ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

Generalized Method of Moments

Semester 1, 2025

Version: March 20, 2025

Contents

1 Method of Moments

One of the oldest principles of estimation is that of the *method of moments*, abbreviated to MOM on occasion, which was originally introduced by Pearson (1894) as a device for estimating the parameters of the family of distributions that he introduced in the companion paper Pearson (1895), making it one of the oldest estimation methods for point estimation.[1] The idea is as follows. Let $f(\cdot;\theta_1,\ldots,\theta_k)$ be the density of a random

[1]Karl Pearson (1857–1936) was arguably the father of mathematical statistics. He made wide ranging contributions across the discipline, as it was at the time. He was a protégé of Sir Francis Galton (1822–1911) who, in turn, was the more significant half-cousin of Charles Darwin (1809–1882). Pearson had wide ranging interests spanning statistics, mathematics, biology, history, law, and languages. On Karl Pearson's retirement in 1933, University College London took the Department of Eugenics and split it into two, with the Department of Statistics headed by his son Egon Sharpe Pearson (1895–1980), the remainder of the Department of Eugenics stayed with the Galton Chair that was taken up by R. A. Fisher (whom we have encountered elsewhere). Neither Karl Pearson nor Fisher were especially happy with this outcome. For his part, Egon Pearson (1895–1980) went on to become a leading British statistician. He took over editorship of Biometrika from his father in 1936, which he continued with for another 30

variable X which has k parameters $\theta_1, \ldots, \theta_k$. Let the r-th raw moment be denoted $\mu'_r = \mathrm{E}[X^r]$.[2] Note that for $r = 1$, μ'_r is just the population mean μ. In general, μ'_r will be a known function of the k parameters $\theta_1, \ldots, \theta_k$; that is, $\mu'_r(\theta_1, \ldots, \theta_k)$. Let $X_1, \ldots, X_n$ denote a simple random sample of size n from the population and let $M'_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ denote the j-th raw sample moment. Form the k equations

$$M'_j = \mu'_j(\theta_1, \ldots, \theta_k), \quad j = 1, \ldots, k,$$

in the k variables $\theta_1, \ldots, \theta_k$ and, assuming that there exists a unique solution to the system of equations, denote it by $(\tilde{\Theta}_1, \ldots, \tilde{\Theta}_k)$. The estimator $(\tilde{\Theta}_1, \ldots, \tilde{\Theta}_k)$, where $\tilde{\Theta}_j$ estimates θ_j, is the *method of moments* estimator of $\theta_1, \ldots, \theta_k$. In essence, the parameters $\theta_1, \ldots, \theta_k$ are estimated by functions of the raw sample moments.

Example 1. MOM in the Uniform Distribution.
A simple random sample of n observations $Y_1, \ldots, Y_n$, is selected from a population in which Y_i, for $i = 1, \ldots, n$, possesses a uniform probability density function over the interval $(0, \theta)$, where θ is unknown. Use the method of moments to estimate the parameter θ.

Solution:
For this uniform distribution,

$$\mathrm{E}[Y] = \int_0^\theta y\theta^{-1}\,\mathrm{d}y = \left[\frac{y^2}{2\theta}\right]_0^\theta = \frac{\theta^2}{2\theta} - \frac{0^2}{2\theta} = \frac{\theta}{2} = \mu'_1(\theta) = \mu(\theta).$$

That is, the mean of a uniform distribution is given by the mid-point of the support of the random variable. As there is only one parameter here we only need a single moment condition to find a method of moments estimator. The corresponding first sample moment is

$$m'_1 = \frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y}.$$

Equating the population and sample moments, and using $\tilde{\theta}$ to denote that value of θ for which equality holds, yields

$$\mathrm{E}[Y] = \mu\left(\tilde{\theta}\right) = m'_1 \implies \frac{\tilde{\theta}}{2} = \overline{Y} \implies \tilde{\theta} = 2\overline{Y}. \qquad \square$$

Example 2. MOM in the Poisson Distribution.
Let $X_1, \ldots, X_n$ be a simple random sample from a Poisson distribution with parameter λ. Similar to the previous example, this is a single parameter problem and so the underlying approach to MOM estimation of λ will be much the same as that followed previously. First, we need to resolve

$$\mathrm{E}[X] = \sum_{x=0}^\infty \frac{xe^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^\infty \frac{xe^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^\infty \frac{e^{-\lambda}\lambda^x}{(x-1)!} = \lambda\sum_{x=1}^\infty \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}$$

years. He was the recipient of numerous honours, including Fellowship of the Econometric Society in 1948. He is, perhaps, best known for the Neyman-Pearson Lemma.

[2] Because μ'_1 is just the population mean, we typically drop the special notation and just write μ rather than μ'_1.

Now make the substitution $y = x - 1$

$$\mathrm{E}[X] = \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} = \lambda,$$

because the sum of Poisson probabilities is unity. Second, we need to form a moment condition by equating this population moment, expressed as a function of the parameter of the distribution, to the corresponding raw sample moment and then solve out for the estimator. The relevant moment condition is

$$\mathrm{E}[X] = \mu\left(\tilde{\lambda}\right) = m_1' \implies \tilde{\lambda} = \overline{X}.$$

That is, the MOM estimator for the mean of a Poisson distribution is simply the sample mean. □

Example 3. MOM in the Negative Exponential Distribution.
Let $X_1, \ldots, X_n$ be a simple random sample from a negative exponential distribution with parameter θ. This is yet another single parameter problem and so the underlying approach to MOM estimation is as in the previous examples followed previously. The relevant density function is $f(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$, $\theta > 0$, for which

$$\mathrm{E}[X] = \int_0^{\infty} x\theta e^{-\theta x}\,\mathrm{d}x = \theta \int_0^{\infty} e^{-\theta x} x\,\mathrm{d}x = \theta \times \Gamma(2)\theta^{-2} = \frac{1}{\theta}.$$

Then the relevant moment condition is

$$\mu\left(\tilde{\theta}\right) = m_1' \implies \frac{1}{\tilde{\theta}} = \overline{X}_n.$$

That is, the MOM estimator for the mean of a negative exponential distribution is simply the reciprocal of sample mean; that is, $\tilde{\theta} = 1/\overline{X}_n$.

Example 4. MOM in the Normal Distribution.
Let $X_1, \ldots, X_n$ be a simple random sample from a Normal distribution with mean μ and variance σ^2. Let $\theta = (\theta_1, \theta_2) = (\mu, \sigma)$. Estimate θ by the method of moments.

Solution:
Recall that $\sigma^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mu_2' - \mu^2$. Equating these population moments to the corresponding sample moments we obtain

$$\begin{aligned} M_1' &= \mu(\tilde{\theta}) = \tilde{\mu}, \\ M_2' &= \mu_2'(\tilde{\theta}) = (\tilde{\sigma})^2 + (\tilde{\mu})^2. \end{aligned}$$

Solving these equations for $\tilde{\theta}$, we see that $\tilde{\mu} = M_1' = \overline{X}_n$ and

$$\tilde{\sigma} = \sqrt{M_2' - (M_1')^2} = \sqrt{n^{-1}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}_n^2} = \sqrt{n^{-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right)}$$
$$= \sqrt{n^{-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2},$$

which we note is not simply $\sqrt{s^2}$, where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ denotes the usual unbiased estimator for a population variance given a simple random sample. Consequently, the method of moments estimator must be a biased estimator. □

Example 5. MOM in a Two Parameter Uniform Distribution.
Let $X_1, \ldots, X_n$ be a simple random sample from a uniform distribution on

$$\left[\mu - \sqrt{2\sigma}\,, \mu + \sqrt{2\sigma}\right]$$

Find the method of moments estimators for the mean μ and standard deviation σ.

Solution:
The moment conditions are as in the previous example, namely

$$\begin{aligned} M_1' &= \mu(\tilde{\theta}) = \tilde{\mu}, \\ M_2' &= \mu_2'(\tilde{\theta}) = (\tilde{\sigma})^2 + (\tilde{\mu})^2\,, \end{aligned}$$

and the solutions to these equations are also identical to those of the previous example: $\tilde{\mu} = \overline{X}_n$ and

$$\tilde{\sigma} = \sqrt{n^{-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}\,.$$

Even though the population distributions are quite different the estimators remain unchanged. It seems that we should be able to do better if we could take these differences into account and, not surprisingly, we can but that remains a story for another time (when we discuss Maximum Likelihood). □

Example 6. MOM and Regression.
If we were pursuing a least squares approach to, say, a simple linear regression model then we would specify the data at hand, being a set of n observations on each of the dependent variable y, an n-vector, and a set of k explanatory variables X, an $n \times k$ matrix. Then, as a linear model is the object of our ambition, we would choose the fitted regression coefficients to be that k-vector β that minimizes the least squares criterion

$$S(\beta) = (y - X\beta)'(y - X\beta).$$

That is, $\hat{\beta} = \operatorname{argmin}_\beta S(\beta)$. To take this sort of idea into a method of moments framework, we need a population moment condition and, in this context, something that mirrors the development of least squares is the assumption $\mathrm{E}\left[u \mid X\right] = 0$ which implies that $\mathrm{E}\left[X'u\right] = 0$ because, by the Law of Iterated Expectations,

$$\mathrm{E}\left[X'u\right] = \mathrm{E}_X\left[\mathrm{E}\left[X'u \mid X\right]\right] = \mathrm{E}_X\left[X'\,\mathrm{E}\left[u \mid X\right]\right] = \mathrm{E}_X\left[X'0\right] = 0.$$

If our moment condition is $\mathrm{E}\left[X'u\right] = 0$ then the corresponding sample moment is

$$\frac{1}{n} X'(y - X\tilde{\beta}) = 0.$$

This is a system of k equations in the k unknowns that are the elements of β and, under the usual assumptions about X having full column rank, we know that there exists a unique solution of the form

$$\tilde{\beta} = (X'X)^{-1}X'y$$

which, as it happens, corresponds to the least squares estimator in this case, although it does not take much for the MOM and least squares estimators to differ. □

Remark 1 (Implicit MOM Optimization Criteria). One feature to take from the previous example is that, corresponding to the least squares normal equation, or first order condition, is a criterion that is optimized, $S(\beta)$ in this case. Although no explicit statement of such was made in the case of the MOM estimator, implicitly there must be some function that can be thought of as such. This will prove important in our later analysis of the generalized method of moments estimator. □

We have seen many examples of method of moments estimation. It has held a significant position in estimation theory for a couple of reasons. First, it is attractive inasmuch as we know that we can estimate population moments consistently using raw sample moments and so, given that our moment conditions result in method of moments estimators that are continuous functions of the sample moments we should be able to bring our machinery from asymptotic theory to bear on MOM problems in a fairly direct way. However, there are problems that need to be addressed. For example, we have assumed that we have as many moment conditions as we have parameters to estimate and that these (probably non-linear) equations will have a unique solution. This amounts to an assumption of identification which is not guaranteed to be true in real world problems. Indeed, it is not even guaranteed that our population will possess all the moments required to satisfy this assumption, although it probably will.

Second, the assumption that our solutions are necessarily continuous functions of the sample moments is not necessarily justified.

Third, in the event that we have more moment conditions than we need, the question arises as to which moments you should be matching to their population counterparts. In practice, we always start at the lowest order moment available to us, the first, and work up. In principle there is nothing in the MOM framework per se that requires this. However, in practice, higher order moments are expensive to estimate. By this is meant that the higher the order moment that you are trying to estimate the greater the number of tail observations you require in order to do this well. But observations in the tails of a distribution tend to be relatively rare events and so, typically, one requires extremely large sample sizes in order to do this well. By way of intuition on this, in order to estimate a mean you only need a single observation whereas you cannot sensibly calculate a variance with fewer than two observations (double the requirement!) Higher order moments are even more greedy for data.

Finally, all of our discussion so far has been predicated on the belief that we only want to use as many moments as we have parameters to estimate. However, that is not necessarily true. Many models in economics and finance generate many more moment conditions than there are parameters to be estimated and it would be wrong, or at least inefficient, to simply discard some subset of the available moment conditions. At the very least one might expect your estimates to be sensitive to the choice of which moments to match. This problem led to the development of what we in economics call the generalized method of moments (GMM). Similar ideas abound in statistics where you might encounter the notion of estimating equations. These in turn can be thought of as a class of extremum or M-estimators (which is short for maximum likelihood-like estimators). All of these ideas are closely related and we simply don't have time to explore them all. Nevertheless, you should be aware that there is much more out there than we have time to cover.

2 Generalized Method of Moments

Although very closely related to the method of moments approach, GMM is sufficiently important in econometrics that it merits its own section. In what follows we shall restrict attention to the case of regression models and much of what we do in this section can be thought of as an extension to both the method of moments, as discussed in the previous section, and our treatment of regression to date.

2.1 GMM for Linear Regression Models

2.1.1 The Basic Setup

Consider the linear regression model

$$y = X\beta + u, \quad \mathrm{E}\left[uu'\right] = \Omega, \tag{1}$$

where there are n observations on the dependent variable y and on the k regressors X. We will assume that X has full column rank but will allow for the possibility that at least some of the explanatory variables may not be pre-determined with respect to the disturbance terms u, so that the assumption $\mathrm{E}\left[u \mid X\right] = 0$ is no longer valid. We shall partition $X = [X_1, X_2]$, where X_1 is comprised of k_1 columns of X that are pre-determined with respect to u and X_2 contains the remaining $k_2 = k - k_1$ columns of X. We will assume that there exists a matrix Z of dimension $n \times p$, $p \geq k_2$, such that $W = [X_1, Z]$ is a matrix of instrumental variables of dimension $n \times (k_1 + p)$, where $n > k_1 + p$, and of full column rank satisfying the condition $\mathrm{E}\left[u_j \mid W_j\right] = 0$, $j = 1, \ldots, n$, where W_j' denotes the j-th row of W. We will further assume that $\mathrm{E}\left[u_i u_j \mid W_i, W_j\right] = \omega_{ij}$, where ω_{ij} denotes the ij-th element of Ω. Given these assumptions we see that

$$\begin{aligned} \mathrm{E}\left[W'(y - X\beta)\right] &= \mathrm{E}\left[W'u\right] = \mathrm{E}_W\left[\mathrm{E}\left[W'u \mid W\right]\right] = \mathrm{E}_W\left[W'\,\mathrm{E}\left[u \mid W\right]\right] = \mathrm{E}_W\left[W'0\right] \\ &= 0 \end{aligned} \tag{2}$$

and

$$\begin{aligned} \mathrm{Var}\left[W'u\right] &= \mathrm{E}\left[W'uu'W\right] = \mathrm{E}_W\left[\mathrm{E}\left[W'uu'W \mid W\right]\right] = \mathrm{E}_W\left[W'\,\mathrm{E}\left[uu' \mid W\right]W\right] \\ &= \mathrm{E}_W\left[W'\Omega W\right]. \end{aligned} \tag{3}$$

The equation specified in (2) is actually a set of $k_1 + p$ equations (or *moment conditions*). Because the expectations of these inner products have zero expectation they are often referred to as *orthogonality conditions*.[3] The set of empirical moment conditions corresponding to (2) is[4]

$$\frac{1}{n}W'(y - X\tilde{\beta}_n) = 0. \tag{4}$$

If it is the case that $p = k_2$ then our model is said to be *exactly identified*, by which we mean that we have exactly the same number of moment conditions as we have parameters

[3] If A is an $m \times n$ and B is an $m \times p$ matrix, the inner product of A and B is defined to be $A'B$. If $n = p = 1$ then we have the inner product of two vectors, which is a scalar quantity, and so the order of A and B does not matter; we see that $A'B = B'A$. This is not true in general because $A'B$ is an $n \times p$ matrix whereas $B'A$ is $p \times n$ matrix and so $A'B \neq B'A$. That is, matrix multiplication is not, in general, commutative.

[4] Observe that we will assume that the scaling by n^{-1} is sufficient to control asymptotic behaviour but, as noted in the Asymptotics handout, there is scope to play with this if necessary.

to estimate, and we are able to solve (4) in exactly the same way as we did in Example 6. Specifically, we see that

$$\tilde{\beta}_n = (W'X)^{-1}W'y, \tag{5}$$

which is the formula for the simplest of *instrumental variables* (IV) estimators. Indeed, $\tilde{\beta}_n$ is sometimes referred to as the *simple instrumental variables estimator*. Based on what we learned in our treatment of the asymptotics of regression models we can see that a full analysis is going to require assumptions along the lines of

1. $\operatorname{plim}_{n\to\infty} n^{-1}W'X = Q_{W'X}$, a non-singular matrix (which may not be symmetric).
2. $\operatorname{plim}_{n\to\infty} n^{-1}W'\Omega W = Q_{W'\Omega W} > 0$, a positive definite matrix.

Coupled with the assumptions already made that will allow us to show the consistency of the estimator and that

$$n^{1/2}(\tilde{\beta}_n - \beta) \xrightarrow{d} \mathrm{N}\left(0, Q_{W'X}^{-1}Q_{W'\Omega W}(Q_{W'X}^{-1})'\right),$$

which yields an asymptotic distribution

$$\tilde{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, n^{-1}Q_{W'X}^{-1}Q_{W'\Omega W}(Q_{W'X}^{-1})'\right).$$

Note the sandwich form of the covariance structure of this distribution, which is typically associated with inefficient estimators that have not taken proper account of Ω in their construction. Obviously, if Ω is unknown then we are going to have to extend our assumptions if we wish to say too much about its estimation or, rather, the estimation of $Q_{W\Omega W}$, which is really what we need.

In the event that $p > k_2$, so that the model is *over-identified*, by which we mean that you have more equations to solve than parameters, then we have a problem because $W'X$ is no longer square and so we need to modify our approach. Nevertheless, we can write the empirical moment condition in the form of a normal equation

$$W'X\tilde{\beta}_n = W'y \tag{6}$$

As discussed in Section 5.3 of the Matrices Handout, provided $W'X$ has full column rank, which it will given our assumptions about X and W each having full column rank, then there exists a left inverse of $W'X$, L say, of dimension $k \times (k_1 + p)$ such that $LW'X = I_k$. Consequently, pre-multiplying both sides of (6) by L yields

$$LW'X\tilde{\beta}_n = LW'y \tag{7}$$

which implies that

$$\tilde{\beta}_n = LW'y = LW'X\beta + LW'u = \beta + LW'u, \tag{8}$$

as $LW'X = I_k$. Under suitable regularity conditions we see that

$$\tilde{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, n^{-1}Q_{LW'\Omega WL'}\right), \tag{9}$$

where $\operatorname{plim}(n^{-1}LW'\Omega WL') = Q_{LW'\Omega WL'} > 0$.

It is worth thinking about the nature of L and some of the choices available. One possibility is to seek a left inverse for $W'X$. As per the discussion in Section 5.3 of the

Matrices Handout, an obvious form of left inverse is $L = (G'G)^{-1}G'$, where $G = W'X$ is of dimension $(k_1 + p) \times k$. Then

$$\begin{aligned} Q_{LW'\Omega WL'} &= \operatorname{plim}\left[(n^{-1}X'WW'X)^{-1}(n^{-1}X'WW'\Omega WW'X)(n^{-1}X'WW'X)^{-1}\right] \\ &= (\operatorname{plim} n^{-1}X'WW'X)^{-1}(\operatorname{plim} n^{-1}X'WW'\Omega WW'X)(\operatorname{plim} n^{-1}X'WW'X)^{-1}. \end{aligned}$$

Note that there is no point trying to simplify this expression because $W'X$ is rectangular rather than square and so individual terms can't move through the matrix inversions. One consequence of this is that the nature of our assumptions change as assumptions about $\operatorname{plim}(n^{-1}W'X)$ and $\operatorname{plim}(n^{-1}W'\Omega W)$ are no longer sufficient for our purposes. Nevertheless, we still have the sandwich form of covariance estimator that is indicative of an inefficient estimator. So maybe that wasn't such a great choice after all.

An alternative choice for L is based on the recognition that we don't actually require L to be a left inverse for $W'X$. Indeed, all that we really need is that $LW'X$ be a non-singular matrix, because then we can solve (7) to yield

$$\tilde{\beta}_n = (LW'X)^{-1}LW'y, \tag{10}$$

which will have as its asymptotic covariance matrix, assuming that we choose L so that $LW'X$ and hence $(LW'X)^{-1}$ is symmetric,

$$(LW'X)^{-1}LW'\Omega WL'(LW'X)^{-1}. \tag{11}$$

The desire for symmetry suggests that we choose $L = X'W\Lambda$, where Λ is yet to be determined but is symmetric and of full rank, so that $X'W\Lambda W'X$ is symmetric and non-singular. Given a choice of this form our estimator becomes

$$\tilde{\beta}_n = (X'W\Lambda W'X)^{-1}X'W\Lambda W'y,$$

with asymptotic covariance matrix

$$(X'W\Lambda W'X)^{-1}X'W\Lambda W'\Omega W\Lambda W'X(X'W\Lambda W'X)^{-1}.$$

If we think back to the discussion of GLS then an obvious choice is $L = X'W(W'\Omega W)^{-1}$.[5] The advantage of this choice for L is that the covariance matrix in the limiting distribution reduces

$$\operatorname{plim} n^{-1}X'W(W'\Omega W)^{-1}W'X.$$

We see that the sandwich structure is no longer there, as $W'X$ is rectangular rather than non-singular, which is suggestive of a more efficient estimator and that we are back to the set of assumptions 1 and 2, as all that we need to control the asymptotic behaviour of $\tilde{\beta}$. For completeness, we note that this efficient GMM estimator is given by

$$\tilde{\beta}_n^{GMM} = (X'W(W'\Omega W)^{-1}W'X)^{-1}X'W(W'\Omega W)^{-1}W'y. \tag{12}$$

[5] An alternative path of reasoning here is the following. Life would simplify a lot of

$$X'W\Lambda W'X = X'W\Lambda W'\Omega W\Lambda W'X,$$

as there would be substantial cancellation and simplification of the expression for the asymptotic variance. This, in turn, would require that $\Lambda = \Lambda W'\Omega W\Lambda \implies I_n = W'\Omega W\Lambda \implies \Lambda = (W'\Omega W)^{-1}$, which is the choice that we have made. As promised, it leads to substantial cancellation and simplification of the expression for the asymptotic variance.

As we showed in Example 6, the least squares approach started with the criterion function $S(\beta)$, which was reduced to a set of normal equations (or first-order conditions), that were then solved to obtained the least squares estimator. The method of moments approach started directly at the level of the first-order conditions that were based on corresponding population moment conditions. Nevertheless, implicit in what we have been doing is a criterion function that is minimized by the resulting GMM estimator. If

$$LW'(y - X\beta) = 0 \implies -2X'W(W'\Omega W)^{-1}W'(y - X\beta) = 0, \tag{13}$$

on substituting for L and scaling by -2 for reasons that will become clear in a moment, then the implicit GMM criterion being minimized is

$$S_n^{GMM}(\beta) = (y - X\beta)'W(W'\Omega W)^{-1}W'(y - X\beta). \tag{14}$$

To see this note that, via a multivariate chain rule,

$$\begin{aligned}\frac{\partial(y - X\beta)'A(y - X\beta)}{\partial\beta} &= \frac{\partial(y - X\beta)'}{\partial\beta}\frac{\partial(y - X\beta)'A(y - X\beta)}{\partial(y - X\beta)} \\ &= -X' \times 2A(y - X\beta) = -2X'A(y - X\beta),\end{aligned} \tag{15}$$

where A is any symmetric matrix not depending on β.[6] We see that the analogue in (13) of the A matrix in (15) is the matrix $W(W'\Omega W)^{-1}W'$, hence (14). We also note that, in this case, the first-order condition from which the estimator is derived becomes

$$X'W(W'\Omega W)^{-1}W'(y - X\beta) = 0.$$

We started this section allowing for regressors that were not predetermined and for disturbances with a non-scalar covariance matrix. As we finish off this section let us consider what happens to GMM when we starting imposing our usual assumptions again. To begin, suppose that all of X could be treated as predetermined, so that we might choose $W = X$, then $L = X'X(X'\Omega X)^{-1}$ and it is easy to show that $\tilde{\beta}_n^{GMM}$ reduces to the OLS estimator $\hat{\beta}_n = (X'X)^{-1}X'y$. We know $\hat{\beta}_n$ to be consistent in this model but it differs from $\check{\beta}_n = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, the GLS estimator, which we know to be fully efficient when Ω is known. A comparison of the formula for the GLS estimator with (5) suggests that we should be able to treat $\check{\beta}_n$ as an IV estimator where the instruments are chosen to be the columns of $\Omega^{-1}X$, which is true. However, the risk here is that one might be tempted to further conclude that if some of the columns of X were not predetermined then we might construct instruments of the form $\Omega^{-1}W$ in the quest for both consistency and efficiency. The problem that may arise is that if Ω is other than diagonal then the set of instruments W that satisfy the moment condition $\mathrm{E}\left[W'\Omega^{-1}(y - X\beta)\right]$ may well be different to those that might satisfy $\mathrm{E}\left[W'(y - X\beta)\right]$. Moreover, there is very little guidance here that can be offered because finding a suitable set of instruments with non-scalar covariance matrix is something that must be done on a case by case basis and will be contingent upon the actual structure of Ω. An interesting discussion of the issues involved is given in Davidson and MacKinnon (2004, Section 9.2). By way of example of when such failings may occur, if there is serial correlation in the disturbances (meaning non-zero off-diagonal elements in Ω) and lagged variables are used as instruments a non-zero expectation may be the outcome. We will not pursue this particular matter further.

[6] If A is asymmetric then we would have $\dfrac{\partial(y - X\beta)'A(y - X\beta)}{\partial(y - X\beta)} = (A + A')x$.

It is important that you are aware of the problem. That said, situations where Ω is known, so that this might be a real world concern, are essentially non-existent, so we shall move on.

In the event that $\Omega = \sigma^2 I_n$, but we still may have endogenous regressors, the GMM criterion divided by σ^2 reduces to

$$S_n^{IV}(\beta) = (y - X\beta)' P_W (y - X\beta).$$

We note that the minimizing value of β is completely unaffected by the presence of the scaling factor σ^2 because it is a positive scalar that is not a function of β. Moreover, the resulting estimator is simply

$$\hat{\beta}_n^{GIVE} = (X' P_W X)^{-1} X' P_W y, \tag{16}$$

which is the *generalized instrumental variable estimator* (GIVE). Note that in an exactly identified model, where W has the same number of columns as X and is also of full column rank then $\hat{\beta}_n^{GIVE}$ reduces to (5). We note that the covariance matrix of the asymptotic distribution of this estimator is simply

$$\sigma^2 (X' P_W X)^{-1}, \tag{17}$$

where clearly an estimator of σ^2 will be required in order to make this feasible. It is the issue of feasibility that we turn to next. As we finish this section we note that the GMM framework incorporates both OLS and IV estimation as special cases, allowing greater generality than either. It has a less comfortable relationship with GLS, although we note that GLS has no way of making any allowance for the possibility of explanatory variables that are other than pre-determined.

2.2 Feasible GMM for Linear Regression Models

As we saw with GLS estimators, the assumption that Ω is known is typically fanciful. Although it was possible to postulate parametric models that might generate a covariance model of the form Ω, what proved perhaps more satisfying was the notion that we could estimate $\operatorname{plim} n^{-1} X' \Omega X$ if we could find a consistent estimator for the regression coefficients so that we might obtain 'consistent' residuals (whatever 'consistent' might mean in this context). In the case of GLS, OLS met our needs because there endogenous regressors were not a problem. However, OLS will clearly fail if we have endogenous regressors as well as a non-scalar covariance structure. However, the same role can be filled by various IV estimators, which will be consistent in the presence of endogenous regressors. Then we can use either White's heteroskedastic robust estimator for the heteroskedasticity of an unknown form or, if correlation between observations is deemed an issue, then use a HAC estimator, as appropriate. We will not pursue this further because the ideas are identical to those that we have encountered previously and so it would reduce to mathematics for mathematics sake alone.

2.2.1 IV Estimators as Two-Step Estimators

It is often the case that people are taught that IV estimators can be thought of as two-step estimators. We posit the existence of extra instruments Z such that it is possible to fit a model of the form

$$X_2 = X_1 \theta_1 + Z\theta_2 + \text{error},$$

which is estimated by OLS. This is the first step. The fitted values $\widehat{X}_2 = X_1\hat{\theta}_1 + Z\hat{\theta}_2$ are then used in the equation of interest

$$y = X_1\beta_1 + X_2\beta_2 + u$$

in place of the X_2 which are correlated with the disturbances u. Thus, we estimate the model

$$y = X_1\beta_1 + \widehat{X}_2\beta_2 + \text{error},$$

again by OLS, this being the second step. It can be shown that this two-step procedure yields identical estimates of β_1 and β_2 as does the IV estimator using X_1 and Z as instruments. The argument, as commonly put, is that the first step, based on regressors X_1 and Z that are (asymptotically) uncorrelated with u, yields value of $\widehat{X}_2$ that are also (asymptotically) uncorrelated with u, thereby 'purging' the equation of interest of the endogeneity problems that plagued it, making OLS the appropriate estimator in the second step.

This is one of those exercises that allow us to use an OLS program to do our calculations for us when we don't otherwise have an appropriate bit of software to do the job for us. That it is not helpful to think of an IV estimator as a two-step estimator, in the way that a feasible GLS estimator might be, can be seen in the fact that although the second step may give us appropriate estimates of the regression coefficients β_1 and β_2, nothing else that it gives us is appropriate. In particular, the estimated standard errors and t statistics that will accompany the standard output of OLS applied to the second step regression are all invalid for inference about the equation of interest. This is because they will be based on the sum of squared residuals from the second stage regression:

$$\hat{e}'\hat{e} = (y - X_1\hat{\beta}_1 - \widehat{X}_2\hat{\beta}_2)'(y - X_1\hat{\beta}_1 - \widehat{X}_2\hat{\beta}_2)$$

It can be shown that, as well as being a biased estimator of the conditional variance of u, $\hat{\sigma}_2 = n^{-1}\hat{e}'\hat{e}$ is an inconsistent estimator of this variance as well. The reason for this is the use of $\widehat{X}_2$ in the construction of the residual vector. In order to obtain a consistent variance estimator, the real X_2 must be used. That is, variance estimation should be based on the residual sum of squares

$$e'e = (y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2)'(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2).$$

Of course, such errors would not be made if one thought of the IV estimator as a single step estimator, because then there would be no scope for thinking of using the second step equation as the basis for variance estimation.

2.2.2 Instrumental Variables Estimators in Simultaneous Equations Models

Consider the problem of estimating the single structural equation model

$$y = Y\beta + Z_1\gamma + u, \tag{18}$$

where Y is a set of m endogenous regressors, Z_1 is a set of k_1 pre-determined regressors and β and γ are vectors of parameters to be estimated. Suppose that the corresponding reduced form for this model is

$$[y, Y] = Z\Pi + [v, V] = [Z_1, Z_2]\begin{bmatrix}\pi_1 & \Pi_1 \\ \pi_2 & \Pi_2\end{bmatrix} + [v, V] \tag{19}$$

where the full column rank matrix $Z = [Z_1, Z_2]$ is a set of $k = k_1 + k_2$ exogenous (or pre-determined) regressors, where $k_2 \geq m$, which is the *order condition for identification*. In order for the structural and reduced form models to be compatible it must be that their coefficients are related by the compatibility conditions

$$\begin{bmatrix} \pi_1 & \Pi_1 \\ \pi_2 & \Pi_2 \end{bmatrix} \begin{bmatrix} 1 \\ -\beta \end{bmatrix} = \begin{bmatrix} \gamma \\ 0 \end{bmatrix} \implies \begin{array}{l} \pi_1 = \Pi_1\beta + \gamma, \\ \pi_2 = \Pi_2\beta. \end{array} \tag{20}$$

Finally the structural disturbance u must be related to the reduced form disturbances according to $u = [v, V][1, -\beta']'$. If we let $\mathrm{Var}\,[u \mid Z] = \sigma_u^2$ and[7]

$$\mathrm{Var}\,[v, V] = \Omega \otimes I_n, \qquad \Omega = \begin{bmatrix} \omega_{11} & \omega_{21}' \\ \omega_{21} & \Omega_{22} \end{bmatrix},$$

so that the rows of $[v, V]$ have common covariance matrix Ω but are independent across observations then the final compatibility condition between the two models is that

$$\sigma_u^2 = [1, -\beta']\Omega[1, -\beta']' = \omega_{11} - 2\omega_{21}'\beta + \beta'\Omega_{22}\beta. \tag{21}$$

In the classical simultaneous equations model, the vector given by $\mathrm{vec}\,[v, V]$ has elements that are assumed to jointly have a multivariate Normal distribution with mean zero given Z, so that $\mathrm{E}\,[\mathrm{vec}\,[v, V] \mid Z] = 0$ and variance as described above.[8] Consequently, conditional on Z, u also follows a Normal distribution with mean zero and conditional variance $\sigma_u^2 I_n$.

There are a few points worth noting about this model. First, the covariance between Y and u is given by

$$\mathrm{E}\,[u\,\mathrm{vec}(Y - Z\Pi)' \mid Z] = \mathrm{E}\,[(v - V\beta)\,\mathrm{vec}(V)' \mid Z] = (\omega_{21} - \Omega_{22}\beta)' \otimes I_n.$$

We see at once that this covariance reduces to zero if $\omega_{21} - \Omega_{22}\beta = 0 \implies \beta = \Omega_{22}^{-1}\omega_{21}$, which is indeed the condition for unbiasedness of OLS in the structural model.

Second, (19) is just a classical linear multivariate regression model (all the regressors are pre-determined) and so we can consistently estimate the parameters Π, call our estimates $\widehat{\Pi}$ say, using the multivariate analog of ordinary least squares, which is $\widehat{\Pi} = (Z'Z)^{-1}Z'[y, Y]$. Consequently, we can obtain consistent estimates of β and γ as functions of the elements of $\widehat{\Pi}$ if we can solve (20), which we can always do uniquely provided that Π_2 has full column rank. This is the *rank condition for identification*.

Finally, suppose that we choose our set of instruments from those available in Z_2. One special case is that of the two-stage least squares (2SLS) estimator, which uses all of Z_2 as instruments so that $W = Z$. From (16) we see that, on writing $X = [Y, Z_1]$ and $\theta = [\beta, \gamma]'$,

$$\hat{\theta}_n^{2SLS} = (X'P_Z X)^{-1}X'P_Z y. \tag{22}$$

Observe that $P_Z X = [P_Z Y, Z_1] = [\widehat{Y}, Z_1] = \ddot{X}$ say, as $P_Z Z_1 = Z_1$ and $\widehat{Y} = P_Z Y$. We shall also write $\hat{y} = P_Z y$. Consequently, we can write

$$\hat{\theta}_n^{2SLS} = (\ddot{X}'\ddot{X})^{-1}\ddot{X}'\hat{y}.$$

[7] A derivations of $\mathrm{Var}\,[v, V]$ is provided in the appendix.

[8] The vec operator is described in Section 8 of the matrices handout. In essence, the vec of a matrix A, written $\mathrm{vec}(A)$, is the vector obtained by stacking the columns of A one atop the other, with the first column at the top of the vector and the last column at the bottom.

From Equation (4.7.5) of the Matrices handout, we can write

$$\hat{\beta}_n^{2SLS} = (\widehat{Y}' M_{Z_1} \widehat{Y})^{-1} \widehat{Y}' M_{Z_1} \hat{y} = (Y' P_Z M_{Z_1} P_Z Y)^{-1} Y' P_Z M_{Z_1} P_Z y \tag{23}$$

and

$$\hat{\gamma}_n^{2SLS} = (Z_1' M_{\widehat{Y}} Z_1)^{-1} Z_1' M_{\widehat{Y}} P_Z y. \tag{24}$$

Considerable simplification can be had on recognizing that

$$P_Z M_{Z_1} P_Z = P_Z (I - P_{Z_1}) P_Z = P_Z - P_Z Z_1 (Z_1' Z_1)^{-1} Z_1' P_Z = P_Z - P_{Z_1},$$

where we have again used the result $P_Z Z_1 = Z_1$. Therefore,

$$\hat{\beta}_n^{2SLS} = (Y'(P_Z - P_{Z_1})Y)^{-1} Y'(P_Z - P_{Z_1}) y \tag{25}$$

An alternative representation for $\hat{\beta}_n^{2SLS}$ is immediately available on applying Equation (4.7.11a) of the Matrices handout to (25). Doing so yields

$$\hat{\beta}_n^{2SLS} = (Y' P_{M_{Z_1} Z_2} Y)^{-1} Y' P_{M_{Z_1} Z_2} y, \tag{26}$$

While (25) is certainly easier to read than is (26), this latter expression makes clear that $P_Z - P_{Z_1}$ is a symmetric, idempotent matrix, which helps simplify any subsequent distributional analysis.[9]

Simplifying $\hat{\gamma}_n^{2SLS}$ is our next task. It is fair to say that that is a bit of a grind to do so directly and so I shall spare you the details. However, if we think of the minimization process as a two-step process then the first step has given $\hat{\beta}_n^{2SLS}$ and in the second step we proceed conditional on our expression for $\hat{\beta}_n^{2SLS}$which allows us to minimize

$$S_n^{2SLS}(\gamma) = (y - Y\hat{\beta}_n^{2SLS} - Z_1\gamma)' P_Z (y - Y\hat{\beta}_n^{2SLS} - Z_1\gamma)$$

with respect to γ. We have seen that the solution to this problem will take the form

$$\hat{\gamma}_n^{2SLS} = (Z_1' P_Z Z_1)^{-1} Z_1' P_Z (y - Y\hat{\beta}_n^{2SLS}) = (Z_1' Z_1)^{-1} Z_1' (y - Y\hat{\beta}_n^{2SLS}),$$

where, clearly, we have treated $y - Y\hat{\beta}_n^{2SLS}$ as the dependent variable in the minimization process and where, again, we have used the result $P_Z Z_1 = Z_1$. It is because of this latter representation that interest in $\hat{\gamma}_n^{2SLS}$ has always come a distant second to the interest in $\hat{\beta}_n^{2SLS}$, even though all coefficients are equally important, as it is the sampling properties of $\hat{\beta}_n^{2SLS}$ that drive everything.

The two-stage least squares estimator has traditionally been popular because it has the property of being relatively efficient among a certain class of IV estimators in this model. We will provide the result in a theorem-proof style.

[9] It is worth noting at this point that subsequent distributional analysis is extremely hard, it took nearly 20 years for the literature to find an exact sampling distribution for $\hat{\beta}_n^{2SLS}$. The problem was eventually cracked by Phillips (1980). (As an aside, Peter Charles Bonest Phillips (1948–) has been one of the world's leading econometric theorists since the early 1970s, his first *Econometrica* having come from his Master's thesis in 1972. Most of his career has been spent at Yale although, even in semi-retirement, he maintains part-time positions in Auckland, Singapore and Southampton. There are few, if any, areas of econometric substance to which Peter hasn't made substantial contributions at some point over his stellar career, including finite sample properties in simultaneous equations models and time series models.) Other significant contributions to this literature include Hillier, Kinal, and Srivastava (1984), Hillier (1985, 1990), in respect of $\hat{\beta}_n^{2SLS}$, and Phillips (1984), Skeels (1995), in respect of $\hat{\gamma}_n^{2SLS}$. Phillips (1983) provides a good introduction to the distribution theory that was required to address these problems, although the original working paper version (Phillips, 1982) is even better. What you will find is that, where there is just a single endogenous regressor, the density of $\hat{\beta}_n^{2SLS}$ can be expressed in terms of confluent hypergeometric functions, which is part of the reason for discussing them earlier in the semester. If we had more time we would have worked through this, but we don't.

Theorem 1 (On the Relative Efficiency of Two-Stage Least Squares)**.** *In the single structural equation* (18)*, with corresponding reduced form* (19) *satisfying the compatibility conditions* (20)–(21)*, the two-stage least squares estimator* $\hat{\theta}_n^{2SLS}$ *defined in* (22) *is asymptotically relatively efficient among that class of estimators for whom the set of instruments is defined to be* $J = [Z_1, Z_{21}]$*, where* $Z_2 = [Z_{21}, Z_{22}]$*, so that* $Z_{21} \subseteq Z_2$*. In the event that* $Z_{21} = Z_2$*,* Z_{22} *is an empty matrix that has no columns.*

Proof. From (17), we know that the asymptotic covariance matrix of $\hat{\theta}_n^{2SLS}$ is of the form $\sigma^2(X'P_ZX)^{-1}$. Any other estimator from the class under consideration will have an asymptotic covariance matrix of the form $\sigma^2(X'P_JX)^{-1}$. Consider the matrix $X'P_ZX - X'P_JX = X'(P_Z - P_J)X$. From Equation (4.7.11a) of the Matrices handout,

$$P_Z - P_J = M_J Z_{22}(Z_{22}' M_J Z_{22})^{-1} Z_{22}' M_J.$$

The important part of this result is that the matrix on the right-hand side of this equality is a symmetric, idempotent matrix and so is positive semi-definite. That is, we can write $P_Z - P_J \geq 0$ or $P_Z \geq P_J$. It follows that $X'P_ZX \geq X'P_JX$ and hence that $(X'P_ZX)^{-1} \leq (X'P_JX)^{-1}$, which establishes the desired result. □

Bibliography

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. Oxford University Press, New York. 9

Hillier, G. H. (1985). On the joint and marginal densities of instrumental variable estimators in a general structural equation. *Econometric Theory* 1, 53–72. 13

Hillier, G. H. (1990). On the normalization of structural equations: Properties of direction estimators. *Econometrica* 58(5), 1181–1194. 13

Hillier, G. H., T. W. Kinal, and V. K. Srivastava (1984). On the moments of ordinary least squares and instrumental variables estimators in a general structural equation. *Econometrica* 52(1), 185–202. 13

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 185, 71–110 (+ 5 plates), ISSN 0264-3820, doi:10.1098/rsta.1894.0003. URL http://rsta.royalsocietypublishing.org/content/185/71 1

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 186, 343–414, ISSN 0264-3820, doi:10.1098/rsta.1895.0010. 1

Phillips, P. C. B. (1980). The exact distribution of instrumental variable estimators in an equation containing n+1 endogenous variables. *Econometrica* 48(4), 861–878. 13

Phillips, P. C. B. (1982). Small sample distribution theory in econometric models of simultaneous equations. Cowles Foundation Discussion Papers 617, Cowles Foundation for Research in Economics, Yale University. 13

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics, Volume I*, Z. Griliches and M. D. Intriligator, editors, chapter 8, 449–516, North Holland, Amsterdam. 13

Phillips, P. C. B. (1984). The exact distribution of exogenous variable coefficient estimators. *Journal of Econometrics* 26(1), 387–398. 13

Skeels, C. L. (1995). Some exact results for estimators of the coefficients on the exogenous variables in a single equation. *Econometric Theory* 11, 484–497. 13

A A Derivation of the Reduced Form Variance

The matrix $[v, V]$ is a matrix of n disturbances from $m+1$ different equations. That is, the matrix contains $n(m+1)$ elements. We can write it as

$$[v, V] = \begin{bmatrix} v_{11} & V_{12} & \dots & V_{1(m+1)} \\ v_{21} & V_{22} & \dots & V_{2(m+1)} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & V_{n2} & \dots & V_{n(m+1)} \end{bmatrix}$$

Here the first subscript denotes the observation number and the second subscript the equation number.[10] So each row contains a set of contemporaneous disturbances from each of the equations and there are n such sets of disturbances, one for each observation. The assumption is the the elements of any given row may be correlated with each other, with common covariance matrix[11]

$$\text{Var}\left[\begin{bmatrix} v_{j1} \\ V_{j2} \\ \vdots \\ V_{j(m+1)} \end{bmatrix}\right] = \Omega = \begin{bmatrix} \omega_{11} & \omega_{12} & \dots & \omega_{1(m+1)} \\ \omega_{12} & \omega_{22} & \dots & \omega_{2(m+1)} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1(m+1)} & \omega_{2(m+1} & \dots & \omega_{(m+1)(m+1)} \end{bmatrix}, \quad j = 1, \dots, n,$$

but that the elements of any given row (observation) are not correlated with each other so that

$$\text{Cov}\left[\begin{bmatrix} v_{i1} \\ V_{i2} \\ \vdots \\ V_{i(m+1)} \end{bmatrix}, \begin{bmatrix} v_{j1} \\ V_{j2} \\ \vdots \\ V_{j(m+1)} \end{bmatrix}\right] = 0_{(m+1)(m+1)}, \quad i \neq j = 1, \dots, n,$$

where $0_{(m+1)(m+1)}$ denotes an $(m+1) \times (m+1)$ matrix of zeros. In the definition of Ω all the elements on the leading diagonal are variances and those on the off-diagonals are covariances. It is convenient to group the elements of Ω into the (1,1) element, ω_{11}, the remaining elements in the first column, which we shall call ω_{21} in a horrible overloading of notation, the remaining elements of the first row, which are then ω_{21}' in the same horrible

[10] Not the equations really have numbers but it is a useful way of distinguishing between them. Our interest is then in estimating the coefficients of the first equation.

[11] Note that I have transposed the elements of the row because we have defined the variance of a vector to be the variance of a column vector. Also, I have exploited the symmetry of Ω to write $\omega_{ij} = \omega_{ji}$ as appropriate.

overloading of notation, and then the remaining $m \times m$ submatrix of elements will be referred to as Ω_{22}, so that we can write

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{21}' \\ \omega_{21} & \Omega_{22} \end{bmatrix}.$$

To obtain the variance of $[v, V]$ we first stack its elements up into a column vector. We do this using the vec operator. Let

$$w = \text{vec}([v, V]) = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \\ V_{12} \\ V_{22} \\ \vdots \\ V_{n2} \\ V_{13} \\ \vdots \\ V_{n(m+1)} \end{bmatrix}.$$

Then, as we have assumed that $\text{E}[[v, V]] = 0_{n(m+1)} \implies \text{E}[w] = 0_{n(m+1)}$, an $n(m+1) \times 1$ zero vector, we can calculate the the variance of w as in Figure 1. As w is an $n(m+1) \times 1$ vector it follows that $\text{Var}[w]$ is an $(m+1) \times (m+1)$ matrix. If you fill in more of the final matrix expression from Figure 1 yourself, which you should, then you see that it has the following structure:

$$\text{Var}[w] = \begin{bmatrix} \omega_{11} I_n & \omega_{12} I_n & \dots & \omega_{1(m+1)} I_n \\ \omega_{12} I_n & \omega_{22} I_n & \dots & \omega_{2(m+1)} I_n \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{1(m+1)} I_n & \omega_{2(m+1)} I_n & \dots & \omega_{(m+1)(m+1)} I_n \end{bmatrix}$$

As you can see this matrix has a peculiar structure whereby each element of the matrix is some number multiplied by a common matrix, in this case I_n. Such structures are known as Kronecker products, named in honour of the great german mathematician Leopold Kronecker (1823–1891). We represent such products using the $\otimes$ symbol, thus $\text{Var}[w] = \Omega \otimes I_n$. More generally, if one has an $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

then the Kronecker product $A \otimes B$, where B is a $p \times q$ matrix, is the $mp \times nq$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}.$$

$$
\begin{aligned}
\operatorname{Var}[w] &= \mathrm{E}\left[ww'\right] = \mathrm{E}\left[\begin{array}{c}\left[\begin{array}{c} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \\ V_{12} \\ V_{22} \\ \vdots \\ V_{n2} \\ V_{13} \\ \vdots \\ V_{n(m+1)} \end{array}\right] \\ \left[\begin{array}{cccccccccc} v_{11} & v_{21} & \cdots & v_{n1} & V_{12} & V_{22} & \cdots & V_{n2} & V_{13} & \cdots & V_{n(m+1)} \end{array}\right] \end{array}\right] \\
&= \begin{bmatrix} \mathrm{E}\left[v_{11}^2\right] & \mathrm{E}\left[v_{11}v_{21}\right] & \cdots & \mathrm{E}\left[v_{11}v_{n1}\right] & \mathrm{E}\left[v_{11}V_{12}\right] & \mathrm{E}\left[v_{11}V_{22}\right] & \cdots & \mathrm{E}\left[v_{11}V_{n2}\right] & \mathrm{E}\left[v_{11}V_{13}\right] & \cdots & \mathrm{E}\left[v_{11}V_{n(m+1)}\right] \\ \mathrm{E}\left[v_{21}v_{11}\right] & \mathrm{E}\left[v_{21}^2\right] & & \mathrm{E}\left[v_{21}v_{n1}\right] & \mathrm{E}\left[v_{21}V_{12}\right] & \mathrm{E}\left[v_{21}V_{22}\right] & & \mathrm{E}\left[v_{21}V_{n2}\right] & \mathrm{E}\left[v_{21}V_{13}\right] & & \mathrm{E}\left[v_{21}V_{n(m+1)}\right] \\ & & & & & \vdots & \vdots & & & & \\ \mathrm{E}\left[V_{n(m+1)}v_{11}\right] & \mathrm{E}\left[V_{n(m+1)}v_{11}\right] & & \mathrm{E}\left[V_{n(m+1)}v_{n1}\right] & \mathrm{E}\left[V_{n(m+1)}V_{12}\right] & \mathrm{E}\left[V_{n(m+1)}V_{22}\right] & \cdots & \mathrm{E}\left[V_{n(m+1)}V_{n2}\right] & \mathrm{E}\left[V_{n(m+1)}V_{13}\right] & \cdots & \mathrm{E}\left[V_{n(m+1)}^2\right] \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{E}\left[v_{11}^2\right] & \mathrm{E}\left[v_{11}v_{21}\right] & \cdots & \mathrm{E}\left[v_{11}v_{n1}\right] & \mathrm{E}\left[v_{11}V_{12}\right] & \mathrm{E}\left[v_{11}V_{22}\right] & \cdots & \mathrm{E}\left[v_{11}V_{n2}\right] & \mathrm{E}\left[v_{11}V_{13}\right] & \cdots & \mathrm{E}\left[v_{11}V_{n(m+1)}\right] \\ \mathrm{E}\left[v_{21}v_{11}\right] & \mathrm{E}\left[v_{21}^2\right] & & \mathrm{E}\left[v_{21}v_{n1}\right] & \mathrm{E}\left[v_{21}V_{12}\right] & \mathrm{E}\left[v_{21}V_{22}\right] & & \mathrm{E}\left[v_{21}V_{n2}\right] & \mathrm{E}\left[v_{21}V_{13}\right] & & \mathrm{E}\left[v_{21}V_{n(m+1)}\right] \\ \vdots & \vdots & & & & \vdots & \vdots & & & & \\ \mathrm{E}\left[V_{n(m+1)}v_{11}\right] & \mathrm{E}\left[V_{n(m+1)}v_{11}\right] & & \mathrm{E}\left[V_{n(m+1)}v_{n1}\right] & \mathrm{E}\left[V_{n(m+1)}V_{12}\right] & \mathrm{E}\left[V_{n(m+1)}V_{22}\right] & \cdots & \mathrm{E}\left[V_{n(m+1)}V_{n2}\right] & \mathrm{E}\left[V_{n(m+1)}V_{13}\right] & \cdots & \mathrm{E}\left[V_{n(m+1)}^2\right] \end{bmatrix} \\
&= \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 & \omega_{12} & 0 & \cdots & 0 & \omega_{13} & \cdots & 0 \\ 0 & \omega_{11} & & 0 & 0 & \omega_{12} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & & & \vdots & & & \vdots & & \\ \vdots & & & & & \vdots & & & \vdots & & \\ 0 & 0 & & \omega_{1(m+1)} & 0 & 0 & \cdots & \omega_{2(m+1)} & 0 & \cdots & \omega_{(m+1)(m+1)} \end{bmatrix}
\end{aligned}
$$

Figure 1: Variance Calculation