered useful in detecting influential observations and multiple outliers (Myers, 1990). Sall (1990) proposed an improved version of the partial regression plot and called it leverage plot. He modified both X and Y axis scale by adding the response mean to $\epsilon_{y|x_2,x_3}$ and X_1 mean to $\epsilon_{x_1|x_2,x_3}$. In his leverage plots, Sall (1990) also included a horizontal line through the response mean value and a 95% confidence curves to the regression line. This modification helps us to view the contribution of other predictor variables in explaining the variability of the response variable by the degree of response shrinkage in the leverage plot. This is very useful in detecting severe multicollinearity. Also based on the position of the horizontal line through response mean and the confidence curves, the following conclusions can be made regarding the significance of the slope:

- Confidence curve crosses the horizontal line = Significant slope
- Confidence curve asymptotic to horizontal line = Border line significance
- Confidence curve does not cross the horizontal line = Non Significant slope

Thus, the leverage plots are considered useful in detecting outliers, multicollinearity, non-linearity, and the significance of the slope. An example of partial leverage plot showing a significant partial regression coefficient is shown in Figure 1.

The partial leverage plot displays three curves: a) the vertical reference line that goes through the response variable mean; b) the partial regression line which quantifies the slope of the partial regression coefficient of the i^{th} variable in the MLR; c) The 95% confidence band for partial regression line. The partial regression parameter estimates for the i^{th} variable in the multiple linear regression and their significance levels are also displayed in the titles. The slope of the partial regression coefficient is considered statistically significant at the 5% level if the response mean line intersects the 95% confidence band. If the response mean line lies within the 95% confidence band without intersecting it, then the partial regression coefficient is considered not significant (Figure 1).

2.2 Partial residual (added-variable or component plus-residual) plot (Larson and McCleary, 1972).

The Partial residual plot is derived as follows:

1) Fit the full regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \quad (4)$$

2) Construct the Partial Residual plot:

$$(\epsilon_i + \beta_1 X_{1i}) = \beta_0 + \beta_1 X_{1i} + \epsilon_i \quad (5)$$

The partial residual plot for X_1 is a simple linear regression between $(\epsilon_i + \beta_1 X_{1i})$ versus X_1 where ϵ_i is the residual of the full regression model. This simple linear regression model has the same slope (β_1) and residual (ϵ) of the multiple linear regression. The partial residual plot display allows to easily evaluating the extent of departures from linearity. These plots are also considered useful in detecting influential outliers and inequality of variance. Mallows (1986) introduced a variation of partial residual plot in which a quadratic term is used both in the fitted model and the plot. This modified partial residual plot is called an augmented partial residual plot. The Augmented Partial residual plot is derived as follows:

1) Fit the full regression model with a quadratic term:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i}^2 + \epsilon_i \quad (6)$$

2) Construct the Augmented Partial Residual plot:

$$(\epsilon_i + \beta_1 X_{1i} + \beta_4 X_{1i}^2) = \beta_0 + \beta_1 X_{1i} + \epsilon_i \quad (7)$$

The augmented partial residual plot for X_1 is a simple linear regression between $(\epsilon_i + \beta_1 X_{1i} + \beta_4 X_{1i}^2)$ versus X_1 where ϵ_i is the residual of the full regression model. The augmented partial residual plot effectively detects the need for a quadratic term or the need for a transformation for X_1 .

An example of augmented partial residual plot showing a significant partial regression coefficient and the regression relationship from a simple regression model are shown in Figure 2. The linear/quadratic regression parameter estimates for the simple and multiple linear regressions and their significance levels are also displayed in the titles. The simple linear regression line describes the relationship between the response and the predictor variable in a simple linear regression. The APR line shows the quadratic regression effect of the i^{th} predictor on the response variable after accounting for the linear effects of other predictors on the response. The APR plot is very effective in detecting significant outliers and non-linear relationships. Significant outliers and/or influential observations are identified and marked on

the APR plot if the absolute 'STUDENT' value exceeds 2.5 or the 'DFFITS' statistic exceeds 1.5. These influential statistics are derived from the MLR model involving all predictor variables. If the correlations among the predictor variables are negligible, the simple and the partial regression lines should have similar slopes.

2.3 VIF plot

Augmented partial residual and partial regression plots in the standard format generally fail to detect the presence of multicollinearity. However, the leverage plot, the partial regression plot expressed in the scale of the original X_i variable, clearly shows the degree of multicollinearity. Stine (1995) proposed overlaying the partial residual and partial regression plots on the same plot to detect the multicollinearity. Thus by overlaying the partial residual and regression plots with the centered X_i values on the X-axis, the degree of multicollinearity can be detected by amount of shrinkage of partial regression residuals. Since the overlaid plot is mainly useful in detecting multicollinearity, I named this plot as VIF plot. An example of VIF plot showing a significant partial regression coefficient and moderate level of multicollinearity is shown in Figure 3.

The VIF plot displays two overlaid curves: a) The first curve shows the relationship between partial residual + response mean and the i^{th} predictor variable b) the second curve displays the relationship between the partial leverage + response mean and the partial i^{th} predictor value + mean of i^{th} predictor value. The slope of the both regression lines should be equal to the partial regression coefficient estimate for the i^{th} predictor. Therefore, both regression lines should be identical in the VIF plot. When there is no high degree multicollinearity, both the partial residual (Symbol 'R') and the partial leverage (Symbol 'E') values should be evenly distributed around the regression line. But, in the presence of severe multicollinearity the partial leverage values, 'E' shrinks and are distributed around the mean of the i^{th} predictor variable. Also, the partial regression for the i^{th} variable shows a non-significant relationship in the partial leverage plots whereas the partial residual plot shows a significant trend for i^{th} variable. Furthermore, the degree of multicollinearity can be measured by the 'VIF' statistic in a MLR model and the 'VIF' statistic for each predictor variable is displayed on the title statement of the 'VIF' plot.

2.4 Simple and delta partial logit plots in binary logistic regression (BLR)

Simple logit plots are very useful in exploring the relationship between a binary response and a single continuous predictor variable in a BLR with a single predictor variable. But these plots are not effective in revealing the complex relationships among the many predictors. However, the partial delta logit plots proposed here are useful in detecting, significant predictors, non-linearity, and multicollinearity. The partial delta logit plot illustrates the effects of a given continuous predictor variable after adjusting for all other predictor variables on the change in the logit estimate when the variable in question is dropped from the BLR. By overlaying the simple logit and partial delta logit plots, many features of the BLR could be revealed. The mechanics of these two logit plots are described using two variable BLR model.

1) Simple logit model for the binary response and the predictor variable X_j

Fit a simple BLR model

$$\text{Logit}(P_i) = \beta_0 + \beta_1 X_1 \tag{8}$$

2) Fit a delta logit model for the binary response and the predictor variable X_1

Obtain the delta logit estimate for a given predictor

Step1: Fit the full BLR model with a quadratic term for X_1

$$\text{Logit}_{(full)}(P_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 \tag{9}$$

Step2: Fit the reduced BLR model

$$\text{Logit}_{(reduced)}(P_i) = \beta_0 + \beta_2 X_2 \tag{10}$$

Step3: Estimate the delta logit: Difference in logit between the full and the reduced model:

$$\Delta \text{logit} = \text{Logit}_{(full)} - \text{Logit}_{(reduced)} \tag{11}$$

Step4: Compute the partial residual for X_1 and add X_1 -mean

$$X_i = a_0 + b_2 X_2 + e_i \tag{12}$$

$$PR_{X_1} = e_i + X_1 \text{ mean} \tag{13}$$

Step 5: Overlay simple logit and partial delta logit plots

Simple logit plot: $\text{Logit}(P_i)$ vs. X_1

Partial delta logit plot: Δlogit vs. PR_{X_1}

Positive or negative slope in the partial delta logit plot shows the significance of the predictor variable in question. Quadratic trend in the partial delta logit plot confirms the need for quadratic term for X_1 in BLR. Clustering of delta logit points near the mean of X_1 in the partial delta logit plot confirms presence of the multicollinearity among the predictors. Large differences between the simple logit and the partial delta logit line illustrate the difference between the simple and the partial effects for a given variable X_1 . See an example of simple and partial delta logit plot in Figure 4.

1.5 Interaction plot in multiple linear regression

The statistical significance of an interaction term ($x_1 * x_2$) in a MLR can be visualized in a 3-d plot between the interaction component, x_1 and x_2 variables. To estimate the interaction component, first fit a full MLR model + the interaction term in question and estimate the predicted value (full model) and estimate the *p-value* for the statistical significance of the interaction term. Then fit a reduced model without the interaction term and estimate the predicted value (reduced). Obtain the interaction component by adding the Y-mean to the difference between the full and the reduced model. Show the interaction effect by plotting the interaction effect in the z-axis and the both x_1 and x_2 variables in the X and y axis. The interaction 3-d plot shows the nature of interaction and the statistical significance of the interaction term is displayed on the title (Figure 5).

1.6 Scatter-plot matrix of simple linear correlations

Examining the correlations among the multi-attributes in a series of simple scatter plots between any two variables is the first step in exploring multivariate data. This scatter plot matrix display is a useful exploratory technique in principal component, exploratory factor and

canonical discriminant analysis. An example of this simple two-dimensional scatter plot matrix showing the correlation between any two attributes is presented Figure 6. The regression line displays significant positive or negative relationship. If the 95% confidence interval lines intersect the y-axis mean (horizontal line) then the observed correlation is considered significant at 5% level. These scatter plots are useful in examining the range variation and the degree of correlations between any two attributes.

3. Model selection and model fit

Selecting the significant predictor variables and model terms are important in multiple linear and logistic regression models. Several step-wise and all possible selection models are available in multiple linear regression models. The MaxR selection method in the SAS software is useful in selecting the best two sub-sets under each variable subgroup and for estimating the C_p and the AIC statistics. The overall model fit plot illustrate the degree of prediction in MLR. The explained variation plot in MLR illustrates the partitioning of the total SS to model and error sums of squares. The receiver operating characteristic curve (ROC) is a graphical display of *sensitivity* versus *1-specificity* illustrating the predictive accuracy of the logistic regression model. The scree plot in PCA and factor analysis is useful in selecting the significant principal components and factors. Bi-plot display of both component (PC, factor, canonical discriminant scores) scores and factor loadings is very effective in studying the relationships within observations, between variables, and the inter-relationship between observations and variables in unsupervised learning methods.

3.1 Model selection in MLR

The $C(p)$ plot (Figure 7) shows the Mallows $C(p)$ statistic against the number of predictor variables for the full model and the best two models for each subset. The Mallows $C(p)$ measures the total squared error for a subset that equals to total error variance plus the bias introduced by not including the important variables in the subset. Additionally, the root mean squared (RMSE) statistic for the full model and best two regression models in each subset is also shown in the $C(p)$ plot. Furthermore, the diameter of the bubbles in the $C(p)$ plot is proportional to the magnitude of RMSE. Thus, the $C(p)$ plot can be used effectively in selecting the best subset in regression models with many (5 to 25) predictor variables.

3.2 Overall model fit in MLR

The overall model fit is illustrated in Figure 8 by displaying the relationship between the observed response variable and predicted values. The N , R^2 , $R^2(\text{adjusted})$, and RMSE statistics that are useful in comparing regression models and the regression model are also included on the plot. If the data contained replicated observations, the deviation from the model includes both 'pure error' and 'deviation from the regression'. The R^2 estimates can be computed from a regression model using the means of the replicated observations as the response. Consequently, the R^2 computed based on the means ($R^2(\text{mean})$) is also displayed in the title statement. If there is no replicated data, $R^2(\text{mean})$ and the R^2 estimate reported by the PROC REG will be identical.

3.3 The explained variation plot in MLR

Figure 9 shows graphically the total and the unexplained variation in the response variable after accounting for the regression model. The ordered and the centered response variable versus the ordered sequence display the total variability in the response. If the ordered response shows a linear trend without any sharp edges at the both ends then response variable has a normal distribution. The unexplained variability in the response variable is given by the residual distribution. The residual variation shows a random distribution without any sudden peaks, trends or patterns if the regression model assumptions are not violated. The differences between the total and residual variability show the amount of variation in the response accounted for by the regression model and are estimated by the R^2 statistic. The predictive potential of the fitted model can be determined by estimating the $R^2_{(\text{prediction})}$ by substituting 'PRESS (i^{th} deleted residual)' for SSE in the formula for the R^2 estimation. The predictive power of the estimated regression model is considered high if the $R^2_{(\text{prediction})}$ estimate is large and closer to the model R^2 . The estimates of $R^2(\text{mean})$ and the $R^2_{(\text{prediction})}$ described previously are also displayed in the 'title' statement. These estimates and the graphical display of explained and unexplained variation help to judge the quality of the model fit.

3.4 The c statistic and ROC curve in BLR:

The ROC curve is constructed by plotting the *sensitivity* (measure of accuracy of predicting events) versus *1-specificity* (measure of error in predicting non-events). The area under the ROC curve is a measure of the classification power of the logistic equation. It varies from 0.5 (the model's predictions are no better than chance) to 1.0 (the model always assigns higher probabilities to correct cases than to incorrect cases). Thus *c statistic* is the percent of all possible pairs of cases in which the model assigns a higher probability to a correct case than to an incorrect case. The area under the ROC curve is equal to the *c-statistic*. The ROC curve rises quickly and the area under the ROC is larger for model with high predictive accuracy. See an example of ROC curve in Figure 10.

3.5 Scree plot in principal component and exploratory factor analysis

In the PCA analysis, the dimensions of standardized multi-attributes define the number of eigenvalues. An eigenvalue greater than 1 indicates that PC accounts for more of the variance than one of the original variables in standardized data. This can be confirmed by visually examining the improved scree plot (Figure 11) of eigenvalues and the parallel analysis of eigenvalues. This enhanced scree plot shows the rate of change in the magnitude of the eigenvalues for an increasing number of PC. The rate of decline levels off at a given point in the scree plot that indicates the optimum number of PC to extract. Also, the intersection point between the scree plot and the parallel analysis plot reveals the optimum number of principal components that could be retained as the significant PC.

3.6 Applications of Bi-plots in un-supervised learning methods

The highlight of presenting the findings of the unsupervised learning methods is studying the bi-plots. In order to display the relationships among the variables, the factor loading for each factor is overlaid on the same plot after being multiplied by the corresponding maximum value of factor score. For example, factor1 loading values are multiplied by the maximum value of

factor1 score, and factor2 loadings are multiplied by the maximum value of factor2 scores. This transformation places both the variables and the observations on the same scale in the bi-plot display since the range of factor loadings are usually shorter (-1 to $+1$) than the factor scores. The correlations among the multivariate attributes used in the factor analysis are revealed by the angles between any two factor loading vectors. For each variable, a factor loading vector is created by connecting the origin (0,0) and the multiplied value of factor1 and factor2 loadings on the bi-plot. The angles between any two variable vectors will be 1) narrower ($< 45^{\circ}$) if the correlations between these two attributes are positive and larger. See an example of bi-plot in Figure 12.

4. Regression diagnostic plots for detecting violations of statistical assumptions

Multiple linear regression models are fairly robust against violation of non-normality especially in large samples. Signs of non-normality are significant skewness (lack of symmetry) and/or kurtosis light-tailedness or heavy-tailedness. The normal probability plot (Figure 13-normal Q-Q plot), along with the normality test statistics, can provide information on the normality of the residual distribution. A fan pattern like the profile of a megaphone, with a noticeable flare either to the right or to the left in the residual plot against predicted value is the indication of significant heteroscedasticity. The Breusch-Pagan test based on the significance of linear model using the squared absolute residual as the response and all combination of variables as predictors is recommended for detecting heteroscedasticity. However, the presence of significant outliers and non-normality may confound with heteroscedasticity and may interfere with the detection. The results of the 'Breusch-Pagan' test and the random pattern of the residuals in the residual plot (Figure 13) both can confirm if the residuals have equal variance. Observations used in the regression modeling are identified as outliers if the absolute STUDENT value exceeds 2.5. Also, observations are identified as influential if the 'DFFITS' statistic value exceeds 1.5. An outlier detection bubble plot between '*student*' and *hat value* identifies the outliers if they falls outside the 2.5 boundary line and detects influential points if the diameter of the bubble plot, which is proportional to DFFITS is relatively big (Figure 13).

5. Model validation

Regression model estimated using the training dataset could be validated by applying the model to an independent validation data and by comparing the model fit. If both models produce similar R^2 and show comparable predictive models, then the estimated regression model could be used for prediction with reasonable accuracy. Model validation could be further strengthened if both training and the validation residual plots show similar pattern. See Fernandez (2002a) for examples of comparing prediction and residual pattern between the training and the validation datasets in multiple linear regression.

6. User-friendly SAS macro applications

The data mining techniques described above can be performed easily by running the SAS data mining macro applications available in the CD-ROM (Fernandez 2002 b). The user-friendly SAS macro applications integrates the statistical and graphical analysis tools available in SAS

systems and provides complete data mining solutions without writing SAS program codes or using the point-and-click approach. Step-by-step instructions for using the SAS macro and interpreting the results are emphasized (Fernandez 2002 a). Thus, by following the step-by-step instructions and downloading the user-friendly SAS macros described in the book, data analysts can perform regression diagnostics quickly and effectively.

7. Summary

The data mining statistical graphical techniques for detecting influential outliers, non-linearity, and multicollinearity using augmented partial residual, partial regression leverage and overlaid augmented partial residual and leverage, VIF PLOT, model selection plot using Cp statistic, plots showing model fit, and explained variation, heteroscedasticity, influential outliers, and departure from normality in multiple linear regression; simple and delta logit plots, ROC curve in binary logistic regression; Scree plot and bi-plot display in principal component and factor analysis are presented here. The instructions for generating these plots using user-friendly SAS macro applications and the instructions for obtaining the macro are reported elsewhere (Fernandez, 2002a).

8. References

1. Belsley, D.A., Kuh, E. and Welsch, R.E. 1980. Regression diagnostics. N.Y. John Wiley.
2. Cook, R.D. And Weisberg, S. (1982) Residuals and Influence in Regression. N.Y. Chapman and Hall.
3. Fernandez, G.C.J 2002a Data mining using SAS applications CRC/Chapman-Hall Publications FL <http://www.ag.unr.edu/gf/dm.html>
4. Fernandez, G.C.J 2002b Data mining using SAS applications - CDROM CRC/Chapman-Hall Publications FL <http://www.ag.unr.edu/gf/dm.html>
5. Larsen W.A. and McCleary S.J. 1972 The use of partial residual plots in Regression analysis. Technometrics 14: 781-790.
6. Mallows, C. L. 1986. Augmented partial residual Technometrics 28: 313-319
7. Mason, R. L. , Gunst, R.F. and Webster, J.T. 1975. Regression analysis and problem of multicollinearity. Commun. Statistics. 4(3): 277-292.
8. Myers, R.H. 1990. Classical and modern regression application. 2nd edition. Duxbury press. CA.
9. Neter, J. Wasserman, W., and Kutner, M.H. 1989. Applied Linear regression Models. 2nd Edition. Irwin Homewood IL.
10. Sall, J 1990. Leverage plots for general linear hypothesis. The Amer. Statistician. Vol.44. 308-315
11. SAS Institute Inc. Data Mining Using Enterprise Miner Software: A Case Study Approach First edition 2000 Carry NC USA.
12. Stine R A. 1995. Graphical Interpretation of Variance Inflation Factors. The American Statistician vol 49: 53-56.

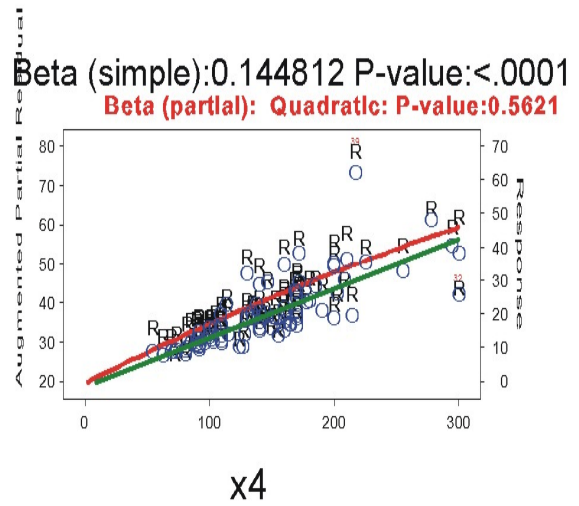
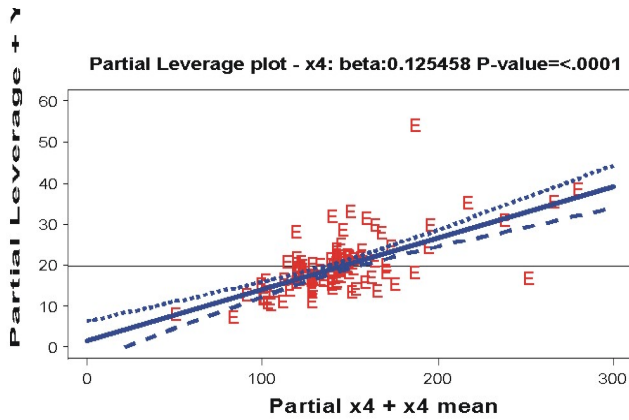
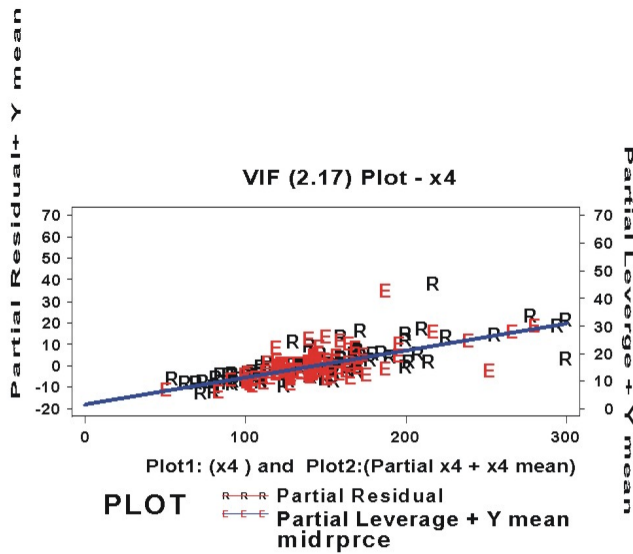


Figure 1 Partial Leverage Plot

Figure 2 Augmented Partial Residual Plot



Logit & Partial delta Logit plots Resp=mpg vs x15
 Chi-square P-value: Linear x15:0.2798 Quadratic x15:0.7963 x15-Linear Odds ratio =0.9987

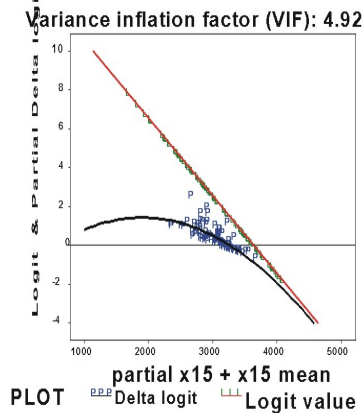


Figure 3 VIF plot

Figure 4. Partial delta logit plot

Testing for significant interaction: $x_4 * x_{15}$ P-value=0.2946

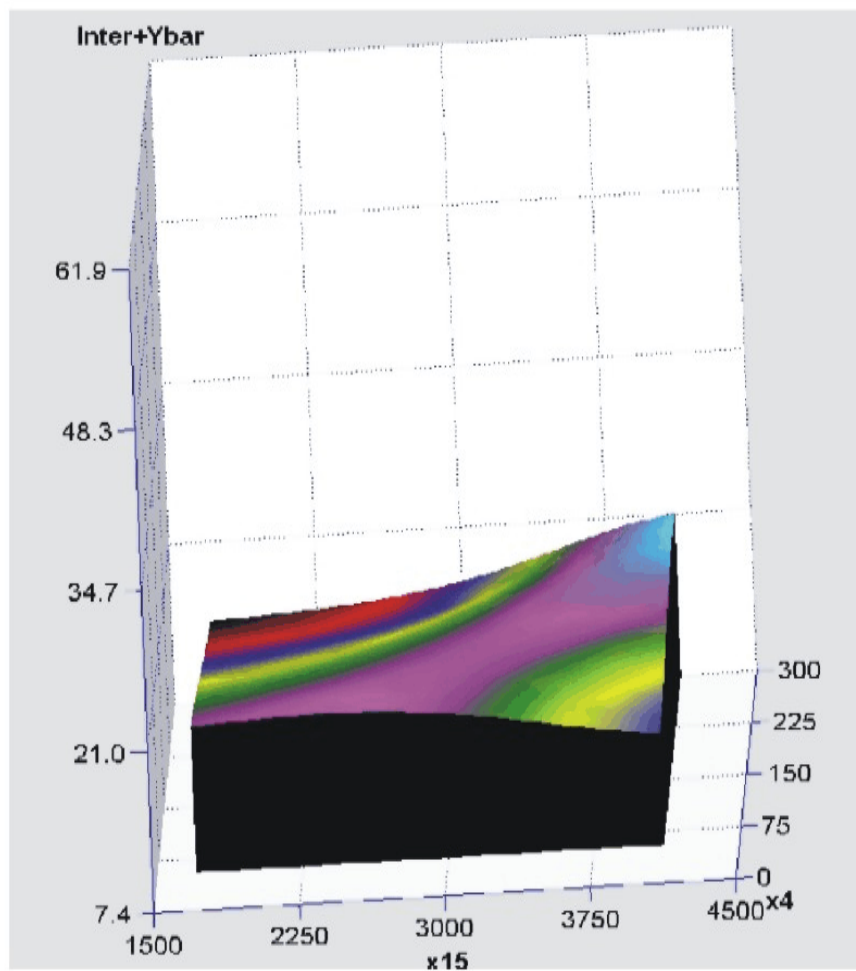


Figure 5. Interaction detection plot in multiple linear regression

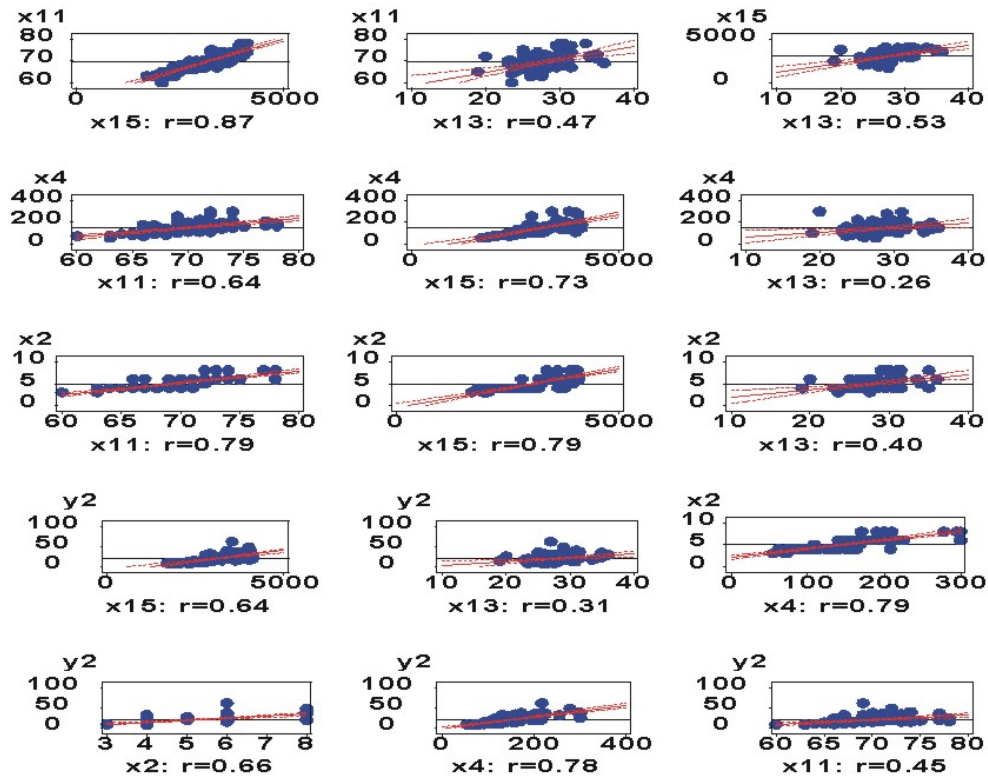


Figure 6 Scatter plot matrix

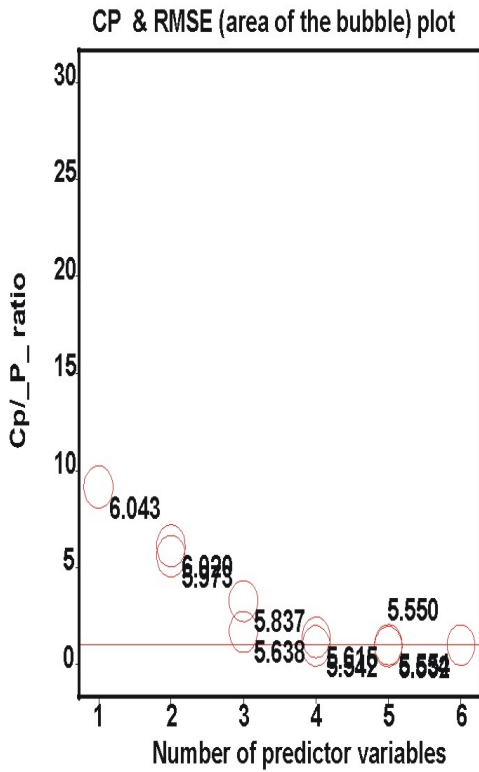


Figure 7 Cp-Model Selection Plot

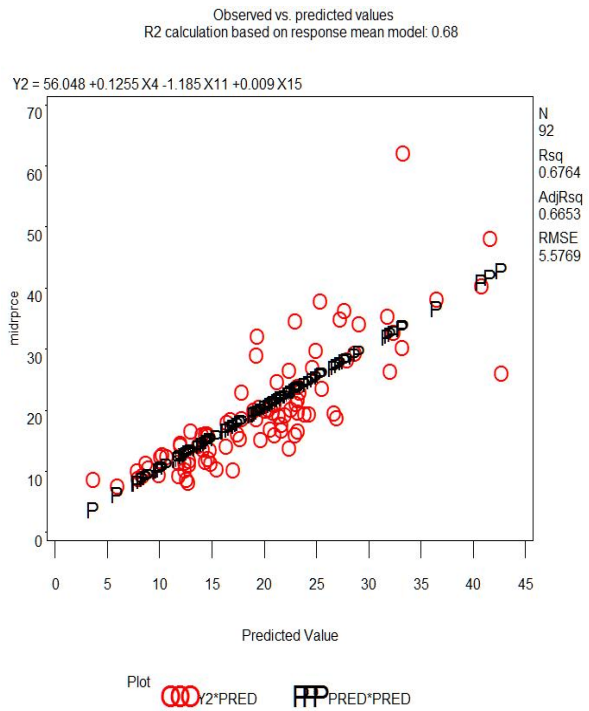


Figure 8 Regression model fit plot

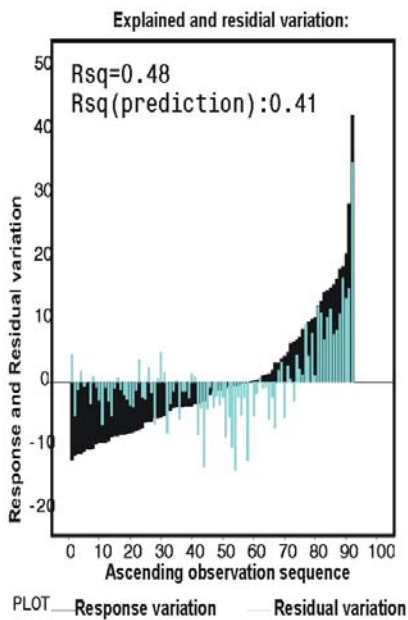


Figure 9 Explained variation plot

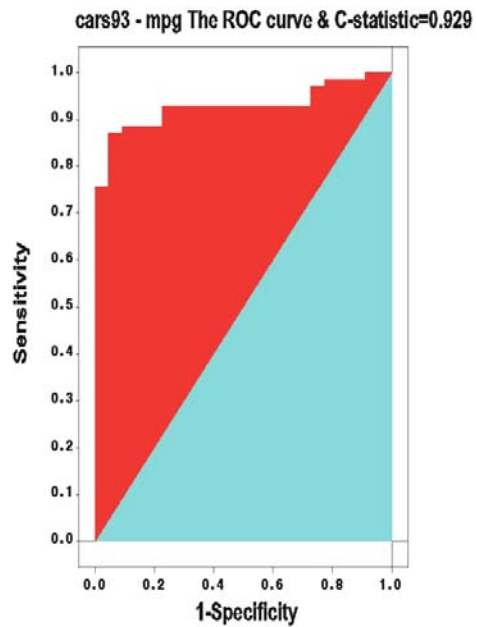


Figure 10 ROC curve

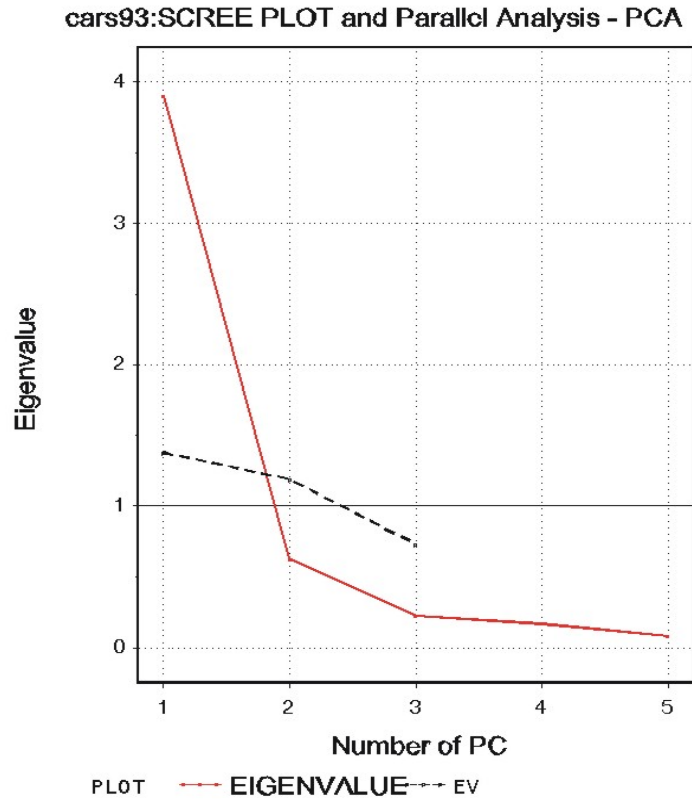
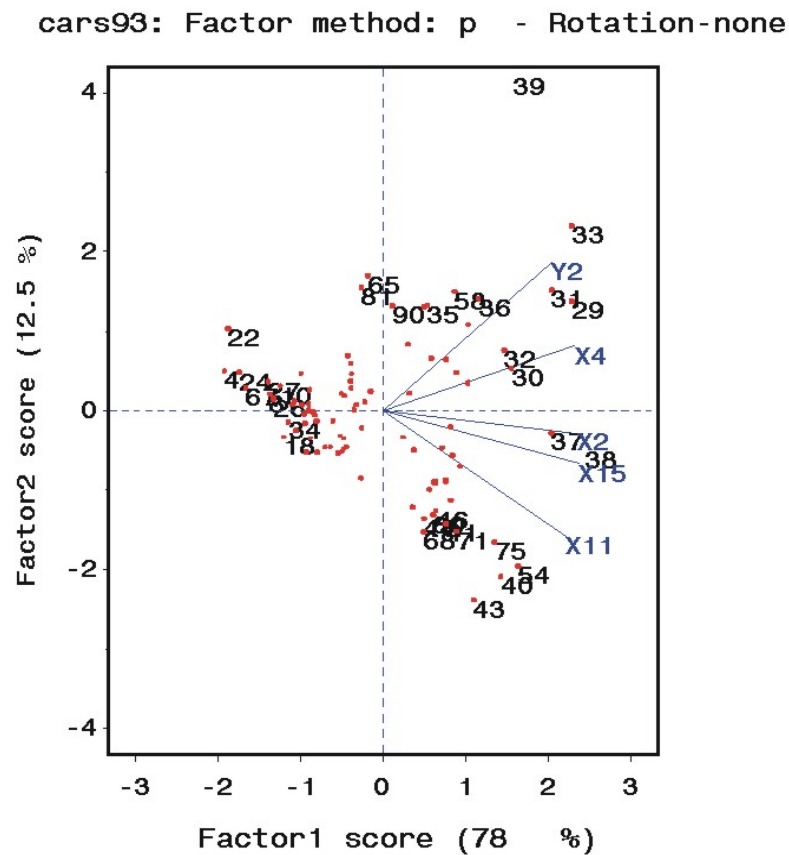


Figure 11 Scree plot in Exploratory factor analysis



New Prairie Press
 Figure 12 Bi-plot display of factor scores and loadings
<https://newprairiepress.org/agstatconference/2004/proceedings/2>

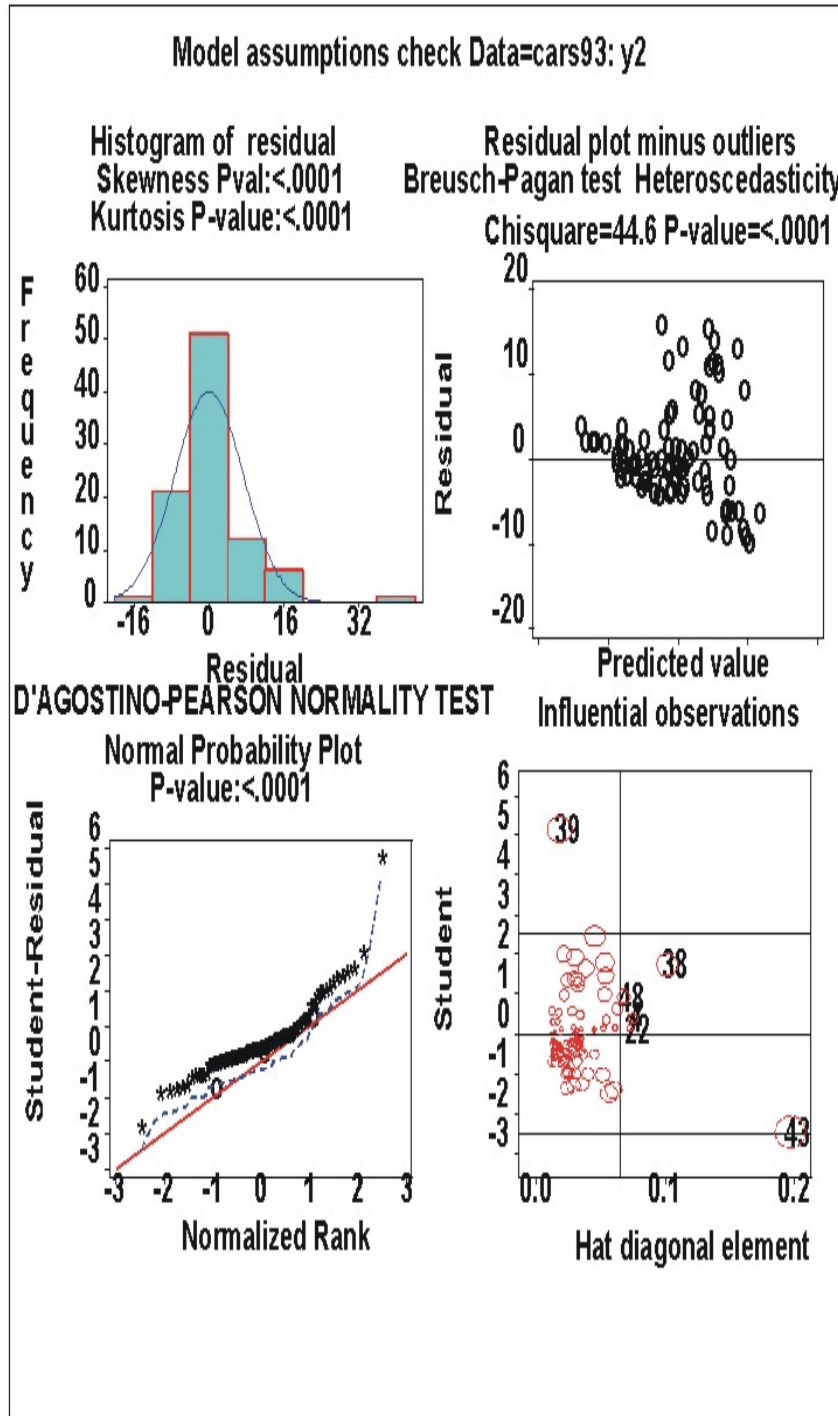


Figure 13 Checking for model violations in multiple linear regression