14-26

The conclusion is the same as before. Note that the same computation done using $\mathbf{V}$ rather than $\mathbf{V}_B$ produces a value of ~~5.1182~~ 5.1162. As before, we observe substantial small sample variation produced by the different estimators.

The latter three test statistics have substantially different values. It is possible to reach different conclusions, depending on which one is used. For example, if the test had been carried out at the 1 percent level of significance instead of 5 percent and LM had been computed using $\mathbf{V}$, then the critical value from the chi-squared statistic would have been 6.635 and the hypothesis would not have been rejected by the LM test. Asymptotically, all three tests are equivalent. But, in a finite sample such as this one, differences are to be expected.[10] Unfortunately, there is no clear rule for how to proceed in such a case, which highlights the problem of relying on a particular significance level and drawing a firm reject or accept conclusion based on sample evidence.

FN 10

~~16~~ 14.6.5 COMPARING MODELS AND COMPUTING MODEL FIT

The test statistics described in Sections ~~16~~14.6.1–~~16~~14.6.3 are available for assessing the validity of restrictions on the parameters in a model. When the models are nested, any of the three mentioned testing procedures can be used. For nonnested models, the computation is a comparison of one model to another based on an estimation criterion to discern which is to be preferred. Two common measures that are based on the same logic as the adjusted R-squared for the linear model are

(KT)

$$\textbf{Akaike information criterion (AIC)} \quad = -2 \ln L + 2K,$$
$$\textbf{Bayes (Schwarz) information criterion (BIC)} = -2 \ln L + K \ln n,$$

where K is the number of parameters in the model. Choosing a model based on the lowest AIC is logically the same as using $\bar{R}^2$ in the linear model; nonstatistical, albeit widely accepted. ~~Another means of comparing nonnested models that is valid in some circumstances is the **Vuong statistic** (see Section 7.3.4).~~

$$\cancel{V = \frac{\sqrt{n}\,\bar{m}}{s_m}, \text{ where } m_i = \ln L_i(1) - \ln L_i(2),}$$

~~and $L_i(j)$ is the contribution of individual i to the log-likelihood under the assumption that model j is correct. We use this test in Example 16.10 to choose between a geometric and a Poisson regression model for a count dependent variable.~~

The AIC and BIC are information criteria, not fit measures as such. This does leave open the question of how to assess the "fit" of the model. Only the case of a linear least squares regression in a model with a constant term produces an R^2, which measures the proportion of variation explained by the regression. The ambiguity in R^2 as a fit measure arose immediately when ~~we moved~~ from the linear regression model to the generalized regression model in Chapter ~~8~~ 9. The problem is yet more acute in the context of the models we consider in this chapter. For example, the estimators of the models for count data in Example ~~16~~ 14.10 make no use of the "variation" in the dependent variable and there is no obvious measure of "explained variation."

A measure of "fit" that was originally proposed for discrete choice models in McFadden (1974), but surprisingly has gained wide currency throughout the empirical

✓

✓

[10] For further discussion of this problem, see Berndt and Savin (1977).

14-27

KT KT

literature is the **likelihood ratio index**, which has come to be known as the **Pseudo R^2**. It is computed as

$$\text{Pseudo}R^2 = 1 - (\ln L)/(\ln L_0)$$

where $\ln L$ is the log-likelihood for the model estimated and $\ln L_0$ is the log-likelihood for the same model with only a constant term. The statistic does resemble the R^2 in a linear regression. The choice of name is for this statistic is unfortunate, however, because even in the discrete choice context for which it was proposed, it has no connection to the fit of the model to the data. In discrete choice settings in which log-likelihoods must be negative, the pseudo R^2 must be between zero and one and rises as variables are added to the model. It can obviously be zero, but is usually bounded below one. In the linear model with normally distributed disturbances, the maximized log-likelihood is

$$\ln L = (-n/2)[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)].$$

With a small amount of manipulation, we find that the pseudo R^2 for the linear regression model is

$$\text{Pseudo}R^2 = \frac{-\ln(1-R^2)}{1+\ln 2\pi + \ln s_y^2},$$

while the "true" R^2 is $1-\mathbf{e}'\mathbf{e}/\mathbf{e}_0'\mathbf{e}_0$. Because s_y^2 can vary independently of R^2—multiplying $\mathbf{y}$ by any scalar, A, leaves R^2 unchanged but multiplies s_y^2 by A^2—although the upper limit is one, there is no lower limit on this measure. This same problem arises in any model that uses information on the scale of a dependent variable, such as the tobit model (Chapter 24). The computation makes even less sense as a fit measure in multinomial 17
18 models such as the ordered probit model (Chapter ~~23~~) or the multinomial logit model.
For discrete choice models, there are a variety of such measures discussed in Chapter ~~23~~. 17
For limited dependent variable and many loglinear models, some other measure that is related to a correlation between a prediction and the actual value would be more useable. Nonetheless, the measure seems to have gained currency in the contemporary literature. [The popular software package, *Stata*, reports the pseudo R^2 with every model fit by MLE, but at the same time, admonishes its users not to interpret it as anything CDRCS
for example, meaningful. See, e.g., http://www.stata.com/support/faqs/stat/pseudor2.html. Cameron and Trivedi (2005) document the pseudo R^2 at length, then give similar cautions about it and urge their readers to seek a more meaningful measure of the correlation between model predictions and the outcome variable of interest. Wooldridge (2002a) dismisses it summarily, and argues that coefficients are more interesting.]

Au: OK to spell out "e.g." in text?

and/

✓
✓

~~16.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION~~

~~The applied literature contains a large and increasing number of models in which one model is embedded in another, which produces what are broadly known as "two-step" estimation problems. Consider an (admittedly contrived) example in which we have the following.~~

~~**Model 1.** Expected number of children = $E[y_1 \mid \mathbf{x}_1, \theta_1]$.~~
~~**Model 2.** Decision to enroll in job training = y_2, a function of $(\mathbf{x}_2, \theta_2, E[y_1 \mid \mathbf{x}_1, \theta_1])$.~~

14-28

the standard table to carry out the test. Unfortunately, in testing H_0 versus H_1 and vice versa, all four possibilities (reject both, neither, or either one of the two hypotheses) could occur. This issue, however, is a finite sample problem. Davidson and MacKinnon show that as $n \rightarrow \infty$, if H_1 is true, then the probability that $\hat{\lambda}$ will differ significantly from 0 approaches 1.

Example 7.2 *J Test for a Consumption Function*

Gaver and Geisel (1974) propose two forms of a consumption function:

$$H_0 : C_t = \beta_1 + \beta_2 Y_t + \beta_3 Y_{t-1} + \varepsilon_{0t}$$

and

$$H_1 : C_t = \gamma_1 + \gamma_2 Y_t + \gamma_3 C_{t-1} + \varepsilon_{1t}.$$

The first model states that consumption responds to changes in income over two periods, whereas the second states that the effects of changes in income on consumption persist for many periods. Quarterly data on aggregate U.S. real consumption and real disposable income are given in Appendix Table F5.1. Here we apply the J test to these data and the two proposed specifications. First, the two models are estimated separately (using observations 1950.2 through 2000.4). The least squares regression of C on a constant, Y, lagged Y, and the fitted values from the second model produces an estimate of λ of 1.0145 with a t ratio of 62.861. Thus, H_0 should be rejected in favor of H_1. But reversing the roles of H_0 and H_1, we obtain an estimate of λ of −10.677 with a t ratio of −7.188. Thus, ~~H_1 is rejected as well.~~ ?

14.6.6 ~~7.3.4~~ VUONG'S TEST AND THE KULLBACK–LEIBLER INFORMATION CRITERION

N

KL

Vuong's (1989) approach to testing **nonnested models** is also based on the likelihood ratio statistic.[8] The logic of the test is similar to that which motivates the likelihood ratio test in general. Suppose that $f(y_i \mid \mathbf{Z}_i, \theta)$ and $g(y_i \mid \mathbf{Z}_i, \gamma)$ are two competing models for the density of the random variable y_i, with f being the null model, H_0, and g being the alternative, H_1. For instance, in Example 7.2, both densities are (by assumption now) normal, y_i is consumption, C_t, $\mathbf{Z}_i$ is $[1, Y_t, Y_{t-1}, C_{t-1}]$, θ is $(\beta_1, \beta_2, \beta_3, 0, \sigma^2)$, γ is $(\gamma_1, \gamma_2, 0, \gamma_3, \omega^2)$, and σ^2 and ω^2 are the respective conditional variances of the disturbances, ε_{0t} and ε_{1t}. The crucial element of Vuong's analysis is that it need not be the case that either competing model is "true"; they may both be incorrect. What we want to do is attempt to use the data to determine which competitor is closer to the truth, that is, closer to the correct (unknown) model. 5.7

We assume that observations in the sample (disturbances) are conditionally independent. Let $L_{i,0}$ denote the ith contribution to the likelihood function under the null hypothesis. Thus, the log likelihood function under the null hypothesis is $\Sigma_i \ln L_{i,0}$. Define $L_{i,1}$ likewise for the alternative model. Now, let m_i equal $\ln L_{i,1} - \ln L_{i,0}$. If we were using the familiar likelihood ratio test, then, the likelihood ratio statistic would be simply $LR = 2\Sigma_i m_i = 2n\overline{m}$ when $L_{i,0}$ and $L_{i,1}$ are computed at the respective maximum likelihood estimators. When the competing models are nested—H_0 is a restriction on H_1—we know that $\Sigma_i m_i \geq 0$. The restrictions of the null hypothesis will never increase

~~[7] For related discussion of this possibility, see McAleer, Fisher, and Volker (1982).~~

[8] Once again, it is necessary to rely on results that we will develop more fully in Chapter 16. But, this discussion of nonnested models is a convenient point at which to introduce Vuong's useful statistic, and we will not be returning to the topic of nonnested models save for a short application in Chapter 24.

the likelihood function. (In the linear regression model with normally distributed disturbances that we have examined so far, the log likelihood and these results are all based on the sum of squared residuals, and as we have seen, imposing restrictions never reduces the sum of squares.) The limiting distribution of the *LR* statistic under the assumption of the null hypothesis is chi squared with degrees of freedom equal to the reduction in the number of dimensions of the parameter space of the alternative hypothesis that results from imposing the restrictions.

Vuong's analysis is concerned with nonnested models for which $\Sigma_i m_i$ need not be positive. Formalizing the test requires us to look more closely at what is meant by the "right" model (and provides a convenient departure point for the discussion in the next two sections). In the context of nonnested models, Vuong allows for the possibility that neither model is "true" in the absolute sense. We maintain the classical assumption that there does exist a "true" model, $h(y_i \mid \mathbf{Z}_i, \alpha)$ where α is the "true" parameter vector, but possibly neither hypothesized model is that true model. The **Kullback–Leibler Information Criterion** (KLIC) measures the distance between the true model (distribution) and a hypothesized model in terms of the likelihood function. Loosely, the KLIC is the log likelihood function under the hypothesis of the true model minus the log likelihood function for the (misspecified) hypothesized model under the assumption of the true model. Formally, for the model of the null hypothesis,

KT

$$\text{KLIC} = E[\ln h(y_i \mid \mathbf{Z}_i, \alpha) \mid h \text{ is true}] - E[\ln f(y_i \mid \mathbf{Z}_i, \theta) \mid h \text{ is true}]. \qquad \cancel{(7\text{-}15)}$$

The first term on the right hand side is what we would estimate with $(1/n)\ln L$ if we maximized the log likelihood for the true model, $h(y_i \mid \mathbf{Z}_i, \alpha)$. The second term is what is estimated by $(1/n)\ln L$ assuming (incorrectly) that $f(y_i \mid \mathbf{Z}_i, \theta)$ is the correct model.[2]

new paragraph

~~In the context of the model in the previous example,~~ Suppose the "true" model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with normally distributed disturbances and $\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{w}$ is the proposed competing model. The KLIC would be the expected log likelihood function for the true model minus the expected log likelihood function for the second model, still assuming that the first one is the truth. By construction, the KLIC is positive. We will now say that one model is "better" than another if it is closer to the "truth" based on the KLIC. If we take the difference of the two KLICs for two models, the true log likelihood function falls out, and we are left with

$$\text{KLIC}_1 - \text{KLIC}_0 = E[\ln f(y_i \mid \mathbf{Z}_i, \theta) \mid h \text{ is true}] - E[\ln g(y_i \mid \mathbf{Z}_i, \gamma) \mid h \text{ is true}].$$

To compute this using a sample, we would simply compute the likelihood ratio statistic, $n\bar{m}$ (without multiplying by 2) again. Thus, this provides an interpretation of the LR statistic. But, in this context, the statistic can be negative—we don't know which competing model is closer to the truth.

More to body of text

[2] Notice that $f(y_i \mid \mathbf{Z}_i, \theta)$ is written in terms of a parameter vector, θ. Because α is the "true" parameter vector, it is perhaps ambiguous what is meant by the parameterization, θ. Vuong (p. 310) calls this the "pseudotrue" parameter vector. It is the vector of constants that the estimator converges to when one uses the estimator implied by $f(y_i \mid \mathbf{Z}_i, \theta)$. In Example ~~7.2~~ 5.2, if H_0 gives the correct model, this formulation assumes that the least squares estimator in H_1 would converge to some vector of pseudo-true parameters. But, these are not the parameters of the correct model—they would be the slopes in the population linear projection of C_t on $[1, Y_t, C_{t-1}]$. ~~(See Section 4.2.2.)~~

14-30

Vuong's general result for nonnested models (his Theorem 5.1) describes the behavior of the statistic

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \sqrt{n}(\overline{m}/s_m), \quad m_i = \ln L_{i,0} - \ln L_{i,1}.$$

~~(7-14)~~

He finds:

(1) Under the hypothesis that the models are "equivalent", $V \xrightarrow{D} N[0, 1]$

(2) Under the hypothesis that $f(y_i \mid \mathbf{Z}_i, \theta)$ is "better", $V \xrightarrow{A.S.} +\infty$

(3) Under the hypothesis that $g(y_i \mid \mathbf{Z}_i, \gamma)$ is "better", $V \xrightarrow{A.S.} -\infty$.

This test is directional. Large positive values favor the null model while large negative values favor the alternative. The intermediate values (e.g., between −1.96 and +1.96 for 95 percent significance) are an inconclusive region. An application appears in Example 14.10.

~~*Example 7.3 Vuong Test for a Consumption Function*~~

~~We conclude Example 7.2 by applying the **Vuong test** to the consumption data. For the linear model, $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with normally distributed disturbances,~~

bold

$$\ln L_i = -1/2\,[\ln\sigma^2 + \ln 2\pi + (y_i - \mathbf{x}_i'\beta)^2/\sigma^2] \qquad \textbf{(7-15)}$$

~~and the maximum likelihood estimators of β and σ^2 are $\mathbf{b}$ and $\mathbf{e}'\mathbf{e}/n$. For the time-series data in Example 7.2, define for H_0, $e_{t0} = C_t - b_1 - b_2Y_t - b_3Y_{t-1}$ and $\mathbf{e}_0'\mathbf{e}_0 = \Sigma_t e_{t0}^2$. Define e_{t1} and $\mathbf{e}_0'\mathbf{e}_1$ likewise for H_1. Then, based on (7-14), we will have~~

$$\hat{m}_t = -1/2[\ln(\mathbf{e}_0'\mathbf{e}_0/\mathbf{e}_1'\mathbf{e}_1) + (e_{t0}^2/(\mathbf{e}_0'\mathbf{e}_0/T) - e_{t1}^2/(\mathbf{e}_0'\mathbf{e}_1/T))], \quad t = 1950.2, \ldots, 2000.4$$

$\leftarrow e_0'e_1 \rightarrow e_1'e_1$

~~(where $T = 203$). The Vuong statistic is −13.604, which once again strongly favors the alternative, H_1.~~

~~7.4 MODEL SELECTION CRITERIA~~

~~The preceding discussion suggested some approaches to model selection based on nonnested hypothesis tests. Fit measures and testing procedures based on the sum of squared residuals, such as R^2 and the Cox test, are useful when interest centers on the within-sample fit or within-sample prediction of the dependent variable. When the model building is directed toward forecasting, within-sample measures are not necessarily optimal. As we have seen, R^2 cannot fall when variables are added to a model, so there is a built-in tendency to overfit the model. This criterion may point us away from the best forecasting model, because adding variables to a model may increase the variance of the forecast error (see Section 5.6) despite the improved fit to the data. With this thought in mind, the **adjusted R^2**,~~

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2) = 1 - \frac{n-1}{n-K}\left(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right), \qquad \textbf{(7-16)}$$

~~has been suggested as a fit measure that appropriately penalizes the loss of degrees of freedom that result from adding variables to the model. Note that $\overline{R}^2$ may fall when a variable is added to a model if the sum of squares does not fall fast enough. (The applicable result appears in Theorem 3.7; $\overline{R}^2$ does not rise when a variable is added to a model unless the t ratio associated with that variable exceeds one in absolute value.)~~

14.7 TWO-STEP MAXIMUM LIKELIHOOD ESTIMATION

The applied literature contains a large and increasing number of applications in which elements of one model are embedded in another, which produces what are known as "two-step" estimation problems. [Among the best known of these is Heckman's (1979) model of sample selection discussed in Example 1.1 and in Chapter 18.] There are two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The first appears in the second model, but not the reverse. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution, $f(y_1, y_2 \,|\, \mathbf{x}1, \mathbf{x}2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ of the two random variables and then maximizing the full log-likelihood function,

MT

$$\ln L(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\sum_{i=1}^{n}\ln f(y_{i1},y_{i2}\,|\,\mathbf{x}_{i1},\mathbf{x}_{i2},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2).$$

A two-step procedure for this kind of model could be used by estimating the parameters of model 1 first by maximizing

$$\ln L_1(\boldsymbol{\theta}_1)=\sum_{i=1}^{n}\ln f_1(y_{i1}\,|\,\mathbf{x}_{i1},\boldsymbol{\theta}_1),$$

and then maximizing the marginal likelihood function for y_2 while embedding the consistent estimator of $\boldsymbol{\theta}_1$, treating it as given. The second step involves maximizing

$$\ln L_2(\hat{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_2)=\sum_{i=1}^{n}\ln f_2(y_{i1}\,|\,\mathbf{x}_{i2},\hat{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_2).$$

There are at least two reasons one might proceed in this fashion. First, it may be straightforward to formulate the two marginal log-likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log-likelihoods may be fairly straightforward, but maximizing the joint log-likelihood may be numerically complicated or difficult.[11] The results given here can be found in an important reference on the subject, Murphy and Topel (2002, first published in 1985).

FN 11

[11] There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

~~There~~ are two parameter vectors, θ_1 and θ_2. The first appears in the second model, although not the reverse. In such a situation, there are two ways to proceed. **Full information maximum likelihood (FIML)** estimation would involve forming the joint distribution $f(y_1, y_2\,|\,\mathbf{x}_1, \mathbf{x}_2, \theta_1, \theta_2)$ of the two random variables and then maximizing the full log-likelihood function,

$$\ln L = \sum_{i=1}^{n} f(y_{i1}, y_{i2}\,|\,\mathbf{x}_{i1}, \mathbf{x}_{i2}, \theta_1, \theta_2).$$

A second, or two-step, **limited information maximum likelihood (LIML)** procedure for this kind of model could be done by estimating the parameters of model 1, because it does not involve θ_2, and then maximizing ~~a conditional~~ log-likelihood function using the estimates from step 1:

$$\ln \hat{L} = \sum_{i=1}^{n} f[y_{i2}\,|\,\mathbf{x}_{i2}, \theta_2, (\mathbf{x}_{i1}, \hat{\theta}_1)].$$

There are at least two reasons one might proceed in this fashion. First, it may be straightforward to formulate the two separate log-likelihoods, but very complicated to derive the joint distribution. This situation frequently arises when the two variables being modeled are from different kinds of populations, such as one discrete and one continuous (which is a very common case in this framework). The second reason is that maximizing the separate log-likelihoods may be fairly straightforward, but maximizing the joint log-likelihood may be numerically complicated or difficult.[11] We will consider a few examples. Although we will encounter FIML problems at various points later in the book, for now we will present some basic results for two-step estimation. Proofs of the results given here can be found in an important reference on the subject, Murphy and Topel (2002).

Suppose, then, that our model consists of the two marginal distributions, $f_1(y_1\,|\,\mathbf{x}_1, \theta_1)$ and $f_2(y_2\,|\,$~~$\mathbf{x}_1$~~$, \mathbf{x}_2, \theta_1, \theta_2)$. Estimation proceeds in two steps.

1. Estimate θ_1 by maximum likelihood in model 1. Let $\hat{\mathbf{V}}_1$ be n times any of the estimators of the asymptotic covariance matrix of this estimator that were discussed in Section 16.4.6. 14
2. Estimate θ_2 by maximum likelihood in model 2, with $\hat{\theta}_1$ inserted in place of θ_1 as if it were known. Let $\hat{\mathbf{V}}_2$ be n times any appropriate estimator of the asymptotic covariance matrix of $\hat{\theta}_2$.

The argument for consistency of $\hat{\theta}_2$ is essentially that if θ_1 *were* known, then all our results for MLEs would apply for estimation of θ_2, and because plim $\hat{\theta}_1 = \theta_1$, asymptotically, this line of reasoning is correct. But the same line of reasoning is not sufficient to justify using $(1/n)\hat{\mathbf{V}}_2$ as the estimator of the asymptotic covariance matrix of $\hat{\theta}_2$. Some correction is necessary to account for an estimate of θ_1 being used in estimation of θ_2. The essential result is the following.

(See point 3 of Theorem D.16.)

[11] There is a third possible motivation. If either model is misspecified, then the FIML estimates of both models will be inconsistent. But if only the second is misspecified, at least the first will be estimated consistently. Of course, this result is only "half a loaf," but it may be better than none.

THEOREM ~~16~~14.8 Asymptotic Distribution of the Two-Step MLE [Murphy and Topel (2002)]

If the standard regularity conditions are met for both log-likelihood functions, then the second-step maximum likelihood estimator of $\boldsymbol{\theta}_2$ *is consistent and asymptotically normally distributed with asymptotic covariance matrix*

$$\mathbf{V}_2^* = \frac{1}{n}[\mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}' - \mathbf{R}\mathbf{V}_1\mathbf{C}' - \mathbf{C}\mathbf{V}_1\mathbf{R}']\mathbf{V}_2],$$

where

$$\mathbf{V}_1 = \text{Asy. Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)] \text{ based on } \ln L_1,$$

$$\mathbf{V}_2 = \text{Asy. Var}[\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)] \text{ based on } \ln L_2 \,|\, \boldsymbol{\theta}_1,$$

$$\mathbf{C} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_1'}\right)\right], \quad \mathbf{R} = E\left[\frac{1}{n}\left(\frac{\partial \ln L_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial \ln L_1}{\partial \boldsymbol{\theta}_1'}\right)\right].$$

The correction of the asymptotic covariance matrix at the second step requires some additional computation. Matrices $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *are estimated by the respective uncorrected covariance matrices. Typically, the BHHH estimators,*

$$\hat{\mathbf{V}}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)\right]^{-1}$$

and

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2'}\right)\right]^{-1}$$

are used. The matrices $\mathbf{R}$ *and* $\mathbf{C}$ *are obtained by summing the individual observations on the cross products of the derivatives. These are estimated with*

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_1'}\right)$$

and

$$\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_{i2}}{\partial \hat{\boldsymbol{\theta}}_2}\right)\left(\frac{\partial \ln f_{i1}}{\partial \hat{\boldsymbol{\theta}}_1'}\right).$$

~~***Example ~~16~~14.5 Two-Step ML Estimation***~~

~~Continuing the example discussed at the beginning of this section, we suppose that y_{i2} is a binary indicator of the choice whether to enroll in the program ($y_{i2} = 1$) or not ($y_{i2} = 0$) and that the probabilities of the two outcomes are~~

$$\cancel{\text{Prob}[y_{i2} = 1 \,|\, \mathbf{x}_{i1}, \mathbf{x}_{i2}] = \frac{e^{\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma E[y_{i1} \,|\, \mathbf{x}_{i1}']}}{1 + e^{\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma E[y_{i1} \,|\, \mathbf{x}_{i1}']}}}$$

A derivation of this useful result is instructive. We will rely on (14-11) and the results of Section 14.4.5.b where the asymptotic normality of the maximum likelihood estimator is developed. The first step MLE of $\boldsymbol{\theta}_1$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_1\left(\hat{\boldsymbol{\theta}}_1\right)}{\hat{\boldsymbol{\theta}}_1} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{\partial \ln f_1(y_{i1}\,|\,\mathbf{x}_{i1},\hat{\boldsymbol{\theta}}_1)}{\partial\hat{\boldsymbol{\theta}}_1}$$

$$= \frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{g}_{i1}\left(\hat{\boldsymbol{\theta}}_1\right) = \bar{\mathbf{g}}_1\left(\hat{\boldsymbol{\theta}}_1\right) = \mathbf{0}.$$

Using the results in that section, we obtained the asymptotic distribution from (14-15),

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) \xrightarrow{d} \left[-\mathbf{H}_{11}^{(1)}(\boldsymbol{\theta}_1)\right]^{-1}\sqrt{n}\,\bar{\mathbf{g}}_1(\boldsymbol{\theta}_1),$$

where the expression means that the limiting distribution of the two random vectors is the same, and

$$\mathbf{H}_{11}^{(1)} = E\left[\frac{1}{n}\frac{\partial^2 \ln L_1(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1'}\right].$$

The second step MLE of $\boldsymbol{\theta}_2$ is defined by

$$\frac{1}{n}\frac{\partial \ln L_2\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)}{\partial\hat{\boldsymbol{\theta}}_2} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{\partial \ln f_2(y_{i2}\,|\,\mathbf{x}_{i1},\mathbf{x}_{i2},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)}{\partial\hat{\boldsymbol{\theta}}_2}$$

$$= \frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{g}_{i2}\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right) = \bar{\mathbf{g}}_2\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right) = \mathbf{0}.$$

Expand the derivative vector, $\bar{\mathbf{g}}_2\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)$, in a linear Taylor series as usual, and use the results in Section 16.4.5.b once again;

$$\bar{\mathbf{g}}_2\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right) = \bar{\mathbf{g}}_2\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right) + \left[\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right)$$

$$+\left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) + o(1/n) = \mathbf{0}.$$

where

$$\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1'}\right] \text{ and } \mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = E\left[\frac{1}{n}\frac{\partial^2 \ln L_2(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2'}\right].$$

To obtain the asymptotic distribution, we use the same device as before,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\right) \xrightarrow{d} \left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]^{-1}\sqrt{n}\,\bar{\mathbf{g}}_2(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$$

$$+\left[-\mathbf{H}_{22}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]^{-1}\left[\mathbf{H}_{21}^{(2)}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]\sqrt{n}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)$$

For convenience, denote $\mathbf{H}_{22}^{(2)} = \mathbf{H}_{22}^{(2)}\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$, $\mathbf{H}_{21}^{(2)} = \mathbf{H}_{21}^{(2)}\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$ and $\mathbf{H}_{11}^{(1)} = \mathbf{H}_{11}^{(1)}\left(\boldsymbol{\theta}_1\right)$. Now substitute the first step estimator of $\boldsymbol{\theta}_1$ in this expression to obtain

14-35

$$\sqrt{n}\left(\hat{\theta}_2-\theta_2\right)\xrightarrow{d}\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\sqrt{n}\,\bar{\mathbf{g}}_2\left(\theta_1,\theta_2\right)$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\sqrt{n}\,\bar{\mathbf{g}}_1\left(\theta_1\right)$$

Consistency and asymptotic normality of the two estimators follow from our earlier results. To obtain the asymptotic covariance matrix for $\hat{\theta}_2$ we will obtain the limiting variance of the random vector in the expression above. The joint normal distribution of the two first derivative vectors has zero means and

preceding

AU: Is "Var" OK italic in DEs on this page?

$$Var\begin{bmatrix}\sqrt{n}\,\bar{\mathbf{g}}_1(\theta_1)\\ \sqrt{n}\,\bar{\mathbf{g}}_2(\theta_2,\theta_1)\end{bmatrix}=\begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix}.$$

Then, the asymptotic covariance matrix we seek is

$$Var\left[\sqrt{n}\left(\hat{\theta}_2-\theta_2\right)\right]=\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\mathbf{\Sigma}_{22}\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\mathbf{\Sigma}_{11}\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]'\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\mathbf{\Sigma}_{21}\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]'\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\mathbf{\Sigma}_{12}\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$

As we found earlier, the variance of the first derivative vector of the log likelihood is the negative of the expected second derivative matrix [see (14-11)]. Therefore $\mathbf{\Sigma}_{22}=\left[-\mathbf{H}_{22}^{(2)}\right]$ and $\mathbf{\Sigma}_{11}=\left[-\mathbf{H}_{11}^{(1)}\right]$. Making the substitution we obtain

$$Var\left[\sqrt{n}\left(\hat{\theta}_2-\theta_2\right)\right]=\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]'\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\mathbf{\Sigma}_{21}\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]'\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$$
$$+\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}\left[\mathbf{H}_{21}^{(2)}\right]\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}\mathbf{\Sigma}_{12}\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}.$$

From (14-15), $\left[-\mathbf{H}_{11}^{(1)}\right]^{-1}$ and $\left[-\mathbf{H}_{22}^{(2)}\right]^{-1}$ are the $\mathbf{V}_1$ and $\mathbf{V}_2$ that appear in Theorem 14.8, which further reduces the expression to

$$Var\left[\sqrt{n}\left(\hat{\theta}_2-\theta_2\right)\right]=\mathbf{V}_2+\mathbf{V}_2\left[\mathbf{H}_{21}^{(2)}\right]\mathbf{V}_1\left[\mathbf{H}_{21}^{(2)}\right]'\mathbf{V}_2-\mathbf{V}_2\mathbf{\Sigma}_{21}\mathbf{V}_1\left[\mathbf{H}_{21}^{(2)}\right]'\mathbf{V}_2-\mathbf{V}_2\left[\mathbf{H}_{21}^{(2)}\right]\mathbf{V}_1\mathbf{\Sigma}_{12}\mathbf{V}_2$$

Two remaining terms are $\mathbf{H}_{21}^{(2)}$ which is the $E[\partial^2\ln L_2(\theta_1,\theta_2)/\partial\theta_2\partial\theta_1']$, which is being estimated by $-\mathbf{C}$ in the statement of the theorem (note (14-11) again for the change of sign) and $\mathbf{\Sigma}_{21}$ which is the covariance of the two first derivative vectors. This is being estimated by $\mathbf{R}$ in Theorem 14.8. Making these last two substitutions produces

$$Var\left[\sqrt{n}\left(\hat{\theta}_2-\theta_2\right)\right]=\mathbf{V}_2+\mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2-\mathbf{V}_2\mathbf{R}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2-\mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{R}'\mathbf{V}_2$$

which completes the derivation.

14-36

Example 16.5 Two-Step ML Estimation

A common application of the two step method is accounting for the variation in a constructed regressor in a second step model. In this instance, the constructed variable is often an estimate of an expected value of a variable that is likely to be endogenous in the second step model. In this example, we will construct a rudimentary model that illustrates the computations.

In Riphahn, Wambach and Million (RWM, 2003), the authors studied whether individuals' use of the German health care system was at least partly explained by whether or not they had purchased a particular type of supplementary health insurance. We have used their data set, German Socioeconomic Panel (GSOEP) at several points. (See, e.g., Example ~~7.ZZ~~ 7.6.) One of the variables of interest in the study is *DocVis*, the number of times the an individual visits the doctor during the survey year. RWM considered the possibility that the presence of supplementary (*Addon*) insurance had an influence on the number of visits. Our simple model is as follows: The model for the number of visits is a Poisson regression (see Section ~~19.XXX~~ 19.2). This is a loglinear model that we will specify as

$$E[DocVis\,|\,\mathbf{x}_2,P_{Addon}] \;=\; \mu(\mathbf{x}_2'\boldsymbol{\beta},\gamma,\mathbf{x}_1'\boldsymbol{\alpha}) \;=\; \exp[\mathbf{x}_2'\boldsymbol{\beta} + \gamma\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})].$$

The model contains not the dummy variable, 1 if the individual has *Addon* insurance and 0 otherwise, which is likely to be endogenous in this equation, but an estimate of $E[Addon\,|\,\mathbf{x}_1]$ from a **logistic probability model** (see Section ~~17.YYY~~ 17.3) for whether the individual has insurance,

KT

AU: Term "logistic probability model" is not in chap. list

$$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha}) \;=\; \frac{\exp(\mathbf{x}_1'\alpha)}{1+\exp(\mathbf{x}_1'\alpha)} \;=\; \text{Prob[Individual has purchased } Addon \text{ insurance} \,|\, \mathbf{x}_1].$$

For purposes of the exercise, we will specify

$$(y_1 = Addon)\quad \mathbf{x}_1 = (constant, Age, Education, Married, Kids)',$$
$$(y_2 = DocVis)\quad \mathbf{x}_2 = (constant, Age, Education, Income, Female)'.$$

As before, to sidestep issues related to the panel data nature of the data set, we will use the 4483 observations in the 1988 wave of the data set, and drop the two observations for which *Income* is zero.

The log likelihood for the logistic probability model is

$$\ln L_1(\boldsymbol{\alpha}) \;=\; \Sigma_i\{(1-y_{i1})\ln[1-\Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})] + y_{i1}\ln\Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})\}.$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i1}(\boldsymbol{\alpha}) = \partial \ln f_1(y_{i1}\,|\,\mathbf{x}_{i1},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_{i1} - \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will maximize this log likelihood with respect to $\boldsymbol{\alpha}$ and then compute $\mathbf{V}_1$ using the BHHH estimator, as in Theorem 14.8. We will also use $\mathbf{g}_{i1}(\boldsymbol{\alpha})$ in computing $\mathbf{R}$.

14-37

The log likelihood for the Poisson regression model is

$$\ln L_2 = \Sigma_i [-\mu(\mathbf{x}_{i2}\boldsymbol{\beta},\gamma,\mathbf{x}_{i1}\boldsymbol{\alpha}) + y_{i2} \ln\mu(\mathbf{x}_{i2}\boldsymbol{\beta},\gamma,\mathbf{x}_{i1}\boldsymbol{\alpha}) - \ln y_{i2}!].$$

The derivatives of this log likelihood are

$$\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) = \partial\ln f_2(y_{i2},\mathbf{x}_{i1},\mathbf{x}_{i2},\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial(\boldsymbol{\beta}',\gamma)' = [y_{i2} - \mu(\mathbf{x}_{i2}\boldsymbol{\beta},\gamma,\mathbf{x}_{i1}\boldsymbol{\alpha})][\mathbf{x}_{i2}', \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]'$$

$$\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) = \partial\ln f_2(y_{i2},\mathbf{x}_{i1},\mathbf{x}_{i2},\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = [y_i - \mu(\mathbf{x}_{i2}\boldsymbol{\beta},\gamma,\mathbf{x}_{i1}\boldsymbol{\alpha})]\gamma\,\Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})[1 \overset{\text{minus}}{-} \Lambda(\mathbf{x}_{i1}'\boldsymbol{\alpha})]\mathbf{x}_{i1}.$$

We will use $\mathbf{g}_{i2}^{(2)}$ for computing $\mathbf{V}_2$ and in computing $\mathbf{R}$ and $\mathbf{C}$ and $\mathbf{g}_{i1}^{(2)}$ in computing $\mathbf{C}$. In particular,

$$\mathbf{V}_1 = [(1/n)\,\Sigma_i\,\mathbf{g}_{i1}(\boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})']^{-1},$$
$$\mathbf{V}_2 = [(1/n)\,\Sigma_i\,\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})']^{-1},$$
$$\mathbf{C} = [(1/n)\,\Sigma_i\,\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\mathbf{g}_{i1}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})'],$$
$$\mathbf{R} = [(1/n)\,\Sigma_i\,\mathbf{g}_{i2}^{(2)}(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\mathbf{g}_{i1}(\boldsymbol{\alpha})'].$$

minus

TB 14.2

Table ~~[illegible]X~~ 14.2 presents the two setp maximum likelihood estimates of the model parameters and estimated standard errors. For the first step logistic model, the standard errors marked H1 vs. V1 comparеs the values computed using the negative inverse of the second derivatives matrix (H1) vs. the outer products of the first derivatives ($\mathbf{V}_1$). As expected with a sample this large, the difference is minor. The latter were used in computing the corrected covariance matrix at the second step. In the Poisson model, the comparison of $\mathbf{V}_2$ to $\mathbf{V}_2$* shows distinctly that accounting for the presence of $\hat{\alpha}$ in the constructed regressor has a substantial impact on the standard errors, even in this relatively large sample. Note that the effect of the correction is to double the standard errors on the coefficients for the variables that the equations have in common, but it is quite minor for *Income* and *Female*, which are unique to the second step model.

Table ~~[illegible]X~~ 14.2 Estimated Logistic and Poisson Models

	Logistic Model for Addon			Poisson Model for DocVis		
	Coefficient	Standard Error (H_1)	Standard Error (V_1)	Coefficient	Standard Error (V_2)	Standard Error (V_2*)
Constant	-6.19246	0.60228	0.58287	0.77808	0.04884	0.09319
Age	0.01486	0.00912	0.00924	0.01752	0.00044	0.00111
Education	0.16091	0.03003	0.03326	-0.03858	0.00462	0.00980
Married	0.22206	0.23584	0.23523			
Kids	-0.10822	0.21591	0.21993			
Income				-0.80298	0.02339	0.02719
Female				0.16409	0.00601	0.00770
$\Lambda(\mathbf{x}_1'\boldsymbol{\alpha})$				3.91140	0.77283	1.87014

The covariance of the two gradients, **R**, may converge to zero in a particular application. When the first and second step estimates are based on different samples, **R** is exactly zero. For example, in our earlier application, **R** is based on two residuals,

$$g_{i1} = \{Addon_i - E[Addon_i|\mathbf{x}_{i1}]\} \text{ and } g_{i2}^{(2)} = \{DocVis_i - E[DocVis_i|\mathbf{x}_{i2},\Lambda_{i1}]\}.$$

The two residuals may well be uncorrelated. This assumption would be checked on a model-by-model basis, but in such an instance, the third and fourth terms in $\mathbf{V}_2$ vanish asymptotically and what remains is the simpler alternative,

$$\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2].$$

(In our application, the sample correlation between g_{i1} and $g_{i2}^{(2)}$ is only 0.015658 and the elements of the estimate of **R** are only about 0.01 times the corresponding elements of **C** – essentially about 99% of the correction in $\mathbf{V}_2^{**}$ is accounted for by **C**.) percent/

✓ It has been suggested that this set of procedures might be more complicated than necessary. [E.g., Cameron and Trivedi (2005, p. 202).] There are two alternative approaches one might take. First, under general circumstances, the asymptotic covariance matrix of the second step estimator could be approximated using the bootstrapping procedure discussed in Section 15.6. We would note, however, if this approach is taken, then it is essential that both steps be "bootstrapped." Otherwise, taking $\hat{\theta}_1$ as given and fixed, we will end up estimating $(1/n)\mathbf{V}_2$, not the appropriate covariance matrix. The point of the exercise is to account for the variation in $\hat{\theta}_1$. The second possibility is to fit the full model at once. That is, use a one step, full information maximum likelihood estimator and estimate θ_1 and θ_2 simultaneously. Of course, this is usually the procedure we sought to avoid in the first place. And with modern software, ✓ this two step method is often quite straightforward. Nonetheless, this is occasionally a possibility. Once again, Heckman's (1979) famous sample selection model provides an illuminating case. The two step and full information estimators for Heckman's model are developed in Section 18.5.3.

~~Chapter~~ ~~28~~ 17~~.) After a bit of manipulation, we find the convenient result that~~

$$\frac{\partial \ln L_2}{\partial \theta_2} = \sum_{i=1}^{n} (y_{i2} - P_i)\hat{\mathbf{x}}^*_{i2} = \sum_{i=1}^{n} v_i \hat{\mathbf{x}}^*_{i2}.$$

~~Once again, any of the three estimators could be used for estimating the asymptotic covariance matrix, but the BHHH estimator is convenient, so we use~~

$$\hat{\mathbf{V}}_2 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{v}_i^2 \hat{\mathbf{x}}^*_{i2}\hat{\mathbf{x}}^{*\prime}_{i2}\right]^{-1}.$$

~~For the final step, we must correct the asymptotic covariance matrix using $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$. What remains to derive—the few lines are left for the reader—is~~

$$\frac{\partial \ln L_2}{\partial \delta} = \sum_{i=1}^{n} v_i[\gamma \exp(\mathbf{x}'_{i1}\delta)]\mathbf{x}_{i1}.$$

~~So, using our estimates,~~

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_i^2[\gamma \exp(\mathbf{x}'_{i1}\hat{\delta})]\hat{\mathbf{x}}^*_{i2}\mathbf{x}'_{i1}, \quad \text{and} \quad \hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{v}_i \hat{\mathbf{x}}^*_{i2}\mathbf{x}'_{i1}.$$

~~We can now compute the correction.~~

~~In many applications, the covariance of the two gradients, **R**, converges to zero. When the first and second step estimates are based on different samples, **R** is exactly zero. For example, in our earlier application, $\mathbf{R} = \sum_{i=1}^{n} u_i v_i \mathbf{x}^*_{i2}\mathbf{x}'_{i1}$. The two "residuals," u and v, may well be uncorrelated. This assumption must be checked on a model-by-model basis, but in such an instance, the third and fourth terms in $\mathbf{V}_2^*$ vanish asymptotically and what remains is the simpler alternative,~~

$$\mathbf{V}_2^{**} = (1/n)[\mathbf{V}_2 + \mathbf{V}_2\mathbf{C}\mathbf{V}_1\mathbf{C}'\mathbf{V}_2].$$

~~We will examine some additional applications of this technique (including an empirical implementation of the preceding example) later in the book. Perhaps the most common application of two-step maximum likelihood estimation in the current literature, especially in regression analysis, involves inserting a prediction of one variable into a function that describes the behavior of another.~~

14 ~~16~~.8 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION AND ROBUST ASYMPTOTIC COVARIANCE MATRICES

Maximum likelihood estimation requires complete specification of the distribution of the observed random variable. If the correct distribution is something other than what we assume, then the likelihood function is misspecified and the desirable properties of the MLE might not hold. This section considers a set of results on an estimation approach that is robust to some kinds of model misspecification. For example, we have found that in a model, if the conditional mean function is $E[y\,|\,\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then certain estimators, such as least squares, are "robust" to specifying the wrong distribution of

14-40

the disturbances. That is, LS is MLE if the disturbances are normally distributed, but we can still claim some desirable properties for LS, including consistency, even if the disturbances are not normally distributed. This section will discuss some results that relate to what happens if we maximize the "wrong" log-likelihood function, and for those cases in which the estimator is consistent despite this, how to compute an appropriate asymptotic covariance matrix for it.[12]

FN 12

14 ~~16~~.8.1 MAXIMUM LIKELIHOOD AND GMM ESTIMATION

Let $f(y_i\,|\,\mathbf{x}_i, \boldsymbol\beta)$ be the true probability density for a random variable y_i given a set of covariates $\mathbf{x}_i$ and parameter vector $\boldsymbol\beta$. The log-likelihood function is $(1/n)\ln L(\boldsymbol\beta\,|\,\mathbf{y}, \mathbf{X}) = (1/n)\sum_{i=1}^{n}\ln f(y_i|\mathbf{x}_i, \boldsymbol\beta)$. The MLE, $\hat{\boldsymbol\beta}_{\mathrm{ML}}$, is the sample statistic that maximizes this function. (The division of $\ln L$ by n does not affect the solution.) We maximize the log-likelihood function by equating its derivatives to zero, so the MLE is obtained by solving the set of empirical moment equations

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(y_i\,|\,\mathbf{x}_i, \hat{\boldsymbol\beta}_{\mathrm{ML}})}{\partial \hat{\boldsymbol\beta}_{\mathrm{ML}}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(\hat{\boldsymbol\beta}_{\mathrm{ML}}) = \bar{\mathbf{d}}(\hat{\boldsymbol\beta}_{\mathrm{ML}}) = \mathbf{0}.$$

The population counterpart to the sample moment equation is

$$E\left[\frac{1}{n}\frac{\partial \ln L}{\partial \boldsymbol\beta}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(\boldsymbol\beta)\right] = E[\bar{\mathbf{d}}(\boldsymbol\beta)] = \mathbf{0}.$$

Using what we know about GMM estimators, if $E[\bar{\mathbf{d}}(\boldsymbol\beta)] = \mathbf{0}$, then $\hat{\boldsymbol\beta}_{\mathrm{ML}}$ is consistent and asymptotically normally distributed, with asymptotic covariance matrix equal to

$$\mathbf{V}_{\mathrm{ML}} = [\mathbf{G}(\boldsymbol\beta)'\mathbf{G}(\boldsymbol\beta)]^{-1}\mathbf{G}(\boldsymbol\beta)'\{\mathrm{Var}[\bar{\mathbf{d}}(\boldsymbol\beta)]\}\mathbf{G}(\boldsymbol\beta)[\mathbf{G}(\boldsymbol\beta)'\mathbf{G}(\boldsymbol\beta)]^{-1},$$

where $\mathbf{G}(\boldsymbol\beta) = \operatorname{plim}\partial\bar{\mathbf{d}}(\boldsymbol\beta)/\partial\boldsymbol\beta'$. Because $\bar{\mathbf{d}}(\boldsymbol\beta)$ is the derivative vector, $\mathbf{G}(\boldsymbol\beta)$ is $1/n$ times the expected Hessian of $\ln L$; that is, $(1/n)E[\mathbf{H}(\boldsymbol\beta)] = \bar{\mathbf{H}}(\boldsymbol\beta)$. As we saw earlier, $\mathrm{Var}[\partial \ln L/\partial\boldsymbol\beta] = -E[\mathbf{H}(\boldsymbol\beta)]$. Collecting all seven appearances of $(1/n)E[\mathbf{H}(\boldsymbol\beta)]$, we obtain the familiar result $\mathbf{V}_{\mathrm{ML}} = \{-E[\mathbf{H}(\boldsymbol\beta)]\}^{-1}$. [All the ns cancel and $\mathrm{Var}[\bar{\mathbf{d}}] = (1/n)\bar{\mathbf{H}}(\boldsymbol\beta)$.] Note that this result depends crucially on the result $\mathrm{Var}[\partial \ln L/\partial\boldsymbol\beta] = -E[\mathbf{H}(\boldsymbol\beta)]$.

n's

14 ~~16~~.8.2 MAXIMUM LIKELIHOOD AND *M* ESTIMATION

The maximum likelihood estimator is obtained by maximizing the function $\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol\beta) = (1/n)\sum_{i=1}^{n}\ln f(y_i, \mathbf{x}_i, \boldsymbol\beta)$. This function converges to its expectation as $n \to \infty$. Because this function is the log-likelihood for the sample, it is also the case (not proven here) that as $n \to \infty$, it attains its unique maximum at the true parameter vector, $\boldsymbol\beta$. (We used this result in proving the consistency of the maximum likelihood estimator.) Since $\operatorname{plim}\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol\beta) = E[\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol\beta)]$, it follows (by interchanging differentiation and the expectation operation) that $\operatorname{plim}\partial\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol\beta)/\partial\boldsymbol\beta = E[\partial\bar{h}_n(\mathbf{y}, \mathbf{X}, \boldsymbol\beta)/\partial\boldsymbol\beta]$. But, if this

[12] The following will sketch a set of results related to this estimation problem. The important references on this subject are White (1982a); Gourieroux, Monfort, and Trognon (1984); Huber (1967); and Amemiya (1985). A recent work with a large amount of discussion on the subject is Mittelhammer et al. (2000). The derivations in these works are complex, and we will only attempt to provide an intuitive introduction to the topic.

14-41

function achieves its *maximum* at β, then it must be the case that plim $\partial \overline{h}_n(\mathbf{y}, \mathbf{X}, \beta)/\partial\beta = \mathbf{0}$. KT

✓✓ An estimator that is obtained by maximizing a criterion function is called an ***M* estimator** KT [Huber (1967)] or an extremum estimator [Amemiya (1985)]. Suppose that we obtain an estimator by maximizing some other function, $M_n(\mathbf{y}, \mathbf{X}, \beta)$ that, although not the log-likelihood function, also attains its unique maximum at the true β as $n \to \infty$. Then the preceding argument might produce a consistent estimator with a known asymptotic distribution. For example, the log-likelihood for a linear regression model with normally distributed disturbances with *different* variances, $\sigma^2\omega_i$, is

$$\overline{h}_n(\mathbf{y}, \mathbf{X}, \beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-1}{2}\left[\ln(2\pi\sigma^2\omega_i) + \frac{(y_i - \mathbf{x}_i'\beta)^2}{\sigma^2\omega_i}\right]\right\}.$$

By maximizing this function, we obtain the maximum likelihood estimator. But we also examined another estimator, simple least squares, which maximizes $M_n(\mathbf{y}, \mathbf{X}, \beta) = -(1/n)\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2$. As we showed earlier, least squares is consistent and asymptotically normally distributed even with this extension, so it qualifies as an M estimator of the sort we are considering here.

Now consider the general case. Suppose that we estimate β by maximizing a criterion function

$$M_n(\mathbf{y}\,|\,\mathbf{X}, \beta) = \frac{1}{n}\sum_{i=1}^{n}\ln g(y_i\,|\,\mathbf{x}_i, \beta).$$

Suppose as well that plim$M_n(\mathbf{y}, \mathbf{X}, \beta) = E[M_n(\mathbf{y}\,|\,\mathbf{X}, \beta)]$ and that as $n \to \infty$, $E[M_n(\mathbf{y}\,|\,\mathbf{X}, \beta)]$ attains its unique maximum at β. Then, by the argument we used earlier for the MLE, plim $\partial M_n(\mathbf{y}\,|\,\mathbf{X}, \beta)/\partial\beta = E[\partial M_n(\mathbf{y}\,|\,\mathbf{X}, \beta)/\partial\beta] = \mathbf{0}$. Once again, we have a set of moment equations for estimation. Let $\hat{\beta}_E$ be the estimator that maximizes $M_n(\mathbf{y}\,|\,\mathbf{X}, \beta)$. Then the estimator is defined by

$$\frac{\partial M_n(\mathbf{y}\,|\,\mathbf{X}, \hat{\beta}_E)}{\partial\hat{\beta}_E} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln g(y_i\,|\,\mathbf{x}_i, \hat{\beta}_E)}{\partial\hat{\beta}_E} = \overline{\mathbf{m}}(\hat{\beta}_E) = \mathbf{0}.$$

Thus, $\hat{\beta}_E$ is a GMM estimator. Using the notation of our earlier discussion, $\mathbf{G}(\hat{\beta}_E)$ is the symmetric Hessian of $E[M_n(\mathbf{y}, \mathbf{X}, \beta)]$, which we will denote $(1/n)E[\mathbf{H}_M(\hat{\beta}_E)] = \overline{\mathbf{H}}_M(\hat{\beta}_E)$. Proceeding as we did above to obtain $\mathbf{V}_{\mathrm{ML}}$, we find that the appropriate asymptotic covariance matrix for the extremum estimator would be

$$\mathbf{V}_E = [\overline{\mathbf{H}}_M(\beta)]^{-1}\left(\frac{1}{n}\boldsymbol{\Phi}\right)[\overline{\mathbf{H}}_M(\beta)]^{-1},$$

where $\boldsymbol{\Phi} = \mathrm{Var}[\partial \log g(y_i\,|\,\mathbf{x}_i, \beta)/\partial\beta]$, and, as before, the asymptotic distribution is normal.

The Hessian in $\mathbf{V}_E$ can easily be estimated by using its empirical counterpart,

$$\mathrm{Est.}[\overline{\mathbf{H}}_M(\hat{\beta}_E)] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln g(y_i\,|\,\mathbf{x}_i, \hat{\beta}_E)}{\partial\hat{\beta}_E\partial\hat{\beta}_E'}.$$

But, $\boldsymbol{\Phi}$ remains to be specified, and it is unlikely that we would know what function to use. The important difference is that in this case, the variance of the first derivatives vector

14-42

need not equal the Hessian, so $\mathbf{V}_E$ does not simplify. We can, however, consistently estimate $\boldsymbol{\Phi}$ by using the sample variance of the first derivatives,

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \ln g(y_i \,|\, \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right]\left[\frac{\partial \ln g(y_i \,|\, \mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'}\right].$$

If this were the maximum likelihood estimator, then $\hat{\boldsymbol{\Phi}}$ would be the OPG estimator that we have used at several points. For example, for the least squares estimator in the heteroscedastic linear regression model, the criterion is $M_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = -(1/n)\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$, the solution is $\mathbf{b}$, $\mathbf{G}(\mathbf{b}) = (-2/n)\mathbf{X}'\mathbf{X}$, and

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}[2\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})][2\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})]' = \frac{4}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'.$$

Collecting terms, the 4s cancel and we are left precisely with the White estimator of (~~8~~9-27)!

14 ~~16~~.8.3 SANDWICH ESTIMATORS

At this point, we consider the motivation for all this weighty theory. One disadvantage of maximum likelihood estimation is its requirement that the density of the observed random variable(s) be fully specified. The preceding discussion suggests that in some situations, we can make somewhat fewer assumptions about the distribution than a full specification would require. The extremum estimator is robust to some kinds of specification errors. One useful result to emerge from this derivation is an estimator for the asymptotic covariance matrix of the extremum estimator that is robust at least to some misspecification. In particular, if we obtain $\hat{\boldsymbol{\beta}}_E$ by maximizing a criterion function that satisfies the other assumptions, then the appropriate estimator of the asymptotic covariance matrix is

$$\text{Est.}\,\mathbf{V}_E = \frac{1}{n}[\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1}\hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\beta}}_E)[\bar{\mathbf{H}}(\hat{\boldsymbol{\beta}}_E)]^{-1}.$$

(K1)

If $\hat{\boldsymbol{\beta}}_E$ is the true MLE, then $\mathbf{V}_E$ simplifies to $\{-[\mathbf{H}(\hat{\boldsymbol{\beta}}_E)]\}^{-1}$. In the current literature, this estimator has been called the **sandwich estimator.** There is a trend in the current literature to compute this estimator routinely, regardless of the likelihood function. It is worth noting that if the log-likelihood is not specified correctly, then the parameter estimators are likely to be inconsistent, save for the cases such as those noted below, so robust estimation of the asymptotic covariance matrix may be misdirected effort. But if the likelihood function is correct, then the sandwich estimator is unnecessary. This method is not a general patch for misspecified models. Not every likelihood function qualifies as a consistent extremum estimator *for the parameters of interest in the model.*

later

One might wonder at this point how likely it is that the conditions needed for all this to work will be met. There are applications in the literature in which this machinery has been used that probably do not meet these conditions, such as the tobit model of Chapter ~~24~~. We have seen one important case. Least squares in the generalized regression model passes the test. Another important application is models of "individual heterogeneity" in cross-section data. Evidence suggests that simple models often overlook unobserved sources of variation across individuals in cross sections, such as

18

unmeasurable "family effects" in studies of earnings or employment. Suppose that the correct model for a variable is $h(y_i \mid \mathbf{x}_i, v_i, \boldsymbol{\beta}, \theta)$, where v_i is a random term that is not observed and θ is a parameter of the distribution of v. The correct log-likelihood function is $\Sigma_i \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \theta) = \Sigma_i \ln \int_v h(y_i \mid \mathbf{x}_i, v_i, \boldsymbol{\beta}, \theta) f(v_i) dv_i$. Suppose that we maximize KT some other **pseudo-log-likelihood function,** $\Sigma_i \ln g(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})$ and then use the sandwich estimator to estimate the asymptotic covariance matrix of $\boldsymbol{\beta}$. Does this produce a consistent estimator of the true parameter vector? Surprisingly, sometimes it does, even though it has ignored the nuisance parameter, θ. We saw one case, using OLS in the GR model with heteroscedastic disturbances. Inappropriately fitting a Poisson model when the negative binomial model is correct—see Chapter ~~26~~ 19—is another case. For some specifications, using the wrong likelihood function in the probit model with proportions data is a third. [These examples are suggested, with several others, by Gourieroux, Monfort, and Trognon (1984).] We do emphasize once again that the sandwich estimator, in and of itself, is not necessarily of any virtue if the likelihood function is misspecified and the other conditions for the M estimator are not met.

~~16~~14.8.4 CLUSTER ESTIMATORS

Micro-level, or individual, data are often grouped or "clustered." A model of production or economic success at the firm level might be based on a group of industries, with multiple firms in each industry. Analyses of student educational attainment might be based on samples of entire classes, or schools, or statewide averages of schools within school districts. And, of course, such "clustering" is the defining feature of a panel data set. We considered several of these types of applications in our analysis of panel data in Chapter ~~9~~ 11. The recent literature contains many studies of clustered data in which the analyst has estimated a pooled model but sought to accommodate the expected correlation across observations with a correction to the asymptotic covariance matrix. We used this approach in computing a robust covariance matrix for the pooled least squares estimator in a panel data model [see (~~9~~11-3) and Example ~~9~~11.1 in Section ~~9~~11.3.2].

For the normal linear regression model, the log-likelihood that we maximize with the pooled least squares estimator is

$$\ln L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2}\frac{(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta})^2}{\sigma^2} \right].$$

[See (~~16~~14-34).] The "cluster-robust" estimator in ~~(9-3)~~ 11-3 can be written

$$\mathbf{W} = \left(\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{X}_i\right)^{-1} \left[\sum_{i=1}^{n} (\mathbf{X}_i'\mathbf{e}_i)(\mathbf{e}_i'\mathbf{X}_i)\right] \left(\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{X}_i\right)^{-1}$$

$$= \left(-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i} \mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1} \left[\sum_{i=1}^{n} \left(\sum_{t=1}^{T_i} \frac{1}{\hat{\sigma}^2}\mathbf{x}_{it}e_{it}\right)\left(\sum_{t=1}^{T_i} \frac{1}{\hat{\sigma}^2}e_{it}\mathbf{x}_{it}'\right)\right] \left(-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{t=1}^{T_i} \mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}$$

$$= \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}'}\right)^{-1} \left[\sum_{i=1}^{n} \left(\sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial\hat{\boldsymbol{\beta}}}\right)\left(\sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial\hat{\boldsymbol{\beta}}'}\right)\right] \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}'}\right)^{-1},$$

14-44

11

where f_{it} is the normal density with mean $\mathbf{x}_{it}'\beta$ and variance σ^2. This is precisely the "cluster-corrected" robust covariance matrix that appears elsewhere in the literature [minus an ad hoc "finite population correction" as in (9-4)].

In the generalized linear regression model (as in others), the OLS estimator is consistent, and will have asymptotic covariance matrix equal to

$$\text{Asy. Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'(\sigma^2\mathbf{\Omega})\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1}.$$

(See Theorem 8.1.) The center matrix in the sandwich for the panel data case can be written

9

$$\mathbf{X}'(\sigma^2\mathbf{\Omega})\,\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i'\mathbf{\Sigma}\mathbf{X}_i,$$

which motivates the preceding robust estimator. Whereas when we first encountered it, we motivated the cluster estimator with an appeal to the same logic that leads to the White estimator for heteroscedasticity, we now have an additional result that appears to justify the estimator in terms of the likelihood function.

Consider the specification error that the estimator is intended to accommodate. Suppose that the observations in group i were multivariate normally distributed with disturbance mean vector $\mathbf{0}$ and unrestricted $T_i \times T_i$ covariance matrix, $\mathbf{\Sigma}_i$. Then, the appropriate log-likelihood function would be

$$\ln L = \sum_{i=1}^{n} \left(-T_i/2 \ln 2\pi - \tfrac{1}{2}\ln|\mathbf{\Sigma}_i| - \tfrac{1}{2}\boldsymbol{\varepsilon}_i'\mathbf{\Sigma}_i^{-1}\boldsymbol{\varepsilon}_i\right),$$

where $\boldsymbol{\varepsilon}_i$ is the $T_i \times 1$ vector of disturbances for individual i. Therefore, we have maximized the wrong likelihood function. Indeed, the β that maximizes this log likelihood function is the GLS estimator, not the OLS estimator. OLS, and the cluster corrected estimator given earlier, "work" in the sense that (1) the least squares estimator is consistent in spite of the misspecification and (2) the robust estimator does, indeed, estimate the appropriate asymptotic covariance matrix.

Now, consider the more general case. Suppose the data set consists of n multivariate observations, $[y_{i,1}, \ldots, y_{i,T_i}]$, $i = 1, \ldots, n$. Each cluster is a draw from joint density $f_i(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\theta})$. Once again, to preserve the generality of the result, we will allow the cluster sizes to differ. The appropriate log likelihood for the sample is

$$\ln L = \sum_{i=1}^{n} \ln f_i(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\theta}).$$

Instead of maximizing ln L, we maximize a pseudo-log-likelihood

$$\ln L_P = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \ln g\left(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\theta}\right),$$

where we make the possibly unreasonable assumption that the same parameter vector, $\boldsymbol{\theta}$ enters the pseudo-log-likelihood as enters the correct one. Assume that it does. Using our familiar first-order asymptotics, the **pseudo-maximum likelihood estimator** KT {MLE}

14-45

will satisfy

$$(\hat{\theta}_{P,ML} - \theta) \approx \left(\frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial^2 \ln f_{it}}{\partial \theta \partial \theta'} \right)^{-1} \left(\frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \frac{\partial \ln f_{it}}{\partial \theta} \right) + (\theta - \beta)$$

$$= \left(\frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} \sum_{t=1}^{T_i} H_{it} \right)^{-1} \left(\sum_{i=1}^{n} w_i \bar{\mathbf{g}}_i \right) + (\theta - \beta),$$

where $w_i = T_i / \sum_{i=1}^{n} T_i$ and $\bar{\mathbf{g}}_i = (1/T_i) \sum_{t=1}^{T_i} \partial \ln f_{it} / \partial \theta$. The trailing term in the expression is included to allow for the possibility that plim $\hat{\theta}_{P,ML} = \beta$, which may not equal θ. [Note, for example, Cameron and Trivedi (2005, p. 842) specifically assume consistency in the generic model they describe.] Taking the expected outer product of this expression to estimate the asymptotic mean squared deviation will produce two terms—the cross term vanishes. The first will be the cluster-corrected matrix that is ubiquitous in the current literature. The second will be the squared error that may persist as n increases because the pseudo-MLE need not estimate the parameters of the model of interest.

We draw two conclusions. We can justify the cluster estimator based on this approximation. In general, it will estimate the expected squared variation of the pseudo-MLE around its probability limit. Whether it measures the variation around the appropriate parameters of the model hangs on whether the second term equals zero. In words, perhaps not surprisingly, this apparatus only works if the estimator is consistent. Is that likely? Certainly not if the pooled model is ignoring unobservable fixed effects. Moreover, it will be inconsistent in most cases in which the misspecification is to ignore latent random effects as well. The pseudo-MLE is only consistent for random effects in a few special cases, such as the linear model and Poisson and negative binomial models discussed in Chapter 25. It is not consistent in the probit and logit models in which this
19 approach often used. In the end, the cases in which the estimator are consistent are rarely, if ever, enumerated. The upshot is stated succinctly by Freedman (2006, p. 302): "The sandwich algorithm, under stringent regularity conditions, yields variances for the MLE that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect. However, it is quite another thing to ignore bias. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter."

14 ~~16~~.9 APPLICATIONS OF MAXIMUM LIKELIHOOD ESTIMATION

11

We will now examine several applications of the maximum likelihood estimator (MLE). We begin by developing the ML counterparts to most of the estimators for the classical and generalized regression models in Chapters 4 through 12. (Generally, the development for dynamic models becomes more involved than we are able to pursue here. The one exception we will consider is the standard model of autocorrelation.) We emphasize, in each of these cases, that we have already developed an efficient, generalized method of moments estimator that has the same asymptotic properties as the MLE under the assumption of normality. In more general cases, we will sometimes find that the GMM estimator is actually preferred to the MLE because of its robustness to failures of the distributional assumptions or its freedom from the necessity to make those assumptions in the first place. However, for the extensions of the classical model based on generalized least sqaures that are treated here, that is not the case. It might be argued that in these cases, the MLE is superfluous. There are occasions when the MLE will be preferred for other reasons, such as its invariance to transformation in nonlinear models and, possibly, its small sample behavior (although that is usually not the case). And, we will examine some nonlinear models in which there is no linear, method of moments counterpart, so the MLE is the natural estimator. Finally, in each case, we will find some useful aspect of the estimator, itself, including the development of algorithms such as Newton's method and the EM method for latent class models.

14 ~~16~~.9.1 THE NORMAL LINEAR REGRESSION MODEL

The linear regression model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

The likelihood function for a sample of n independent, identically and normally distributed disturbances is

$$L = (2\pi\sigma^2)^{-n/2} e^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(2\sigma^2)}. \qquad (\overset{14}{\cancel{16}}\text{-}32)$$

KT

FN 13

The transformation from ε_i to y_i is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, so the **Jacobian** for each observation, $|\partial\varepsilon_i/\partial y_i|$, is one.[13] Making the transformation, we find that the likelihood function for the n observations on the observed random variables is

$$L = (2\pi\sigma^2)^{-n/2} e^{(-1/(2\sigma^2))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}. \qquad (\overset{14}{\cancel{16}}\text{-}33)$$

To maximize this function with respect to $\boldsymbol{\beta}$, it will be necessary to maximize the exponent or minimize the familiar sum of squares. Taking logs, we obtain the log-likelihood function for the classical regression model:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}. \qquad (\overset{14}{\cancel{16}}\text{-}34)$$

The necessary conditions for maximizing this log-likelihood are

$$\begin{bmatrix} \dfrac{\partial \ln L}{\partial\boldsymbol{\beta}} \\ \dfrac{\partial \ln L}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}. \qquad (\overset{14}{\cancel{16}}\text{-}35)$$

The values that satisfy these equations are

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b} \quad \text{and} \quad \hat{\sigma}^2_{\text{ML}} = \frac{\mathbf{e}'\mathbf{e}}{n}. \qquad (\overset{14}{\cancel{16}}\text{-}36)$$

[13] See (B-41) in Section B.5. The analysis to follow is conditioned on $\mathbf{X}$. To avoid cluttering the notation, we will leave this aspect of the model implicit in the results. As noted earlier, we assume that the data generating process for $\mathbf{X}$ does not involve $\boldsymbol{\beta}$ or σ^2 and that the data are well behaved as discussed in Chapter 4.

14-47

The slope estimator is the familiar one, whereas the variance estimator differs from the least squares value by the divisor of n instead of $n - K$.[14]

FN 14 delete

The Cramér–Rao bound for the variance of an unbiased estimator is the negative inverse of the expectation of

$$\begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial\beta\partial\beta'} & \dfrac{\partial^2 \ln L}{\partial\beta\partial\sigma^2} \\[2ex] \dfrac{\partial^2 \ln L}{\partial\sigma^2\partial\beta'} & \dfrac{\partial^2 \ln L}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\dfrac{\mathbf{X}'\varepsilon}{\sigma^4} \\[2ex] -\dfrac{\varepsilon'\mathbf{X}}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\varepsilon'\varepsilon}{\sigma^6} \end{bmatrix}. \qquad 14\ (\cancel{16}\text{-}37)$$

In taking expected values, the off-diagonal term vanishes, leaving

$$[\mathbf{I}(\beta, \sigma^2)]^{-1} = \begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4/n \end{bmatrix}. \qquad 14\ (\cancel{16}\text{-}38)$$

The least squares slope estimator is the maximum likelihood estimator for this model. Therefore, it inherits all the desirable *asymptotic* properties of maximum likelihood estimators.

We showed earlier that $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$ is an unbiased estimator of σ^2. Therefore, the maximum likelihood estimator is biased toward zero:

$$E[\hat{\sigma}^2_{\text{ML}}] = \frac{n-K}{n}\sigma^2 = \left(1 - \frac{K}{n}\right)\sigma^2 < \sigma^2. \qquad 14\ (\cancel{16}\text{-}39)$$

Despite its small-sample bias, the maximum likelihood estimator of σ^2 has the same desirable asymptotic properties. We see in (~~16~~ 14-39) that s^2 and $\hat{\sigma}^2$ differ only by a factor $-K/n$, which vanishes in large samples. It is instructive to formalize the asymptotic equivalence of the two. From (~~16~~ 14-38), we know that

$$\sqrt{n}(\hat{\sigma}^2_{\text{ML}} - \sigma^2) \xrightarrow{d} N[0, 2\sigma^4].$$

It follows that

$$z_n = \left(1 - \frac{K}{n}\right)\sqrt{n}(\hat{\sigma}^2_{\text{ML}} - \sigma^2) + \frac{K}{\sqrt{n}}\sigma^2 \xrightarrow{d} \left(1 - \frac{K}{n}\right)N[0, 2\sigma^4] + \frac{K}{\sqrt{n}}\sigma^2.$$

But $K/\sqrt{n}$ and K/n vanish as $n \to \infty$, so the limiting distribution of z_n is also $N[0, 2\sigma^4]$. Because $z_n = \sqrt{n}(s^2 - \sigma^2)$, we have shown that the asymptotic distribution of s^2 is the same as that of the maximum likelihood estimator.

The standard test statistic for assessing the validity of a set of linear restrictions in the linear model, $\mathbf{R}\beta - \mathbf{q} = \mathbf{0}$, is the F ratio,

$$F[J, n-K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})}{J}.$$

With normally distributed disturbances, the F test is valid in any sample size. There remains a problem with nonlinear restrictions of the form $\mathbf{c}(\beta) = \mathbf{0}$, since the counterpart to F, which we will examine here, has validity only asymptotically even with normally distributed disturbances. In this section, we will reconsider the Wald statistic and examine two related statistics, the likelihood ratio statistic and the Lagrange multiplier

[14] As a general rule, maximum likelihood estimators do not make corrections for degrees of freedom.

statistic. These statistics are both based on the likelihood function and, like the Wald statistic, are generally valid only asymptotically.

No simplicity is gained by restricting ourselves to linear restrictions at this point, so we will consider general hypotheses of the form

$$H_0 : \mathbf{c}(\boldsymbol{\beta}) = \mathbf{0},$$
$$H_1 : \mathbf{c}(\boldsymbol{\beta}) \neq \mathbf{0}.$$

KT

The **Wald statistic** for testing this hypothesis and its limiting distribution under H_0 would be

$$W = \mathbf{c}(\mathbf{b})'\{\mathbf{C}(\mathbf{b})[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}(\mathbf{b})'\}^{-1}\mathbf{c}(\mathbf{b}) \xrightarrow{d} \chi^2[J], \qquad (\overset{14}{\cancel{16}}\text{-}40)$$

where

$$\mathbf{C}(\mathbf{b}) = [\partial\mathbf{c}(\mathbf{b})/\partial\mathbf{b}']. \qquad (\overset{14}{\cancel{16}}\text{-}41)$$

KT

The **likelihood ratio (LR) test** is carried out by comparing the values of the log-likelihood function with and without the restrictions imposed. We leave aside for the present how the restricted estimator $\mathbf{b}_*$ is computed (except for the linear model, which we saw earlier). The test statistic and it's limiting distribution under H_0 are

its

$$\text{LR} = -2[\ln L_* - \ln L] \xrightarrow{d} \chi^2[J]. \qquad (\overset{14}{\cancel{16}}\text{-}42)$$

The log-likelihood for the regression model is given in $(\overset{14}{\cancel{16}}\text{-}34)$. The first-order conditions imply that regardless of how the slopes are computed, the estimator of σ^2 without restrictions on $\boldsymbol{\beta}$ will be $\hat{\sigma}^2 = (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})/n$ and likewise for a restricted estimator $\hat{\sigma}^2_* = (\mathbf{y}-\mathbf{X}\mathbf{b}_*)'(\mathbf{y}-\mathbf{X}\mathbf{b}_*)/n = \mathbf{e}'_*\mathbf{e}_*/n$. The **concentrated log-likelihood**[15] will be KT

FN 15

$$\ln L_c = -\frac{n}{2}[1 + \ln 2\pi + \ln(\mathbf{e}'\mathbf{e}/n)]$$

and likewise for the restricted case. If we insert these in the definition of LR, then we obtain

$$\text{LR} = n\ln[\mathbf{e}'_*\mathbf{e}_*/\mathbf{e}'\mathbf{e}] = n\,(\ln\hat{\sigma}^2_* - \ln\hat{\sigma}^2) = n\ln(\hat{\sigma}^2_*/\hat{\sigma}^2). \qquad (\overset{14}{\cancel{16}}\text{-}43)$$

KT

The **Lagrange multiplier (LM) test** is based on the gradient of the log-likelihood function. The principle of the test is that if the hypothesis is valid, then at the restricted estimator, the derivatives of the log-likelihood function should be close to zero. There are two ways to carry out the LM test. The log-likelihood function can be maximized subject to a set of restrictions by using

$$\ln L_{\text{LM}} = -\frac{n}{2}\left[\ln 2\pi + \ln\sigma^2 + \frac{[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]/n}{\sigma^2}\right] + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\beta}).$$

[15] See Section E4.3.

14-49

The first-order conditions for a solution are

$$\begin{bmatrix} \dfrac{\partial \ln L_{\mathrm{LM}}}{\partial \beta} \\ \dfrac{\partial \ln L_{\mathrm{LM}}}{\partial \sigma^2} \\ \dfrac{\partial \ln L_{\mathrm{LM}}}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{X}'(\mathbf{y}-\mathbf{X}\beta)}{\sigma^2} + \mathbf{C}(\beta)'\lambda \\ \dfrac{-n}{2\sigma^2} + \dfrac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma^4} \\ \mathbf{c}(\beta) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \qquad 14\ (\cancel{16}\text{-}44)$$

The solutions to these equations give the restricted least squares estimator, $\mathbf{b}_*$; the usual variance estimator, now $\mathbf{e}_*'\mathbf{e}_*/n$; and the Lagrange multipliers. There are now two ways to compute the test statistic. In the setting of the classical linear regression model, when we actually compute the Lagrange multipliers, a convenient way to proceed is to test the hypothesis that the multipliers equal zero. For this model, the solution for λ_* is $\lambda_* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})$. This equation is a linear function of the least squares estimator. If we carry out a *Wald* test of the hypothesis that λ_* equals $\mathbf{0}$, then the statistic will be

$$\mathrm{LM} = \lambda_*'\{\text{Est. Var}[\lambda_*]\}^{-1}\lambda_* = (\mathbf{Rb}-\mathbf{q})'[\mathbf{R}s_*^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}). \qquad 14\ (\cancel{16}\text{-}45)$$

The disturbance variance estimator, s_*^2, based on the restricted slopes is $\mathbf{e}_*'\mathbf{e}_*/n$.

An alternative way to compute the LM statistic often produces interesting results. In most situations, we maximize the log-likelihood function without actually computing the vector of Lagrange multipliers. (The restrictions are usually imposed some other way.) An alternative way to compute the statistic is based on the (general) result that under the hypothesis being tested,

$$E[\partial \ln L/\partial \beta] = E[(1/\sigma^2)\mathbf{X}'\varepsilon] = \mathbf{0}$$

FN 16

and[16]

$$\text{Asy. Var}[\partial \ln L/\partial \beta] = -E[\partial^2 \ln L/\partial\beta\partial\beta']^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad 14\ (\cancel{16}\text{-}46)$$

We can test the hypothesis that at the restricted estimator, the derivatives are equal to zero. The statistic would be

$$\mathrm{LM} = \frac{\mathbf{e}_*'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n} = nR_*^2. \qquad 14\ (\cancel{16}\text{-}47)$$

In this form, the LM statistic is n times the coefficient of determination in a regression of the residuals $e_{i*} = (y_i - \mathbf{x}_i'\mathbf{b}_*)$ on the full set of regressors.

FN 17

With some manipulation we can show that $W = [n/(n-K)]JF$ and LR and LM are approximately equal to this function of F.[17] All three statistics converge to JF as n increases. The linear model is a special case in that the LR statistic is based only on the unrestricted estimator and does not actually require computation of the restricted least squares estimator, although computation of F does involve most of the computation of $\mathbf{b}_*$. Because the log function is concave, and $W/n \geq \ln(1 + W/n)$, Godfrey (1988) also shows that $W \geq \mathrm{LR} \geq \mathrm{LM}$, so for the linear model, we have a firm ranking of the three statistics.

[16] This makes use of the fact that the Hessian is block diagonal.

[17] See Godfrey (1988, pp. 49–51). N