Chapter 16

Nonlinear Regression

16.1 Model and Assumptions

16.1.1 Model

In the prequel we have studied the linear regression model in some detail. Such models are linear with respect to the parameters given the explanatory variables and have additive disturbances. The function may be nonlinear in terms of the explanatory variable, as in a polynomial function, but with a simple redefinition of the variables is linear in the variables given the parameters. Strictly speaking such models are bilinear in the parameters and explanatory variables. In the space of functions of the explanatory variables that have parametric representations this is a very small (measure zero) subspace.

We now consider the more general class of models that are intrinsically nonlinear. They are necessarily nonlinear with respect to the parameters since nonlinearity with respect to the variables only can be handled with the linear model by a simple redefinition of variables. For identification reasons, we restrict our attention to functions that are additive with respect to the disturbances. Notationally, such a model can be written

\[ y_t = h(x_{t1}, x_{t2}, \ldots, x_{tk}; \theta_1, \theta_2, \ldots, \theta_p) + u_t \]

\[ = h(x_t, \theta) + u_t \]

\[ = h_t(\theta) + u_t \]

for \( t = 1, 2, \ldots, n \), where \( h(\cdot) \) is a scalar-valued function that is nonlinear with respect to the \( p \times 1 \) vector of parameters \( \theta \). The \( k \times 1 \) vector \( x_t \) comprise the independent variables of the model, and may or may not enter nonlinearly. The use of subscript-\( t \) does not necessarily indicate time-series data but is used for notational economy.
16.1.2 Some Examples

Some simple examples will make the meaning of the types of nonlinearities entertained more transparent. Consider the simple Cobb-Douglas production function with multiplicative disturbances

$$ Q_t = A K_t^\alpha L_t^\beta u_t $$

which can be transformed by taking logs on both sides to become

$$ \ln Q_t = \ln A + \alpha \ln K_t + \beta \ln L_t + \ln u_t $$

and

$$ q_t = a + \alpha k_t + \beta l_t + \epsilon_t $$

is a linear model with $q_t = \ln Q_t$, $k_t = \ln K_t$, $l_t = \ln L_t$, $\epsilon_t = \ln u_t$, and $a = \ln A$. Thus although the model appears nonlinear it is linear after a transformation and redefinition of variables.

A closely related model that is not so easily handled is Cobb-Douglas with additive errors:

$$ Q_t = A K_t^\alpha L_t^\beta + u_t $$

Here any transformation to make the model linear in the parameters will also involve $u_t$ in a nonlinear fashion that also involves the parameters. It is intrinsically nonlinear. Although this may seem like a rather artificial distinction, we learned in the misspecification chapter that distinguishing between additive and multiplicative error forms can be important.

A generalization of the Cobb-Douglas model is the CES model, with additive errors

$$ y_t = \beta_0 [\beta_1 x_t^{\beta_2} + (1 - \beta_1) x_t^{\beta_2}]^{-\beta_3/\beta_2} + u_t $$

which is also intrinsically nonlinear as in the last paragraph. But the same model with multiplicative errors:

$$ y_t = \beta_0 [\beta_1 x_t^{\beta_2} + (1 - \beta_1) x_t^{\beta_2}]^{-\beta_3/\beta_2} \cdot u_t $$

is now also intrinsically nonlinear, since taking logs will not make the nonlinear function linear in the parameters. There are a host of other models, in fact anything other than the bilinear form, that will also be intrinsically nonlinear.

16.1.3 Basic Assumptions

The nonlinear regression equation can be considered the specification of the model. In addition, we will make stochastic assumptions analogous to those introduced for the linear model. For reasons that will become apparent below, we cannot avoid the problem of stochastic regressors so we introduce the first four assumptions for the conditional zero mean case used in Chapter 11

(i) $E[u_t] = 0$

(ii) $E[u_t^2] = \sigma^2$
16.2. NONLINEAR LEAST SQUARES (NLS)

(iii) \((x_t, u_t)\) jointly i.i.d.
(iv) \(E[u_t | x_t] = 0\) and \(E[u_t^2 | x_t] = \sigma^2\)

Note that the unconditional moment assumptions (i) and (ii) are redundant given (iv) but are explicitly retained for completeness. The additional assumptions needed to obtain asymptotic behavior will be introduced below when needed.

16.1.4 Some Notation

In vector notation, the model can be written

\[
y = h(\theta) + u
\]

where \(y = (y_1, y_2, ..., y_n)'\), \(u = (u_1, u_2, ..., u_n)'\), and \(h(\theta) = (h_1(\theta), h_2(\theta), ..., h_n(\theta))'\) are all \(n \times 1\) vectors. We denote \(\Theta \in \mathbb{R}^p\) as the relevant parameter space for \(\theta\) and introduce the notation

\[
[H(\theta)]' = \left( \frac{\partial h_1(\theta)}{\partial \theta} : \frac{\partial h_2(\theta)}{\partial \theta} : \ldots : \frac{\partial h_n(\theta)}{\partial \theta} \right)
\]

\[
Q(\theta) = E \left[ \frac{\partial h_t(\theta)}{\partial \theta} \frac{\partial h_t(\theta)}{\partial \theta'} \right]
\]

\[
M = E \left[ u_t^2 \frac{\partial h_t(\theta)}{\partial \theta} \frac{\partial h_t(\theta)}{\partial \theta'} \right].
\]

Rather obviously, the linear regression model is a very special case where \(h(\theta) = X\beta, H(\theta) = X, \) and \(\theta = \beta\). In addition, we will sometimes use the notation \(z = (y, x')'\) for all the observable variables.

16.2 Nonlinear Least Squares (NLS)

16.2.1 Least Squares

In the \((p + 1)\)-dimensional space of \(y_t\) and \(\theta\), the function \(h_t(\theta)\) defines a \(p\)-dimensional hyper-surface, given \(x_t\). Analogous to the linear model, we seek to find estimates that make the hyper-surface close to the values of \(y_t\) by some measure. Specifically, we find \(\hat{\theta}\) such that

\[
\psi_n(\theta) = n^{-1} \sum_{t=1}^{n} (y_t - h_t(\theta))^2
\]

is minimized so

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \psi_n(\theta)
\]

where, as mentioned above, \(\Theta\) is the set from which we are choosing.
As with the linear model, in practice, we find the estimates as the solution to the first-order conditions

\[ 0 = \frac{\partial \psi_n(\theta)}{\partial \theta} = -\frac{2}{n} \sum_{t=1}^{n} (y_t - h_t(\theta)) \frac{\partial h_t(\theta)}{\partial \theta} = -\frac{2}{n} H(\theta)'(y - h(\theta)). \]

Note that as a consequence of \( h(\theta) \) being nonlinear in \( \theta \), \( H(\theta) \) will also be a function of \( \theta \), and we must find the roots of a \( p \)-equation nonlinear system. This is the essential complication in the nonlinear model relative to the linear.

### 16.2.2 Gauss-Newton Method

In Newton’s method for solving a nonlinear equation, a linear approximation to the equation is solved and use iteratively to obtain a solution to the nonlinear problem. For the nonlinear least-squares problem we will take an analogous approach. We first expand \( h_t(\theta) \) in a linear Taylor series (about some initial estimate \( \theta^i \)) to obtain

\[ h(x_t, \theta) = h(x_t, \theta^i) + \frac{h(x_t, \theta^i)}{\partial \theta} (\theta - \theta^i) + R_t^i \]

where \( R_t^i \) is the remainder term. Substitution of this expression into the nonlinear regression and rearranging yields

\[ y_t - h(x_t, \theta^i) = \frac{h(x_t, \theta^i)}{\partial \theta} (\theta - \theta^i) + R_t^i + u_t. \]

for \( t = 1, 2, ..., n. \)

In vector/matrix notation we have

\[ y - h(\theta^i) = H(\theta^i)(\theta - \theta^i) + R^i + u \]

where \( R^i \) is the vector of remainder terms for all observations. Treating \( \theta^i \) as given, we define \( y^i = y - h(\theta^i) \), \( H^i = H(\theta^i) \), \( \delta^i = (\theta - \theta^i) \), and \( u^i = R^i + u. \)

We now can write the model in the form

\[ y^i = H^i \delta^i + u^i \]

which has every appearance of the linear regression model. Applying least squares to this linear model yields (assuming \( H^i \) has full column rank)

\[ \hat{\delta}^i = (H^i'H^i)^{-1}H^i'y^i \]

or

\[ \hat{\theta} - \theta^i = (H(\theta^i)'H(\theta^i))^{-1}H(\theta^i)'(y - h(\theta^i)) \]
where $\hat{\theta} = \theta^i + \hat{\delta}^i$ represent least squares estimates of $\theta$ after linearizing the system around the initial values $\theta^i$.

Define $\theta^{i+1} = \hat{\theta}$, then we can rearrange the least square estimation equation for the linearized model as the iterative equations

$$\theta^{i+1} = \theta^i + (H'H)^{-1}H'(y - h)\big|_{\theta^i}.$$

where $|_{\theta^i}$ indicates that $H$ and $h$ have been evaluated at $\theta^i$. Note that at each step, we are simply regressing $y - h(\theta^i)$ on $H(\theta^i)$. Suppose we iterate until $\theta^{i+1}$ is unchanged from $\theta^i$, then at convergence, $\theta^{i+1} = \theta^i$ and

$$0 = (H'H)^{-1}H'(y - h)\big|_{\theta^i},$$

or, since $|H'H| \neq 0$,

$$0 = H'(y - h)\big|_{\theta^i},$$

and the first order conditions are satisfied. This procedure is called the Gauss-Newton procedure and is an application of Newton’s method since we are linearizing the first-order conditions and finding the solution and iterating.

Several practical issues arise in implementing the Guass-Newton procedure or any nonlinear interactive estimator for that matter. First is the possibility that multiple roots may exist for the first-order conditions. In such cases bounds can sometimes be placed on the estimates on the basis of what is economically reasonable, such as values that lie between 0 and 1, for example. And in some models it is possible to state categorically that the objective function is globally concave so there is only one root, such as in variations on the binary probability model.

A second issue is the question of convergence. The iterations of Newton’s method are known to converge at a quadratic rate in the neighborhood of the solution if the first derivative of the function to be solved exists. Specifically, for $\theta$ scalar and $\hat{\theta}$ the solution, this means $(\theta^{i+1} - \hat{\theta}) = O((\theta^i - \hat{\theta})^2)$ so if $(\theta^i - \hat{\theta})$ is small then $(\theta^{i+1} - \hat{\theta})$ is much smaller and convergence is assured. It is entirely possible that convergence will not occur in certain regions of the parameter space. An example of such a problematic region would occur if the function had a saddle point which was both a maximum in one direction and a minimum in another. The bounds discussed in the previous paragraph can sometimes be useful in ruling out such regions.

Sometimes initial values can be taken from related preliminary estimation approaches. One frequently encounters models and estimators where the researcher starts from initial consistent estimates and then iterates. For example, we could use values from the linear regression of the log-linear form of the Cobb-Douglas model to obtain starting values for use in nonlinear least squares applied to the nonlinear form. Such an approach is also useful in avoiding the multiple root problem.

A third issue is the problem of obtaining derivatives. A first-best solution would be to have analytic derivatives that are explicitly evaluated in the iterations. Unfortunately, obtaining such derivatives and coding them into the
algorithm is sometime infeasible, tedious, or very complicated. In such cases numerical approximations to the derivatives may be used. It should be noted, however, that such an approach pays a price in terms of precision of the estimates and/or their estimated standard errors, which both use these approximate derivatives.

16.2.3 Consistency

We now turn to the properties of the estimator that come from the nonlinear least squares exercise. It is clear from the above that the regressors in each iteration, namely $H(\theta^i)$, are stochastic. As a result, the small sample properties of the estimator are problematic and we must rely on the large-sample asymptotic behavior for purposes of inference.

The following consistency result follows from verification of the high-level conditions given in Theorem 4.5 from Chapter 4.

**Theorem 16.1.** Suppose (i) $\Theta$ compact; (ii) $h(x; \theta)$ continuous for $\theta \in \Theta$ with probability 1; (iii) $\exists d(x)$ with $(h(x; \theta) - h(x; \theta^o))^2 < d(x) \forall \theta \in \Theta$ and $E[d(x)] < \infty$; (iv) $h(x; \theta) = h(x; \theta^o)$ for all $x$ iff $\theta = \theta^o$, then $\hat{\theta} \overset{p}{\to} \theta^o$. □

**Proof:** See Appendix to this chapter. □

Strictly speaking, these low-level conditions are not assured, in general, and should be verified for the nonlinear model at hand. The first two assumptions will be relatively innocuous and easily verified for most models. The third assumption, which is used to prove uniform convergence in probability, can be tedious to verify but will not usually present problems.

The fourth assumption assures global identifiability of the parameter vector within $\Theta$. This global identification means that we can obtain consistent estimates regardless of the starting values used. It may be unrealistic since it is possible that $(h(x; \theta) - h(x; \theta^o)) = 0$ has solutions other than $\theta^o$. This condition may be weakened to local identification by substituting the condition: $h(x; \theta) = h(x; \theta^o)$ iff $\theta = \theta^o$ for all $x$ and $\theta$ in a neighborhood of $\theta^o$. In this case, consistency would follow if the initial value is itself consistent or is assured to be in the neighborhood specified. This is a standard problem in nonlinear estimation.

16.2.4 Asymptotic Normality

The following asymptotic normality result follows from verifying the high-level conditions of Theorem 4.6 from Chapter 4. Note that this result does not presuppose any particular distribution of $u_t$.

**Theorem 16.2.** Suppose (i) $\hat{\theta} \overset{p}{\to} \theta^o$; (ii) $\theta^o$ interior to $\Theta$; (iii) $h(x; \theta)$ twice continuously differentiable in a neighborhood $N$ of $\theta^o$; (iv) $\exists D(z)$ s.t. $\|\frac{\partial^2(y-h(x, \theta))}{\partial \theta \partial \theta'}\| < D(z)$ for all $\theta \in N$ and $E[D(z)] < \infty$; (v) $E \left[ \frac{\partial h(x, \theta^o)}{\partial \theta} \frac{\partial h(x, \theta^o)}{\partial \theta'} \right] = Q$ exists and is nonsingular, then $\sqrt{n}(\hat{\theta} - \theta^o) \overset{d}{\to} N(0, \sigma^2 Q^{-1})$. □
16.3. MAXIMUM LIKELIHOOD ESTIMATION

Proof: See Appendix to this chapter. □

We do not need all the assumptions of the consistency proof only the result (i), which can occur under more general conditions. Condition (ii) is needed to exclude boundary points for which the limiting distribution cannot be normal since it is truncated even if the distribution is collapsing. Condition (iii) is a regularity condition and easily verified. Condition (iv) is analogous to (iii) in the consistency theorem except now we are talking about estimating the matrix $Q$ consistently. Condition (v) is needed for the limiting covariance to be invertible and usually follows from local identification.

16.3 Maximum Likelihood Estimation

16.3.1 Normality and Likelihood Function

As with the linear model, there is a one-to-one relationship between least squares estimation of $\theta$ and maximum likelihood estimation. Suppose $u_t \sim i.i.d. \ N(0, \sigma^2) ind. x_t$ then $y_t | x_t \sim N(h_t(x_t, \theta), \sigma^2)$.

Due to independence we can write the joint density or likelihood function as

$$f(y|x, \theta, \sigma^2) = L(\theta, \sigma^2 | y, x) = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - h_t(\theta))^2 \right)$$

and the log-likelihood function is

$$\mathcal{L}(\theta, \sigma^2 | y, x) = -\frac{n}{2} \ln 2\pi \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - h_t(\theta))^2$$

$$= -\frac{n}{2} \ln 2\pi \sigma^2 - \frac{n}{2\sigma^2} \psi_n(\theta).$$

Obviously, maximizing this function with respect to $\theta$ is identical to minimizing $\psi_n(\theta)$ with respect to $\theta$. Thus, the maximum likelihood estimator of $\theta$ is $\hat{\theta}$, the NLS estimator. Given the regularity conditions are met, this means that $\hat{\theta}$ is CUAN and efficient within this class.

For estimating $\sigma^2$, we find

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (y_t - h_t(\hat{\theta}))^2$$

$$= \psi_n(\hat{\theta})$$
is the maximum likelihood estimator of $\sigma^2$. It can be shown to be generally consistent under the assumptions of Theorem 16.1 and asymptotically normal under general conditions. Moreover, it will attain the efficiency bound within the class of CUAN estimators of $\sigma^2$.

### 16.3.2 Newton’s Method

Consider the direct maximization of the log-likelihood using Newton’s method. Suppose $\hat{\theta}$ is the MLE, then solving the first-order conditions yields

$$0 = \frac{\partial \ln L(\theta)}{\partial \theta} - \frac{n}{2\sigma^2} \frac{\partial \psi_n(\theta)}{\partial \theta}.$$

Expanding this in a linear Taylor’s series around an initial value, say $\theta^i$, yields

$$0 = -\frac{n}{2\sigma^2} \frac{\partial \psi_n(\theta^i)}{\partial \theta} - \frac{n}{2\sigma^2} \frac{\partial^2 \psi_n(\theta^i)}{\partial \theta \partial \theta'} (\theta - \theta^i) + R.$$

Ignoring the remainder term and solving for $\hat{\theta} = \theta^{i+1}$ yields

$$\theta^{i+1} = \theta^i - \left(\frac{\partial^2 \psi_n(\theta^i)}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial \psi_n(\theta^i)}{\partial \theta}$$

where

$$\frac{\partial \psi_n(\theta^i)}{\partial \theta} = -\frac{2}{n} \sum h_t(\theta^i) (y_t - h_t(\theta^i))$$

$$\frac{\partial^2 \psi_n(\theta^i)}{\partial \theta \partial \theta'} = \frac{2}{n} \sum_{t=1}^n \left\{ (y_t - h_t(\theta^i)) \frac{\partial^2 h_t(\theta^i)}{\partial \theta \partial \theta'} - \frac{\partial h_t(\theta^i)}{\partial \theta} \frac{\partial h_t(\theta^i)}{\partial \theta'} \right\}.$$

Note that the weight matrix for this approach differs from the Gauss-Newton case.

### 16.3.3 Method of Scoring

A variation on Newton’s method that is frequently advocated is the method of scoring. First we note that the weight matrix is an average of an i.i.d. random matrix and will converge in probability to its expectation. At the target, this expectation can be written as the average of a simpler expression,

$$\frac{\partial^2 \psi_n(\theta^0)}{\partial \theta \partial \theta'} \xrightarrow{p} E[2 \frac{\partial h_t(\theta^0)}{\partial \theta} \frac{h_t(\theta^0)}{\partial \theta'}] = 2Q.$$
In the method of scoring we estimate the expectation of this simpler expression directly with an average and use that as the weight matrix. Thus we have

\[ \theta^{i+1} = \theta^i - \left( \frac{2}{n} \sum_{t=1}^{n} \frac{\partial h_t(\theta)}{\partial \theta} \frac{\partial h_t(\theta)}{\partial \theta'} \right) \frac{1}{n} \frac{\partial \psi_n(\theta^i)}{\partial \theta} \]

\[ = \theta^i + \left[ \sum_{t=1}^{n} \frac{h_t(\theta^i)}{\partial \theta} \frac{\partial h_t(\theta^i)}{\partial \theta'} \right]^{-1} \sum_{t=1}^{n} \frac{\partial h_t(\theta^i)}{\partial \theta'} (y_t - h_t(\theta)) \]

This is, of course, the Gauss-Newton method. Experience suggests that this may be better behaved since it imposes restrictions of the model on the weight matrix that follow from examination of the expectation.

### 16.3.4 Controlling Step-Size

The approach above essentially approximates the nonlinear minimization by a quadratic and then minimizes that. If the problem is truly quadratic it will converge in one step. If not, then the step will hopefully move us toward the minimum. Sometimes it does not and instead overshoots the target or gets into a cycle where no progress occurs. This suggests that other techniques be adopted. However, the above quadratic approximation is known to work quite well near a unique minimum so it should be incorporated into any proposed solution.

It is useful to think of the step and its size in terms of gradients. Now \( \frac{\partial \psi_n(\theta)}{\partial \theta} \) is the gradient of the minimand and in the space of the parameters can be thought of as a direction vector that tells us how much the minimand changes for a unit increase in each of the parameters. In the neighborhood of the minimum this direction will point uphill away from the minimum and its negative will be the direction of steepest descent for a unit step-size. Premultiplication by the weight matrix is introduced to allow for the second-order curvature and modifies the direction accordingly. It is the optimal step-size when we approximate the problem by a quadratic.

The problem is that the derivative implicitly involves a unit step-size and that can result in overshooting or even cycling if the function is locally very non-quadratic. A solution is to introduce a step-size parameter

\[ \theta^{i+1} = \theta^i + \lambda (H' H)^{-1} H' (y - h) |_{\theta^i} \]

where 0 \( \leq \lambda \leq 1 \). There is a great deal of artwork in the choice of \( \lambda \). For many problems the unit step-size works well. If it does not a simple solution is to use the halving step-size algorithm. In this approach, you start with \( \lambda = 1 \) and if the proposed step-size does not result in a reduction in the minimand you use \( \lambda = .5 \) and try again. You continue with this halving until a reduction is obtained. If a reduction cannot be obtained then the minimand has a flat spot and the minimum is not unique.
16.4 Inference

16.4.1 Standard Normal Tests

First we consider testing a scalar null hypothesis. Now, the convergent iterate \( \hat{\theta} \) will, by Theorem 16.2, satisfy
\[
\sqrt{n}(\hat{\theta} - \theta^o) \xrightarrow{d} N(0, \sigma^2 Q^{-1})
\]
where
\[
Q(\theta) = E\left[ \frac{\partial h_t(\theta)}{\partial \theta} \frac{\partial h_t(\theta)}{\partial \theta'} \right]\\
Q = Q(\theta^0).
\]
Thus,
\[
\frac{\sqrt{n}(\hat{\theta}_j - \theta^o_j)}{\sqrt{\sigma^2 [Q^{-1}]_{jj}}} \xrightarrow{d} N(0, 1)
\]
where \([Q^{-1}]_{jj}\) indicates, as in previous chapters, the \(jj\)-th element of the bracketed matrix.

A null hypothesis might provide \( \theta^o \) which leaves us needing estimates of \( \sigma^2 \) and \( Q \). An obvious estimator of \( \sigma^2 \) is
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (y_t - h(x_t, \hat{\theta}))^2.
\]
\[
= \psi_n(\hat{\theta})
\]

An alternative is
\[
s^2 = \frac{1}{n - p} \sum_{t=1}^{n} (y_t - h(x_t, \hat{\theta}))^2
\]
\[
= \frac{n}{n - p} \hat{\sigma}^2.
\]

A natural estimator for \( Q \) is
\[
\hat{Q} = \frac{1}{n} H(\hat{\theta})' H(\hat{\theta})
\]
\[
= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial h_t(\hat{\theta})}{\partial \theta} \frac{\partial h_t(\hat{\theta})}{\partial \theta'}
\]
which is the weight matrix and is provided as a byproduct of Gauss-Newton estimation.

Consistency of the variance estimator should follow from the conditions of Theorem 16.1. We need an additional smoothness assumption to guarantee consistency of the covariance matrix estimator. Specifically, we have
Theorem 16.3. If the hypotheses of Theorems 16.1 and 16.2 are satisfied and
\( \exists \delta(z) \) s.t. \( \left| \frac{\partial h(x, \theta)}{\partial \theta} \right| < \delta(z) \) for all \( \theta \) in a neighborhood \( N \) of \( \theta^0 \) and \( E[\delta(z)] < \infty \), then \( \hat{\sigma}^2 \rightarrow_p \sigma^2 \), \( s^2 \rightarrow_p \sigma^2 \) and \( \hat{Q} \rightarrow_p Q \). □

Proof: See Appendix to this chapter. □

Thus, a consistent estimator of the limiting covariance matrix is \( s^2 \hat{Q}^{-1} \).

Substituting this into the above ratio, we find

\[
\frac{\sqrt{n}(\hat{\theta}_j - \theta^0_j)}{\sqrt{s^2 \hat{Q}^{-1}}_{jj}} = \frac{(\hat{\theta}_j - \theta^0_j)}{\sqrt{s^2 [H(\hat{\theta})/H(\theta)]_{jj}^{-1}}} \xrightarrow{d} N(0, 1),
\]

under the null hypothesis \( H_0 : \theta_j = \theta^0_j \). Note that the denominator of the middle expression above is the usual least-squares calculated standard error from the final iteration.

Under an alternative hypothesis, say \( H_1 : \theta_j = \theta^1_j \neq \theta^0_j \), we have

\[
\frac{\sqrt{n}(\hat{\theta}_j - \theta^0_j)}{\sqrt{s^2 \hat{Q}^{-1}}_{jj}} = \frac{\sqrt{n}(\hat{\theta}_j - \theta^1_j)}{\sqrt{s^2 [H(\hat{\theta})/H(\theta)]_{jj}^{-1}}} + \frac{\sqrt{n}(\theta^1_j - \theta^0_j)}{\sqrt{s^2 [H(\hat{\theta})/H(\theta)]_{jj}^{-1}}} = N(0, 1) + \frac{\sqrt{n}(\theta^1_j - \theta^0_j)}{\sqrt{s^2 [Q^{-1}]_{jj}}} + o_p(1)
\]

so we see that the statistic will diverge in the positive/negative direction at the rate \( \sqrt{n} \) depending on whether \( (\theta^1_j - \theta^0_j) \) is positive/negative. The test will therefore be consistent.

16.4.2 Likelihood Ratio Statistics

Suppose that we seek to test the null hypothesis \( H_0 : r(\theta^0) = 0 \) where \( r(\cdot) \) is a \( q \times 1 \) continuously differentiable vector function. Since we have the maximum likelihood estimator we can directly apply the results of Chapter 5. There we introduced the trinity of tests for a vector restriction. They were the likelihood ratio, Wald and Lagrange multiplier tests. It is instructive to present these tests in some detail for the present case as with the linear regression model.

First, we consider the likelihood ratio statistic. Suppose \( \hat{\theta} \) denotes the unrestricted maximum likelihood estimator used in the Wald-type test and \( \tilde{\theta} \) is the restricted maximum likelihood estimator, then \( r(\hat{\theta}) = 0 \). Let \( \hat{u} = y - h(\hat{\theta}) \) and \( \tilde{u} = y - h(\tilde{\theta}) \) denote the unrestricted and restricted residuals, then \( \hat{\sigma}^2 = \hat{u}'\hat{u}/n \) and \( \tilde{\sigma}^2 = \tilde{u}'\tilde{u}/n \) are the unrestricted and restricted maximum likelihood estimators of \( \sigma^2 \).

Now substitution of \( \hat{\sigma}^2 \) into the unrestricted log-likelihood function yields
the concentrated unrestricted log-likelihood function

\[ L^u = -\frac{n}{2} \ln 2\pi \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} \psi_n(\theta) \]

\[ = -\frac{n}{2} \ln 2\pi \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} e_u' e_u \]

\[ = -\frac{n}{2} \ln 2 - \frac{n}{2} \ln \hat{\sigma}^2 - \frac{n}{2}. \]

Similarly the concentrated restricted log-likelihood is

\[ L^r = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \tilde{\sigma}^2 - \frac{n}{2}. \]

Thus we have the likelihood ratio statistic and its usual limiting distribution

\[ LR = 2(L^u - L^r) \]

\[ = 2(-\frac{n}{2} \ln \hat{\sigma}^2 + \frac{n}{2} \ln \tilde{\sigma}^2) \]

\[ = n \ln \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \xrightarrow{d} \chi^2_q. \]

There is a close relationship between the LR statistic and the familiar F-statistic comparing restricted and unrestricted sums-of-squares. Note that

\[ \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = \frac{\tilde{\mathbf{u}}' \tilde{\mathbf{u}} / n}{\mathbf{u}' \mathbf{u} / n} = 1 + \frac{\tilde{\mathbf{u}}' \mathbf{u} - \mathbf{u}' \mathbf{u}}{\mathbf{u}' \mathbf{u} / (n-p)} \]

\[ = 1 + \frac{q}{n-p} F \]

The likelihood ratio statistic is obviously a nonlinear monotonic function of the F-statistic. And in the limit the relationship becomes linear. Expand the statistic as a function of \( \tilde{\sigma}^2 \) about \( \hat{\sigma}^2 \) to obtain

\[ LR = n \frac{1}{\tilde{\sigma}^2 / \hat{\sigma}^2} \left( \tilde{\sigma}^2 - \hat{\sigma}^2 \right) + O_p(n(\tilde{\sigma}^2 - \hat{\sigma}^2)^2) \]

\[ = n \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)}{\hat{\sigma}^2} + o_p(1) \]

\[ = \frac{\tilde{\mathbf{u}}' \mathbf{u} - \mathbf{u}' \mathbf{u}}{\hat{\sigma}^2} + o_p(1) \]

since \( (\tilde{\sigma}^2 - \hat{\sigma}^2) = O_p(1/n) \) under the null. If we divided the numerator of the leading term by the number of restrictions, we would have the nonlinear regression analog to the F-statistic in the linear regression model except we are using the restricted estimator of \( \sigma^2 \) in the denominator, which will make no difference asymptotically, under the null.
16.4.3 Wald-Type Tests

The asymptotic normality result may be utilized to obtain a quadratic form based test of the Wald type. Using the delta method we find that

$$\sqrt{n}r(\hat{\theta}) \xrightarrow{d} N(0, \sigma^2 RQ^{-1}R')$$

and

$$nr(\hat{\theta})'[\sigma^2 RQ^{-1}R']^{-1}r(\hat{\theta}) \xrightarrow{d} \chi^2_q$$

where $R(\theta) = \frac{\partial r(\theta)}{\partial \theta'}$, $R = R(\theta^0)$, and $RQ^{-1}R'$ nonsingular. Consistent estimates of $\sigma^2$ and $Q$ are introduced above, and $\hat{R} = R(\hat{\theta})$ will be consistent for $R$. Thus, under the null hypothesis, we have the feasible Wald-type statistic

$$W = nr(\hat{\theta})'[\hat{\sigma}^2 \hat{R}\hat{Q}^{-1}\hat{R}']^{-1}r(\hat{\theta})$$

$$= r(\hat{\theta})'[\hat{\sigma}^2 \hat{R}(\hat{H}\hat{H})^{-1}\hat{R}']^{-1}r(\hat{\theta}) \xrightarrow{d} \chi^2_q.$$ 

Under the alternative hypothesis $H_1 : r(\theta^0) \neq 0$, the statistic will diverge in the positive direction at the rate $n$.

It is also possible to restate the Wald-type test in terms of sums-of-squares of the restricted and unrestricted models and thereby see the close connection between LR and W tests. Without loss of generality, we can reparameterize the model so the null restriction is a zero restriction. Let $\theta_1$ denote the first $p - q$ elements of the parameter vector and $\theta_2$ the final $q$ elements. Let $\theta_2^* = r(\theta_1, \theta_2)$ and suppose the $q \times 1$ restriction function is one-to-one with $\theta_2$. Then we can obtain the inverse function $\theta_2 = r^{-1}(\theta_1, \theta_2^*)$ and the regression equation becomes

$$y_t = h(x_t, \theta_1, \theta_2) + u_t$$

$$= h(x_t, \theta_1, r^{-1}(\theta_1, \theta_2^*)) + u_t$$

$$= h^*(x_t, \theta_1, \theta_2^*) + u_t$$

and the null hypothesis becomes $H_0 : \theta_2^* = 0$. Note that this renormalization of the restriction will leave the likelihood ratio statistic unchanged since MLE is invariant to normalizations. Also note that $R = (0, I_q)$, the derivative matrix for the renormalized restrictions, will now be a constant and nonstochastic.

Using the approach of Section 8.4.3 for the linear model on the nonlinear
model we find

\[ r(\hat{\theta})'[\hat{\sigma}^2 R(\hat{H}'\hat{H})^{-1} R']^{-1} r(\hat{\theta}) = \hat{\theta}_2[\hat{\sigma}^2 \hat{H}' M_1 \hat{H} \hat{\theta}_2]^{-1} \]

\[ = \frac{1}{\hat{\sigma}^2} \hat{\theta}_2' \hat{H}_2' M_1 \hat{H}_2 \hat{\theta}_2 \]

\[ = \frac{1}{\hat{\sigma}^2} (M_1 u - \hat{u})'(M_1 u - \hat{u}) + o_p(1) \]

\[ = \frac{1}{\hat{\sigma}^2} (u'M_1 u - 2 \hat{u}' M_1 u + \hat{u}' \hat{u}) + o_p(1) \]

\[ = \frac{1}{\hat{\sigma}^2} (u'M_1 u - \hat{u}' \hat{u}) + o_p(1) \]

where \( \hat{H} = (\hat{H}_1 : \hat{H}_2) \), \( M_1 = I_n - \hat{H}_1(\hat{H}_1' \hat{H}_1)^{-1} \hat{H}_1' \) and expansion of \( h(\hat{\theta}) \) about \( \theta^0 \) yields

\[ M_1 \hat{u} = M_1(y - h(\hat{\theta})) \]

\[ = M_1(y - h(\theta^0) + \hat{H}_1(\hat{\theta}_1 - \theta^0) + \hat{H}_2 \hat{\theta}_2 + O_p(\|\hat{\theta} - \theta^0\|^2)) \]

\[ = M_1 u + M_1 \hat{H}_2 \hat{\theta}_2 + o_p(1). \]

Thus the leading behavior of the \( W \) statistic is given by

\[ \frac{1}{\hat{\sigma}^2} (\hat{u}' \hat{u} - \hat{u}' \hat{u}) \overset{d}{\longrightarrow} \chi_q^2 \]

which, if divided by \( q \), is the nonlinear analog to the \( F \)-statistic and differs from the \( LR \) only in the denominator. In fact, as asserted more generally in Chapter 5, the \( W \) and \( LR \) tests, beyond having the same limiting distribution, are asymptotically identical in the sense that they will accept and reject together, under the null, in large samples.

### 16.4.4 Lagrange Multiplier Test

The Lagrange multiplier test examines the unrestricted scores (first derivatives of the log-likelihood function) evaluated at the restricted estimates. If the restriction is correct then the average of the scores will converge in probability to zero and appropriately normalized be asymptotically normal. For this model we have

\[ \frac{\partial \ln L(\bar{\theta})}{\partial \theta} = -\frac{1}{2\sigma^2} \frac{\partial \psi_n(\bar{\theta})}{\partial \theta}. \]

For simplicity, we consider the case where the restrictions have been rewritten in zero form as discussed above, so \( \theta = (\hat{\theta}_1', 0')' \).
Now, by definition, the restricted estimates yield
\[ 0 = \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_1} \]
\[ = \frac{1}{\sigma^2} \sum_{t=1}^{n} (y_t - \check{h}_t(\tilde{\theta})) \frac{\partial \check{h}_t(\tilde{\theta})}{\partial \theta_1}, \]
so for the LM statistic, we only consider the final \( r \) elements of the score
\[ \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_2} = \frac{1}{\sigma^2} \sum_{t=1}^{n} (y_t - h_t(\bar{\theta})) \frac{\partial h_t(\bar{\theta})}{\partial \theta_2}. \]
Premultiplying both of the above scores by \( 1/\sqrt{n} \) and expanding around \( \theta_1^0 \) yields respectively
\[ 0 = \frac{1}{\sigma^2} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \frac{\partial h_t(\theta^0)}{\partial \theta_1} \right] 
+ \frac{1}{n} \sum_{t=1}^{n} \left( u_t \frac{\partial^2 h_t(\theta^0)}{\partial \theta_1 \partial \theta_1} \frac{\partial h_t(\theta^0)}{\partial \theta_1} \right) \sqrt{n}(\check{\theta}_1 - \theta_1^0) + o_p(1) \]
\[ = \frac{1}{\sigma^2} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \frac{\partial h_t(\theta^0)}{\partial \theta_1} - \frac{1}{\sigma^2} Q_{11} \sqrt{n}(\check{\theta}_1 - \theta_1^0) + o_p(1) \]
and
\[ \frac{1}{\sigma^2} \frac{1}{\sqrt{n}} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_2} = \frac{1}{\sigma^2} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \frac{\partial h_t(\theta^0)}{\partial \theta_2} \right] 
+ \frac{1}{n} \sum_{t=1}^{n} \left( u_t \frac{\partial^2 h_t(\theta^0)}{\partial \theta_2 \partial \theta_1} \frac{\partial h_t(\theta^0)}{\partial \theta_1} \right) \sqrt{n}(\check{\theta}_1 - \theta_1^0) + o_p(1) \]
\[ = \frac{1}{\sigma^2} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \frac{\partial h_t(\theta^0)}{\partial \theta_2} - \frac{1}{\sigma^2} Q_{21} \sqrt{n}(\check{\theta}_1 - \theta_1^0) + o_p(1). \]
Solving the first for \( \sqrt{n}(\check{\theta}_1 - \theta_1^0) \) and substituting into the second yields
\[ \frac{1}{\sqrt{n}} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_2} = \frac{1}{\sigma^2} \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \frac{\partial h_t(\theta^0)}{\partial \theta_2} \right] 
+ \frac{1}{n} \sum_{t=1}^{n} \left( u_t \frac{\partial^2 h_t(\theta^0)}{\partial \theta_2 \partial \theta_1} \frac{\partial h_t(\tilde{\theta})}{\partial \theta_1} \right) \sqrt{n}(\check{\theta}_1 - \theta_1^0) + o_p(1) \]
\[ \rightarrow_d N(0, \frac{1}{\sigma^2} \overline{Q}_{22}) \]
where \( \overline{Q}_{22} = (Q_{22} - Q_{21}Q_{11}^{-1}Q_{21}) \). And
\[ \frac{1}{n} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_2'} \sigma^2 \overline{Q}_{22}^{-1} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_2'} \rightarrow_d \chi^2_{r}. \]
For $\tilde{Q}_{22} = \frac{1}{n} \tilde{H}_2^T \tilde{M}_1 \tilde{H}_2 \rightarrow_p Q_{22}$ and tilde indicates the restricted estimates were used to evaluate the matrix, consider the feasible analog

$$\frac{1}{n} \partial \ln L(\hat{\theta}) \sigma^2 \tilde{Q}_{22}^{-1} \partial \ln L(\hat{\theta}) = \frac{1}{n} \frac{1}{\sigma^2} u^T \tilde{M}_1 \tilde{H}_2 (\tilde{H}_2^T \tilde{M}_1 \tilde{H}_2)^{-1} \frac{1}{\sigma^2} \tilde{H}_2^T \tilde{M}_1 u$$

$$= \frac{1}{\sigma^2} [u^T \tilde{M}_1 \tilde{H}_2 \tilde{H}_2^T \tilde{M}_1 \tilde{H}_2 - 1 \tilde{H}_2^T \tilde{M}_1 u]$$

$$= \frac{1}{\sigma^2} [u^T \tilde{u} - \hat{u}^T \hat{u}] + o_p(1)$$

under the null hypothesis. The leading term is the same as that of the likelihood ratio statistic.

### 16.5 Example

We consider the estimation of a production function. Specifically, we consider the generic problem

$$Q_i = h(K_i, L_i, u_i; \theta)$$

where $Q_i$ is firm output, $K_i$ is firm input of capital services, $L_i$ is firm input of labor services, $u_i$ is a random error, and $\theta$ is the parameter vector. The data for the problem are 25 observations taken from Kmenta.

We first entertain the possibility that the specification should be Cobb-Douglas with multiplicative errors. In this case, we have

$$Q_i = \theta_1 K_i^{\theta_2} L_i^{\theta_3} e^{u_i}$$

or in log-linear form

$$q_i = a + \alpha k_i + \beta l_i + u_i$$

where $q_i = \ln Q_i$, $k_i = \ln K_i$, and $l_i = \ln L_i$. The regression results for this data are:

$$q_i = 2.4811 + 0.6401 k_i + 0.2573 l_i + e_i.$$

Since there is a possible issue with respect to whether the error is multiplicative or additive, and hence the possibility of heteroskedasticity, we calculate the White-test for the regression and obtain $W = 3.1106$. This yields a $p$-value of .68294 and there seems to be no evidence that heteroskedasticity impacts the inferences.

It is of interest in any production function whether constant returns to scale apply. For this specification, this implies $\alpha + \beta = 1$. For the log-linear model this can best be tested following a reparameterization:

$$(q_i - l_i) = a + \alpha (k_i - l_i) + (\alpha + \beta - 1) l_i + u_i$$

$$= a + \alpha (k_i - l_i) + \beta^* l_i + u_i.$$
The null hypothesis of $\alpha + \beta = 1$ becomes $\beta^* = 0$ for the reparameterized model. The regression results for this model are

$$(q_i - l_i) = 2.4811 + 0.6401(k_i - l_i) - 0.1026l_i + u_i.$$ 

The $t$-ratio for $\beta^*$ for testing its' being zero is -2.037, which is significant if we consult the asymptotically appropriate standard normal tables. If we consult the tables for the $t$-distribution it is not quite significant. So there is marginal evidence of non-constant returns to scale.

Now we turn to the Cobb-Douglas model with additive errors:

$$Q_i = AK_\alpha L_\beta^* + u_i.$$ 

Estimation by NLS yields

$$Q_i = 11.366K_i^{0.6307}L_i^{2.753} + e_i$$

with $SSE = 1258.06$ and the standard errors of the coefficient estimates respectively (1.620, 0.380, 0.0282). Note that coefficient estimates and the $t$-ratios for $\alpha$ and $\beta$ are very similar to those for the log-linear model. The White-test yields $W = 10.087$, which has a p-value of .0728. This is not significant but does indicate that there is weak evidence of heteroskedasticity and hence the non-linear model might have a misspecified error term. On this basis one might prefer the log-linear model.

A test for constant returns to scale can be also be conducted for the non-linear model. We reparameterize to obtain

$$Q_i = A(K_i/L_i)^\alpha L_i^{\alpha + \beta} + u_i$$

$$= A(K_i/L_i)^\alpha L_i^{\beta^*} + u_i.$$ 

The null that $\alpha + \beta = 1$ therefore becomes $\alpha^* = 1$. Performing this regression yields the same estimates and standard errors for $A$ and $\alpha$, while the estimate and standard error for $\beta^*$ are 0.9060 and 0.05268. A $t$-ratio for testing the null that $\beta^* = 1$ is $-1.7845$, which is not significant for a two-sided alternative.

Finally, we consider estimation of the more general CES production function. Specifically, we consider

$$Q_i = \theta_1(\theta_2K_i^{\theta_3} + (1 - \theta_2)L_i^{\theta_3})^{\theta_4/\theta_3} + u_i.$$ 

Estimation by NLS yields

$$Q_i = 11.213(0.4053K_i^{-0.5963} + (1 - 0.4053)L_i^{-0.5963})^{-0.82718/0.5963} + e_i$$

with $SSE = 979.958$ and the standard errors of the coefficient estimates respectively (1.426, 0.1436, 0.3017, 0.0560). The White test is 16.93, which is marginally significant and indicates that heteroskedasticity may be compromising the inferences on the coefficients.
Constant returns to scale for this model implies $\theta_3 = 1$, which is not rejected when we calculate the appropriate $t$-ratio. The CES approaches the Cobb-Douglas model in the limit as $\theta_3 \to 0$. The CES is not well-defined in this limiting case, however, some indication of which model works best can be obtained by forming the likelihood-ratio statistic. We have

$$LR = n \ln \left( \frac{\hat{\sigma}^2}{\sigma^2} \right)$$

$$= 25 \ln \left( \frac{57.1844}{46.665} \right)$$

$$= 5.082$$

which is significant at the 5% level for a $\chi^2_1$, which is the appropriate limiting distribution under the null. This indicates that the CES is preferable to the Cobb-Douglas.

16.A Appendix

We will first prove consistency by verifying the conditions of Theorem 4.5 from the Appendix of Chapter 4. For purposes of the proof below, we define $\psi_n(\theta) = n^{-1} \sum_{t=1}^{n} (y_t - h(x_t, \theta))^2$ and $\psi_0(\theta) = E[(y - h(x, \theta))^2]$ for $\theta \in \Theta$.

Proof of Theorem 16.1. Now (C1), compactness of $\Theta$, is assured by condition (i). Adding and subtracting $h(x_t, \theta^0)$ inside the square for $\psi_0(\theta)$, expanding the square, and taking expectations yields

$$\psi_0(\theta) = E[(y - h(x, \theta^0)) + (h(x, \theta^0) - h(x, \theta))^2]$$

$$= \sigma^2 + E[(h(x, \theta^0) - h(x, \theta))^2].$$

Identification (C4) follows since, by condition (iv) the expectation will be zero only if $\theta = \theta^0$. And (C3), continuity of $\psi_0(\theta)$, is assured since continuity of $h(x, \theta)$ and condition (iii) mean Newey and McFadden’s Lemma 2.4, presented in the Appendix to Chapter 5, is satisfied for $a(z, \theta) = (h(x, \theta^0) - h(x, \theta))^2$.

This leaves us with the uniform convergence condition (C2) to verify. Adding and subtracting $h(x_t, \theta^0)$ inside the square in $\psi_n(\theta)$ and expanding the square yields

$$\psi_n(\theta) = n^{-1} \sum_{t=1}^{n} [(y_t - h(x_t, \theta^0)) + (h(x_t, \theta^0) - h(x_t, \theta))^2]$$

$$= n^{-1} \sum_{t=1}^{n} u_t^2 - n^{-1} 2 \sum_{t=1}^{n} u_t (h(x_t, \theta) - h(x_t, \theta^0))$$

$$+ n^{-1} \sum_{t=1}^{n} (h(x_t, \theta^0) - h(x_t, \theta))^2.$$
Thus, we have

\[
\psi_n(\theta) - \psi_0(\theta) = n^{-1} \sum_{t=1}^{n} (u_t^2 - \sigma^2) + n^{-1} 2 \sum_{t=1}^{n} u_t (h(x_t, \theta) - h(x_t, \theta^0)) \\
+ n^{-1} \sum_{t=1}^{n} \{(h(x_t, \theta^0) - h(x_t, \theta))^2 - E[(h(x, \theta^0) - h(x, \theta))^2] \} \\
= n^{-1} \sum_{t=1}^{n} g_t + n^{-1} \sum_{t=1}^{n} b_t + n^{-1} \sum_{t=1}^{n} c_t
\]

where the definitions of \(g_t, b_t, \) and \(c_t\) are obvious. Now, by the triangle inequality,

\[
|\psi_n(\theta) - \psi_0(\theta)| \leq \left| n^{-1} \sum_{t=1}^{n} g_t \right| + \left| n^{-1} \sum_{t=1}^{n} b_t \right| + \left| n^{-1} \sum_{t=1}^{n} c_t \right|
\]

and

\[
\sup_{\theta \in \Theta} |\psi_n(\theta) - \psi_0(\theta)| \leq \sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} g_t \right| + \sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} b_t \right| + \sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} c_t \right|
\]

Clearly, the first term converges in probability to zero since it is not a function of \(\theta\) and is the average of an \(i.i.d.\) variable with expectation zero. And the third term converges in probability to zero by continuity and condition (iii) and the Lemma as above.

This leaves the second term. Let \(a(z_t, \theta) = 2u_t (h(x_t, \theta) - h(x_t, \theta^0)) = b_t\), which is continuous at each \(\theta \in \Theta\) for \(z_t = (x_t', u_t)'.\) Now, by condition (iii),

\[
|a(z, \theta)| = |2u(h(x, \theta) - h(x, \theta^0))| \leq \left| 2u(d(x)) \right|^{1/2} = d'(x)
\]

for all \(\theta \in \Theta\). And, by condition (iii), \(d'(x)^2 = 4u^2d(x))\) has finite unconditional expectation \(E[d'(x)^2] = E[4u^2d(x)] = 4\sigma^2 E[d(x)]\), since \(E[u^2|x| = \sigma^2\). But \(E[d'(x)^2] < \infty\) implies \(E[d'(x)] < \infty\) and all the conditions of the Lemma are met. Therefore, since \(E[a(z, \theta)] = 0,\)

\[
\sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} a(z_t, \theta) - E[a(z, \theta)] \right| = \sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} b_t \right| \xrightarrow{p} 0,
\]

and the second term also converges in probability to zero. Thus \(\sup_{\theta \in \Theta} |\psi_0(\theta) - \psi_0(\theta)| \xrightarrow{p} 0\), which verifies uniform convergence.

We will now prove the asymptotic limiting normality of the nonlinear regression estimator by verifying the conditions of Theorem 4.6. In the following, \(\nabla_\theta\) indicates the gradient operator which yields the first derivative of its object with respect to \(\theta\) and \(\nabla_{\theta\theta}\) yields the second derivative matrix of a scalar object.
Proof of Theorem 16.2. Consistency and interiority (N1) are assured directly by conditions (i) and (ii) of Theorem 16.2. And local twice continuous differentiability (N2) follows from condition (iii). For the nonlinear regression problem we have

\[ \nabla_{\theta} \psi_n(\theta) = -\frac{2}{n} \sum_{t=1}^{n} \frac{\partial h_t}{\partial \theta} (y_t - h_t(\theta)) \]

so by i.i.d., \( E[u|x] = 0 \), and \( E[u|x] = \sigma^2 \) we have

\[ \sqrt{n} \nabla_{\theta} \psi_n(\theta^0) = -\frac{2}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial h_t}{\partial \theta} u_t \xrightarrow{d} N(0, 4\sigma^2Q) \]

so (N5) is satisfied with \( V = 4\sigma^2Q \). Furthermore, we have

\[ \nabla_{\theta\theta} \psi_n(\theta) = \frac{n}{n-1} \sum_{t=1}^{n} \frac{\partial^2 (y_t - h(x_t, \theta))^2}{\partial \theta \partial \theta'} \]

and for \( a(z_t, \theta) = \frac{\partial^2 (y_t - h(x_t, \theta))^2}{\partial \theta \partial \theta'} \) the conditions of the Lemma are satisfied locally by i.i.d., (iii), and (iv), whereupon

\[ \sup_{\theta \in N} \| \nabla_{\theta\theta} \psi_n(\theta) - \Psi(\theta) \| \xrightarrow{p} 0 \]

for continuous \( \Psi(\theta) = E[\frac{\partial^2 (y_t - h(x_t, \theta))^2}{\partial \theta \partial \theta'}] \) and (N3) is satisfied. Finally, (N4) is satisfied since \( Q = Q(\theta^0) \) nonsingular by (v). Thus all the conditions of Theorem 4.6 are satisfied and since \( \hat{Q} = \Psi(\theta^0) = 2Q \) and \( V = 4\sigma^2Q \) we have the result given in Theorem 16.2.

Proof of Theorem 16.3. In the proof of Theorem 16.1 above, we showed that \( \psi(\theta) \) converges uniformly to \( \psi_0(\theta) = \psi_n(\theta) = \psi_0(\theta) \) so \( \psi_n(\theta) - \psi_0(\theta) \xrightarrow{p} 0 \), whereupon by the Slutsky theorem \( \hat{\sigma}^2 = \psi_n(\theta) \xrightarrow{p} \psi_0(\theta^0) = \sigma^2 \). Let \( a(z_t, \theta) = \frac{\partial h(x_t, \theta)}{\partial \theta} \), which is continuous (ii) of Theorem 16.1, and uniformly bounded for \( \theta \in N \) by the assumptions of this theorem. Then by Lemma 2.4,

\[ \sup_{\theta \in N} \left\| \frac{1}{n} \sum_{t=1}^{n} \frac{\partial h_t}{\partial \theta} \frac{\partial h_t}{\partial \theta'} - Q(\theta) \right\| \xrightarrow{p} 0 \]

and \( Q(\theta) = E \left[ \frac{\partial h_t}{\partial \theta} \frac{\partial h_t}{\partial \theta'} \right] \) continuous. Thus, \( \hat{Q} = Q(\theta^0) \xrightarrow{p} 0 \) and by the Slutsky theorem, \( \hat{Q} \xrightarrow{p} Q(\theta^0) = Q \).
Chapter 17

Discrete Dependent Variable Models

Sometimes we are interested in modeling economic processes where the outcome is dichotomous or binary in nature. For example, the decision by a worker to join a union or not depending on a number of conditions such as union dues, relative salaries, and education level. Another example would be the decision to retire from a job or not depending on age and relative incomes. In fact, any situation where the outcomes are two states that depend on economics factors is in this category. In order to quantify the states, we arbitrarily assign a 1 to one state and a 0 to the other state. These values would be the dependent variables of what statisticians would call a Bernouli process. Note that the integer nature of the dependent variables is different from what we had in mind in the regression models where our leading example had a continuous disturbance and hence continuous dependent variable.

17.1 Linear Probability Model

Suppose we are interested in modeling the response of the state for any observation to the changes in the attributes that characterize the observation. Let $y_i$ denote the dichotomous dependent variable, which takes on a value of 1 or 0, and $x_i$ a $k \times 1$ vector of values which reflect the attributes of interest. The most straightforward approach might seem to be a linear regression as was considered before

$$y_i = x'_i \beta + u_i \tag{17.1}$$

for $i = 1, 2, ..., n$, where $u_i$ is the disturbance term and $\beta$ is a vector of response parameters. Consistent with the standard regression model, we suppose that $y_i$ and $x_i$ are jointly i.i.d. and $E[u_i | x_i] = 0$. Due to the Bernouli nature of the
process, we find
\[
\begin{align*}
\Pr(y_i = 1|x_i) &= \Pr(y_i = 1|x_i) \cdot 1 + \Pr(y_i = 0|x_i) \cdot 0 \\
&= E[y_i|x_i] \\
&= x'_i\beta,
\end{align*}
\]

For obvious reasons, this model is known as the linear probability model and has been widely used due to its seeming simplicity and the easy availability of standard least squares software.

The problem with this approach is that the validity of standard least squares techniques depends on the usual assumptions being met, which is unfortunately not the case here. The most obvious problem is that, due to the integer nature of the dependent variable, the disturbances cannot be normally distributed. This means the favored finite sample approaches using t- and F-distributions are not appropriate. As we found in Chapter 11, non-normality does not present unsurmountable problems provided the sample is sufficiently large and the usual stochastic assumptions are met. Moreover, since \( y_i \) is generated by a Bernoulli process,
\[
[u_i^2|x_i] = \text{var}[y_i|x_i] = x'_i\beta \cdot (1-x'_i\beta)
\]
and the disturbances must, by their very nature, be conditionally heteroskedastic. As with non-normality, heteroskedasticity does not present unsurmountable problems since we learned how to handle such situations, but it does complicate matters.

Most importantly, as a probability measure, \( \Pr(y_i = 1|x_i) \), must satisfy the constraint
\[
1 \geq x'_i\beta \geq 0
\]
for all observations. Although the true coefficients may satisfy this restriction, the corresponding least-squares estimates generally will not. That is, \( x'_i\beta \leq 1 \) and \( 0 \geq x'_i\beta \) for some \( i \) and \( j \) is possible. This occurs because, for the true disturbances, \( u_i = y_i - x'_i\beta \geq 0 \) for \( y_i = 1 \) and \( u_i = y_i - x'_i\beta \leq 0 \) for \( y_i = 0 \). Least squares will systematically alter the values of \( \beta \) so that the squares of some of the more extreme values of \( \epsilon_i = y_i - x'_i\beta \) will be smaller but at the expense of making some of the smaller \( \epsilon_i \) larger with an opposite sign. Thus the OLS estimates will be biased and inconsistent.

This is most easily seen in the bivariate case. Suppose \( \alpha + \beta x_i \) for \( \beta > 0 \) satisfies the condition for all \( x_i \) in the sample, which lie between \( x_l \) and \( x_u \) \((x_l \leq x_u)\). The corresponding values of \( y_i \) will be either \( 0 \) or \( 1 \), with the the \( 0 \)'s predominating for smaller values of \( x_i \) and the \( 1 \)'s predominating for the larger. However they are distributed, the regression line will pass through the middle of the scatter for both the \( 0 \)'s and the \( 1 \)'s and hence have a steeper slope and less positive intercept. This is illustrated in the following diagram.
17.2 Nonlinear Probability Models

17.2.1 Latent Variable Model

A more reasonably behaved model can be motivated through the use of latent or unobserved variables. Consider a yes-no decision by a consumer as to whether to purchase a particular item or not, or the decision by a worker to join the union or not. We might think that the decision depends on the marginal utility gained by the yes decision versus the no decision. If the marginal gain exceeds some threshold then the yes decision is preferable and we will observe such an action. Let $y^*_i$ denote this gain, which results from a linear data generating process such as we utilized for the linear regressions model. Specifically, we have

$$y^*_i = \mathbf{x}'_i \beta^* + u^*_i$$

(17.4)

where $(\mathbf{x}'_i, u^*_i)$ are jointly i.i.d. and $\beta^*$ is the $k \times 1$ vector of parameters of interest. We also assume $u^*_i$ independent of $\mathbf{x}_i$, $E[u^*_i] = 0$, and $E[u^*_i u^*_i] = \sigma^2$.

The problem with this regression model is that $y^*_i$ is not directly observable instead we observe $y_i$ which results from a decision process based on $y^*_i$. This is an example of latent variable models, which play an extremely important role in applied microeconometrics.

If $\beta^*$ includes an intercept then, without loss of generality, we can subsume the threshold into the intercept and so assume the threshold is zero, is which case we observe

$$y_i = \begin{cases} 1 & \text{if } y^*_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(17.5)
Let \( w_i = -u_i^* / \sigma \) then \( w_i \) is \( i.i.d. \) independent of \( x_i \) with \( [w_i] = 0 \), and \( E[w_i^2] = 1 \), whereupon we can rewrite (17.4) as

\[
y_i = \begin{cases} 
1 & \text{if } w_i \leq x_i'(\beta^* / \sigma) \\
0 & \text{otherwise}
\end{cases}
\]

Thus, due to independence,

\[
Pr(y_i = 1|x_i) = Pr(w_i \leq x_i'(\beta^* / \sigma))
= \int_{-\infty}^{x_i'(\beta^* / \sigma)} f_w(w)dw
= F_w(x_i'(\beta^* / \sigma))
\]

where \( F_w(\cdot) \) is the CDF of \( w \). Of course, \( Pr(y_i = 0|x_i) = 1 - F_w(x_i'(\beta^* / \sigma)) \) since the alternatives are mutually exclusive.

### 17.2.2 Nonlinear Regression Model

This finding can be used to write a nonlinear regression model. As above, we can write

\[
E[y_i|x_i] = Pr(y_i = 1|x_i) \cdot 1 + Pr(y_i = 0|x_i) \cdot 0
= F_w(x_i'(\beta^* / \sigma)).
\]

Unexpected this equation yields

\[
y_i = F_w(x_i'(\beta / \sigma) + u_i \quad (17.6)
\]

for \( i = 1, 2, ..., n \), where \( \beta = \beta^* / \sigma \) and \( u_i = y_i - F_w(x_i'(\beta / \sigma)) \), which by construction has zero conditional mean. The regression model given in (17.5) is an example of a single-index model, which may arise in other contexts. The conditional expectation of \( y_i \) depends on \( x_i \) only through the scalar linear function \( x_i'(\beta / \sigma) \) feeding into the nonlinear CDF function \( F_w(\cdot) \).

Note that this regression solves the need to satisfy the constraint \( 1 \geq Pr(y_i = 1|x_i) \geq 0 \) by construction since \( F_w(\cdot) \) is a CDF. Also note that we only see \( \beta^* \) in ratio with \( \sigma \), which means that the coefficients are only identifiable up to a scale factor. This is a normalization of the CDF to reflect unit variance. In the absence of the normalization, the parameter vector \( \beta^* \) and standard error \( \sigma \) are not separately identifiable.

One might be tempted to apply the results from the previous chapter at this point and estimate the model and perform inference via nonlinear least squares. Unfortunately we are still left with the problem of heteroskedasticity since

\[
E[w_i^2] = [w_i^2|x_i] = \text{var}[y_i|x_i] = F_w(x_i'(\beta / \sigma) \cdot (1 - F_w(x_i'(\beta / \sigma))). \quad (17.7)
\]

This means that any inferences based on the standard statistic yielded by NLS will be suspect. Although we could possible obtain robust standard errors or
Theorem 17.1. Suppose (i) \( B \) compact, (ii) \( Q = E[xx'] \) nonsingular, and (iii) \( F_w(\cdot) \) everywhere continuous and strictly monotonically increasing, then \( \beta \xrightarrow{p} \beta^0 \).

Proof. The proof proceeds by verifying the four conditions (C1)-(C4) of Theorem 4.5. The compactness condition (C1) is assured by Assumption (i) of the current theorem. For the present problem, we have \( \psi_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - F_w(x'_i \beta))^2 \) and

\[
\psi(\beta) = E[(y - F_w(x' \beta))^2]
= E[((y - F_w(x' \beta) - F_w(x' \beta^0))^2]
= E[F_w(x' \beta^0) \cdot (1 - F_w(x' \beta^0))] + E[(F_w(x' \beta) - F_w(x' \beta^0))^2].
\]

The first term does not depend on \( \beta \) and the second attains a minimum of zero at \( \beta = \beta^0 \). By Assumption (ii) of the current theorem, for any \( c \neq 0 \), \( c'Q = E[c'xx'] = \int c'xx'f_x(x)dx \neq 0 \), which means \( c'x \neq 0 \) for \( x \) with positive probability measure. Let \( c = (\beta - \beta^0) \), then \( \beta \neq \beta^0 \) implies, for some \( x \) with positive measure, \( x' \beta \neq x' \beta^0 \) and hence \( F_w(x' \beta) \neq F_w(x' \beta^0) \) due to strict monotonicity from Assumption (iv). It follows that \( (F_w(x' \beta) - F_w(x' \beta^0))^2 > 0 \) for some \( x \) with positive measure so \( E[(F_w(x' \beta) - F_w(x' \beta^0))^2] > 0 \) and \( \psi_0(\beta) \) attains a unique minimum at \( \beta = \beta^0 \), which means the identification condition (C4) is satisfied. This leaves us with the uniform convergence condition (C3) and the continuity condition (C3) to verify. Let \( a(z, \beta) = (y - F_w(x' \beta))^2 \) then \( a(z, \beta) \) everywhere continuous by assumption (iii) and \( a(z, \beta) \leq 1 \) by construction. Thus the conditions of Newey and McFadden’s Lemma 2.4, presented in the Appendix to Chapter 5, are satisfied whereupon \( \psi_0(\beta) = E[a(z, \beta)] = E[(y - F_w(x' \beta))^2] \) is continuous and sup_{\beta \in B} \[ n^{-1} \sum_{i=1}^{n} a(z_i, \beta) - E[a(z, \beta)] \] \xrightarrow{p} 0, so conditions (C3) and (C4) are satisfied.

Assumption (ii) is the same as introduced for identification in the stochastic regressor version of the linear least squares case so has no added cost. Assumptions (iii) and (iv) on the CDF are met by the most common choices of distribution for the latent disturbances: the normal and logistic.

17.2.3 Weighted Nonlinear Least Squares (WNLS)

Given its consistency, it is straightforward to develop the asymptotic normality of the NLS estimator of the binomial probability model. Since the errors are
heteroskedastic, we should utilize a robust covariance estimator, which corrects the standard errors for the heteroskedasticity, in conducting inference. Rather than work out the details of such an approach, we will correct for heteroskedasticity directly in the estimation step as with weighted least squares. Suppose that

$$
\lambda_i^2 = E[u_i^2|x_i] = F_w(x_i'\beta') \cdot (1 - F_w(x_i'\beta'))
$$

were known then we can transform the regression equation (17.5) to obtain the transformed model

$$
\frac{1}{\lambda_i} y_i = \frac{1}{\lambda_i} F_w(x_i'\beta) + \frac{1}{\lambda_i} u_i
$$

or

$$
\tilde{y}_i = \tilde{F}_w(x_i'\beta) + \tilde{u}_i
$$

where \( \tilde{y}_i = \frac{1}{\lambda_i} y_i \), \( \tilde{F}_w(x_i'\beta) = \frac{1}{\lambda_i} F_w(x_i'\beta) \), and \( \tilde{u}_i = \frac{1}{\lambda_i} u_i \). The transformed disturbance \( \tilde{u}_i \) will have zero mean and unit conditional variance and hence be homoskedastic. We then proceed to perform NLS of this transformed model, which should have the same properties as the least squares estimator above under similar assumptions.

The complication with this approach, of course, is that we don’t know \( \lambda_i^2 \) since \( \beta \) is the target of our estimation exercise. It seems natural to proceed with a two-step approach at this point. In the first step we perform NLS to obtain \( \hat{\beta} \), which is then used to calculate

$$
\hat{\lambda}_i^2 = F_w(x_i'\hat{\beta}) \cdot (1 - F_w(x_i'\hat{\beta}))
$$

and the feasible transformed model

$$
\hat{\tilde{y}}_i = \hat{F}_w(x_i'\hat{\beta}) + \hat{\tilde{u}}_i
$$

where \( \hat{\tilde{y}}_i = \frac{1}{\hat{\lambda}_i} y_i \), \( \hat{F}_w(x_i'\hat{\beta}) = \frac{1}{\hat{\lambda}_i} F_w(x_i'\hat{\beta}) \), and \( \hat{\tilde{u}}_i = \frac{1}{\hat{\lambda}_i} u_i \). We now apply NLS taking \( \hat{\lambda}_i \) as given, to obtain \( \hat{\beta} \) as the solution to the first-order conditions

$$
0 = -2 \sum_{i=1}^{n} \frac{1}{\hat{\lambda}_i} (y_i - \hat{F}_w(x_i'\hat{\beta})) \frac{\partial \hat{F}_w(x_i'\hat{\beta})}{\partial \beta}
$$

$$
= -2 \sum_{i=1}^{n} \frac{1}{\hat{\lambda}_i^2} (y_i - F_w(x_i'\hat{\beta})) \frac{\partial F_w(x_i'\hat{\beta})}{\partial \beta}
$$

$$
= -2 \sum_{i=1}^{n} \frac{1}{F_w(x_i'\hat{\beta}) \cdot (1 - F_w(x_i'\hat{\beta}))} (y_i - F_w(x_i'\hat{\beta})) f_w(x_i'\hat{\beta}) x_i
$$

since \( \frac{\partial F_w(x_i'\beta)}{\partial \beta} = f_w(x_i'\beta) x_i \) where \( f_w(\cdot) \) is the PDF of \( w \). This estimator should be asymptotically efficient relative to the NLS estimator above and equivalent to an estimator based on the true values of \( \lambda_i^2 \). There might be some appeal for using the new more efficient estimator to recalculate \( \hat{\lambda}_i^2 \), \( \hat{\tilde{y}}_i \), \( \hat{F}_w \), and \( \hat{\tilde{u}}_i \) and
17.2. NONLINEAR PROBABILITY MODELS

We will bypass the details of the asymptotic behavior of these approaches for reasons that become obvious in a few paragraphs.

17.2.4 Maximum Likelihood Estimation (MLE)

Given that we have completely specified the distributional structure of the model it is natural to develop its maximum likelihood estimator. Since \( \Pr(y_i = 1|x_i) = F_w(x_i'\beta) \) and \( \Pr(y_i = 0|x_i) = 1 - F_w(x_i'\beta) \) then the joint probability structure of all the observations or the likelihood function is given by

\[
L_n(\beta) = \left( \prod_{y_i=1} F_w(x_i'\beta) \right) \left( \prod_{y_i=0} [1 - F_w(x_i'\beta)] \right)
\]

and the log-likelihood is

\[
\mathcal{L}_n(\beta) = \frac{1}{n} \ln(L(z)) = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i \ln(F_w(x_i'\beta)) + (1 - y_i) \ln(1 - F_w(x_i'\beta)) \right]
\]

(17.9)

where \( f(y|x; \beta) = F_w(x'\beta)^y [1 - F_w(x'\beta)]^{1-y} \) is the density of \( y \) given \( x \). The maximum likelihood estimator is taken as the solution to the first-order conditions

\[
0 = \frac{\partial \mathcal{L}(z; \hat{\beta})}{\partial \beta}
\]

(17.10)

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ y_i \frac{1}{1 - F_w(x_i'\beta)} f_w(x_i'\beta)x_i - (1 - y_i) \frac{1}{F_w(x_i'\beta)} f_w(x_i'\beta)x_i \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{y_i}{F_w(x_i'\beta)(1 - F_w(x_i'\beta))} (1 - F_w(x_i'\beta)) f_w(x_i'\beta)x_i - \frac{1 - y_i}{F_w(x_i'\beta)(1 - F_w(x_i'\beta))} F_w(x_i'\beta) f_w(x_i'\beta)x_i \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{F_w(x_i'\beta)(1 - F_w(x_i'\beta))} (y_i - F_w(x_i'\beta)) f_w(x_i'\beta)x_i.
\]

Comparison with (17.7) reveals that \( \hat{\beta} \) solving these equations will be the same as the weighted NLS estimator resulting when we iterate to convergence. So the
iterated weighted NLS approach is a feasible numerical approach for obtaining the MLE of this model.

In order to establish general consistency of MLE we need to add a uniform boundedness in probability condition, which will be shown to be superfluous for the specific Logit and Probit cases considered in the next section.

**Theorem 17.2.** Suppose (i) \( B \) compact, (ii) \( Q = E[xx'] \) nonsingular, and (iii) \( F_w(\cdot) \) everywhere continuous and strictly positive in a neighborhood \( N \) of \( \partial f \) and (ii) above. By (iii) above both \( u \) and \( \partial f \) are integrable functions of functions of \( \beta \) with a suitable redifinition of terms.

Finally, (iv) and (v) above are simply conditions (v) and (vi) of Theorem 5.2

### 17.3 Probit and Logit

In order to proceed with any of the above approaches we have to specify the density of the latent disturbance. An obvious choice is to assume \( u^* \sim N(0, \sigma^2) \) so \( w \sim N(0,1) \), whereupon

\[
f_w(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{17.11}
\]

\[
F_w(\cdot) = \Phi(\cdot) = .5[1 + \text{erf}(\frac{z}{\sqrt{2}})]
\]
where $\varphi(z)$ and $\Phi(z)$ denote the standard normal PDF and CDF. The nonlinear regression model based on the standard normal CDF is called the Probit model. An attractive alternative is to assume $u^* \sim \text{Logistic}(0, \sigma^2)$ so $w \sim \text{Logistic}(0, 1)$, whereupon

$$f_w(w) = \frac{e^{w/k}}{k(1 + e^{w/k})^2}$$

$$F_w(z) = \frac{1}{1 + e^{-w/k}}$$

for $k = \sqrt{3}/\pi$. We will call the regression model based on this second formulation the standard Logit model.

The standard logistic density is bell-shaped like the normal but slightly more peaked and thicker-tailed but thinner in the shoulders as seen in Figure 17.2.

For our current purposes, interest naturally focuses on the CDF’s of the distributions which are even more similar than the densities, as seen in Figure 17.3. In consequence of this similarity of CDF’s, we might expect the estimates to be similar from the two alternatives, when they are normalized in the same fashion. A frequent justification for using the Logit approach rather than Probit is that the CDF is available in closed form although this rationalization may be called into question.
Other choices of normalization for the CDF and hence $\beta^*$ are possible, of course, which will rescale the parameter vector. An alternative scaling of the logistic is to define $w_i = -(\pi/\sqrt{3}) u_i^* / \sigma$, whereupon the above density and CDF will still apply but now $k = 1$ and $\beta = \beta^* \cdot (\pi/\sqrt{3}) / \sigma$. This is by far the most commonly used scaling for the logistic distribution and we will call it the common Logit model. Note that the estimates and their standard errors will be rescaled as a result of this choice. Specifically, they will be $(\pi/\sqrt{3}) = 1.8138$ times larger. The likelihood will, of course, be the same for either scaling as will the usual ratios used for inference.

Aside from the natural appeal of using normally distributed disturbances in the latent variable model with the Probit approach and the analytic tractability of the Logit approach, both approaches yield much simpler conditions for consistency and asymptotic normality.

**Theorem 17.4.** Suppose the model is either Logit or Probit and (i) $B$ compact, (ii) $Q = \text{E}[xx']$ nonsingular, then $\hat{\beta} \xrightarrow{p} \beta^0$.

**Proof.** For both the Logit and Probit models, condition (iii) of Theorem 17.2 is assured by the functional form and we show below that this along with assumption (ii) implies the boundedness condition (iv) of Theorem 17.2 is met, so we only need to verify assumptions (i) and (ii). For the Logit, we use the mean value expansion of $F_w(w)$ about $w = x'\beta = 0$ to obtain

$$
\ln(F_w(x'\beta)) = \ln(F_w(0)) + \lambda(x'\bar{\beta})x'\beta
$$

where $1 \geq \lambda(w) = f_w(w)/F_w(w) = 1 - F_w(w) \geq 0$ and $\bar{\beta} = k\beta$ for $0 \leq k \leq 1$. 

Figure 17.3: Standard Normal and Logistic Distributions
By the triangle inequality, we have

\[
\ln(F_w(x'z)) \leq \ln(F_w(0)) + \left| \lambda(x'z) \right| = \ln(F_w(0)) + \lambda(x'z)
\]

since \(\lambda(w) \geq 0\)

\[
\leq \ln(F_w(0)) + |x'z| \quad \text{since } 1 \geq \lambda(w)
\]

\[
\leq \ln(F_w(0)) + \|x\| \|z\| \quad \text{by the Cauchy-Schwarz inequality}
\]

Since the Logistic is symmetric \(1 - F_w(w) = F_w(-w)\) has the same bounding function. Thus, since \(0 \leq y \leq 1\) and likewise \(1 - y\), then

\[
\ln f(y|x; \beta) = |y \ln(F_w(x'z)) + (1 - y) \ln(1 - F_w(x'z))|
\]

\[
\leq 2(\ln(F_w(0)) + \|x\| \|z\|)
\]

\[
\leq 2(\ln(F_w(0)) + b \|x\|) = d(z)
\]

where \(b = \sup_{z \in \mathcal{B}} \|z\| \geq 0\) exists since \(\mathcal{B}\) compact and \(\|z\|\) continuous. And existence of second moments of \(x\) assure this bounding function will have a finite expectation.

For the Probit model, we use the same mean value expansion to obtain

\[
\ln(\Phi(x'z)) = \ln(\Phi(0)) + \lambda(x'z)x'z
\]

where \(\lambda(w) = \varphi(w)/\Phi(w) \geq 0\) and \(\bar{z} = k\beta\) for \(0 \leq k \leq 1\). The \(\lambda(w)\) is known as the inverse Mills ratio and is well studied. In particular, it is convex and asymptotes to 0 as \(w \to \infty\) and \(|w|\) as \(w \to -\infty\). Consequently, it is easy to see that \(\lambda(w) \leq 1 + |w|\), since \(\lambda(0) = .798\). Using the triangle inequality and replicating the above arguments, we have

\[
\ln(\Phi(x'z)) \leq \ln(\Phi(0)) + \lambda(x'z)|x'z|
\]

\[
\leq \ln(\Phi(0)) + (1 + \left| x'z \right|) |x'z|
\]

\[
\leq \ln(\Phi(0)) + (1 + \|x\| \|z\|) \|x\| \|z\| \quad \text{by Cauchy-Schwarz}
\]

\[
\leq \ln(\Phi(0)) + (1 + \|x\| \|z\|) \|x\| \|z\| \quad \text{since } \bar{z} = k\beta \text{ and } 0 \leq k \leq 1
\]

And, again, due to symmetry we have

\[
\ln f(y|x; \beta) = |y \ln(\Phi(x'z)) + (1 - y) \ln(1 - \Phi(x'z))|
\]

\[
\leq 2(\ln(\Phi(0)) + \|x\| \|z\|) \|x\| \|z\|
\]

\[
\leq 2(\ln(F_w(0)) + (1 + b \|x\|)b \|x\|) = d(z)
\]

where \(b\) is the same as for the Logit. Existence of second moments for \(x\) insure that this bounding function has finite expectation.

**Theorem 17.5.** Suppose the model is either Logit or Probit and
\((i)\) \(\beta_0\) interior \(\mathcal{B}\), which is compact, and \((ii)\) \(\mathcal{Q} = \mathbb{E}[xx']\) nonsingular, then \(\sqrt{n}(\beta - \beta_0) \longrightarrow_d \mathcal{N}(0,\sigma^{-1})\).
Proof. By Theorem 17.4 assumptions (i) and (ii) imply the consistency and interiority of condition (i) of Theorem 17.3. Obviously, conditions (ii) and (iii) are satisfied for both models due to the functional forms. For both Logit and Probit, it is straightforward to show that

\[
\vartheta(\beta) = E[\partial \ln f(y|x; \beta) / \partial \beta \cdot \partial \ln f(y|x; \beta) / \partial \beta']
\]

\[
= \frac{1}{F_w(x'|\beta)^2(1 - F_w(x'|\beta))^2} (y - F_w(x'|\beta))^2 f_w(x'|\beta)^2 xx' | x
\]

\[
= E\left[ \frac{f_w(x'|\beta)^2}{F_w(x'|\beta)(1 - F_w(x'|\beta))} xx' \right]
\]

\[
= E[\lambda(x'|\beta)\lambda(-x'|\beta) xx']
\]

For both Logit and Probit, \( \lambda(x'|\beta)\lambda(-x'|\beta) \geq 0 \) is bounded from above so the expectation exists. Moreover, since \( \lambda(w)\lambda(-w) \) is bounded away from zero for \( w \) on any open interval, for \( w \), we can show positive definiteness of \( Q = E[xx'] \) implies positive definiteness of \( \vartheta(\beta) = E[\lambda(x'|\beta)\lambda(-x'|xx')] \) and condition (iv) is satisfied. Let \( \lambda_w(w) = d\lambda(w)/dw \), then we can show that

\[
\frac{\partial^2 \ln f(y|x; \beta)}{\partial \beta \partial \beta'} = [y\lambda_w(x'|\beta) + (1 - y)\lambda_w(-x'|\beta)] xx'
\]

and note that \( \lambda_w(x'|\beta) \geq 0 \) is bounded for both Logit and Probit and hence \( [y\lambda_w(x'|\beta) + (1 - y)\lambda_w(-x'|\beta)] \geq 0 \) is also bounded. Thus \( \|[y\lambda_w(x'|\beta) + (1 - y)\lambda_w(-x'|\beta)] xx'\| \leq c ||xx'|| \) for some \( c \), which has finite expectation and condition (v) of Theorem 17.3 is satisfied. \( \square \)
Chapter 18

Generalized Method of Moments

18.1 Method of Moments

18.1.1 Estimation by Analogy

Suppose we are interested in some characteristic or property of the distribution of some random variable $z$. For example, the mean or some quantile such as the median. Typically, given knowledge of the distribution, we can write the property as a function of the parameters of the distribution. For example, the mean of a log-normal distribution is $\exp(\mu + \sigma^2/2)$, where $\mu$ is the location parameter and $\sigma^2$ is the scale. And estimates of the property may be obtained by estimating the parameters of the model and plugging them into the function which defines the property of interest. Given sufficient smoothness to the function and regularity of the distribution, then maximum likelihood estimators of the parameters will be asymptotically efficient and likewise the asymptotically efficient estimator of the property is the function of the maximum likelihood parameter estimators.

A limitation of this approach is the requirement that we know the underlying distribution. If we misspecify the parametric form of the distribution, we may introduce bias and inconsistency into our estimators. An alternative would be to base our estimates on the empirical distribution of the random variable taken from a sample. The empirical distribution takes the observations $(z_1, z_2, \ldots, z_n)$ as points of a discrete distribution with probability $1/n$ assigned to each point. We take the properties of interest of the sample distribution as our estimators. The population mean of this distribution is the sample mean and the population variance is the sample variance. Likewise the population quantiles will be the sample quantiles. This approach is known as estimation by analogy and has a long tradition in statistics.

The essential appeal of this approach is that it will minimize the extent to
which we need to parameterize the distribution. For example, in a regression model, we will not need to specify the distribution of the disturbances or the regressors but just assume certain regularity properties such as i.i.d., existence of moments, and/or boundedness. This is a nonparametric or possibly semiparametric approach which allows us to focus our attention on the part of the model that is of economic interest and not have misspecification of the distribution contaminate our analysis.

18.1.2 Method of Moments Estimation

We consider first a model of an i.i.d. scalar random variable. Suppose we seek to estimate $p$ unknown parameters $\theta' = (\theta_1, \theta_2, ..., \theta_p)$ of the distribution $f_z(z; \theta)$ of the random variable $Z$. The first $p$ population moments of the random variable can be expressed as $p$ known functions of the $p$ parameters

$$
\mu_1 = E[Z] = h_1(\theta_1, \theta_2, ..., \theta_p)
$$
$$
\mu_2 = E[Z^2] = h_2(\theta_1, \theta_2, ..., \theta_p)
$$
$$
\vdots
$$
$$
\mu_p = E[Z^p] = h_p(\theta_1, \theta_2, ..., \theta_p)
$$
or in vector notation, 

$$
\mu = h(\theta).
$$

If this mapping is isomorphic then the inverse function

$$
\theta = h^{-1}(\mu)
$$

exists and the parameters are identified by the moments.

Given a sample of observations $(z_1, z_2, ..., z_n)$, the inverse function suggests an estimator for the parameters. Let

$$
\hat{\mu}_j = \sum_{t=1}^{n} z_t^j
$$

be the sample $j$-th moment and $\hat{\mu}' = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_p)$ then our estimator is

$$
\hat{\theta} = h^{-1}(\hat{\mu}).
$$

This estimator is known as a method of moments estimator. It is not known as “the” method of moments estimator because we might have used other moments than the first $p$ raw moments.

If the inverse function exists and is continuous at the true parameter value $\theta^0$, then this method of moments estimator will be consistent. Given the i.i.d. assumption and existence of the $p$ moments, then the sample moments will be consistent estimates of the corresponding population moments, so $\hat{\mu} \to_p \mu^0$ where $\mu^0$ indicates the true values of the population moments. By the continuity
18.1. METHOD OF MOMENTS

Theorem, the consistency will go through the continuous inverse function, so \( \hat{\theta} = h^{-1}(\hat{\mu}) \rightarrow_{p} h^{-1}(\mu^0) = \theta^0 \).

The above approach would seem to be parametric. However, it is often the case that the distribution can be rewritten to have two sets of parameters: those associated with moments, say \( \theta \), and others associated with the shape of the distribution, say \( \eta \). The moment conditions above may identify the first set of parameters, in which case, the second set will be left unspecified. To the extent that the identified parameters \( \theta \) are the same across all possible distributions, the problem becomes semiparametric with the unspecified shape of the distribution being nonparametric. We will see the distinction in some of the examples below.

18.1.3 Just-Identified Models

This approach generalizes to a vector of random variables and moments of functions of the random variables. Suppose \( z \) is a \( k \times 1 \) vector random variable. We have \( p \) functions of the data and a \( p \times 1 \) vector of unknown parameters of interest, denoted

\[
g_j(z, \theta) = g_j(z_1, z_2, \ldots, z_k; \theta_1, \theta_2, \ldots, \theta_p)
\]

for \( j = 1, 2, \ldots, m \). These functions are assumed to have zero mean at the true value of the parameters so

\[
E[g_j(z, \theta^0)] = 0
\]

for \( j = 1, 2, \ldots, p \). In vector notation, we have

\[
E[g(z, \theta^0)] = 0
\]

where \( g(z, \theta)' = (g_1(z, \theta), g_2(z, \theta), \ldots, g_m(z, \theta)) \). There is no loss of generality by assuming the zero mean since we can always subtract this component and include it in \( g(z, \theta) \). For example in the previous method of moments equations we can subtract \( Z^j \) from \( h_j(\theta) \) in each equation.

Suppose the mean of this function

\[
\gamma(\theta) = E[g(z, \theta)]
\]

exists and \( \gamma(\theta) = 0 \) has a unique solution at \( \theta = \theta^0 \). Since the number of parameters and equations are equal, we say the model is just-identified by the moment restriction. Then a natural estimator of \( \theta^0 \) is provided by the analogous solution \( \hat{\theta} \) of the sample moments

\[
0 = \frac{1}{n} \sum_{t=1}^{n} g(z_t, \hat{\theta})
= \bar{g}_n(\hat{\theta})
\]

where notation is obvious.
The solution, as we will see in the examples below, is sometimes simple and sometimes not so simple. In particular if the zero condition is linear in $\theta$, the solution is straightforward. If the condition is nonlinear then the solution may not be available in closed form and finding it may necessitate iterative procedures. The consistency will be easier to establish in the linear cases as well. General asymptotic behavior for this just-identified approach are covered as a special case of the more general approach studied in the sequel.

18.1.4 Some Examples

As an simple example of the the advantages of considering moments and the complications from nonlinearity, consider the 2-parameter exponential distribution:

$$f(z) = \lambda e^{-\lambda(z-\gamma)}$$

where $\gamma \geq 0$ is the point at which the distribution begins to have non-zero support and $\lambda$ is a scale parameter. This distribution is widely used in life-to-failure analysis in manufacturing. Maximum likelihood might seem an attractive option for estimating the parameters but with one parameter defining the lower boundry of support the usual assumptions are not satisfied and MLE is not the solution to first-order conditions. Consequently, the limiting distributions will not be normal.

An attractive alternative, in this case, is to use moments to identify the parameters and proceed with method of moments. The first two raw moments can be shown to be

$$\mu_1 = \mu = \gamma + \frac{1}{\lambda}$$
$$\mu_2 = \sigma^2 + \mu^2 = \frac{1}{\lambda^2} + \left(\gamma + \frac{1}{\lambda}\right)^2$$

where $\sigma^2$ is the variance. These moment equations have the unique solution

$$\gamma = \mu_1 - \sqrt{\mu_2 - \mu_1^2}$$
$$\lambda = \frac{1}{\sqrt{\mu_2 - \mu_1^2}}.$$ 

The method of moments estimators of $\hat{\theta}' = (\hat{\gamma}, \hat{\lambda})$ are obtained by plugging the sample raw moments into these nonlinear equations for the population moments. These equations, being nonlinear, will have more complicated finite sample properties. The asymptotic limiting behavior, however, is easily obtained using the delta method and the limiting distribution of the raw moments, which is multivariate normal.

As an example of a linear in the parameters, just-identified model, we consider the instrumental variables model from Chapter 11. Our basic stochastic assumption was that $w' = (y, x', z')$ are jointly i.i.d. for $x$ and $z$ ($p \times 1$) vectors
and $0 = E[z u]$ for $u = (y - x' \beta)$. Let $g(w, \theta) = z(y - x' \beta)$ for $\theta$ then

$$0 = E[g(w, \theta)]$$
$$= E[z(y - x' \beta)]$$
$$= E[zy] - E[zx'] \beta$$
$$= \sum_{zy} - \sum_{zx} \beta$$
$$= \gamma(\beta)$$

for $\theta = \beta$, $\sum_{zx} = E[zx']$, and $\sum_{zy} = E[zy]$. If $\sum_{zx} = P$ is nonsingular as assumed in chapter 11, then this is a just-identified problem with solution given by

$$\beta = \sum_{zx}^{-1} \sum_{zy}.$$

This example is semiparametric since the shape parameters of the distributions of $z$, $x$, and $u$ are unspecified. We only require existence of the second moments $\sum_{zx}$ (nonsingular), $\sum_{zy}$, and $\sum_{zu} = E[z u] = 0$.

Using sample analogs, we have the method of moments estimator $\widetilde{\beta}$ as the solution to

$$0 = \frac{1}{n} \sum_{t=1}^{n} g(w_t, \hat{\theta})$$
$$= \frac{1}{n} \sum_{t=1}^{n} [z_t (y_t - x_t' \hat{\beta})]$$
$$= \frac{1}{n} \sum_{t=1}^{n} z_t y_t - \frac{1}{n} \sum_{t=1}^{n} z_t x_t' \hat{\beta}$$
$$= \frac{1}{n} Z' y - \frac{1}{n} Z' X \hat{\beta}$$

where $X$, $Z$, and $y$ are the obvious matrices of observations on $x$, $z$, and $y$. The solution is the IV estimator

$$\tilde{\beta} = \left( \frac{1}{n} Z' X \right)^{-1} \frac{1}{n} Z' y$$
$$= \left( Z' X \right)^{-1} Z' y$$

which is seen as a sample analog and will exist with probability one due to the nonsingularity of $P$. Although $g_n(\hat{\theta})$ is linear in $\hat{\theta} = \beta$, the solution is nonlinear in $x$ and $z$ and the finite-sample distribution becomes complicated. The large-sample asymptotic behavior is considered at length in Chapter 11 and falls out as a special case of the overidentified models considered below.
CHAPTER 18. GENERALIZED METHOD OF MOMENTS

18.2 Generalized Method of Moments (GMM)

18.2.1 Over-Identified Models

Suppose we have more zero mean conditions than the number of parameters being estimated. For example, if we have more than \( p \) instruments in the IV problem above. Let \( m \) denote the length of the vector \( g(z, \theta) \) in the zero mean condition

\[
0 = E[g(z, \theta^0)]
\]

and \( m > p \), then we say the model is over-identified. When \( m = p \), the model is just-identified and the method of moments discussion above applies.

If the elements of \( g(z, \theta) \) are linearly independent with respect to \( \theta \) then the method of moments approach discussed above presents complications. Specifically, we assume the second matrix

\[
\Omega(\theta) = E[g(z, \theta)g(z, \theta)']
\]

is nonsingular, since if the covariance were singular we could eliminate some of the elements of \( g(z, \theta) \) as linear combinations of others.

We can attempt, as before to find \( \hat{\theta} \) that yield \( 0 = \bar{g}_n(\hat{\theta}) \). Unfortunately, this system is over-determined with respect to \( \theta \). Although the mean is zero in the population at \( \theta^0 \), the sample mean will have positive definite covariance \( \Omega(\theta)/n \) and hence

\[
0 \neq \lambda\bar{g}_n(\theta)
\]

with probability one for any \( \lambda \neq 0 \) and any choice of \( \theta \), including even \( \theta^0 \). We can set \( p \) of the sample moments to zero but the probability of any of the remainder being a linear combination of the zero moments is zero.

A possible solution is to discard some of the equations so that \( m = p \) and proceed to obtain a just-identified method of moments estimator. The obvious question with this approach is which equations to discard and which to retain. Plus, we have learned that additional restrictions on the distribution of estimators typically reduces their variability. More will be said about efficiency gains below.

18.2.2 GMM Estimation

Although we cannot set all the sample moments \( \bar{g}_n(\theta) \) to zero at the same time, it is possible to make them all small, in some sense. Note that \( \bar{g}_n(\theta^0) \) has mean zero and covariance \( \Omega/n \), so by convergence in quadratic mean, it will be arbitrarily close to zero with probability one. In choosing \( \theta \) that makes \( \bar{g}_n(\theta) \) small, there will be trade-offs between the elements, making some elements larger and others smaller as we vary our choice of \( \theta \). One approach might be to minimize the sum of squares of \( \bar{g}_n(\theta) \), namely \( \bar{g}_n(\theta)'\bar{g}_n(\theta) \). A more general approach would be to weight these tradeoffs, as in minimizing \( \bar{g}_n(\theta)'W\bar{g}_n(\theta) \), where \( W \) is a positive definite weight matrix.
Of course some choices of $W$ turn out to be better than others. And the optimal ones will have to be estimated. So we consider the more general minimization problem where the weight matrix is a consistent estimator $\hat{W}$ of $W$. Specifically, treating $\hat{W}$ as given, we find $\hat{\theta}$ such that

$$\psi_n(\theta) = \frac{1}{2} \bar{g}_n(\theta)^\prime \hat{W} \bar{g}_n(\theta)$$

is minimized so

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \psi_n(\theta)$$

where, as mentioned above, $\Theta$ is the set from which we are choosing. This is called the generalized method of moments (GMM) approach and was first proposed by Lars Hansen. For mathematical convenience and without loss of generality, we halved the criterion function, since the minimum will be the same whether we halve or not.

As with the least squares models (linear and nonlinear), we find the estimates as the solution to the first-order conditions

$$0 = \frac{\partial \psi_n(\hat{\theta})}{\partial \theta} = \left( \frac{\partial \bar{g}_n(\hat{\theta})}{\partial \theta} \right)^\prime \hat{W} \bar{g}_n(\hat{\theta})$$

$$= \bar{G}_n(\hat{\theta})^\prime \hat{W} \bar{g}_n(\hat{\theta})$$

where $\bar{G}_n(\theta) = \frac{\partial \bar{g}_n(\theta)}{\partial \theta}$. Note that $\bar{g}_n(\theta)$ may be linear in $\theta$, in which case $\bar{G}_n(\theta) = \bar{G}_n$ is not a function of $\theta$ and first order conditions will also be linear in $\hat{\theta}$ and easily obtained. We will consider an IV example below that satisfies this simplification.

The method of moments estimator falls out as a special case of GMM when $m = p$ and the model is just-identified by the moment conditions. In this case, we can choose $\hat{\theta}$ such that $\bar{g}_n(\hat{\theta}) = 0$, which obviously minimizes the criterion function. Note that the weight matrix $\hat{W}$ has no role in determining the estimator for this case, which will be reflected in the limiting covariance results below.

### 18.2.3 Newton-Raphson Method

If $\bar{g}_n(\theta)$ is nonlinear in $\theta$, then $\bar{G}_n(\theta)$ will be a function of $\theta$ and the first-order conditions will be nonlinear in $\hat{\theta}$ and finding a solution may require iterative techniques. In the nonlinear regression case, we linearized the regression function and proceeded to do least squares. Here we will linearize the $\bar{g}_n(\theta)$ function and use the linearized version to do GMM estimation.

First, we expand $\bar{g}_n(\theta)$ in a Taylor series (about some initial estimate $\theta^i$) to obtain

$$\bar{g}_n(\theta) = \bar{g}_n(\theta^i) + \bar{G}_n(\theta^i)(\theta - \theta^i) + R^i$$

where $R^i$ is the remainder term. We linearize by treating $\theta^i$ as given, defining $\delta = (\theta - \theta^i)$, and setting the remainder term to zero to obtain

$$\bar{g}_n^*(\delta) = \bar{g}_n^*(\theta^i) + \bar{G}_n(\theta^i)\delta$$
Using this linearization in the criterion function, we seek to minimize \( \frac{1}{2} \hat{g}_n^\ast(\delta)^\prime \hat{W} \hat{g}_n^\ast(\delta) \) with respect to \( \delta \). Our estimate \( \hat{\delta} \) is given as the solution to the first-order conditions

\[
0 = \left( \frac{\partial \hat{g}_n^\ast(\delta)}{\partial \delta} \right)^\prime \hat{W} \hat{g}_n^\ast(\delta) = \bar{G}_n(\theta^i)^\prime \hat{W} \left[ \hat{g}_n(\theta^i) + \bar{G}_n(\theta^i) \hat{\delta} \right].
\]

Assuming \( \bar{G}_n(\theta^i)^\prime \hat{W} \bar{G}_n(\theta^i) \) has full column rank, then

\[
\hat{\delta} = -\left( \bar{G}_n(\theta^i)^\prime \hat{W} \bar{G}_n(\theta^i) \right)^{-1} \bar{G}_n(\theta^i)^\prime \hat{W} \hat{g}_n(\theta^i).
\]

is the unique solution to the linearized GMM system.

Define \( \theta^{i+1} = \theta^i + \hat{\delta} \), then we can rearrange the solution for the linearized model as the iterative equations

\[
\theta^{i+1} = \theta^i - \left( \bar{G}_n(\theta^i)^\prime \hat{W} \bar{G}_n(\theta^i) \right)^{-1} \bar{G}_n(\theta^i)^\prime \hat{W} \hat{g}_n |_{\theta^i}.
\]

or, since \( \bar{G}_n(\theta^i)^\prime \hat{W} \bar{G}_n(\theta^i) \mid_{\theta^i} \) nonsingular,

\[
0 = \bar{G}_n(\theta^i)^\prime \hat{W} \hat{g}_n \mid_{\theta^i}
\]

and the first order conditions (18.5) are satisfied.

In the literature this procedure for finding a solution to the GMM problem is known as the Newton-Raphson method. Strictly speaking, Newton’s method, as was shown in Chapter 16, yields the iterations

\[
\theta^{i+1} = \theta^i - \left( \frac{\partial^2 \psi_n(\theta^i)}{\partial \theta \partial \theta^i} \right)^{-1} \frac{\partial \psi_n(\theta^i)}{\partial \theta},
\]

assuming the second derivatives exist. Now if \( g(z, \theta) \) is twice continuously differentiable, then for our current problem

\[
\frac{\partial^2 \psi_n(\theta)}{\partial \theta \partial \theta^i} = \left( \frac{\partial^2 \hat{g}_n(\theta)}{\partial \theta^i \partial \theta^j} \right)^\prime \hat{W} \hat{g}_n(\theta) + \bar{G}_n(\theta)^\prime \hat{W} \left( \frac{\partial \hat{g}_n(\theta)}{\partial \theta^i} \right).
\]

When evaluated at \( \theta^0 \), the first term becomes asymptotically negligible in probability, since \( \hat{g}_n(\theta^0) \rightarrow_p 0, \hat{W} \rightarrow_p W, \) and \( \partial^2 \hat{g}_n(\theta^0)/\partial \theta_i \partial \theta_j = O_p(1) \) for \( \text{E}[\partial^2 g(z, \theta^0)/\partial \theta_i \partial \theta_j] \) finite. Thus at the target value \( \theta^0 \) we find

\[
\frac{\partial^2 \psi_n(\theta^0)}{\partial \theta \partial \theta^i} - \bar{G}_n(\theta^0)^\prime \hat{W} \bar{G}_n(\theta^0) \rightarrow_p 0
\]
which suggests using the simpler matrix in the iterations. This Newton-Raphson approach is the analog to the Gauss-Newton procedure for nonlinear least squares where we also eliminated a term in the second derivative matrix since it was asymptotically negligible at the truth.

All of the practical issues identified for the Gauss-Newton procedure for the nonlinear regression model in Chapter 16 apply equally to the current problem when the first-order conditions are nonlinear. Multiple roots or solutions are a potential problem. Convergence may be a problem though we know that convergence is quadratic near the solution. Good preliminary estimates may go a long way toward avoiding both of these problems. And using numeric derivatives, while attractive, can lead to substantial losses in the precision of the estimates. Refer back to Chapter 16 for more information on these issues.

18.2.4 A Simple Over-Identified Example

Consider again the linear instrumental variables problem from Chapter 11, but suppose that \( z \) is an \( m \times 1 \) vector and \( m > p \). Then we have more instruments than regressors and the model is over-identified. For this model we have

\[
\bar{g}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} z_t(y_t - x_t'\beta)
\]

\[
= \frac{1}{n} Z'(y - X\beta)
\]

\[
\bar{G}_n(\theta) = -\frac{1}{n} Z'X
\]

for \( \theta = \beta \). For some choice of nonsingular \( \hat{W} \) the first-order conditions for GMM become

\[
0 = (-\frac{1}{n} Z'X)'\hat{W}(\frac{1}{n} Z'(y - X\hat{\beta}))
\]

\[
= \frac{1}{n} X'Z\hat{W}(\frac{1}{n} Z'(y - X\hat{\beta}))
\]

which has the solution

\[
\hat{\beta} = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'y.
\]

This estimator obviously differs from the usual IV estimator and equally obviously depends on the choice of weight matrix \( \hat{W} \). If we choose \( W = \frac{1}{n} Z'Z \), then the estimator is the 2SLS estimator.

This GMM estimator can be framed as an IV estimator to gain some insight. Define \( \tilde{Z} = Z\hat{W}Z'X \) then

\[
\hat{\beta} = (\tilde{Z}'\hat{\beta})^{-1}\tilde{Z}'\tilde{y}
\]

and the GMM estimator is seen as an IV estimator with a particular choice of instruments. The instruments \( \tilde{Z} = ZA \) are a linear transformation of the
original instruments \( Z \) using the \( p \times m \) transformation matrix \( A = \hat{W}Z'X \). Once again an obvious issue, which will be taken up later, is the appropriate choice of \( \hat{W} \). For the 2SLS weights, the instruments will be fitted values from a first stage regression of \( x_t \) on \( z_t \).

18.3 Asymptotic Behavior of GMM

18.3.1 Consistency

We now turn to the properties of the estimator that comes from the GMM exercise. The estimator is generally nonlinear, hence the iterative solution to the first-order conditions. As a result, the small sample properties of the estimator are problematic and we must generally rely on the large-sample asymptotic behavior for purposes of inference.

The following consistency result follows from verification of the high-level conditions given in Theorem 4.5 from Chapter 4.

**Theorem 18.1.** Suppose \( \hat{W} \rightarrow_p W \) and (i) \( \theta^o \in \Theta \) compact; (ii) \( g(z; \theta) \) continuous for \( \theta \in \Theta \) with probability 1; (iii) \( \exists d(z) \) with \( ||g(z; \theta)|| < d(z) \forall \theta \in \Theta \) and \( E[d(z)] < \infty \); (iv) \( W \) positive definite and \( E[g(z; \theta)] = 0 \) iff \( \theta = \theta^o \), then \( \hat{\theta} \rightarrow_p \theta^o \). □

**Proof:** See Appendix to this chapter. □

Strictly speaking, these low-level conditions are not assured, in general, and should be verified for the model at hand. The first two assumptions will be relatively innocuous and easily verified for most models. The third assumption, which is used to prove uniform convergence in probability, can be tedious to verify but will not usually present problems. In some cases such as the linear IV problem it is straightforward.

The fourth assumption assures global identifiability of the parameter vector within \( \Theta \). This global identification means that we can obtain consistent estimates regardless of the starting values used. It may be unrealistic when \( g_n(\theta) \) is nonlinear, since it is possible that \( W\gamma(\theta) = 0 \) has multiple solutions. This condition may be weakened to local identification by substituting the condition: \( W\gamma(\theta) = 0 \) iff \( \theta = \theta^o \) for all \( \theta \in \Theta \). In this case, consistency would follow if the initial value is itself consistent or is assured to be in the neighborhood specified with probability one.

18.3.2 Asymptotic Normality

The structure of the GMM problem allows us to obtain normality with only first derivatives assumed. Accordingly, we will not appeal to Theorem 4.6 but instead prove the result directly.

**Theorem 18.2.** Suppose \( \hat{W} \rightarrow_p W \) and (i) \( \hat{\theta} \rightarrow_d \theta^o \), \( \theta^o \) interior to \( \Theta \); (ii) \( g(z; \theta) \) continuously differentiable in a neighborhood \( N \) of \( \theta^o \); (iii) \( \exists D(z) \) s.t.
18.3. ASYMPTOTIC BEHAVIOR OF GMM

$\frac{\partial g(z; \theta)}{\partial \theta} < D(z)$ for all $\theta \in \mathcal{N}$ and $E[D(z)] < \infty$; (iv) $\Gamma' W \Gamma$ nonsingular for $\Gamma = E \left[ \frac{\partial g(z; \theta^0)}{\partial \theta} \right]$, and (v) $\sqrt{n} \hat{g}_n(\theta^0) \xrightarrow{d} N(0, \Omega)$; $\sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, (\Gamma' W \Gamma)^{-1} \Gamma' W \Omega W \Gamma (\Gamma' W \Gamma)^{-1})$.

**Proof:** See Appendix to this chapter. \qed

We do not need all the assumptions of the consistency proof only interiority and consistency of $\hat{\theta}$ in (i), which can occur under more general conditions. Condition (ii) is a regularity condition and easily verified. Condition (iii) is analogous to (ii) in the consistency theorem except now we are talking about estimating the matrix $\Gamma$ consistently. Condition (iv) is needed for the inverse in the limiting covariance and usually follows from local identification. Condition (v) only requires existence of moments of $g(z; \theta^0)$ and is needed to invoke the central limit theorem and have asymptotic normality.

For the just-identified case, $m = p$ and the matrix $\Gamma$ is square nonsingular so the limiting covariance matrix simplifies to $\Gamma^{-1} \Omega \Gamma^{-1} = (\Gamma' \Omega^{-1} \Gamma)^{-1}$, which does not depend on the weight matrix used and attains the lower bound developed in the next subsection.

### 18.3.3 Optimal Weight Matrix

The consistency and normality results above apply for any weight matrix such that $\hat{W} \rightarrow_p W$. The results in Theorem 18.2 allows us to establish the weight matrix that would minimize, in some sense, the limiting covariance matrix.

When the weight matrix is chosen optimally, then the form of the covariance matrix simplifies considerably. For the present case $W = \Omega^{-1}$, whereupon the covariance becomes $(\Gamma' \Omega^{-1} \Gamma)^{-1}$, yields the smallest covariance in the sense given below. Since $\Omega$ is positive definite, we can find square upper-triangular $U$ such that $\Omega = U'U$ and $\Omega^{-1} = U^{-1}U'$. Define $\Gamma^* = U^{-1} \Gamma$ and $W^* = UWU'$ then the difference

$$(\Gamma' \Omega^{-1} \Gamma) - (\Gamma' W \Gamma)(\Gamma' W \Omega W \Gamma)^{-1}(\Gamma' W \Gamma)$$

$$= (\Gamma' U^{-1} U^{-1} \Gamma) - (\Gamma' U^{-1} U W W' U^{-1} \Gamma)$$

$$\times (\Gamma' U^{-1} U W W' U^{-1} \Gamma)^{-1} (\Gamma' U^{-1} U W W' U^{-1} \Gamma)$$

$$= (\Gamma^* \Gamma^*) - (\Gamma^* W \Gamma^*) (\Gamma^* W \Gamma^*)^{-1} (\Gamma^* W \Gamma^*)$$

$$= \Gamma^* [I_m - W^* \Gamma^* (\Gamma^* W \Gamma^*)^{-1} (\Gamma^* W \Gamma^*)] \Gamma^*$$

The component in brackets is symmetric idempotent so the entire expression is positive semidefinite. This means the first matrix in the difference on the left-hand side of (18.7) exceeds the second by a positive semidefinite matrix. And the inverse of the second, which is the covariance for any $W$, exceeds the first, which is the covariance for $W = \Omega^{-1}$, by a positive semidefinite matrix. Thus $(\Gamma' \Omega^{-1} \Gamma)^{-1}$ is a lower bound for the limiting covariance matrix for the GMM estimator.
18.3.4 Efficient Two-Step Estimator

In order to obtain the efficient GMM estimator we need a consistent estimator of \( \Omega = E \left[ g(z; \theta^0)g(z; \theta^0)' \right] \). A natural estimator is

\[
\tilde{\Omega} = \Omega_n(\tilde{\theta}) = \frac{1}{n} \sum_{t=1}^{n} g(z_t; \tilde{\theta})g(z_t; \tilde{\theta})'
\]

(18.8)

where \( \tilde{\theta} \) is a preliminary consistent estimator of \( \theta^0 \). This estimator is consistent and will also be useful in consistently estimating the efficient limiting covariance matrix. We will also require a consistent estimator of \( \Gamma \), which is provided by \( \bar{G}_n(\hat{\theta}) \). Consistency of both is assured by the following theorem.

**Theorem 18.3.** Suppose the hypotheses of Theorem 18.2 are satisfied, and \( \exists \delta(z) \) s.t. \( \|g(z; \theta)\|_2^2 < \delta(z) \) for all \( \theta \) in a neighborhood \( N \) of \( \theta^0 \) and \( E[\delta(z)] < \infty \), then \( \bar{G}_n(\theta) \rightarrow \Gamma \) and \( \Omega_n(\theta) \rightarrow_p \Omega \). \( \square \)

**Proof:** See Appendix to this chapter. \( \square \)

The conditions of this theorem only slightly strengthen those of Theorem 18.2. We can take \( D(z) \) from Theorem 18.2 and set \( \delta(z) = D(z)^2 \) for the bounding function here and add the assumption that \( E[D(z)^2] < \infty \).

This leaves us with the task of finding an initial consistent estimator of \( \theta \). A common practice is to set \( \tilde{W} = W = I_m \) and proceed with GMM with this weight matrix as a first step, so

\[
\tilde{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{2} g_n(\theta)'g_n(\theta).
\]

By Theorem 18.1 this estimator will be consistent and by Theorem 18.3, \( \tilde{\Omega} \) based on it will also be consistent. We next obtain a second step GMM estimator using \( \tilde{\Omega} \) as the weight matrix, so

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{2} g_n(\theta)'\Omega_n(\tilde{\theta})^{-1}g_n(\theta).
\]

This two-step estimator will attain the lower bound for limiting covariance. Specifically

\[
\sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, (\Gamma'\Omega^{-1}\Gamma)^{-1}).
\]

This estimator is called the \textit{two-step efficient GMM estimator}.

This two-step estimator will depend on the choice of the weight matrix in the first step. It is possible to avoid this shortcoming by recomputing the estimated weight matrix \( \tilde{\Omega} = \Omega_n(\tilde{\theta}) \) using the second step estimator \( \hat{\theta} \) and then restimating GMM using this updated estimator. We continue to iterate in this fashion until \( \hat{\theta} \) converges, whereupon \( \tilde{\Omega} \) will also have converged. This approach is called the \textit{iterated efficient GMM estimator}. Note that the limiting distribution will be the same as the two-step procedure since \( \tilde{\Omega} \) will, in the end, still be consistent. It has an advantage in that the estimator is invariant with
respect to the scale of the data and the initial weighting matrix. And there is some evidence that the iterations result in better finite sample properties.

Another variation is to include the weight matrix in the minimization problem. Define

$$\Omega_n(\theta) = \sum_{t=1}^{n} g(z_t; \theta) g(z_t; \theta)'$$

and then take the estimator of $\theta$ as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{2} \bar{g}_n(\theta)' \Omega_n^{-1}(\theta) \bar{g}_n(\theta).$$

This estimator is called the continuously updating efficient GMM estimator. It is somewhat more complicated but has the same limiting distribution as the two-step and iterated GMM estimators. The usual asymptotic limiting distribution can be thought of as the leading term in an expansion with additional terms that are of smaller order in $n$. When account is taken of these higher-order terms there is some theoretical evidence that the continuously updating GMM estimator performs better. Such an analysis is beyond the scope of this chapter.

### 18.4 GMM Inference

#### 18.4.1 Standard Normal Tests

First we consider testing a scalar null hypothesis such as $H_1: \theta_j = \theta_j^o$. Now, the convergent iterate $\hat{\theta}$ will, by Theorem 15.2, yield

$$\sqrt{n}(\hat{\theta} - \theta^o) \overset{d}{\rightarrow} N(0, C^{-1}).$$

for $C = (\Gamma'W\Gamma)^{-1}\Gamma'W\Omega W\Gamma (\Gamma'W\Gamma)^{-1}$. Thus,

$$\frac{\sqrt{n}(\hat{\theta}_j - \theta_j^o)}{\sqrt{\hat{C}_{jj}}} \overset{d}{\rightarrow} N(0, 1)$$

where the $jj$-th subscript indicates the $jj$-th element of the matrix. Substituting the consistent estimators from Theorem 18.3, we have

$$\frac{\sqrt{n}(\hat{\theta}_j - \theta_j^o)}{\sqrt{\hat{C}_{jj}}} \overset{d}{\rightarrow} N(0, 1)$$

(18.9)

for $\hat{C} = (\hat{G}_n(\hat{\theta})' \hat{W} \hat{G}_n(\hat{\theta}))^{-1} \hat{G}_n(\hat{\theta})' \hat{W} \hat{W} \hat{G}_n(\hat{\theta}) (\hat{G}_n(\hat{\theta})' \hat{W} \hat{G}_n(\hat{\theta}))^{-1} \rightarrow_p C$.

Under an alternative hypothesis, say $H_1: \theta_j = \theta_j^1 \neq \theta_j^2$, we have

$$\frac{\sqrt{n}(\hat{\theta}_j - \theta_j^2)}{\sqrt{\hat{C}_{jj}}} = \frac{\sqrt{n}(\hat{\theta}_j - \theta_j^1)}{\sqrt{\hat{C}_{jj}}} + \frac{\sqrt{n}(\theta_j^1 - \theta_j^2)}{\sqrt{\hat{C}_{jj}}} = N(0, 1) + \frac{\sqrt{n}(\theta_j^1 - \theta_j^2)}{\sqrt{\hat{C}_{jj}}} + o_p(1)$$
so we see that the statistic will diverge in the positive/negative direction at the rate \( \sqrt{n} \) depending on whether \((\hat{\theta}_j^{\dagger} - \hat{\theta}_j^{\circ})\) is positive/negative. The test will therefore be consistent.

### 18.4.2 Wald-Type Tests

Suppose that we seek to test the null hypothesis \( H_0 : r(\theta^0) = 0 \) where \( r(\cdot) \) is a \( q \times 1 \) continuously differentiable vector function. The asymptotic normality result may be utilized to obtain a quadratic form based test of the Wald type. Using the delta method we find that

\[
\sqrt{n} r(\hat{\theta}) = \sqrt{n} r(\theta^0) + R(\theta^0)^{1/2} (\hat{\theta} - \theta^0) + o_p(1)
\]

\[\rightarrow N(0, RC R')\]

where \( R(\theta) = \partial r(\theta)/\partial \theta' \), \( R = R(\theta^0) \), and \( \theta^* \) between \( \hat{\theta} \) and \( \theta^0 \), so

\[
nr(\hat{\theta})'[RCR']^{-1}nr(\hat{\theta}) \rightarrow \chi^2_q \]

for \( RCR' \) nonsingular. A consistent estimate of \( C \) is introduced above, and \( \hat{R} = R(\hat{\theta}) \) will be consistent for \( R \). Thus, under the null hypothesis, we have the feasible Wald-type statistic

\[ W_{GMM} = n r(\hat{\theta})'[\hat{R} \hat{C} \hat{R}']^{-1}r(\hat{\theta}) \rightarrow \chi^2_q \]

Under the alternative hypothesis \( H_0 : r(\theta^0) \neq 0 \), the statistic will diverge in the positive direction at the rate \( n \).

### 18.4.3 LR-Type Test

The likelihood-ratio statistic compared the value of the log-likelihood function at restricted and unrestricted estimates of the parameter. We can construct a similar statistic using the criterion functions for the GMM case when based on a consistent estimator of the optimal weight matrix \( W = \Omega^{-1} \).

First, we must work out the details of the restricted GMM estimator. Define

\[ \tilde{\theta} = \arg\min_{\theta \in \Theta} \psi_n(\theta) \text{ s.t. } r(\theta) = 0 \]

and the corresponding Lagrangian objective function

\[ \varphi(\theta, \lambda) = \frac{1}{2} \bar{g}_n(\theta)' \bar{W} \bar{g}_n(\theta) + \lambda' r(\theta) \]

where \( \lambda \) is a \( q \times 1 \) vector of Lagrangian multipliers. The first-order conditions for optimizing this function, evaluated at the solution \( \tilde{\theta} \) are

\[
\frac{\partial \varphi(\hat{\theta}, \lambda)}{\partial \theta} = \tilde{G}_n(\tilde{\theta})' \tilde{W} \tilde{g}_n(\tilde{\theta}) + R(\tilde{\theta})' \lambda = 0 \]

\[ (18.11) \]

\[
\frac{\partial \varphi(\hat{\theta}, \lambda)}{\partial \lambda} = r(\tilde{\theta}) = 0. \]

\[ (18.12) \]
The solution \( \tilde{\theta} \) is the \textit{restricted GMM estimator}. An alternative would be to use the restriction to substitute out some of the parameters and then proceed with the usual GMM approach on the simplified problem. Notice that we used, at least asymptotically, the same weight matrix for both the restricted and unrestricted criterion functions.

In the following, we will define \( V = \Gamma' W T \). Since \( \tilde{\theta} \to_p \theta^0 \), then \( \bar{G}_n(\tilde{\theta}) \to_p \Gamma \), and \( R = R(\tilde{\theta}) \to_p R(\theta^0) \). Now premultiply the first equation (18.10) by \( RV^{-1} \) to obtain

\[
0 = RV^{-1} \bar{G}_n(\tilde{\theta})' \hat{W} \bar{g}_n(\tilde{\theta}) + RV^{-1} R(\tilde{\theta})' \lambda
\]

which can be solved for \( \lambda \) to yield

\[
\lambda = -[RV^{-1} R']^{-1} RV^{-1} \bar{G}_n(\tilde{\theta})' \hat{W} \bar{g}_n(\tilde{\theta})
\]

\[
= -[RV^{-1} R']^{-1} RV^{-1} \Gamma W \bar{g}_n(\tilde{\theta}) + o_p(1/\sqrt{n})
\]

Now define \( \hat{g}_n = \hat{g}_n(\theta^0) \), \( G_n = G_n(\theta^0) \), and expand \( \hat{g}_n(\theta) \) around \( \theta^0 \) to obtain

\[
\lambda = -[RV^{-1} R']^{-1} RV^{-1} \Gamma W \hat{g}_n
\]

\[
= -[RV^{-1} R']^{-1} RV^{-1} \Gamma W \hat{g}_n(\tilde{\theta} - \theta^0) + o_p(1/\sqrt{n})
\]

\[
= -[RV^{-1} R']^{-1} RV^{-1} \Gamma W \hat{g}_n + o_p(1/\sqrt{n})
\]

since \( \Gamma W \hat{G}_n \to_p V \), \( (\tilde{\theta} - \theta^0) = \theta^0(1/\sqrt{n}) \) and \( 0 = r(\theta^0) \), whereupon \( 0 = r(\theta) = R(\theta - \theta^0) + o_p(1/\sqrt{n}) \).

Expanding the first term in the first equation (18.10) of the first-order conditions around \( \theta \) yields

\[
\bar{G}_n(\tilde{\theta})' \hat{W} \bar{g}_n(\tilde{\theta}) = \bar{G}_n(\theta)' \hat{W} \bar{g}_n(\theta) + \bar{G}_n(\tilde{\theta})' \hat{W} \bar{G}_n(\theta)(\tilde{\theta} - \theta) + o_p(1/\sqrt{n})
\]

\[
= \bar{G}_n(\theta)' \hat{W} \bar{G}_n(\theta)(\tilde{\theta} - \theta) + o_p(1/\sqrt{n})
\]

\[
= \Gamma' W T (\tilde{\theta} - \theta) + o_p(1/\sqrt{n})
\]

since \( \bar{G}_n(\tilde{\theta})' \hat{W} \bar{g}_n(\tilde{\theta}) = 0 \). Substituting this and the expression for \( \lambda \) back into the first equation (18.10) of the first-order conditions yields

\[
0 = \Gamma' W T (\tilde{\theta} - \theta) - R'[RV^{-1} R']^{-1} RV^{-1} \Gamma W \bar{g}_n + o_p(1/\sqrt{n})
\]

\[
= V(\tilde{\theta} - \theta) - R'[RV^{-1} R']^{-1} RV^{-1} \Gamma W \bar{g}_n + o_p(1/\sqrt{n}).
\]
Solving for \((\tilde{\theta} - \hat{\theta})\) yields
\[
(\tilde{\theta} - \hat{\theta}) = V^{-1}R'[RV^{-1}R']^{-1}RV^{-1}\Gamma W g_n + o_p\left(\frac{1}{\sqrt{n}}\right)
\]
\[
= V^{-1}R'[RV^{-1}R']^{-1}R(\theta - \theta^0) + o_p\left(\frac{1}{\sqrt{n}}\right)
\]
\[
= V^{-1}R'[RV^{-1}R']^{-1}r(\hat{\theta}) + o_p\left(\frac{1}{\sqrt{n}}\right),
\]
(18.13)

since \(\sqrt{n}(\tilde{\theta} - \theta^0) = V^{-1}\Gamma W \sqrt{n}g_n + o_p(1)\) and \(r(\hat{\theta}) = R(\theta^0)(\tilde{\theta} - \theta^0) + o_p(1/\sqrt{n})\). Furthermore, we find
\[
\sqrt{n}(\tilde{\theta} - \theta^0) = [I_p - V^{-1}R'[RV^{-1}R']^{-1}R] \sqrt{n}(\tilde{\theta} - \theta^0) + o_p(1)
\]
\[
\rightarrow_d N(0, [V^{-1} - V^{-1}R'[RV^{-1}R']^{-1}RV^{-1}]).
\]

Note that this covariance matrix has rank \(p - q\), which reflects the fact that \(\tilde{\theta}\) satisfies the \(q\) restrictions.

Expanding \(g_n(\theta)\) around \(\hat{\theta}\), we obtain
\[
g_n(\tilde{\theta}) = g_n(\hat{\theta}) + \tilde{G}_n(\tilde{\theta})(\tilde{\theta} - \hat{\theta}) + o_p\left(\frac{1}{\sqrt{n}}\right)
\]
which may be substituted into \(\psi_n(\theta) = \frac{1}{2}g_n(\theta)'W g_n(\theta)\) to yield
\[
\psi_n(\tilde{\theta}) = \frac{1}{2}g_n(\tilde{\theta})'W g_n(\tilde{\theta}) + g_n(\tilde{\theta})'W \tilde{G}_n(\tilde{\theta})(\tilde{\theta} - \hat{\theta})
\]
\[
+ \frac{1}{2}(\tilde{\theta} - \hat{\theta})'G_n(\tilde{\theta})'W \tilde{G}_n(\tilde{\theta})(\tilde{\theta} - \hat{\theta}) + o_p(\frac{1}{n}).
\]

Recognizing that the first term on the right-hand side is \(\psi_n(\tilde{\theta})\), the second term is zero, reorganizing, and multiplying by \(2n\) yields
\[
2n(\psi_n(\tilde{\theta}) - \psi_n(\hat{\theta})) = n(\tilde{\theta} - \hat{\theta})'G_n(\tilde{\theta})'W \tilde{G}_n(\tilde{\theta})(\tilde{\theta} - \hat{\theta}) + o_p(1)
\]
\[
= n(\tilde{\theta} - \hat{\theta})'\Gamma WT(\tilde{\theta} - \hat{\theta}) + o_p(1)
\]
\[
= n(\tilde{\theta} - \hat{\theta})'V(\tilde{\theta} - \hat{\theta}) + o_p(1).
\]

Now, substitute for \((\tilde{\theta} - \hat{\theta})\) from above and define \(LR_{GMM} = 2n[\psi_n(\tilde{\theta}) - \psi_n(\hat{\theta})]\) to obtain
\[
LR_{GMM} = nr(\tilde{\theta})'[RV^{-1}R']^{-1}RV^{-1}VV^{-1}R'[RV^{-1}R']^{-1}r(\hat{\theta}) + o_p(1)
\]
\[
= nr(\tilde{\theta})'[RV^{-1}R']^{-1}r(\hat{\theta}) + o_p(1).
\]

The quadratic form above is quite similar to that for the Wald-type statistic, except for the matrix in the middle, which is \(C\) for \(W_{GMM}\) and \(V^{-1}\).
for $L_{RGMM}$. If we use asymptotically efficient weights then $W = \Omega^{-1}$ and 
$C = (\Gamma' \Omega^{-1} \Gamma)^{-1} = V^{-1}$, whereupon

$$L_{RGMM} = W_{GMM} + o_p(1) \xrightarrow{d} \chi^2_q$$

And we see that, when based on an asymptotically efficient weight matrix, the $L_{RGMM}$ and $W_{GMM}$ tests are asymptotically equivalent tests and will reject and not together in large samples, under the null hypothesis. If we do not use the efficient weight matrix then the two will not be asymptotically equivalent and the $L_{RGMM}$ will not have the posited limiting chi-squared distribution.

One special case of the $L$-type test has received enormous attention in the literature. Suppose the unrestricted model is just-identified then the unrestricted GMM estimator becomes method of moments and is the solution to $g_n(\hat{\theta}) = 0$, so $\psi_n(\hat{\theta}) = 0$ and $L_{RGMM} = n\psi_n(\hat{\theta}) \xrightarrow{d} \chi^2_q$. Now consider the over-identified unrestricted model above with $m$ equations and $p$ parameters and suppose we add $m - p$ parameters to the equations so that the model is just-identified and use our model with $p$ parameters as the restricted model. Since we do not need to perform the just-identified unrestricted estimation, we can treat $\hat{\theta}$ as the restricted estimator and define

$$J = 2n\psi_n(\hat{\theta}) \xrightarrow{d} \chi^2_{m-p}$$

This is called the $J$-test and is a test of the all the over-identifying restrictions. There are $p$ parameters and $m > p$ moment conditions. Any $p$ of the moment conditions would identify the parameters so we do not distinguish between identifying and over-identifying moment conditions – all are treated symmetrically. The null hypothesis is that all the moment conditions are true. If any are not satisfied we expect to get larger values of the statistic.

### 18.4.4 LM-Type Test

Having established the limiting distribution of the restricted estimator, we can develop a GMM version of the Lagrange multiplier test in a straightforward fashion. In the maximum likelihood case, Lagrange multiplier tests are also called score tests and are quadratic forms involving the unrestricted first derivatives of the likelihood function evaluated at the restricted estimates. We first consider the GMM case for a general weight matrix and then the case where the weight matrix is asymptotically optimal.

Analogous to maximum likelihood, consider the first derivative of the unrestricted criterion function evaluated at the restricted estimates

$$\frac{\partial \psi_n(\hat{\theta})}{\partial \theta} = G_n(\hat{\theta})' \hat{W} g_n(\hat{\theta}),$$

which will generally be nonzero. Now expand around $\hat{\theta}$ to find

$$g_n(\hat{\theta}) = g_n(\bar{\theta}) + G_n(\hat{\theta})(\bar{\theta} - \hat{\theta}) + o_p(1/\sqrt{n}) \\
\bar{G}_n(\hat{\theta}) = \bar{G}_n(\bar{\theta}) + O_p(1/\sqrt{n})$$
which can be substituted into the derivative to yield
\[
\frac{\partial \psi_n(\theta)}{\partial \theta} = \hat{G}_n(\theta)' \hat{W} g_n(\theta) + (\hat{G}_n(\theta) - \hat{G}_n(\bar{\theta}))' \hat{W} g_n(\theta) + (\hat{G}_n(\bar{\theta}) - \hat{G}_n(\bar{\theta}))' \hat{W} \hat{G}_n(\bar{\theta})(\bar{\theta} - \hat{\theta}) + o_p(1/\sqrt{n})
\]
\[
= \hat{G}_n(\bar{\theta})' \hat{W} g_n(\theta) + (\hat{G}_n(\theta) - \hat{G}_n(\bar{\theta}))' \hat{W} g_n(\theta) + \hat{G}_n(\bar{\theta})' \hat{W} \hat{G}_n(\bar{\theta})(\bar{\theta} - \hat{\theta}) + o_p(1/\sqrt{n})
\]
since the first term is zero and the second and fourth are \(o_p(1/\sqrt{n})\). Now \(\hat{G}_n(\bar{\theta})' \hat{W} \hat{G}_n(\bar{\theta}) \rightarrow_p V\), so we have
\[
\sqrt{n} \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} = V \sqrt{n}(\bar{\theta} - \hat{\theta}) + o_p(1)
\]
\[
= R'[RV^{-1}R']^{-1}R \sqrt{n}(\bar{\theta} - \theta^0) + o_p(1) \quad (18.15)
\]
\[
\rightarrow_d N(0, R'[RV^{-1}R']^{-1}RRC'[RV^{-1}R']^{-1}R)
\]
where we have substituted for \((\bar{\theta} - \hat{\theta})\) from (18.12).

Note that \(R\) is \(q \times p\) so the covariance has rank \(q < p\) so this is a singular multivariate normal distribution. Thus we will form the quadratic using a generalized inverse of the covariance matrix to obtain the Lagrange multiplier test statistic
\[
LM_{GMM} = n \left( \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} \right)^\top \left( R'[RV^{-1}R']^{-1}RRC'[RV^{-1}R']^{-1}R \right) \frac{\partial \psi_n(\bar{\theta})}{\partial \theta}
\]
\[
\rightarrow \chi^2_q.
\]
And the \(LM_{GMM}\) test has the same limiting distribution as the \(W_{GMM}\) test. Using a generalized inverse effectively eliminates \(p - q\) linear dependent elements of \(\frac{\partial \psi_n(\bar{\theta})}{\partial \theta}\) and forms a quadratic form with the \(q\) remaining linear independent elements.

It can be shown that the \(LM_{GMM}\) test is also asymptotically equivalent to the \(W_{GMM}\). First, we observe that \(V^{-1}R'[RRC']^{-1}RV^{-1}\) is a generalized inverse of the covariance of \(n^{1/2} \frac{\partial \psi_n(\bar{\theta})}{\partial \theta}\) in (18.14). Substituting for the generalized inverse and \(\sqrt{n} \frac{\partial \psi_n(\bar{\theta})}{\partial \theta}\) in the quadratic form yields
\[
LM_{GMM} = n \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} \left( V^{-1}R'[RRC']^{-1}RV^{-1} \right) \frac{\partial \psi_n(\bar{\theta})}{\partial \theta}
\]
\[
= n(\bar{\theta} - \theta^0)' R'[RRC']^{-1}R(\bar{\theta} - \theta^0) + o_p(1)
\]
\[
= n \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} \left( R'[RRC']^{-1}R \right) + o_p(1)
\]
\[
= W_{GMM} + o_p(1).
\]
Thus we see that the \(LM_{GMM}\) and \(W_{GMM}\) tests are asymptotically equivalent and will reject or not together in large samples, under the null hypothesis. Note
that this equivalence holds regardless of the weight matrix used. They will both be equivalent to the $LR_{GMM}$ test if the weight matrix for both restricted and unrestricted GMM is asymptotically the optimal choice $W = \Omega^{-1}$, so $C = (\Gamma'\Omega^{-1}\Gamma)^{-1} = V^{-1}$.

The above formulation of the $LM_{GMM}$ test avoids the necessity of finding a generalized inverse numerically, which can sometimes present problems in practice as generalized inverse algorithms are not as accurate as classical inversion algorithms. And recall that the quadratic form will be the same for any choice of generalized inverse. The generalized inverse can be further simplified if the weight matrix is asymptotically optimal so $C = V^{-1}$. In this case, $V^{-1}$ is a generalized inverse and the quadratic form becomes

$$LM_{GMM} = n \frac{\partial \psi_n(\hat{\theta})}{\partial \theta'} V^{-1} \frac{\partial \psi_n(\hat{\theta})}{\partial \theta} \rightarrow_d \chi^2_q.$$ 

A feasible version of this statistic is obtained by substituting the consistent estimator $\tilde{V} = V_n(\tilde{\theta})$ as the covariance matrix to yield

$$LM_{GMM} = n \frac{\partial \psi_n(\tilde{\theta})}{\partial \theta'} \tilde{V}^{-1} \frac{\partial \psi_n(\tilde{\theta})}{\partial \theta} \rightarrow_d \chi^2_q.$$ 

Given the asymptotic equivalence of the tests one might wonder why all three have been given consideration. The reasons are the same as they were for maximum likelihood. The Wald-type test only requires unrestricted estimates and the LM-type test only requires the restricted estimates. If one or the other is simpler to obtain then the corresponding test is the natural choice. The LR-type test requires both restricted and unrestricted estimates but is invariant with respect to our choice for writing the restrictions. Plus the J-test falls out as a special case of the LR-type test. A cynical observer might suggest that all three persist in use because it gives us some choice which allow us to better match our preconceptions.

### 18.4.5 Robust Covariance Estimation

Up until this point we have considered the case where the underlying $z$ are jointly *i.i.d.* This assumption is satisfactory in many applications including cross-section and panel data cases. The limiting covariance estimators for this case are asymptotically appropriate and include the White-type heteroskedastic consistent covariance estimators for the OLS, NLS, and IV estimators as special cases, among others. Time-series models that otherwise satisfy the *i.i.d.* assumption will be similarly covered.

In the case of potentially serially dependent data the framework established above and all the various results apply with some modification. This does require additional assumptions and more attention to the estimated optimal weight matrix. The proofs on consistency only explicitly involved the *i.i.d.* assumption when Lemma 2.4 of Newey and West was invoked. But this Lemma still applies if the data are strictly stationary and ergotic. Accordingly, we
replace the assumption that $z_t$ are i.i.d. with the assumption that they are strictly stationary and ergotic. Strict stationarity implies the joint distribution of $(z_\tau, z_{\tau+1}, \ldots, z_{\tau+h})$ does not depend on $\tau$ for any $m$. Ergodicity means the sample means are consistent: $n^{-1} \sum_{t=1}^n a(z_t) \to E[a(z)]$ for measurable functions $a(z)$ with $E[|a(z)|] < \infty$. This modification allows the consistency theorem its proof to be applied.

The asymptotic normality result requires some additional concepts when serial dependence is allowed. Complications arise in satisfying Condition (v) of Theorem 18.2. For the i.i.d. case we can appeal to the Lindberg-Levy central limit theorem to obtain the result. For the dependent case, we need some modifications to the notation.

To begin with, the covariance of a single observation on $g(z_t, \theta^0)$ is no longer the same as the covariance of the limiting distribution of $\sqrt{n} \hat{g}_n(\theta^0)$. Specifically, assuming all the needed covariances exist,

$$\sqrt{n} \hat{g}_n(\theta^0) = \frac{1}{\sqrt{n}} \sum_{t=1}^n g(z_t, \theta^0) = \frac{1}{\sqrt{n}} \sum_{t=1}^n w_t,$$

for $w_t = g(z_t, \theta^0)$, has covariance matrix

$$\Omega_n = E \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^n w_t \frac{1}{\sqrt{n}} \sum_{t=1}^n w_t' \right]$$

$$= \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n E[w_t w_s']$$

$$= \frac{1}{n} \left( \sum_{t=1}^n E[w_t w_t'] + \sum_{t=1}^n \sum_{s \neq t} E[w_t w_s'] \right).$$

Now, define $\Omega_{t-s} = E[w_t w_s']$ as the auto-covariance matrices, where due to stationarity they only depend on the difference between the indices, then we have

$$\Omega_n = \frac{1}{n} \left[ n\Omega_0 + \sum_{t=1}^{n-1} \sum_{s=t+1}^{n-1} \Omega_{t-s} + \Omega'_{t-s} \right]$$

$$= \frac{1}{n} \left[ n\Omega_0 + \sum_{s=1}^{n-1} \frac{n-s}{n}(\Omega_s + \Omega'_s) \right]$$

$$= \Omega_0 + \sum_{s=1}^{n-1} \frac{n-s}{n}(\Omega_s + \Omega'_s)$$

$$= \Omega_0 + \sum_{s=1}^{n-1} \left(1 - \frac{s}{n}\right)(\Omega_s + \Omega'_s).$$
18.4. GMM INFERENCE

This structure reflects the fact that we only have covariances between $w_\tau$ and $w_{\tau-s}$ for $\tau = s + 1, \ldots, n$.

Intuitively, a stochastic process $\{w_t\}_{t=-\infty}^{\infty}$ is ergotic if any two collections of the random variables partitioned far enough apart are essentially independent. This means that the covariances in the sums above will be growing smaller as the subscript grows larger. This is of course required in order for the covariance of the normalized partial sum $\sqrt{n}\hat{g}_n(\theta^0)$ to be finite. Along these lines, we suppose the infinite sequence is convergent, so

$$\Omega = \lim_{n \to \infty} \Omega_n.$$ 

In the i.i.d. case $\Omega = \Omega_0$, the own covariance and the Lindberg-Levy central limit theorem applies. In the present context, we will assume that a central limit theorem for dependent processes applies and

$$\sqrt{n}\hat{g}_n(\theta^0) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} w_t \to_d N(0, \Omega).$$

Obtaining this result requires more than strict stationarity and ergodicity. A stronger form of asymptotic independence is required of the sequence such as $\alpha$-mixing or that it be a martingale difference. We will omit the technical details that justify the above assumption. With this assumption the results of Theorem 18.2 will still apply.

In order to conduct inference based on the results of the theorem, we will need consistent estimates of the components of the covariance matrix. The previously introduced estimator $\hat{\Gamma} = \hat{G}_n(\hat{\theta})$ will continue to be consistent for $\Gamma$ under the assumptions of Theorem 18.2. Whether conducting inference or efficient GMM estimation, we will need a consistent estimator of $\Omega$. The problem is that $\Omega_n$ involves the $n$ covariance matrices $(\Omega_0, \Omega_1, \ldots, \Omega_{n-1})$ so consistent estimation of all the covariances is infeasible. Fortunately, these covariances grow smaller with the size of the index so we can restrict our attention to the first $m+1$. An estimate of the $s-$th covariance is provided by

$$\hat{\Omega}_s = \frac{1}{n-s} \sum_{t=s+1}^{n} (g(z_t, \hat{\theta})g(z_t, \hat{\theta})') \to_p \Omega_s,$$

which will be consistent under the assumptions of Theorem 18.3. Newey and West suggest an estimator of $\Omega$ based on $m+1$ of these estimated covariances

$$\hat{\Omega} = \hat{\Omega}_0 + \sum_{s=1}^{m} \left(1 - \frac{s}{m+1}\right)(\hat{\Omega}_s + \hat{\Omega}'_s)$$

where $m \to \infty$ but $m/n \to 0$. They show that this estimator will be consistent under general conditions which will not be presented here.

An important issue with this covariance estimator is the choice of $m$, which is a tuning parameter. There is a tradeoff between increased bias as $m$ is made
smaller and increased variance a $m$ is made larger. In their original work Newey and West suggested $m = o(n^{1/4})$ should be satisfied. Recent work on the problem suggests $m = cn^{1/3}$ for the optimal choice. Stock and Watson suggest $c=.75$ as a likely choice. There has been much work on data driven choices of $m$, which are beyond the scope of this chapter.

18.5 Example

Keynes conjectured that the marginal propensity to consume (MPC) across individuals would be constant between 0 and 1. He also conjectured that the average propensity to consume (APC) would fall at higher income levels. Consistent with these conjectures, economist proposed the following linear model for consumption:

$$C = \alpha + \beta Y$$

where $C$ is consumption and $Y$ is disposable income. The marginal propensity to consume is $dC/dY = \beta$ and is constant as income rises while the average propensity to consume $C/Y = \alpha/Y + \beta$ will fall as income rises, provided $\alpha > 0$. Graphically, for a particular income level $Y$, the MPC is the slope of the line $C = \alpha + \beta Y$ and the APC is the slope of a ray passing from the origin passing through the line for that level.

Dusenberry proposed an empirical specification of this model at the aggregate level

$$C_t = \alpha + \beta Y_t + u_t$$

where $u_t$ is an unknown disturbance term. Early empirical studies gave strong support to this model, which fit the data very well. Based on cross-sectional and aggregate data it seemed that income was the primary determinant of consumption.

Based on this model and expectations that income would rise following World War II, predictions were that aggregate consumption would grow more slowly than income as APC began to fall. Consequently, there were predictions of secular stagnation and a renewed and prolonged depression following the war as government expenditures were scaled back. These dire predictions were wrong. Consumption rose at much the same rate as income. In fact Kuznets showed that APC was very stable at the aggregate level from decade to decade. This contradiction of the Keynesian conjecture in aggregate time series is known as the “consumption puzzle”.

Two early proposals, based on forward-looking consumers, that attempted to explain this puzzle were the life-cycle hypothesis of Ando, Brumberg, and Modigliano and the permanent income hypothesis of Friedman. We will analyze the latter since it will highlight many desirable features of GMM. Friedman hypothesized that there were two components to current income $Y$ a permanent component $Y^P$, which is expected income, and a transitory component $Y^T$, which are temporary and unpredictable deviations from the expectations so

$$Y = Y^P + Y^T.$$
Friedman argued that consumers will make their consumption decisions on the basis of permanent income only via the relationship
\[ C = \beta Y^P \]
which yields \( APC = C/Y = \beta(Y^P/Y) \). If high-income households have higher transitory income than low-income households, then APC is lower in high-income households. Over the long-run, income variation is due mainly (if not solely) to variation in permanent income, which implies a stable APC. Note that there is no intercept in this model.

Friedman considered the following econometric specification
\[ C_t = \alpha + \beta Y^P_t + u_t \]
where \( \alpha \) is expected to be small. He argued that a proper measure of permanent income is the present discounted value all current and expected future streams of income. He further argued that a reasonable predictor of permanent income is a weighted average of current and past streams of income with higher weights on recent income. A parsimonious specification that is consistent with this is the geometric lag specification
\[ Y^P_t = (1 - \lambda)\left[ Y_t + \lambda Y_{t-1} + \lambda^2 Y_{t-2} + \lambda^3 Y_{t-3} + \ldots \right] \]
for \( 1 > \lambda > 0 \). The problem with this specification is that it introduces a large (infinite?) number of regressor with nonlinear restrictions on their coefficients.

An apparent simplification in the model is achieved by considering the Koyck lag transformation
\[
C_t - \lambda C_{t-1} = (1 - \lambda)\alpha + \beta(Y^P_t - \lambda Y^P_{t-1}) + u_t - \lambda u_{t-1} \\
= (1 - \lambda)\alpha + (1 - \lambda)\beta Y^P_t + u_t - \lambda u_{t-1}
\]
or following some rearrangement
\[
C_t = (1 - \lambda)\alpha + (1 - \lambda)\beta Y_t + \lambda C_{t-1} + u_t - \lambda u_{t-1} \\
= \alpha^* + \beta^* Y_t + \lambda C_{t-1} + u_t^*
\]
where notation is obvious. This reparameterization suggests linear regression analysis by ordinary least squares, which was the path taken by Friedman. We are interested in whether or not \( \alpha^* = 0 \) and/or \( \lambda = 0 \). If \( \alpha^* > 0 \) then the permanent income model will not explain stable APC over time. And if \( \lambda = 0 \), then the permanent income model reduces to the Dusenberry specification.

For out econometric analysis, we utilize quarterly aggregate data from 1952.1 to 1961.2, which is relevant for the time of Friedman’s analysis. We begin by performing OLS on the Dusenberry specification and obtain (usual homoskedastic estimated standard errors in parenthesis)
\[
\hat{C}_t = 1.53492 + 0.87199 Y_t \\
(2.51204) \quad (0.00789)
\]
with $R^2 = 0.99705$ and $DW = 0.64705$. The intercept being non-zero, but not at all significant with a prob-value of $-0.54528$. The slope coefficient is in the hypothesized range and is significantly less than 1. The Durbin-Watson statistic is in the left-hand 1% tail, which indicates positive first-order serial correlation. The presence of serial correlation and possible heteroskedasticity means the standard errors are likely misstated. We could obtain standard errors that are robust against such problems but will proceed on to more complete specifications first.

Next we consider the Koyck transformation and estimate by OLS to obtain (usual homoskedastic estimated standard errors in parenthesis)

$$\hat{C}_t = 2.38178 + 0.43141Y_t + 0.50523C_{t-1}$$

with $R^2 = 0.99799$ and $DW = 1.15820$. Assuming the classical least squares assumption are satisfied, then under the null hypothesis that $\lambda = 0$, $\hat{\lambda}/s.e.(\hat{\lambda}) = 0.50523/1.141 = 4.42787$ asymptotically has a two-sided prob-value of approximately zero. Since $\lambda \geq 0$, then the alternative hypothesis is one-sided and the one-sided prob-value is still approximately zero. We strongly reject the null. Moreover, the size of $\hat{\lambda}$ indicates that there is a substantial contribution of lagged income to permanent income. The Dusenberry specification looks to be inadequate. The intercept is an estimate of $\alpha^*$ and the coefficient on $Y_t$ is an estimate of $\beta^*$. Using the definitions of $\beta^*$ and $\alpha^*$, we find $\hat{\beta} = \hat{\beta}^*/(1 - \hat{\lambda}) = 0.87194$ and $\hat{\alpha} = \hat{\alpha}^*/(1 - \hat{\lambda}) = 4.8393$, which interestingly enough for $\hat{\beta}$ is almost the same as the bivariate results. We will not calculate standard errors since the regression has obvious complication as set out below.

The problem is that, in the presence of serially correlated errors, a lagged dependent variable can be expected to be correlated with the regressors and the OLS estimates will be biased and inconsistent. The Durbin-Watson value 1.15820 is biased towards zero in the presence of lagged dependent variables so we construct the Durbin h-test which has a value of 3.6501 and under the null of no serial correlation, is a realization of a standard normal and so has a prob-value of approximately zero. Moreover, even if the original errors $u_t$ were serially uncorrelated, $u_t^* = u_t - \lambda u_{t-1}$ will be serially correlated for $\lambda > 0$.

The obvious solution to the problems revealed in the previous paragraph is to utilize instrumental variables techniques. The question is what to use as instruments. With some trepidation, we suppose that the errors in the determination of consumption are orthogonal to the determinates of income. Accordingly, the regressor $Y_t$ should not present problems and likewise $Y_{t-1}$. Moreover, $Y_{t-1}$ is a primary determinate of $C_{t-1}$, so is a good choice as an instrument. Thus the regression vector is $x_t^* = (1, Y_t, C_{t-1})$ and the instrument vector is $z_t^* = (1, Y_t, Y_{t-1})$.

Instrumental variable applied to this problem yields (Newey-West standard errors in parenthesis)

$$\hat{C}_t = 1.88517 + 0.79065Y_t + 0.09187C_{t-1}$$

(2.94087)  (0.20584)  (0.24232)
with $R^2 = 0.99721$ and $DW = 0.69252$. First notice the much smaller value of $\lambda$ and large value of its estimated standard error, which under the null of $\lambda = 0$ yields a ratio of $0.37911$, which asymptotically has a one-sided prob-value of $0.35348$. We cannot reject the null hypothesis that $\lambda = 0$. Note that the implied estimates $\hat{\beta} = \hat{\beta}^*/(1 - \hat{\lambda}) = 0.87063$ and $\hat{\alpha} = \hat{\alpha}^*/(1 - \hat{\lambda}) = 2.07587$, are again almost the same as the bivariate results. The corresponding Newey-West estimated standard errors for these transformations are $0.01102$ and $3.00932$. Testing the null hypothesis that $\beta = 1$, yields $(\hat{\beta} - 1)/s.e.(\hat{\beta}) = 11.739$, which asymptotically has a one-sided prob-value of zero. It is significantly less than one and more than zero. Testing the null that $\alpha = 0$, yields $\hat{\alpha}/s.e.(\hat{\alpha}) = 0.68981$, which has a two-side prob-value of $0.49499$, so we cannot reject the null. These are very different results than for least squares for $\hat{\lambda}$ and cast serious doubt on the permanent income hypothesis.

We now consider adding another instrument. For $\lambda > 0$, then if the permanent income model is correct, then $C_t$ will not only be correlated with $Y_t$ but also $Y_{t-1}$. This suggests that we can use $Y_{t-1}$ and $Y_{t-2}$ as instruments for $C_{t-1}$. Thus we now have $z'_t = (1, Y_t, Y_{t-1}, Y_{t-2})$ and the problem is overidentified. We have four equation

$$g(w_t, \theta) = \begin{pmatrix}
C_t - [(1 - \lambda)\alpha + (1 - \lambda)\beta Y_t + \lambda C_{t-1}] \\
Y_t(C_t - [(1 - \lambda)\alpha + (1 - \lambda)\beta Y_t + \lambda C_{t-1}]) \\
Y_{t-1}(C_t - [(1 - \lambda)\alpha + (1 - \lambda)\beta Y_t + \lambda C_{t-1}]) \\
Y_{t-2}(C_t - [(1 - \lambda)\alpha + (1 - \lambda)\beta Y_t + \lambda C_{t-1}])
\end{pmatrix}$$

where $w'_t = (C_t, C_{t-1}, Y_t, Y_{t-1}, Y_{t-2})$ and $\theta' = (\alpha, \beta, \lambda)$. Since GMM is designed to deal with nonlinear problems we estimate $\alpha$ and $\beta$ directly. Moreover, there will be no need to use the nonlinear transformation to obtain estimated standard errors. In view of the serial correlation resulting from the Koyck transformation and the evident serial correlation in the original bivariate regression we will report Newey-West standard errors.

We report three sets of GMM estimates for this over-identified case plus, for comparison purposes, the above results for the just-identified. The one-step estimator uses $(\frac{1}{n}Z'Z)^{-1}$ as the weight matrix. This means that the one-step estimates will be nonlinear 2SLS estimates with the corresponding Newey-West estimated standard errors. We use these as the starting point for second-step efficient GMM and iterated efficient GMM estimation. Results are reported in Table 18.1 below.

<table>
<thead>
<tr>
<th>GMM Procedure</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Just-identified (NLS)</td>
<td>4.81393</td>
<td>0.87194</td>
<td>0.50523</td>
</tr>
<tr>
<td></td>
<td>(5.64892)</td>
<td>(0.01545)</td>
<td>(0.09459)</td>
</tr>
<tr>
<td>Just-identified (NLIV)</td>
<td>2.07587</td>
<td>0.87063</td>
<td>0.09187</td>
</tr>
<tr>
<td></td>
<td>(3.00932)</td>
<td>(0.01102)</td>
<td>(0.24232)</td>
</tr>
<tr>
<td>First-step (NL2SLS)</td>
<td>1.54904</td>
<td>0.87142</td>
<td>0.02693</td>
</tr>
<tr>
<td></td>
<td>(3.33215)</td>
<td>(0.01179)</td>
<td>(0.25067)</td>
</tr>
<tr>
<td>Second-Step (GMM)</td>
<td>3.63698</td>
<td>0.86538</td>
<td>0.07297</td>
</tr>
<tr>
<td></td>
<td>(3.09981)</td>
<td>(0.01111)</td>
<td>(0.2433)</td>
</tr>
<tr>
<td>Iterated (GMM)</td>
<td>4.86906</td>
<td>0.86101</td>
<td>0.06734</td>
</tr>
<tr>
<td></td>
<td>(3.14683)</td>
<td>(0.01126)</td>
<td>(0.24077)</td>
</tr>
</tbody>
</table>
In looking at the above table, one is immediately struck by how similar the results are from estimator to estimator, once we correct for the problem of a RHS variable correlated with the errors. The first-step NL2SLS estimates are not that different from the just-identified IV estimates. Adding the instrument did not make much difference for IV estimation. The second-step and iterated estimates are also not that different from each other, but are little changed from the first-step NL2SLS estimates. The estimates of \( \lambda \) are larger but still not significantly different from zero, in either case, at any choice of significance. The estimates of \( \alpha \) are larger than before but are still insignificantly different from zero. The estimates of \( \beta \) are in every case significantly greater that zero and less than one.

This analysis should make it clear that the permanent income hypothesis, while possibly explaining the consumption puzzle, has a difficult time explaining the behavior of the data. If we eliminate \( C_t \) as a regressor, as the above results suggest, then we are back in the Dusenberry specification, which was also found wanting relative to the behavior of the data. A different model will have to be used to explain the consumption puzzle. Along the lines of the adequacy of the Koyck lag specification, we have a J-test result for the second-step and iterated GMM estimators. The values are respectively 2.79864 and 2.98906, which for a chi-squared with one degree of freedom yield prob-values of 0.09434 and 0.08388. So the omnibus test of model adequacy does not reject the specification. This is probably because we have not been creative enough in setting up implied moment conditions for the Dusenberry model.

### 18.A Appendix

We will first prove consistency by verifying the conditions of Theorem 4.5 from the Appendix of Chapter 4. For purposes of the proof below, we define \( \psi_n(\theta) = n^{-1} \sum_{t=1}^{n} (y_t - h(x_t, \theta))^2 \) and \( \psi_0(\theta) = E[(y - h(x, \theta))^2] \) for \( \theta \in \Theta \).

**Proof of Theorem 18.1.** Define \( \gamma_0(\theta) = E[g(z, \theta)] \), which is guaranteed to exist for \( \theta \in \Theta \) by (iii), and

\[
\psi_0(\theta) = \frac{1}{2} \gamma_0(\theta)'W\gamma_0(\theta).
\]

Now (C1), compactness of \( \Theta \), is assured by condition (i). Conditions (ii) and (iii) mean Newey and McFadden’s Lemma 2.4, presented in the Appendix to Chapter 5, is satisfied for \( a(z, \theta) = g(z, \theta^0) \). Thus \( \gamma_0(\theta) \) is continuous whereupon (C3) is met and moreover \( \sup_{\theta \in \Theta} \| \tilde{g}_n(z, \theta) - \gamma_0(\theta) \| \to_p 0 \). Identification condition (C4) follows since, by condition (iv) the quadratic form \( \gamma_0(\theta)'W\gamma_0(\theta) \geq 0 \) will be zero only if \( \theta = \theta^0 \).

This leaves us with the uniform convergence condition (C2) to verify. Adding and subtracting \( \gamma_0(\theta) \) and \( W \) inside the quadratic form in \( \psi_n(\theta) \) and expanding.
yields

\[
\psi_n(\theta) = \frac{1}{2} \bar{g}_n(\theta)' \tilde{W} \bar{g}_n(\theta)
\]

\[
= \frac{1}{2} \left( (\bar{g}_n(\theta) - \gamma_0(\theta)) + \gamma_0(\theta) \right)' W (\bar{g}_n(\theta) - \gamma_0(\theta)) + \gamma_0(\theta)
\]

\[
+ \frac{1}{2} \left( (\bar{g}_n(\theta) - \gamma_0(\theta)) + \gamma_0(\theta) \right)' (\tilde{W} - W) (\bar{g}_n(\theta) - \gamma_0(\theta)) + \gamma_0(\theta)
\]

\[
= \frac{1}{2} \gamma_0(\theta)' W \gamma_0(\theta) + (\bar{g}_n(z, \theta) - \gamma_0(\theta))' W \gamma_0(\theta)
\]

\[
+ \frac{1}{2} (\bar{g}_n(z, \theta) - \gamma_0(\theta))' W (\bar{g}_n(z, \theta) - \gamma_0(\theta))
\]

\[
+ \frac{1}{2} \gamma_0(\theta)' (\tilde{W} - W) \gamma_0(\theta) + (\bar{g}_n(z, \theta) - \gamma_0(\theta))' (\tilde{W} - W) \gamma_0(\theta)
\]

\[
+ \frac{1}{2} (\bar{g}_n(z, \theta) - \gamma_0(\theta))' (\tilde{W} - W) (\bar{g}_n(z, \theta) - \gamma_0(\theta)).
\]

Since \( \Theta \) compact and \( \gamma_0(\theta) \) continuous, then \( \gamma_0(\theta) \) bounded for \( \theta \in \Theta \). By the triangle and Cauchy-Swartz inequalities we have

\[
|\psi_n(\theta) - \psi_0(\theta)| \leq \left| (\bar{g}_n(z, \theta) - \gamma_0(\theta))' W \gamma_0(\theta) \right|
\]

\[
+ \frac{1}{2} \left| (\bar{g}_n(z, \theta) - \gamma_0(\theta))' W (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \right|
\]

\[
+ \left| (\bar{g}_n(z, \theta) - \gamma_0(\theta))' (\tilde{W} - W) \gamma_0(\theta) \right|
\]

\[
+ \frac{1}{2} \left| (\bar{g}_n(z, \theta) - \gamma_0(\theta))' (\tilde{W} - W) (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \right|
\]

\[
\leq \| (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \| \| W \| \| \gamma_0(\theta) \|
\]

\[
+ \frac{1}{2} \| (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \|^2 \| W \|
\]

\[
+ \| (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \| \| W - \tilde{W} \| \| \gamma_0(\theta) \|
\]

\[
+ \frac{1}{2} \| (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \|^2 \| \tilde{W} - W \|.
\]

Since \( \gamma_0(\theta) \) bounded, \( (\bar{g}_n(z, \theta) - \gamma_0(\theta)) \) converges uniformly in probability to zero, and \( \tilde{W} \to_p W \) for any \( \theta \in \Theta \), then \( \sup_{\theta \in \Theta} |\psi_n(\theta) - \psi_0(\theta)| \to_p 0 \) and \( \text{(C2)} \) is satisfied. \( \square \)

We now prove asymptotic normality of the GMM estimator but do not use Therem 4.6.

Proof of Theorem 18.2. By (i) \( \hat{\theta} \in \mathcal{N} \) with probability one and we can expand
the first-order conditions in a Taylor’s series around \( \theta^0 \) to obtain

\[
0 = \bar{G}_n(\hat{\theta})' \bar{W} g_n(\hat{\theta})
\]

\[
= \bar{G}_n(\hat{\theta})' \bar{W} \left[ g_n(\theta^0) + \bar{G}_n(\hat{\theta})(\hat{\theta} - \theta^0) \right]
\]

\[
= \bar{G}_n(\hat{\theta})' \bar{W} \bar{g}_n(\theta^0) + \bar{G}_n(\hat{\theta})' \bar{W} \bar{G}_n(\hat{\theta})(\hat{\theta} - \theta^0)
\]

for \( \tilde{\theta} \) between \( \hat{\theta} \) and \( \theta^0 \). Now, by (ii), (iii) and Lemma 2.4, \( \sup_{\theta \in \mathcal{X}} \| \bar{G}_n(\hat{\theta}) - \Gamma(\theta) \| \to_p 0 \) for \( \Gamma(\theta) = E \left[ \frac{\partial g(z; \theta)}{\partial \theta} \right] \). Thus \( \bar{G}_n(\hat{\theta}) \to_p \Gamma(\theta) = \Gamma \) and \( \bar{G}_n(\hat{\theta})' \bar{W} \bar{G}_n(\hat{\theta}) \to_p \Gamma' \bar{W} \Gamma \) which is nonsingular by (iv). So with probability one, we can solve for \( (\hat{\theta} - \theta^0) \) and scale by \( \sqrt{n} \) to obtain

\[
\sqrt{n}(\hat{\theta} - \theta^0) = -[\bar{G}_n(\hat{\theta})' \bar{W} \bar{G}_n(\hat{\theta})]^{-1} \bar{G}_n(\hat{\theta})' \bar{W} \sqrt{n} \bar{g}_n(\theta^0)
\]

\[
= -(\Gamma' \bar{W} \Gamma)^{-1} \Gamma' \bar{W} \sqrt{n} \bar{g}_n(\theta^0)
\]

\[
+ \left[ (\Gamma' \bar{W} \Gamma)^{-1} \Gamma' \bar{W} - \bar{G}_n(\hat{\theta})' \bar{W} \bar{G}_n(\hat{\theta}) \right]^{-1} \bar{G}_n(\hat{\theta})' \bar{W} \sqrt{n} \bar{g}_n(\theta^0)
\]

\[
= (\Gamma' \bar{W} \Gamma)^{-1} \Gamma' \bar{W} \sqrt{n} \bar{g}_n(\theta^0) + o_p(\| \sqrt{n} \bar{g}_n(\theta^0) \|)
\]

since all the estimates in the brackets are consistent for their targets. By (v), \( \sqrt{n} \bar{g}_n(\theta^0) = O_p(1) \), so the remainder term is \( o_p(1) \) and we have the result of the theorem. \( \Box \)

We now prove consistency of a couple of covariance component estimators.

**Proof of Theorem 18.3.** As seen in the previous proof, Condition (iii) of Theorem 18.2 assures \( \bar{G}_n(\hat{\theta}) \to_p \Gamma(\theta^0) = \Gamma \). Define \( \Omega_n(\theta) = \sum_{i=1}^n g(z_i; \theta)g(z_i; \theta)' \) and \( \Omega(\theta) = E[g(z; \theta)g(z; \theta)'] \), then the local uniform boundedness condition in this theorem implies \( \sup_{\theta \in \mathcal{X}} \| \Omega_n(\theta) - \Omega(\theta) \| \to_p 0 \) for \( \theta \) in a neighborhood of \( \theta^0 \). With probability one, then \( \hat{\theta} \) lies in the neighborhood so \( \| \Omega_n(\hat{\theta}) - \Omega(\hat{\theta}) \| \to_p 0 \) and \( \hat{\Omega} = \Omega(\hat{\theta}) \to_p \Omega(\theta^0) = \Omega \). \( \Box \)