Multicollinearity

Violation of the rank condition

$Y = X\beta + u$

$\text{Rank}(X) = K < n.$

Perfect multicollinearity

$\text{Rank}(X) = r < K \Rightarrow |X'X| = 0$ and $(X'X)^{-1}$ is not defined. Thus, the OLS estimator b is not defined.

Example 1: Dummy variable regression

Wage regression with Sex dummies. If M and F dummies are defined and the wage regression is run with M, F, and an intercept + other relevant X variables; we have the case of a perfect multicollinearity. This is because $M + F = 1$ which is the intercept. Consequently, the X matrix is not of full column rank and $X'X$ becomes singular.

- Need to drop either the intercept or one of the Sex dummies.

Ex 2: Fixed coefficient type production function.

$\frac{X_2}{X_3} = a$ (constant)

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \quad \text{is the regression}$$
$$= \beta_1 + X_3(\beta_3 + \beta_2 \cdot a) + u$$
$$= \beta_1 + \beta X_3 + u \quad \text{where } \beta = (\beta_3 + \beta_2 \cdot a)$$

$$\beta = \beta_3 + \beta_2 \cdot a$$

- Thus, only β which is a linear combination of β_2 and β_3 can be estimated.
- Alternately, either X_2 or X_3 can be used as a regressor but not both.
- Can β_3 and β_2 be estimated if a is known? No. If X_3 is used as a regressor, you get an estimate of $\beta = \beta_3 + \beta_2 a$ which cannot be solved for β_3 and β_2.

** A Formal proof with K regressors

Since the rank of X is $r < K$, we can partition X as

$$\underset{n\times K}{X} = [\underset{n\times r}{X_r} \quad \underset{n\times s}{X_s}] \qquad r+s=K$$

where each of the X variable in X_s can be expressed as a linear combination of X_r.

That is,

$$\left.\begin{aligned} X_s(1) &= X_r W_1 \\ &\vdots \\ X_s(s) &= X_r W_s \end{aligned}\right\} \quad \text{$X_s(1)$ denotes the 1st variable in X_s, and so on.}$$

$$\Rightarrow X_s = (\underset{n\times 1}{X_s(1)} \;\; -- \;\; \underset{n\times 1}{X_s(s)}) = X_r[W_1 \cdots W_s]$$

$$= X_r \underset{r\times s}{W}$$

Thus, $$X = [X_r \quad X_s] = [X_r \quad X_r W] = X_r[I_r \quad W]$$
$$= X_r \underset{r\times K}{\underset{\sim}{Z}} \quad (\text{say})$$

Now $Y = X\beta + u = X_r Z \beta + u$

$= X_r \beta_r + u$ where $\beta_r = Z \cdot \beta$ which is a linear combination of β.

OLS

$b_r = (X_r' X_r)^{-1} X_r' Y$ is <u>BLUE</u>

- It uses only r X-variables — the ones which are <u>not</u> linearly dependent on others.

Thus, although information on X is available, only X_r of them are used.

- Cannot estimate β. Can you solve β from $\underbrace{\beta_r}_{}= Z \cdot \beta$? [Z is a $r \times K$ matrix].

estimated from OLS (so it is known).

Assume that Z is also known.

Is $\beta = Z^{-1} \beta_r$??

- Any use of estimating β_r ?

If you want to forecast Y.

Near Exact Multicollinearity:

$|X'X| \neq 0$ but very close to zero,

$\Rightarrow \sum_{i=1}^{K} X_i \lambda_i \approx 0$ where λ_i are not all zero.

i.e., $X_1 \lambda_1 + X_2 \lambda_2 + \cdots + X_K \lambda_K \approx 0$

$$\Rightarrow \quad X_K = X_1 (-\lambda_1/\lambda_K) + X_2 \left(-\frac{\lambda_2}{\lambda_K}\right) + \cdots + X_{K-1}\left(-\frac{\lambda_{K-1}}{\lambda_K}\right) + V_K \quad (*)$$

assuming $\lambda_K \neq 0$. In the above equation the V_K is the difference between the l.h.s. and the r.h.s. If $\sum_i X_i \lambda_i = 0$, $V_K = 0$.

Now think of $(*)$ as a regression equation.

- The better the fit of this auxiliary regression (i.e., higher the R^2), more severe is the multicollinearity
- Can use any of the X variable on the l.h.s. $\Rightarrow$ (K-1) diffrent regression (one being the intercept) can be run and (K-1) different R^2's can be obtained.

<u>Symptom</u>: If R^2 for the regression $(Y = X\beta + u)$ is less than at least one of R^2_K (from the auxiliary regression) — there is severe MC (Klein).

Note that since $|X'X| \neq 0$ ——— the OLS is defined.

Effects of MC:

Solution of b — the OLS estimator — exists and b is BLUE. But the estimated variance of b increases with the degree of MC ($\Rightarrow$ low t values).

To prove this, partition X as

$$X = [\underset{\downarrow\, i\text{th } X}{X_i} \quad \underset{\searrow \text{ all other } X \text{ (except } i)}{X_{-i}}]$$

$$\Rightarrow \quad X'X = \begin{bmatrix} X_i' X_i & X_i' X_{-i} \\ X_{-i}' X_i & X_{-i}' X_{-i} \end{bmatrix}$$

The leading term in $(X'X)^{-1}$ is

$$\left[X_i' X_i - X_i' X_{-i} (X_{-i}' X_{-i})^{-1} X_{-i}' X_i\right]^{-1}$$

$$= \left[X_i' M_{-i} X_i\right]^{-1} \quad \text{where } M_{-i} = I - X_{-i}(X_{-i}' X_{-i})^{-1} X_{-i}'$$

Thus, $V(b_i) = \sigma^2 / (X_i' M_{-i} X_i)$

Now $X_i' M_{-i} X_i$ is RSS_i in the regression X_i on X_{-i}.

[Note that $e'e = u'Mu = Y'MY$ is the RSS of the regression Y on X].

Since $R_i^2 = 1 - \dfrac{RSS_i}{TSS_i} \Rightarrow RSS_i = TSS_i(1 - R_i^2)$

$$V(b_i) = \frac{\sigma^2}{TSS_i(1 - R_i^2)}$$

as MC $\uparrow$, $R_i^2 \uparrow \Rightarrow V(b_i) \uparrow \Rightarrow t \downarrow$.

Since σ^2 is not Known, $V(b_i)$ is to be estimated from

$$\text{Est } V(b_i) = \frac{s^2}{TSS_i(1-R_i^2)}$$

Now $s^2 = \frac{e'e}{n-K}$ and $R^2 = 1 - \frac{e'e}{y'y}$

$$\Rightarrow s^2 = \frac{(1-R^2)\cdot y'y}{n-K}$$

Therefore, $$\text{Est. } V(b_i) = \frac{(1-R^2)}{(1-R_i^2)} \cdot \frac{\frac{1}{n} y'y}{\frac{1}{n} x_i' x_i} \frac{1}{n-K}$$

$$= \frac{(1-R^2)}{(1-R_i^2)} \cdot \frac{s_y^2}{s_{x_i}^2} \cdot \frac{1}{n-K}.$$

- If $R_i^2 \uparrow$, Est. $V(b_i) \uparrow$ $(n-k) \uparrow \cdots \downarrow$
- If $R^2 \uparrow$, $\cdots\cdots \downarrow$
- If $s_{x_i}^2 \uparrow$ $\cdots\cdots \downarrow$
- Low variance in X_i and high MC have same effect.
- Condition Index (Belsley, Kuh and Welsh (1980)) Regression Diagnostic

$$\eta_k = \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_{min}}}, \quad k = 1, \cdots K$$

where λ's are the eigen values of the $X'X$ matrix.

If the condition index exceeds 20 "danger level", Can easily exceed 100!!

This is automatic is SAS.

tion of β_1 and β_2 given by the dashed line is low. Consequently, if the researcher's interest centers on this linear combination, the multicollinearity need not be of concern. This might happen, for example, if the estimated equation is to be used for prediction purposes and the multicollinearity pattern is expected to prevail in the situations to be predicted.

REMEDIES

(2) INCORPORATE ADDITIONAL INFORMATION

There are several possibilities here, most of which should be considered even in the absence of multicollinearity.

(a) *Obtain more data* Because the multicollinearity problem is essentially a data problem, additional data that do not contain the multicollinearity feature could solve the problem. Even getting additional data with the same multicollinearity character would help, since the larger sample size would provide some additional information, helping to reduce variances.

(b) *Formalize relationships among regressors* If it is believed that the multicollinearity arises not from an unfortunate data set but from an actual approximate linear relationship among some of the regressors, this relationship could be formalized and the estimation could then proceed in the context of a simultaneous equation estimation problem.

(c) *Specify a relationship among some parameters* Economic theory may suggest that two parameters should be equal, that the sum of several elasticities should be unity, or, in general, that there exists a specific relationship among some of the parameters in the estimating equation. Incorporation of this information, via methods discussed in chapter 12, will reduce the variances of the estimates. As an example, consider specifying that the coefficients of a lag structure take the form of a Koyck distributed lag (i.e., they decline geometrically), as discussed in section 9.3.

(d) *Drop a variable* A popular means of avoiding the multicollinearity problem is by simply omitting one of the collinear variables. If the true coefficient of that variable in the equation being estimated is zero, this is a correct move. If the true coefficient of that variable is *not* zero, however, a specification error is created. As noted in section 6.2, omitting a relevant variable causes estimates of the parameters of the remaining variables to be biased (unless some of these remaining variables are uncorrelated with the omitted variable, in which case their parameter estimates remain unbiased). The real question here is whether, by dropping a variable, the econometrician can reduce the variance of the remaining estimates by enough to compensate for this bias introduced. This suggests the use of the MSE criterion in undertaking a decision to drop a variable. This approach should not be adopted cavalierly, since, as noted by Drèze (1983, p. 296), "setting a coefficient equal to zero because it is estimated with poor precision amounts to elevating ignorance to arrogance."

(e) *Incorporate estimates from other studies* If an extraneous estimate of the coefficient of one of the variables involved in the multicollinearity is avail-

able, it can be used, via
to alleviate the high vari
is done, however, care
relevant. For example,
alleviate time series m
long-run version of ma
for time series studies.

(f) *Form a princip*
grouped together to fo
of variables by itself.
variables included in t
pretation; otherwise th
in undertaking a study
researcher might find
ing activity are high
readily be interpreted
not confuse the mean
structing such a com
variables in question.

(g) *Shrink the OL*
zero vector, a resear
each individual para
lent to incorporating
close to the zero vec
estimator and the St

- Leamer (1983b, p
dence (high varia
of weak evidence
seem not to unde
sample size." He
amusing account
described in term
- The Ballentine p
3.2, in the gener
between the X a
blue or green a
mate β_x and β_z
mates are larger
- In addition to
associated with

able, it can be used, via the mixed estimation technique described in chapter 12, to alleviate the high variance problem occasioned by the multicollinearity. If this is done, however, care must be taken to ensure that the extraneous estimate is relevant. For example, estimates from cross-sectional studies are often used to alleviate time series multicollinearity, but cross-section estimates relate to the long-run version of many parameters, rather than the short-run version relevant for time series studies.

(f) *Form a principal component* The variables that are collinear could be grouped together to form a composite index capable of representing this group of variables by itself. Such a composite variable should be created only if the variables included in the composite have some useful combined economic interpretation; otherwise the empirical results will have little meaning. For example, in undertaking a study of the effect of marketing activity on consumer demand, a researcher might find that variables representing different dimensions of marketing activity are highly collinear; some combination of these variables could readily be interpreted as a "marketing variable" and its use in the model would not confuse the meaning of the empirical results. The most popular way of constructing such a composite index is to use the first principal component of the variables in question.

(g) *Shrink the OLS estimates* By shrinking the OLS estimates towards the zero vector, a researcher may be able to reduce the risk (the sum of the MSEs of each individual parameter estimate) of the estimates. Implicitly, this is equivalent to incorporating the *ad hoc* stochastic prior information that the true β is close to the zero vector. The two most popular means of doing this are the ridge estimator and the Stein estimator.

GENERAL NOTES

11.2 Consequences

- Leamer (1983b, pp. 300–3) stresses the fact that collinearity as a cause of weak evidence (high variances) is indistinguishable from inadequate data variability as a cause of weak evidence. Goldberger (1989, p. 141) speculates that the reason practitioners seem not to understand this is because there is no fancy polysyllabic name for "small sample size." He suggests the term "micronumerosity" be used, and provides a very amusing account of how all of the ills and manifestations of multicollinearity can be described in terms of micronumerosity.
- The Ballentine portrays the multicollinearity phenomenon succinctly. Consider figure 3.2, in the general notes to chapter 3. Multicollinearity is reflected by a large overlap between the X and Z circles. This could create a large red area at the expense of the blue or green areas. These blue and green areas reflect the information used to estimate β_x and β_z; since less information is used, the variances of these parameter estimates are larger.
- In addition to creating high variances of coefficient estimates, multicollinearity is associated with the undesirable problem that calculations based on the data matrix are