Linear Models, Problems

John Fox

McMaster University

Draft: Please do not quote without permission

Revised January 2003

Glossary

added-variable plot A diagnostic graph for showing leverage and influence of observations on a regression coefficient.

Breusch-Pagan test A score test for non-constant error variance (heteroscedasticity).

collinearity (multicollinearity) Strong linear relationships among the columns of the model matrix in a linear model, reducing the precision of coefficient estimates.

component-plus-residual plot A diagnostic graph for nonlinearity, plotting partial residuals against an explanatory variable.

Durbin-Watson statistics Test statistics for serial correlation of the errors in a time-series regression.

heteroscedastic-consistent covariance matrix An estimate of the covariance matrix of the least-squares regression coefficients that is consistent even when the error variance is not constant.

high-leverage observation An observation that can exert substantial influence on the least-squares estimates by virtue of its unusual combination of values of the explanatory variables; hat-values are measures of leverage.

influential observation An observation whose removal substantially changes the regression coefficients; dfbeta, dfbetas, and Cook's distances are measures of influence.

outlier In a linear model, a value of the response variable that is conditionally unusual given the values of the explanatory variables; studentized residuals may be used to locate outliers.

variance-inflation factor (VIF) A measure of the impact of collinearity on the precision of estimation of a coefficient.

Despite its broad direct applicability, the normal linear model makes strong assumptions about the structure of the data. If these assumptions are not satisfied, data analysis based on the linear model may be suboptimal or even wholly misleading. The use of linear models in data analysis should therefore be accompanied by diagnostic techniques meant to reveal inadequacies in the model or weaknesses in the data. In favorable instances, diagnostics that reveal such problems also point toward their solution. This article deals with a number of problems that can afflict linear models and least-squares estimation: usual data, non-normal errors, non-constant error variance, non-independent errors, nonlinear relationships, collinearity, and measurement error in the explanatory variables.

I. The Normal Linear Model: Structure and Assumptions

Most of applied statistics is based, directly or indirectly, on the normal linear model fit by least-squares: Linear models for simple and multiple regression, analysis of variance, and analysis of covariance (dummy regression) have broad direct application. Various extensions and generalizations of the linear model — for example, generalized linear models (logistic regression, Poisson regression, etc.), robust regression, additive regression, Cox regression for survival data, linear structural-equation models, linear models fit by generalized least squares — retain central features of the normal linear model. Finally, many kinds of statistical models are fit by adaptations of linear least squares, such as the method of iteratively reweighted least-squares (IRLS) commonly employed to fit generalized linear models. In this article, I use the terms "linear model" and "regression" more or less interchangeably.

The normal linear model may be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

where

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is a column vector of observations on the response variable.
- $\mathbf{X} = \{x_{ij}\}$ is an $n \times p$ model (or design) matrix of regressors. The regressors comprising the columns of $\mathbf{X}$ may be quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial terms, dummy regressors or other contrasts representing categorical explanatory variables, interaction regressors, and so on.
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a column vector of regression coefficients. Usually, β_1 is a constant or intercept term, in which case $x_{i1} = 1$.
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is a column vector of errors. The errors are assumed to be normally and independently distributed, with zero expectations and common variance σ^2.
- $\mathbf{N}_n$ is the n-variable multivariate-normal distribution.
- $\mathbf{0}$ is a column vector of n zeroes.
- $\mathbf{I}_n$ is the order-n identity matrix.

If the values of the x's are random rather than fixed, then the errors are additionally assumed to be independent of the x's, and the x's are assumed to be measured without error. Fixed x's occur in experimental research, where the explanatory variables are under the direct control of the researcher; in observational research, the x's are usually construed as random. Assuming that

random x's are independent of the errors implies that omitted explanatory variables (which are components of the errors) are independent (or somewhat more weakly, uncorrelated) with the explanatory variables that appear explicitly in the model.

The central task of linear-model analysis is to estimate the parameters of the model — that is, the regression coefficients (the β's) and the error variance (σ^2). Under the model, including the assumption of normality, the least squares estimate of $\boldsymbol{\beta}$, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, has a variety of desirable, and even optimal properties. This solution implies that the model matrix $\mathbf{X}$ is of full column rank p; otherwise the least-squares coefficients $\mathbf{b}$ are not unique. The least-squares residuals are $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$, and an unbiased estimate of the error variance is $s^2 = \mathbf{e}^T\mathbf{e}/(n-p)$.

II. Diagnostics and Remedies

As mentioned, the normal linear model incorporates strong assumptions about the data, including assumptions of linearity, constant error variance (homoscedasticity), normality, and independence. Linearity follows from the assumption that the average error is zero, for then $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. The assumption of linearity is to be broadly construed, since through transformations, polynomial regressors, and so on, the linear model may incorporate what are conventionally thought of as nonlinear partial relationships. That is, the linear model is linear in the parameters but not necessarily in the explanatory variables.

Not all of the assumptions of the model are checkable. For example, when $\mathbf{X}$ is random, we assume that it is independent of the errors $\boldsymbol{\varepsilon}$, but because the least-squares residuals $\mathbf{e}$ are necessarily uncorrelated with $\mathbf{X}$, the assumption is partly reflected in the estimated model. We can detect certain departures from the assumptions (such as certain forms of nonlinearity, non-constant error variance, non-normality, or serial correlation of the errors), but not a correlation between the errors and a particular column of $\mathbf{X}$, for example, induced by an omitted (confounding or lurking)

explanatory variable. Although there are tests for omitted explanatory variables (such as Ramsey's RESET test — see, e.g., Godfrey, 1988), these necessarily make assumptions about the nature of the omission. Likewise, except when there are replicated observations at each unique row of $\mathbf{X}$, possible departures from linearity are so diverse as to preclude effective, fully general, methods for detecting them.

The consequences of violating the assumptions of the normal linear model range in seriousness from benign to fatal. For example, depending partly upon the configuration of the model matrix, moderate violation of the assumption of constant error variance may have only a minor impact on the efficiency of estimation. In contrast, substantial violation of the assumption of linearity suggests that the wrong mean function is fit to the data, rendering the results of the analysis entirely meaningless.

Diagnostic methods, which are the focus of this article, are designed to discover violations of the assumptions of the model and data conditions that threaten the integrity of the analysis. Two general approaches to linear-model problems are numerical diagnostics, including statistical hypothesis tests, and graphical diagnostics. Tests determine, for example, whether there is evidence in the data to reject one or another assumption of the model, or to determine whether there is evidence that particular observations do not belong with the others. Graphical methods seek to reveal patterns that are indicative of problems with either the model or the data, and often are useful in suggesting ways to improve the data analysis, for example, by transformation of the variables or other respecification of the model. The two approaches, numerical diagnostics and graphs, are often complementary — for example using a test to check a pattern discerned in a graph, or drawing a graph to display numerical diagnostics. This article emphasizes graphical displays; more information on tests may be found, for example, in Godfrey (1988).

With a few exceptions, space considerations preclude the presentation of examples in this article.

Likewise, the presentation is limited to basic and widely available methods, excluding some newer and more advanced techniques, and the generalization of diagnostics to other kinds of statistical models. Examples, advanced methods, and generalizations are copiously available in the sources cited in the bibliography.

III. Unusual Data

Least-squares estimates, and other statistics associated with them, such as coefficient standard errors, correlation coefficients, and the regression standard error (standard deviation of the residuals), can be seriously influenced by unusual data. It is useful to distinguish among outliers, high-leverage observations, and influential observations. In the context of a linear model, an outlier is an observation whose y-value is unusual conditional on the values of the x's. (Thus, a univariate outlier on y or an x need not be a regression outlier.) A high-leverage observation, in contrast, is one that has an unusual combination of x-values. Observations that combine outlyingness and leverage influence the values of the least-squares coefficients, in the sense that the removal of such observations changes the coefficients substantially. Other regression quantities, however, can be strongly affected under other circumstances. For example, a high-leverage, in-line observation decreases the standard errors of (some of) the regression coefficients; such an observation may in certain circumstances be regarded as a source of precision in estimation and in others as providing an illusion of precision. Similarly, an outlying observation at a low-leverage point will not affect the coefficients much, but will decrease the correlation, increase the residual standard error, and increase the standard errors of the coefficients.

Standard numerical diagnostics for outliers, leverage, and influence, along with a graphical method for examining leverage and influence on the coefficients are described in the follow subsections.

A. Leverage: Hat-Values

The fitted values $\widehat{\mathbf{y}}$ in linear least-squares regression are a linear transformation of the observed response variable: $\widehat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{Hy}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat-matrix (because it transforms $\mathbf{y}$ to $\widehat{\mathbf{y}}$). The matrix $\mathbf{H}$ is symmetric ($\mathbf{H} = \mathbf{H}^T$) and idempotent ($\mathbf{H} = \mathbf{H}^2$), and thus its ith diagonal entry, h_{ii}, gives the sum of squared entries in its ith row or column. Because the ith row of $\mathbf{H}$ represents the weight associated with the ith observed response in determining each of the fitted values, $h_i = h_{ii}$, called the hat-value, summarizes the potential contribution of observation i to the fitted values collectively, and is a suitable measure of the leverage of this observation in the least-squares fit.

Because $\mathbf{H}$ is a projection matrix (projecting $\mathbf{y}$ orthogonally onto the subspace spanned by the columns of $\mathbf{X}$), the average hat-value is p/n. By rule of thumb, hat-values are considered noteworthy when they exceed twice (or in small samples, three times) the average value. It is best, however, to examine hat-values graphically — for example in an index plot (a scatterplot with hat-values on the vertical axis and observation indices on the horizontal axis) — using the numerical cutoffs to aid interpretation of the graph.

B. Outliers: Studentized Residuals

Although we regard residuals as estimates of the errors, and although the residuals are the key to outlying observations, it is a sad fact that even when the standard assumptions hold, the least-squares residuals, though normally distributed, are dependent and heteroscedastic: $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{Hy} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$. In most circumstances, the diagonal entries of $\mathbf{H}$ are unequal and the off-diagonal entries are nonzero. One approach is to standardize the residuals, as $e_i^* = e_i/(s\sqrt{1-h_i})$, but the standardized residuals do not have a simple distribution.

The more useful studentized residuals, however, are each t-distributed (under the model) with

$n - p - 1$ degrees of freedom: $\text{rstudent}_i = e_i/(s_{(-i)}\sqrt{1-h_i})$, where $s_{(-i)}$ is the estimated standard deviation of the errors computed from a regression with observation i deleted. (There are efficient computational formulas for the studentized residuals that do not require refitting the model removing each observation in turn.)

Although the studentized residuals are each distributed as t, different studentized residuals are correlated, and there is moreover a problem of simultaneous inference when, as is usually the case, our attention is drawn to the largest $|\text{rstudent}_i|$. A solution is to employ a Bonferroni adjustment for the outlier test, multiplying the usual two-sided p-value for the biggest absolute studentized residual by the number of observations. It is also helpful in this context to examine plots of the studentized residuals (see the section below on non-normal errors).

C. Influence: dfbeta, dfbetas, Cook's Distances

Although, as noted above, unusual data can exert influence on all regression outputs, I concentrate here on the coefficient estimates. A straightforward approach to the influence of individual observations is to ask how much the coefficients change when an observation is deleted. The change in the regression coefficients attending the deletion of observation i can be computed without literally refitting the model: $\mathbf{dfbeta}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i/(1-h_i)$, where $\mathbf{x}_i$ is the ith row of the model matrix (written as a column vector); dfbeta_{ij} gives the influence of observation i on coefficient j. A standardized version of this statistic, called dfbetas_{ij}, simply divides dfbeta_{ij} by the standard error of b_j.

There are $n \times p$ of each of dfbeta_{ij} and dfbetas_{ij}, and so their examination can be tedious, even in graphs such as index plots. Several similar measures have been proposed to examine the influence of individual observations on the vector of coefficient estimates. The most popular such measure is Cook's distance, $D_i = (\mathbf{b} - \mathbf{b}_{(-i)})^T\mathbf{X}^T\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})/ps^2$, where $\mathbf{b}$ is the least-squares

coefficient vector computed on the full data set and $\mathbf{b}_{(-i)}$ is the coefficient vector with observation i omitted. As in the case of rstudent and dfbeta, Cook's D can be calculated without refitting the model. Cook's distance is invariant with respect to rescaling of columns of $\mathbf{X}$, and can be conveniently viewed in a graphical display such as an index plot.

D. Joint Influence: Added-Variable Plots

Single-observation deletion statistics such as dfbeta, dfbetas, and Cook's distances, can fail to detect observations that are jointly influential. A good graphical diagnostic that works in such instances, and more generally, is the added-variable plot, also called a partial-regression plot. One such plot is constructed for each coefficient in the model. The collection of p added-variable plots represents projections of the $p-1$ dimensional regression surface onto a series of planes.

For example, for b_2, we begin by regressing y on all columns of the model matrix excluding the second, $\mathbf{X}_{(-2)}$, obtaining residuals $\mathbf{y}^*_{(2)} = \mathbf{y} - \mathbf{X}^T_{(-2)}(\mathbf{X}^T_{(-2)}\mathbf{X}_{(-2)})^{-1}\mathbf{X}^T_{(-2)}\mathbf{y}$. Then we compute residuals for the regression of the second column, $\mathbf{x}_{(2)}$, on the others: $\mathbf{x}^*_{(2)} = \mathbf{x}_{(2)} - \mathbf{X}^T_{(-2)}(\mathbf{X}^T_{(-2)}\mathbf{X}_{(-2)})^{-1}\mathbf{X}^T_{(-2)}\mathbf{x}_{(2)}$. (I have skipped over the first column of the model matrix because it is typically a column of ones, representing the regression constant; one can construct an added-variable plot for the constant, but the other coefficients are almost always of greater interest.) The added variable plot graphs $\mathbf{y}^*_{(2)}$, on the vertical axis, against $\mathbf{x}^*_{(2)}$, on the horizontal axis. The least-squares slope corresponding to this scatterplot is the coefficient b_2 from the original least-squares fit; the residuals are the original least-squares residuals e_i; and the standard error of the slope for the simple regression of $\mathbf{y}^*_{(2)}$ on $\mathbf{x}^*_{(2)}$ is (except for degrees of freedom) the standard error of the coefficient b_2 in the original model. The added-variable plot therefore shows the leverage and influence of the observations in determining b_2, along with the precision of the estimate b_2. An example appears in Figure 1.

Fig. 1 about here.

IV. Non-Normal Errors

Violations of the assumption of normally distributed errors can threaten the efficiency of estimation (for example, in the case of heavy-tailed error distributions) or can compromise the interpretability of the least-squares fit, which estimates the conditional mean of y as a function of the x's (for example, in the case of skewed errors). As explained in the preceding section, least-squares residuals have some different properties from the errors; nevertheless, examining the distribution of the residuals can be informative about the distribution of the errors.

Quantile-comparison plots of studentized residuals help us to focus on the tails of the distribution. Such plots graph ordered residuals (usually on the vertical axis) against approximate expected quantiles of a reference distribution — either the normal distribution or the t distribution with $n - p - 1$ degrees of freedom. If the residuals follow the reference distribution, then the plot will be approximately linear. Systematic nonlinear patterns are indicative, for example, of skewed residuals or heavy-tailed residuals. Enhancing the plot with a bootstrapped confidence envelope, computed assuming the truth of the fitted model, provides a visual guide to the extremity of departures from normality. See Figure 2 for an illustration.

Fig. 2 about here.

There are several methods for testing departures from normality. One approach is to imbed the normal linear model in a more general model that indexes non-normality by one or more parameters. Consider, for example, the Box-Cox regression model

$$y_i^{(\lambda)} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\varepsilon_i \sim \text{NID}(0, \sigma^2)$$

where

$$y_i^{(\lambda)} = \begin{cases} (y_i^\lambda - 1)/\lambda & \text{for } \lambda \neq 0 \\ \log_e y_i & \text{for } \lambda = 0 \end{cases}$$

is a modified power transformation (and the y_i are all positive). This model, with parameters $\beta_1, \beta_2, \ldots, \beta_p, \sigma^2$, and λ, is estimated by the method of maximum likelihood. A value of $\widehat{\lambda}$ significantly different from 1 is indicative of non-normal errors (corrected by the Box-Cox transformation).

V. Non-Constant Error Variance

Non-constant error variance (or heteroscedasticity) also threatens the efficiency of least-squares estimation as well as the validity of statistical inference. We could, in principle, examine the spread of residuals around the fitted regression surface, but because this surface is usually high-dimensional, it is, except in the simplest cases, impractical to do so. An alternative is to look for common patterns of non-constant residual variation. One such pattern is a tendency for residual spread to increase with the level of the response, a pattern that may be discernible in a plot of residuals (for example, rstudent_i or $|\text{rstudent}_i|$) against fitted values ($\widehat{y}_i$).

A simple score test for the dependence of spread on level (called the Breusch-Pagan test, but independently developed by Cook and Weisberg) is to regress squared standardized residuals, $e_i^2/\widehat{\sigma}^2$, on the fitted values, where $\widehat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/n$ is the maximum-likelihood estimator of the error variance. Half the regression sum of squares from this auxiliary regression is distributed as chi-square with one degree of freedom under the null hypothesis of constant error variance. The test can be extended to other, more general, models for the error variance.

Once discovered, non-constant error variance can often be corrected by transformation of the response variable. In other instances, weighted-least-squares (WLS) estimation can be employed to obtain efficient coefficient estimates and accurate standard errors. Finally, it is possible to correct

the standard errors of the ordinary least-squares estimates for unmodelled heteroscedasticity. A commonly employed approach, due to White (1980; see Long and Ervin, 2000, for variations), estimates the covariance matrix of the least-squares coefficients consistently even in the presence of heteroscedasticity:

$$\widetilde{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widetilde{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\widetilde{\mathbf{\Sigma}}$ =diag(e_i^2). The estimator $\widetilde{V}(\mathbf{b})$ is used in place of the usual $\widehat{V}(\mathbf{b}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$. Discovering the form of non-constant error variance and taking it into account is advantageous, however, because doing so improves the efficiency of estimation.

VI. Non-Independence

The potential ill effects of non-independent errors are similar to those of non-constant error variance. Departures from independence can occur in a variety of ways. Usually, the assumption of independence must be justified by the method of data collection. Common sources of non-independence include clustered data (for example, hierarchical data, such as data collected on children within schools), longitudinal data collected on individuals over time, and time-series data.

How one checks for non-independence depends upon the structure of the data. In time-series regression, for example, it is common to examine the autocorrelations of the residuals, where the autocorrelation at lag t is $r_t = \sum_{i=t+1}^{n} e_i e_{i-t} / \sum_{i=1}^{n} e_i^2$. Tests for serial correlation in the errors may be based on the Durbin-Watson statistics $D_t = \sum_{i=t+1}^{n} (e_i - e_{i-t})^2 / \sum_{i=1}^{n} e_i^2$. The D_t have a complex sampling distribution that depends upon the configuration of the model matrix for the regression, but there are good approximations available for the p-values, which may also be estimated by bootstrapping.

Remedies for non-independence depend upon its source. For example, serial correlation of errors

in time-series regression may be handled by generalized least-squares estimation. Likewise, linear and nonlinear mixed-effects models provide efficient estimates for clustered and longitudinal data.

VII. Nonlinearity

Nonlinearity is in a sense the most serious violation of the assumptions of the linear model because it implies that we are fitting the wrong equation to the data. As employed here, 'nonlinearity' encompasses all departures from the specified functional form of the model; as mentioned, the linear model of equation (1) encompasses specifications that are nonlinear in the explanatory variables but linear in the parameters, such as polynomial regression models, models including interactions, models with dummy regressors, and so on. As the other problems discussed in this article, however, nonlinearity can vary in degree from trivial to serious. If there are one or two explanatory variables, we can look for nonlinearity directly, in a smoothed two or three-dimensional scatterplot of the data, but the higher-dimensional problem is much less tractable.

If combinations of values of the explanatory variables are replicated in the data, then we can test for lack of fit by contrasting the model fit to the data with a model that simply partitions the observations according to the unique rows of the model matrix. This strategy can be adapted as an approximation in other cases by partitioning the space of the explanatory variables into regions.

An alternative general approach is to look for nonlinearity of a particular form, such as a nonlinear partial relationship between the response and each explanatory variable. This approach can be implemented both in graphical diagnostics and in tests.

A frequently useful graphical diagnostic is the component-plus-residual (or partial residual) plot. Suppose that the coefficient β_1 in the linear model is a regression constant, and focus on the partial relationship between y and x_2 conditional on the other x's. The partial residual associated with x_2 is $\mathbf{e}_{2|3,\ldots,p} = b_2\mathbf{x}_2 + \mathbf{e}$, where $\mathbf{e}$ is the vector of least-squares residuals and $\mathbf{x}_2$ is the second column

of the model matrix. The component-plus-residual plot is a scatterplot of the partial residuals, on the vertical axis, against the values of x_2, on the horizontal axis. Smoothing the plot with a nonparametric-regression line aids interpretation. Once we discern the nature of a nonlinear partial relationship between the response and a particular predictor x_j, we can determine how to model it — for example, by a transformation of x_j or by a polynomial in x_j. An illustrative component-plus-residual plot appears in Figure 3. Fig. 3 about here.

When there are strong nonlinear relationships among the explanatory variables, component-plus-residuals plots can prove misleading. More sophisticated versions of these diagnostics based on polynomial or nonparametric regression are more robust.

We can test for nonlinearity in a partial relationship by embedding the linear model in a more general model. For example, the Box-Tidwell regression model includes parameters for the power transformation of one or more predictors. In a similar spirit, we could replace the linear model with a semiparametric additive model in which certain terms are modelled nonparametrically, contrasting the fit of this model with the linear model by an approximate incremental F-test. Likewise, if an explanatory variable is discrete, we can treat it as a set of dummy regressors, contrasting the resulting fit with the original linear model.

VIII. Collinearity and Other Sources of Imprecision

Suppose that we fit a linear model in which β_1 is the regression constant. The sampling variance of the least-squares estimate b_2 of β_2 is

$$V(b_2) = \frac{\sigma^2}{(n-1)s_2^2} \times \frac{1}{1-R_2^2} \tag{2}$$

where σ^2 is the error variance (estimated by s^2 in a application), n is the sample size, s_2^2 is the sample variance of x_2, and R_2^2 is the squared multiple correlation from the regression of x_2 on the other x's. Thus, imprecise estimates are the product of large error variance, small samples, homogeneous explanatory variables, and strong linear relationships among (some of) the explanatory variables, termed collinearity or multicollinearity. The second factor in equation (2), $1/(1 - R_2^2)$, called the variance-inflation factor (VIF), expresses the harm produced by collinearity. The square-root of the VIF is the expansion of the size of the confidence interval for β_2 due to collinearity.

Variance-inflation factors are applicable to one-degree-of-freedom effects, but not, for example, to sets of related dummy regressors or polynomial terms. Write the linear model as

$$\mathbf{y} = \beta_1\mathbf{1} + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\beta}_3\mathbf{X}_3 + \boldsymbol{\varepsilon}$$

where $\mathbf{1}$ is a vector of ones, $\mathbf{X}_2$ contains the columns of the model matrix pertaining to a particular multiple-degree-of-freedom effect, and $\mathbf{X}_3$ contains the remaining columns of the model matrix. Then the generalized variance inflation factor is $\text{GVIF} = \det\mathbf{R}_{22}\det\mathbf{R}_{33}/\det\mathbf{R}$, where $\mathbf{R}_{22}$ is the correlation matrix among the columns of $\mathbf{X}_2$, $\mathbf{R}_{33}$ is the correlation matrix among the columns of $\mathbf{X}_3$, $\mathbf{R}$ is the correlation matrix for $[\mathbf{X}_2, \mathbf{X}_3]$ (i.e., the full model matrix omitting the constant regressor), and det stands for the determinant. The GVIF gives the inflation in the squared (hyper)volume of the joint confidence region for the coefficients in $\boldsymbol{\beta}_2$ due to collinearity between the columns of $\mathbf{X}_2$ and $\mathbf{X}_3$.

There are other diagnostics for collinearity — based, for example, on eigenvalue-eigenvector or singular-value decompositions of the model matrix or of the cross-products or correlations among the columns of the model matrix — but variance-inflation factors are particularly easy to interpret. Once collinear relationships are located, they can be explored by regressing one column (or set of

columns) of the model matrix on the others.

What to do about collinearity is a thornier issue. Unlike most of the other problems discussed in this article, collinearity is usually best understood as a problem with the data rather than with the model: Because of strong correlations among the x's, the data are uninformative about certain partial relationships between the response and explanatory variables. Common approaches to collinearity, such as variable selection and biased estimation, make the problem *seem* to disappear by (often surreptitiously) altering the questions asked of the data.

IX. Measurement Error in the Explanatory Variables

Consider the following regression model

$$y_i = \beta_1 \xi_i + \beta_2 x_{i2} + \varepsilon_i$$

where the 'latent' explanatory variable ξ is not directly observed (hence the Greek letter); x_2 is directly observed and measured without error; and the regression error ε behaves according to the usual assumptions — in particular, $E(\varepsilon_i) = 0$, and the errors are uncorrelated with one-another and with the explanatory variables. To simplify the exposition, but without loss of generality, suppose that that all of the variables, y, ξ, and x_2 (along with ε) have means of 0, eliminating the regression constant from the model. Imagine that we have a fallible observable indicator x_1 of ξ,

$$x_{i1} = \xi_i + \delta_i$$

where δ is the measurement-error component of x_1. Let us suppose further that the measurement errors are 'well-behaved': $E(\delta) = 0$; $E(\xi\delta) = 0$ (the measurement errors and 'true scores' are un-

correlated); $E(x_2\delta) = 0$ (the measurement errors in x_1 are uncorrelated with the other explanatory variable); and $E(\varepsilon\delta) = 0$ (the measurement and regression errors are uncorrelated).

Under these circumstances, it is not hard to show that

$$\begin{aligned} \beta_1 &= \frac{\sigma_{y1}\sigma_2^2 - \sigma_{12}\sigma_{y2}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2 - \sigma_\delta^2\sigma_2^2} \\ \beta_2 &= \frac{\sigma_{y2}\sigma_1^2 - \sigma_{12}\sigma_{y1}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} - \frac{\beta_1\sigma_{12}\sigma_\delta^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{aligned} \tag{3}$$

but that the population analogs of the least-squares coefficients (i.e., the quantities that the least-squares coefficients estimate) are

$$\begin{aligned} \beta_1' &= \frac{\sigma_{y1}\sigma_2^2 - \sigma_{12}\sigma_{y2}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ \beta_2' &= \frac{\sigma_{y2}\sigma_1^2 - \sigma_{12}\sigma_{y1}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{aligned} \tag{4}$$

In these equations, the σ's are variances and covariances: For example, σ_{y1} is the covariance of y and x_1; σ_1^2 is the variance of x_1; and σ_δ^2 is the measurement-error variance (which I suppose is nonzero).

Comparing equations (3) and (4) reveals that the least-squares coefficients are biased are estimators of β_1 and β_2: Because the denominator of β_1 in equation (3) must be positive, while $-\sigma_\delta^2\sigma_2^2$ is necessarily negative, the least-squares estimand β_1' is 'attenuated' towards zero.

The bias in β_2' has a different (and more interesting) characterization: Let us write $\beta_2' = \beta_2 + bias$, where

$$bias = \frac{\beta_1\sigma_{12}\sigma_\delta^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

The bias can be in either direction, but as the measurement-error variance σ_δ^2 grows, the least-squares estimand β_2' is driven towards σ_{y2}/σ_2^2, which is the population analog of the least-squares

coefficient for the regression of y on x_2 alone. That is, a large measurement-error component in x_1 makes it a weak statistical control, in effect removing it from the regression equation. Of course, when there are several explanatory variables all subject to measurement error, the effects are more complex.

What can be done about measurement error in the predictors? Under certain circumstances — for example, when there are multiple indicators of latent explanatory variables — it is possible to derive unbiased estimators that take account of the measurement errors (see, e.g., Bollen, 1989).

Bibliography

Atkinson, A. C. (1985) *Plots, Transformations and Regression.* Oxford University Press, Oxford.

Atkinson, A. C., and Riani, M. (2000) *Robust Diagnostic Regression Analysis.* Springer, New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980) *Regression Diagnostics.* Wiley, New York.

Bollen, K. A. (1989) *Structural Equations With Latent Variables.* Wiley, New York.

Chatterjee, S., and Hadi, A. S. (1988) *Sensitivity Analysis in Linear Regression.* Wiley, New York.

Cook, R. D. (1998) *Regression Graphics.* Wiley, New York.

Cook, R. D., and Weisberg, S. (1982) *Residuals and Influence in Regression.* Chapman and Hall, New York.

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

Duncan. O. D. (1961) A socioeconomic index for all occupations. In *Occupations and Social Status* (A. J. Reiss, Jr., ed.), pp. 109–138. Free Press, New York.

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

Fox, J. (1997) *Applied Regression, Linear Models, and Related Methods.* Sage, Thousand Oaks CA.

Fox, J., and Suschnigg, C. (1989) A Note on Gender and the Prestige of Occupations. *Canadian*

Journal of Sociology **14**, 353–360.

Godfrey, L. G. (1988) *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge University Press, Cambridge.

Long, J. S., and Ervin, L. H. (2000) Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician* **54**, 217–224.

White, H. (1980) A Heteroskedastic Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity. *Econometrica* **48**, 817–838.

Figure Captions

Figure 1. Added-variable plot for income in the regression of occupational prestige on the income and educational levels of occupations. The regression was employed by Duncan (1961) in the construction of a prestige scale for occupations. The two labelled observations, ministers and railroad conductors, serve to decrease the income slope. The least-squares line on the plot represents the regression plane in the direction of income controlling for education.

Figure 2. A t quantile-comparison plot for the studentized residuals from Duncan's occupational-prestige regression. The broken lines represent a point-wise bootstrapped 95-percent confidence envelope. The residual distribution looks heavy-tailed in comparison with the t-distribution, and one observation (minister) is slightly outside the confidence envelope.

Figure 3. Component-plus-residual plot for percent women in the regression of occupational prestige on the income level, education level, and percentage of women in 102 Canadian occupations. The broken line in the plot represents the edge of the regression plane, which is nearly horizontal in the direction of percent women. The solid line is for a nonparametric-regression smooth, which suggests a weak quadratic partial relationship. The data are discussed by Fox and Suschnigg (1989).

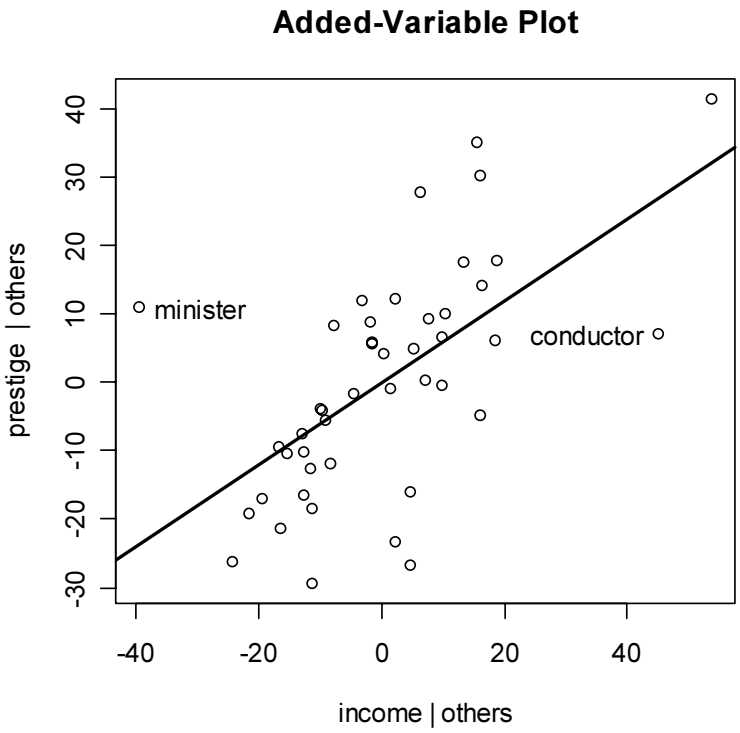
Added-Variable Plot
prestige | others
income | others
minister
conductor
-40
-20
0
20
40
-30
-20
-10
0
10
20
30
40

Figure 1:

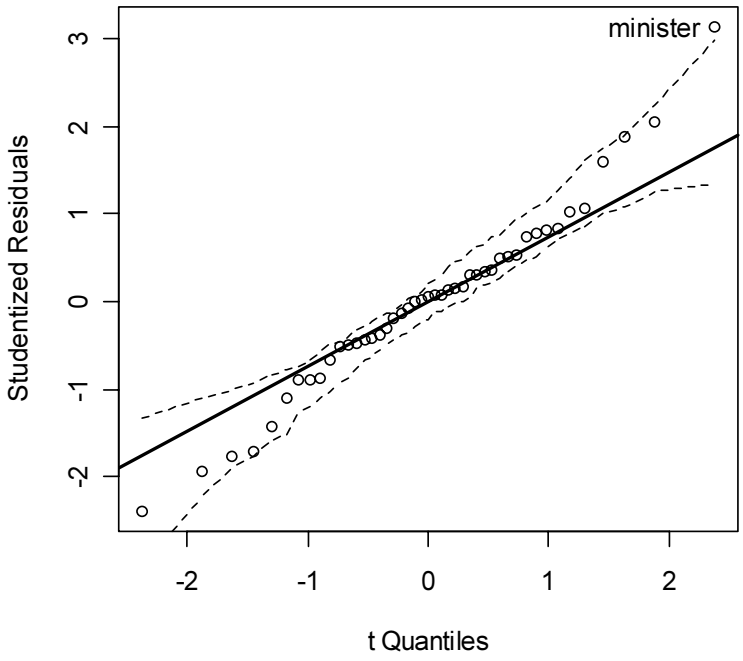

Figure 2:

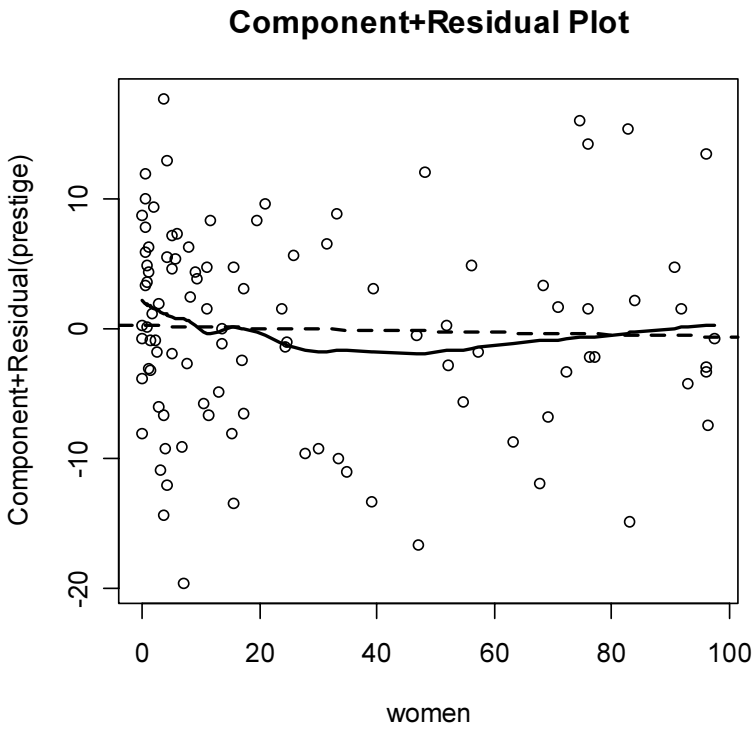
Component+Residual Plot
Component+Residual(prestige)
10
0
-10
-20
0
20
40
60
80
100
women

Figure 3: