Chapter 9

Multicollinearity

9.1 The Nature of Multicollinearity

9.1.1 Extreme Collinearity

The standard OLS assumption that $(\, x_{i1}, x_{i2}, \ldots, x_{ik} \,)$ not be linearly related means that for any $(\, c_1, c_2, \ldots, c_k \,)$

$$x_{ik} \neq c_1 x_{i1} + c_2 x_{i2} + \cdots + c_{k-1} x_{i,k-1} \tag{9.1}$$

for some i. If the assumption is violated, then we can find $(\, c_1, c_2, \ldots, c_{k-1} \,)$ such that

$$x_{ik} = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_{k-1} x_{i,k-1} \tag{9.2}$$

for all i. Define

$$\mathbf{X}_1 = \begin{pmatrix} x_{12} & \cdots & x_{1k} \\ x_{22} & \cdots & x_{2k} \\ \vdots & & \vdots \\ x_{n2} & \cdots & x_{nk} \end{pmatrix}, \ \mathbf{x}_k = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kn} \end{pmatrix}, \text{ and } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{k-1} \end{pmatrix}.$$

Then extreme collinearity can be represented as

$$\mathbf{x}_k = \mathbf{X}_1 \mathbf{c}. \tag{9.3}$$

We have represented extreme collinearity in terms of the last explanatory variable. Since we can always re-order the variables this choice is without loss of generality and the analysis could be applied to any non-constant variable by moving it to the last column.

9.1.2 Near Extreme Collinearity

Of course, it is rare, in practice, that an exact linear relationship holds. Instead, we have

$$x_{ik} = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_{k-1} x_{i,k-1} + v_i \tag{9.4}$$

or, more compactly,

$$\mathbf{x}_k = \mathbf{X}_1 \mathbf{c} + \mathbf{v}, \tag{9.5}$$

where the v's are small relative to the x's. If we think of the v's as random variables they will have small variance (and zero mean if $\mathbf{X}$ includes a column of ones).

A convenient way to algebraically express the degree of collinearity is the sample correlation between x_{ik} and $w_i = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_{k-1} x_{i,k-1}$, namely

$$r_{x,w} = \frac{\text{cov}(\, x_{ik}, w_i \,)}{\sqrt{\text{var}(x_{i,k})\, \text{var}(w_i)}} = \frac{\text{cov}(\, w_i + v_i, w_i \,)}{\sqrt{\text{var}(w_i + v_i)\, \text{var}(w_i)}} \tag{9.6}$$

Clearly, as the variance of v_i grows small, this value will go to unity. For near extreme collinearity, we are talking about a high correlation between at least one variable and some linear combination of the others.

We are interested not only in the possibility of high correlation between x_{ik} and the linear combination $w_i = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_{k-1} x_{i,k-1}$ for a particular choice of $\mathbf{c}$ but for any choice of the coefficient. The choice which will maximize the correlation is the choice which minimizes $\sum_{i=1}^{n} w_i^2$ or least squares. Thus $\widehat{\mathbf{c}} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{x}_k$ and $\widehat{\mathbf{w}} = \mathbf{X}_1 \widehat{\mathbf{c}}$ and

$$(r_{x,\widehat{w}})^2 = R_{k\cdot}^2 \tag{9.7}$$

is the R^2 of this regression and hence the maximal correlation between x_{ki} and the other x's.

9.1.3 Absence of Collinearity

At the other extreme, suppose

$$R_{k\cdot}^2 = r_{x,\widehat{w}} = \text{cov}(\, x_{ik}, \widehat{w}_i \,) = 0. \tag{9.8}$$

That is, x_{ik} has zero correlation with all linear combinations of the other variables for any ordering of the variables. In terms of the matrices, this requires $\widehat{\mathbf{c}} = 0$ or

$$\mathbf{X}_1'\mathbf{x}_k = 0. \tag{9.9}$$

regardless of which variable is used as $\mathbf{x}_k$. This is called the case of orthogonal regressors, since the various x's are all orthogonal. This extreme is also very rare, in practice. We usually find some degree of collinearity, though not perfect, in any data set.

9.2 Consequences of Multicollinearity

9.2.1 For OLS Estimation

We will first examine the effect of x_{k1} being highly collinear upon the estimate $\widehat{\beta}_k$. Now let

$$\mathbf{x}_k = \mathbf{X}_1\mathbf{c} + \mathbf{v} \tag{9.10}$$

The OLS estimates are given by the solution of

$$\begin{aligned} \mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}'(\,\mathbf{X}_1 : \mathbf{x}_k\,)\widehat{\boldsymbol{\beta}} \\ &= (\,\mathbf{X}'\mathbf{X}_1 \colon \mathbf{X}'\mathbf{x}_k\,)\widehat{\boldsymbol{\beta}} \end{aligned} \tag{9.11}$$

Applying Cramer's rule to obtain $\widehat{\beta}_k$ yields

$$\widehat{\beta}_k = \frac{|\mathbf{X}'\mathbf{X}_1 : \mathbf{X}'\mathbf{y}|}{|\mathbf{X}'\mathbf{X}|} \tag{9.12}$$

However, as the collinearity becomes more extreme, the columns of $\mathbf{X}$ (the rows of $\mathbf{X}'$) become more linearly dependent and

$$\lim_{\mathbf{v}\to 0} \widehat{\beta}_1 = \frac{0}{0} \tag{9.13}$$

which is indeterminant.

Now, the variance-covariance matrix is

$$\begin{aligned}\sigma^2(\,\mathbf{X}'\mathbf{X}\,)^{-1} &= \sigma^2\frac{1}{|\mathbf{X}'\mathbf{X}|}\,\mathrm{adj}(\,\mathbf{X}'\mathbf{X}\,) \\ &= \sigma^2\frac{1}{|\mathbf{X}'\mathbf{X}|}\,\mathrm{adj}\left[\begin{pmatrix}\mathbf{X}_1' \\ \mathbf{x}_k'\end{pmatrix}(\,\mathbf{X}_1 : \mathbf{x}_k\,)\right] \\ &= \sigma^2\frac{1}{|\mathbf{X}'\mathbf{X}|}\,\mathrm{adj}\begin{pmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{x}_k \\ \mathbf{X}_1'\mathbf{x}_k & \mathbf{x}_k'\mathbf{x}_k\end{pmatrix}. \qquad (9.14)\end{aligned}$$

The variance of $\widehat{\beta}_k$ is given by the (k,k) element, so

$$\mathrm{var}(\,\widehat{\beta}_k\,) = \sigma^2\frac{1}{|\mathbf{X}'\mathbf{X}|}\,\mathrm{cof}(k,k) = \sigma^2\frac{1}{|\mathbf{X}'\mathbf{X}|}|\mathbf{X}_1'\mathbf{X}_1|. \qquad (9.15)$$

Thus, for $|\mathbf{X}_1'\mathbf{X}_1| \neq 0$, we have

$$\lim_{\mathbf{v}\to 0}\mathrm{var}(\,\widehat{\beta}_k\,) = \frac{\sigma^2|\mathbf{X}_1'\mathbf{X}_1|}{0} = \infty. \qquad (9.16)$$

and the variance of the collinear terms becomes unbounded.

It is instructive to give more structure to the variance of the last coefficient estimate in terms of the sample correlation $R_{k\cdot}^2$ given above. First we obtain the covariance of the OLS estimators other than the intercept. Denote $\mathbf{X} = (\ell : \mathbf{X}^*)$ where ℓ is an $n \times 1$ vector of ones and $\mathbf{X}^*$ are the nonconstant columns of $\mathbf{X}$, then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix}\ell'\ell & \ell'\mathbf{X}^* \\ \mathbf{X}^{*\prime}\ell & \mathbf{X}^{*\prime}\mathbf{X}^*\end{bmatrix}. \qquad (9.17)$$

Using the results for the inverse of a partitioned matrix we find that the lower right-hand $k-1 \times k-1$ submatrix of the inverse is given by

$$\begin{aligned}(\mathbf{X}^{*\prime}\mathbf{X}^* - \mathbf{X}^{*\prime}\ell(\ell'\ell)^{-1}\ell'\mathbf{X}^*)^{-1} &= (\mathbf{X}^{*\prime}\mathbf{X}^* - n\overline{\mathbf{x}}^*\overline{\mathbf{x}}^{*\prime})^{-1} \\ &= [(\mathbf{X}^* - \ell\overline{\mathbf{x}}^{*\prime})'(\mathbf{X}^* - \ell\overline{\mathbf{x}}^{*\prime})]^{-1} \\ &= (\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}\end{aligned}$$

where $\overline{\mathbf{x}}^* = \ell'\mathbf{X}^*/n$ is the mean vector for the nonconstant variables and $\overline{\mathbf{X}} = \mathbf{X}^* - \ell\overline{\mathbf{x}}^{*\prime}$ is the demeaned or deviation form of the data matrix for the nonconstant variables.

We now denote $\overline{\mathbf{X}} = (\overline{\mathbf{X}}_1 : \overline{\mathbf{x}}_k)$ where $\overline{\mathbf{x}}_k$ is last column $(k-1)^{\text{th}}$, then

$$\overline{\mathbf{X}}'\overline{\mathbf{X}} = \begin{bmatrix} \overline{\mathbf{X}}_1'\overline{\mathbf{X}}_1 & \overline{\mathbf{X}}'\overline{\mathbf{x}}_k \\ \overline{\mathbf{x}}_k'\overline{\mathbf{X}} & \overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k \end{bmatrix}. \tag{9.18}$$

Using the results for partitioned inverses again, the (k,k) element of the inverse of $(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}$ is given by,

$$\begin{aligned} (\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_k'\overline{\mathbf{X}_1}(\overline{\mathbf{X}}_1'\overline{\mathbf{X}}_1)^{-1}\overline{\mathbf{X}}_1'\overline{\mathbf{x}}_k)^{-1} &= 1/(\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_k'\overline{\mathbf{X}}_1(\overline{\mathbf{X}}_1'\overline{\mathbf{X}}_1)^{-1}\overline{\mathbf{X}}_1'\overline{\mathbf{x}}_k) \\ &= 1/\overline{\mathbf{e}}_k'\overline{\mathbf{e}}_k \\ &= 1/(\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k \cdot \overline{\mathbf{e}}_k'\overline{\mathbf{e}}_k/\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k) \\ &= 1/(\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k(1 - \frac{SSE_k}{SST_k})) \\ &= 1/(\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k(1 - R_{k\cdot}^2)) \end{aligned}$$

where $\overline{\mathbf{e}}_k = (I_n - \overline{\mathbf{X}}_1(\overline{\mathbf{X}}_1'\overline{\mathbf{X}}_1)^{-1}\overline{\mathbf{X}}_1')\overline{\mathbf{x}}_k$ are the OLS residuals from regressing the demeaned x_k's on the other variables and SSE_k, SST_k, and $R_{k\cdot}^2$ are the corresponding statistics for this regression. Thus we find

$$\begin{aligned} \text{var}(\widehat{\beta}_k) &= \sigma^2[(\,\mathbf{X}'\mathbf{X}\,)^{-1}]_{kk} = \sigma^2/(\overline{\mathbf{x}}_k'\overline{\mathbf{x}}_k(1 - R_{k\cdot}^2)) \\ &= \sigma^2/(\textstyle\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2(1 - R_{k\cdot}^2)) \\ &= \sigma^2/(n \cdot \frac{1}{n}\textstyle\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2(1 - R_{k\cdot}^2)). \end{aligned} \tag{9.19}$$

and the variance of $\widehat{\beta}_k$ increases with the noise σ^2 and the correlation $R_{k\cdot}^2$ of x_k with the other variables, and decreases with the sample size n and the signal $\frac{1}{n}\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2$.

Thus, as the collinearity becomes more and more extreme:

- The OLS estimates of the coefficients on the collinear terms become indeterminant. This is just a manifestation of the difficulties in obtaining $(\mathbf{X}'\mathbf{X})^{-1}$.
- The OLS coefficients on the collinear terms become infinitely variable. Their variances become very large as $R_{k\cdot}^2 \to 1$.
- The OLS estimates are still BLUE and with normal disturbances BUE. Thus, any unbiased estimator will be afflicted with the same problems.

Collinearity does not effect our estimate s^2 of σ^2. This is easy to see, since we have shown that

$$(\,n-k\,)\frac{s^2}{\sigma^2} \sim \chi^2_{n-k} \tag{9.20}$$

regardless of the values of $\mathbf{X}$, provided $\mathbf{X'X}$ still nonsingular. This is to be contrasted with the $\widehat{\beta}$ where

$$\widehat{\beta} \sim N(\,\beta, \sigma^2(\,\mathbf{X'X}\,)^{-1}\,) \tag{9.21}$$

clearly depends on $\mathbf{X}$ and more particularly the near non-invertibility of $\mathbf{X'X}$.

9.2.2 For Inferences

Provided collinearity does not become extreme, we still have the ratios $(\widehat{\beta}_j - \beta_j)/\sqrt{s^2 d^{jj}} \sim t_{n-k}$ where $d^{jj} = [(\,\mathbf{X'X}\,)^{-1}]_{jj}$. Although $\widehat{\beta}_j$ becomes highly variable as collinearity increases, d^{jj} grows correspondingly larger, thereby compensating. Thus under $H_0: \ \beta_j = \beta_j^0$, we find $(\widehat{\beta}_j - \beta_j^0)/\sqrt{s^2 d^{jj}} \sim t_{n-k}$, as is the case in the absence of collinearity. This result that the null distribution of the ratios is not impacted as collinearity becomes more extreme seems not to be fully appreciated in most texts.

The inferential price extracted by collinearity is loss of power. Under $H_0: \ \beta_j = \beta_j^1 \neq \beta_j^0$, we can write

$$(\widehat{\beta}_j - \beta_j^0)/\sqrt{s^2 d^{jj}} = (\widehat{\beta}_j - \beta_j^1)/\sqrt{s^2 d^{jj}} + (\beta_j^1 - \beta_j^0)/\sqrt{s^2 d^{jj}}. \tag{9.22}$$

The first term will continue to follow a t_{n-k} distribution, as argued in the previous paragraph, as collinearity becomes more extreme. However, the second term, which represents a "shift" term, will grow smaller as collinearity becomes more extreme and d^{jj} becomes larger. Thus we are less likely to shift the statistic into the tail of the ostensible null distribution and hence less likely to reject the null hypothesis. Formally, $(\widehat{\beta}_j - \beta_j^0)/\sqrt{s^2 d^{jj}}$ will have a noncentral t distribution, but the noncentrality parameter will become smaller and smaller as collinearity becomes more extreme.

Alternatively the inferential impact can be seen through the impact on the confidence intervals. Using the standard approach discussed in the previous chapter, we have $[\widehat{\beta}_j - a\sqrt{s^2 d^{jj}}, \widehat{\beta}_j + a\sqrt{s^2 d^{jj}})$ as the 95% confidence interval, where a is the critical value for a .025 tail. Note that as collinearity

becomes more extreme and d^{jj} becomes larger, the width of the interval becomes larger as well. Thus we see that the estimates are consistent with a larger and larger set of null hypothesis as the collinearity strengthens. In the limit it is consistent with any null hypothesis and we have zero power.

We should emphasize that collinearity does not always cause problems. The shift term in (9.xx) can be written

$$(\beta_j^1 - \beta_j^0)/\sqrt{s^2 d^{jj}} = \sqrt{n}(\beta_j^1 - \beta_j^0)/\sqrt{\sigma^2/(\frac{1}{n}\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2(1 - R_{k.}^2))}$$

which clearly depends on other factors than the degree of collinearity. The size of the shift increases with the sample size $\sqrt{n}$, the difference between the null and alternative hypotheses $(\beta_j^1 - \beta_j^0)$, and the signal noise ratio $(\frac{1}{n}\sum_{i=1}^n (x_{ik} - \overline{x}_k)^2/\sigma^2$. The important question is not whether collinearity is present or extreme but whether is is extreme enought to eliminate the power of our test. This is also a phenonmenon that does not seem to be fully appreciated or well-enough advertised in most texts.

We can easily tell when collinearity is *not* a problem if the coefficients are significant or we reject the null hypothesis under consideration. Only if apparently important variables are insignificantly different from zero or have the wrong sign should we consider the possibility that collinearity is causing problems.

9.2.3 For Prediction

If all we are interested in is prediction of y_p given $x_{p1}, x_{p2}, \ldots, x_{pk}$, then we are not particularly interested in whether or not we have isolated the individual effects of each x_{ij}. We are interested in predicting the total effect or variation in y.

A good measure of how well the linear relationship captures the total effect or variation is the R^2 statistic. But the R^2 value is related to s^2 by

$$R^2 = 1 - \frac{e'e}{(y - \overline{y})'(y - \overline{y})} = 1 - (n - k)\frac{s^2}{\text{var}(y)}, \tag{9.23}$$

which does not depend upon the collinearity of $\mathbf{X}$.

Thus, we can expect our regressions to predict well, despite collinearity and insignificant coefficients, provided the R^2 value is high. This depends, of course, upon the collinearity continuing to persist in the future. If the collinearity does not continue, then prediction will become increasingly uncertain. Such uncertainty will be reflected, however, by the estimated standard errors of the forcast and hence wider forecast intervals.

9.2.4 An Illustrative Example

As an illustration of the problems introduced by collinearity, consider the consumption equation

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 W_t + u_t, \tag{9.24}$$

where C_t is consumption expenditures at time t, Y_t is income at time t and W_t is wealth at time t. Economic theory suggests that the coefficient on income should be slightly less than one and the coefficient on wealth should be positive. The time-series data for this relationship are given in the following table:

C_t	Y_t	W_t
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

Table 9.1: Consumption Data

Applying least squares to this equation and data yields

$$C_t = \underset{(6.752)}{24.775} + \underset{(0.823)}{0.942} Y_t - \underset{(0.081)}{0.042} W_t + e_t,$$

where estimated standard errors are given in parenthesis. Summary statistics for the regression are: $SSR = 324.446$, $s^2 = 46.35$, and $R^2 = 0.9635$. The coefficient estimate for the marginal propensity to consume seems to be a reasonable value however it is not significantly different from either zero or one. And the coefficient on wealth is negative, which is not consistent with economic theory. Wrong signs and insignificant coefficient estimates on a priori important variables are the classic symptoms of collinearity. As an indicator of the possible collinearity the squared correlation between Y_t and W_t is .9979, which suggests near extreme collinearity among the explanatory variables.

9.3 Detecting Multicollinearity

9.3.1 When Is Multicollinearity a Problem?

Suppose the regression yields significant coefficients, then collinearity is not a problem—even if present. On the other hand, if a regression has insignificant coefficients, then this may be due to collinearity or that the variables, in fact, do not enter the relationship.

9.3.2 Zero-Order Correlations

If we have a trivariate relationship, say

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \tag{9.25}$$

we can look at the zero-order correlation between x_2 and x_3. As a rule of thumb, if this (squared) value exceeds the R^2 of the original regression, then we have a problem of collinearity. If r_{23} is low, then the regression is likely insignificant.

In the previous example, $r^2_{WY} = 0.9979$, which indicates that Y_t is more highly related to W_t than C_t and we have a problem. In effect, the variables are so closely related that the regression has difficulty untangling the separate eefects of Y_t and W_t.

In general ($k > 3$), when one of the zero-order correlations between xs is large relative to R^2 we have a problem.

9.3.3 Partial Regressions

In the general case ($k > 3$), even if all the zero-order correlations are small, we may still have a problem. For while x_1 my not be strongly linearly related to any single x_i ($i \neq 1$), it may be very highly correlated with some linear combination of xs.

To test for this possibility, we should run regressions of each x_i on all the other xs. If collinearity is present, then one of these regressions will have a high R^2 (relative to R^2 for the complete regression).

For example, when $k = 4$ and

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t \tag{9.26}$$

is the regression, then collinearity is indicated when one of the partial regressions

$$\begin{aligned} x_{t2} &= \alpha_1 + \alpha_3 x_{t3} + \alpha_4 x_{t4} \\ x_{t3} &= \gamma_1 + \gamma_2 x_{t2} + \gamma_4 x_{t4} \quad (9.27) \\ x_{t4} &= \delta_1 + \delta_2 x_{t2} + \delta_3 x_{t3} \quad (9.28) \end{aligned}$$

yields a large R^2 relative to the complete regression.

9.3.4 The F Test

The manifestation of collinearity is that estimators become insignificantly different from zero, due to the inability to untangle the separate effects of the collinear variables. If the insignificance is due to collinearity, the total effect is not confused, as evidenced by the fact that s^2 is unaffected.

A formal test, accordingly, is to examine whether the total effect of the insignificant (possibly collinear) variables is significant. Thus, we perform an F test to test the joint hypothesis that the individually insignificant variables are all insignificant.

For example, if the regression

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t \quad (9.29)$$

yields insignificant (from zero) estimates of β_2, β_3 and β_4, we use an F test of the joint hypothesis $\beta_2 = \beta_3 = \beta_4 = 0$. If we reject this joint hypothesis, then the total effect is strong, but the individual effects are confused. This is evidence of collinearity. If we accept the null, then we are forced to conclude that the variables are, in fact, insignificant.

For the consumption example considered above, a test of the null hypothesis that the collinear terms (income and wealth) are jointly zero yields an F-statistic value of 92.40 which is very extreme under the null when the variable has an $F_{2,7}$. Thus the variables are individually insignificant but are jointly significant, which indicates that collinearity is, in fact, a problem.

9.3.5 The Condition Number

Belsley, Kuh, and Welsh (1980), suggest an approach that considers the invertibility of $\mathbf{X}$ directly. First, we transform each column of $\mathbf{X}$ so that they are of similar scale in terms of variability by dividing each column to

unit length:

$$\mathbf{x}_j^* = \mathbf{x}_j / \sqrt{\mathbf{x}_j' \mathbf{x}_j} \tag{9.30}$$

for $j = 1, 2, ..., k$. Next we find the eigenvalues of the moment matrix of the so-transformed data matrix by finding the k roots of :

$$\det(\mathbf{X}^{*\prime}\mathbf{X}^* - \lambda \mathbf{I}_k) = 0. \tag{9.31}$$

Note that since $\mathbf{X}^{*\prime}\mathbf{X}^*$ is positive semi-difinite the eigenvalues will be between zero and one with values of zero in the event of singularity and close to zero in the event of close to singularity. The condition number of the matrix is taken as the ratio of the largest to smallest of the eigenvalues:

$$c = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{9.32}$$

Using an analyis of a number of problems BKW suggest that collinearity is a possible issue when $c \geq 20$. For the example the condition number is 166.245, which indicates a very poorly conditioned matrix. Although this approach tells a great deal about the invertibility of $\mathbf{X}'\mathbf{X}$ and hence the signal, it tells us nothing about the noise level relative to the signal.

9.4 Correcting For Collinearity

9.4.1 Additional Observations

Professor Goldberger has quite aptly described multicollinearity as "micronumerosity" or not enough observations. Recall that the shift term depends on the difference between the null and alternative, the signal-noise ratio, and the sample size. For a given signal-noise ratio, unless collinearity is extreme, it can always be overcome by increasing the sample size sufficiently. Moreover, we can sometimes gather more data that, hopefully, will not suffer the collinearity problem. With designed experiments, and cross-sections, this is particularly the case. With time series data this is not feasible and in any event gathering more data is time-conusming and expensive.

9.4.2 Independent Estimation

Sometimes we can obtain outside estimates. For example, in the Ando-Modigliani consumption equation

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 W_t + u_t, \tag{9.33}$$

we might have a cross-sectional estimate of β_1, say $\widehat{\beta}_1$. Then,

$$(C_t - \widehat{\beta}_1 Y_t) = \beta_0 + \beta_2 W_t + u_t \tag{9.34}$$

becomes the new problem. Treating $\widehat{\beta}_1$ as known allows estimation of β_2 with increases precision. It would not reduce the precision of the estimate of β_1 which would simply be the cross-sectional estimate. The implied error term, moreover, is more complicated since $\widehat{\beta}_1$ may be correlated with W_t. Mixed estimation approaches should be used to handle this approach carefully. Note that this is another way to gather more data.

9.4.3 Prior Restrictions

Consider the consumption equation from Klein's Model I:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} \beta_3 W_t + \beta_4 W_t' + u_t, \tag{9.35}$$

where C_t is the consumption expenditure, P_t is profits, W_t is the private wage bill and W_t' is the governement wage bill.

Due to market forces, W_t and W_t' will probably move together and collinearity will be a problem for β_3 and β_4. However, there is no prior reason to discriminate between W_t and W_t' in their effect on C_t. Thus it is reasonable to suppose W_t and W_t' impact C_t in the same way. That is, $\beta_3 = \beta_4$. The model is now

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} \beta (W_t + W_t') + u_t, \tag{9.36}$$

which should avoid the collinearity problem.

9.4.4 Ridge Regression

One manifestation of collinearity is that the effected estimates, say $\widehat{\beta}_1$, will be extreme with a high probability. Thus,

$$\sum_{i=1}^{k} \widehat{\beta}_i^2 = \widehat{\beta}_1^2 + \widehat{\beta}_2^2 + \cdots + \widehat{\beta}_k^2 = \widehat{\beta}' \widehat{\beta} \tag{9.37}$$

will be large with a high probability.

By way of treating the disease by treating its symptoms, we might restrict $\widehat{\beta}' \widehat{\beta}$ to be small. Thus, we might reasonably

$$\min_{\widehat{\beta}} (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta}) \qquad \text{subject to } \widehat{\beta}' \widehat{\beta} \leq m. \tag{9.38}$$

Form the Lagrangian (since $\widehat{\beta}'\widehat{\beta}$ is large, we must impose teh restriction with equality).

$$\begin{aligned}\mathcal{L} &= (\mathbf{y}-\mathbf{X}\widetilde{\beta})'(\mathbf{y}-\mathbf{X}\widetilde{\beta})+\lambda(m-\widetilde{\beta}'\widetilde{\beta}) \\ &= \sum_{t=1}^{n}\left(y_t-\sum_{i=1}^{k}\widetilde{\beta}_i x_{ti}\right)^2++\lambda(m-\sum_{i=1}^{k}\widetilde{\beta}_i^2). \end{aligned} \tag{9.39}$$

The first-order conditions yield

$$\frac{\partial\mathcal{L}}{\partial\beta_j}=-2\sum_t\left(y_t-\sum_i\widetilde{\beta}_i x_{ti}\right)x_{tj}+2\lambda\widetilde{\beta}_i^2=0, \tag{9.40}$$

or

$$\begin{aligned}\sum_t y_t x_{tj} &= \sum_t\sum_i x_{ti}x_{tj}\widetilde{\beta}_i+\lambda\widetilde{\beta}_j \\ &= \sum_i\widetilde{\beta}_i\sum_t x_{ti}x_{tj}+\lambda\widetilde{\beta}_j, \end{aligned} \tag{9.41}$$

for $j=1,2,\ldots,k$. In matrix form, we have

$$\mathbf{X}'\mathbf{y}=(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)\widetilde{\beta}. \tag{9.42}$$

So, we have

$$\widetilde{\beta}=(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{y}. \tag{9.43}$$

This is called ridge regression.

Substition yields

$$\begin{aligned}\widetilde{\beta} &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'(\mathbf{X}\beta+\mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'(\mathbf{X}\beta+(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{u}) \end{aligned} \tag{9.44}$$

and

$$\mathrm{E}(\widetilde{\beta})=(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{X}\beta=\mathbf{P}\beta, \tag{9.45}$$

so ridge regression is biased. Rather obviously, as λ grows large, the expectation "shrinks" towards zero so the bias is towards zero. Next, we find that

$$\mathrm{Cov}(\widetilde{\beta})=\sigma^2(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_n)^{-1}=\sigma^2\mathbf{Q}<\sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{9.46}$$

If $\mathbf{u} \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_n)$, then

$$\widetilde{\beta} \sim \mathrm{N}(\mathbf{P}\beta, \sigma^2 \mathbf{Q}) \tag{9.47}$$

and inferences are possible only for $\mathbf{P}\beta$ and hence the complete vector.

The rather obvious question in using ridge regression is what is the best choice for λ? We seek to trade off the increased bias against the reduction in the variance. This may be done by considering the mean square error (MSE) which is given by

$$\begin{aligned} \mathrm{MSE}(\widetilde{\beta}) &= \sigma^2 \mathbf{Q} + (\mathbf{P} - \mathbf{I}_k)\beta\beta'(\mathbf{P} - \mathbf{I}_k) \\ &= (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_n)^{-1}\{\sigma^2 \mathbf{X}'\mathbf{X} + \lambda^2 \beta\beta'\}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_n)^{-1}. \end{aligned}$$

We might choose to minimize the determinant or trace of this function. Note that either is an decreasing function of λ through the inverses and an increasing function through the term in brackets. Note also that the minimand depends on the true unkown β, which makes it infeasible.

In practice, it is useful to obtain what is called a ridge trace, which plots out the estimates, estimated standard error, and estimated square root of mean squared error (SMSE) as a function of λ. Problematic terms will frequently display a change of sign and a dramatic reduction in the SMSE. If this phenonmenon occurs at a sufficiently small value of λ, then the bias will be small and inflation in SMSE relative to the standard error will be small and we can conduct inference in something like the usual fashion. In particular, if the estimate of a particular coefficient seems to be significantly different from zero despite the bias toward zero, we can reject the null that it is zero.

Chapter 10

Stochastic Explanatory Variables

10.1 Nature of Stochastic X

In previous chapters, we made the assumption that the x's are nonstochastic, which means they are not random variables. This assumption was motivated by the control variables in controlled experiments, where we can choose the values of the independent variables. Such a restriction allows us to focus on the role of the disturbances in the process and was most useful in working out the stochastic properties of the estimators and other statistics. Unfortunately, economic data do not usually come to us in this form. In fact, the independent variables are random variables much like the dependent variable whose values are beyond the control of the researcher.

Consequently we will restate our model and assumptions with an eye toward stochastic x. The model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \text{ for } i = 1, 2, ..., n. \tag{10.1}$$

The assumptions with respect to unconditional moments of the disturbances are the same as before:

(i) $\mathrm{E}[u_i] = 0$

(ii) $\mathrm{E}[u_i^2] = \sigma^2$

(iii) $\mathrm{E}[u_i u_j] = 0$, $j \neq i$

The assumptions with respect to x must be modified. We replace the assumption of x nonstochastic with an assumption regarding the joint stochastic behavior of u_i and $\mathbf{x}_i$. Several alternative assumptions will be introduced regarding the degree of dependence between $\mathbf{x}_i$ and u_i. For stochastic x, the assumption of linearly independent x's implies that the covariance matrix of the x's has full column rank and is hence positive definite. Stated formally, we have:

(iv) (u_i,$\mathbf{x}_i$) jointly *i.i.d.* with {dependence assumption}

(v) $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] = \mathbf{Q}$ p.d.

Notice that the assumption of normality, which was introduced in previous chapters to facilitate inference, was not reintroduced. Thus we are effectively relaxing both the nonstochastic regressor and normality assumptions at the same time. The motivation for dispensing with the normality assumption will become apparent presently.

We will now examine the various alternative assumptions that will be entertained with respect to the degree of dependence between $\mathbf{x}_i$ and u_i.

10.1.1 Independent X

The strongest assumption we can make relative to this relationship is that $\mathbf{x}_i$ are stochastically independent of u_i. This means that the distribution of $\mathbf{x}_i$ depends in no way on the value of u_i and visa versa. Note that

$$\mathrm{cov}\,(\,g(\mathbf{x}_i), h(u_i)\,) = 0, \tag{10.2}$$

for any functions $g(\cdot)$ and $h(\cdot)$ in this case.

10.1.2 Conditional Zero Mean

The next strongest assumption is $\mathrm{E}[u_i|\mathbf{x}_i] = 0$, which implies

$$\mathrm{cov}\,(\,g(\mathbf{x}_i), u_i)\,) = 0, \tag{10.3}$$

for any funtion $g(\cdot)$. This assumption is motivated by the assumption that our model is simply a statement of conditional expectation, $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta$, and may or may not be accompanied by a conditional second moment assumption such as $\mathrm{E}[u_i^2|\mathbf{x}_i] = \sigma^2$. Note that independence along with the unconditional statements $\mathrm{E}[u_i] = 0$ and $\mathrm{E}[u_i^2] = \sigma^2$ imply conditional zero mean and constant conditional variance, but not the reverse.

10.1.3 Uncorrelated X

A less strong assumption is that x_{ij} and u_i are uncorrelated, that is,

$$\text{cov}\,(\,x_{ij}, u_i\,) = \text{E}(\,x_{ij}, u_i\,) = 0. \tag{10.4}$$

The properties of $\widehat{\boldsymbol{\beta}}$ are less accessible in this case. Note that conditional zero mean always implies uncorrelated, but not the reverse. It is possible to have a random variables that are uncorrelated but neither has constant conditional mean given the other. In general, the conditional second moment will also be nonconstant.

10.1.4 Correlated X

A priori information sometimes suggests the possibility that x_{ij} is correlated with u_i. That is, that is,

$$\text{E}(\,x_{ij}, u_i\,) \neq 0. \tag{10.5}$$

As we shall see below, this can have quite serious implications for the OLS estimates.

An example is the case of simultaneous equations models that we will examine later. A second example occurs when our right-hand side variables are measured with error. Suppose

$$y_i = \alpha + \beta x_i + u_i \tag{10.6}$$

is the true model but

$$x_i = x_i + v_i \tag{10.7}$$

is the only available measurement of x_i. If we use x_i^* in our regression, then we are estimating the model

$$\begin{aligned} y_i &= \alpha + \beta(\,x_i^* - v_i\,) + u_i \\ &= \alpha + \beta x_i^* + (\,u_i - \beta v_i\,). \end{aligned} \tag{10.8}$$

Now, even if the measurement error v_i were independent of the disturbance u_i, the right-hand side variable x_t^* will be correlated with the effective disturbance $(u_i - \beta v_i)$.

10.2 Consequences of Stochastic X

10.2.1 Consequences for OLS Estimation

Recall that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}X'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\
&= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
&= \boldsymbol{\beta}+(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}'\mathbf{u} \\
&= \boldsymbol{\beta}+\left(\frac{1}{n}\sum_{j=1}^{n}x_jx_j'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_iu_i\right)
\end{aligned} \tag{10.9}$$

We will now examine the bias and consistency properties of the estimators under the alternative dependence assumptions.

Uncorrelated X

Suppose that $\mathbf{x}_t$ is only assumed to be uncorrelated with u_i. Rewrite the second term in (10.9) as

$$\begin{aligned}
\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_iu_i\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j'\right)^{-1}\mathbf{x}_i\right]u_i \\
&= \frac{1}{n}\sum_{i=1}^{n}w_iu_i.
\end{aligned}$$

Note that w_i is a function of both $\mathbf{x}_i$ and $\mathbf{x}_j$ and is nonlinear in $\mathbf{x}_i$ for $j=i$. Now u_i is uncorrelated with the $\mathbf{x}_j$ for $j\neq i$ by independence and the level of $\mathbf{x}_i$ by the assumption but is not necessarily uncorrelated with the nonlinear function of $\mathbf{x}_i$. Thus,

$$\mathrm{E}[\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{u}]\neq\mathbf{0}, \tag{10.10}$$

in general, whereupon

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}]=\boldsymbol{\beta}+\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\neq\boldsymbol{\beta}. \tag{10.11}$$

Similarly, we find $\mathrm{E}[s^2]\neq\sigma^2$. Thus both $\widehat{\boldsymbol{\beta}}$ and s^2 will be biased, although the bias may disappear asymptotically as we will see below. Note that sometimes these moments are not well defined.

Now, each element of $x_i x_i'$ and $\mathbf{x}_i u_i$ are *i.i.d.* random variables with expectations Q and 0, respectively. Thus, the law of large numbers guarantees that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i \xrightarrow{p} \mathrm{E}\,\mathbf{x}_i\mathbf{x}_i = \mathbf{Q}, \tag{10.12}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i \xrightarrow{p} \mathrm{E}\,\mathbf{x}_i u_i = \mathbf{0}. \tag{10.13}$$

It follows that

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty}\widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{u} \\
&= \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{u} \\
&= \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot 0 = \boldsymbol{\beta}.
\end{aligned} \tag{10.14}$$

Similarly, we can show that

$$\operatorname*{plim}_{n\to\infty} s^2 = \sigma^2. \tag{10.15}$$

Thus both $\widehat{\boldsymbol{\beta}}$ and s^2 will be consistent.

Conditional Zero Mean

Suppose $\mathrm{E}[u_i|\mathbf{x}_i] = 0$. Then,

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} + \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}] \\
&= \boldsymbol{\beta} + \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\mathbf{u}|\mathbf{X})] \\
&= \boldsymbol{\beta} + \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot\mathbf{0}] = \boldsymbol{\beta}.
\end{aligned}$$

and OLS is unbiased. The Gauss-Markov theorem continues to hold in the sense that least-squares is BLUE in the class of estimators that is linear in $\mathbf{y}$ with the linear transformation matrix a function only of X. For the variance estimator, we can show unbiasedness, in a similar fashion,

$$\begin{aligned}
\mathrm{E}\,s^2 &= \mathrm{E}\,\mathbf{e}'\mathbf{e}/(n-k) \\
&= \frac{1}{n-k}\,\mathrm{E}\,\mathbf{u}'(\mathrm{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{u} \\
&= \sigma^2,
\end{aligned} \tag{10.16}$$

provided that $\mathrm{E}[u_i^2|\mathbf{x}_i] = \sigma^2$. Naturally, since conditional zero mean implies uncorrelatedness, then we have the same consistency results, namely

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} \qquad \text{and} \qquad \operatorname*{plim}_{n\to\infty} s^2 = \sigma^2. \tag{10.17}$$

Independent X

Suppose $\mathbf{x}_i$ is independent of u_i. Then, provided $\mathrm{E}[u_i] = 0$ and $\mathrm{E}[u_i^2] = \sigma^2$, we have conditional zero mean and the corresponding unbiasedness results

$$\mathrm{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{E}\,s^2 = \sigma^2, \tag{10.18}$$

together with the BLUE property and the consistency results

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} \qquad \text{and} \qquad \operatorname*{plim}_{n\to\infty} s^2 = \sigma^2. \tag{10.19}$$

Correlated X

Suppose that $\mathbf{x}_i$ is correlated with u_i. That is,

$$\mathrm{E}\,\mathbf{x}_i u_i = \mathbf{d} \neq \mathbf{0}. \tag{10.20}$$

Obviously, since $\mathbf{x}_i$ is correlated with u_i there is no reason to believe $\mathrm{E}[\{(X'X)^{-1}X'\}\mathbf{u}] = 0$ so the OLS estimator will be biased. Turning to the possibility of consistency, we see, by the law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i u_i \xrightarrow{p} \mathrm{E}\,\mathbf{x}_i u_i = \mathbf{d}, \tag{10.21}$$

whereupon

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot\mathbf{d} \neq \boldsymbol{\beta} \tag{10.22}$$

since Q^{-1} is nonsingular and $\mathbf{d}$ is nonzero. Thus OLS is also inconsistent.

10.2.2 Consequences for Inferences

In previous chapters, the assumption of normal disturbances was introduced to facilitate interences. Together with the nonstochastic regressor assumption, it implied that the distribution of the least squares estimator,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

which is linear in the disturbances, has an unconditional normal distribution. If the x's are random variables, however, the unconditional distribution of the estimator will not be normal, in general, since the estimator is a rather complicated function of *both* X and $\mathbf{u}$. For example, if the x's are also normal, the estimator will be non-normal. Fortunately, we can appeal to the central limit theorem for help in large samples.

We shall develop the large-sample asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ only for the case of independence. The limiting behavior is identical for the conditional zero mean case with constant conditional variance. Now,

$$\operatorname*{plim}_{n\to\infty}(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\,) = \mathbf{0} \tag{10.23}$$

in this case so in order to have a nondegenerate distribution we consider

$$\sqrt{n}(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\,) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\sqrt{n}\frac{1}{n}\mathbf{X}'\mathbf{u}. \tag{10.24}$$

The typical element of

$$\frac{1}{n}X'\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i \tag{10.25}$$

is

$$\frac{1}{n}\sum_{i=1}^{n}x_{ij}u_i. \tag{10.26}$$

Note the $\mathbf{x}_{ij}u_i$ are *i.i.d.* random variables with

$$\mathrm{E}\,x_{ij}u_i = \mathbf{0} \tag{10.27}$$

and

$$\mathrm{E}(\,x_{ij}u_i\,)^2 = \mathrm{E}_x\,x_{ij}\,\mathrm{E}_u\,u_i = \sigma^2 q_{jj}, \tag{10.28}$$

where q_{jj} is the jj-th element of $\mathbf{Q}$. Thus, according to the central limit theorem,

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}x_{ij}u_i \xrightarrow{d} N(0,\sigma^2 q_{ii}). \tag{10.29}$$

In general,

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i = \frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{u} \xrightarrow{d} N(\mathbf{0},\sigma^2\mathbf{Q}). \tag{10.30}$$

Since $\frac{1}{n}X'X$ converges in probability to the fixed matrix $\mathbf{Q}$, we have

$$\sqrt{n}(\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\,) \xrightarrow{d} \mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{u} \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}). \tag{10.31}$$

For inferences,

$$\sqrt{n}(\widehat{\beta}_j - \beta_j) \xrightarrow{d} N(0, \sigma^2 q_{jj}) \tag{10.32}$$

and

$$\frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{\sqrt{\sigma^2 q_{jj}}} \xrightarrow{d} N(0,1). \tag{10.33}$$

Unfortunately, neither σ^2 nor Q^{-1} are available. We can use s^2 as a consistent estimate of σ^2 and

$$\widehat{Q} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = n(\mathbf{X}'\mathbf{X})^{-1} \tag{10.34}$$

as a consistent estimate of Q. Substituting, we have

$$\frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{\sqrt{s^2\widehat{q}^{jj}}}\frac{\sqrt{n}(\widehat{\beta}_j - \beta_j)}{\sqrt{s^2[n(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} \xrightarrow{d} N(0,1). \tag{10.35}$$

Thus, the usual statistics we use in conducting t-tests are asymptotically standard normal. This is particularly convenient since the t-distribution converges to the standard normal. The small-sample inferences we learned for the nonstochastic regressor case are appropriate in large samples for the stochastic regressor case.

In a similar fashion, we can show that the approach introduced in previous chapters for inference on complex hypotheses that had an F-distribution under normality with nonstochastic regressors continue to be appropriate in large samples with non-normality and stochastic regressors. For example, consider again the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{10.36}$$

with $\mathrm{H}_0\colon \boldsymbol{\beta}_2 = 0$ and $\mathrm{H}_1\colon \boldsymbol{\beta}_2 \neq 0$. Regression on this unrestriced model yields SSE_u while regression on the restricted model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}$ yields the SSE_r. We form the statistic $[(SSE_r - SSE_u)/k_2]/[SSE_u/(n-k)]$,

where k is the unrestricted numer of regressors and k_2 is the number of restrictions. Under normality and nonstochastic regressors this statistic will have a $F_{k_2,n-k}$ distribution. Note that asymptotically, as n becomes large, the denominator converges to σ^2 and the $F_{k_2,n-k}$ distribution converges to a $\chi^2_{k_2}$ distribution (divided by k_2). But this would be the limiting distribution of this statistic even if the regressors are nonstochastic and the disturbances non-normal.

10.3 Correcting for Correlated X

10.3.1 Instruments

Consider

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{10.37}$$

and premultiply by $\mathbf{X}'$ to obtain

$$\begin{aligned} \mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{u} \\ \frac{1}{n}\mathbf{X}'\mathbf{y} &= \frac{1}{n}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}'\mathbf{u}. \end{aligned} \tag{10.38}$$

If $\mathbf{X}$ is uncorrelated with $\mathbf{u}$, then the last term disappears in large samples:

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{y} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}, \tag{10.39}$$

which may be solved to obtain

$$\begin{aligned} \boldsymbol{\beta} &= \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right) \frac{1}{n}\mathbf{X}'\mathbf{y} \\ &= \operatorname*{plim}_{n\to\infty} (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y} = \operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}}. \end{aligned} \tag{10.40}$$

Of course, if $\mathbf{X}$ is correlated with $\mathbf{u}$, say $\mathrm{E}\,\mathbf{x}_i u_i = \mathbf{d}$, then

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{u} = \mathbf{d} \qquad \text{and} \qquad \operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}. \tag{10.41}$$

Suppose we can find iid variables $\mathbf{z}_i$ that are independent and hence uncorrelated with u_i, then

$$\mathrm{E}\,\mathbf{z}_i u_i = \mathbf{0}. \tag{10.42}$$

Also, suppose that the $\mathbf{z}_i$ are correlated with $\mathbf{x}_i$ so

$$\mathrm{E}\,\mathbf{z}_i\mathbf{x}_i = \mathbf{P}, \qquad \mathrm{E}\,\mathbf{z}_i\mathbf{z}_i = \mathbf{M}, \tag{10.43}$$

and P is nonsingular. Such variables are known as instruments for the variables $\mathbf{x}_i$.

10.3.2 Instrumental Variable (IV) Estimation

Suppose that we premultiply (10.37) by

$$\mathbf{Z}' = (z_1, z_2, \ldots, z_n) \tag{10.44}$$

to obtain

$$\begin{aligned} \mathbf{Z}'\mathbf{y} &= \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u} \\ \frac{1}{n}\mathbf{Z}'\mathbf{y} &= \frac{1}{n}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}\mathbf{Z}'\mathbf{u}. \end{aligned} \tag{10.45}$$

But since $\mathrm{E}\,\mathbf{z}_i u_i = 0$, then

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{u} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i u_i = \mathbf{0}, \tag{10.46}$$

so

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{y} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} \tag{10.47}$$

or

$$\begin{aligned} \boldsymbol{\beta} &= \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1} \frac{1}{n}\mathbf{Z}'\mathbf{y} \\ &= \operatorname*{plim}_{n\to\infty} (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \end{aligned} \tag{10.48}$$

Now,

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \tag{10.49}$$

is known as the instrumental variable (IV) estimator. Note that OLS is an IV estimator with $\mathbf{X}$ chosen as the instruments.

10.3.3 Properties of the IV Estimator

We have just shown that

$$\operatorname*{plim}_{n\to\infty} \widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}, \tag{10.50}$$

so the IV estimator is consistent. In small samples, since

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\ &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u}, \end{aligned} \tag{10.51}$$

we generally have bias since we are only assured that $\mathbf{z}_i$ is uncorrelated with u_i, but not that $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'$ is uncorrelated with $\mathbf{u}$.

Asymptotically, if $\mathbf{z}_i$ is independent of u_i, then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} (\mathbf{0}, \sigma^2\mathbf{P}^{-1}\mathbf{M}\mathbf{P}^{-1}), \tag{10.52}$$

where, as above,

$$\mathbf{P} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{X} \qquad \text{and} \qquad \mathbf{M} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{Z}. \tag{10.53}$$

Let

$$\widetilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}} \tag{10.54}$$

be the IV residuals. Then

$$\widetilde{s}^2 = \frac{\widetilde{\mathbf{e}}'\widetilde{\mathbf{e}}}{n-k} = \sum_{i=1}^{n} \frac{\widetilde{e}_t^2}{n-k} \tag{10.55}$$

is consistent.

The ratios

$$\frac{\widetilde{\beta}_j - \beta_j}{\sqrt{\widetilde{s}^2[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}]_{jj}}} \xrightarrow{d} N(0,1), \tag{10.56}$$

where the denominator is printed by IV packages as the estimated standard errors of the estimates.

10.3.4 Optimal Instruments

The instruments $\mathbf{z}_i$ cannot be just any variables that are independent of and uncorrelated with u_i. They should be as closely related to $\mathbf{x}_i$ as possible while remaining uncorrelated with u_i.

Looking at the asymptotic covariance matrix $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}^{-1}$, we can see as $\mathbf{z}_i$ and $\mathbf{x}_i$ become unrelated and hence uncorrelated, that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{X} = \mathbf{P} \tag{10.57}$$

goes to zero. The inverse of P consequently grows large and $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}^{-1}$ will become large. Then the consequence of using $\mathbf{z}_i$ that are not close to $\mathbf{x}_i$ is imprecise estimates. In fact, we can speak of optimal instruments as being all of $\mathbf{x}_i$ except the part that is correlated with u_i.

10.4 Detecting Correlated X

10.4.1 An Incorrect Procedure

With other problems of OLS we have examined the OLS residuals for signs of the problem. In the present case, where u_i being correlated with $\mathbf{x}_i$ is the problem, we might naturally see if our proxy for u_i, the OLS residuals e_t, are correlated with $\mathbf{x}_i$. Thus,

$$\sum_{i=1}^{n} \mathbf{x}_i e_t = \mathbf{X}'\mathbf{e} \tag{10.58}$$

might be taken as an indication of any correlation between $\mathbf{x}_i$ and u_i. Unfortunately, one of the properties of OLS guarantees that

$$\mathbf{X}'\mathbf{e} = \mathbf{0}. \tag{10.59}$$

Thus, this procedure will not work.

10.4.2 A Priori Information

Typically, we know that $\mathbf{x}_i$ is correlated with u_i as a result of the structure of the model. For example, in the errors in variables models. In such cases, the candidates for instruments also usually are evident.

10.4.3 Example: Simultaneous Equations

Consider the consumption equation

$$C_t = \alpha + \beta Y_t + u_t, \tag{10.60}$$

where income, Y_t, is defined by the identity

$$Y_t = C_t + G_t. \tag{10.61}$$

Substituting (10.60) into (10.61), we obtain

$$Y_t = \alpha + \beta Y_t + u_t + G_t, \tag{10.62}$$

and solving for Y_t,

$$Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} G_t + \frac{1}{1-\beta} u_t. \tag{10.63}$$

Rather obviously, Y_t is linearly related and hence correlated with u_t. A candidate as an instrument for Y_t is the exogenous variable G_t.

10.4.4 An IV Approach

As a test of whether $\mathbf{x}_i$ is correlated with u_i and hence

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta},$$

the most common procedure is to compare the OLS and IV estimates. If OLS is consistent, we expect to find between no difference between the two. If not, then

$$\operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta} = \operatorname*{plim}_{n\to\infty} \widetilde{\boldsymbol{\beta}}$$

and the difference will show up.

More formally,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \xrightarrow{d} \left[\mathbf{0}, \sigma^2(\mathbf{P}^{-1}\mathbf{M}\mathbf{P}^{-1} - \mathbf{Q}^{-1})\right], \tag{10.64}$$

where

$$\begin{aligned} \mathbf{P} &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{X}, \\ \mathbf{M} &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{Z}'\mathbf{Z}, \\ \mathbf{Q} &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{X}. \end{aligned}$$

This procedure is known as the Wu test.

Chapter 11

Nonscalar Covariance

11.1 Nature of the Problem

11.1.1 Model and Ideal Conditions

Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{11.1}$$

where $\mathbf{y}$ is $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is the $n \times k$ matrix of observations on the explanatory variables, and $\mathbf{u}$ is the vector of unobservable disturbances.

The ideal conditions are

(i) $\mathrm{E}[\mathbf{u}] = \mathbf{0}$

(ii & iii) $\mathrm{E}[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I}_n$

(iv) $\mathbf{X}$ full column rank

(v) $\mathbf{X}$ nonstochastic

(vi) $[\mathbf{u} \sim N(\mathbf{0},\sigma^2\mathbf{I}_n)]$

11.1.2 Nonscalar Covariance

Nonscalar covariance means that

$$\mathrm{E}[\mathbf{u}\mathbf{u}'] = \sigma^2\Omega, \;\; \mathrm{tr}(\Omega) = n \tag{11.2}$$

an n-by-n positive definite matrix such that $\Omega \neq \mathbf{I}_n$. That is,

$$\mathrm{E}\left[\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} (u_1, u_2, \ldots, u_n)\right] = \sigma^2 \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1n} & \omega_{2n} & \cdots & \omega_{nn} \end{bmatrix} \tag{11.3}$$

A covariance matrix can be nonscalar either by having non-constant diagonal elements or non-zero off diagonal elements or both.

11.1.3 Some Examples

Serial Correlation

Consider the model

$$y_t = \alpha + \beta x_t + u_t, \tag{11.4}$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t, \tag{11.5}$$

and $\mathrm{E}[\varepsilon_t] = 0$, $\mathrm{E}[\varepsilon_t^2] = \sigma^2$, and $\mathrm{E}[\varepsilon_t \varepsilon_s] = 0$ for all $t \neq s$. Here, u_t and u_{t-1} are correlated, so Ω is not diagonal. This is a problem that afflicts a large fraction of time series regressions.

Heteroscedasticity

Consider the model

$$C_i = \alpha + \beta Y_i + u_i \qquad i = 1, 2, \ldots, n, \tag{11.6}$$

where C_i is consumption and Y_i is income for individual i. For a cross-section, we might expect more variation in consumption by high-income individuals. Thus, $\mathrm{E}[u_i^2]$ is not constant. This is a problem that afflicts many cross-sectional regressions.

Systems of Equations

Consider the joint model

$$\begin{aligned} y_{t1} &= x_{t1}'\beta_1 + u_{t1} \\ y_{t2} &= x_{t2}'\beta_2 + u_{t2}. \end{aligned}$$

If u_{t1} and u_{t2} are correlated, then the joint model has a nonscalar covariance. If the error terms u_{t1} and u_{t2} are viewed as omitted variables then it is obvious to ask whether common factors have been omitted and hence the terms are correlated.

11.2 Consequences of Nonscalar Covariance

11.2.1 For Estimation

The OLS estimates are

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \end{aligned} \tag{11.7}$$

Thus,

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\mathbf{u}] = \boldsymbol{\beta}, \tag{11.8}$$

so OLS is still unbiased (but not BLUE since (ii & iii) not satisfied).

Now

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{11.9}$$

so

$$\begin{aligned} \mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\mathbf{u}\mathbf{u}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

The diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ can be either larger or smaller than the corresponding elements of $(\mathbf{X}'\mathbf{X})^{-1}$. In certain cases we will be able to establish the direction of the inequality.

Suppose

$$\begin{aligned} &\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q} \text{ p.d.} \\ &\frac{1}{n}\mathbf{X}'\Omega\mathbf{X} \xrightarrow{p} \mathbf{M} \end{aligned} \tag{11.10}$$

then $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n}(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}'\Omega\mathbf{X}(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1} \xrightarrow{p} \frac{1}{n}\mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}^{-1} \xrightarrow{p} 0$ so

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} \tag{11.11}$$

since $\widehat{\boldsymbol{\beta}}$ unbiased and the variances go to zero.

11.2.2 For Inference

Suppose

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\Omega) \tag{11.12}$$

then

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}). \tag{11.13}$$

Thus

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} \nsim N(0,1) \tag{11.14}$$

since the denominator may be either larger or smaller than $\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$. And

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} \nsim t_{n-k} \tag{11.15}$$

We might say that OLS yields biased and inconsistent estimates of the variance-covariance matrix. This means that our statistics will have incorrect size so we over- or under-reject a correct null hypothesis.

11.2.3 For Prediction

We seek to predict

$$y_* = \mathbf{x}_*'\boldsymbol{\beta} + u_* \tag{11.16}$$

where $*$ indicates an observation outside the sample. The OLS (point) predictor is

$$\widehat{y}_* = \mathbf{x}_*'\widehat{\boldsymbol{\beta}} \tag{11.17}$$

which will be unbiased (but not BLUP). Prediction intervals based on $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ will be either too wide or too narrow so the probablility content will not be the ostenisble value.

11.3 Correcting For Nonscalar Covariance

11.3.1 Generalized Least Squares

Since Ω positive definite we can write

$$\Omega = \mathbf{P}\mathbf{P}' \tag{11.18}$$

for some $n \times n$ nonsingular matrix $\mathbf{P}$ (typically upper or lower triangular). Multiplying (11.1) by $\mathbf{P}^{-1}$ yields

$$\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\mathbf{u} \tag{11.19}$$

or

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^* \tag{11.20}$$

where $\mathbf{y}^* = \mathbf{P}^{-1}\mathbf{y}$, $\mathbf{X}^* = \mathbf{P}^{-1}\mathbf{X}$, and $\mathbf{u}^* = \mathbf{P}^{-1}\mathbf{u}$.

Perform OLS on the transformed model yields the generalized least squares or GLS estimator

$$\begin{aligned}\overline{\boldsymbol{\beta}} &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* \\ &= ((\mathbf{P}^{-1}\mathbf{X})'\mathbf{P}^{-1}\mathbf{X})^{-1}(\mathbf{P}^{-1}\mathbf{X})'\mathbf{P}^{-1}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{P}^{-1'}\mathbf{P}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}^{-1\prime}\mathbf{P}^{-1}\mathbf{y}.\end{aligned}$$

But $\mathbf{P}^{-1'}\mathbf{P}^{-1} = \mathbf{P}'^{-1}\mathbf{P}^{-1} = \Omega^{-1}$ whereupon we have the alternative representation

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}. \tag{11.21}$$

This estimator is also known as the Aitken estimator. Note that GLS reduces to OLS when $\Omega = \mathbf{I}_n$.

11.3.2 Properties with Known Ω

Suppose that Ω is a known, fixed matrix, then

- $\mathrm{E}[\mathbf{u}^*] = \mathbf{0}$
- $\mathrm{E}[\mathbf{u}^*\mathbf{u}^{*\prime}] = \mathbf{P}^{-1}\,\mathrm{E}[\mathbf{u}\mathbf{u}']\mathbf{P}^{-1\prime} = \sigma^2\mathbf{P}^{-1}\Omega\mathbf{P}^{-1\prime} = \sigma^2\mathbf{P}^{-1}\mathbf{P}\mathbf{P}'\mathbf{P}^{-1\prime} = \sigma^2\mathbf{I}_n$
- $\mathbf{X}^* = \mathbf{P}^{-1}\mathbf{X}$ nonstochastic
- $\mathbf{X}^*$ has full column rank

so the transformed model satisfies the ideal model assumptions (i)-(v).

Applying previous results for the ideal case to the transformed model we have

$$\mathrm{E}[\overline{\boldsymbol{\beta}}] = \boldsymbol{\beta} \tag{11.22}$$

$$\mathrm{E}[(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1} \tag{11.23}$$

and the GLS estimator is unbiased and BLUE. We assume the transformed model satisfies the asymptotic properties studied in the previous chapter. First, suppose

$$\frac{1}{n}\mathbf{X}'\Omega^{-1}\mathbf{X} = \frac{1}{n}\mathbf{X}^{*\prime}\mathbf{X}^* \xrightarrow{p} \mathbf{Q}^* \text{ p.d.} \tag{a}$$

then $\overline{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. Secondly, suppose

$$\frac{1}{\sqrt{n}}\mathbf{X}'\Omega^{-1}\mathbf{u} = \frac{1}{\sqrt{n}}\mathbf{X}^{*\prime}\mathbf{u}^* \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}^*) \tag{b}$$

then $\sqrt{n}(\overline{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}^{*-1})$. Inference and prediction can proceed as before for the ideal case.

11.3.3 Properties with Unknown Ω

If Ω is unknown then the obvious approach is to estimate it. Bear in mind, however, that there are up to $n(n+1)/2$ possible different parameters if we have no restrictions on the matrix. Such a matrix cannot be estimated consistently since we only have n observations and the number of parameters is increasing faster than the sample size. Accordingly, we look at cases where $\Omega = \Omega(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ a $p \times 1$ finite-length vector of unknown paramters. The three examples will fall into this category.

Suppose we have an estimator $\widehat{\boldsymbol{\lambda}}$ (possibly consistent) then we obtain $\widehat{\Omega} = \Omega(\widehat{\boldsymbol{\lambda}})$ and the feasible GLS estimator

$$\begin{aligned} \widehat{\overline{\boldsymbol{\beta}}} &= (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{u}. \end{aligned}$$

The small sample properties of this estimator are problematic since $\widehat{\Omega} = \widehat{\mathbf{P}}\widehat{\mathbf{P}}'$ will generally be a function of $\mathbf{u}$ so the regressors of the feasible transformed model $\widehat{\mathbf{X}}^* = \widehat{\mathbf{P}}^{-1}\mathbf{X}$ become stochastic. The feasible GLS will be biased and non-normal in small samples even if the original disturbances were normal.

It might be supposed that if $\widehat{\boldsymbol{\lambda}}$ is consistent that everything will work out in large samples. Such happiness is not assured since there are possibly $n(n+1)/2$ possible nonzero elements in Ω which can interact with the x's in a pathological fashion. Suppose that (a) and (b) are satisfied and furthermore

$$\frac{1}{n}[\mathbf{X}'\Omega(\widehat{\boldsymbol{\lambda}})^{-1}\mathbf{X} - \mathbf{X}'\Omega(\boldsymbol{\lambda})^{-1}\mathbf{X}] \xrightarrow{p} \mathbf{0} \tag{c}$$

and

$$\frac{1}{\sqrt{n}}[\mathbf{X}'\Omega(\widehat{\boldsymbol{\lambda}})^{-1}\mathbf{u} - \mathbf{X}'\Omega(\boldsymbol{\lambda})^{-1}\mathbf{u}] \xrightarrow{p} \mathbf{0} \tag{d}$$

then

$$\sqrt{n}(\widehat{\overline{\boldsymbol{\beta}}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}^{*-1}). \tag{11.24}$$

Thus in large samples, under (a)-(d), the feasible GLS estimator has the same asymptotic distribution as the true GLS. As such it shares the optimality properties of the latter.

11.3.4 Maximum Likelihood Estimation

Suppose

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\Omega) \tag{11.25}$$

then

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\Omega) \tag{11.26}$$

and

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2, \Omega; \mathbf{y}, \mathbf{X}) &= f(\mathbf{y}|\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \Omega) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}\left|\Omega\right|^{1/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\Omega^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \end{aligned}$$

Taking Ω as given, we can maximize $L(\cdot)$ w.r.t. $\boldsymbol{\beta}$ by minimizing

$$\begin{aligned} (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\Omega^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{P}'^{-1}\mathbf{P}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta})'(\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}). \end{aligned} \tag{11.27}$$

Thus OLS on the transformed model or the GLS estimator

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y} \tag{11.28}$$

is MLE and BUE since it is unbiased.

11.4 Seemingly Unrelated Regressions

11.4.1 Sets of Regression Equations

We consider a model with G agents and a behavioral equation with n observations for each agent. The equation for agent j can be written

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{u}_j, \tag{11.29}$$

where $\mathbf{y}_j$ is $n \times 1$ vector of observations on the dependent variable for agent j, $\mathbf{X}_j$ is the $n \times k$ matrix of observations on the explanatory variables, and $\mathbf{u}_j$ is the vector of unobservable disturbances. Writing the G sets of equations as one system yields

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_G \end{pmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_G \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_G \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_G \end{pmatrix} \tag{11.30}$$

or more compactly

$$\overline{\mathbf{y}} = \overline{\mathbf{X}}\boldsymbol{\beta} + \overline{\mathbf{u}} \tag{11.31}$$

where the definitions are obvious.

The individual equations satisfy the usual OLS assumptions

$$\mathrm{E}[\mathbf{u}_j] = \mathbf{0} \tag{11.32}$$

and

$$\mathrm{E}[\mathbf{u}_j \mathbf{u}_j'] = \sigma_j^2 \mathbf{I}_n \tag{11.33}$$

but due to common ommited factors we must allow for the possibility that

$$\mathrm{E}[\mathbf{u}_j \mathbf{u}_\ell'] = \sigma_{j\ell} \mathbf{I}_n \quad j \neq \ell. \tag{11.34}$$

In matrix notation we have

$$\mathrm{E}[\overline{\mathbf{u}}] = \mathbf{0} \tag{11.35}$$

and

$$\mathrm{E}[\overline{\mathbf{u}}\overline{\mathbf{u}}'] = \Sigma \otimes \mathbf{I}_n = \sigma^2 \Omega \tag{11.36}$$

where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1G} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{G1} & \sigma_{G2} & \cdots & \sigma_G^2 \end{bmatrix}. \tag{11.37}$$

11.4.2 SUR Estimation

We can estimate each equation by OLS

$$\widehat{\boldsymbol{\beta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}_j \tag{11.38}$$

and as usual the estimators will be unbiased, BLUE for linearity w.r.t. $\mathbf{y}_j$, and under normality

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}_j, \sigma_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}). \tag{11.39}$$

This procedure, however, ignores the covariances between equations. Treating all equations as a combined system yields

$$\overline{\mathbf{y}} = \overline{\mathbf{X}}\boldsymbol{\beta} + \overline{\mathbf{u}} \tag{11.40}$$

where

$$\overline{\mathbf{u}} \sim (\mathbf{0}, \Sigma \otimes \mathbf{I}_n) \tag{11.41}$$

is non-scalar. Applying GLS to this model yields

$$\begin{aligned} \overline{\boldsymbol{\beta}} &= (\overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{y}} \\ &= (\overline{\mathbf{X}}'(\Sigma^{-1} \otimes \mathbf{I}_n)\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}'(\Sigma^{-1} \otimes \mathbf{I}_n)\overline{\mathbf{y}} \end{aligned}$$

This estimator will be unbiased and BLUE for linearity in $\overline{\mathbf{y}}$ and will, in general, be efficient relative to OLS.

If $\overline{\mathbf{u}}$ is multivariate normal then

$$\overline{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{X}})^{-1}). \tag{11.42}$$

Even if $\overline{\mathbf{u}}$ is not normal then, with reasonable assumptions about the behavior of $\overline{\mathbf{X}}$, we have

$$\sqrt{n}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathbf{d}} \mathbf{N}(\mathbf{0}, [\lim \frac{1}{n}(\overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{X}})]^{-1}). \tag{11.43}$$

11.4.3 Diagonal Σ

There are two special cases in which the SUR esimator simplifies to OLS on each equation. The first case is when Σ is diagonal. In this case

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_G^2 \end{bmatrix} \tag{11.44}$$

and

$$\begin{aligned} \overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{X}} &= \begin{bmatrix} \mathbf{X}_1' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_G' \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2}\mathbf{I}_n & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2}\mathbf{I}_n & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_G^2}\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_G \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma_1^2}\mathbf{X}_1'\mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2}\mathbf{X}_2'\mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_G^2}\mathbf{X}_G'\mathbf{X}_G \end{bmatrix}. \end{aligned}$$

Similarly,

$$\overline{\mathbf{X}}'(\Sigma \otimes \mathbf{I}_n)^{-1}\overline{\mathbf{y}} = \begin{bmatrix} \frac{1}{\sigma_1^2}\mathbf{X}_1'\mathbf{y}_1 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2}\mathbf{X}_2'\mathbf{y}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_G^2}\mathbf{X}_G'\mathbf{y}_G \end{bmatrix} \tag{11.45}$$

whereupon

$$\overline{\boldsymbol{\beta}} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2 \\ \vdots \\ (\mathbf{X}_G'\mathbf{X}_G)^{-1}\mathbf{X}_G'\mathbf{y}_G \end{bmatrix}. \tag{11.46}$$

So the estimator for each equation is just the OLS estimator for that equation alone.

11.4.4 Identical Regressors

The second case is when each equation has the same set of regressor, i.e. $\mathbf{X}_j = \mathbf{X}$ so

$$\overline{\mathbf{X}} = \mathbf{I}_G \otimes \mathbf{X}. \tag{11.47}$$

And

$$\begin{aligned} \overline{\boldsymbol{\beta}} &= [(\mathbf{I}_G \otimes \mathbf{X}')(\Sigma^{-1} \otimes \mathbf{I}_n)(\mathbf{I}_G \otimes \mathbf{X})]^{-1}(\mathbf{I}_G \otimes \mathbf{X}')(\Sigma^{-1} \otimes \mathbf{I}_n)\overline{\mathbf{y}} \\ &= (\Sigma^{-1} \otimes \mathbf{X}'\mathbf{X})^{-1}(\Sigma^{-1} \otimes \mathbf{X}')\overline{\mathbf{y}} \\ &= [\Sigma \otimes (\mathbf{X}'\mathbf{X})^{-1}](\Sigma^{-1} \otimes \mathbf{X}')\overline{\mathbf{y}} \\ &= [\mathbf{I}_G \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\overline{\mathbf{y}} \\ &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_1 \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_2 \\ \vdots \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_G \end{bmatrix}. \end{aligned}$$

In both these cases the other equations have nothing to add to the estimation of the equation of interest because either the omitted factors are unrelated or the equation has no additional regressors to help reduce the sum- of-squared errors for the equation of interest.

11.4.5 Unknown Σ

Note that for this case Σ comprises $\boldsymbol{\lambda}$ in the general form $\Omega = \Omega(\boldsymbol{\lambda})$. It is finite-length with $G(G+1)/2$ unique elements. It can be estimated consistently using

OLS residuals. Let

$$\mathbf{e}_j = \mathbf{y}_j - \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j$$

denote the OLS residuals for agent j. Then by the usual arguments

$$\widehat{\sigma}_{j\ell} = \frac{1}{n} \sum_{i=1}^{n} e_{ij} e_{i\ell}$$

and

$$\widehat{\Sigma} = (\widehat{\sigma}_{j\ell})$$

will be consistent. Form the feasible GLS estimator

$$\widehat{\overline{\boldsymbol{\beta}}} = (\overline{\mathbf{X}}' (\widehat{\Sigma}^{-1} \otimes \mathbf{I}_n) \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}' (\widehat{\Sigma}^{-1} \otimes \mathbf{I}_n) \overline{\mathbf{y}}$$

which can be shown to satisfy (a)-(d) and will have the same asymptotic distribution as $\overline{\boldsymbol{\beta}}$. This estimator will be obtained in two steps: the first step is to estimate all equations by OLS and thereby obtain the estimator $\widehat{\Sigma}$, in the second step we obtain the feasible GLS estimator.