## 8. Parameter Significance Tests for Spatial Regression

Before developing significance tests of parameters for spatial regression, it is appropriate to begin by stating a few general properties of maximum-likelihood estimators that will be crucial for the analysis below. (These properties are developed in more detail in Section A3.7 of the Appendix.) Here we employ the following notational conventions. First, observe that since both log likelihood functions and sampling distributions of maximum-likelihood estimators depend on the given sample size, *n*, we now make this explicit by writing $L_n$ and $\hat{\theta}_n$, and replace expression (7.1.9) with the more sample-explicit form:

(8.1) $$L_n(\hat{\theta}_n \mid y) = \max_{\theta \in \Theta} L_n(\theta \mid y)$$

In addition, note that the symbol, $\theta$, in (8.1) is treated as a *variable* which denotes possible parameter values. The desired estimator, $\hat{\theta}_n$ , is then distinguished as the value of $\theta$ that maximizes the log-likelihood function, $L_n(\theta)$. But it is also important to distinguish the *true value* of $\theta$, which we now denote by $\theta_0$. In particular, note that all distributional properties of the random vector, $\hat{\theta}_n$, will necessarily depend on the *true* distribution of *y*, say with density $f(y \mid \theta_0)$.

In these terms, the single most important property of maximum-likelihood estimators is their *consistency*, namely that for sufficiently large sample sizes, the estimator, $\hat{\theta}_n$, is very likely to be close to the true value, $\theta_0$. More precisely, as *n* becomes large, the chance of $\hat{\theta}_n$ being further from $\theta_0$ than any arbitrarily small amount, $\varepsilon$, shrinks to zero, i.e.,

(8.2) $$\lim_{n \to \infty} \Pr\left( \| \hat{\theta}_n - \theta_0 \| > \varepsilon \right) = 0 \quad \text{for all } \varepsilon > 0$$

This is expressed more compactly by saying that $\hat{\theta}_n$ *converges in probability* to $\theta_0$ and is written as

(8.3) $$\hat{\theta}_n \underset{prob}{\rightarrow} \theta_0$$

This consistency property ensures that given enough sample information, maximum-likelihood estimators will eventually "learn" the true values of parameters. Without such a guarantee, it is hard to consider any estimator as being statistically reliable.

The single most useful tool for establishing such consistency results is the classical *Law of Large Numbers* (LLN), which states that for any sequence of independently and identically distributed (*iid*) random variables, $(X_1,..,X_n)$, from a statistical population, *X*,

with mean, $E(X)=\mu$, the sample mean, $\bar{X}_n$ converges in probability to this population mean as *n* increases, i.e.,

(8.4) $$\bar{X}_n = \tfrac{1}{n}\sum_{i=1}^{n} X_i \underset{prob}{\rightarrow} E(X)=\mu$$

Since this law is one of the two most important results in statistics (and will be used several times below), it is worth pointing out that unlike the other major result, namely the Central Limit Theorem, assertion (8.4) is obtainable by elementary means that are completely intuitive. To do so, recall first from expression (3.1.18) of Part II that we have already shown that $\bar{X}_n$ is an *unbiased estimator* of $\mu$, i.e., that for *n*,

(8.5) $$E(\bar{X}_n)=\mu$$

Moreover, if we let $\text{var}(X)=\sigma^2$, then it was shown in expression (3.1.19) of Part II that for all *n*,

(8.6) $$\text{var}(\bar{X}_n) = E[(\bar{X}_n-\mu)^2] = \sigma^2/n$$

In particular, this implies that the expected squared deviation, $E[(\bar{X}_n-\mu)^2]$, of $\bar{X}_n$ from $\mu$ must *shrink to zero* as *n* becomes large. But since the mean (center of mass) of $\bar{X}_n$ is always the same, namely at $\mu$, this implies that the probability distribution of $\bar{X}_n$ must eventually concentrate around $\mu$, as shown schematically in Figure 8.1 below:

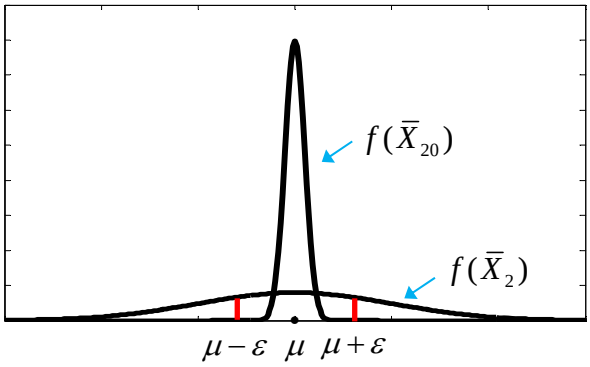



**Figure 8.1 Law of Large Numbers**

Here sample sizes $n=2,20$ are shown with $\sigma^2=1$, so that the respective sample-mean variances are given by 1/2 and 1/20.[1] For the particular epsilon interval $[\mu-\varepsilon,\mu+\varepsilon]$

[1] The densities plotted here are for $X \sim N(\mu,1)$.

shown, it is clear that almost all of the probability mass for $\overline{X}_{20}$ is already inside this interval. So even without a formal proof, it should be clear that $\overline{X}_n$ must converge in probability to $\mu$.[2]

Given the general consistency property in (8.3), the second major property of maximum-likelihood estimators is that their sampling distributions are always *asymptotically normal* with means given by the *true* parameter values. This can be expressed somewhat more formally (in a manner analogous to the Central Limit Theorems in Section 3 of Part II) by asserting that for sufficiently large sample sizes, *n*,

$$\hat{\theta}_n \approx_d N[\theta_0, \text{cov}(\hat{\theta}_n)] \tag{8.7}$$

where the relevant *covariance matrix*, $\text{cov}(\hat{\theta}_n)$, is here left undefined, and will be developed in more detail below. Note in particular from (8.7) that $\hat{\theta}_n$ is always an *asymptotically unbiased* estimator of $\theta_0$, i.e., that [3]

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta_0 \tag{8.8}$$

It is these asymptotic properties that make it possible to construct approximate *significance tests* for parameters even without knowing the exact distributions of maximum-likelihood estimators.

With respect to significance tests for SEM and SLM in particular, recall that all such tests in Figure 7.7 above use *z-values* [as in expression (7.5.7) above] rather than the standard *t-values* used for parameter significance tests in OLS models [as in expression (7.3.26) of Part II]. The reason for this is that even though we are assuming multi-normally distributed errors in both SEM and SLM, the exact distributions of estimators $(\hat{\beta}, \hat{\sigma}^2, \hat{\rho})$ for these models are *not* necessarily normal, or even expressible in closed form. So we *must* appeal to the *asymptotic normality* of such estimators to carry out significance tests, and it is for this reason that *z-values* are used. (See Section 8.4.1 below for further discussion of *z*-values versus *t*-values).

It should be evident here that (with the notable exception of the Central Limit Theorems developed in Section 3 of Part II) the present asymptotic analysis is the most technically challenging material in this NOTEBOOK. In view of this, our present objective is simply to illustrate these results by examples where these general asymptotic properties reduce to more familiar results obtainable by elementary means. We start with the classic example

[2] A formal proof amounts simply to *Chebyschev's Inequality*, which shows in the present case that for any $k > 0$, $\Pr(|\overline{X}_n - \mu| > (k/\sqrt{n})\sigma) \le 1/k^2$. So as long as $k$ increases more slowly than $\sqrt{n}$, both $k/\sqrt{n}$ and $1/k^2$ can be made arbitrarily small.

[3] It might seem obvious from (8.3) that condition (8.8) should hold. But in fact these two conditions are generally quite independent (i.e., each can hold without the other).

of estimating the mean of a univariate normal random variable in Section 8.1 below, and then proceed to a multivariate example in Section 8.2. This second example involves the General Linear Model, and will provide a useful conceptual framework for the SEM and SLM results to follow.

## 8.1 A Basic Example of Maximum Likelihood Estimation and Inference

To illustrate the general methods of parameter inference for maximum likelihood estimation it is instructive to begin with the single-parameter case of a *normally distributed* random variable, $Y \sim N(\mu,\sigma^2)$, with known variance, $\sigma^2$, but with *unknown* mean, $\mu$.[4] For a given random sample, $y=(y_1,..,y_n)$, the log likelihood of $\mu$ given $y$ then takes the familiar form [recall expression (3.2.6) and (3.2.7) in Part II]:

$$(8.1.1)\qquad L_n(\mu \mid y,\sigma^2) = \log\left(\prod_{i=1}^n f(y_i \mid \mu)\right) = \log\left(\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2}\right)$$

$$= \sum_{i=1}^n \left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \tfrac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right]$$

$$= -n\log\left(\sigma\sqrt{2\pi}\right) - \tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2$$

If we now use the simplifying notation $L_n'$ for first derivatives,[5] and solve the usual *first-order condition* for a maximum with respect to $\mu$, we see that

$$(8.1.2)\qquad 0 = L_n'(\mu \mid y) = \frac{d}{d\mu} L_n(\mu \mid y) = \tfrac{1}{\sigma^2}\sum_{i=1}^n (y_i-\mu) = \tfrac{1}{\sigma^2}\sum_{i=1}^n y_i - \tfrac{n}{\sigma^2}\mu$$

$$\Rightarrow \sum_{i=1}^n y_i - n\mu = 0 \;\Rightarrow\; \boxed{\hat{\mu}_n = \tfrac{1}{n}\sum_{i=1}^n y_i = \bar{y}_n}$$

and thus that $\hat{\mu}_n$ is precisely the *sample mean*, $\bar{y}_n$. So the main advantage of this example is that the *sampling distribution* of this particular maximum-likelihood estimator is obtainable by elementary methods.

[4] In fact, this is one of the prime examples used by early contributors to Maximum Likelihood Estimation, including Gauss (1896) and Edgeworth (1908), as well as in the subsequent definitive work of Fisher (1922). For an interesting discussion of these early developments see Hald, A (1999) "On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares", *Statistical Science*, 14: 214-222

[5] Be careful *not* to confuse this use of primes with that of vector and matrix transposes, like $A'$,

### 8.1.1 Sampling Distribution by Elementary Methods

Note first that *consistency* of this estimator is precisely the Law of Large Numbers in (8.4) with $X$ replaced by the random variable $Y$ in this case. As for the asymptotic normality condition in (8.7), we have a much sharper result for the sample mean. In particular, it follows as a very special case of the *Linear Invariance* property (Section 3.2.2 of Part II) of the multi-normal random vector, $(Y_1,..,Y_n)$, that the sample mean, $\overline{Y}_n = \Sigma_{i=1}^n(\frac{1}{n})Y_i$, is *exactly* normally distributed. In particular, if the *true* mean of $Y$ is $E(y) = \mu_0$, so that $Y \sim N(\mu_0,\sigma^2)$, then by linear invariance we obtain the exact sampling distribution of $\hat{\mu}_n$,

(8.1.3) $$\hat{\mu}_n = \overline{Y}_n \sim N(\mu_0,\sigma^2/n)$$

### 8.1.2 Sampling Distribution by General Maximum-Likelihood Methods

Given these well-known results for $\hat{\mu}_n$, we now consider how they would be obtained within the general theory of maximum-likelihood estimation. In the present case, the general asymptotic normality result in (8.7) asserts that

(8.1.4) $$\hat{\mu}_n \approx_d N[\mu_0, \text{var}(\hat{\mu}_n)]$$

which is clearly consistent with (8.1.3). From a practical perspective, our main objective will be to show how the large-sample variance, $\text{var}(\hat{\mu}_n)$, is calculated and to verify that it is precisely, $\sigma^2/n$. In doing so, we shall also illustrate the general strategy for analyzing the large sample properties of maximum-likelihood estimators. This will not only yield an asymptotic approximation to the variance of such estimators, but will also show why they are both consistent and asymptotically normal. (A more detailed development is give in Section A3.7 of the Appendix.)

***Standardized Likelihood Functions***

The key observation to be made here is that by replacing data values, $y_i$, with their associated random variables, $Y_i$, the log-likelihood function in (8.1.1) can be viewed as a sum of *iid* random variables, $X_i(\mu) \equiv \log f(Y_i \mid \mu)$,

(8.1.5) $$L_n(\mu \mid Y_1,..,Y_n) = \sum_{i=1}^n \log f(Y_i \mid \mu) = \sum_{i=1}^n X_i(\mu)$$

[where we now suppress the given parameter, $\sigma^2$, except when needed]. So if we divide both sides by *n* and let $\overline{L}_n = \frac{1}{n}L_n$, then this is seen to be the *sample mean* , $\overline{X}_n(\mu)$, for a sample of size *n* from the random variable, $X(\mu) \equiv \log f(Y \mid \mu)$, i.e.,

(8.1.6) $$\bar{L}_n(\mu \mid Y_1,..,Y_n) = \tfrac{1}{n}\sum_{i=1}^{n} \log f(Y_i \mid \mu) = \tfrac{1}{n}\sum_{i=1}^{n} X_i(\mu) = \bar{X}_n(\mu)$$

Thus, if we now denote the common mean of these random variables by

(8.1.7) $$\bar{L}(\mu) = E[X(\mu)] = E[\log f(Y \mid \mu)]$$

[where the expectation is with respect to $Y \sim N(\mu_0, \sigma^2)$], then it follows from the LLN that $\bar{L}_n(\mu \mid Y_1,..,Y_n)$ *converges in probability* to this mean, i.e.,

(8.1.8) $$\bar{L}_n(\mu \mid Y_1,..,Y_n) \underset{prob}{\rightarrow} \bar{L}(\mu)$$

Notice also that since $1/n$ is simply a positive constant, this transformation of $L_n$ has no effect on maxima. So the maximum-likelihood estimator, $\hat{\mu}_n$, for sample data, $(y_1,..,y_n)$, must still be given by

(8.1.9) $$\hat{\mu}_n = \max_{\mu \in \mathbb{R}} \bar{L}_n(\mu \mid y_1,..,y_n)$$

For purposes of analysis, this scaled version of $L_n$ thus constitutes a perfectly good "likelihood" function, and is usually designated as the *standardized likelihood function*. In these terms, the LLN ensures that these standardized likelihood functions, $\bar{L}_n(\cdot \mid y_1,..,y_n)$, must have a unique limiting form, $\bar{L}(\cdot)$, given by (8.1.8) which may be designated as the *limiting likelihood function*. This implies that essentially all large-sample properties of maximum-likelihood estimators can be studied in terms of this limiting form, and in particular, that the large-sample distribution of $\hat{\mu}_n$ can be obtained.

In the present case, we can learn a great deal by simply computing this limiting likelihood function. To do so, recall from (8.1.1) and (8.1.7) that

(8.1.10) $$\bar{L}(\mu) = E[\log f(Y \mid \mu)] = E\left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \tfrac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2\right]$$

$$= \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \tfrac{1}{2\sigma^2} E\left[(Y-\mu)^2\right]$$

$$= -\log\left(\sigma\sqrt{2\pi}\right) - \tfrac{1}{2\sigma^2} E\left[Y^2 - 2\mu Y + \mu^2\right]$$

$$= -\log\left(\sigma\sqrt{2\pi}\right) - \tfrac{1}{2\sigma^2}\left[E(Y^2) - 2\mu E(Y) + \mu^2\right]$$

But since $E(Y) = \mu_0$ and $E(Y^2) = \text{var}(Y) + [E(Y)]^2 = \sigma^2 + \mu_0^2$, it then follows that

$$(8.1.11) \qquad \bar{L}(\mu) = -\log(\sigma\sqrt{2\pi}) - \tfrac{1}{2\sigma^2}\left[(\sigma^2 + \mu_0^2) - 2\mu\,\mu_0 + \mu^2\right]$$

$$= -[\log(\sigma\sqrt{2\pi}) - \tfrac{1}{2}] - \tfrac{1}{2\sigma^2}(\mu_0^2 - 2\mu\,\mu_0 + \mu^2)$$

$$= c_\sigma - \tfrac{1}{2\sigma^2}(\mu_0 - \mu)^2$$

where $c_\sigma = \tfrac{1}{2} - \log(\sigma\sqrt{2\pi})$ is a constant depending only on $\sigma$. So we see that in the present case, $\bar{L}$ is a simple quadratic function, as shown by the solid black curve in Figure 8.2 below, where we have used the parameter values, $\mu_0 = 1$ and $\sigma^2 = 1$.

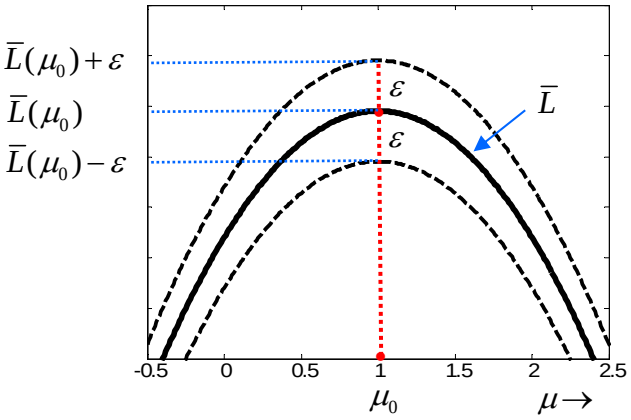



**Figure 8.2. Limit Curve and ε-Band**

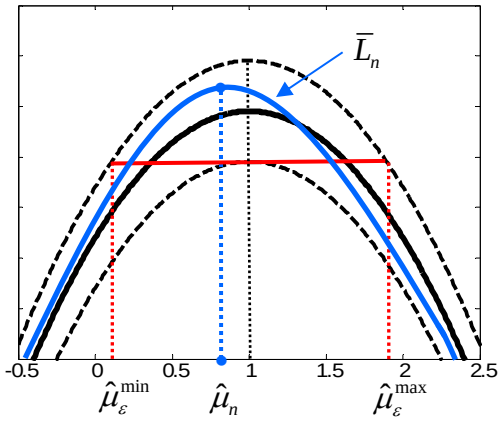



**Figure 8.3. Estimate Interval for ε-Band**

Notice also that this limiting function achieves its maximum at precisely the *true* mean value, $\mu_0$, which can be seen from the following first order condition (that is shown in the Appendix to hold for all limiting likelihood functions):

$$(8.1.12) \qquad \bar{L}'(\mu) = \tfrac{1}{\sigma^2}(\mu_0 - \mu) \;\Rightarrow\; \bar{L}'(\mu_0) = 0$$

Next observe from expression (8.1.8) that for any given value of $\mu$ on the horizontal axis, the likelihood values, $\bar{L}_n(\mu \mid y_1,..y_n)$, should eventually be very close to the limiting likelihood value, $\bar{L}(\mu)$, for all sufficiently large data samples, $(y_1,..,y_n)$. However, this does *not* imply that the entire function, $\bar{L}_n(\cdot \mid y_1,..y_n)$, will be close to the limiting function, $\bar{L}(\cdot)$. Here one requires a "uniform" version of probabilistic convergence (as detailed in the Appendix). For the present, it suffices to say that under mild regularity conditions, one can ensure uniform convergence in probability on any given interval containing the true mean, $\mu_0$, such as the interval, $I = [-0.5, 2.5]$, about $\mu_0 = 1$ shown Figure 8.2. What this means is that as sample sizes increase, realized likelihood

functions, $\bar{L}_n(\cdot \mid y_1,..y_n)$, will eventually be contained in any given $\varepsilon$-band on interval *I* (such as the one shown) with probability approaching one. One such realization, $\bar{L}_n$, is shown (schematically) by the blue curve in Figure 8.3,[6] with corresponding maximum-likelihood estimate, $\hat{\mu}_n$, also shown.

### *Consistency of* $\hat{\mu}_n$

This convergence property of likelihood functions, $\bar{L}_n$, in turn implies consistency of their associated maximum-likelihood estimates, $\hat{\mu}_n$. To see this, note that in order to stay inside this $\varepsilon$-band, each function, $\bar{L}_n$, evaluated at $\mu_0$ must achieve a value, $\bar{L}_n(\mu_0)$ in the interval of values $[\bar{L}(\mu_0)-\varepsilon, \bar{L}(\mu_0)+\varepsilon]$ shown on the left in Figure 8.2. Thus the *maximum* value of $\bar{L}_n$ must be at least $\bar{L}(\mu_0)-\varepsilon$, which means that this maximum (tangency) point on $\bar{L}_n$ must lie somewhere above the horizontal red line shown in Figure 8.3, as illustrated by the example in the figure. But since this maximum is by definition achieved at $\bar{L}_n(\hat{\mu}_n)$, this in turn implies that $\hat{\mu}_n$ must lie somewhere in the corresponding $\varepsilon$*-containment interval*, $[\hat{\mu}_\varepsilon^{\min}, \hat{\mu}_\varepsilon^{\max}]$, of $\mu$-values on the horizontal axis. Finally, observe that as $\varepsilon$-bands are chosen to be smaller, their corresponding $\varepsilon$-containment intervals must eventually shrink to the single value, $\mu_0$. So for sufficiently large sample sizes, *n*, we see that maximum-likelihood estimates, $\hat{\mu}_n$, must eventually be arbitrarily close to $\mu_0$ (with probability approaching one), and thus that such estimators satisfy *consistency*. While this consistency argument is certainly more complex than the direct appeal to the Law of Large Numbers for the simple case of $\hat{\mu}_n \equiv \bar{Y}_n$, it serves to illustrate the approach used for *all* maximum-likelihood estimators. Moreover, it helps to provide some geometric intuition for the *large-sample variance* of such estimators, to which we now turn.

### *Large-Sample Variance of* $\hat{\mu}_n$

First observe that by taking the *second derivative*, $\bar{L}'' = (d/d\mu)\bar{L}'$, of the limiting likelihood function and evaluating this at $\mu_0$, we see from (8.1.12) that

$$(8.1.13) \qquad \bar{L}''(\mu) = \tfrac{d}{d\mu}\bar{L}'(\mu) = -\tfrac{1}{\sigma^2} \;\Rightarrow\; \bar{L}''(\mu_0) = -\tfrac{1}{\sigma^2}$$

But for sufficiently large sample sizes, *n*, the scaled likelihood functions, $\bar{L}_n$, were seen to be *uniformly* close to $\bar{L}$ in the neighborhood of $\mu_0$, so that their *shapes* should be

[6] Here it is worth noting from expression (8.1.1) that like the limiting curve, $\bar{L}$, all such realizations, $\bar{L}_n$, in the present case must be smooth quadratic functions (such as the one shown).

similar to $\bar{L}$ in this neighborhood. Thus it is reasonable to expect that (8.1.13) should hold approximately for such functions, i.e., that

(8.1.14) $$\bar{L}_n''(\mu_0) \approx \bar{L}''(\mu_0) = -\tfrac{1}{\sigma^2}$$

But this in turn implies from the definition of $\bar{L}_n$ that the original log-likelihood functions, $L_n$, must satisfy:

(8.1.15) $$L_n(\mu_0) = n\bar{L}_n(\mu_0) \;\Rightarrow\; L_n''(\mu_0) = n\bar{L}_n''(\mu_0) \approx -\tfrac{n}{\sigma^2}$$

By inverting this expression and multiplying by -1, we see that

(8.1.16) $$-L_n''(\mu_0)^{-1} \;\approx\; \tfrac{\sigma^2}{n}$$

Finally, since we happen to *know* that the right hand side is precisely the *variance* of $\hat{\mu}_n = \bar{Y}_n$, we see that this variance is well approximated by the negative inverse of the second derivative of the log-likelihood function, $L_n$, evaluated at the true mean, $\mu_0$, i.e.,

(8.1.17) $$\text{var}(\hat{\mu}_n) \approx -L_n''(\mu_0)^{-1}$$

While all of this might seem to be purely coincidental, it is shown in the Appendix that this relation is *always true* for maximum-likelihood estimators. More importantly for our present purposes, this geometric argument actually suggests *why* this should be so. To begin with, note that while the first derivative, $\bar{L}'(\mu)$, of the limiting likelihood function reveals its *slope* at each point, $\mu$, the second derivative, $\bar{L}''(\mu)$ reveals its *curvature*, i.e., rate of change of slope at $\mu$. So $\bar{L}''(\mu_0)$ corresponds geometrically to the curvature of the limiting likelihood function at the true mean, $\mu_0$. With this in mind we now illustrate the effects of such curvature in Figures 8.4 and 8.5 below.

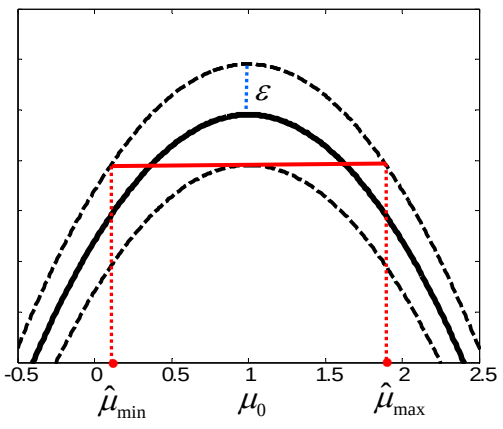



**Figure 8.4. Estimate Interval for $\sigma^2 = 1$**

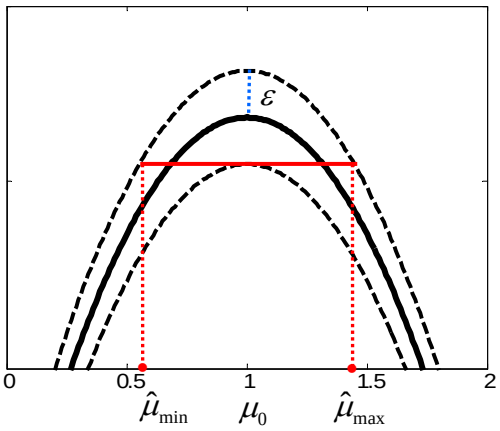



**Figure 8.5. Estimate Interval for $\sigma^2 = 1/4$**

Figure 8.4 simply repeats the relevant features of Figure 8.3. Here we used the variance parameter, $\sigma^2 = 1$, which implies from (8.1.13) that $\bar{L}''(\mu_0) = -1$. Note in particular that this negative sign reflects the *concavity* of $\bar{L}$ required for a maximum at $\mu_0$. In Figure 8.5 we have used the same value of $\varepsilon$ for the $\varepsilon$-band around function, $\bar{L}$, but have now reduced the variance parameter to $\sigma^2 = 1/4$. This in turn is seen to yield more extreme curvature, $\bar{L}''(\mu_0) = -4$, at $\mu_0$. But the key point to notice is that this sharper curvature necessarily compresses the corresponding $\varepsilon$*-containment interval* that delimits the feasible range of maximum-likelihood estimates, $\hat{\mu}_n$, for large *n*. By comparing these two figures, one can see that the permissible deviations from $\mu_0$ in Figure 8.5 are only about half those in Figure 8.4 (decreasing from about 0.8 to 0.4). This in turn implies that the permissible *squared* deviations are only about a quarter as large. Moreover, the constancy of curvature in the present example implies that this same relation must hold for all $\varepsilon$, and thus that the *expected* squared deviations of $\hat{\mu}_n$ should also be about a quarter as large. But this is precisely the relative *variance* of $\hat{\mu}_n$ at each level of curvature. In short, we see that for large samples, *n*, with log-likelihoods close to the limiting likelihood, the desired variance of $\hat{\mu}_n$ is indeed (inversely) proportional to negative curvature, as in (8.1.17).

Finally, it should be noted that while the constancy of curvature in this example makes such relations easier to see, this is of course a very special case. More generally, all that can be said is that for *sufficiently large* samples, *n*, almost all realizations of $\hat{\mu}_n$ will be so close to $\mu_0$ that curvature can be treated as constant over the relevant range of $\hat{\mu}_n$.

### *Asymptotic Normality of* $\hat{\mu}_n$

To motivate asymptotic normality of this estimator, note first that the exact sample-mean form of $\hat{\mu}_n$ allowed a direct application of the CLT by simply rescaling this estimator. However, since the exact form of such estimators is rarely available, general arguments for asymptotic normality are necessarily more subtle in nature. But if one assumes that likelihood functions are sufficiently smooth, then one can appeal to local approximation arguments. In particular, if the standardized likelihood function, $\bar{L}_n$, is *continuously twice differentiable* (as in normal-density cases) then the tangent line defined by the first derivative, $\bar{L}_n'$, at the true parameter value, $\mu_0$, yields the best local linear approximation to values, $\bar{L}_n'(\mu)$, in the neighborhood of $\mu_0$. This tangent line [with slope, $\bar{L}_n''(\mu_0)$], is shown (in red) in Figure 8.6 below. But by the consistency result above, it follows that for sufficiently large *n* we may assume that the estimator $\hat{\mu}_n$ is itself close to $\mu_0$, and thus that $\bar{L}_n'$ is approximately linear on the interval between $\mu_0$ and $\hat{\mu}_n$, as shown in the figure.

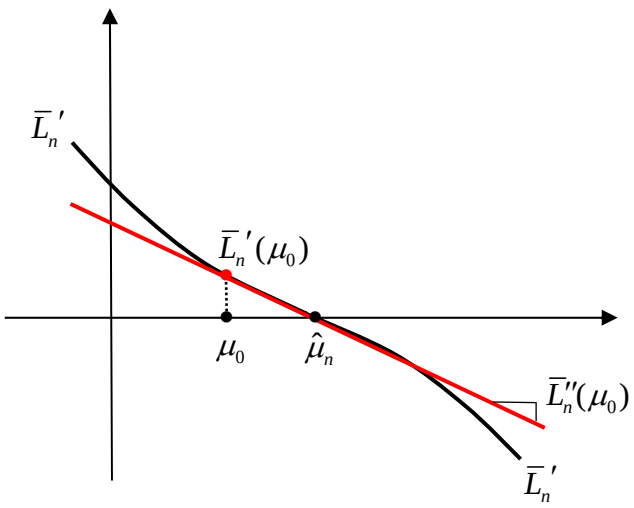



**Figure 8.6. Local Linear Approximation**

More formally, this implies that the slope, $\bar{L}_n''(\mu_0)$, can be approximated by the ratio of differences,

$$\frac{\bar{L}_n'(\hat{\mu}_n) - \bar{L}_n'(\mu_0)}{\hat{\mu}_n - \mu_0} \approx \bar{L}_n''(\mu_0) \tag{8.1.18}$$

So by appealing once more to (8.1.14), we see that this ratio should be approximately *constant* for all large *n*, i.e., that

$$\frac{\bar{L}_n'(\hat{\mu}_n) - \bar{L}_n'(\mu_0)}{\hat{\mu}_n - \mu_0} \approx \bar{L}''(\mu_0) \tag{8.1.19}$$

Finally, noting that by definition, $\bar{L}_n'(\hat{\mu}_n) = 0$, it follows by letting $a_0 = -1/\bar{L}''(\mu_0)$ that

$$\hat{\mu}_n - \mu_0 \approx a_0 \bar{L}_n'(\mu_0) \tag{8.1.20}$$

Note in particular from (8.1.14) that this relation holds for the quadratic likelihood function (8.1.1) with $a_0 = \sigma^2$.

More generally, the local linear relation in (8.1.20) implies that limiting distribution of $\hat{\mu}_n - \mu_0$ must be essentially the same as that of $\bar{L}_n'(\mu_0)$. But the limiting distribution of $\bar{L}_n'(\mu_0)$ is readily obtainable. To see this, recall from the definition of $\bar{L}_n$ in expression (8.1.6) that if we let $L(\mu \mid y) = \log f(y \mid \mu)$ [and employ the linearity properties of derivatives] then

(8.1.21) $$\bar{L}_n(\mu_0 \mid y_1,..,y_n) = \tfrac{1}{n}\sum_{i=1}^n \log f(y_i \mid \mu_0) = \tfrac{1}{n}\sum_{i=1}^n L(\mu_0 \mid y_i)$$

$$\Rightarrow \bar{L}'_n(\mu_0) = \bar{L}'_n(\mu_0 \mid y_1,..,y_n) = \tfrac{1}{n}\sum_{i=1}^n L'(\mu_0 \mid y_i)$$

Thus $\bar{L}'_n(\mu_0)$ is itself seen to be a sample mean of classical form with *iid* random variables, $L'(\mu_0 \mid Y_i)$, $i = 1,..,n$. But since the classical the CLT shows that a simple rescaling of sample means must be asymptotically normally distributed, it follows from (8.1.19) that essentially the same rescaling will apply to $\hat{\mu}_n - \mu_0$ . As detailed in Section A3.7.3 of the Appendix, the appropriate scale factor turns out to be $\sqrt{n}$. So by recalling from (8.1.17) that $\text{var}(\hat{\mu}_n) \approx -\bar{L}''_n(\mu_0)^{-1}$ , we obtain the limiting result

(8.1.22) $$\sqrt{n}\left(\hat{\mu}_n - \mu_0\right) \approx_d N\left(0, -\bar{L}''_n(\mu_0)^{-1}\right)$$

Thus to apply this large-sample distribution, all that remains to be done is to compute the large-sample variance in (8.1.22).

***Computation of Large-Sample Variance***

Since direct computation of the variance in (8.1.22) requires knowledge of $\mu_0$, we must estimate this quantity. But since $\hat{\mu}_n$ is a consistent estimator of $\mu_0$, it is natural to use the estimated variance:

(8.1.23) $$\widehat{\text{var}}(\hat{\mu}_n) = -L''_n(\hat{\mu}_n)^{-1}$$

In practice this can be calculated by numerically approximating the second derivative of the log-likelihood function at the maximum, i.e., $L''_n(\hat{\mu}_n) = L''_n(\hat{\mu}_n \mid y_1,..,y_n)$. However, when this second derivative can be computed as an *explicit* function of the sample data, $(y_1,..,y_n)$, it is often more appropriate to use mean values.

By way of motivation, it should be noted that perhaps the weakest link in the chain of arguments above was the supposition that curvature of the likelihood function, $\bar{L}_n(\mu_0)$, at the true mean is well approximated by that of the limiting likelihood function, $\bar{L}(\mu_0)$, i.e., that $\bar{L}''_n(\mu_0) \approx \bar{L}''(\mu_0)$, as in expression (8.1.14) above. Since averaging produces smoothing effects, it should thus be more reasonable to suppose that

(8.1.24) $$E\left[\bar{L}''_n(\mu_0)\right] = E\left[\bar{L}''_n(\mu_0 \mid Y_1,..,Y_n)\right] \approx \bar{L}''(\mu_0)$$

and then to approximate the large-sample variance by:[7]

(8.1.25) $$\text{var}(\hat{\mu}_n) \approx -\left(E[L_n''(\mu_0)]\right)^{-1} = -\left(E[L_n''(\mu_0 \mid Y_1,..,Y_n)]\right)^{-1}$$

Finally we note that the negative expected curvature value used in (8.1.20) is of much wider interest, and (in honor of its discoverer) is usually designated as *Fisher information*,[8]

(8.1.26) $$\mathcal{I}_n(\mu_0) = -E\left[L_n''(\mu_0 \mid Y_1,..,Y_n)\right]$$

In these terms, the variance approximation in (8.1.20) can be rewritten as

(8.1.27) $$\text{var}(\hat{\mu}_n) \approx [\mathcal{I}_n(\mu_0)]^{-1}$$

Note in particular that since higher values of $\mathcal{I}_n(\mu_0)$ mean lower variances of $\hat{\mu}_n$ and thus sharper estimates of $\mu_0$, this measure does indeed reflect the amount "information" in $L_n$ about $\mu_0$. For computational purposes, we must again substitute $\hat{\mu}_n$ for $\mu_0$, to obtain the *large-sample variance estimate*,

(8.1.28) $$\widehat{\text{var}}(\hat{\mu}_n) = [\mathcal{I}_n(\hat{\mu}_n)]^{-1}$$

In subsequent sections it will be shown that explicit expressions for Fisher information can be obtained for both SEM and SLM. So we shall employ this expectation version of variance estimates in our analyses of these models. Thus, to avoid any possible confusion of (8.1.28) with (8.1.23) above, we now follow the standard convention of designating the more direct estimate of negative curvature in (8.1.23) as *observed Fisher information*,

(8.1.29) $$\mathcal{I}_n^{obs}(\mu_0) = -L_n''(\mu_0)$$

and thus redesignate (8.1.23) as *observed large-sample variance estimate*,

(8.1.30) $$\widehat{\text{var}}_{obs}(\hat{\mu}_n) = [\mathcal{I}_n^{obs}(\hat{\mu}_n)]^{-1}$$

[7] As shown in Section A3.7.3 of the Appendix, it is this expected-curvature expression that is used in formal convergence proofs. So while both approximations are used in practice, the main advantage of the (8.1.17) approach is that it allows the role of geometric curvature to be seen more easily.
[8] Alternative definitions of Fisher Information are developed in Section A3.7.3 of the Appendix.

## 8.2 Sampling Distributions for General Linear Models with Known Covariance

We next develop a *multi-parameter* example in which the sampling distributions of parameter estimates can again be obtained by elementary methods. Here we start with following *General Linear Model*

(8.2.1) $$Y = X\beta + \varepsilon, \;\; \varepsilon \sim N(0,V)$$

where the covariance matrix, *V*, is assumed to be *known*. As in Section 7.2.2, this in turn implies that

(8.2.2) $$Y \sim N(X\beta,V)$$

Before proceeding with this case, it is worth noting that (8.2.2) is in fact a direct extension of our previous model in Section 8.2.1. In particular, that model is seen to be the special case in which, $V = \sigma^2 I_n$ with $\sigma^2$ known, and in which $X = 1_n$. Thus $\beta$ reduces to a single parameter, $\beta = (\beta_0) = \mu$, in this case, and we see that:

(8.2.3) $$Y \sim N(\mu 1_n, \sigma^2 I_n) \;\Rightarrow\; Y_i \underset{iid}{\sim} N(\mu,\sigma^2)\,,\; i = 1,..,n$$

So it is perhaps not surprising that the same methods above can be applied to this more general version.[9]

### 8.2.1 Sampling Distribution by Elementary Methods

As mentioned above, the only difference here is that we are now in a *multi-parameter* setting where sampling distributions must be obtained for the vector of maximum-likelihood estimators in (7.2.18) above, i.e., for

(8.2.4) $$\hat{\beta}_n = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

But since this is simply a linear transformation of the random vector, $Y$, we can obtain the sampling distribution of $\hat{\beta}$ by again appealing directly to expression (3.2.2) of the *Linear Invariance Theorem.* To do so, note simply that if we let

(8.2.5) $$A = (X'V^{-1}X)^{-1}X'V^{-1}$$

so that $\hat{\beta}_n = AY$, then it follows at once from (3.2.2) together with (8.2.2) above that

(8.2.6) $$\hat{\beta}_n \sim N(AX\beta, AVA')\;,$$

[9] Here we ignore questions of consistency, which involve a somewhat more complex application of the Law of Large Numbers [as for example in Theorem 10.2 in Green (2003)].

and thus that $\hat{\beta}_n$ is *exactly* multi-normally distributed. Moreover, as we have already shown in expressions (7.3.21) and (7.3.22) of Part II, $\hat{\beta}_n$ has *mean vector*

$$(8.2.7) \qquad E(\hat{\beta}_n) = AX\beta = (X'V^{-1}X)^{-1}(X'V^{-1}X)\beta = \beta$$

and *covariance matrix*,

$$\begin{aligned}(8.2.8) \qquad \text{cov}(\hat{\beta}_n) &= \text{cov}\left[(X'V^{-1}X)^{-1}X'V^{-1}\varepsilon\right] \\ &= (X'V^{-1}X)^{-1}X'V^{-1}\,\text{cov}(\varepsilon)V^{-1}X\,(X'V^{-1}X)^{-1} \\ &= (X'V^{-1}X)^{-1}X'V^{-1}VV^{-1}X\,(X'V^{-1}X)^{-1} \\ &= (X'V^{-1}X)^{-1}(X'V^{-1}X)(X'V^{-1}X)^{-1} \\ &= (X'V^{-1}X)^{-1}\end{aligned}$$

so that the *exact sampling distribution* of $\hat{\beta}_n$ is given by

$$(8.2.9) \qquad \boxed{\hat{\beta}_n \sim N[\beta,(X'V^{-1}X)^{-1}]}$$

As in the single-parameter case above, this distribution allows us to construct significance tests for all $\beta_j$ coefficients.

### 8.2.2 Sampling Distribution by General Maximum-Likelihood Methods

Here we shall focus only on those aspects of the Maximum-Likelihood approach that are needed for calculating the desired sampling distribution of $\hat{\beta}_n$, as in (8.2.9) above. Thus, as a direct extension of expression (8.1.4), we start by assuming that $\hat{\beta}_n$ is both *asymptotically multi-normal* and *asymptotically unbiased*, i.e.,

$$(8.2.10) \qquad \hat{\beta}_n \approx_d N[\beta,\text{cov}(\hat{\beta}_n)]$$

So the main task is to estimate the covariance matrix, $\text{cov}(\hat{\beta}_n)$, in (8.2.10). To do so, recall from expression (8.1.17) that the desired asymptotic variance estimate was obtained in terms of the second derivative of the log-likelihood function evaluated at the true parameter value. Exactly the same result is true in the multi-parameter case, except that here we must calculate *partial derivatives* of the log-likelihood function with respect to all parameters. The details of partial derivatives for both the scalar and multi-dimensional case are developed in Sections A2.5 through A2.7 in the Appendix to Part II

(which we here designate as Appendix A2). In the present case, the log-likelihood function is precisely the same as that in expression (7.2.14) above with $\sigma^2 \equiv 1$, i.e.,

$$(8.2.11) \qquad L(\beta \mid y) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|V| - \tfrac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)$$

As shown for the OLS case in Section A2.7.3 of Appendix A2, maximizing this function with respect to parameter vector, $\beta$, amounts to setting all partial derivatives of $L(\beta \mid y)$ equal to zero, where the vector of partial derivatives is called the *gradient vector* of $L(\beta \mid y)$, and is written as:

$$(8.2.12) \qquad \nabla_\beta L(\beta \mid y) = \begin{pmatrix} \frac{\partial}{\partial \beta_1} L(\beta \mid y) \\ \vdots \\ \frac{\partial}{\partial \beta_k} L(\beta \mid y) \end{pmatrix}$$

But since $\beta$ only appears in the last term of (8.2.11), it follows that this first order condition for a maximum reduces to:

$$\begin{aligned} (8.2.13) \qquad 0 = \nabla_\beta L(\beta \mid y) &= \nabla_\beta [-\tfrac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)] \\ &= -\tfrac{1}{2}\nabla_\beta [y'V^{-1}y - 2y'V^{-1}X\beta + 2\beta'X'V^{-1}X\beta] \\ &= X'V^{-1}y - X'V^{-1}X\beta \end{aligned}$$

where the last line follows from expressions (A2.7.7) and (A2.7.11) of Appendix A2. Notice that solving this expression for $\beta$ yields precisely the maximum-likelihood estimate in (8.2.4) above. But our present interest in the matrix of *second* partial derivatives of $L$ at the true value of $\beta$, say $\beta_0$,[10]

$$(8.2.14) \qquad \nabla_{\beta\beta} L(\beta_0 \mid y) = \left[\nabla_\beta [\nabla_\beta L(\beta \mid y)]\right]_{\beta=\beta_0} = \begin{pmatrix} \frac{\partial^2}{\partial \beta_1^2} L(\beta_0 \mid y) & \cdots & \frac{\partial^2}{\partial \beta_1 \partial \beta_k} L(\beta_0 \mid y) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \beta_k \partial \beta_1} L(\beta_0 \mid y) & \cdots & \frac{\partial^2}{\partial \beta_k^2} L(\beta_0 \mid y) \end{pmatrix}$$

which (as in Section A2.7 of Appendix A2) is designated as the *Hessian matrix* for $L(\beta \mid y)$ evaluated at $\beta_0$. So by the last line of (8.2.13) [together with (A2.7.7) in Appendix A2][11], we see that

$$(8.2.15) \qquad \nabla_{\beta\beta} L(\beta_0 \mid y) = \nabla_\beta [X'V^{-1}y - X'V^{-1}X\beta]_{\beta=\beta_0}$$

[10] Note that the *intercept* coefficient in $\beta$ is here designated as "$\beta_1$" precisely to avoid notational conflicts with this *true* coefficient vector, $\beta_0$.

[11] In particular, it follows from (A2.7.7) that for any *symmetric* matrix, $B = (b_1', .., b_n')'$, we must have $\nabla_x Bx = (\nabla_x b_1'x, .., \nabla_x b_n'x) = B' = B$.

$$= -\nabla_{\beta}(X'V^{-1}X\beta)$$

$$= -X'V^{-1}X$$

But, if we again replace data vector, $y$, by its corresponding random vector, $Y$, and take expectations (under $\beta_0$) then we see in this case that

(8.2.16) $$E[\nabla_{\beta\beta} L(\beta_0 \mid Y)] = -X'V^{-1}X$$

Thus by (8.2.8) this is now seen to imply that

(8.2.17) $$\text{cov}(\hat{\beta}_n) = (X'V^{-1}X)^{-1} = \left(-E[\nabla_{\beta\beta} L(\beta_0 \mid Y)]\right)^{-1}$$

So if *Fisher information* in (8.1.21) is here replaced by the corresponding *Fisher Information matrix*,

(8.2.18) $$\mathcal{I}_n(\beta_0) = -E[\nabla_{\beta\beta} L(\beta_0 \mid Y)]$$

then it follows from (8.2.17) that the covariance of $\hat{\beta}_n$ is precisely the inverse of the Fisher Information matrix, i.e.,

(8.2.19) $$\boxed{\text{cov}(\hat{\beta}_n) = \mathcal{I}_n(\beta_0)^{-1}}$$

While this is of course a very special case in which covariance is *exactly* inverse Fisher Information, it serves to motivate the general results to follow.

### 8.3 Asymptotic Sampling Distributions for the General Case

Given this multi-parameter example, it can be shown that the asymptotic sampling distributions for general maximum likelihood estimates are essentially of the same form. In particular, if the log-likelihood function for $n$ samples, $y = (y_1,..,y_n)'$ from a distribution with $k$ unknown parameters, $\theta = (\theta_1,..,\theta_k)'$, is denoted by $L(\theta \mid y)$ [as in (7.1.8) above], and if the maximum-likelihood estimator for $\theta$ is denoted by $\hat{\theta}_n$ [as in (7.1.9), with sample size, $n$, made explicit], then (under mild regularity conditions) $\hat{\theta}_n$ is both *asymptotically multi-normal* and *asymptotically unbiased*, i.e.,

(8.3.1) $$\hat{\theta}_n \approx_d N[\theta_0, \text{cov}(\hat{\theta}_n)]$$

where $\theta_0$ is the true value of $\theta$. So expressions (8.1.4) and (8.2.10) are both seen to be instances of this general expression. Moreover, the asymptotic covariance matrix,

$\text{cov}(\hat{\theta}_n)$, takes the same form as expression (8.2.17) through (8.2.19). In particular, if we again denote the relevant *Fisher Information matrix* by

(8.3.2) $$\mathcal{I}_n(\theta_0) = -E[\nabla_{\theta\theta} L(\theta_0 \mid Y)]$$

then the asymptotic covariance of $\hat{\beta}_n$ is precisely the inverse of the Fisher Information matrix, i.e.,

(8.3.3) $$\text{cov}(\hat{\theta}_n) = \mathcal{I}_n(\theta_0)^{-1}$$

So in these terms, the explicit form of (8.3.1) is simply

(8.3.4) $$\hat{\theta}_n \approx_d N[\theta_0, \mathcal{I}_n(\theta_0)^{-1}]$$

Finally, this distribution is again made operational by appealing to the consistency of $\hat{\theta}_n$ to replace $\theta_0$ with $\hat{\theta}_n$ and write

(8.3.5) $$\boxed{\hat{\theta}_n \approx_d N[\theta_0, \mathcal{I}_n(\hat{\theta}_n)^{-1}]}$$

But taking this approximation to be the relevant sampling distribution for $\hat{\theta}_n$, one can proceed with a range of statistical analyses regarding the nature of $\theta_0$. But rather than developing testing procedures within this general framework, it is more convenient to do so in the specific contexts of SEM and SAR, to which we now turn.

## 8.4 Parameter Significance Tests for SEM

For the SE-model in (7.3.1) through (7.3.3), it follows at once that the relevant *parameter vector* is given by $\theta = (\beta, \rho, \sigma^2)$,[12] with likelihood function,

(8.4.1) $$L(\theta \mid y) = L(\beta, \sigma^2, \rho \mid y)$$

$$= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2}\log|B_\rho| - \tfrac{1}{2\sigma^2}(y - X\beta)' B_\rho' B_\rho (y - X\beta)$$

where $B_\rho = I_n - \rho W$. For this model, if we now designate the sum of diagonal elements of any matrix, $A = (a_{ij})$, as the *trace* of $A$, written $tr(A) = \Sigma_i a_{ii}$, and if (for notational

[12] For notational simplicity, we here drop transposes and implicitly assume that both $\theta$ and $\beta$ are column vectors.

simplicity) we let $G_\rho = WB_\rho^{-1} = B_\rho^{-1}W$,[13] then it can be shown [see Ord (1975) and Appendix B in Doreian (1980)] that the expected value of the Hessian matrix for $L(\theta|Y)$ evaluated at the true value of $\theta$ is given by [14]

$$(8.4.2) \qquad E[\nabla_{\theta\theta} L(\theta|Y)] = E(\nabla_{\theta\theta} L) = \begin{pmatrix} E(\nabla_{\beta\beta}L) & E(\nabla_{\beta\sigma^2}L) & E(\nabla_{\beta\rho}L) \\ E(\nabla_{\sigma^2\beta}L) & E(\nabla_{\sigma^2\sigma^2}L) & E(\nabla_{\sigma^2\rho}L) \\ E(\nabla_{\rho\beta}L) & E(\nabla_{\rho\sigma^2}L) & E(\nabla_{\rho\rho}L) \end{pmatrix}$$

$$= -\begin{pmatrix} \frac{1}{\sigma^2} X'B_\rho' B_\rho X & 0 & 0 \\ 0' & \frac{n}{2\sigma^4} & \frac{1}{\sigma^2} tr(G_\rho) \\ 0' & \frac{1}{\sigma^2} tr(G_\rho) & tr[G_\rho(G_\rho + G_\rho')] \end{pmatrix}$$

where for simplicity we now drop the subscripts "0" denoting "true" values. It then follows at once from (8.3.2) and (8.3.3) that *asymptotic covariance matrix* of the maximum-likelihood estimators, $\hat{\theta} = (\hat{\beta}, \hat{\rho}, \hat{\sigma}^2)$,[15] is given by

$$(8.4.3) \qquad \text{cov}(\hat{\theta}) = \mathcal{I}_n(\theta)^{-1} = -E[\nabla_{\theta\theta} L(\theta|Y)]^{-1}$$

$$= \begin{pmatrix} \frac{1}{\sigma^2} X'B_\rho' B_\rho X & 0 & 0 \\ 0' & \frac{n}{2\sigma^4} & \frac{1}{\sigma^2} tr(G_\rho) \\ 0' & \frac{1}{\sigma^2} tr(G_\rho) & tr[G_\rho(G_\rho + G_\rho')] \end{pmatrix}^{-1}$$

Thus the desired *asymptotic sampling distribution* of $(\hat{\beta}, \hat{\sigma}^2, \hat{\rho})$ for SEM is given by

$$(8.4.4) \qquad \begin{pmatrix} \hat{\beta} \\ \hat{\sigma}^2 \\ \hat{\rho} \end{pmatrix} \sim N\left[ \begin{pmatrix} \beta \\ \sigma^2 \\ \rho \end{pmatrix}, \begin{pmatrix} \frac{1}{\sigma^2} X'B_\rho' B_\rho X & 0 & 0 \\ 0' & \frac{n}{2\sigma^4} & \frac{1}{\sigma^2} tr(G_\rho) \\ 0' & \frac{1}{\sigma^2} tr(G_\rho) & tr[G_\rho(G_\rho + G_\rho')] \end{pmatrix}^{-1} \right]$$

[13] The last equality follows from the fact that $WB_\rho^{-1} = W(\Sigma_{n=0}^{\infty} \rho^n W^n) = (\Sigma_{n=0}^{\infty} \rho^n W^n)W = B_\rho^{-1}W$. Because of this, $G_\rho$ is defined both ways in the literature [compare for example Ord (1975) and Doreian (1980)].

[14] Note in the last line of (8.4.2) that 0 denotes a *zero vector* of the same length as $\beta$, together with its transpose, $0'$.

[15] Again for notational simplicity, we now drop the sample-size subscripts (*n*) on estimators.

Before applying this distribution to construct specific tests, it is of interest to notice from the pattern of zeros inside this covariance matrix that further simplifications are possible here. In particular this covariance matrix is seen to be the inverse of a *block diagonal* matrix. But just as in the case of simple diagonal matrices, matrix multiplication shows that the inverse of any $(2\times 2)$ block diagonal matrix is given by

$$(8.4.5) \qquad \begin{pmatrix} A & \\ & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & \\ & B^{-1} \end{pmatrix}$$

so that (8.4.3) can be rewritten as

$$(8.4.6) \qquad \text{cov}(\hat{\theta}) = \begin{bmatrix} \sigma^2(X'B_\rho'B_\rho X)^{-1} & \\ & \begin{pmatrix} \frac{n}{2\sigma^4} & \frac{1}{\sigma^2}tr(G_\rho) \\ \frac{1}{\sigma^2}tr(G_\rho) & tr[G_\rho(G_\rho+G_\rho')] \end{pmatrix}^{-1} \end{bmatrix}$$

In particular, this shows that $\hat{\beta}$ is *uncorrelated* with either $\hat{\sigma}^2$ or $\hat{\rho}$, so that by the general properties of multi-normal distributions, we may conclude that $\hat{\beta}$ is completely *independent* of $(\hat{\sigma}^2,\hat{\rho})$. This in turn implies that the joint distribution of $(\hat{\beta},\hat{\rho},\hat{\sigma}^2)$ in (8.4.4) can be factored into a product of the marginal distributions for $\hat{\beta}$ and $(\hat{\sigma}^2,\hat{\rho})$. With respect to $\hat{\beta}$ in particular, this marginal distribution is seen to be of the form

$$(8.4.7) \qquad \hat{\beta} \sim N[\beta, \sigma^2(X'B_\rho'B_\rho X)^{-1}]$$

But since the covariance expression can be rewritten as

$$(8.4.8) \qquad \sigma^2(X'B_\rho'B_\rho X)^{-1} = \left(X'[\sigma^2(B_\rho'B_\rho)^{-1}]^{-1}X\right)^{-1} = (X'V_\rho^{-1}X)^{-1}$$

where $V_\rho = \sigma^2(B_\rho'B_\rho)^{-1}$, it follows from expressions (8.2.8) [together with (6.1.6) and (6.1.7)] that expression (8.4.6) is precisely the instance of the GLS in Section 7.6.2 above for the case of an SE-model with known spatial dependency parameter given by $\rho$. In other words, the independency property of SEM allows all analyses of $\hat{\beta}$ to be carried out using the GLS model in (8.2.9) for any given values of the other parameters, $(\sigma^2,\rho)$. As we shall see below, this simplification is *not* true for SLM.

### 8.4.1 Parametric Tests for SEM

To develop appropriate tests of parameters for SEM, recall that all unknown true parameter values $(\beta,\sigma^2,\rho)$ in the above expressions are estimated using $(\hat{\beta},\hat{\sigma}^2,\hat{\rho})$. So in these terms, the estimated asymptotic covariance for purposes of statistical inference is

obtained by substituting these estimated values $(\hat{\beta},\hat{\sigma}^2,\hat{\rho})$ into (8.4.6). For notational convenience, we denote this *estimated covariance matrix* by $S_{SEM}$, which is now seen to have the block diagonal form:

(8.4.9)
$$S_{SEM} = \begin{bmatrix} \hat{\sigma}^2(X'B'_{\hat{\rho}}B_{\hat{\rho}}X)^{-1} & \\ & \hat{\sigma}^4\begin{pmatrix} n/2 & \hat{\sigma}^2 tr(G_{\hat{\rho}}) \\ \hat{\sigma}^2 tr(G_{\hat{\rho}}) & \hat{\sigma}^4 tr[G_{\hat{\rho}}(G_{\hat{\rho}}+G'_{\hat{\rho}})] \end{pmatrix}^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} s^2_{\hat{\beta}_0} & \cdots & s_{\hat{\beta}_0\hat{\beta}_k} & & \\ \vdots & \ddots & \vdots & & \\ s_{\hat{\beta}_k\hat{\beta}_0} & \cdots & s^2_{\hat{\beta}_k} & & \\ & & & s^2_{\hat{\sigma}^2} & s_{\hat{\sigma}^2\hat{\rho}} \\ & & & s_{\hat{\rho}\hat{\sigma}^2} & s^2_{\hat{\rho}} \end{bmatrix}$$

Before using these estimates, notice from (8.4.6) that we have here factored out the quantity, $\hat{\sigma}^4$, in the lower diagonal matrix [see also Ord (1975), expression (19) in Doreian (1980), and expression (5.48) of Upton and Fingleton (1985) – which is class reference number 18]. The reason for this can be seen from the first row of the (unfactored) lower block-diagonal matrix in (8.4.6), which will have all elements close to zero when variance, $\sigma^2$, is large. So to avoid possible numerical stability problems when *computing* the inverse of this matrix, it is convenient to introduce such a factorization.

To apply these estimates, we first consider the most important set of parameter tests, namely tests for beta coefficients, $\beta_j$, $j = 0,1,..,k$, in the linear term, $X\beta$ (where as usual, the intercept, $\beta_0$, tends to be of less interest than the slope coefficients, $\beta_1,..,\beta_k$, for explanatory variables). While such analyses are is conceptually similar to those for Geo-Regression in Section 7.3.2 above, there are sufficient differences to warrant a more careful development here. For parameter, $\beta_j$, it follows at once from (8.4.7) that the marginal distribution of the estimator, $\hat{\beta}_j$, must be normal and of the form

(8.4.10) $$\hat{\beta}_j \sim N[\beta_j, \text{var}(\hat{\beta}_j)] \quad , \quad j = 0,1,..,k$$

where $\text{var}(\hat{\beta}_j)$ is by definition the $j^{th}$ diagonal element of the covariance matrix, $\sigma^2(X'B'_\rho B_\rho X)^{-1}$. Here we have estimated this quantity by diagonal element, $s^2_{\hat{\beta}_j}$, of matrix, $S_{SEM}$, in (8.4.9). So by using this estimate, (8.4.10) can be rewritten as,

(8.4.11) $$\hat{\beta}_j \sim N(\beta_j, s^2_{\hat{\beta}_j}) \ , \quad j = 0,1,..,k$$

This is the operational form of the sampling distribution that we shall employ for testing the significance of $\beta_j$. To employ the standard normal tables, one must first standardize $\hat{\beta}_j$ to obtain the corresponding *z-statistic*:

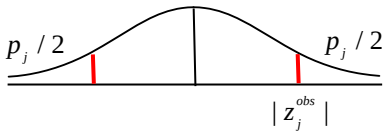
(8.4.12) $$z_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim N(0,1), \quad j = 0,1..,k$$

where $s_{\hat{\beta}_j} = \sqrt{s^2_{\hat{\beta}_j}}$ denotes the standard deviation of $\hat{\beta}_j$. So under the null hypothesis, $\beta_j = 0$, it follows from (8.4.12) that the *z-score*, $z_j = \hat{\beta}_j / s_{\hat{\beta}_j}$, must be standard normal, i.e., that

(8.4.13) $$z_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \sim N(0,1) \ , \quad j = 0,1,..,k$$

Finally, if the *observed z-score* value is denoted by $z_j^{obs}$,[16] then the *p-value* for this (two-sided) test is thus given by

(8.4.14) $$p_j = \Pr\left(|\, z_j \,| > |\, z_j^{obs} \,|\right)$$




We shall illustrate this test for the Eire example below. But before doing so, it is important to point out that this test treats $s^2_{\hat{\beta}_j}$ in (8.4.11) as a "known" quantity, and in particular *ignores all variation* in this estimator. But in OLS, for example, this estimator can be shown to be both chi-square distributed and independent of $\hat{\beta}_j$, so that the ratio, $\hat{\beta}_j / s^2_{\hat{\beta}_j}$, is *t-distributed*. Thus, the relevant tests of $\beta_j$ coefficients in OLS are *t-tests*. But in more general settings such as SEM, the distribution of $\hat{\beta}_j / s^2_{\hat{\beta}_j}$ is unknown. So the standard "fall back" position is to treat $s^2_{\hat{\beta}_j}$ as a constant (by appealing implicitly to its large-sample consistency property), and to employ the normal distribution in (8.4.11) for testing purposes.

The consequence of this convention is to *inflate* significance levels (i.e., reduce p-values) to some degree. In the OLS case, this can be seen by noting that *t*-distributions have fatter

[16] Here "observed" means the actual value calculated by maximum-likelihood estimation.

tails than the standard normal distribution, thus increasing the p-value in (8.4.14) for any given observed value, $z_j^{obs}$. Because of this, some analysts prefer to use a more conservative *t*-test based on the number of parameters in the model (as in the OLS case). In particular, if we now re-designate the ratio in (8.4.13) as a *pseudo t-statistic*,

$$(8.4.15) \qquad t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

and let $T_v$ denote the *t-distribution* with *v degrees of freedom*, then [following Davidson and MacKinnon (1993, Section 3.6)] one can construct a corresponding *pseudo t-test* of the null hypothesis, $\beta_j = 0$, by assuming that $t_j$ is *t*-distributed with degrees of freedom equal to *n* minus the number of model parameters. In this case, the relevant parameter vector $(\beta, \sigma^2, \rho)$ is of length, $k+3$, so that[17]

$$(8.4.16) \qquad \boxed{t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \sim T_{n-(k+3)}}$$

The appropriate p-value for this test is then given by the probability in (8.4.14) with respect to the *t*-distribution in (8.4.16), and will be dented by $p_j^{t-pseudo}$. While these values are not reported in the screen output of **sem.m** or **slm.m** (as for the Eire example in Figure 7.7), we shall report them here just to illustrate the types of significance inflation that can occur.

But before turning to the Eire example, we first construct a test of the other key parameter of the SE-model, namely the *spatial dependence parameter*, $\rho$.[18] Here it is important to note that unlike the simple (and appealing) *rho statistic*, $\hat{\rho}_W$, in Section 4.1.1 above, which unfortunately yields an inconsistent estimate of $\rho$, the present maximum-likelihood estimator, $\hat{\rho}$, of $\rho$ is *consistent* (assuming of course that the SE-model is correctly specified). So from a theoretical perspective, formal hypothesis tests based on this estimator are of great interest. Here the same testing procedure for $\beta_j$ coefficients can be applied with the appropriate changes. First, it again follows from (8.4.4) that the estimator, $\hat{\rho}$, is normally distributed, so that as a parallel to (8.4.11), we now have

$$(8.4.17) \qquad \hat{\rho} \sim N\left(\rho, s_{\hat{\rho}}^2\right)$$

[17] Given that $\hat{\beta}$ in (7.3.11) is functionally independent of $\sigma^2$, one could in principle use $v = n-(k+2)$ in (8.4.16). However, we here adopt the (conservative) approach of using *all* parameters to calculate *v*.

[18] Note that while the *error variance*, $\sigma^2$, is also a model parameter, and indeed is also asymptotically normally distributed by (8.4.4), one is rarely interested in testing specific hypotheses about $\sigma^2$. So following standard convention, we simply report the estimated value, $\hat{\sigma}^2$, in screen outputs like Figure 7.7.

where $s_{\hat{\rho}} = \sqrt{s^2_{\hat{\rho}}}$. Moreover, since the natural null hypothesis for this parameter is again, $\rho = 0$ (here denoting the absence of spatial autocorrelation), we have the corresponding *z-score*,

$$(8.4.18) \qquad z_{\hat{\rho}} = \frac{\hat{\rho}}{s_{\hat{\rho}}} \sim N(0,1)$$

under this hypothesis. So the presence of non-zero spatial autocorrelation (either positive or negative) is now gauged by the two-sided p-value,

$$(8.4.19) \qquad p_{\hat{\rho}} = \Pr\left(|z_{\hat{\rho}}| > |z^{obs}_{\hat{\rho}}|\right)$$

where $z^{obs}_{\hat{\rho}}$ is again the observed z-score value. While this is always the default test employed in spatial regression software, it should be noted that (as in Section 4 above) a one-sided test for *positive spatial autocorrelation* is generally of more relevance. But we choose to employ the (more conservative) two-sided test to maintain comparability with other software. Finally, as with tests of $\beta$ coefficients above, we shall also report the p-value, $p^{t-pseudo}_{\hat{\rho}}$, for the corresponding *pseudo t-test* that captures at least some of the statistical variation in the estimator, $\hat{\rho}$.

### 8.4.2 Application to the Irish Blood Group Data

To apply these results to the Eire case, we first note that the estimated covariance matrix, $S_{SEM}$, in (8.4.9) is one of the outputs of **sem.m**, denoted by **cov**. In the present case, this matrix for the Eire data take the form in Figure 8.6 below, where each row and column is labeled by its corresponding parameter estimator $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, \hat{\rho})$:

| | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\sigma}^2$ | $\hat{\rho}$ |
|---|---|---|---|---|
| $\hat{\beta}_0$ | 1.9464 | - 0.3061 | 0 | 0 |
| $\hat{\beta}_1$ | - 0.3061 | 0.7809 | 0 | 0 |
| $\hat{\sigma}^2$ | 0 | 0 | 0.3874 | - 0.0211 |
| $\hat{\rho}$ | 0 | 0 | - 0.0211 | 0.0112 |

**Figure 8.6. SEM Parameter Covariance Matrix**

This clearly illustrates the block diagonal structure of $S_{SEM}$. To relate this covariance matrix to the SEM results in Figure 7.7, we focus on the important "Pale" coefficient, $\beta_1$. By recalling that the standard error of $\hat{\beta}_1$ in (8.4.12) is given from Figure 8.6 by

(8.4.20) $$s_{\hat{\beta}_1} = \sqrt{s^2_{\hat{\beta}_1}} = \sqrt{0.7809} = 0.8837$$

we see that the *z-score* for a two-sided test of the hypothesis, $\beta_1 = 0$, is given by[19]

(8.4.21) $$z_1 = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{1.5532}{.8837} = \boxed{1.7577}$$

as in Figure 7.7. Finally, the desired *p-value* for this test is given by

(8.4.22) $$p_1 = \Pr(|z_1| \geq 1.7577) = 2 \cdot \Phi(-1.7577) = \boxed{0.0788}$$

as in Figure 7.7. To compare this results with the corresponding *pseudo t-test* for $\beta_1$, observe first that in this case there are 4 parameters, $(\beta_0, \beta_1, \sigma^2, \rho)$, so that the for the $n = 26$ counties in Eire, the appropriate two-sided p-value is calculated with respect to a *t*-distribution with $v = 26 - 4 = 22$ degrees of freedom, yielding

(8.4.23) $$p_1^{t-pseudo} = \boxed{0.0927}$$

This is still weakly significant ($\leq .10$), but is noticeably less significant than the result in (8.4.22) based on the normal distribution. However, it should be noted that the sample size, $n = 26$, in this Eire example is quite small. So in larger sample sizes, where the tails of $T_{n-(k+3)}$ are much closer to those of $N(0,1)$, this difference will be far less noticeable.

Finally, for completeness, we also calculate the corresponding tests for the spatial dependency parameter, $\rho$. As in (8.4.21), the relevant *z*-score in (8.4.18) is seen from Figures 7.7 and 8.6 to be

(8.4.24) $$z_{\hat{\rho}} = \frac{\hat{\rho}}{s_{\hat{\rho}}} = \frac{0.7885}{\sqrt{0.0112}} = \boxed{7.467}$$

with corresponding p-values for the *z-test* and *pseudo t-test* given by:

(8.4.25) $$p_{\hat{\rho}} = \Pr(|z_{\hat{\rho}}| \geq 7.467) = \boxed{8.2 \cdot 10^{-14}} \quad \text{and} \quad p_{\hat{\rho}}^{t-pseudo} = \boxed{1.82 \times 10^{-7}}$$

So while the pseudo *t*-test again yields a somewhat weaker result, both p-values are vanishingly small,[20] and confirm that spatial autocorrelation in this model is strongly present.

[19] Note that since all of the following calculation examples are done to a much higher degree of precision than the numbers shown, the results on the right hand sides will not agree "exactly" with the indicated operations on the left-hand sides.

[20] Note that the reported value in Figure 7.7 is *not zero*, but rather is simply smaller than the number of decimal places allowed in this (default) printing format.

## 8.5 Parametric Significance Tests for SLM

Using the same notation as above, recall that the log-likelihood function for the SL-model in (6.2.2) through (6.2.4) is given by

(8.5.1) $L(\theta \mid y) = L(\beta, \sigma^2, \rho \mid y)$

$$= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2}\log|V_\rho| - \tfrac{1}{2\sigma^2}(y - X_\rho\beta)'V_\rho^{-1}(y - X_\rho\beta)$$

where $X_\rho = B_\rho^{-1}X$ and where $\rho$ is now the spatial dependency parameter for the dependent variable, $y$, rather than for the residual errors, $\varepsilon$. If for notational simplicity we let $H(\rho,\beta) = tr[G_\rho(G_\rho + G_\rho')] + \sigma^{-2}\beta' X' G_\rho' G_\rho X\beta$, and again let $G_\rho = WB_\rho^{-1}$, then the same analysis of this log-likelihood function as in (8.4.2) and (8.4.3) above [see Appendix B in Doreian (1980)] yields the corresponding covariance matrix for SLM:

(8.5.2) $$\text{cov}(\hat\beta, \hat\sigma^2, \hat\rho) = \begin{pmatrix} \frac{1}{\sigma^2}X'X & 0 & \frac{1}{\sigma^2}X'G_\rho X\beta \\ 0' & \frac{n}{2\sigma^4} & \frac{1}{\sigma^2}tr(G_\rho) \\ \frac{1}{\sigma^2}\beta' X' G_\rho' X & \frac{1}{\sigma^2}tr(G_\rho) & H(\rho,\beta) \end{pmatrix}^{-1}$$

Thus the appropriate *asymptotic sampling distribution* for SLM is given by:

(8.5.3) $$\begin{pmatrix} \hat\beta \\ \hat\sigma^2 \\ \hat\rho \end{pmatrix} \sim N\left[ \begin{pmatrix} \beta \\ \sigma^2 \\ \rho \end{pmatrix}, \begin{pmatrix} \frac{1}{\sigma^2}X'X & 0 & \frac{1}{\sigma^2}X'G_\rho X\beta \\ 0' & \frac{n}{2\sigma^4} & \frac{1}{\sigma^2}tr(G_\rho) \\ \frac{1}{\sigma^2}\beta' X' G_\rho' X & \frac{1}{\sigma^2}tr(G_\rho) & H(\rho,\beta) \end{pmatrix}^{-1} \right]$$

The key difference from SEM is that the present covariance matrix, $\text{cov}(\hat\beta, \hat\sigma^2, \hat\rho)$, is *not block diagonal*. The essential reason for this can be seen by comparing the reduced forms for SEM and SLM in (6.1.8) and (6.2.6), respectively, which we now reproduce:

(8.5.4) $$Y = X\beta + u \;\;,\;\; u \sim N(0, \sigma^2 V_\rho)$$

(8.5.5) $$Y = X_\rho\beta + u \;\;,\;\; u \sim N(0, \sigma^2 V_\rho)$$

These are seen to differ only in that $X$ for SEM is replaced by $X_\rho = B_\rho^{-1}X$ for SLM. So the difference here is that $\rho$ in SLM is *directly influencing the mean* of $Y$ while in SEM it is not [i.e., $E(Y \mid X) = B_\rho^{-1}X\beta$ rather than $E(Y \mid X) = X\beta$]. It is this linkage between $\rho$ and $\beta$ in SLM that creates non-zero covariances between $\hat\rho$ and the $\hat\beta$ components.

### 8.5.1 Parametric Tests for SLM

If we now substitute consistent maximum-likelihood estimates $(\hat{\beta},\hat{\sigma}^2,\hat{\rho})$ for the true parameter values in (8.5.2), and again factor out $\hat{\sigma}^4$ (for numerical stability in calculating the inverse), then the estimated covariance matrix, $S_{SLM}$, is seen to have the form:

(8.5.6)
$$S_{SLM} = \hat{\sigma}^4 \begin{pmatrix} \hat{\sigma}^2 X'X & 0 & \hat{\sigma}^2 X' G_{\hat{\rho}} X\hat{\beta} \\ 0' & \frac{n}{2} & \hat{\sigma}^2 tr(G_{\hat{\rho}}) \\ \hat{\sigma}^2 \hat{\beta}' X' G'_{\hat{\rho}} X & \hat{\sigma}^2 tr(G_{\hat{\rho}}) & \hat{\sigma}^4 H(\hat{\rho},\hat{\beta}) \end{pmatrix}^{-1}$$

Given this estimated matrix, it follows by using the same notation as for SEM that *all relations* in expressions (8.4.11) through (8.4.19) continue to hold, where $(\{\hat{\beta}_j : j = 0,1,..,k\},\hat{\sigma}^2,\hat{\rho})$ is now the vector of maximum-likelihood estimates for SLM rather than SEM, and where the standard deviations, $(\{s_{\hat{\beta}_j} : j = 0,1,..,k\}, s_{\hat{\sigma}^2}, s_{\hat{\rho}})$, are now the square roots of the diagonal elements of $S_{SLM}$ rather than $S_{SEM}$. So aside from these differences, all *z*-tests and pseudo *t*-tests are *identical* in form.

### 8.5.2 Application to the Irish Blood Group Data

As with **sem.m**, the MATLAB program **slm.m** offers an optional output of the $S_{SLM}$ matrix, designated by **cov**, which in a manner to Figure 8.6 above, now has the form:

| | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\sigma}^2$ | $\hat{\rho}$ |
|---|---|---|---|---|
| $\hat{\beta}_0$ | 10.327 | 0.8259 | 0.4129 | - 0.3589 |
| $\hat{\beta}_1$ | 0.8259 | 0.3366 | 0.0381 | - 0.0331 |
| $\hat{\sigma}^2$ | 0.4129 | 0.0381 | 0.2172 | - 0.0145 |
| $\hat{\rho}$ | - 0.3589 | - 0.0331 | - 0.0145 | 0.0126 |

**Figure 8.7. SLM Parameter Covariance Matrix**

Notice in particular that all elements of this covariance matrix are *nonzero*. So while there appear to be no direct links between $\beta$ and $\sigma^2$ in the Fisher information matrix for SLM [recall (8.3.2)], there are indirect links as seen in its inverse. More generally, only *block-diagonal* patterns of zeros in the Fisher information matrix ensure independence in the multi-normal case.

With these observations, the relevant test statistics can again be obtained from (the right panel of) Figure 7.7 together with the diagonal elements of $S_{SLM}$ in Figure 8.7. For $\beta_1$ we see in this case that

$$(8.5.7) \qquad z_1 = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{2.0142}{\sqrt{.3366}} = \boxed{3.472}$$

which in turn yields the following p-value for the *z-test* in Figure 7.7, together with the corresponding *pseudo t-test*:

$$(8.5.8) \qquad p_1 = \Pr(|z_1| \geq 3.472) = 2 \cdot \Phi(-3.472) = \boxed{0.00052} \quad \text{and} \quad p_1^{t-pseudo} = \boxed{0.0022}$$

So again we see that the significance level for the z-test is inflated. But even for the more conservative pseudo *t*-test, the "Pale" effect is here vastly more significant than for SEM, as was seen graphically in Figure 7.8 above.

Turning finally to the spatial dependency parameter, $\rho$, we may again use Figures 7.7 and 8.7 to obtain the following *z-score*,

$$(8.5.9) \qquad z_{\hat{\rho}} = \frac{\hat{\rho}}{s_{\hat{\rho}}} = \frac{0.7264}{\sqrt{0.0126}} = \boxed{6.467}$$

and corresponding *p-values*,

$$(8.5.10) \qquad p_{\hat{\rho}} = \Pr(|z_{\hat{\rho}}| \geq 6.467) = \boxed{9.99 \cdot 10^{-11}} \quad \text{and} \quad p_{\hat{\rho}}^{t-pseudo} = \boxed{1.66 \times 10^{-6}}$$

So even though the "Rippled Pale" in Figure 7.8 fits the Blood Group data far better than the "Pale" itself, these results show that there remains a great deal of spatial autocorrelation that is not accounted for by this single explanatory variable.