Response Plots for Linear Models

David J. Olive *

Southern Illinois University

June 9, 2010

Abstract

For linear models, the response and residual plots can be used to check the model for goodness and lack of fit. The plots can also be used to visualize the linear model in the background of the data, to check for linearity or nonlinearity, constant variance or nonconstant variance, to choose a response transformation and to compare competing models such as the full model versus a reduced model or least squares versus a robust regression estimator. The plots can be modified to incorporate information from regression diagnostics and robust regression.

KEY WORDS: Outliers, response transformations, robust regression.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (E-mail: *dolive@math.siu.edu*). This work was supported by the National Science Foundation under grant DMS 0600933.

1. INTRODUCTION

The *linear model* is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of response 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 errors. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$, and $Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Multiple linear regression, analysis of covariance and many analysis of variance (anova) models are linear models.

Suppressing the dependence on i, assume that the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ has unknown probability density function (pdf) $f(y - \boldsymbol{x}^T\boldsymbol{\beta})$, a location family with location parameter = sufficient predictor = $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Then $Y|SP$ has pdf $f(y - SP)$, the error distribution has pdf $f(y)$, and the linear model is $Y = SP + e$ with conditional mean function $E(Y|SP) = SP$ and conditional variance function $V(Y|SP)$. The constant variance assumption is $V(Y|SP) \equiv \sigma^2$, and if the error distribution is normal, $e \sim N(0, \sigma^2)$, then $Y|SP \sim N(SP, \sigma^2)$. The estimated sufficient predictor $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} = \hat{Y}$ is the fitted value, and the residual $r = Y - \hat{Y}$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Note that $Y = ESP + r$.

A residual plot can be defined to be a plot of $\boldsymbol{a}^T\boldsymbol{w}$ versus r where $\boldsymbol{a}$ is a known vector and $\boldsymbol{w}$ is a vector of potential predictors. Then a response plot is similar to a residual plot, but uses the response Y on the vertical axis instead of the residual r.

These plots are crucial in the dimension reduction literature. In a 1D regression model, the response Y is independent of the predictors $\boldsymbol{x}$ given the sufficient predictor $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Then the response plot of the estimated sufficient predictor $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$

versus Y is also called an estimated sufficient summary plot, and is used to visualize the 1D regression model in the background of the data. See Cook (1998, p. 10). When there is a single nontrivial predictor x, a scatterplot of x versus Y is a response plot. Brillinger (1983) recognized that the response plot can be used to visualize the conditional mean function $E(Y|SP)$ of a 1D regression model. The response plot has long been used to visualize the coefficient of determination R^2 in multiple linear regression. See Chambers, Cleveland, Kleiner and Tukey (1983, p. 280). Response plots are called marginal model plots and model checking plots by Cook and Weisberg (1997, 1999, p. 396).

Unless otherwise stated, we will consider response and residual plots that use the fitted values $\boldsymbol{x}^T\hat{\boldsymbol{\beta}} = \hat{Y}$ on the horizontal axis. Now $Y = \hat{Y} + r$, and ignoring the residuals gives the line $Y = \hat{Y}$, so the plotted points scatter about the identity line with unit slope and zero intercept. Since the vertical deviations from the identity line are the residuals $r_i = Y_i - \hat{Y}_i$, the response plot simultaneously shows the response, fitted values and residuals. The estimated conditional mean function is the identity line: $\widehat{E(Y|SP)} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.

Let the *iid error model* be the linear model where the zero mean constant variance errors are iid from a unimodal distribution that is not highly skewed. Then the zero mean assumption holds without loss of generality if the linear model contains a constant and $E(e)$ exists. Under the iid error model, if the fitted values take on many values, then the plotted points should scatter about the identity line in a (roughly) evenly populated band while the plotted points in the residual plot should scatter about the $r = 0$ line in a (roughly) evenly plotted band. Deviations from the evenly populated band suggest that something is wrong with the iid error model.

Many anova models satisfy the iid error model, but often the fitted values do not take on many values. Consider the one way anova model with k treatments and $m \geq 5$ replications per treatment. The plotted points still scatter about the identity or $r = 0$ line, but there are k dot plots corresponding to the k treatments. The dot plots should have similar spread and shape if the one way anova model assumptions are reasonable.

If the zero mean errors are iid from a highly skewed unimodal distribution, then the estimated conditional mean function is still the identity line in the response plot. Hence the plotted points still scatter about the identity line and the $r = 0$ line. It may be necessary to add a scatterplot smoother such as lowess as a visual aid. If the scatterplot smoother is close to the identity line and $r = 0$ line in the response and residual plots, respectively, then linearity is reasonable. Cook and Weisberg (1997) suggest a graphical check on the constant variance assumption.

Unless otherwise stated, assume the iid error model is a useful approximation for the data. The response and residual plots are best for $n > 5p$. If $p < n < 2p$, then the linear model is seriously "overfitting" the data, and the two plots may only be useful for detecting large deviations from the linear model. For $n > 5p$, the following four sections use the plots for checking goodness and lack of fit, linearity or nonlinearity, constant or nonconstant variance, for detecting outliers and influential cases, and for comparing competing models such as the full model versus a reduced model or OLS versus a robust regression estimator. The response plots can be used to choose a response transformation. Information from regression diagnostics and robust regression can be incorporated into the plots. The appendix reviews some concerns about robust regression.

2. GOODNESS AND LACK OF FIT

Cook and Weisberg (1997, 1999, pp. 74, 397) suggest using the response plot for assessing the linear model while Olive and Hawkins (2005) suggest that the response and residual plots are crucial for visualizing the model and for examining lack of fit. If the normal error model is appropriate, then $Y|SP \approx N(ESP, \hat{\sigma}^2)$. The response plot suggests that the iid error model is reasonable if the plotted points follow the identity line in an evenly populated band or if the dot plots have similar shape and spread. To use the response plot to visualize the model, mentally examine a narrow vertical strip about $\hat{Y} = \hat{y}$. The cases in the strip have numbers near $\hat{y}$ on average for linear models with $E(e) = 0$. The response plot suggests that linearity is reasonable but the constant variance assumption is not reasonable if the plot is linear and the cases in different narrow bands have clearly different variability.

Example 1. Wood (1973) provides data where the octane number is predicted from 3 feed compositions and the log of a combination of process conditions. The OLS response and residual plots in Figure 1 suggest that the model is linear but the constant variance assumption may not be reasonable. There appear to be three groups of data. Tremendous profit can be gained by raising the octane number by one point, and the two cases with the largest fitted values $\hat{Y} \approx 97$ were of the greatest interest.

Example 2. SAS Institute (1985, p. 126) uses clover data to illustrate the one way anova model. The response variable is the nitrogen content of red clover plants inoculated with six strains of bacteria, and each strain has five replicates. Figure 2 shows the response and residual plots. Moore (2007, p. 634) states that the one way anova F

test is approximately correct if $\max(S_1, ..., S_k) \leq 2 \min(S_1, ..., S_k)$ where S_i is the sample standard deviation of the ith group. Replace the standard deviations by the ranges of the dot plots when examining the response and residual plots. Linearity seems reasonable, but the approximately constant variance assumption may not hold.

Suppose the full model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a useful approximation to the data. To check a submodel or reduced model, Olive and Hawkins (2005) suggest making the response and residual plots for the full and submodel. Also make an RR plot of the residuals from the submodel versus the residuals from the full model, and make an FF plot which replaces the residuals with the fitted values. If the submodel is good, then its response and residual plots will look similar to those of the full model and the plotted points in the FF and RR plots will cluster tightly about the identity line.

Marginal residual plots of the individual predictors versus the residuals are also useful for checking lack of fit. Cook and Weisberg (1997) and Sheather (2009, pp. 193-195) show that marginal response plots are useful for checking goodness of fit.

For 1D regression, including generalized linear models and many survival analysis models, it is well known that near replicates of $\boldsymbol{x}$ can be used to form lack of fit tests. If $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$, then cases corresponding to points in a narrow vertical slice of the response plot are near replicates of $\boldsymbol{x}^T\boldsymbol{\beta}$, and narrow slices of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ tend to contain far more cases than narrow slices of $\boldsymbol{x}$. This point has been missed in much of the literature. See, for example, Miller, Neill and Sherfey (1999).

There is an enormous literature on tests and diagnostics for goodness and lack of fit for linear models. See references in Cheng and Wu (1994), Peña and Slate (2006) and Su and Yang (2006). Diagnostics for heteroscedasticity include Cook and Weisberg (1983,

1997).

3. GENERALIZED LEAST SQUARES

In this section the assumptions of the iid error model are relaxed. The *generalized least squares* (GLS) model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a linear model with $E(\boldsymbol{e}) = \boldsymbol{0}$, but $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix. The *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and $w_i = 1/v_i$. The *GLS estimator*

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{1}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Following Freedman (2005, p. 54), the *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{2}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$.

The GLS estimator can be transformed to a linear model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$. One transformation uses the numerically unstable Cholesky decomposition: there is a nonsingular $n \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. Let $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$.

An alternative transformation uses the spectral theorem and is a special case of the singular value decomposition: there is a symmetric, nonsingular $n \times n$ matrix $\boldsymbol{R}$ such that $\boldsymbol{V} = \boldsymbol{RR}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$. This method has better computational properties than transformation based on the Cholesky decomposition.

The response and residual plots can be made for the transformed model to check the linearity and constant variance assumptions as in the previous section, assuming that the distribution of ϵ is not highly skewed. If the plots are good, then the GLS model may be a reasonable approximation for the data. Similar plots can be made for FGLS since the FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{RR}$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix.

The plots based on the transformed model give both a check on linearity and on whether the model using $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$) gives a good approximation of the data, provided that $n > 5(p + q + 1)$ where $q = 0$ for GLS. Plots based on the GLS residuals and fitted values should be similar to those based on OLS since, under regularity conditions, both OLS and GLS provide consistent estimators of $\boldsymbol{\beta}$. Then the plotted points scatter about the identity and $r = 0$ lines, but usually not in evenly populated bands. Hence these plots can not be used to check whether the GLS model with $\boldsymbol{V}$ is a good approximation to the data.

If the plots for the transformed model show high leverage points or outliers while the response plot based on OLS is linear without outliers, then the GLS model may be poor. Then it may be better to use the consistent but inefficient OLS estimator along with the sandwich estimator.

Sheather (2009, ch. 9, ch. 10) makes the residual plots based on the Cholesky

decomposition and shows that many linear models with serially correlated errors (e.g. AR(1) errors) and many linear mixed models can be fit with FGLS. Houseman, Ryan and Coull (2004) also use the Cholesky decomposition. Montgomery, Peck and Vining (2006, pp. 182-183) make residual plots based on the spectral theorem.

Additional diagnostics are given by Banejee and Frees (1997), Cook and Weisberg (1983), De Gruttola, Ware and Louis (1987), Polasek (1984) and Shi and Chen (2009). Long and Ervin (2000) discuss methods for obtaining standard errors when the constant variance assumption is violated.

4. DETECTION OF OUTLIERS AND INFLUENTIAL CASES

For detection of outliers and influential cases for linear models, including anova models, it is crucial to make the response and residual plots. If $n > 5p$ and the plotted points do not scatter about the identity line and the $r = 0$ line in evenly populated bands, then the iid error model assumptions may not hold. Departures from these ideal shapes are often easily detected when there are one or more groups of outliers even if OLS is used.

Huber and Ronchetti (2009, p. 154) note that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. Fit the model to the bulk of the data. If the fit passes through the cluster, then the cases may be good leverage points, otherwise they may be

outliers.

For multiple linear regression, Rousseeuw and Leroy (1987, p. 8) suggest that regression diagnostics and residuals from robust regression are two methods for detecting outliers. Information from regression diagnostics can be incorporated into the response and residual plots by highlighting cases with large diagnostic values. To summarize the information from two or more regression estimators, make the residual and response plots for each estimator. An RR plot is a scatterplot matrix of the residuals from the different estimators. An FF plot replaces the residuals by the fitted values and includes the response on the top or bottom row of the scatterplot matrix. Note that this row will contain the response plots from the different estimators.

Response and residual plots are rarely shown in the outlier literature for linear models, but the plots are very effective for suggesting that something is wrong with the iid error model. The plots often show two or more groups of data, and outliers often cause an obvious tilt in the residual plot. Influence diagnostics such as Cook's distances CD_i from Cook (1977) and the weighted Cook's distances WCD_i from Peña (2005) are also sometimes useful. In the following example, cases in the plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(\text{WCD}_\text{i})| > 4.5\text{MAD}(\text{WCD}_\text{i})$ are highlighted with crosses, where the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(\text{w}_\text{i})|)$.

Example 3. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over

five feet! In Figure 3, notice that the OLS fit passes through the outliers, but the response plot is resistant to Y–outliers since Y is on the vertical axis. Also notice that only two of the outliers had large Cook's distance and only one case had a large WCD_i.

Figure 4 shows the response plots for OLS, the *R/Splus* resistant estimator `ltsreg`, and for `hbreg` and $\hat{\boldsymbol{\beta}}_{k,B}$ which are described in the appendix. Notice that only the fit from $\hat{\boldsymbol{\beta}}_{k,B}$ (BBFIT) did not pass through the outliers.

The CD_i and WCD_i are the most effective when there is a single cluster about the identity line. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail. For the Wood (1973) data in Figure 1, none of the cases had large CD_i or WCD_i.

Example 4. Dunn and Clark (1974, p. 129) study the effects of four fertilizers on wheat yield using a Latin square design. The row blocks were 4 types of wheat, and the column blocks were 4 plots of land. Each plot was divided into 4 subplots. Case 14 had a yield of 64.5 while the next highest yield was 35.5. For the response plot in Figure 5, note that both Y and $\hat{Y}$ are large for the high yield. Also note that $\hat{Y}$ underestimates Y by about 10 for this case.

Example 5. Snedecor and Cochran (1967, p. 300) give a data set with 5 types of soybean seed. The response frate = number of seeds out of 100 that failed to germinate. Five blocks were used. The response and residual plots in Figure 6 suggest that one case is not fit well by the model. On further examination of the data, there seems to be a block treatment interaction, which is not allowed by the completely randomized block design.

Example 6. Rousseeuw and Leroy (1987, pp. 242-245) give a modified wood data

set with 4 nontrivial predictors and 4 planted outliers. Figure 7 shows an FF plot for the data, using OLS, least absolute deviations (L_1), `lmsreg` (ALMS), `ltsreg` (ALTS) and the Olive (2005) MBA resistant estimator. The four planted outliers have the smallest values of the response, and can not be detected by the OLS response and residual plots. They can be detected by the ALMS and MBA response plots.

No method can detect outliers for all data sets where the bulk of the data follows the iid error model, but the graphical method of using the OLS response and residual plots to detect outliers is competitive with alternative methods. The method works well for both multiple linear regression and anova models. The RR and FF plots can be useful for multiple linear regression when the OLS plots fail.

The literature on detecting outliers and influential cases is massive. Hawkins and Olive (2002) used the RR and residual plots while Olive (2005) suggested using response, residual, FF and RR plots. Olive (2004) highlighted cases with large Cook's distances in the response and residual plots. Other references include Barnett and Lewis (1994), Belsley, Kuh and Welsch (1980), Cook (1977, 1986), Cook and Critchley (2000), Cook and Weisberg (1982), Fox (1991) and Rousseeuw and van Zomeren (1990).

The appendix reviews concerns about robust regression estimators. Also see Carroll, Ruppert and Stefanski (1999), Hawkins and Olive (2002), Huber and Ronchetti (2009), McKean, Sheather and Hettmansperger (1993), Velilla (1998) and Wang and Suter (2003).

5. RESPONSE TRANSFORMATIONS

The applicability of a linear model can be expanded by allowing response transfor-

mations. If $n > 5p$ and the iid error model is a useful approximation for the data for the response transformation $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$, then the plotted points in the response plot will scatter about the identity line in a roughly evenly populated band.

An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i.$$

Two families of transformations are frequently used. Assume that all of the values of the "response" Z_i are positive. A power transformation has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$, and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$ for anova models. For linear models with at least one continuous predictor, add $\pm 1/3$ and possibly $\pm 1/4$ and $\pm 2/3$ to Λ_L. The *modified power transformation family*

$$Y_i = t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda}$$

for $\lambda \neq 0$, and $t_0(Z_i) = \log(Z_i)$ for $\lambda = 0$ where $\lambda \in \Lambda_L$.

There are several reasons to use a coarse grid Λ_L of powers. First, several of the powers correspond to simple transformations such as the log, square root, and reciprocal. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge in probability to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring modified power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

Box and Cox (1964) gave a numerical method for selecting the response transformation for the modified power transformations. Although the method gives a point

estimator $\hat{\lambda}_o$, often an interval of "reasonable values" is generated (either graphically or using a profile likelihood to make a confidence interval), and $\hat{\lambda} \in \Lambda_L$ is used if it is also in the interval.

There are three recent graphical methods for response transformations. Following Olive (2004), compute the fitted values $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response". Then make the transformation plot of $\hat{W}$ versus W for each $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the iid error model is reasonable for $(\hat{W}, W)$. If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, then examine the residual plots, consult subject matter experts and use the simplest or most reasonable transformation. This graphical method is not restricted to OLS.

After selecting the transformation, the usual checks should be made. A variant of the method would plot the residual plot or both the response and the residual plot for each of the values of λ. Residual plots do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55).

The method proposed by Cook and Olive (2001) is similar but replaces $\hat{W}$ by $\hat{Z}$. Cook and Weisberg (1994) show that the inverse response plot of Z versus $\hat{Z}$ can be used to visualize a strictly monotone transformation t. This method needs more assumptions on the predictor distribution, but does not need to restrict $t(Z)$ to a parametric family such as the power transformations.

In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis is for the fitted values that result from using $t_\lambda(Z)$ as the "response" in the software.

Example 7. Kuehl (1994, p. 128) gives data for counts of hermit crabs in six different

coastline habitats, where C is the count of crabs and the "response" $Z = C + 1/6$. Each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. The one way anova model $W_{ij} = t_\lambda(Z_{ij}) = \mu_i + e_{ij} = \eta + \tau_i + e_{ij}$ was fit for $i = 1, ..., 6$ with $n_i = 25$, and $j = 1, ..., n_i$. Each of the six habitats was a level with 25 replicates. Figure 8 shows the five transformation plots. The transformation $Y = \log(Z)$ is used since the six dot plots have roughly the same shape and spread. The transformations $1/Z$ and $1/\sqrt{Z}$ do not handle the 0 counts well, while the transformations $\sqrt{Z}$ and Z have variance that increases with the mean.

Example 2 continued. None of the transformation plots for the clover data removed the nonconstant variance. Thus the one way anova F test should be replaced by a method that can handle nonconstant variance.

Example 8. Box and Cox (1964) analyze data from a 3^3 experiment on the behavior of yarn under cycles of repeated loadings. Here Z = number of cycles until failure while the three predictors are the length, amplitude and load. A constant and the three main effectts were used. For this data set, there is one value of the response for each of the 27 treatment level combinations. Figure 9 shows four of the five transformation plots. The plotted points curve away from the identity line in three of the four plots. The plotted points for the log transformation follow the identity line with roughly constant variance.

This transformation plot is the response plot where $Y = \log(Z)$. To visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose that log(cycles to failure) given fit = 6 is of interest. Mentally examine the plot about a narrow vertical strip about $\hat{Y} = 6$, perhaps from 5.75 to 6.25. The

cases in the narrow strip have a mean close to 6 since they fall close to the identity line. Similarly, when $\hat{Y} = \hat{y}$ for $\hat{y}$ between 4.5 and 8.5, the cases have log(cycles to failure) near $\hat{y}$, on average. Cases 19 and 20 had the largest Y values with long length, short amplitude of loading cycle and low load. Cases 8 and 9 had the smallest Y values with low length, high amplitude and high load.

For experimental design models, interest is often in finding the combination of predictors that result in the largest or smallest values of the response. This example illustrates that the response plot is useful for finding combinations of levels with desirable values of the response.

6. DISCUSSION

The graphical techniques discussed in this paper are useful because they are not tied to a specific estimator such as OLS and can be used even if all of the predictors are categorical. Also, the response and residual plots are easily made with good statistical software packages. Figures 1 and 7-9 were made with *Splus* while Figures 3-6 were made with *R*. Using the response and residual plots speeds up the process of finding a linear model that is a useful approximation of the data, and both plots should be made before performing inference.

Use the response plot to visualize the linear model $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ in the background of the data, to check goodness of fit, to check linearity, to check the strength of the relationship as measured by R^2 or the signal to noise ratio, to detect outliers and to check if a nonlinearity is monotone (useful for response transformations). For anova models with replication, use the plot to check which populations have similar means.

The response plot can be used to show that the linear model provides a simple, useful approximation for the relationship between the response variable Y and the predictors $\boldsymbol{x}$, and to explain the linear model to consulting clients, students or researchers. If the identity line fits the data better than any horizontal line, then the predictors $\boldsymbol{x}$ are needed in the model ($H_A : \boldsymbol{\beta} \neq \mathbf{0}$). If a horizontal line fits the data about as well as the identity line, then the H_A may be statistically significant but not practically significant.

Use the residual plot to check for nonconstant variance, to visualize $e|\boldsymbol{x}^T\boldsymbol{\beta}$, to check for lack of fit, to check for nonlinearity and to check whether the error distribution is highly skewed. For anova models with replication, use the residual plot to judge whether the p dot plots have similar shape and spread. A marginal residual plot of a predictor x_j or a potential predictor w_j versus r is useful to check whether x_j is independent of e, to check if x_j^2 should be added to the model and to check if w_j or both w_j and w_j^2 should be added to the model.

Extensions to other models have been made. The response plot of the ESP versus Y is useful for visualizing 1D regression models in the background of the data while the plot of the fitted values versus Y is useful for checking models, including nonlinear regression and time series models, than may not be 1D regression models. For multivariate regression with k response variables, make the residual and response plots for each of the k variables.

APPENDIX: CONCERNS ABOUT ROBUST REGRESSION

For multiple linear regression, an enormous number of alternatives to OLS have been suggested. Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggest that high breakdown robust regression estimators are unstable, take too long to compute, do

not provide an adequate remedy for the ill effects of outliers, and that their statistical and computational properties are not adequately understood. This appendix further addresses some of these concerns.

If d of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma = d/n$. Then the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. High breakdown regression estimators have $\gamma \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in general position: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. *For the remainder of this appendix, assume that the clean data are in general position.*

The computational complexity of the "brand name" high breakdown estimators is too high. The least trimmed sum of absolute deviations (LTA) and the least median of squares (LMS) estimators have $O(n^p)$ complexity. The least trimmed sum of squares (LTS), least quantile of differences, repeated median and regression depth complexities are far higher, and there may be no known method for computing S, τ, projection based, constrained M and MM estimators. See Rousseeuw and Leroy (1987) for references. Čížek (2006, 2008) showed that LTS and LTA are $\sqrt{n}$ consistent. Kim and Pollard (1990) showed that LMS is $n^{1/3}$ consistent.

Since the above estimators take too long to compute, they are replaced by practical estimators that have not been shown to be both consistent and high breakdown. Often practical "robust estimators" generate a sequence of K trial fits called *attractors*: $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$. Then some criterion is evaluated and the attractor $\boldsymbol{b}_A$ that minimizes the criterion is used as the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $\boldsymbol{b}_{0,j}$ be the jth start and

compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. Then $\boldsymbol{b}_{k,j}$ is the jth attractor for $j = 1, ..., K$. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$, i.e., the attractors are the starts. Elemental starts are the fits from randomly selected "elemental sets" of p cases.

Many criteria for screening the attractors have been suggested. Suppose $c_n \approx n/2$ and $c_n/n \to 1/2$. The LMS(c_n) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c_n) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b})$. The LTA($c_n$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used and concentration is used to produce the attractors, then the best attractor is not consistent if K and k are fixed and free of n. Hence no matter how the attractor is chosen, the resulting estimator is not consistent. The proof is simple given the results of He and Portnoy (1992) who show that if a start $\boldsymbol{b}$ is a consistent estimator of $\boldsymbol{\beta}$, then the attractor is a consistent estimator of $\boldsymbol{\beta}$. Also the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The classical estimator applied to a randomly drawn elemental set is an inconsistent estimator, so the K starts and the K attractors are inconsistent. The final estimator is an attractor and thus inconsistent. The breakdown value of the estimator is bounded above by $K/n \to 0$. If concentration is iterated to convergence so that k is not fixed, then it is not known

whether the attractor is inconsistent, but the breakdown value $\leq K/n$. Hence the widely used elemental concentration algorithm estimators are not high breakdown estimators.

Olive (2005) showed that $\hat{\boldsymbol{\beta}}$ is high breakdown if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$) stays bounded under high contamination. (Notice that if $\|\hat{\boldsymbol{\beta}}\| = \infty$, then median$(|r_i|) = \infty$, and if $\|\hat{\boldsymbol{\beta}}\| = M$ then median$(|r_i|)$ is bounded if fewer than half of the cases are outliers.) Let Q_L denote the LMS, LTS or LTA criterion. Then an estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for d near $n/2$. Concentration insures that Q_{LTS} for the attractor is no larger than Q_{LTS} for the start. Hence concentration applied to a high breakdown start results in a high breakdown attractor.

High breakdown estimators are not necessarily useful for detecting outliers. Suppose $\gamma < 0.5$. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ will be bounded if n is large enough. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but still have small residuals.

Now assume that the multiple linear regression model contains a constant β_1. Let $\boldsymbol{b}_k$ be the attractor from using concentration on the high breakdown start consisting of OLS applied to the c_n cases with Y's closest to the median of the Y_i and let $\hat{\boldsymbol{\beta}}_{k,B} = 0.9999\boldsymbol{b}_k$. Since Olive (2005) showed that the start is a high breakdown estimator, $\hat{\boldsymbol{\beta}}_{k,B}$ is a high breakdown estimator of $\boldsymbol{\beta}$.

Let $\hat{\boldsymbol{\beta}}_C$ be a consistent estimator of $\boldsymbol{\beta}$ while $\hat{\boldsymbol{\beta}}_A$ is an estimator that is useful for outlier detection such as the *R/Splus* `lmsreg` or `ltsreg` estimators. The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Let $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_{k,B}) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_{k,B}$.

It is possible to get a high breakdown estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ using $a = 1$ if the probability that $\hat{\boldsymbol{\beta}}_A$ gets arbitrarily close to $\boldsymbol{\beta}$ as $n \to \infty$ goes to zero. (The multiplier 0.9999 guarantees that $\hat{\boldsymbol{\beta}}_{k,B}$ is biased.) For example, use an elemental concentration algorithm with K and k fixed. However, the sample sizes need to be large before the estimator behaves like $\hat{\boldsymbol{\beta}}_C$. Salibian-Barrera, Willems and Zamar (2008) modified this idea, using $a = 1$ and OLS followed by concentration as $\hat{\boldsymbol{\beta}}_C$, but replaced Q_L by the τ estimator criterion. It is not known whether the resulting "fast τ estimator" is consistent or high breakdown.

The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ uses three attractors: $\hat{\boldsymbol{\beta}}_C$, $\hat{\boldsymbol{\beta}}_{k,B}$ and $\hat{\boldsymbol{\beta}}_A$. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether `ltsreg` is consistent. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS or LTA estimator corresponding to Q_L.

Theorem 1. Suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the MLR model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_{k,B})$ is bounded for γ near 0.5, the `hbreg` estimator is high breakdown. Let $Q_L^* = Q_L$ for

LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $aQ_L^*(\hat{\boldsymbol{\beta}}_L) \leq aQ_L^*(\hat{\boldsymbol{\beta}}_D)$ for any estimator $\hat{\boldsymbol{\beta}}_D$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_{k,B}))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

Want a near 1 so that `hbreg` has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but want a large enough so that `hbreg` performs like $\hat{\boldsymbol{\beta}}_C$ for moderate n. The `hbreg` estimator was implemented with $a = 1.4$ using Q_{LTA}, $\hat{\boldsymbol{\beta}}_C$ = OLS and $\hat{\boldsymbol{\beta}}_A$ = `ltsreg`. Simulations were run with the x_{ij} (for $i > 1$) and e_i iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The `hbreg` estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 1 for $p = 5$ and 100 runs. ALTS denotes `ltsreg`, HB denotes `hbreg` and BB denotes $\hat{\boldsymbol{\beta}}_{k,B}$.

As implemented, the `hbreg` estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate n and to have outlier resistance comparable to competing practical resistant estimators. Although `hbreg` can be used to make response and residual plots, a more effective method for detecting outliers is to make an FF plot including OLS, ALTS, $\hat{\boldsymbol{\beta}}_{k,B}$ and some competing resistant estimators such as ALMS and MBA. See example 3 where the OLS, ALTS and `hbreg` fits passed through the outliers.

REFERENCES

Banejee, M., and Frees, E. W. (1997), "Influence Diagnostics for Linear Longitudinal Models," *Journal of the American Statistical Association*, 92, 999-1005.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data,* 3rd ed., New York, NY: Wiley.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* New York, NY: Wiley.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,* 26, 211-246.

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., Pacific Grove, CA: Wadsworth, 97-114.

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

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1999), "Comment on 'Regression Depth' by P.J. Rousseeuw and M. Hubert," *Journal of the American Statistical Association,* 94, 410-411.

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

Cheng, K. F., and Wu, J. W. (1994), "Testing Goodness of Fit for a Parametric Family of Link Functions," *Journal of the American Statistical Association*, 89, 657-664.

Čížek, P. (2006), "Least Trimmed Squares Under Dependence," *Journal of Statistical*

Planning and Inference, 136, 3967-3988.

Čížek, P., (2008), "General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models," *Econometric Theory*, 24, 1500-1529.

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,* New York, NY: Wiley.

Cook, R. D., and Critchley, F. (2000), "Identifying Outliers and Regression Mixtures Graphically," *Journal of the American Statistical Association,* 95, 781-794.

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. (1982), *Residuals and Influence in Regression,* London, UK: Chapman & Hall.

Cook, R. D., and Weisberg, S. (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1-10.

Cook, R. D., and Weisberg, S. (1994), "Transforming a Response Variable for Linearity," *Biometrika,* 81, 731-737.

Cook, R. D., and Weisberg, S. (1997), "Graphics 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,* New York, NY: Wiley.

De Gruttola, V., Ware, J. H., and Louis, T. A. (1987), "Influence Analysis of Generalized Least Squares Estimators," *Journal of the American Statistical Association,* 82, 911-917.

Dunn, O. J., and Clark, V. A. (1974), *Applied Statistics: Analysis of Variance and Regression*, New York, NY: Wiley.

Fox, J. (1991), *Regression Diagnostics,* Newbury Park, CA: Sage.

Freedman, D. A. (2005), *Statistical Models Theory and Practice,* New York, NY: Cambridge University Press.

Hawkins, D. M., and Olive, D. J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

Houseman, E. A., Ryan, L. M., and Coull, B. A. (2004), "Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Errors," *Journal of the American Statistical Association*, 99, 383-394.

Huber, P. J., and Ronchetti, E. M. (2009), *Robust Statistics,* 2nd ed., Hoboken, NJ: Wiley.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Kuehl, R. O. (1994), *Statistical Principles of Research Design and Analysis*, Belmont, CA: Duxbury Press.

Long, J. S., and Ervin, L. H. (2000), "Using Heteroscedasticity Consistent Standard

Errors in the Linear Regression Model," *The American Statistician*, 54, 217-224.

McKean, J. W., Sheather, S. J., and Hettmansperger, T. P. (1993), "The Use and Interpretation of Residuals Based on Robust Estimation," *Journal of the American Statistical Association,* 88, 1254-1263.

Miller, F. R., Neill, J. W., and Sherfey, B. W. (1999), "Implementation of a Maxmin Power Criterion to Select Near Replicates for Regression Lack-of-Fit Tests," *Journal of the American Statistical Association*, 94, 610-620.

Montgomery, D. C., Peck, E. A., and Vining, G. (2006), *Introduction to Linear Regression Analysis,* 4th ed., Hoboken, NJ: Wiley.

Moore, D. S. (2007), *The Basic Practice of Statistics*, New York, NY: W. H. Freeman.

Olive, D. J. (2004), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Basel, Switzerland: Birkhäuser, 221-233.

Olive, D. J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

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

Peña, D. (2005), "A New Statistic for Influence in Regression," *Technometrics*, 47, 1-12.

Peña, E. A., and Slate, E. H. (2006), "Global Validation of Linear Model Assumptions," *Journal of the American Statistical Association*, 101, 341-354.

Polasek, W. (1984), "Regression Diagnostics for General Linear Regression Models," *Journal of the American Statistical Association*, 79, 336-340.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection,*

New York, NY: Wiley.

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

Salibian-Barrera, M., Willems, G., and Zamar, R. H. (2008), "The Fast τ-Estimator of Regression," *Journal of Computational and Graphical Statistics*, 17, 659-682.

SAS Institute, (1985), *SAS User's Guide: Statistics*, Version 5, Cary, NC: SAS Institute.

Sheather, S. J. (2009), *A Modern Approach to Regression with R,* New York, NY: Springer.

Shi, L., and Chen, G. (2009), "Influence Measures for General Linear Models with Correlated Errors," *The American Statistician*, 63, 40-42.

Snedecor, G. W., and Cochran, W. G. (1967), *Statistical Methods*, 6th ed., Ames, IA: Iowa State College Press.

Su, Z., and Yang, S.-S., (2006), "A Note on Lack-of-Fit Tests for Linear Models Without Replication," *Journal of the American Statistical Association*, 101, 205-210.

Tukey, J. W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

Velilla, S. (1998), "A Note on the Behavior of Residual Plots in Regression," *Statistics and Probability Letters,* 37, 269-278.

Wang, H. Z., and Suter, D. (2003), "Using Symmetry in Robust Model Fitting," *Pattern Recognition Letters*, 24, 2953-2966.

Wood, F. S. (1973), "The Use of Individual Effects and Residuals in Fitting Equations to Data," *Technometrics*, 15, 677-695.

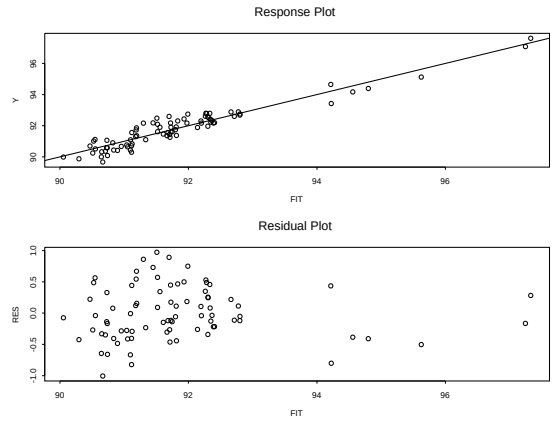

Figure 1: Octane Data

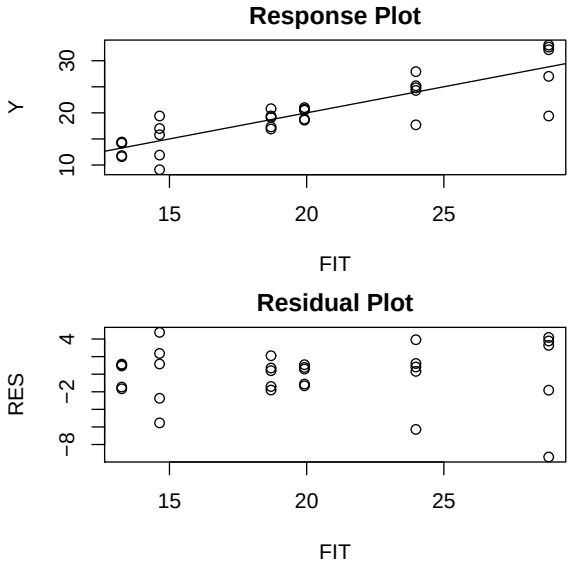

Figure 2: One Way Anova Data

Table 1: MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

n	method	mn or sd	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
25	HB	mn	1.0743	1.0832	0.9733	1.0385	1.0302
		sd	0.4392	0.4869	0.5075	0.5207	0.5169
	OLS	mn	1.0390	1.0170	1.0045	1.0044	1.0031
		sd	0.2120	0.2318	0.2457	0.2144	0.2651
	ALTS	mn	1.0590	1.1511	1.0074	1.0694	1.0554
		sd	0.4538	0.4787	0.5071	0.5291	0.5223
	BB	mn	1.0648	0.5685	0.5757	0.5989	0.5702
		sd	0.4725	0.4418	0.4526	0.4475	0.4341
400	HB	mn	0.9937	0.9981	0.9993	1.0106	0.9972
		sd	0.0440	0.0553	0.0504	0.0469	0.0496
	OLS	mn	0.9937	0.9981	0.9993	1.0106	0.9972
		sd	0.0440	0.0553	0.0504	0.0469	0.0497
	ALTS	mn	0.9901	0.9972	1.0112	1.0410	1.0043
		sd	0.1519	0.1397	0.1544	0.1635	0.1595
	BB	mn	0.9914	0.8762	0.8879	0.8845	0.8606
		sd	0.1251	0.1265	0.1307	0.1280	0.1183

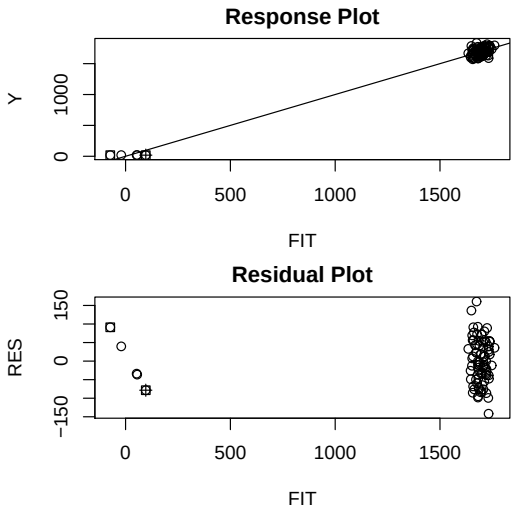

Figure 3: Buxton Data

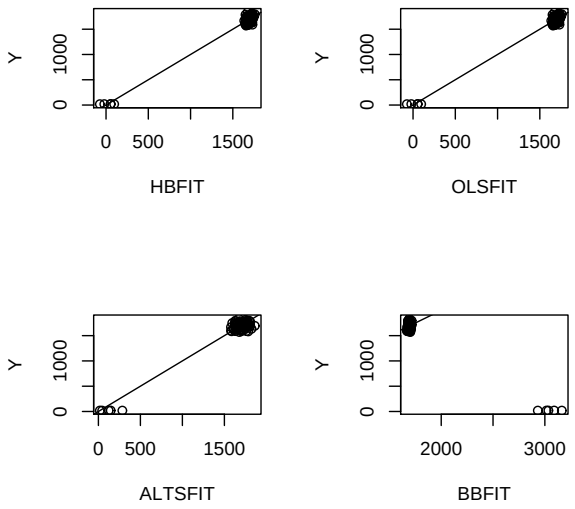

Figure 4: Response Plots Comparing Robust Regression Estimators

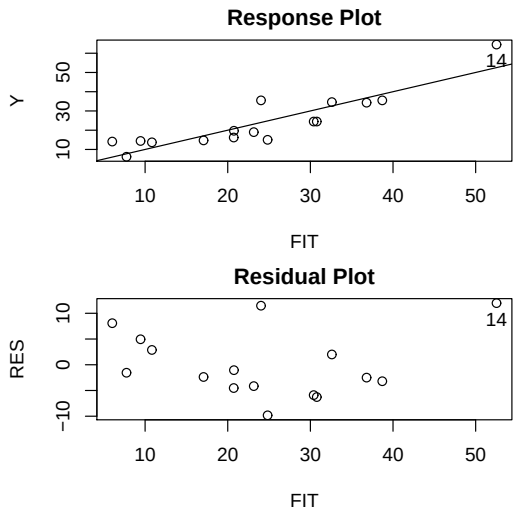

Figure 5: Latin Square Data

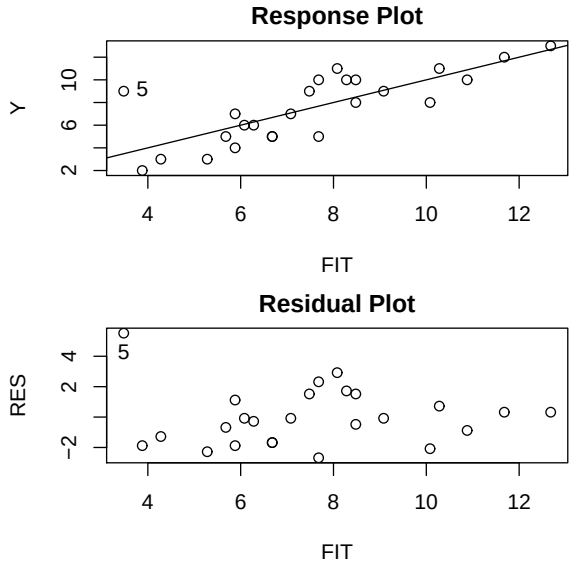

Figure 6: One Way Block Design Does Not Fit All of the Data

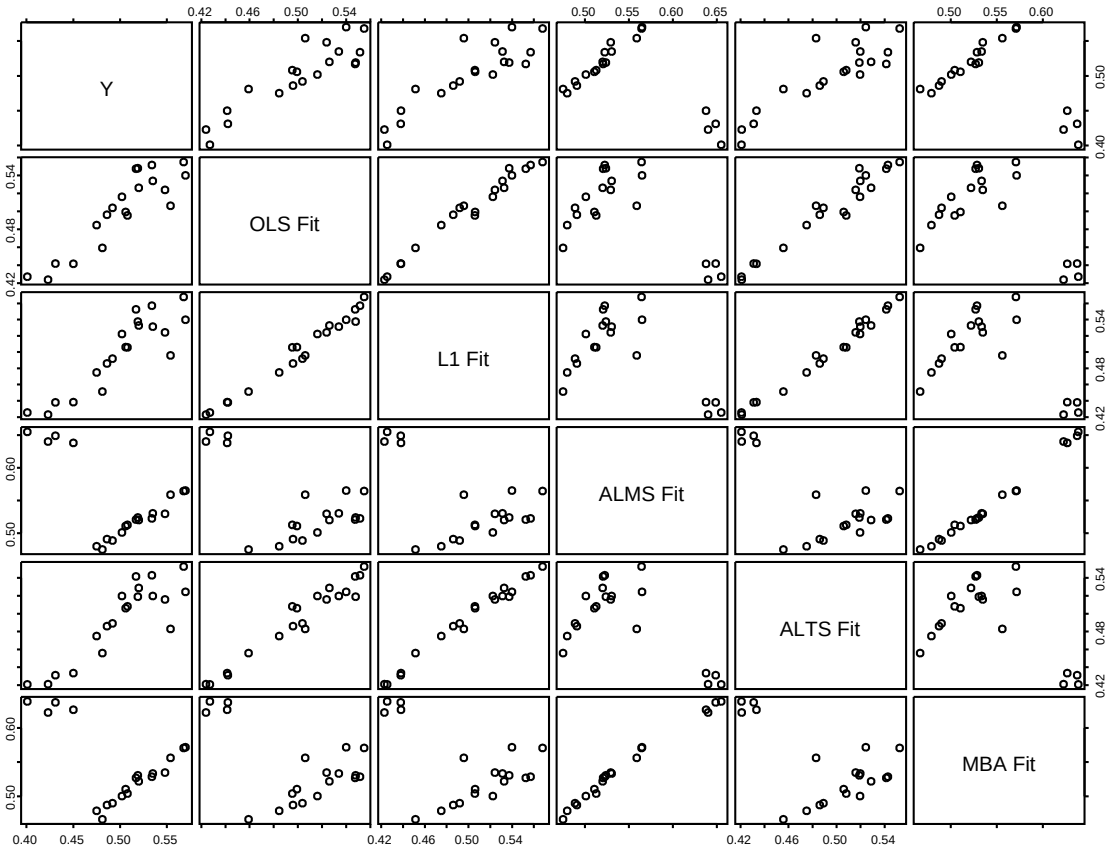

Figure 7: FF Plot for Wood Data

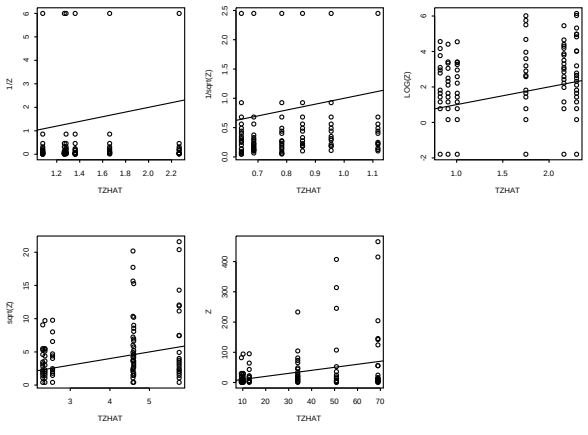

Figure 8: Transformation Plots for Crab Data

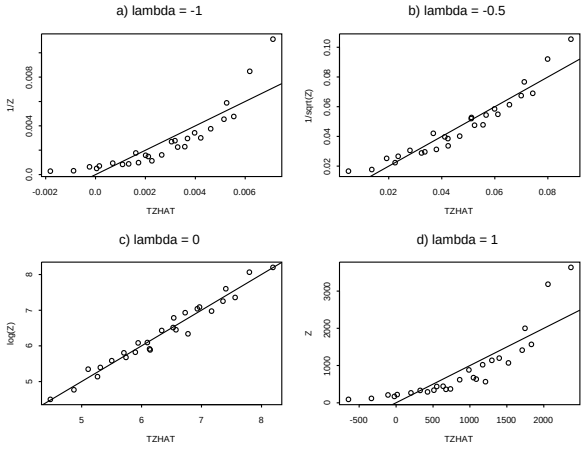

Figure 9: Transformation Plots for Textile Data