A Simple Plot for Model Assessment

David J. Olive[*]

Southern Illinois University

September 16, 2005

Abstract

Regression is the study of the conditional distribution $y|\boldsymbol{x}$ of the response y given the predictors $\boldsymbol{x}$. Consider models of the form $y = m(\boldsymbol{x}) + e$. Many anova, categorical, multi–index, nonlinear regression, nonparametric regression, semiparametric and time series models have this form. Assume that the model is a good approximation of the data and that $\hat{m}$ is a good estimator of m. Then the plotted points in a *fit response plot* of $\hat{m}$ versus y will scatter about the identity line. Applications of this plot include checking the goodness of fit of the model and detecting influential cases and outliers.

KEY WORDS: Diagnostics; Dimension Reduction; Goodness of Fit; Outliers; Regression; Single Index Models.

[*] David J. Olive is Associate Professor, Department of Mathematics, Mailcode 4408, Southern Illinois University, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This research was supported by NSF grant DMS 0202922.

1 INTRODUCTION

Regression is the study of the conditional distribution $y|\boldsymbol{x}$ of the response y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. Many important statistical models have the form

$$y_i = m(\boldsymbol{x}_i) + e_i \equiv m_i + e_i \tag{1.1}$$

for $i = 1, ..., n$ where the zero mean error e_i is independent of $\boldsymbol{x}_i$. Additional assumptions on the errors are often made. Let $\hat{m}$ be an estimator of m, let the ith predicted or fitted value $\hat{y}_i = \hat{m}_i = \hat{m}(\boldsymbol{x}_i)$, and let the ith residual $r_i = y_i - \hat{y}_i$.

The above class of models is very rich. Many anova, categorical, nonlinear regression, nonparametric regression, semiparametric and time series models have this form. An additive error *single index model* uses

$$y = h(\boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{1.2}$$

The *multiple linear regression model* is an important special case where h is the identity function: $h(\boldsymbol{\beta}^T \boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$. See Härdle, Hall, and Ichimura (1993), Naik and Tsai (2001), Simonoff and Tsai (2002), Stoker (1986) and Weisberg and Welsh (1994). An additive error *multi–index model* has the form

$$y = h(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}) + e \tag{1.3}$$

where $k \geq 1$ is as small as possible. See Hristache, Juditsky, Polzehl, and Spokoiny (2001).

Another important special case of model (1.1) is the *response transformation model* where

$$z_i \equiv t^{-1}(y_i) = t^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_i + e_i)$$

and thus

$$y_i = t(z_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i. \tag{1.4}$$

Koenker and Geling (2001) note that if z_i is an observed survival time, then many *survival models* have the above form, including the Cox (1972) *proportional hazards model.*

There are several important regression models that do not have additive errors including generalized linear models. If

$$y = g(\boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{1.5}$$

then the regression has 1–dimensional structure while

$$y = g(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}, e) \tag{1.6}$$

has k–dimensional structure if $k \geq 1$ is as small as possible. These models do not necessarily have additive errors although the additive error single index and multi–index models are important exceptions. See Li and Duan (1989), Li (1991, 2000), Cook (1998, p. 49), Cook and Weisberg (1999, p. 414) and Cook and Ni (2005).

Several authors have suggested that plotting $\boldsymbol{a}^T \boldsymbol{x}$ versus y for various choices of $\boldsymbol{a}$ is useful for model assessment. Cook and Weisberg (1997, 1999, ch. 17) call this plot a *model checking plot.* In particular, plot each predictor x_j versus y, and also plot $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus y if model (1.5) holds. Suppose that ordinary least squares (OLS) is used to fit the multiple linear regression model, then the *forward response plot* of $\hat{y}$ versus y can be used to visualize the coefficient of determination R^2. See, for example, Chambers, Cleveland, Kleiner and Tukey (1983, p. 280). For models of the form (1.5), an *estimated sufficient summary plot* (ESSP) is a plot of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus y, and is used to visualize the conditional distribution of $y|\boldsymbol{x}$. An amazing result is that if the OLS fitted values $\hat{y}_{OLS}$ are produced from regressing y on $\boldsymbol{x}$, then the plot of $\hat{y}_{OLS}$ versus y is an ESSP and can be often used to visualize g even if g is unknown or misspecified. See Brillinger (1977, 1983), Li and Duan (1989) and Cook and Weisberg (1999, p. 432).

Residual plots are also used for model assessment, but residual plots emphasize lack of fit and can not be used to visualize $y|\boldsymbol{x}$. Cook and Weisberg (1997) discuss why plots such as the ESSP and model checking plot that emphasize the goodness of fit are needed.

In addition to the usual model checks, we suggest making a plot of $\hat{m}$ versus y, called a *fit response plot* or *FY plot*. To understand the information contained in the FY plot, first consider a plot of m_i versus y_i. Ignoring the error in the model $y_i = m_i + e_i$ gives $y = m$ which is the equation of the *identity line* with unit slope and zero intercept. The vertical deviations from the identity line are $y_i - m_i = e_i$. The reasoning for the FY plot is very similar. The line $y = \hat{m}$ is the identity line and the vertical deviations from the line are the residuals $y_i - \hat{m}_i = y_i - \hat{y}_i = r_i$. Suppose that the model $y_i = m_i + e_i$ is a good approximation to the data and that $\hat{m}$ is a good estimator of m. If the identity line is added to the plot as a visual aid, then the plotted points will scatter about the line and the variability of the residuals can be examined.

For a given data set, it will often be useful to generate the FY plot, residual plots, and model checking plots. An advantage of the FY plot is that if the model is not a good approximation to the data or if the estimator $\hat{m}$ is poor, then detecting deviations from the identity line is simple. Similarly, residual variability is easier to judge against a line than a curve. On the other hand, model checking plots such as the ESSP may provide information about the form of the conditional mean function $E(y|\boldsymbol{x}) = m(\boldsymbol{x})$.

Many numerical diagnostics for detecting outliers and influential cases on the fit have been suggested, and often this research generalizes results from Cook (1977, 1986) to various models of form (1.1). Information from these diagnostics can be incorporated into the FY plot by highlighting cases that have large values of the diagnostic. Section 2 provides some examples illustrating applications of the FY plot.

2 Examples

When the bulk of the data follows model (1.1), the following *rules of thumb* for the FY plot are useful for finding influential cases and outliers. 1) Look for points with large absolute residuals and for points far away from $\overline{y}$. 2) Also look for gaps separating the data into clusters.

Rule 1) suggests that the FY plot can be viewed as a graphical analog of Cook's distance for the model. If the multiple linear regression (MLR) model is used along with Cook's distance D_i (Cook 1977), assume that OLS is used to fit the model and to make the FY plot $\hat{y}$ versus y. Then D_i tends to be large if $\hat{y}_i$ is far from the sample mean $\overline{y}$ and if the corresponding absolute residual $|r_i|$ is not small. If $\hat{y}_i$ is close to $\overline{y}$ then D_i tends to be small unless $|r_i|$ is large. An exception to these rules of thumb occurs if a group of cases form a cluster and the OLS fit passes through the cluster. Then the D_i's corresponding to these cases tend to be small even if the cluster is far from $\overline{y}$.

To see why gaps are important, suppose that OLS was used to obtain $\hat{y} = \hat{m}$. Recall that the squared correlation r^2 between y and $\hat{y}$ is equal to the coefficient of determination R^2. Even if an alternative MLR estimator is used, r^2 over emphasizes the strength of the

MLR relationship when there are two clusters of data since much of the variability of y is due to the smaller cluster. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{y}_w$ versus y using the entire data set. If the identity line passes through the bulk of the data but not the cluster, then the cluster points may be outliers.

Now suppose that the MLR model is incorrect. If OLS is used in the FY plot, then the plot is also an ESSP, and if $y = g(\boldsymbol{\beta}^T \boldsymbol{x}, e)$, then the plot can be used to visualize g for many data sets. Hence the plotted points may be very far from linear. The plotted points in FY plots created from other MLR estimators may not be useful for visualizing g, but will also often be far from linear.

An advantage of the FY plot over numerical diagnostics is that while it depends strongly on the model m, defining numerical diagnostics for different fitting methods can be difficult while the FY plot is simply a plot of $\hat{m}$ versus y. For the MLR model, the FY plot can be made from any good MLR estimator such as OLS.

The most important example is the MLR model. For this model, the FY plot is the forward response plot. If the MLR model holds and the errors e_i are iid with zero mean and constant variance σ^2, then the plotted points should scatter about the identity line with no other pattern.

Example 1. Buxton (1920) gives 20 measurements (in mm) of 88 men. We chose to predict stature using an intercept, head length, nasal height, bigonal breadth and cephalic index. Observation 9 was deleted since it had missing values. Five individuals, numbers 62-66, were reported to be about 0.75 inches tall with head lengths well over five feet! For these cases, stature was incorrectly recorded as head length and 18 or 19 mm given for stature, making the cases massive outliers with enormous leverage. Figure 1 shows the forward response plot and residual plot based on the OLS MLR model. Although an index plot of Cook's distance D_i may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the FY plot with $D_i > \min(0.5, 2p/n)$ were highlighted. Notice that the OLS fit passes through the outliers, but the FY plot is resistant to y-outliers since y is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{y}$, only two of the outliers had large Cook's distance. FY plots using other MLR estimators such as `lmsreg` were similar. The Buxton data set can be downloaded from the web site (www.math.siu.edu/olive/ol-bookp.htm).

The above example is typical of many "benchmark" outlier data sets for MLR robust regression. In these data sets traditional OLS diagnostics such as index plots of Cook's distances and residuals often fail to detect the outliers, but the combination of the FY plot and residual plot is usually able to detect the outliers.

Two important models can be transformed into the MLR model $y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ where the iid errors have constant variance $\text{Var}(e_i) = \sigma^2$. The first model is the response transformation model (1.4). If the transformation $y_i = t(z_i)$ is correct, then the plotted points $\hat{y}_i$ versus y_i in the FY plot will follow the identity line. Cook and Olive (2001) use a similar idea and provide several examples.

The second model is the generalized linear squares (GLS) model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Following Seber and Lee (pp. 66-68), assume $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix. Hence there is a nonsingular $n \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. The OLS model uses $\boldsymbol{V} = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{V}$ is diagonal for weighted least squares. Let $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$. Then

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. Let $\hat{\boldsymbol{\beta}}_{OLS}$ be the OLS estimator of $\boldsymbol{\beta}$ obtained from the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and let $\hat{\boldsymbol{\beta}}_{GLS}$ be the GLS estimator. Recall that $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression of $\boldsymbol{Z}$ on $\boldsymbol{U}$ (without an intercept). Let $\hat{\boldsymbol{Y}}_{OLS}$ and $\hat{\boldsymbol{Y}}_{GLS}$ denote the vectors of OLS and GLS fitted values.

Several plots could be made to check the GLS model. The plotted points of $\hat{y}_{i,OLS}$ versus y_i will follow the identity line, but the variability will depend on the value of $\hat{y}_{i,OLS}$. Moreover, we cannot check whether $\boldsymbol{V}$ is providing a useful model. A solution is to make the OLS forward response and residual plots for the OLS regression of $\boldsymbol{Z}$ on $\boldsymbol{U}$. If the GLS model is a useful approximation to the data, then OLS model in $\boldsymbol{Z}$ and $\boldsymbol{U}$ should be appropriate. Hence the plotted points $\hat{z}_i$ versus z_i in the forward response plot should follow the identity line in an evenly populated band while the plotted points $\hat{z}_i$ versus $r_{z,i} = z_i - \hat{z}_i$ in the residual plot should follow the line $r_{z,i} = 0$ in an evenly populated band.

An *FF plot* is a scatterplot matrix of the response and the fitted values from several models. The following example shows that the FF plot can be useful for comparing the various models. Olive and Hawkins (2005) used similar ideas to compare competing submodels found from numerical methods for variable selection.

Example 2. The lynx data is a well known time series of $n = 114$ cases concerning the number z_t of lynx trapped in a section of Northwest Canada from 1821 to 1934. Following Lin and Pourahmadi (1998), let the response $Y_t = \log_{10}(z_t)$. Moran (1953) suggested the autoregressive AR(2) model $\hat{Y}_t = 1.05 + 1.41 Y_{t-1} - .77 Y_{t-2}$. Tong (1977) suggested the AR(11) model $\hat{Y}_t = 1.13 Y_{t-1} - .51 Y_{t-2} + .23 Y_{t-3} - .29 Y_{t-4} + .14 Y_{t-5} - .14 Y_{t-6} + .08 Y_{t-7} - .04 Y_{t-8} + .13 Y_{t-9} + .19 Y_{t-10} - .31 Y_{t-11}$. Brockwell and Davis (1991, p. 550) suggested the AR(12) model $\hat{Y}_t = 1.123 + 1.084 Y_{t-1} - .477 Y_{t-2} + .265 Y_{t-3} - .218 Y_{t-4} + .180 Y_{t-9} - .224 Y_{t-12}$. Tong (1983) suggested the following two self–exciting autoregressive models. The SETAR(2,7,2) model uses $\hat{Y}_t = .546 + 1.032 Y_{t-1} - .173 Y_{t-2} + .171 Y_{t-3} - .431 Y_{t-4} + .332 Y_{t-5} - .284 Y_{t-6} + .210 Y_{t-7}$ if $Y_{t-2} \leq 3.116$ and $\hat{Y}_t = 2.632 + 1.492 Y_{t-1} - 1.324 Y_{t-2}$, otherwise. The SETAR(2,5,2) model uses $\hat{Y}_t = .768 + 1.064 Y_{t-1} - .200 Y_{t-2} + .164 Y_{t-3} - .428 Y_{t-4} + .181 Y_{t-5}$ if $Y_{t-2} \leq 3.05$ and $\hat{Y}_t = 2.254 + 1.474 Y_{t-1} - 1.202 Y_{t-2}$, otherwise.

The FF plot and the lynx data is an interesting example for teaching alternative time series models, measures of model adequacy and diagnostics for outliers and influential cases. The FF plot shown in Figure 2 used the fitted values and Y_t for $13 \leq t \leq 114$. The top row of the FF plot gives the FY plots of the five different models. Note that the fitted values from the AR(11) and AR(12) models were very similar, as were those from the two SETAR models. The correlations between Y_t and the fitted values were 0.9099, 0.9056, 0.8871, 0.8226 and 0.8739 for the AR(2), AR(11), AR(12), SETAR(2,7,2) and SETAR(2,5,2) models, respectively. A common way to judge time series models is to leave out the last k cases and then see how well the model predicts these cases. The FY plot should be examined to see whether the plotted points follow the identity line and whether the fitted model systematically fits some values of Y_t better than others (e.g., sometimes larger values of t are fit better than smaller values).

The fact that the FY plot is useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. The FY plot is not limited to models of the form (1.1). The plot can be made as long as fitted values $\hat{m}$ can be obtained from the model. If $\hat{m}_i \approx y_i$ for $i = 1, ..., n$ then the plotted points will scatter about the identity line.

For multiple linear regression, the forward response plot and the residual plot of fitted values versus residuals should always be made. The plotted points in the forward response plot should be linear and the plotted points in the residual plot should cluster about the horizontal axis with an ellipsoidal or rectangular pattern with no trend. Otherwise the assumptions needed for OLS inference fail to hold.

The FY plot should be used to complement rather than to replace existing techniques for model assessment and exploratory data analysis. For example, scan the raw data and make a scatterplot matrix of the predictors $\boldsymbol{x}$ and the response y to discover gross outliers and useful information about the conditional distribution of $y|\boldsymbol{x}$. The literature for goodness of fit tests is massive. See, for example, references in Kauermann and Tutz (2001) and Stute and Zhu (2005).

3 References

Brillinger, D.R. (1977), "The Identification of a Particular Nonlinear Time Series," *Biometrika,* 64, 509-515.

Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

Brockwell, P.J., and Davis, R.A. (1991), *Time Series: Theory and Methods*, Springer–Verlag, NY.

Buxton, L. H. D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston.

Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R.D. (1986), "Assessment of Local Influence," *Journal of the Royal Statistical Society,* B, 48, 133-169.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regression Through Graphics,* John Wiley and Sons, Inc., NY.

Cook, R.D., and Ni, L. (2005), "Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach," *Journal of the American Statistical Association*, 100, 410-428.

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Cook, R.D., and Weisberg, S. (1997), "Graphs for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* John Wiley and Sons, Inc., NY.

Cox, D.R. (1972) "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single Index Models," *The Annals of Statistics,* 21, 157-178.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics,* 29, 1537-1566.

Kauermann, G., and Tutz, G. (2001), "Testing Generalized Linear and Semiparametric Models Against Smooth Alternatives," *Journal of the Royal Statistical Society, B*, 63, 147-166.

Koenker, R., and Geling, O. (2001), "Reappraising Medfly Longevity: a Quantile Regression Survival Analysis," *Journal of the American Statistical Association,* 96, 458-468.

Li, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association,* 86, 316-342.

Li, K.C. (2000), *High Dimensional Data Analysis via the SIR/PHD Approach,* Unpublished Manuscript Available from (http://www.stat.ucla.edu/ kcli/).

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Lin, T.C., and Pourahmadi, M. (1998), "Nonparametric and Nonlinear Models and Data Mining in Time Series: A Case-Study on the Canadian Lynx Data," *Journal of the Royal Statistical Society, C,* 47, 187-201.

Moran, P.A.P (1953), "The Statistical Analysis of the Sunspot and Lynx Cycles," *Journal of Animal Ecology*, 18, 115-116.

Naik, P.A., and Tsai, C. (2001), "Single-Index Model Selections," *Biometrika,* 88, 821-832.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., John Wiley and Sons, NY.

Simonoff, J.S., and Tsai, C. (2002), "Score Tests for the Single Index Model," *Technometrics,* 44, 142-151.

Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

Stute, W. and Zhu, L. (2005), "Nonparametric Checks for Single-Index Models," *The Annals of Statistics,* 33, 1048-1084.

Tong, H. (1977), "Some Comments on the Canadian Lynx Data," *Journal of the Royal Statistical Society, A*, 140, 432-468.

Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, 21, Springer–Verlag, Heidelberg.

Weisberg, S., and Welsh, A.H. (1994), "Adapting for the Missing Link," *The Annals of Statistics,* 22, 1674-1700.

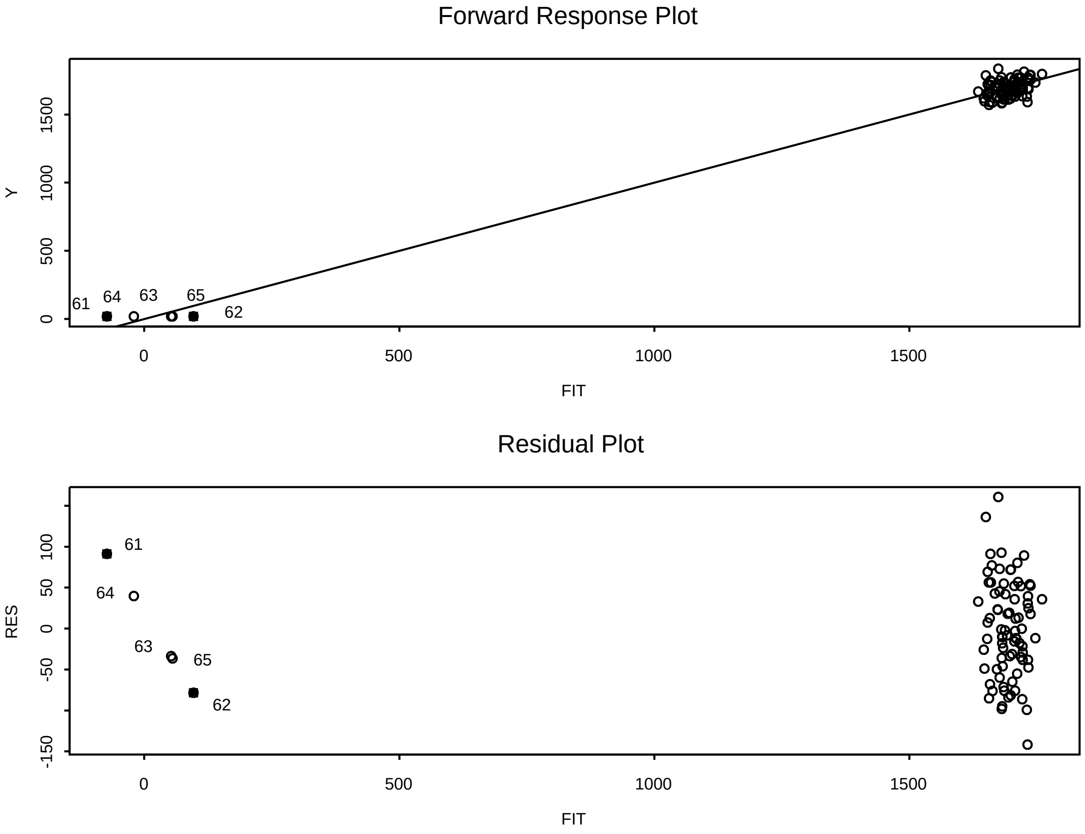

Figure 1: Plots for Buxton Data

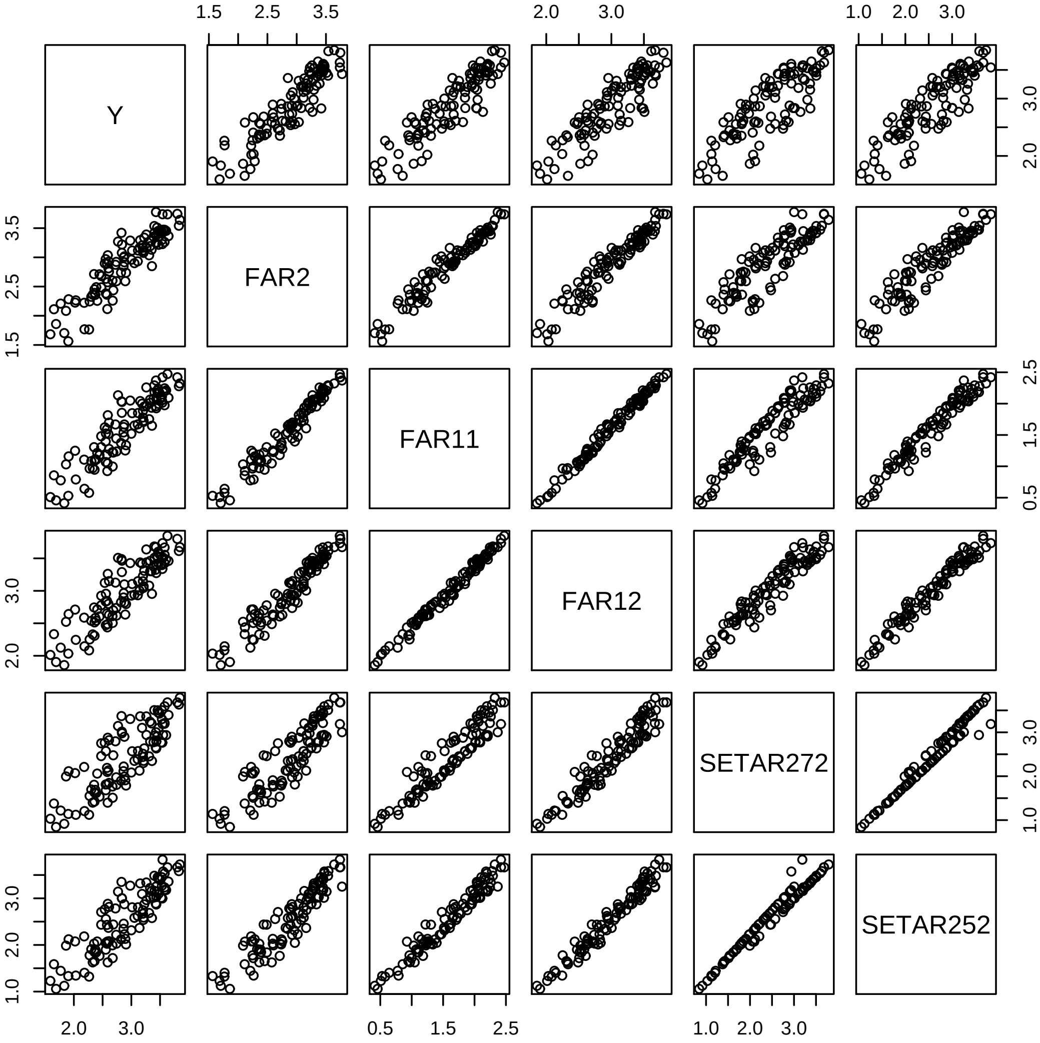

Figure 2: FF Plot for Lynx Data