LINEAR MODELS WITH DOMINANT VARIABLES

Peter S. Lufkin, SysteMetrics, Inc.

1.0 INTRODUCTION

Most researchers experienced in applied regression analysis have encountered problems with dominant variables. This is the paradoxical situation where one (or more) independent variables fits the model too well, or as Kennedy (1980) describes it, "the (dominant) accounts for so much of the variation in a dependent variable that the influence of other variables cannot be estimated." Typically, the dominant has an obvious and very strong relationship with the dependent variable, so that the problem lies with estimating the equation rather than with theoretical difficulties.

The statistical symptoms of the dominant variable problem are usually clear: a high multiple correlation coefficient ($R^2$); a very efficient estimate of one parameter - the dominant - in contrast to estimators with large standard errors for the remaining explanatory variables; and parameter estimates often with inappropriate magnitudes and incorrect signs. Such results are virtually useless when the main purpose of the model is to provide specific parameter estimates.

Traditional approaches to the problem include simply doing nothing, omitting the dominant, or redefining the dependent (Rao and Miller, 1971). These are relatively drastic solutions based on avoiding the dominant variable or ignoring its influence. New estimation strategies may be forthcoming as it is understood that collinearity between the dominant variable and the other explanatory variables is the critical reason for the effect of the dominant. In this paper we demonstrate that such collinearity is largely responsible for the dominant variable problem, and discuss one alternative estimation method based on orthogonalising the dominant.

Attempting to estimate equations with dominant variables can lead to spurious conclusions regarding variable selection. Explanatory variables actually relevant to an equation may seem superfluous because of switched signs and high standard errors. This is often the case in models where outputs are regressed on raw materials. For example, Rao and Miller discuss the following equation of textile production; where $(Q)$ is the value of output, $(L)$ is the wage bill, $(R)$ represents fixed capital, and $(M)$ is the value of raw materials.

$$\log Q = -0.408 -0.059 \log K -0.002 \log L +1.094 \log M$$

$$(1.59) \quad (1.37) \quad (.039) \quad (16.8)$$

$$R^2 = .997$$

Although the model succeeds in explaining almost all of the variation in the dependent, only the raw material parameter has a significant $t$ value (in parentheses). The labor and capital parameters have the wrong signs and their low $t$ scores indicate their explanatory power is questionable. The analyst ignorant of the obvious relevance of the labor and capital variables might even drop them from the equation.

We can take a more detailed view of the problem with a hospital output model explaining variations in the total days of care provided by hospitals. The equation is specified so that total patient days ($DAYS$) is a function of hospital location ($URBAN$), complexity of the caseload ($MIX$), rates of surgery ($SURG$), seasonal trends ($SUM$), and a size variable - the number of beds ($BEDS$). A dummy variable ($REV$) has been included to measure the impact of a physician review program intended to reduce unnecessarily long hospital stays. The equation is estimated for a large sample of short-stay, non-teaching hospitals. The results (Table 1, model 1) show the same confusion of signs and low significance evident in the textiles example. We expected, for instance, that urban location ($URBAN$) would have a positive impact on total days ($DAYS$), and knew from past studies that case mix ($MIX$) and surgical rates ($SURG$) have a considerable effect. Only $URBAN$ and $BEDS$ have significant parameters, despite the theoretical importance of the other explanatory variables.

2.0 COLLINEARITY WITH THE DOMINANT VARIABLE

The sources of the dominant variable problem are evident when we examine collinearities in the hospital model. Referring to Appendix 1, we find that $BEDS$ is the dominant variable: the covariance between $DAYS$ and $BEDS$ is over ten times higher than for any other variables; and the correlation of $BEDS$ with $DAYS$ is .93. $BEDS$ is uncorrelated with the program variable $REV$.

The impact of collinearity with the dominant variable ($BEDS$) is apparent when the dominant is dropped from the equation (Table 1, model 2); Parameters change for each variable correlated with $BEDS$ - the coefficients for $URBAN$ and $MIX$ both have the correct sign, and precision for $URBAN$ and $SURG$ coefficients have increased; parameters remain unchanged for variables uncorrelated with the dominant - the REV coefficient changes very little. The general consequences of omitting a relevant variable are noted in Appendix 2 and discussed in most econometric texts. See for example, Kmenta, (1971) and Maddala, (1977).

The impact of collinearity between individual variables is emphasized when a dominant variable is present. Even weak correlations with the dominant led to dramatic changes in the estimated parameters: a correlation of .25 with $BEDS$ caused the $URBAN$ parameter to increase by a magnitude of 4, and change signs when $BEDS$ was dropped from the equation; the casemix parameter ($MIX$) flipped signs, also with a correlation with $BEDS$ of .25; the $SURG$ variable has the strongest
correlation with BEDS (.55) and its parameter increased from 2.5 to 61.6.

These results underscore the role collinear-{
arity among the explanatory variables plays in the
dominant variable problem. Without such collinear-
arity, dominant variables like BEDS would have
no effect on the other parameters, irregardless
of their strong covariation with the dependent.
This is a fairly simple fact which has gone un-
noticed in most discussions of dominant variables.

3.0 COLLINEARITY DIAGNOSTICS IN PROC REG

Recognizing that collinearity is at the
heart of the matter, we can attempt to more
rigorously assess its effects using the TOL, VIF,
and COLLIN diagnostics available in the SAS pro-
cedure REG. These were estimated for the hospital
model and are reported in Appendix 3.

The Variance Inflation Factor (VIF) and the

closely related Tolerance (TOL) measures are
based on auxiliary regressions of each independ-
ent variable of the remaining independents. An
R2 indicative of a strong linear dependency, say
.9, would be reflected in a VIF of 10 (calculated
as 1/1-R2). TOL is simply 1-R2, so that for the
same example it would equal .1. Referring again
to Appendix 3, neither of these statistics indi-
cates serious collinearities.

COLLIN invokes a useful diagnostic procedure
described by Belsley, Kuh and Welsh (1980). This
method provides two types of information: The
condition index indicates the presence of a near
dependency among the columns of the data matrix;
and the variance-decomposition proportions Iden-
tify the variables involved. Belsley et al.
suggest, as a rough rule of thumb, that a condi-
tion index greater than 30 accompanied by vari-
ance proportions of at least .5 for two or more
coefficients, would indicate a collinear rela-
tionship with the potential to degrade the
estimated parameters (p.112-113). The highest
condition index for the hospital model is 23,
which suggests a moderate linear dependency.
Reading across the variance proportion table we
find that the intercept and MIX are highly rela-
ted, with over 90% of the variation in their
estimated parameters associated with the alarming
condition index. This is due to the relatively
small variation of MIX around a mean of 1. We do
not find, however, any indication of degrading
collinearities involving the dominant BEDS, de-
spite the evidence discussed in the previous
section.

This discussion makes it clear that the
collinearity diagnostics supplied by PROC REG
have limited usefulness for identifying harmful
near dependencies in dominant variable equations.
There are a number of reasons for this. First,
the VIF and TOL do not reflect the extra sensi-
tivity of the parameters to collinearity with a
dominant variable. The linear dependencies in-
dicated by these measures for BEDS were mild, but
nevertheless capable of severely affecting parame-
ter estimates. Second, a more general short-
coming of VIF and TOL is that even when serious
collinearity is evident, they do not help to
identify the specific variables involved. And

third, the variance-decomposition method may not
work when a dominant variable is involved. The
procedure is sensitive to "essential" scaling im-
balance, where the variation introduced by one
variable is much greater than the variation con-
tributed by the other independents (Noltemeyer, Kuh,
Welsh, p.156). In this situation potential near
dependencies involving the dominant may be diffi-
cult to detect using the variance-decomposition
methodology.

More useful information is provided by the
parameter estimates of auxiliary regression of
the dominant against the other independent vari-
ables. Once the covariance matrix has indicated
a probable dominant variable, this regression can
identify the variables most strongly affected.

BEDS=.45.8 +21.6 URBAN -20.1 MIX +2.50 SURG
(.38) (6.3) (1.5) (20)
-1.53 Sum -.703 (.43) (.11)
r2 = .33

The auxiliary regression for BEDS is reported
above. As the correlation matrix suggested, SURG
is most strongly related to BEDS, with UNBAN also
significant. Coefficients of this equation not
only indicate the strength of collinearity, but
also show its direction and provide an estimate
of the bias to the remaining parameters when BEDS
is omitted.

4.0 ESTIMATING EQUATIONS WITH DOMINANT VARIABLES

The dominant variable is only a problem for
certain models, depending on their intended use.
If the analyst is primarily concerned with pre-
dictions instead of parameter estimates, a domi-

nant variable model can provide accurate fore-
casts. But if specific parameter estimates are
important, as in descriptive models or evaluation
models, then the influence of the dominant should
be recognized and possibly ameliorated.

There are at least four strategies for
estimating dominant variable equations: ignore
the issue; omit the dominant variable; redefine
the dependent in a form less susceptible to the
influence of the dominant variable; or substitute
a proxy for the dominant. Again, selection of
the appropriate approach depends heavily on the
purpose of the model, e.g. prediction, descrip-
tion, or evaluation.

IGNORE THE DOMINANT: This is a reasonable ap-
proach when the model is used for forecasting.
The dominant variable may degrade the estimates
of specific parameters, but the model can still
be capable of accurate prediction. This is true
only if any collinear relationships remain stable
into the prediction period.

Ignoring the influence of the dominant may
also be appropriate when variables of particular
interest are independent of the dominant variable.
As we noted earlier, the coefficients for vari-
ables unrelated to the dominant are unaffected by
including or omitting the dominant from the equa-
tion. For instance, the program variable REV in
the hospital model is nearly independent of the
dominant BEDS. If we were interested solely in
the parameter of REV, then the problems with
coefficients for the other variables would be irrelevant. Often in evaluation models explanatory variables serve only as statistical controls. If the intervention variable is orthogonal to the dependent then there is no problem.

However, it is rare that any two explanatory variables are completely independent, and even very slight collinearity can cause considerable effects to estimators and their variances when a dominant is involved. Thus, it may not be appropriate to "ignore the dominant" for any equation where variables of interest are even slightly collinear with the dominant.

Omit the Dominant Variable: In some situations omitting the dominant variable is the simplest way to improve estimates of the remaining coefficients. For example, an economist's production function often does not include a variable for raw materials. Consider the re-estimate of the textile model discussed earlier:

$$\log Q = -206 + .413 \log K + .708 \log L \quad R^2 = .93$$

After the dominant (raw materials) has been dropped, the parameters for labor (L) and capital (K) have the proper signs and magnitudes (Rao and Miller, p. 42). The R² indicates that relatively little explanatory power has been lost by the deletion; and overall the model is much more useful than its predecessor containing the dominant, raw materials.

However, there are reservations to this equation. First, the remaining parameters are biased to the extent that labor and capital are collinear with materials. But perhaps more important, the equation leaves out a critical variable in its description of the production process. This is a theoretical problem: One might legitimately ask how textiles can be made without raw materials.

Aside from theoretical difficulties, excluding the dominant can often reduce the explanatory power of a model to the extent that the estimates of the remaining parameters are suspect. This may not be the case for models with other strong explanatory variables like the textile equation, but it is a definite problem for equations like the hospital model.

As previously discussed in Table 1, model 2, the mean squared error (MSE) of the hospital equation increased almost sixfold when BEDS was dropped. The parameters of the remaining variables, like the textile model, now have appropriate signs and magnitudes, but are also biased and inconsistent. The standard error of REV has almost doubled. Examining a scatter plot of the standardized residuals (Appendix 4) makes it clear that without BEDS the error term is no longer distributed around mean 0 and that the equation is misspecified.

Omitting the dominant variable from the equation is a strategy that should be followed with considerable caution. It is inappropriate for forecast models because of the increased MSE and potential bias which comes from excluding a relevant variable. It is not very helpful for evaluation models because significant tests for intervention variables (like REV, for example) are underestimated, even if the intervention variable is independent of the dominant. But for descriptive models omitting the dominant does provide at least rough estimates of parameters that were otherwise masked by collinearity with the dominant. In models where the dominant adds little new information, such as the textile example, estimates of the "omitted dominant" equation may be acceptable.

Redefine the Dependent: The influence of the dominant variable can be diminished by using a rate, rather than a total measure, as the dependent variable. This is an intuitively appealing way to include the dominant in the equation, but reduce its almost lock-step covariance with the dependent. For example, the hospital model was re-estimated using average days of care instead of total days of care. The results are reported in Table 1, model 3.

In this model the impact of BEDS is indeed reduced, but so is the overall power of the model—the rate equation explains less than 10% of the variation of the average length of stay. The standardized residuals are not centered on zero but are distributed in a linear pattern indicating that this equation is seriously misspecified. It is clear that variables important for explaining variation in the dependent are missing.

As an alternative to the original equation, the rate model can be extremely useful, given two rather obvious conditions. First, the model must still be meaningful; does it make a difference to the analyst that he or she is forecasting (describing, evaluating) averages instead of totals? Second, it is important that the independent variables have sufficient explanatory power relative to the new dependent. If the available independent do provide an adequately specified model then the rate strategy is the optimal solution to the dominant variable problem: the corrupting influence of the dominant has been diminished without dropping the dominant and biasing the remaining coefficients. Thus, redefining the dependent should be the first alternative when the equation is adequately specified and relevant to the analyst's needs. This is especially true for descriptive and evaluative models because of the importance of accurate parameter estimates.

Unfortunately, the rate model may not always be sufficiently specified. This is true for the hospital model, where the low explained variation and the skewed distribution of the residuals clearly indicate a mispecified equation. (See Appendix 4).

Proxy for the Dominant: A final approach for estimating a dominant variable equation involves selecting a proxy for the dominant. This solution is appropriate for problems like the hospital equation, where the other alternatives we have discussed are impractical. Consider the hospital model: the original equation provides unreasonable parameter estimates for variables collinear with the dominant; omitting the dominant eliminates its influence on the remaining parameters, but leaves them biased, inconsistent and i-
effcient: redefining the dependent as a rate takes away most of the model's explanatory power leaving the equation obviously misspecified. Replacing BEDS with a suitable proxy could at least improve the efficiency of the parameters previously estimated by dropping the dominant.

Ideally, a proxy for the dominant can be found that minimizes collinearity with the other independents, yet is still strongly related to the dependent. The trick, of course, is to find such a variable. A logical choice for the proxy is an alternative form of the dominant variable, purged of its collinearity—and thus its dominance—with the other independent variables. This is done by regressing the dominant on the other independents and retrieving the residuals. These residuals define the proxy that can then be substituted into the original equation.

An equation using this type of proxy has four interesting characteristics, the last one being the most important. A more detailed discussion of these characteristics can be found in Appendix 2.

1. The parameter estimates are not different from those in the "omitted dominant" equation. This reminds us that, in general, adding or deleting an orthogonal variable (such as the proxy) will not affect the parameter estimates of the other independents.

2. The parameter estimates have the same biases and inconsistencies as the "omitted dominant" equation. The strength and direction of the bias can be explored by regressing the dominant on the other independents (the "auxiliary" regression described in an earlier section).

3. The coefficient for the proxy is identical to the coefficient for the dominant variable in the original equation.

4. The residual variance is less than for the equation that omitted the dominant. We have the same biased parameters, but they are now more precise. Other things being equal, the standard errors of parameters will be smaller, and their t-values larger.

The hospital model was re-estimated using a proxy for BEDS, and the results are reported in Table 1, model 4. As was expected, the estimated parameters have not changed from the "omitted dominant" equation. BEDS has the same coefficient as in the original equation (Table 1, model 1). The greatest change brought about by the proxy was the improvement in the precision of the estimates. The standard errors for all of the parameter estimates are smaller than for any of the other hospital equations. This result is especially important for models where the significance of specific estimators, such as intervention effects like REV, are critical. When redefining the dependent as a rate is not a viable alternative, the proxy model will at least provide more efficient estimates than the "omitted dominant" strategy.

5.0 SUMMARY

In this paper we have taken an applied approach to estimating dominant variable equations. Using a simple economic production model and a hospital output model it was demonstrated that collinearity among the independent variables is necessary for a dominant variable problem to exist. If there is no such collinearity then the dominant is not "dominant" at all, and is just a very significant explanatory variable.

It was also shown that some of the collinearity diagnostics in Proc Reg fail to indicate the potential harm of collinearity with the dominant. This is because normally moderate collinearities can, nevertheless, seriously impact parameter estimates when a dominant variable is involved. The effects of collinearity seem emphasized in the dominant variable situation.

The variance-decomposition methodology was particularly disappointing in diagnosing collinearity with the dependent. We suggest that correlation and covariance matrices are less sophisticated, but more useful indicators of the dominant variable and its correlates. An auxiliary regression of the dominant on the other regressors will estimate the strength and direction of these collinearities.

Four approaches to estimating dominant variable equations were discussed. Of the three traditional approaches—ignoring the problem, omitting the dependent, and redefining the dependent—the latter was considered the optimal choice. By redefining the dependent variable as a rate instead of a total the severe covariance of the dominant with the dependent is diminished, without dropping the dominant and biasing the remaining parameters. When the rate model was not feasible, as in the case of the hospital model, a proxy for the dominant can be used to improve the efficiency of the estimators provided by the "omitted dominant" equation.

REFERENCES


Consider the following equation, where \(X_2\) is the dominant variable.

\[
\gamma = X_1 \beta_1 + X_2 \beta_2 + \epsilon
\]

If we drop \(X_2\) and estimate

\[
\hat{\gamma} = X_1 \hat{\beta}_1 + \epsilon
\]

Then the estimate of \(\beta_1\) will be biased, as in the general case of omitting a relevant variable. This is shown with the familiar proof.

\[
\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'\hat{\gamma} = (X_1'X_1)^{-1}(X_1'X_1)\beta_1 + (X_1'X_1)^{-1}X_1'\epsilon
\]

or \(E(\hat{\beta}_1) = \beta_1 + \beta_2 P\), where \(P = (X_1'X_1)^{-1}X_1\) is the matrix of auxiliary regression coefficients of \(X_2\) on \(X_1\). These parameters are useful as a measure of collinearity between \(X_2\) and \(X_1\), and they provide estimates of the direction and magnitude of the bias of \(\hat{\beta}_1\) if \(X_2\) is dropped from the equation.

If \(X_1\) and \(X_2\) are independent, \(P=0\), then \(\hat{\beta}_1\) will be unbiased. However, the estimate of the residual variance \(S_2^2\) will still include the variance of the omitted variable, \(X_2\). Thus \(S_2^2\) will be greater than the true residual \(\sigma^2\) from (1) and will bias the estimated variance of the parameters upward, so that tests of significance and confidence intervals for \(\beta_1\) will be overly conservative.

The effects of the dominant variable on estimates of \(\beta_1\) can be eliminated by "purging" \(X_2\) of its collinearity with \(X_1\). The new values of \(X_2^*\) would replace \(X_2\) in the original equation

\[
\gamma = X_1 \beta_1 + X_2^* \beta_2^* + \epsilon
\]

or \(E(\hat{\beta}_1) = \beta_1 + \beta_2 P\), where \(P = (X_1'X_1)^{-1}X_1\) is the matrix of auxiliary regression coefficients of \(X_2\) on \(X_1\). These parameters are useful as a measure of collinearity between \(X_2\) and \(X_1\), and they provide estimates of the direction and magnitude of the bias of \(\hat{\beta}_1\) if \(X_2\) is dropped from the equation.

If \(X_1\) and \(X_2\) are independent, \(P=0\), then \(\hat{\beta}_1\) will be unbiased. However, the estimate of the residual variance \(S_2^2\) will still include the variance of the omitted variable, \(X_2\). Thus \(S_2^2\) will be greater than the true residual \(\sigma^2\) from (1) and will bias the estimated variance of the parameters upward, so that tests of significance and confidence intervals for \(\beta_1\) will be overly conservative.

The effects of the dominant variable on estimates of \(\beta_1\) can be eliminated by "purging" \(X_2\) of its collinearity with \(X_1\). The new values of \(X_2^*\) would replace \(X_2\) in the original equation

\[
\gamma = X_1 \beta_1 + X_2^* \beta_2^* + \epsilon
\]

or \(E(\hat{\beta}_1) = \beta_1 + \beta_2 P\), where \(P = (X_1'X_1)^{-1}X_1\) is the matrix of auxiliary regression coefficients of \(X_2\) on \(X_1\). These parameters are useful as a measure of collinearity between \(X_2\) and \(X_1\), and they provide estimates of the direction and magnitude of the bias of \(\hat{\beta}_1\) if \(X_2\) is dropped from the equation.

If \(X_1\) and \(X_2\) are independent, \(P=0\), then \(\hat{\beta}_1\) will be unbiased. However, the estimate of the residual variance \(S_2^2\) will still include the variance of the omitted variable, \(X_2\). Thus \(S_2^2\) will be greater than the true residual \(\sigma^2\) from (1) and will bias the estimated variance of the parameters upward, so that tests of significance and confidence intervals for \(\beta_1\) will be overly conservative.

The effects of the dominant variable on estimates of \(\beta_1\) can be eliminated by "purging" \(X_2\) of its collinearity with \(X_1\). The new values of \(X_2^*\) would replace \(X_2\) in the original equation

\[
\gamma = X_1 \beta_1 + X_2^* \beta_2^* + \epsilon
\]

or \(E(\hat{\beta}_1) = \beta_1 + \beta_2 P\), where \(P = (X_1'X_1)^{-1}X_1\) is the matrix of auxiliary regression coefficients of \(X_2\) on \(X_1\). These parameters are useful as a measure of collinearity between \(X_2\) and \(X_1\), and they provide estimates of the direction and magnitude of the bias of \(\hat{\beta}_1\) if \(X_2\) is dropped from the equation.

If \(X_1\) and \(X_2\) are independent, \(P=0\), then \(\hat{\beta}_1\) will be unbiased. However, the estimate of the residual variance \(S_2^2\) will still include the variance of the omitted variable, \(X_2\). Thus \(S_2^2\) will be greater than the true residual \(\sigma^2\) from (1) and will bias the estimated variance of the parameters upward, so that tests of significance and confidence intervals for \(\beta_1\) will be overly conservative.
the partitioned $X'X$ matrix:

$$
\begin{bmatrix}
X_{11} & X_{12} \\
X_{21} & X_{22}
\end{bmatrix}
$$

solving for $\beta_2$ we find

$$
\hat{\beta}_2 = (X_{22}X_{22})^{-1}X_{22}Y
$$

and since $Q_2$ is idempotent,

$$
\bar{\beta}_2 = (X_{22}X_{22})^{-1}X_{22}Y = (X_{22}X_{22})^{-1}X_{22}Q_2Y
$$

which is identical to the expression for $\hat{\beta}_2$, where $X_{22} = Q_2 X_2$.

Fourth, the estimate of the residual variance for the proxy model (3) will be less than for the omitted variable equation (2), because $X_2$ is no longer left in the error term. Thus, while there is no change in the parameter estimates between (2) and (3), the variances of the estimates are lower for the proxy model (3).

**APPENDIX 3**

**COLLINEARITY DIAGNOSTICS**

**APPENDIX 4**

**SCATTER PLOTS OF RESIDUALS BY THE DEPENDENT FOR ALTERNATIVE ESTIMATIONS OF THE HOSPITAL MODEL**

**MODEL 1:** Full Model, includes the Dominant

**MODEL 2:** Omit the Dominant

**MODEL 3:** Redefine the Dependent as a Rate

**NOTE:** The explained variation and the scatter plots for the proxy equation (model 4) and the full model equation (model 1) are identical.