A Predictive Approach to Model Selection and Multicollinearity*

Edward Greenberg
Department of Economics
Washington University
One Brookings Drive
St. Louis MO 63130

Robert P. Parks
Department of Economics
Washington University
One Brookings Drive
St. Louis MO 63130

November 8, 1993

Abstract

We argue for the adoption of a predictive approach to model specification. Specifically, we derive the difference between means and the ratio of determinants of covariance matrices when a subset of explanatory variables is included or excluded from a regression. For several special cases these measures are shown to be related to widely used tools for studying model specification. Results for a set of simulated data and for two economic applications are presented as examples.

*Thanks to Professor Siddhartha Chib for comments. An earlier version of this paper was presented at the Midwest Econometrics Group meeting at Notre Dame in September 1991.

1 Introduction

This paper addresses the question of when it might be reasonable to simplify a linear regression model by not including a set of variables in an equation that contains another set of variables whose inclusion is clearly warranted on the basis of theory or previous empirical evidence. We are thus concerned with the problem of model selection, which we approach from a predictive Bayesian viewpoint. This problem, of course, has been intensively and extensively studied from many other viewpoints. Since nonexperimental data in general, and economics data in particular, are often highly correlated, model specification is closely related to the problem of multicollinearity.

Our approach is to compare the predictive density for an equation with and without the set of variables in question and to argue that the set may be safely omitted if the omission has little or no effect on the predictive density. This approach can be utilized in more general settings; for example, Marriott et al. [15] take a very similar approach in the context of a model with serially correlated errors. The predictive density seems to us the best instrument for making this kind of choice because it is defined in terms of observable values of the dependent variable, about which an investigator is likely to have information. In contrast, little or nothing is known about values of regression coefficients in almost all econometric studies. The pervasive use of the null hypothesis value of zero for a coefficient is a sign of this ignorance.

Gaver and Geisel [8] and Amemiya [2] review a number of methods that have been proposed to deal with the model selection problem, both in the sampling-theory and Bayesian frameworks. We show below that our approach is closely related to several of these methods. Zellner's [17] Bayesian approach to model selection requires that a prior probability of the truth of each model be specified. Although our method does not require the specification of such probabilities and may therefore be somewhat lacking in Bayesian rigor, our analysis and examples show that the method has practical advantages and provides useful information. In addition, we note below the relationship between the Bayes factor and our measures.

Emphasis on the predictive distribution has been championed by Geisser and co-authors; see, for example, Geisser [9] and Johnson and Geisser [12]. Johnson and Geisser's [12] application of predictive methods to the problem of detecting influential observations provides new insights into the problem. They find that some of the measures they examine have been proposed in the sampling-theory approach. Geisser and Eddy [10] propose a predictive sample reuse technique that involves the computation of a predictive density, where the density is evaluated as the product of the conditional densities at the observed sample points. Likelihoods evaluated at the sample points are then compared for each model under consideration. This approach does not require specifying probabilities of models. Like most of the other approaches to model selection, however, the method attempts to rank the models without regard to the point at which the prediction will take place. As will be seen below, our method can be used to examine prediction at particular points. We do not, however, attempt to reduce the problem of model selection to ranking by one criterion, because we believe that there may be much to be learned from examination of a richer information set.

Other statisticians have argued persuasively for a predictive approach. For example, Aitchison and Dunsmore [1] note at page *x*:

> Now the purpose of such [parametric] inference statements is surely to convey to some second party information about what is likely to happen if the experiment is performed again, or perhaps repeated a number of times. It is surprising therefore that greater thought has not been given to the more direct practical type of inference, where statements are required for what is likely to occur when future experiments are performed.

A quotation from Clayton, Geisser, and Jennings [4] is pertinent to the model selection problem:

> Econometric and other statistical models are often simplifications of extremely complicated phenomena, and it is a mistake to assume that any particular model is actually a true representation of the underlying process. What is hoped is that the model may be an adequate description and perhaps useful for some purpose. Hence it is often puzzling why there has been so much effort, especially in the softer social sciences, devoted to "testing" parameters of a model as if they were true entities and not, as in most instances, convenient artifices. A more substantial enterprise than testing should be model selection, i.e. selecting one of several alternative models such that the selected model (irrespective of its truth) would serve best some purpose of the investigator (descriptive or predictive).

In our approach, comparing predictions of various models at the values of the independent variables observed in the sample would help evaluate the descriptive power of a model, while comparing predictions at a value of the independent variables for which a prediction is to be made would evaluate its predictive power.

Finally, Jaynes [11], in discussing the approach of physicists to the problem of hypothesis testing, puts the matter clearly (in what follows λ is a parameter value that is being compared to the presently accepted value of λ_0) :

> When we retain the null hypothesis, our reason is not that it has emerged from the test with a high posterior probability, or even that it has accounted well for the data. H_0 is retained for the totally different reason that if the most sensitive available test fails to detect its existence, the new effect $(\lambda - \lambda_0)$ can have no observable consequences. That is, we are still free to adopt the alternative H_1 if we wish to; but then we shall be obliged to use a value of λ so close to the previous λ_0 that all our resulting predictive distributions will be indistinguishable from those based on H_0.

For example, there can be no practical value in adding a set of variables that leaves the predictive variance unchanged and changes the predictive mean only in the third decimal place of a variable that is observed to one decimal place, even if the regression coefficient is highly significant. Examination of the effects on the predictive density is therefore a way of avoiding the problem, which arises in hypothesis testing, of including a variable whose coefficient is highly significant, perhaps because of a large sample size, but which has only a negligible effect on the dependent variable.

We conclude that, for the Bayesian, the natural basis on which to judge whether a model is satisfactory is the predictive density. Since this density is defined over observations, the researcher can judge whether the differences between one model and another are minor or major. Although it is usually impossible in a parametric model to observe parameters directly, the data are observed. A Bayesian's statements about the world are conditioned on the data, and the predictive density is such a statement. Hence, in considering model selection, it is natural for the Bayesian to turn to the predictive density to decide whether to include a subset of variables. A difference between predictive densities indicates whether two models imply any real differences about the world, although neither may in fact be true.

Our main results are derived in Section 2; Section 3 presents examples; and Section 4 contains our conclusions.

2 Predictive Densities

We divide the explanatory variables in a regression into two sets, X_1 and X_2. The full model is specified as

$$y = X_1\beta_1 + X_2\beta_2 + e, \tag{1}$$

and the predictive distribution for y_0 given $X_0 = (X_{01}, X_{02})$ is compared with the predictive distribution for y_0 given only X_{01} from the partial model

$$y = X_1\beta_1 + e_1. \tag{2}$$

Dimensions for vectors and matrices are as follows: y, e, and e_1 are $T \times 1$, X_1 is $T \times k_1$, X_2 is $T \times k_2$, β_1 is $k_1 \times 1$, β_2 is $k_2 \times 1$, y_0 is $n \times 1$, X_0 is $n \times (k_1 + k_2)$, X_{01} is $n \times k_1$, and X_{02} is $n \times k_2$. We assume that $e \sim \mathcal{N}(0, \sigma^2 I)$ and $e_1 \sim \mathcal{N}(0, \sigma_1^2 I)$.

Throughout the paper we assume that all variables have zero means. In most cases this is achieved by defining variables as deviations from their sample means. When the data are a pooled cross section-time series, the variables are defined as deviations from cross-section and time means. Degrees of freedom are adjusted accordingly.

Zellner [17, pages 72–74] shows that the predictive density conditional on X_0 for the model

$$y = X\beta + u, \tag{3}$$

where X is $T \times k$, β is $k \times 1$, u is $T \times 1$, $u \sim \mathcal{N}(0, \sigma^2 I)$ and the (noninformative) prior for (β, σ) is proportional to $1/\sigma$, is a multivariate Student t form with

$$E(y_0|X_0) = X_0\hat{\beta} \tag{4}$$

when $\nu = T - k - 1 > 1$ and covariance matrix

$$\text{Cov}(y_0|X_0) = \frac{\nu}{\nu - 2}s^2[I + X_0(X'X)^{-1}X_0'] \tag{5}$$

when $\nu > 2$. In these equations, νs^2 is the sum of squared residuals, and $\hat{\beta}$ is the least squares estimator of β in (3).

In view of our introductory remarks, it is possible that previous empirical work has established prior information regarding β_1. In that case, the above results are easily extended to the case where $\beta_1 \sim \mathcal{N}(\bar{\beta}_1, A)$, and β_2 may have a diffuse or an informative prior.

Applying (4) to (1) and (2), it follows that

$$E(y_0|X_0) - E(y_0|X_{01}) = (X_{02} - X_{01}\hat{\Gamma}_1)\hat{\beta}_2 = \Delta'\hat{\beta}_2, \tag{6}$$

where $\hat{\Gamma}_1 = (X_1'X_1)^{-1}X_1'X_2$ and $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1}X'y$. Note the role played by $\Delta' = X_{02} - X_{01}\hat{\Gamma}_1$. It equals zero when $X_{02} = X_{01}\hat{\Gamma}_1$, which implies that the point at which the prediction is to be made lies on the regression of X_2 on X_1 in the original sample. Thus, a value of Δ far from zero indicates that the prediction is to be made at a value of X that is quite different from those in the original sample. This relationship is the Bayesian counterpart of the principle that multicollinearity does not affect prediction if it persists in the period for which predictions are to be made.

From the sampling-theory viewpoint, $\Delta'\beta_2$ is the bias in the least squares prediction of y when the true model is (1), but the statistician estimates on the basis of (2); $\Delta'\hat{\beta}_2$ can therefore be viewed as an estimate of the bias. When $n = 1$, (5) is proportional to the sampling-theory estimator for the variance of the prediction. A tradeoff of bias and variance plays a role in several sampling-theory approaches to model selection. The above comments can be extended to cover the case of general linear restrictions of the form $R\beta = r$, of which $\beta_2 = 0$ is a special case.

The effect of including X_2 on the predictive covariance matrix is indicated by the generalized variance ratio (GVR), defined as the ratio of the determinants of the two covariance matrices:

$$\begin{aligned} \text{GVR} = \frac{|\text{Cov}(y_0|X_{01})|}{|\text{Cov}(y_0|X_0)|} \quad &= \quad \left(\frac{\nu_1 - k_2 - 2}{\nu_1 - 2}\right)^n \left(1 + \frac{\Delta R^2}{1 - R^2}\right)^n \\ &\times \quad \frac{|I + X_{01}(X_1'X_1)^{-1}X_{01}'|}{|I + X_0(X'X)^{-1}X_0'|}, \end{aligned} \tag{7}$$

where R^2 is determined from the regression of y on X_1 and X_2, ΔR^2 is the increase in R^2 from adding X_2 to the regression of y on X_1, and $\nu_1 = T - k_1 - 1$. When $n = 1$ the GVR reduces to the ratio of the two predictive variances.

It may be seen that (7) is the product of three terms, the first two of which are independent of the point at which the prediction is to be made. The first depends on the degrees of freedom for the two equations and is always less than one. The second, which is greater than or equal to one, is a function of the improvement in the fit of the regression when X_2 is included. It can be written in terms of the two R^2,

$$\left(\frac{1-R_1^2}{1-R^2}\right)^n,$$

where R_1^2 is based on X_1 only, or in terms of the F statistic for testing the hypothesis that $\beta_2 = 0$,

$$\left(1 + F\frac{k_2}{T-k_1-k_2-1}\right)^n.$$

The third term depends on the point at which the prediction is to be made. It is less than or equal to one, as may be seen from rewriting it as

$$\frac{|I + X_{01}(X_1'X_1)^{-1}X_{01}'|}{|I + X_{01}(X_1'X_1)^{-1}X_{01}' + \Delta'[X_2'(I-H_1)X_2]^{-1}\Delta|}, \tag{8}$$

where $H_1 = X_1(X_1'X_1)^{-1}X_1'$. It is instructive to consider several special cases.

Suppose first that $\Delta = 0$ and $n = 1$. Then the third term equals one, and GVR can be written in terms of degrees of freedom and adjusted R^2s; it is therefore related to several of the variable selection rules described in Judge, et al. [13, section 21.2] or in Amemiya [2]. For example, C_p is a linear transformation of GVR:

$$C_p = \frac{(T-k_1-2)(T-k_1-k_2-1)}{T-k_1-k_2}\mathrm{GVR} + 2k_1 - T.$$

Suppose next that we are considering prediction at the observed design matrix, i.e. $X_0 = X$. Then by applying Theorem A.8.2 in Judge, et al. [13, page 950] it is easy to see that the third term in (7) equals 2^{-k_2}. Again (7) depends only on degrees of freedom and the R^2s, and the GVR is thus related to the selection rules mentioned above.

The GVR can also be related to the posterior odds criterion and the Bayes factor; see Zellner [18], [19], Zellner and Siow [20], or Smith and Spiegelhalter [16]. This may be regarded as the ratio of the predictive densities for the two models evaluated at the observed sample points. For the problem we are considering, with the additional assumption that both models have the same prior probability, the posterior odds ratio for model (1) relative to model (2) is

$$\left(\pi^{k_2/2}(\nu_1-k_2)^{k_2/2}\frac{\Gamma(\nu_2/2)}{\Gamma(\nu_1/2)}\right)\left(1+\frac{\Delta R^2}{1-R^2}\right)^{\nu_1/2}\left(\frac{|X_1'X_1|}{|X'X|}\right)^{1/2}s^{k_2}.$$

The first of these terms involves degrees of freedom, the second the relative R^2 (as in the GVR), the third the design matrices, and the fourth term, which has no counterpart in the GVR, is a function of the residual variance from the full model.

Another special case is to set $X_{01} = 0$, $k_2 = 1$, and $n = 1$. Then

$$\frac{\mathrm{Var}(y_0|0)}{\mathrm{Var}(y_0|0, X_{02})} = \left(\frac{T - (k_1 + 4)}{T - (k_1 + 3)}\right) \left(1 + \frac{\Delta R^2}{1 - R^2}\right) \left(1 + \frac{X_{02}^2 / X_2' X_2}{1 - R_{2\cdot 1}^2}\right)^{-1},$$

where $R_{2\cdot 1}^2$ is the R^2 in the regression of X_2 on X_1. The term $1/(1 - R_{2\cdot 1}^2)$ is called the "variance inflation factor" in the multicollinearity literature. If $X_{02}^2 = X_2' X_2 / \nu$ and $\nu_1/\nu \approx 1$, then GVR ≈ 1 when the F statistic for testing the hypothesis that $\beta_2 = 0$ is approximately equal to the variance inflation factor.

As a final special case, assume that $X_0 = x_i'$, where x_i' is the ith row of X; i.e. the prediction is to take place at the observed value of the ith observation. Then the third term in (7) can be written in terms of the ith diagonal element of the "hat" matrices, $X(X'X)^{-1}X'$ and $X_1(X_1'X_1)^{-1}X_1'$, as

$$\frac{1 + x_{1i}'(X_1'X_1)^{-1}x_{1i}}{1 + x_i'(X'X)^{-1}x_i} = \frac{1 + h_{1i}}{1 + h_i}. \tag{9}$$

In Belsley, Kuh, and Welsch [3] the h_i are extensively used to find observations that have a large effect on parameter estimates (influential observations). See also Johnson and Geisser [12]. Thus, in terms of this diagnostic tool, (7) can be written as the product of a degrees of freedom term, a term in the relative goodness of fit of the two models, and a term in the relative degree of influence of x_i in the two models. Let us consider this third term in a bit more detail. Since it is less than or equal to 1, we have $h_i \geq h_{1i}$. When $n = 1$, we have from (8) and (9),

$$\begin{aligned} h_i - h_{1i} &= \Delta'[X_2'(I - H_1)X_2]^{-1}\Delta \\ &= \frac{(x_{2i}' - x_{1i}'\hat{\Gamma}_1)(x_{2i}' - x_{1i}'\hat{\Gamma}_1)'}{\sum_i (x_{2i}' - x_{1i}'\hat{\Gamma}_1)(x_{2i}' - x_{1i}'\hat{\Gamma}_1)'}. \end{aligned}$$

Thus, predicting at a point where Δ is large tends to reduce the variance of the partial model relative to that of the full model.

Two other matters are addressed before we turn to empirical examples: the information criterion and the relation between Δ and GVR. Johnson and Geisser [12], who are concerned with a predictive view of detecting influential observations, propose the use of the Kullback-Liebler (K-L) divergence to measure discrepancies between densities; see Kullback and Liebler [14]. In the case of two normal distributions, the K-L divergence for comparing f_1 and f_2 is given by

$$I(f_2, f_1) = \frac{1}{2}[(\mu_2 - \mu_1)'\Sigma_1^{-1}(\mu_2 - \mu_1) + \mathrm{tr}\Sigma_2\Sigma_1^{-1} - \ln|\Sigma_2\Sigma_1^{-1}| - M], \tag{10}$$

where the μ_i are the means and the Σ_i the covariance matrices of the M dimensional normal densities being compared. The information measure has the virtue of combining differences between means and covariance matrices into one measure.

To explain the appearance of the graphs presented below, we conclude this section with a few remarks on the relation between $\Delta'\hat{\beta}_2/\sigma_{\hat{y}_{01}}$ and GVR when $n = 1$, where

$$\sigma_{\hat{y}_{01}} = \sqrt{\frac{\nu_1}{\nu_1 - 2}} s_1 \sqrt{1 + X_{01}(X_1'X_1)^{-1}X_{01}'}$$

is the standard error of prediction for model (2) at X_{01}. Consider first the case $k_2 = 1$. The first two terms in (7) are denoted by K, which permits us to write, using (8),

$$\mathrm{GVR} = \frac{K}{1 + a(\Delta'\hat{\beta}_2/\sigma_{\hat{y}_{01}})^2}, \tag{11}$$

where

$$a = \frac{\nu_1 s_1^2}{(\nu_1 - 2)\hat{\beta}_2^2 X_2'(I - H_1)X_2}.$$

Thus the points in a plot of GVR on $(\Delta'\hat{\beta}_2/\sigma_{\hat{y}_{01}})^2$, where the variation is in $\Delta/\sigma_{\hat{y}_{01}}$, lie on equation (11).

For general k_2 and $n = 1$, we establish an inequality to show that a similar equation acts as an upper bound to GVR for a given value of the standardized difference in means. Let

$$\mathrm{GVR} = \frac{K}{1 + b\Delta' M_1 \Delta/\sigma_{\hat{y}_0}^2},$$

where $b = \nu_1 s_1^2/(\nu_1 - 2)$ and $M_1 = [X_2'(I - H_1)X_2]^{-1}$. The square of the standardized difference in means is $\Delta'\hat{\beta}_2\hat{\beta}_2'\Delta/\sigma_{\hat{y}_{01}}^2$. Since the ranks of M_1 and $\hat{\beta}_2\hat{\beta}_2'$ are k_2 and 1, respectively, and both are symmetric there exist matrices R and Q, $Q'Q = I$, such that $R'M_1R = I = Q'R'M_1RQ$ and $'R'\hat{\beta}_2\hat{\beta}_2'RQ = \begin{pmatrix} \omega & 0 \\ 0 & 0 \end{pmatrix}$, where ω is a scalar. Then, defining u by $\Delta = RQu$, we have $\Delta'\hat{\beta}_2\hat{\beta}_2'\Delta = \omega u_1^2$ and $\Delta' M_1 \Delta = \sum_1^{k_2} u_i^2$. Therefore,

$$\begin{aligned} \mathrm{GVR} &= \frac{K}{1 + b\Delta' M_1 \Delta/\sigma_{\hat{y}_{01}}^2} && (12) \\ &= \frac{K}{1 + b\sum u_i^2/\sigma_{\hat{y}_{01}}^2} && (13) \\ &\leq \frac{K}{1 + bu_1^2/\sigma_{\hat{y}_{01}}^2} && (14) \\ &= \frac{K}{1 + \frac{b}{\omega}\Delta'\hat{\beta}_2'\hat{\beta}_2\Delta/\sigma_{\hat{y}_{01}}^2}. && (15) \end{aligned}$$

3 Examples

In the examples that follow we present several graphical and numerical devices to compare predictive means and GVRs of various models. These include summary statistics, scatter plots of GVR and $\Delta'\hat{\beta}_2/\sigma_{\hat{y}_{01}}$, K-L statistics, and overlap statistics. Before defining the latter, we note that Marriott et al. [15] employ a somewhat different graphical device. They present scatter plots of (in our notation) $|y_i - E(y_i|y, X)|$ on $[\text{Var}(y_i|y, X)]^{1/2}$. The expectation and variance for each observation are computed for several specifications of ARMA(p, q) errors in a linear regression model. Models for which the points are close to zero in both dimensions are considered preferable.

Our overlap statistic measures the extent to which 95% confidence intervals computed at the sample X values for a model containing all explanatory variables overlap confidence intervals from a model containing a subset. Let a confidence interval for the full model be (a, b) and that of the subset model be (c, d). Then the overlap of the intervals is r/s, where $s = \max(b, d) - \min(a, c)$ and r is defined as:

$$r = \begin{cases} d - c & \text{if } c > a \text{ and } b > d, \\ b - a & \text{if } a > c \text{ and } d > b, \\ \min(d - a, b - c) & \text{otherwise.} \end{cases} \tag{16}$$

If r/s is large, there is little difference in the position and length of the confidence intervals generated by the models, and if it is small there is a large difference. It is easy to see that $0 \leq r/s \leq 1$. Note that the Bayesian and sampling-theory confidence intervals coincide since we have assumed diffuse priors for the parameters.

3.1 Simulated data

Our first example is based on constructed data. The model is

$$y_i = b_1 x_{1t} + x_{2t} + x_{3t} + x_{4t} + u_t, \quad t = 1, \ldots, 30,$$

where u_t is a pseudo-random normal variate with mean 0 and variance 100; x_{jt} for $j = 1, 2, 3, 5$ are pseudo-random variates with mean 100 and variance 400;

$$x_{4t} = (x_{2t} + x_{3t})/2 + w_t,$$

where w_t is a pseudo-random normal variate with mean 0 and variance 1/900; and b_1 takes the values 0.1, 1.0, 2.0, and 5.0. Note that x_{5t} does not enter the model generating the dependent variable but is included in the regressions.

For this example we use changes in predictive densities to compare various models, rather than specify an X_1 and X_2, since there is no relevant theory. We calculate the mean posterior prediction for each observation, its standard error, the difference in the predictions from the 'full' model (using x_1 to x_5), the K-L statistics, and the overlap statistics. The various statistics are calculated for the full model and for models with 1, 2, 3, and 4 variables deleted. Three lists of the variables were specified to determine the effects of deleting variables in different orders (with the last variable deleted first). They are:

1. x_1, x_2, x_3, x_4, x_5;

2. x_1, x_2, x_5, x_3, x_4;

3. x_2, x_3, x_4, x_1, x_5.

Figure 1 (page 22) shows the effects of deleting variables from list 1 with $b_1 = .1$; the data points are labeled with the number of variables removed from the model, and GVR is plotted against $\Delta'\hat{\beta}_2/\sigma_{\hat{y}_{01}}$, the standardized difference in the mean of the posterior distribution. The standardization is by the standard deviation of the reduced model (hence the standardization changes between models). From Figure 1a we see that the curve defined by (11) is clearly indicated; the figure also reveals only very small changes in standardized means and a GVR that is close to one for most observations. Figure 1b displays the effect of deleting 2 variables (x_5 and x_4), while Figures 1c and 1d, respectively, show the additional effects of removing x_3 and x_2. Note the large changes in both the GVR and the mean differences when first x_3 and then x_2 are deleted. These effects are consistent with the true model: x_5 does not appear in it, and x_4 is almost a linear combination of x_2 and x_3.

In Figure 2a (page 23) we combine the graphs of Figure 1. Although the scale of the GVR axis is now compressed, the graph clearly reveals the slight effect on both standardized mean difference and GVR from dropping first x_5 and then x_4. In contrast, the points labeled ‘3’ (resp. ‘4’) show that removing x_3 (x_2) has a noticeable effect. Figure 2b shows the effects of removing the variables in a different order. It can be clearly seen that removing x_4 has little effect; removing x_3 has a noticeable effect; removing x_5 has no effect; and removing x_2 has a pronounced effect. Again, these results are consistent with our knowledge of the true model. In Figures 2c and 2d, we set $b_1 = 5$ and compare deleting variables according to lists 1 and 2. It can be seen that the size of the coefficient of x_1 has no effect on results. We shall see below, however, that this coefficient does play a role when variables are removed according to list 3.

In some applications it may be of interest to know whether particular variables affect particular observations in an important way. Accordingly, we present Figure 3 (page 24), which repeats Figure 2 except that the points are unlabeled and lines connect the observations. These graphs show how the GVR and normalized deviations change for a particular observation as the model is varied. For example, the standardized difference of the observation on the far left of each graph increases as more variables are removed, while the line on the right shows that the standardized difference decreases for that observation as the fourth variable is removed.

Figure 4 (page 25) compares the results of removing variables according to lists 1 and 3. Figures 4a and 4c are identical to Figures 1a and 1c. Figures 4b and 4d reveal again

Table 1: Overlap statistics, List 1, All b_1

	Omitted variables			
Statistic	x_5	x_4, x_5	x_3, x_4, x_5	x_2, x_3, x_4, x_5
Mean	0.965446	0.923488	0.341212	0.223044
Max	0.979343	0.977273	0.373777	0.241762
Q3	0.977273	0.951598	0.351753	0.234447
Med	0.971485	0.932216	0.349555	0.228276
Q1	0.959846	0.901245	0.340885	0.223308
Min	0.920340	0.823271	0.187333	0.126265

Table 2: Overlap statistics, List 2

	Omitted variables			
Statistic	x_5	x_1, x_5	x_1, x_4, x_5	x_1, x_3, x_4, x_5
	$b_1 = .1$			
Mean	0.965446	0.7057	0.710283	0.313172
Max	0.979343	0.798292	0.816929	0.342713
Q3	0.977273	0.763943	0.78585	0.331103
Med	0.971485	0.752413	0.765924	0.322065
Q1	0.959846	0.666576	0.650816	0.316346
Min	0.92034	0.411592	0.450845	0.141848
	$b_1 = 5$			
Mean	0.965446	0.096244	0.098293	0.090441
Max	0.979343	0.103856	0.107239	0.100961
Q3	0.977273	0.100083	0.104496	0.098135
Med	0.971485	0.098411	0.101674	0.095215
Q1	0.959846	0.097233	0.100281	0.093283
Min	0.92034	0.013165	-0.01965	-0.06969

what we already know: removing x_5 and x_4 has little effect on the predictive distribution. Note, however, the role played by the coefficient of x_1. In Figure 4b, where the coefficient is 0.1, removing x_1 has little effect, whereas Figure 4d clearly shows that removing the same variable when the coefficient is 5.0 has a large effect. Hence, the effect depends both on the size of coefficient and the order in which the variable is removed.

We present summaries of the overlap statistics for various models in Tables 1 to 3. Table 1 reveals that the average overlap remains over .9 when x_5 and x_4 are excluded, then drops precipitously when x_3 is excluded, and drops somewhat less when x_2 is also excluded. Table 2 indicates the sensitivity of results to the size of b_1: the mean overlap drops much more for b_1 equal to 5.0 then when it equals 0.1. Table 3 reveals a large effect on the average overlap from excluding x_3 when x_4 has already been eliminated.

Figure 5 (page 26) plots the confidence intervals that were used to construct the overlap statistics for $b_1 = 0.1$ and list 1. The horizontal axis is an observation label, and the vertical axis is the confidence interval for each data point. There are 5 confidence intervals for each data point—one for the full model, and one for each model deleting one, two, three,

Table 3: Overlap Statistics, List 3, All b_1

	Omitted variables			
Statistic	x_4	x_3, x_4	x_3, x_4, x_5	x_2, x_3, x_4, x_5
Mean	0.932858	0.347052	0.341212	0.223044
Max	0.990849	0.373699	0.373777	0.241762
Q3	0.974193	0.353109	0.351753	0.234447
Med	0.940615	0.347558	0.349555	0.228276
Q1	0.902701	0.343195	0.340885	0.223308
Min	0.823524	0.276110	0.187333	0.126265

Table 4: K-L Statistics, List 1, All b_1

Omitted Variables	KLA[a]	KL1[b]	Variance[c]
x_2, x_3, x_4, x_5	30890.56	61780.80	1992.89
x_3, x_4, x_5	11571.12	23141.99	746.49
x_4, x_5	60.92	121.68	3.91
x_5	1.74	3.39	0.10

[a] Reference distribution is full model.
[b] Reference distribution is partial model.
[c] Variance of predicted y_t.

or four variables. The observations are ordered by their mean prediction, which is labeled with a small horizontal bar. Deletion of relevant variables should increase the confidence intervals, and deletion of irrelevant (x_5) or collinear (x_4) variables should not change the confidence intervals much. The figure reveals that deleting the first two variables in list 1 does not have much effect, but deleting x_3 and then x_2 does, as is shown by the much wider confidence intervals. The confidence interval graphs tell the same story that we have seen in Figures 1, 2, and 3, and in the overlap tables—influential variables show large effects.

Although the K-L statistics of Tables 4, 5, and 6 yield conclusions similar to those reported above about the importance of the variables, it is difficult to know what represents a large change in these numbers since the K-L statistic has a minimum of 0 and no maximum. In addition, the K-L statistic depends upon which model is treated as the reference distribution. (In these tables, KLA denotes the full model treated as the reference, and KL1 denotes the submodel treated as the reference.) For comparison purposes we also present the variance of the difference in the predicteds. For all values of b_1 and list 1 the KL statistics are contained in Table 4. Table 5 indicates the effects of different values of b_1. For the third list, the values are the same across values of b_1; see Table 6. Note that both KL statistics and the variance jump more than two orders of magnitude when an important variable is dropped. Thus, in this example all three measures provide similar information.

We conclude from this example that the nature of the model that generated the data is revealed clearly by scatter plots of GVR and $\Delta'\hat{\beta}_2$, the overlap statistics and graphs, and the K-L values. For the next two examples, we confine our attention to the scatter plots and the

Table 5: K-L Statistics, List 2

Omitted Variables	KLA[a]	KL1[b]	Variance [c]
	b_1 = .5		
x_1, x_3, x_4, x_5	14652.74	29305.17	945.29
x_1, x_4, x_5	1241.84	2483.43	80.08
x_1, x_5	1201.03	2401.90	77.46
x_5	1.74	3.39	0.10
	b_1 = 1		
x_1, x_3, x_4, x_5	20645.80	41291.28	1331.94
x_1, x_4, x_5	5449.11	10897.99	351.52
x_1, x_5	5424.98	10849.80	349.98
x_5	1.74	3.39	0.10
	b_1 = 2		
x_1, x_3, x_4, x_5	42112.03	84223.75	2716.86
x_1, x_4, x_5	23009.03	46017.83	1484.42
x_1, x_5	23005.10	46010.04	1484.18
x_5	1.74	3.39	0.10
	b_1 = 5		
x_1, x_3, x_4, x_5	182351.70	364703.00	11764.57
x_1, x_4, x_5	148851.70	297703.10	9603.30
x_1, x_5	148803.30	297606.40	9600.20
x_5	1.74	3.39	0.10

[a]Reference distribution is full model.
[b]Reference distribution is partial model.
[c]Variance of predicted y_t.

Table 6: K-L Statistics, List 3, All b_1

Omitted Variables	KLA[a]	KL1[b]	Variance[c]
x_2, x_3, x_4, x_5	30890.56	61780.80	1992.89
x_3, x_4, x_5	11571.12	23141.99	746.49
x_3, x_4	10741.51	21482.87	692.98
x_4	57.12	114.15	3.67

[a]Reference distribution is full model.
[b]Reference distribution is partial model.
[c]Variance of predicted y_t.

overlap tables.

3.2 Investment Model

This section is based on the work of Fazzari, Hubbard, and Petersen [7] and Fazzari and Petersen [6]. The reader should examine these sources for further details; we present only a brief summary of their model. These authors are concerned with the role of cash flow variables as explanations for firm investment in new plant and equipment. Taking a Tobin-q hypothesis as the basic model, they study whether the addition of a cash flow variable (CF_t) helps to explain the fixed investment to capital ratio (I_t/K_t) in a cross-section of firms. In addition, they examine a variant of the basic model in which sales and lagged sales (S_t, S_{t-1}) are also included. Firms are grouped by their dividend payout ratios as a method of determining which are likely to be cash constrained and therefore particularly sensitive to cash flows. We utilize a sample of 443 observations, consisting of 37 to 48 firms over 10 years; these are firms that paid little or no dividends during the year.[1] Individual firm and time dummy variables are included to eliminate unobservable firm differences and common time effects, but are not reported upon. The x_i values for within sample predictions are therefore net of firm and time means.

Tables 7–10 present regression results for several variants of the model. Of particular interest is the statistical significance of CF_t in an equation that includes q_t, S_t, and S_{t-1}. When we turn to the predictive analysis, however, CF_t seems to be somewhat less central. Figures 6 and 7 (pages 27 and 28) display GVR against standardized differences in means at the sample values of X. As above, a '1', '2', or '3' represents an observation, and '1' indicates that CF_t has been removed from the full model, '2' indicates that CF_t and S_{t-1} have been removed, and '3' indicates that CF_t, S_{t-1}, and S_t have been removed. Note that there is little effect in either predictive means or GVR when CF_t is omitted. For only two observations does the standardized predictive mean change by more than one standard deviation. It may be seen that removing S_{t-1} has little effect on most observations, but a considerable effect is evident when S_t is removed. Thus, although the coefficient of CF_t is highly significant in the full model, its presence in an equation that already contains q_t, S_t, and S_{t-1} does not greatly affect the predictive densities evaluated at the observed X matrix. Figure 7 reveals that one observation is clearly an outlier. It can also be seen that for a number of observations GVR decreases upon removing S_{t-1} when CF_t has already been removed. By analogy to the simulated data, it appears that S_{t-1} for those observations is an inconsequential variable. Further analysis might be done to separate those firms to determine how they differ from the rest of the firms which have a GVR increase.

The overlap statistics reported in Tables 11–13 suggest a similar conclusion. Table 11 compares confidence intervals based on the full model and the full model without CF_t. The average overlap is .93, and the median is .96; 75% of the observations yield an overlap

[1] We are grateful to Professor Steven Fazzari for making these data available to us.

Table 7: Regression analysis: q model

SSE	DFE	MSE	RMSE	RSQUARE
10.16142	384	0.023467	0.153191	0.248068

VARIABLE	BETA	STDERR	TRATIO	PROBT
q_t	0.007861	0.000777	10.11823	0

$\hat{\Gamma}_1$			
	S_t	S_{t-1}	CF_t
q_t	0.095145	0.073975	0.008587

Table 8: Regression analysis: q_t and S_t model

SSE	DFE	MSE	RMSE	RSQUARE
8.369626	383	0.019374	0.139191	0.380658

VARIABLE	BETA	STDERR	TRATIO	PROBT
q_t	0.003706	0.000828	4.477794	9.664E-6
S_t	0.04367	0.004541	9.616853	0

$\hat{\Gamma}_1$		
	S_{t-1}	CF_t
q_t	0.017864	0.003188
S_t	0.589738	0.056753

Table 9: Regression analysis: q_t, S_t, and S_{t-1} model

SSE	DFE	MSE	RMSE	RSQUARE
7.531083	382	0.017474	0.132187	0.44271

VARIABLES	BETA	STDERR	TRATIO	PROBT
q_t	0.004714	0.000799	5.897797	7.452E-9
S_t	0.076962	0.006457	11.91904	0
S_{t-1}	-0.05645	0.008149	-6.92744	1.57E-11

$\hat{\Gamma}_1$	
	CF_t
q_t	0.005022
S_t	0.117323
S_{t-1}	-0.102710

Table 10: Regression analysis: Full model

SSE	DFE	MSE	RMSE	RSQUARE
7.066109	381	0.016433	0.128191	0.477117

VARIABLES	BETA	STDERR	TRATIO	PROBT
q_t	0.003625	0.000802	4.521252	7.959E-6
S_t	0.051514	0.00788	6.537244	1.78E-10
S_{t-1}	-0.03418	0.008944	-3.821080	.000152
CF_t	0.216901	0.040776	5.319348	1.679E-7

Table 11: Overlap statistics, q_t, S_t and S_{t-1} model

N		443	
Mean	0.932002	Sum	412.8767
Std Dev	0.069894	Variance	0.004885
Skewness	-4.42321	Kurtosis	31.55166

Quintiles

Max	0.969835
Q3	0.969775
Med	0.960839
Q1	0.919392
Min	0.211606

Table 12: Overlap statistics: q_t and S_t model

N		443	
Mean	0.880007	Sum	389.843
Std Dev	0.099071	Variance	0.009815
Skewness	-3.94582	Kurtosis	19.79354

Quintiles

Max	0.945768
Q3	0.921175
Med	0.921019
Q1	0.885834
Min	0.057009

Table 13: Overlap statistics: q_t model

N	443	Sum Wgts	443
Mean	0.785133	Sum	347.8138
Std Dev	0.115624	Variance	0.013369
Skewness	-3.11602	Kurtosis	11.11508

Quintiles

Max	0.868775
Q3	0.837181
Med	0.836919
Q1	0.788278
Min	0.107075

Table 14: Definitions of money demand data (All variables in logs)

Variable	Definition
RTB	Treasury Bill Rate
RSD	Savings Deposit Rate
RSD_{-1}	RSD lagged 1 period
RSD_{-2}	RSD lagged *i* periods
GP	Real Gross National Product
W	Real Wealth
VCC	Value of credit card transactions
PP	Inflation rate
RTB_{-1}	RTB lagged 1 period
RTB_{-2}	RTB lagged 2 periods

greater than .92. These statistics suggest that the CF_t variable has a great effect on the confidence intervals for only a small number of observations.

On the substantive question of the importance of CF_t for investment decisions, it should be noted that we did not examine the effects of changing the deletion order of the variables. Moreover, we have worked only with those firms that paid low dividends; CF_t may have appeared to be more influential if we had included firms with more varied dividend behavior.

3.3 Money Demand Model

In this section we examine a money demand model in the spirit of the work of Cooley and LeRoy [5]. The object of their research is to examine the extent to which the coefficients of focus variables (i.e. those in X_1) depend on the other variables, termed the "doubtful variables," included in the equation. They take as their example money demand over the period 1952.2 to 1978.4. The dependent variable is the logarithm of real money minus the logarithm of real GNP. The focus variables are the Treasury bill rate and the savings and loan passbook rate. The doubtful variables are real GNP, the inflation rate, the real value of credit card transactions, real wealth, and lagged values of these variables. We were not able to duplicate completely the variables used by these authors and estimate over a slightly different time period. Definitions of the variables are presented in Table 14.

Like Cooley and LeRoy, we are interested in how coefficients change as the specification changes. Rather than consider all possible regressions, however, we present results in which the variables are deleted from the bottom up in the order of the list in Table 14. As may be seen in Figure 8 (page 29), the variables seem to fall into three groups as doubtful variables are deleted; within each group the GVR is approximately constant, but it varies from group to group. Lagged values of RSD and RTB constitute the first group; PP is the second group; and VCC, W, and GP is the third. The effects of changing specifications on the coefficients of the focus variables are presented in Table 15. The only consistent sign is that of the sum of the coefficients of RSD, and its lagged values.

Table 15: Coefficients of focus variables

Variables Included	Coefficients of RTB	Coefficients of RSD	Sum of Coefficients[a] RTB	Sum of Coefficients[a] RSD
RTB and RSD	.0159	-.0318	.0159	-.0318
Above and RSD_{-1}	.0155	-.0914	.0155	-.0315
Above and RSD_{-2}	.0154	-.0242	.0154	-.0314
Above and GP	.0210	-.0114	.0210	-.0070
Above and *W*	-.0008	.0013	-.0008	-.0077
Above and VCC	-.0040	.0061	-.0040	-.0068
Above and PP	-.0047	.0049	-.0047	-.0069
Above and RTB_{-1}	.0034	.0044	-.0064	-.0060
Above and RTB_{-2}	.0046	.0046	.0045	-.0061

a. Current and Lagged Values

Table 16: Overlap means and standard deviations as variables are deleted

Variables Deleted	Mean	S.D.
RTB_{-2}	.975	.020
Above and RTB_{-1}	.908	.065
Above and PP	.904	.069
Above and VCC	.814	.085
Above and *W*	.502	.074
Above and GP	.407	.053
Above and RSD_{-1}	.404	.056
Above and RSD_{-2}	.404	.058

Figure 8 reveals little change in the means of the predicted densities as the specification is changed. The first four variables deleted (moving from the bottom up in Table 14) result in changes that remain in the one-standard deviation range. As *W* and GP, RSD_{-2}, and RSD_{-1} are deleted, the differences in means remain in the two-standard deviation band except for one observation. Thus it is clear that the mean of the predictive densities is little affected by the inclusion of the doubtful variables. Table 16 shows that the overlap proportions vary sharply between groups (due to the changes in standard errors), but not very greatly within groups. A sharp break occurs when *W* is omitted, dropping the average overlap from .814 to .502.

In conclusion, although Table 15 reveals rather large changes in coefficients as variables are added, we find that the predicted means are not greatly affected. We thus agree with Cooley and Leroy that little can be said about values of the coefficients of the focus variables from these data. It is, however, noteworthy that the length of confidence intervals depends greatly on the doubtful variables because the coefficient variances are quite sensitive to them.

4 Conclusions

The examples of the previous section suggest that examination of changes in predicted means and of GVRs is a useful supplement to other methods of investigating model specification. Although these concepts arise naturally in the Bayesian view of inference, they have sampling-theory interpretations and are related to concepts that already appear in that literature. The graphical tools we have introduced, however, furnish useful information that does not appear elsewhere. In future work, we plan to investigate the usefulness of the predictive approach to other aspects of model specification.

References

[1] J. Aitchison and I. R. Dunsmore. *Statistical Prediction Analysis*. Cambridge University Press, Cambridge, 1975.

[2] Takeshi Amemiya. Selection of regressors. *International Economic Review*, 21(2):331–354, June 1980.

[3] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics*. John Wiley & Sons, New York, 1980.

[4] Murray K. Clayton, Seymour Geisser, and Dennis E. Jennings. A comparison of several model selection procedures. In P. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques*, chapter 27, pages 425–439. Elsevier Science Publishers B. V., Amsterdam, 1986.

[5] T. F. Cooley and S. F. LeRoy. Identification and estimation of money demand. *American Economic Review*, 71(5):825–844, December 1981.

[6] Steven Fazzari and Bruce Petersen. Investment smoothing with working capital: New evidence on the impact of financial constraints. Washington University, 1992.

[7] Steven M. Fazzari, R. Glenn Hubbard, and Bruce C. Petersen. Financing constraints and corporate investment. *Brookings Papers on Economic Activity*, (1):141–206, 1988.

[8] Kenneth M. Gaver and Martin S. Geisel. Discriminating among alternative models: Bayesian and non-Bayesian methods. In Paul Zarembka, editor, *Frontiers in Econometrics*, chapter Two, pages 49–77. Academic Press, New York, 1974.

[9] S. Geisser. On the prediction of observables: A selective update. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 203–230. Elsevier Science Publishers B. V. (North-Holland), New York, 1985.

[10] S. Geisser and W. F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160, 1979.

[11] E. T. Jaynes. What is the question? In R. D. Rosenkrantz, editor, *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, pages 376–400. Kluwer Academic Publishers, Dordrecht, 1983.

[12] W. Johnson and S. Geisser. A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, 78:137–144, 1983.

[13] George G. Judge, W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee. *The Theory and Practice of Econometrics.* John Wiley & Sons, New York, second edition, 1985.

[14] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[15] John Marriott, Nalini Ravishanker, Alan Gelfand, and Jeffrey Pai. Bayesian analysis of ARMA processes: Complete sampling based inference under full likelihoods. .

[16] A. F. M. Smith and D. J. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, B*, 42(2):213–220, 1980.

[17] Arnold Zellner. *An Introduction to Bayesian Inference in Econometrics.* John Wiley & Sons, New York, 1971.

[18] Arnold Zellner. Jeffreys-Bayes posterior odds ratio and the Akaike information criterion for discriminating between models. *Economics Letters*, 1:337–342, 1978.

[19] Arnold Zellner. Posterior odds ratios for regression hypotheses: General considerations and some specific results. In Arnold Zellner, editor, *Basic Issues in Econometrics*, chapter 3.6, pages 275–305. University of Chicago Press, Chicago, 1984.

[20] Arnold Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 585–603. University Press, Valencia, Spain, 1980.

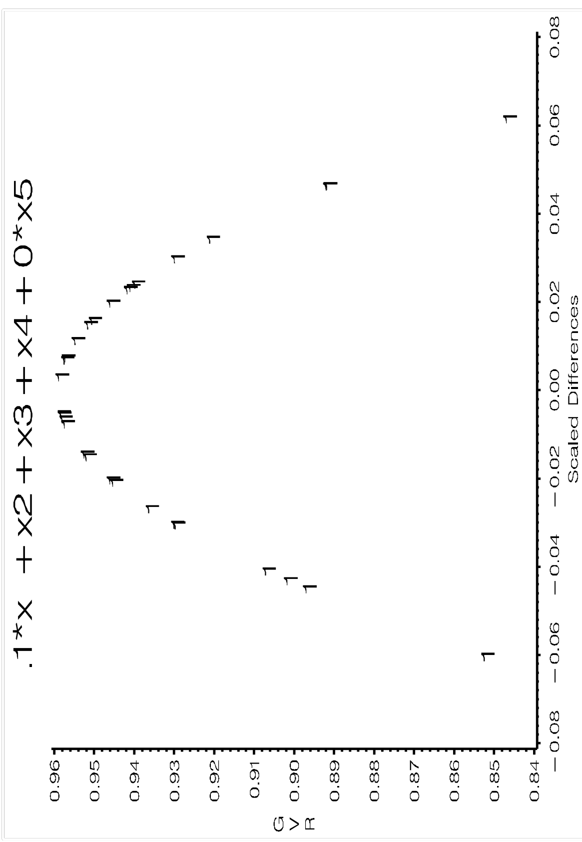


Figure 1A

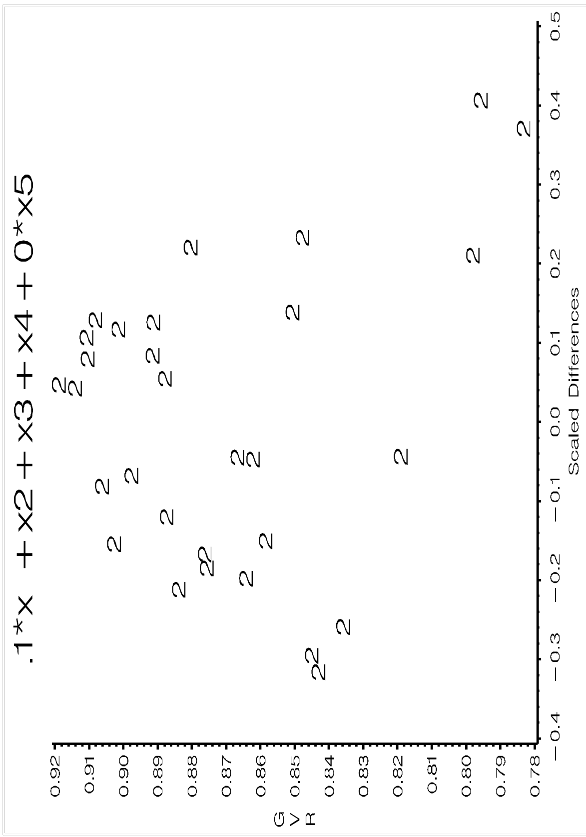


Figure 1B

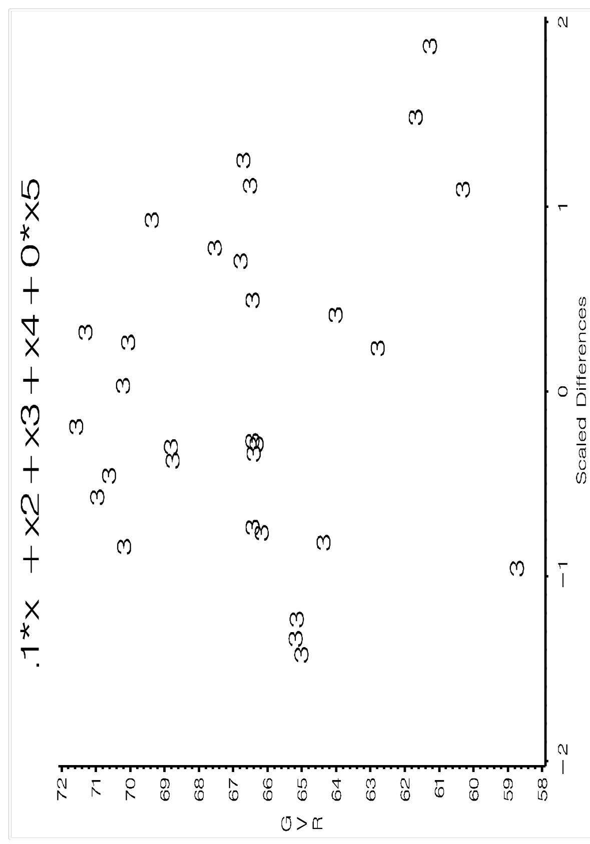


Figure 1C

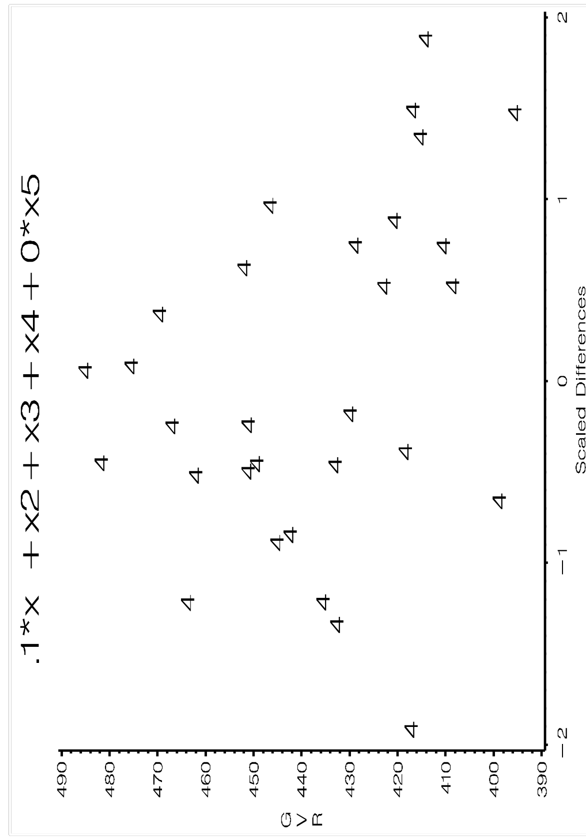


Figure 1D

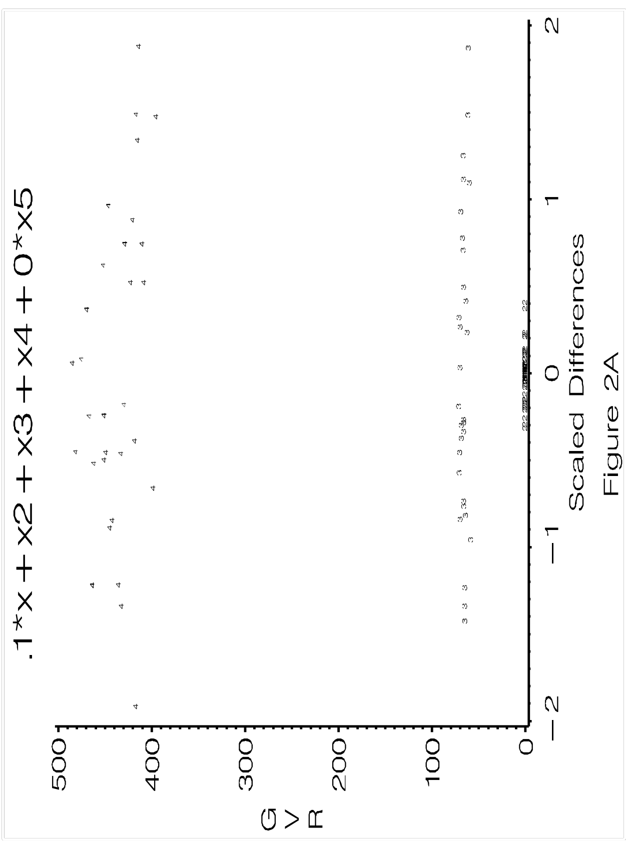


Figure 2A

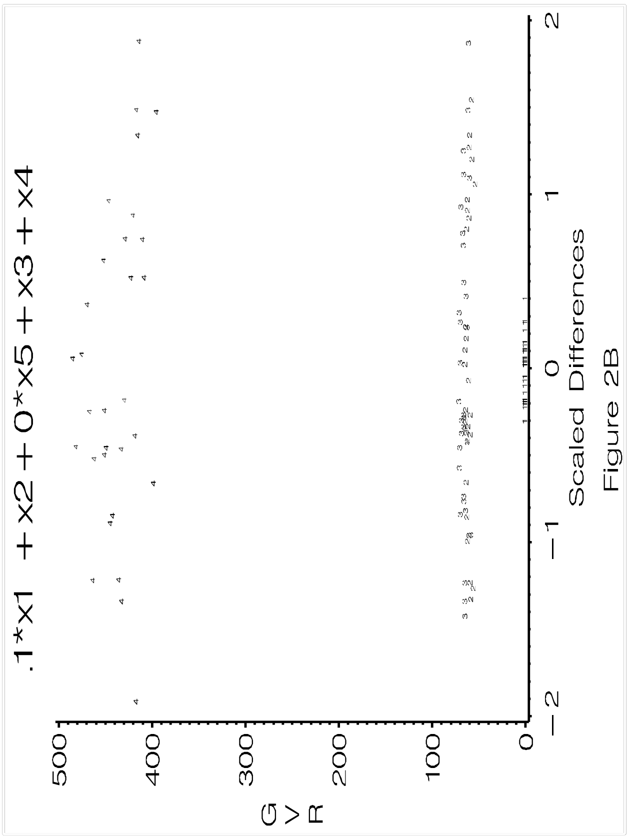


Figure 2B

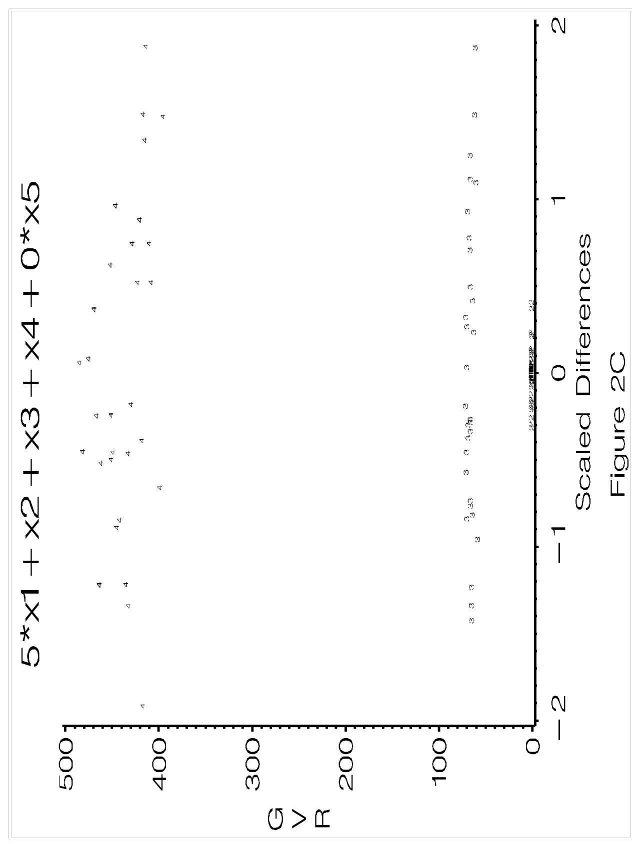


Figure 2C

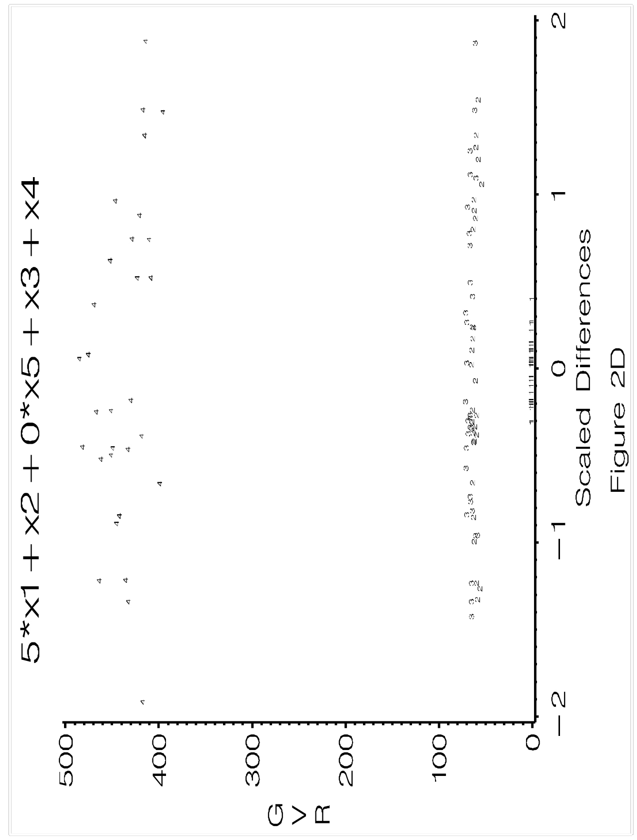
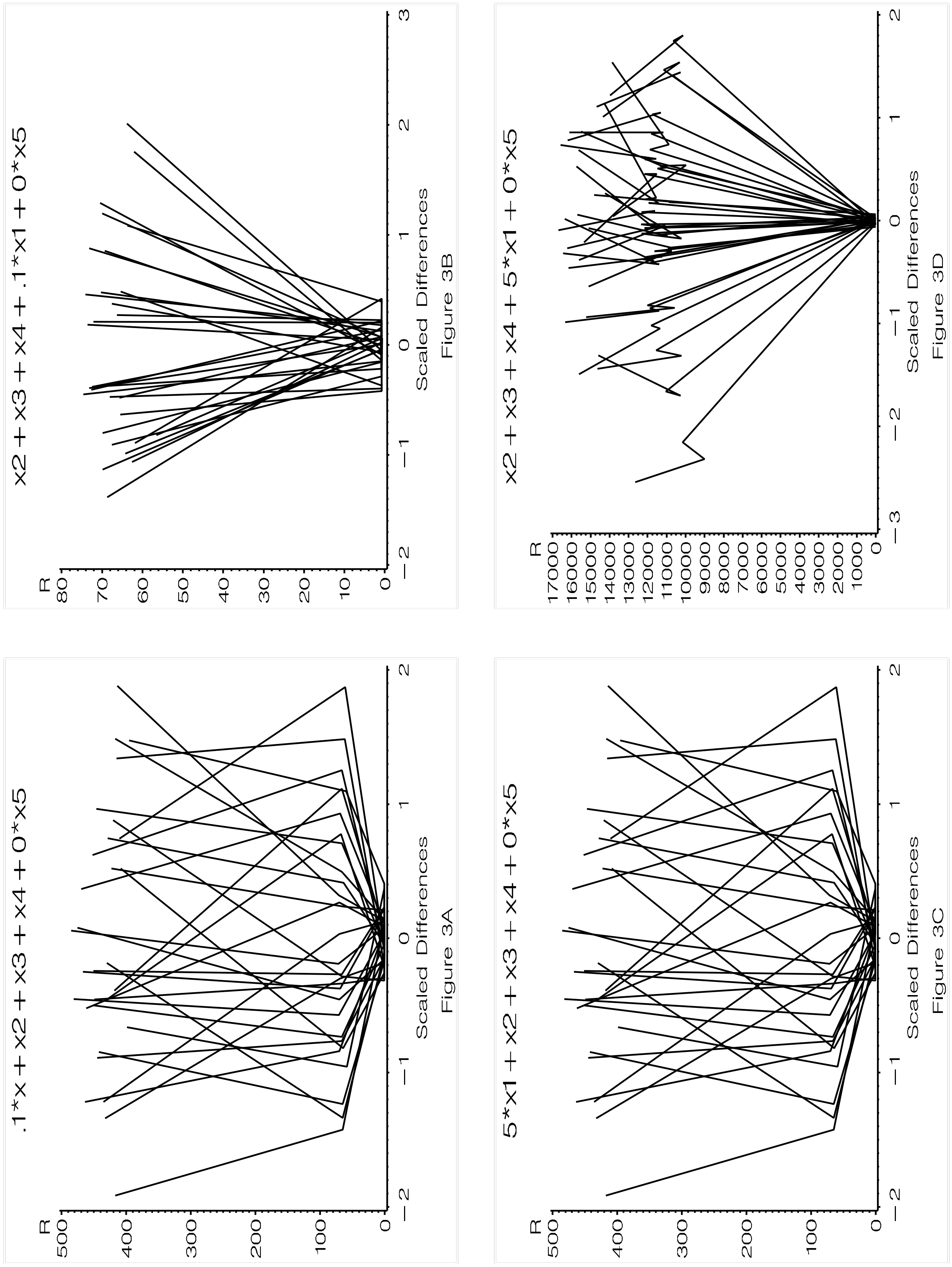


Figure 2D


.1*x + x2 + x3 + x4 + 0*x5
R
500
400
300
200
100
0
−2
−1
0
1
2
Scaled Differences
Figure 3A
x2 + x3 + x4 + .1*x1 + 0*x5
R
80
70
60
50
40
30
20
10
0
−2
−1
0
1
2
3
Scaled Differences
Figure 3B
5*x1 + x2 + x3 + x4 + 0*x5
R
500
400
300
200
100
0
−2
−1
0
1
2
Scaled Differences
Figure 3C
x2 + x3 + x4 + 5*x1 + 0*x5
R
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
−3
−2
−1
0
1
2
Scaled Differences
Figure 3D

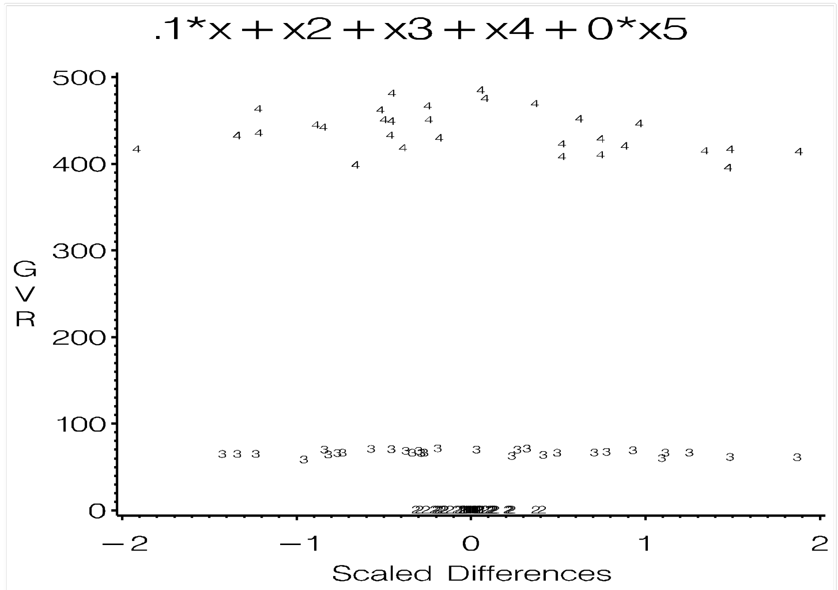


Figure 4A

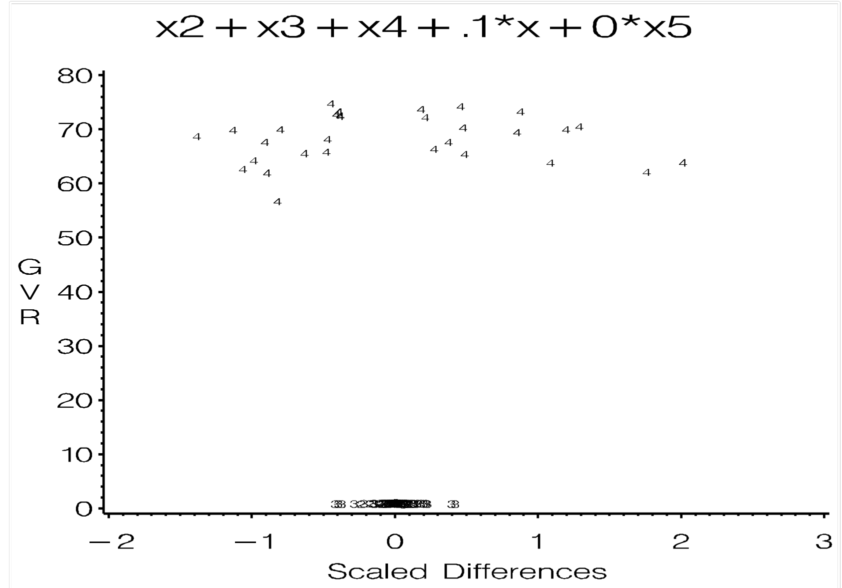


Figure 4B

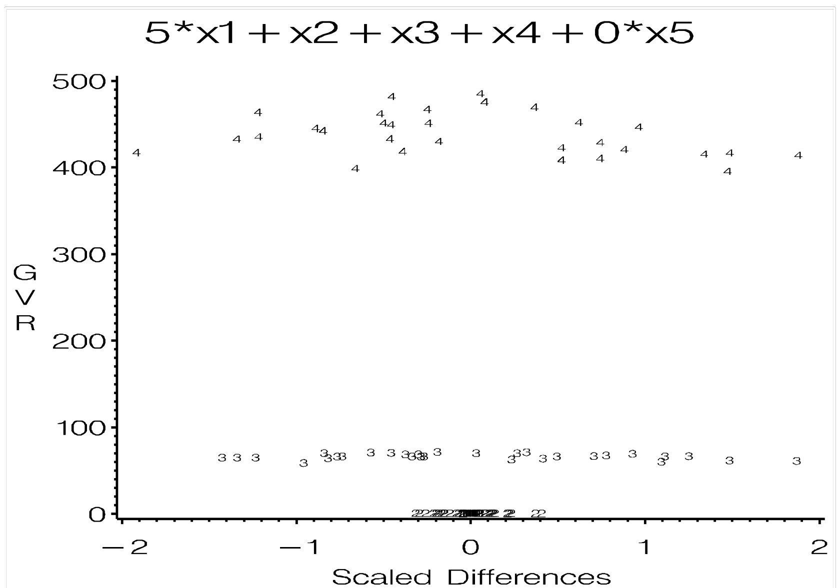


Figure 4C

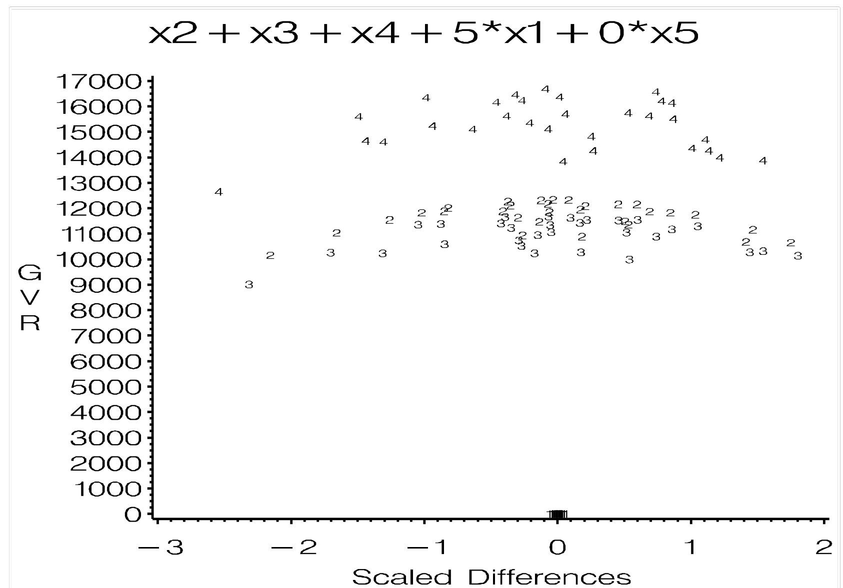
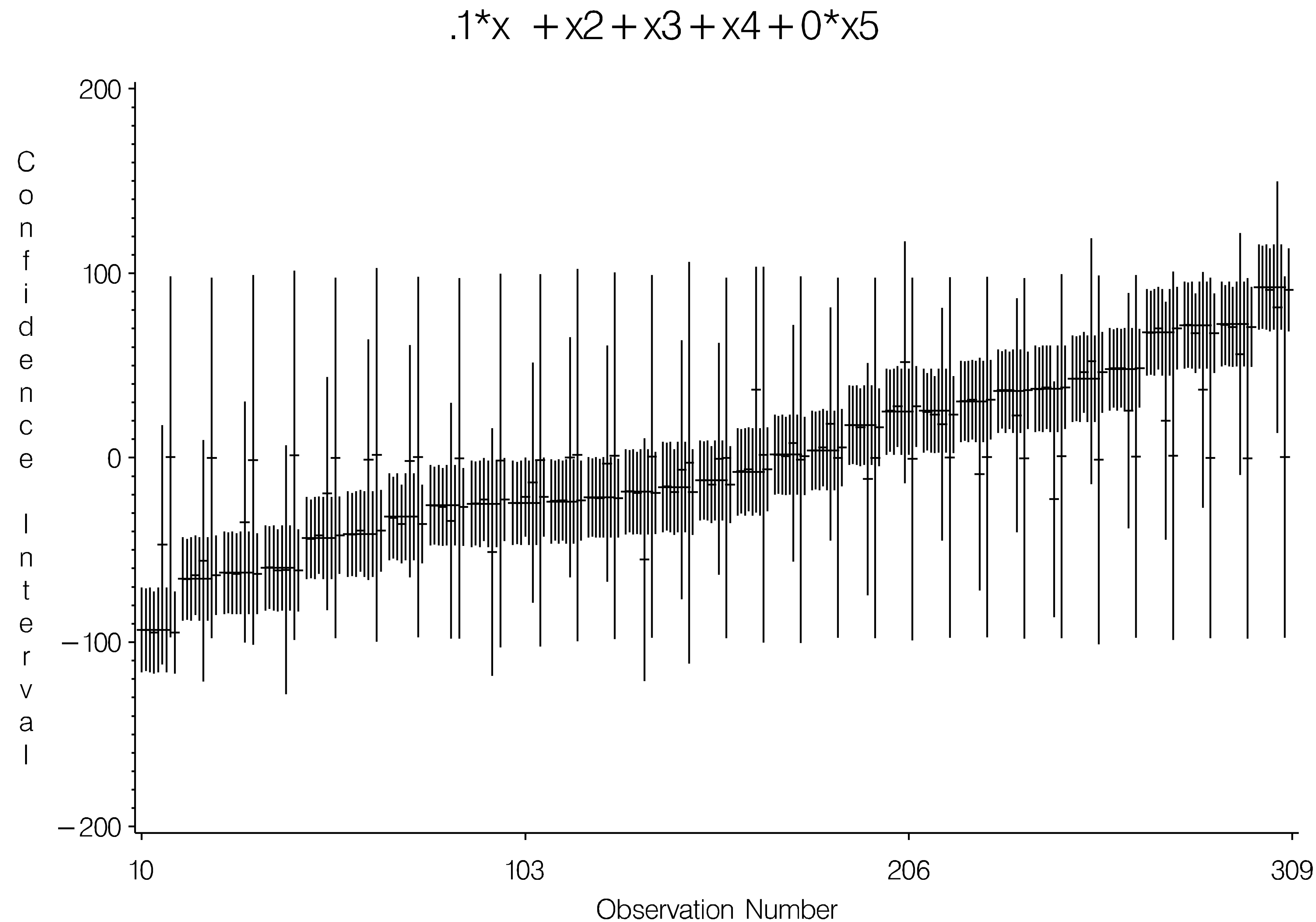


Figure 4D


.1*x +x2+x3+x4+0*x5
Confidence Interval
200
100
0
−100
−200
10
103
206
309
Observation Number
Figure 5

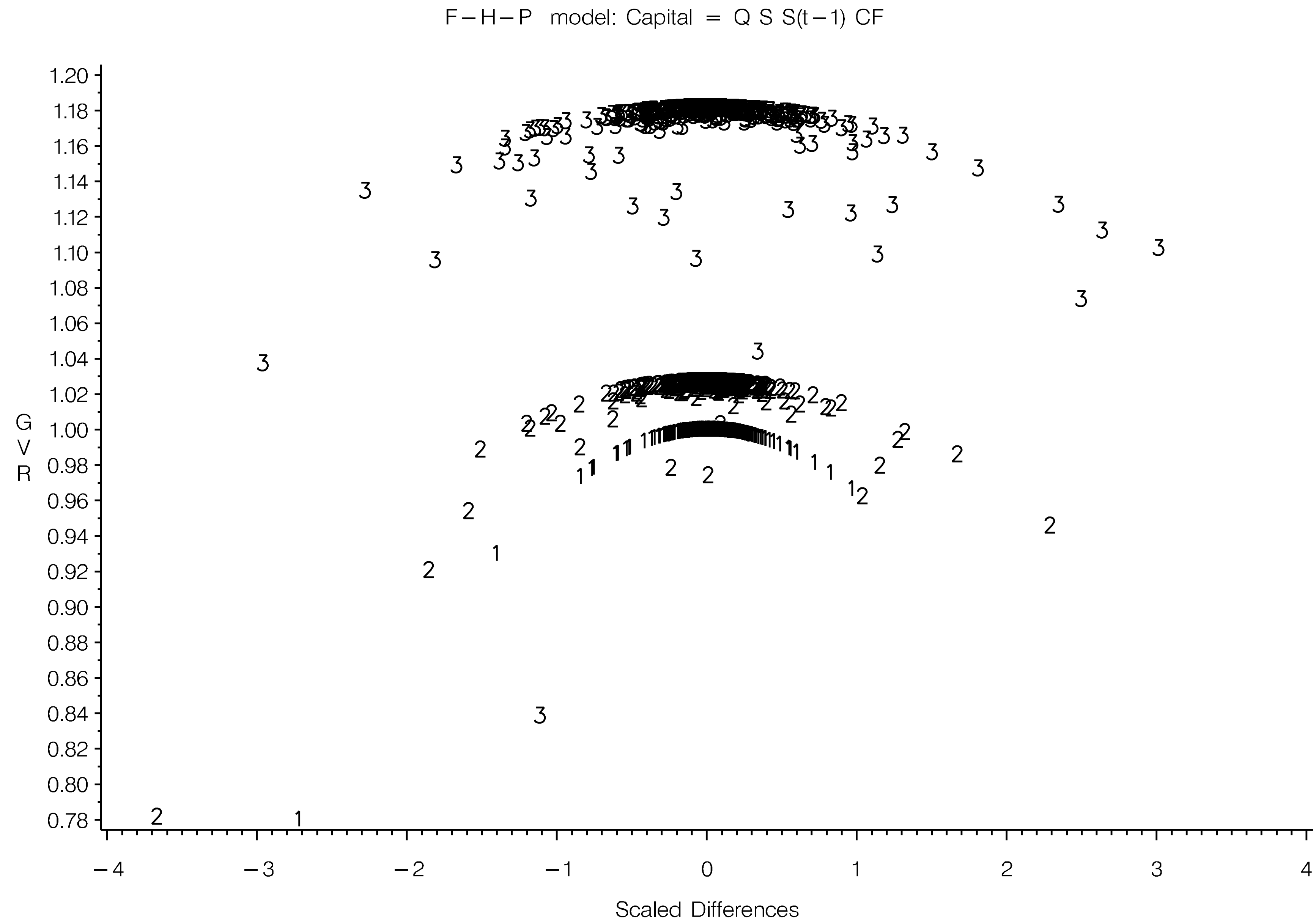

F − H − P model: Capital = Q S S(t − 1) CF
G V R
1.20
1.18
1.16
1.14
1.12
1.10
1.08
1.06
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
− 4
− 3
− 2
− 1
0
1
2
3
4
Scaled Differences

Figure 6

F − H − P model: Capital = Q S S(t − 1) CF

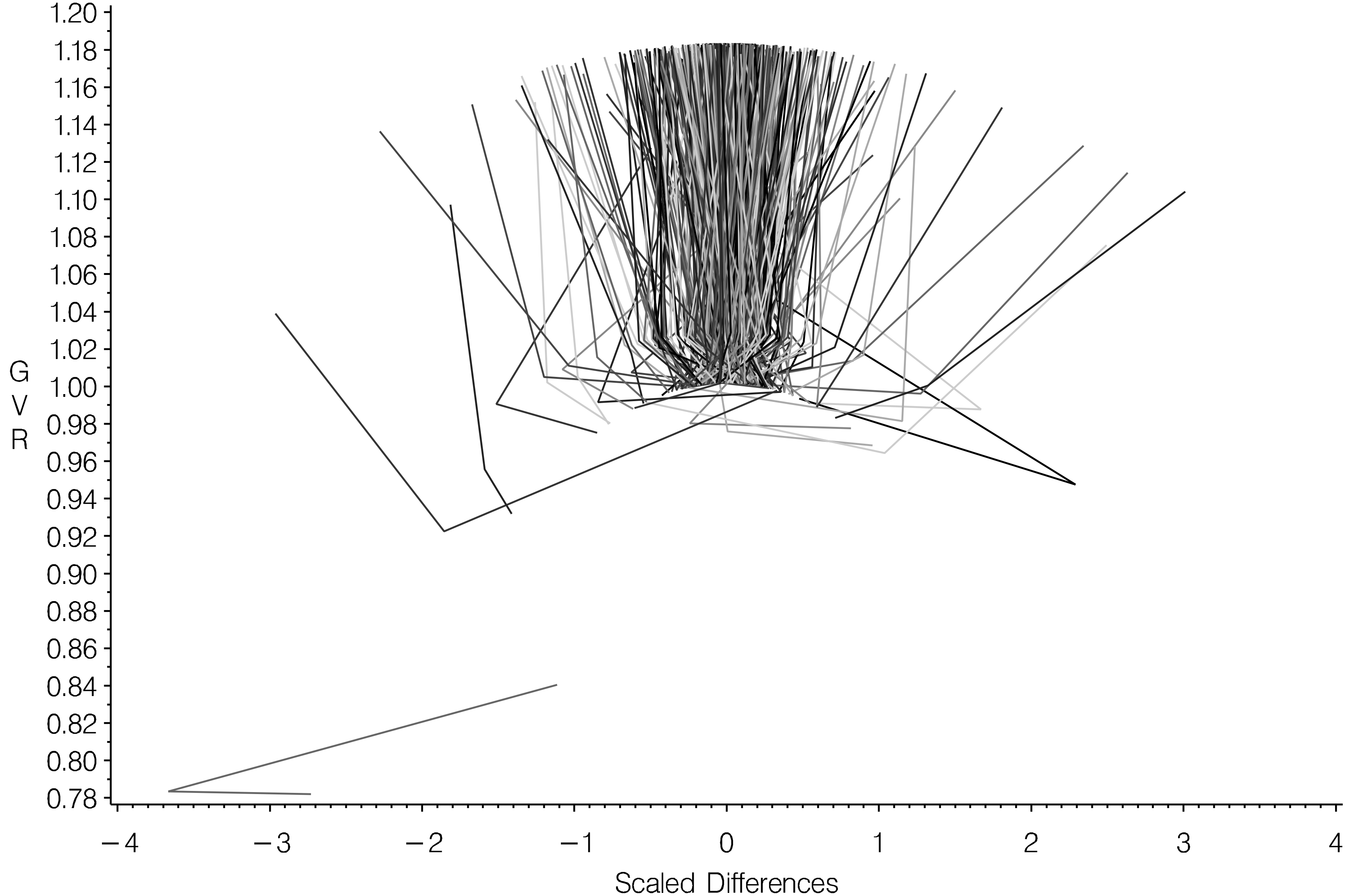


Figure 7

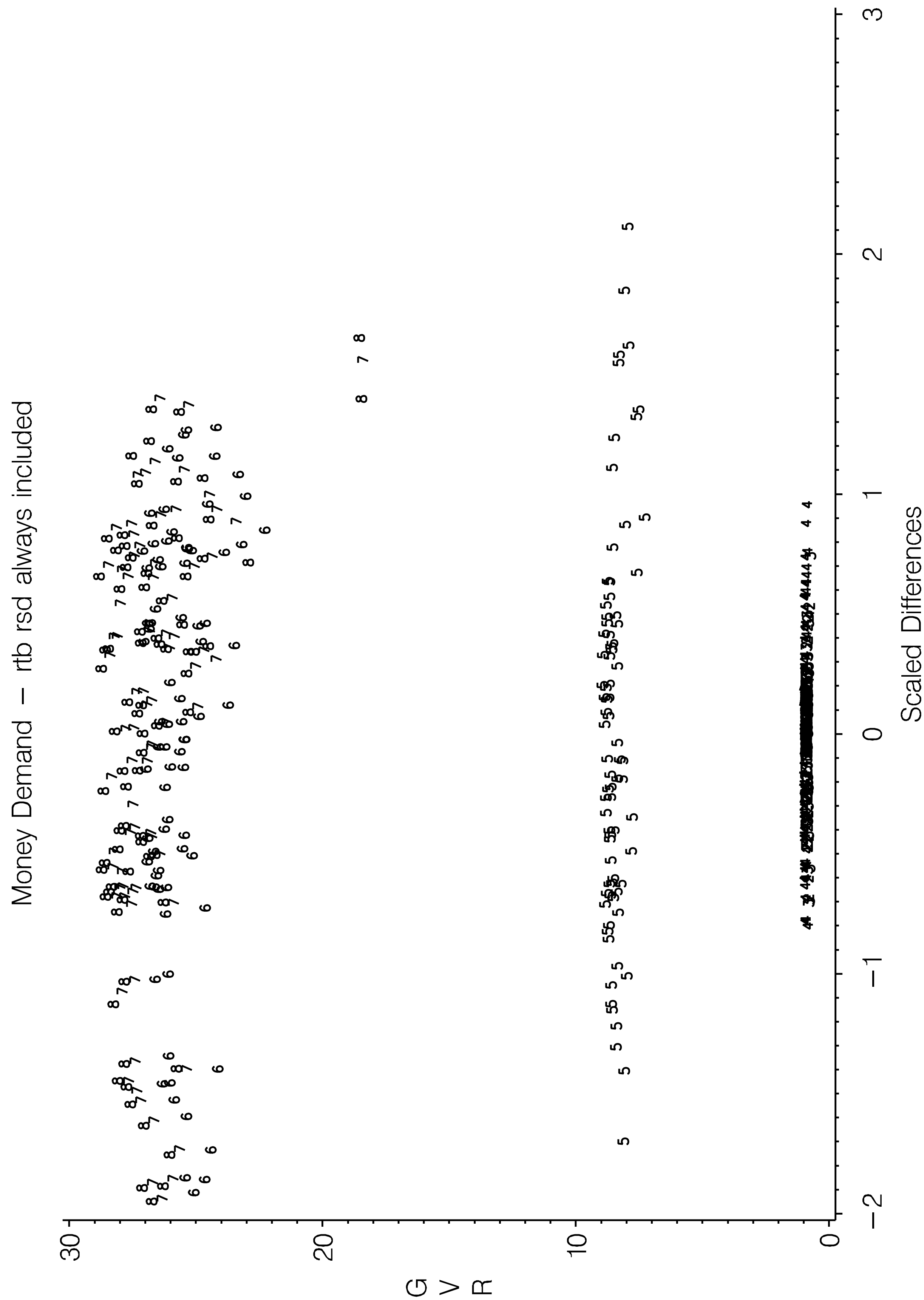

Money Demand – rtb rsd always included
G V R
Scaled Differences

Figure 8