QUASI-MAXIMUM LIKELIHOOD

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. General Misspecification
3. Model Selection Tests
4. QMLE in the Linear Exponential Family
5. Generalized Estimating Equations for Panel Data

1. INTRODUCTION

- Traditionally, maximum likelihood estimation, usually conditional on a set of explanatory variables, is studied under the assumption that the underlying density function is correctly specified.
- In econometrics, White (1982, *Econometrica*) popularized the notion that the underlying density might be completely misspecified, resulting in the notion of **quasi-maximum likelihood** (sometimes **pseudo-maximum likelihood**). This raises several questions.

- (1) If the model is generally misspecified, how to we interpret our estimates? (Technically, how do we interpret the probability limits of the quasi-MLEs?)
- (2) If we admit that our model is likely to be misspecified, how do we perform statistical inference?

- (3) How might we detect whether a particular density is misspecified? (Generally, we can nest a particular model in a more general model and perform Wald and Lagrange Multiplier tests.)
- (4) If we maintain the notion that all models are, at best, approximations, are there ways to choose between competing models that are nonnested? This leads to **model selection tests**.

- In some cases, maximizing a log-likelihood function that is not correct in its entirety can nevertheless consistently estimate parameters in a feature that is correctly specified, usually the condition mean but sometimes the conditonal mean and conditional variance. The work of Gourieroux, Monfort, and Trognon (1984a, Econometrica) was very influential.

• The same ideas can be applied to cross section and panel data. For panel data, we already studied partial maximum likelihood where, say, a density for $D(\mathbf{y}_{it}|\mathbf{x}_{it}),\ t = 1, \ldots, T$ is assumed to be correct, but the joint distribution is left unspecified. But the model for $D(\mathbf{y}_{it}|\mathbf{x}_{it})$ might be wrong, or maybe $E(\mathbf{y}_{it}|\mathbf{x}_{it})$ is correctly specified but other features of $D(\mathbf{y}_{it}|\mathbf{x}_{it})$ are wrong.

2. GENERAL MISSPECIFICATION

- We assume standard regularity conditions that make the asymptotic analysis simple, namely, that the **quasi-log-likelihood function**, log $f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})$, is twice continuously differentiable in $\boldsymbol{\theta}$.
- The key difference now is that we do not assume a "true" value of theta, which we called $\boldsymbol{\theta}_o$. Instead, we postulate the existence of a unique solution to the population problem

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})], \tag{1}$$

which we denote this value by $\boldsymbol{\theta}^*$. Often called the **pseudo-true value**.

• White (1982, 1994) discusses the interpretation of $\boldsymbol{\theta}^*$ as providing the best approximation to the true density in the parametric class $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$, where closeness is measured in terms of the Kullback-Leibler information criterion; see also the appendix to Chapter 13.

• Still let $\hat{\boldsymbol{\theta}}$ denote the solution to

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}). \tag{2}$$

Now call this the **quasi-maximum likelihood estimator** (**QMLE**) (or sometimes “pseudo” replaces “quasi”).

- Consistency of $\hat{\theta}$ for θ^* follows in the same way as when the model is correctly specified, provided θ^* is unique, the objective function is continuous in θ (and we can relax to that continuity "with probability one"), and other regularity conditions hold.
- Asymptotic inference concerning θ^* is more interesting. For one, the information matrix equality would only hold by fluke. Second, except in cases where particular features of the distribution are correctly specified, calulations of expected Hessians (conditional on covariates) are generally incorrect.

• In the absense of further information, there is only one legitimate estimator of $Avar(\hat{\boldsymbol{\theta}})$:

$$\widehat{Avar(\hat{\boldsymbol{\theta}})} = \left(\sum_{i=1}^{N} \mathbf{H}_i(\hat{\boldsymbol{\theta}})\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'\right)\left(\sum_{i=1}^{N} \mathbf{H}_i(\hat{\boldsymbol{\theta}})\right)^{-1}, \qquad (3)$$

where, as before, $\mathbf{s}_i(\boldsymbol{\theta})$ is the $P \times 1$ score vector and $\mathbf{H}_i(\boldsymbol{\theta})$ is the $P \times P$ Hessian.

• As usual, this estimator is "legitimate" in the sense that, when divided by N, the right hand side converges in probability to $Avar[\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)] = \mathbf{A}^{*-1}\mathbf{B}^*\mathbf{A}^{*-1}$, where $\mathbf{A}^* \equiv -E[\mathbf{H}_i(\boldsymbol{\theta}^*)]$ and $\mathbf{B}^* \equiv E[\mathbf{s}_i(\boldsymbol{\theta}^*)\mathbf{s}_i(\boldsymbol{\theta}^*)']$.

- Asymptotic t statistics for testing hypotheses about the θ_j^* are easily obtained because the asymptotic standard errors of the $\hat{\theta}_j$ are the square roots of the diagonal elements of the estimated asymptotic variance matrix.
- Score tests also need to use the sandwich form, where $\hat{\boldsymbol{\theta}}$ is replaced by $\tilde{\boldsymbol{\theta}}$, the restricted estimate. Fortunately, even though $\tilde{\mathbf{A}}$ might not be positive definite, the sandwich estimator is always at least positive semidefinite because because $\tilde{\mathbf{B}}$ is always at least positive semidefinite.

• Inference based on the quasi-likelihood ratio statistic is not tractable: the LR statistic no longer has a limiting chi-square distribution and its limiting distribution depends on unknown parameters.

• As an example, consider the probit model but where $P(y_i = 1|\mathbf{x}_i) \neq \Phi(\mathbf{x}_i\boldsymbol{\theta})$ for all $K \times 1$ vectors $\boldsymbol{\theta}$, so the probit model is misspecified. Let $\hat{\boldsymbol{\theta}}$ be obtained by maximizing the probit log-likelihood.

• Under weak conditions $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^* \in \mathbb{R}^K$ where $\Phi(\mathbf{x}\boldsymbol{\theta}^*)$ provides the "best" approximation to $P(y_i = 1|\mathbf{x}_i = \mathbf{x})$ in the sense of minimizing the Kullback-Leibler distance.

- For a continuous explanatory variable, say x_j, we would estimate the partial effect of x_j on $P(y_i = 1|\mathbf{x}_i = \mathbf{x})$ as the partial derivative $\hat{\theta}_j \phi(\mathbf{x}\hat{\boldsymbol{\theta}})$, which consistently estimates $\theta_j^* \phi(\mathbf{x}\boldsymbol{\theta}^*)$.
- To get valid confidence intervals for θ_j^* and partial effects such as $\phi(\mathbf{x}\boldsymbol{\theta}^*)$, we need to use the fully robust sandwich estimator (along with the delta method). Or, use the nonparametric bootstrap.
- Viewing the probit model as an appoximation to the true response probability is really no different than thinking of the linear probability model as an approximation. Probit might be a better approximation.

- In Stata:

```
probit y x1 x2 ... xK, robust
```

- Need to understand that this command does not produce valid inference of the index parameters in a "heteroskedastic probit." That is, if we write

$$y_i = 1[\mathbf{x}_i\boldsymbol{\theta}_o + e_i > 0] \tag{4}$$

$$D(e_i|\mathbf{x}_i) = Normal[0, h(\mathbf{x}_i)], \tag{5}$$

but then use standard probit, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_o$. (In fact, in many cases $\boldsymbol{\theta}^*$ would not be very close to $\boldsymbol{\theta}_o$.)

- As far as we know, the only interpretation of $\boldsymbol{\theta}^*$ is as the "best" approximation to $P(y = 1|\mathbf{x})$ using the misspecified model $P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\theta})$.
- Whether probit provides a good approximation is generally difficult to say. But we can do inference on the pseudo-true parameters and partial effects.

• We can also allow for complete density misspecification in the context of partial (pooled) MLE. We must allow for a general estimate of the Hessian: for each i,

$$\mathbf{H}_i(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \mathbf{H}_{it}(\hat{\boldsymbol{\theta}}) \tag{6}$$

• Without assuming that $f_t(\mathbf{y}_t|\mathbf{x}_t;\boldsymbol{\theta})$ is correctly specified for each t it makes little sense to discuss dynamic completeness, and the scores are generally serially correlated (when evaluated now at $\boldsymbol{\theta}^*$).

- Without further analysis one should use, in the sandwich,

$$\mathbf{s}_i(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \mathbf{s}_{it}(\hat{\boldsymbol{\theta}}) \tag{7}$$

so that terms $\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{ir}(\hat{\boldsymbol{\theta}})'$ for $t \neq r$ are accounted for, as in pooled MLE with correctly specified marginal densities.

• Packages such as Stata properly compute the asymptotic variance estimate when the “cluster” option is used. For example, for panel data, the command

```
probit y x1 x2 ... xK, cluster(id)
```

can be used to make inference robust not just to incomplete dynamics, but also to misspecification of the marginal distribution.

- Exercise: If $P(y_{it} = 1|\mathbf{x}_{it}) = \Phi(\mathbf{x}_{it}\boldsymbol{\theta}_o)$, but this model is not dynamically complete, provide an estimator of $Avar(\hat{\boldsymbol{\theta}})$ (for the pooled MLE) that is valid and uses only first derivatives.
- Similar comments hold for other models estimated by pooled MLE.

3. MODEL SELECTION TESTS

- Properties of MLE under general misspecification can be used to derive a **model selection test** due to Vuong (1988).
- The test is intended to allow one to choose between competing models. Here we treat the case where the two models are, in a sense to be made precise, **nonnested**. (When one model is a special case of the other, the score approach provides a much simpler way to test an attractive null model against a more general alternative.)

- If we are content with just choosing the model with the “best fit” given the data at hand, then it is legitimate to choose the model with the largest value of the log-likelihood.
- Having the largest log likelihood – more precisely, the largest *expected* log-likelihood – is necessary but not sufficient for a model to be correctly specified. (Cannot compare all possible models.) A density model cannot be correctly specified if it delivers (asymptotically) a lower average log-likelihood than another model.
- Comparing log-likelihood values is analogous to comparing *R*-squareds in a regression context.

• As suggested by Vuong (1988), it is useful to attach statistical significance to the difference in log likelihoods. For nonnested models, this turns out to be very easy.

• Let $f_1(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_1)$ and $f_2(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_2)$ be competing models for the density of $D(\mathbf{y}_i|\mathbf{x}_i)$, where both may be misspecified. Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be the quasi-maximum likelihood estimators converging to $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$, respectively. Let $\mathcal{L}_m = \sum_{i=1}^{N} \ell_{im}(\hat{\boldsymbol{\theta}}_m)$ be the quasi-log likelihood evaluated at the relevant estimate for $m = 1, 2$. Then

$$(\mathcal{L}_1 - \mathcal{L}_2)/N \xrightarrow{p} E[\log f_1(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_1^*)] - E[\log f_2(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_2^*)], \tag{8}$$

where the expected values are over the joing distribution $(\mathbf{x}_i, \mathbf{y}_i)$.

• We can actually say more. Using a mean value expansion and the $\sqrt{N}$-consistency of $\hat{\boldsymbol{\theta}}_m^*$ for $\boldsymbol{\theta}_m^*$, it can be shown that

$$N^{-1/2}(\mathcal{L}_1 - \mathcal{L}_2) = N^{-1/2} \sum_{i=1}^{N} [\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)]$$

$$= N^{-1/2} \sum_{i=1}^{N} [\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)] + o_p(1). \qquad (9)$$

(See Problem 13.13.)

• Key to obtaining a simple model specification test because it shows that the estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ do not affect that asymptotic distribution of $N^{-1/2}(\mathcal{L}_1 - \mathcal{L}_2)$.

• Therefore, we can obtain an asymptotic normal distribution for $N^{-1/2}(\mathcal{L}_1 - \mathcal{L}_2)$ under the null hypothesis

$$H_0 : E[\ell_{i1}(\boldsymbol{\theta}_1^*)] = E[\ell_{i2}(\boldsymbol{\theta}_2^*)]. \tag{10}$$

• Under this null,

$$N^{-1/2} \sum_{i=1}^{N} [\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)] \xrightarrow{d} Normal(0, \eta^2) \tag{11}$$

where $\eta^2 = Var[\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)]$.

• A consistent estimator of η^2 is just the sample variance of the individual differences, $\hat{d}_i \equiv \ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)$:

$$\hat{\eta}^2 \equiv N^{-1} \sum_{i=1}^{N} (\hat{d}_i - \bar{\hat{d}})^2. \tag{12}$$

• Voung's model selection (VMS) statistic is

$$N^{-1/2}(\mathcal{L}_1 - \mathcal{L}_2)/\hat{\eta} = \frac{N^{-1/2} \sum_{i=1}^{N} \hat{d}_i}{\left[N^{-1} \sum_{i=1}^{N} (\hat{d}_i - \bar{\hat{d}})^2 \right]^{1/2}} \xrightarrow{d} Normal(0,1), \tag{13}$$

where the limiting standard normal distribution holds under H_0.

• If we use $(N-1)^{-1}$ in place of N^{-1} in computing $\hat{\eta}^2$ we get the standard t statistic for testing a zero mean for $\hat{d}_i$. (But we act as if $\hat{d}_i = d_i$, which is justified asymptotically.)

• To make thecomputations simple, compute $\hat{d}_i$ for each i and then regress $\hat{d}_i$on 1, $i = 1, \ldots, N$, to test that the mean is different from zero.

• Need to understand the scope of its application, including the underlying null hypothesis. Cannot use the VMS statistic and its limiting standard normal distribution for testing nested models under correct specification. Recall that the LR statistic is simply $LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r)$, and, under the null, LR has a limiting χ^2_Q distribution, where Q is the number of restrictions. The important point is that, if the models are nested and correctly specified, then $\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*) = \ell_i(\boldsymbol{\theta}_o) - \ell_i(\boldsymbol{\theta}_o) = 0$.

• This degeneracy, namely, that $\eta^2 = 0$, makes the VMS statistic useless.

• The sense in which the models must be nonnested to apply Vuong's approach is that

$$P[\ell_{i1}(\boldsymbol{\theta}_1^*) \neq \ell_{i2}(\boldsymbol{\theta}_2^*)] > 0. \tag{14}$$

In other words, the log-likelihoods evaluated at the psuedo true values $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ must differ for a nontrivial set of outcomes on $(\mathbf{x}_i, \mathbf{y}_i)$. This not only rules out models that are obviously nested, but it rules out other degeneracies, too.

• For example, if y_i is a count variable, $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iK})$, and we specify different Poisson distributions – the first with mean function $\exp(\mathbf{x}_i\boldsymbol{\theta})$ and the second with mean function $(\mathbf{x}_i\boldsymbol{\theta})^2$ – these models are nonnested provided that the mean of y_i given $\mathbf{x}_i$ actually depends on the nonconstant elements in $\mathbf{x}_i$.

• But if $E(y_i|\mathbf{x}_i) = E(y_i)$, then $f_1(y|\mathbf{x};\boldsymbol{\theta}_1^*)$ and $f_2(y|\mathbf{x};\boldsymbol{\theta}_2^*)$ are Poisson distributions with the same (constant) means, and the limiting standard normal distribution for Vuong's statistic fails.

- On the other hand, if the cpmpeting models are Poisson and geometric, even with the same mean function, say $\exp(\mathbf{x}_i\boldsymbol{\theta})$, the models are nonnested no matter what because the Poisson and geometric distributions differ even if they both have constant means.
- Because the models must be nonnested, $E[\ell_{i1}(\boldsymbol{\theta}_1^*)] = E[\ell_{i2}(\boldsymbol{\theta}_2^*)]$ can only hold if *both* models are misspecified. If one model were correctly specified, yet the densities differed, then we would have a strict inequality in favor of the correctly specified model.

- Summary: Vuong's test: it applies to nonnested models where the null hypothesis is that both models are misspecified yet fit equally well.
- If we reject model 2 in favor of model 1 because *VSM* is statistically greater than zero, then we can only conclude that model 1 fits better in the sense that $E[\ell_{i1}(\boldsymbol{\theta}_1^*)] > E[\ell_{i2}(\boldsymbol{\theta}_2^*)]$. It does *not* mean that model 1 is correctly specified (although it could be).
- There are many models that can fit better than a given model, and clearly not all can be correct.

Example: Probit versus Logit for Labor Force Participation

```
. probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------

. predict phat_p
(option pr assumed; Pr(inlf))

. gen ll_p = inlf*log(phat_p) + (1 - inlf)*log(1 - phat_p)
```

```
. logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Logistic regression                                     Number of obs   =        753
                                                        LR chi2(7)      =     226.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -401.76515                             Pseudo R2       =     0.2197

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0213452   .0084214    -2.53   0.011    -.0378509   -.0048394
        educ |   .2211704   .0434396     5.09   0.000     .1360303    .3063105
       exper |   .2058695   .0320569     6.42   0.000     .1430391    .2686999
     expersq |  -.0031541   .0010161    -3.10   0.002    -.0051456   -.0011626
         age |  -.0880244    .014573    -6.04   0.000     -.116587   -.0594618
     kidslt6 |  -1.443354   .2035849    -7.09   0.000    -1.842373   -1.044335
     kidsge6 |   .0601122   .0747897     0.80   0.422     -.086473    .2066974
       _cons |   .4254524   .8603696     0.49   0.621    -1.260841    2.111746
------------------------------------------------------------------------------

. predict phat_l
(option pr assumed; Pr(inlf))

. gen ll_l = inlf*log(phat_l) + (1 - inlf)*log(1 - phat_l)
```

```
. gen diffll = ll_p - ll_l

. reg diffll

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  0,   752) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  .128152988   752  .000170416           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  .128152988   752  .000170416           Root MSE      =  .01305

------------------------------------------------------------------------------
      diffll |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0006148   .0004757     1.29   0.197    -.0003191    .0015487
------------------------------------------------------------------------------

. * Probit fits better than logit, but not in a statistically
. * significant sense.

. * Even if it did, would the partial effects at interesting values be
. * much affected?
```

Example: Lognormal versus Truncated Normal for Positive Hours

```
. use mroz

. tab inlf

   =1 if in |
  lab frce, |
       1975 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        325       43.16       43.16
          1 |        428       56.84      100.00
------------+-----------------------------------
      Total |        753      100.00

. * Compute Vuong test for truncated normal versus lognormal.

. gen lhours = log(hours)
(325 missing values generated)
```

```
. reg lhours nwifeinc educ exper expersq age kidslt6 kidsge6

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  7,   420) =   11.90
       Model |  66.3633428     7  9.48047755           Prob > F      =  0.0000
    Residual |  334.513835   420  .796461511           R-squared     =  0.1655
-------------+------------------------------           Adj R-squared =  0.1516
       Total |  400.877178   427   .93882243           Root MSE      =  .89245

------------------------------------------------------------------------------
      lhours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0019676   .0044436    -0.44   0.658    -.0107021    .0067668
        educ |  -.0385626   .0202098    -1.91   0.057    -.0782876    .0011624
       exper |    .073237   .0179004     4.09   0.000     .0380514    .1084225
     expersq |   -.001233   .0005378    -2.29   0.022    -.0022902   -.0001759
         age |  -.0236706    .007248    -3.27   0.001    -.0379175   -.0094237
     kidslt6 |   -.585202   .1186066    -4.93   0.000    -.8183386   -.3520654
     kidsge6 |  -.0694175   .0373355    -1.86   0.064    -.1428053    .0039703
       _cons |   7.896267   .4260789    18.53   0.000     7.058755     8.73378
------------------------------------------------------------------------------

. predict xb1
(option xb assumed; fitted values)

. predict u1, resid
(325 missing values generated)
```

```
. di sqrt(421/428)*.89245
.88512184

. * It is important to make sure we compute the LLF for the lognormal
. * distribution, which means subtracting log(hours):

. gen llf1 = log(normalden(u1/.88512184)) - log(.88512184) - lhours
(325 missing values generated)

. sum llf1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        llf1 |       428   -8.162678    .8146383  -12.79851   -6.26466

. di 428*-8.162678
-3493.6262

. * So the log likelihood for the positive part is -3,493.63
```

```
. truncreg hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0)
(note: 325 obs. truncated)

Truncated regression
Limit:   lower =          0                      Number of obs =    428
         upper =       +inf                      Wald chi2(7)  =  59.05
Log likelihood = -3390.6476                      Prob > chi2   = 0.0000

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
    nwifeinc |   .1534399   5.164279     0.03   0.976    -9.968361    10.27524
        educ |  -29.85254   22.83935    -1.31   0.191    -74.61684    14.91176
       exper |   72.62273   21.23628     3.42   0.001     31.00039    114.2451
     expersq |  -.9439967   .6090283    -1.55   0.121     -2.13767    .2496769
         age |  -27.44381   8.293458    -3.31   0.001    -43.69869   -11.18893
     kidslt6 |  -484.7109   153.7881    -3.15   0.002      -786.13   -183.2918
     kidsge6 |  -102.6574   43.54347    -2.36   0.018    -188.0011   -17.31379
       _cons |   2123.516   483.2649     4.39   0.000     1176.334    3070.697
-------------+----------------------------------------------------------------
sigma        |
       _cons |    850.766   43.80097    19.42   0.000     764.9177    936.6143
------------------------------------------------------------------------------
```

```
. predict xb2, xb

. gen u2 = hours - xb2

. gen llf2 = log(normalden(u2/850.766 )) - log(850.766 )
         - log(norm(xb2/ 850.766))

. replace llf2 = . if ~inlf
(325 real changes made, 325 to missing)

. sum llf2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        llf2 |       428   -7.922074    .7561236  -15.55169  -6.853047

. di 428*-7.922074
-3390.6477
```

```
. gen diff = llf2 - llf1
(325 missing values generated)

. reg diff

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  0,   427) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  203.606866   427  .476831069           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  203.606866   427  .476831069           Root MSE      =  .69053

------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2406037    .033378     7.21   0.000     .1749981    .3062094
------------------------------------------------------------------------------

. * The truncated normal fits substantially better, and we can reject the
. * lognormal.
```

```
. * Compute fitted values.

. * Truncated normal:

. gen yh2 = xb2 + 850.766*(normden(xb2/ 850.766)/norm(xb2/ 850.766))

. replace yh2 = . if hours == 0
(325 real changes made, 325 to missing)

. * lognormal:

. gen yh1 = exp(xb1 + (.88512184)^2/2)

. replace yh1 = . if hour == 0
(325 real changes made, 325 to missing)
```

```
. corr hours yh1
(obs=428)

             |    hours      yh1
-------------+------------------
       hours |   1.0000
         yh1 |   0.3579   1.0000

. di .3579^2
.12809241

. corr hours yh2
(obs=428)

             |    hours      yh2
-------------+------------------
       hours |   1.0000
         yh2 |   0.3723   1.0000

. di .3723^2
.13860729

. * So the truncated normal fits the conditional mean, E(hours|x,hours > 0),
. * somewhat better, too.
```

Panel Data

- Voung's approach applies directly to panel data methods when two complete densities have been specified for $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT} | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$. After all, the previous approach applies for any situation with random sampling in the cross section and completely specified densities.
- It may be computationally hard because estimation of models for $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT} | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ can be computationally hard (for example, CRE probit versus CRE logit).

• Can also be extended to partial (pooled) MLEs, provided we properly account for the time series dependence. For each t, let $f_{t1}(\mathbf{y}_t|\mathbf{x}_t;\boldsymbol{\theta}_1)$ and $f_{t2}(\mathbf{y}_t|\mathbf{x}_t;\boldsymbol{\theta}_2)$ be competing models of the conditional density in each time period. The partial log likelihoods are

$$\ell_{im}(\boldsymbol{\theta}_m) = \sum_{t=1}^{T} \log f_{tm}(\mathbf{y}_{it}|\mathbf{x}_{it};\boldsymbol{\theta}_m) = \sum_{t=1}^{T} \ell_{itm}(\boldsymbol{\theta}_m), \ m = 1,2. \tag{15}$$

- The same null hypothesis, $E[\ell_{i1}(\boldsymbol{\theta}_1^*)] = E[\ell_{i2}(\boldsymbol{\theta}_2^*)]$, makes sense in the PMLE setting (and is the weakest sense in which the models fit equally well).
- Moreover, the key result

$$N^{-1/2}\sum_{i=1}^{N}[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)] = N^{-1/2}\sum_{i=1}^{N}[\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)] + o_p(1) \qquad (16)$$

still holds under the null. Assuming $P[\ell_{i1}(\boldsymbol{\theta}_1^*) \neq \ell_{i2}(\boldsymbol{\theta}_2^*)] > 0$ is satisfied, the variance η^2 is positive.

• However, in estimating η^2, we must account for the serial dependence in

$$\{\ell_{it1}(\boldsymbol{\theta}_1^*) - \ell_{it2}(\boldsymbol{\theta}_2^*) : t = 1, \ldots, T\}. \tag{17}$$

• Let $\hat{d}_{it} = \ell_{it1}(\hat{\boldsymbol{\theta}}_1) - \ell_{it2}(\hat{\boldsymbol{\theta}}_2)$ denote the difference in estimated log likelihoods for each t, and let $\hat{\lambda}_t = N^{-1}\sum_{i=1}^{N}\hat{d}_{it}$. Then $\hat{\eta}^2$ is easily obtained as

$$\hat{\eta}^2 = N^{-1}\sum_{i=1}^{N}\left\{\sum_{t=1}^{T}(\hat{d}_{it} - \hat{\lambda}_t)^2 + \sum_{t=1}^{T}\sum_{r\neq t}^{T}(\hat{d}_{it} - \hat{\lambda}_t)(\hat{d}_{ir} - \hat{\lambda}_t)\right\}. \tag{18}$$

• This variance estimator allows for the possibility that the mean difference in log likelihoods varies across t under the null, but that the averages across t are the same.

• If the null hypothesis is the stronger version, $E[\ell_{it1}(\boldsymbol{\theta}_1^*)] = E[\ell_{it2}(\boldsymbol{\theta}_2^*)]$ for $t = 1, \ldots, T$, then $\hat{\lambda}_t$ can be replaced with the average of $\hat{d}_{it}$ across i and t, say $\hat{\lambda}$. In this case, the test statistic is simply the t statistic $\hat{\lambda}/\text{se}(\hat{\lambda})$, where $\text{se}(\hat{\lambda})$ is the heteroskedasticity and serial correlation robust standard error from the pooled regression $\hat{d}_{it}$ on 1, $t = 1, \ldots, T; i = 1, \ldots, N$.

• Vuong's model selection test is different from other tests in the context of nonnested models. The Cox (1961, 1962) approach tests a specified model against a nonnested alternative, and a key component of the test is the average difference in log-likelihoods, $(\mathcal{L}_1 - \mathcal{L}_2)/N$. But with Cox's approach, one model is taken to be the correct model under the null hypothesis.

• Usually the procedure is carried out with each model in turn assumed to be true under H_0.

4. QMLE IN THE LINEAR EXPONENTIAL FAMILY

- In some cases, we are willing to believe some feature of a distribution is correctly specified, but allow other features to be misspecified.
- Here we cover the case where the conditional mean, $E(y|\mathbf{x})$, is correctly specified, but other features of the distribution need not be.
- The main question is: For what set of density functions will a quasi-MLE consistently estimate the parameters in a correctly specified conditional mean? Another important issue concerns appropriate inference.

• Motivation: Suppose $E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o)$ for a known function $m(\cdot, \cdot)$ for some $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$. We know that, under identification and weak regularity conditions, nonlinear least squares is consistent for $\boldsymbol{\theta}_o$.

• The NLS estimator is easily seen to be the quasi-MLE if we specify, say,

$$D(y|\mathbf{x}) = Normal(m(\mathbf{x}, \boldsymbol{\theta}_o), 1). \tag{19}$$

• The point is this: $D(y|\mathbf{x})$ may differ in essentially arbitrary ways from normality with a unit variance, yet the QMLE under this assumption is consistent for the parameters in $E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o)$.

- The homoskedastic normal density (with variance fixed at some value) is a member of the **linear exponential family** (**LEF**).
- It turns out that any member of the LEF has the feature that a correctly specified mean is identified by the associated QMLE, even when the rest of the distribution is misspecified. These results were obtained by Gourieroux, Monfort, and Trognon (1984a), or GMT (1984a).

- A log-likelihood in the LEF written as a function of the mean as

$$\log f(y|\mu) = a(\mu) + b(y) + yc(\mu), \tag{20}$$

for functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$.

- Notice in the last term y appears linearly.

• Let M denote the set of possible values of the mean. GMT (1984a) show that $\mu_o \equiv E(y_i)$ solves

$$\max_{\mu \in M}[a(\mu) + E(y_i)c(\mu)] \;\; = \max_{\mu \in M}[a(\mu) + \mu_o c(\mu)]. \tag{21}$$

• The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are easily obtained for the normal, Bernoulli, Poisson, exponential, and other cases; GMT (1984a) contains a summary table.

• For the Bernoulli distribution,

$$\begin{aligned} \log f(y|\mu) &= (1-y) \cdot \log(1-\mu) + y \cdot \log(\mu) \\ &= \log(1-\mu) + y \cdot \log[\mu/(1-\mu)], \;\; 0 < \mu < 1. \end{aligned} \tag{22}$$

Therefore, $a(\mu) = \log(1-\mu)$ and $b(y) = 0$, and $c(\mu) = \log[\mu/(1-\mu)]$.

- For some distributions to fit into the LEF, notably the gamma and negative binomial, a nuisance parameter must be fixed at a specific value; see GMT (1984a) for details.
- In the most popular examples, it is easy to directly verify that μ_o maximizes $a(\mu) + \mu_o c(\mu)$.
- For now, focus on the meaning of the result.

• Consider the Bernoulli case but where y_i is *any* random variable with support in the unit interval, $[0, 1]$. y_i can be discrete, continuous, or have both features. For example, we could have $P(y_i = 0) > 0$ but $P(y_i = y) = 0$ for $y \in (0, 1]$, or y_i might take on values in $\{0, 1/m_i, 2/m_i, \ldots, 1\}$ for some positive integer m_i.

• Key point: Regardless of the nature of y_i – except that $0 \leq y_i \leq 1$ – which means its mean μ_o is in $(0, 1)$, μ_o, maximizes the expected value of the Bernoulli log-likelihood.

- Maximizing the log-likelihood of a density in the LEF for a random sample always leads to sample average as the estimate for μ_0. (This is typically shown in basic statistics courses for the Bernoulli, geometric, Poisson, exponential, and normal densities.)
- Under random sampling, we know the sample average is generally consistent for $\mu_0 = E(y)$ for any distribution of y provided $E(|y|) < \infty$. So the QMLE in these cases is robust for estimating μ_0.

• In practice, we are interested in conditional means, which we parameterize as $m(\mathbf{x}, \boldsymbol{\theta})$. Then the conditional quasi-log-likelihood function becomes

$$\log f(y|m(\mathbf{x}, \boldsymbol{\theta})) = a(m(\mathbf{x}, \boldsymbol{\theta})) + b(y) + yc(m(\mathbf{x}, \boldsymbol{\theta})). \tag{23}$$

Because the mean is now assumed to be correctly specified, we assume there is $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$ such that $E(y_i|\mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\theta}_o)$.

• A simple iterated expectations argument shows that $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[a(m(\mathbf{x}_i,\boldsymbol{\theta})) + y_i c(m(\mathbf{x}_i,\boldsymbol{\theta}))], \tag{24}$$

regardless of the actual distribution $D(y_i|\mathbf{x}_i)$.

• For emphasis: The nature of y_i need not even correspond to the chosen density. For example, y_i could be a nonnegative, continuous variable, and we use the Poisson quasi-log-likelihood, which is in the LEF. The Poisson QMLE is consistent for the conditional mean parameters provided the mean – with the leading case being an exponential function – is correctly specified.

• A useful characterizations of QMLE in the LEF is based on the score. It can be shown that the score has the form

$$\mathbf{s}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta})'[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]/v(m(\mathbf{x}_i, \boldsymbol{\theta})) \tag{25}$$

where $\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta})$ is the $1 \times P$ gradient of the mean function and, importantly, $v(\mu)$ is the variance function associated with the chosen LEF density. For the standard normal, $v(\mu) = 1$, for the Bernoulli, $v(\mu) = \mu(1 - \mu)$, for the Poisson $v(\mu) = \mu$, and for the exponential, $v(\mu) = \mu^2$.

• The structure of the score shows immediately that the QMLE is Fisher consistent if: $E(y_i|\mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\theta}_o)$ then $E[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{0}$, which in turn implies that the unconditional mean of the score is zero.

• We can also use the score to compute the expected Hessian conditional on $\mathbf{x}_i$:

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = -E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o) / v(m(\mathbf{x}_i, \boldsymbol{\theta}_o)). \quad (26)$$

Further,

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = E(u_i^2|\mathbf{x}_i) \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \boldsymbol{\theta}_o) / [v(m(\mathbf{x}_i, \boldsymbol{\theta}_o))]^2 \quad (27)$$

where $u_i \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)$.

• From the previous expressions, it follows immediately that the conditional information matrix equality holds if $E(u_i^2|\mathbf{x}_i) = v(m(\mathbf{x}_i, \boldsymbol{\theta}_o))$, that is

$$Var(y_i|\mathbf{x}_i) = v(m(\mathbf{x}_i, \boldsymbol{\theta}_o)). \tag{28}$$

• In other words, *if* the chosen LEF density has has a conditional variance equal to the actual $Var(y_i|\mathbf{x}_i)$, then we can use the ususal MLE standard errors and inference (even if features of the distribution other than the first two conditional moments are misspecified).

• In the Bernoulli case, $v(m) = m(1 - m)$, and in the Poisson case, $v(m) = m$.

• For example, in a Poisson regression analysis, if $Var(y_i|\mathbf{x}_i) = E(y_i|\mathbf{x}_i)$ and the mean function is correctly specified, we can act as if we are using MLE rather than quasi-MLE, even if higher-order conditional moments of y_i do not match up with the Poisson distribution.

• If $Var(y_i|\mathbf{x}_i)$ is unrestricted, the information matrix equality will not hold, and then the fully robust sandwich estimator

$$\widehat{Avar(\hat{\boldsymbol{\theta}})} = \left(\sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'\right)\left(\sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\right)^{-1}. \quad (29)$$

• The above formula is allowed because we are assuming correct specification of the conditional mean. Here

$$\mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})' \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) / v(m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})). \tag{30}$$

• If we did not want to assume correct specification of the conditional mean we would be in setup of QMLE with general misspecification.

- Quasi-MLE in the LEF is closely related to the so-called **generalized linear model** (**GLM**) literature in statistics. The terminology and some particulars differ, and the early GLM literature did not recognize the robustness of the approach for estimating conditional mean parameters.
- In modern applications the key feature is that they both use quasi-MLE to estimate parameters of a conditional mean.

• The GLM approach is more restrictive in that the conditional mean is assumed to have an **index structure**. In particular, the mean is assumed to have the form $m(\mathbf{x}, \boldsymbol{\theta}) = r(\mathbf{x}\boldsymbol{\theta})$ where the "index" $\mathbf{x}\boldsymbol{\theta}$ is linear in parameters and $r(\cdot)$ is a function of the index.

• An important component of the GLM apparatus is the **link function**, which implicitly defines the mean function. If we let η denote the index $\mathbf{x}\boldsymbol{\theta}$, then the link function $g(\cdot)$ is such that $\eta = g(\mu)$. The link function is strictly monotonic and therefore has an inverse, and so $\mu = g^{-1}(\eta)$ or, in the notation of conditional mean functions, $m(\mathbf{x}, \boldsymbol{\theta}) = g^{-1}(\mathbf{x}\boldsymbol{\theta})$.

• The name "generalized linear model" comes from the underlying linearity of the index function, and then the link function introduces nonlinearity.

• In most applications, it is more natural to specify the conditional mean function because we want the mean function to be consistent with the nature of y_i, and y_i is the outcome we hope to explain.

• Directly specifying $m(\mathbf{x}, \boldsymbol{\theta})$ does not wed one to the the index structure, although, in most applications, $m(\mathbf{x}, \boldsymbol{\theta})$ has an index form. If, say, $m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta})$ then the link function is $g(\mu) = \log(\mu)$ for $\mu > 0$. If $m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta})/[1 + \exp(\mathbf{x}\boldsymbol{\theta})]$ then $g(\mu) = \log[\mu/(1 - \mu)]$ for $0 < \mu < 1$. See McCullagh and Nelder.

• The GLM literature recognizes that assuming $Var(y_i|\mathbf{x}_i)$ corresponds to the LEF assumption is too restrictive for many applications.

• Middle ground between $Var(y_i|\mathbf{x}_i)$ conforming to LEF and $Var(y_i|\mathbf{x}_i)$ unrestricted is

$$Var(y_i|\mathbf{x}_i) = \sigma_o^2 v(m(\mathbf{x}_i, \boldsymbol{\theta}_o)) \tag{31}$$

for some $\sigma_o^2 > 0$, which is often called the **dispersion parameter**.

• Call this the **GLM variance assumption** (because this assumption was key in the original GLM literature).

• When $\sigma_o^2 > 1$ then we say there is **overdispersion** (relative to the chosen density); **underdispersion** is when $\sigma_o^2 < 1$, and both cases arise in practice.

• Under the GLM variance assumption, it is straightforward to estimate σ_o^2. Let $u_i \equiv y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)$ be the additive "errors," so that $E(u_i|\mathbf{x}_i) = 0$ and $Var(u_i|\mathbf{x}_i) = Var(y_i|\mathbf{x}_i)$. Because

$$E(u_i^2|\mathbf{x}_i) = \sigma_o^2 v(m(\mathbf{x}_i, \boldsymbol{\theta}_o)) \equiv \sigma_o^2 v_i,$$

$$E(u_i^2/v_i) = E[E(u_i^2/v_i|\mathbf{x}_i)] = E[E(u_i^2|\mathbf{x}_i)/v_i] = E(\sigma_o^2 v_i/v_i) = \sigma_o^2. \qquad (32)$$

- By the usual analogy principle argument,

$$\hat{\sigma}^2 = (N-P)^{-1} \sum_{i=1}^{N} \hat{u}_i^2 / \hat{v}_i \tag{33}$$

is consistent for σ_o^2, where $\hat{u}_i \equiv y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ are the residuals, $\hat{v}_i \equiv v(m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}))$ are the estimated conditional variances from the LEF density.

• Degrees-of-freedom adjustment is common (but, of course, does not affect consistency). In the GLM literature, the standardized residuals $\hat{u}_i/\sqrt{\hat{v}_i}$ are called the **Pearson residuals** and the estimate based on $(\hat{u}_i/\sqrt{\hat{v}_i})^2$ is the **Pearson dispersion estimator**.

• Under the GLM variance assumption, it is easily seen that the generalized information matrix equality is satisfied. In fact, a conditional version holds, which we can called the **generalized conditional information matrix equality** (**GCIME**):

$$E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] = -\sigma_o^2 E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \sigma_o^2 \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o). \tag{34}$$

• Can use M-estimation results to obtain

$$\widehat{Avar(\hat{\boldsymbol{\theta}})} = \hat{\sigma}^2\left(\sum_{i=1}^{N}\mathbf{A}(\mathbf{x}_i,\hat{\boldsymbol{\theta}})\right)^{-1} = \hat{\sigma}^2\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}m_i(\hat{\boldsymbol{\theta}})'\nabla_{\boldsymbol{\theta}}m_i(\hat{\boldsymbol{\theta}})/v_i(\hat{\boldsymbol{\theta}})\right)^{-1}. \quad (35)$$

• Packages that have GLM commands usually allow this estimator as an option, along with a fully robust version or the "MLE" version. In Stata, the command is "glm."

• Structure essentially the same as weighted NLS estimator. In fact, can show that the QMLE is asymptotically equivalent to the WNLS estimator using the weight function $1/v(m(\mathbf{x}_i,\breve{\boldsymbol{\theta}}))$ for a preliminary consistent estimator $\breve{\boldsymbol{\theta}}$.

• For example, suppose y_i is a binary response that follows a probit model. Rather than use probit we could first estimate $\boldsymbol{\theta}_o$ by NLS to obtain $\breve{\boldsymbol{\theta}}$. Then, estimate $\boldsymbol{\theta}_o$ by WNLS using weighting function $1/[\Phi(\mathbf{x}_i\breve{\boldsymbol{\theta}})[1 - \Phi(\mathbf{x}_i\breve{\boldsymbol{\theta}})]$. The WNLS estimator is $\sqrt{N}$-equivalent to the MLE. (Of course, there is no reason to take such an approach; it is computationally more difficult than MLE.)

• The binary response case is the one LEF situation where it makes no sense to talk about $E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i)$ being correctly specified but some other feature of $D(y_i|\mathbf{x}_i)$ being misspecified.

• Under the GLM variance assumption, the QMLE has an important efficiency property: it is the efficient estimator in the class of estimators that use only

$$E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o). \tag{36}$$

• If we use the assumption $Var(y_i|\mathbf{x}_i) = \sigma_o^2 v(m(\mathbf{x}_i, \boldsymbol{\theta}_o))$ to provide more information on $\boldsymbol{\theta}_o$ then we can get more efficient estimators, but those would be less robust because they would generally be inconsistent under the conditional mean assumption only.

- The LEF can be extended to multiple responses. That is, $\mathbf{y}_i$ can be a $G \times 1$ vector. A particularly useful log likelihood in the LEF is the multinomial. The multinomial quasi-log likelihood can be used for estimating multiple fractional response models (such as expenditure or cost shares).
- For modeling the mean and variance together, say $m(\mathbf{x}_i, \boldsymbol{\theta})$ and $v(\mathbf{x}_i, \boldsymbol{\theta})$ (or multivariate versions), the normal QMLE is attractive. The QMLE in this case is consistent when the mean and variance are correctly specified with arbitrary misspecification of the rest of the distribution.

5. POOLED QMLE IN THE LEF FOR PANEL DATA

- We can extend the QMLE in the LEF to panel data. The simplest approach is to specify conditional mean functions $m_t(\mathbf{x}_t, \boldsymbol{\theta})$, $t = 1, \ldots, T,$ along with an LEF density, and then to proceed with estimation by ignoring any time dependence.
- The pooled quasi-likelihood is

$$\ell_i(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log[f_t(y_{it}|\mathbf{x}_{it}; \boldsymbol{\theta})] \tag{37}$$

where $f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta})$ is in the LEF.

- Most of the time the mean would depend on time by allowing certain parameters to change over time, such as $m_t(\mathbf{x}_t, \boldsymbol{\theta}) = \exp(\alpha_t + \mathbf{x}_t\boldsymbol{\beta})$.
- Correct specification of the mean for each t means that, for some $\boldsymbol{\theta}_o$,

$$E(y_{it}|\mathbf{x}_{it}) = m_t(\mathbf{x}_{it}, \boldsymbol{\theta}_o), \; t = 1, \ldots, T. \tag{38}$$

- This does *not* imply strict exogeneity of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$.
- The score for each t is:

$$\mathbf{s}_{it}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} m_t(\mathbf{x}_{it}, \boldsymbol{\theta})'[y_{it} - m_t(\mathbf{x}_{it}, \boldsymbol{\theta})]/v(m_t(\mathbf{x}_{it}, \boldsymbol{\theta})). \tag{39}$$

• The **pooled QMLE** (or **partial QMLE**) is generally found by solving $\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{s}_{it}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

• Generally, the scores $\{\mathbf{s}_{it}(\boldsymbol{\theta}_o) : t = 1, \ldots, T\}$ are serially correlated, which means fully robust inference is needed. (That is, a sandwich estimator $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ with $\hat{\mathbf{B}}$ allowing any kind of serial dependence in $\{\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\}$.

• With pooled MLE, we saw an important case where the scores are not serially correlated: the distribution $D(\mathbf{y}_{it}|\mathbf{x}_{it})$ is dynamically complete in the sense that it also equals $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{y}_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{x}_{i1})$.

• With QMLE in the LEF, we are only assuming a correctly specified mean. If the conditional mean is dynamically complete in the sense that

$$E(y_{it}|\mathbf{x}_{it}) = E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1}), \tag{40}$$

then it is easily seen, using $\mathbf{s}_{it}(\boldsymbol{\theta}_o) = \nabla_{\boldsymbol{\theta}} m_t(\mathbf{x}_{it}, \boldsymbol{\theta}_o)' u_{it}/v(m_t(\mathbf{x}_{it}, \boldsymbol{\theta}_o))$ with $u_{it} \equiv y_{it} - m_t(\mathbf{x}_{it}, \boldsymbol{\theta}_o)$, that

$$E[\mathbf{s}_{it}(\boldsymbol{\theta}_o)|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1})] = \mathbf{0}, \tag{41}$$

and so the scores evaluated at $\boldsymbol{\theta}_o$ are serially uncorrelated.

• We need not assume any other features of the LEF density, such as the conditional variance, are correctly specified.

• Assuming the scores are serially uncorrelated but without any assumptions on $Var(y_{it}|\mathbf{x}_{it})$, the appropriate asymptotic variance estimator of the pooled QMLE is

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_{\theta}\hat{m}_{it}'\nabla_{\theta}\hat{m}_{it}/\hat{v}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^{2}\nabla_{\theta}\hat{m}_{it}'\nabla_{\theta}\hat{m}_{it}/\hat{v}_{it}^{2}\right)^{-1} \cdot \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_{\theta}\hat{m}_{it}'\nabla_{\theta}\hat{m}_{it}/\hat{v}_{it}\right)^{-1}. \tag{42}$$

• The “glm” option in Stata is “robust.”

• The estimate that allows unrestricted serial correlation, in addition to any variances, is

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\hat{m}_{it}'\nabla_{\boldsymbol{\theta}}\hat{m}_{it}/\hat{v}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\hat{u}_{it}\hat{u}_{ir}\nabla_{\boldsymbol{\theta}}\hat{m}_{it}'\nabla_{\boldsymbol{\theta}}\hat{m}_{it}/\hat{v}_{it}\hat{v}_{ir}\right)^{-1}$$
$$\cdot\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\hat{m}_{it}'\nabla_{\boldsymbol{\theta}}\hat{m}_{it}/\hat{v}_{it}\right)^{-1}. \qquad (43)$$

- The sandwich estimator in equation (43) can be computed using "cluster" options – we will see specific examples later.
- This variance estimate still assumes the condtional mean is correctly specified (for each t).
- Notice that, for QMLE in the LEF, the serial correlation issue essentially comes down to (conditional) serial correlation in the errors, $\{u_{it} : t = 1, \ldots, T\}$, where $u_{it} \equiv y_{it} - E(y_{it}|\mathbf{x}_{it})$.

6. GENERALIZED ESTIMATING EQUATIONS FOR PANEL DATA

- We can always used a pooled QMLE and make inference fully robust to serial correlation, but the pooled estimator may be imprecise, especially with lots of serial correlation. Can we do better?
- If we make the stronger assumption of strict exogeneity of the regressors, it is possible to obtain a more efficient estimator than pooled QMLE: multivariate weighted nonlinear least squares (MWNLS).
- But the relative efficiency of the MWNLS approach relies on being able to properly find the $T \times T$ conditional variance matrix, $Var(\mathbf{y}_i|\mathbf{x}_i)$.

• The **generalized estimating equations** (**GEE**) approach to panel data is MWNLS where one explicitly recognizes that the chosen model for $Var(\mathbf{y}_i|\mathbf{x}_i)$ is incorrect.

• Once we maintain strict exogeneity of $\{\mathbf{x}_{it} : t = 1, 2, \ldots, T\}$, think of GEE as having the same starting point as QMLE in the LEF: the mean function is chosen to be consistent with the nature of y_{it} (for example, nonnegative or fractional), and then the LEF density is chosen to be well-defined when evaluated at y_{it}.

- As with pooled QMLE, the variance associated with the chosen LEF is allowed to be misspecified. GEE uses the nominal (or “working”) variance assumption in a multivariate weighted least squares procedure.
- But GEE takes a further step: rather than ignoring serial dependence in the implied errors over time – as in pooled QMLE – it uses simple correlation structures in implementing MWNLS, recognizing that these structures are almost certainly wrong.

• General GEE approach starts with

$$E(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta}_o), \text{ some } \boldsymbol{\theta}_o \in \boldsymbol{\Theta}, \tag{44}$$

where $\mathbf{y}_i$ and $\mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta}_o)$ are $T \times 1$ and $\mathbf{x}_i$ is the collection of all regressors across all time periods.

• Notice that the vector $E(\mathbf{y}_i|\mathbf{x}_i)$ has elements $E(y_{it}|\mathbf{x}_i)$. In most cases, we restrict how $E(y_{it}|\mathbf{x}_i)$ depends on $\mathbf{x}_i$, for example, $E(y_{it}|\mathbf{x}_i) = E(y_{it}|\mathbf{x}_{it})$. In practice, this means that we assume time-varying regressors are strictly exogenous.

• In other words, like random effects (or general GLS) in a linear model, GEE assumes strictly exogenous explanatory variables.

• Let $\mathbf{V}(\mathbf{x}_i, \boldsymbol{\theta})$ be the $T \times T$ diagonal matrix with the LEF variances down the diagonal. So, for the Bernoulli, the t^{th} diagonal element is $v_t(\mathbf{x}_i, \boldsymbol{\theta}) \equiv m_t(\mathbf{x}_i, \boldsymbol{\theta})[1 - m_t(\mathbf{x}_i, \boldsymbol{\theta})]$, where we need $0 < m_t(\mathbf{x}_i, \boldsymbol{\theta}) < 1$. For the Poisson, the diagonal elements are $v_t(\mathbf{x}_i, \boldsymbol{\theta}) = m_t(\mathbf{x}_i, \boldsymbol{\theta})$, $t = 1, \ldots, T$.

• In GEE, the **working correlation matrix** is typically chosen to be constant (that is, not as a function of $\mathbf{x}_i$).

• In other words, define the standardized errors as $e_{it} \equiv u_{it}/\sqrt{v_{it}}$ (the population versions of the Pearson residuals), where $u_{it} = y_{it} - E(y_{it}|\mathbf{x}_{it})$, $v_{it} = v_t(\mathbf{x}_i, \boldsymbol{\theta}_o)$. Note that as long as the mean is correctly specified then $E(e_{it}|\mathbf{x}_i) = 0$. Further, *if* the LEF variance is correctly specified, then $Var(e_{it}|\mathbf{x}_i) = 1$ (or, under the GLM variance assumption $Var(y_{it}|\mathbf{x}_i) = \sigma_o^2 v_t(\mathbf{x}_i, \boldsymbol{\theta}_o)$, $Var(e_{it}|\mathbf{x}_i) = \sigma_o^2$).

• GEE nominally acts as if $Corr(e_{it}, e_{is}|\mathbf{x}_i) = Corr(e_{it}, e_{is})$. Even if we assume the GLM variance assumption, $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ is virtually never constant when y_{it} has discreteness, or if it is limited in some other way (say, $y_{it} \geq 0$).

• In other words, GEE explicitly recognizes that the correlation matrix we "work with" may not equal $Corr(e_{it}, e_{is}|\mathbf{x}_i)$.

• Further, if the GLM variance assumption is wrong, $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ would depend on $\mathbf{x}_i$ except by fluke because $Var(e_{it}|\mathbf{x}_i)$ would depend on $\mathbf{x}_i$.

• If no restrictions are imposed on the constant correlations, it is an **unstructured working correlation matrix**:

$$\mathbf{R}(\boldsymbol{\rho}) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1T} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2T} \\ \rho_{13} & \rho_{23} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & 1 & \rho_{T-1,T} \\ \rho_{1T} & \rho_{2T} & \cdots & \rho_{T-1,T} & 1 \end{pmatrix} \tag{45}$$

• Given an initial consistent estimator $\check{\theta}$ (almost certainly the pooled QMLE), we can estimate each ρ_{ts} as

$$\begin{aligned} \hat{\rho}_{ts} &= \text{Sample Correlation}(\check{u}_{it}/\sqrt{\check{v}_{it}}\,, \check{u}_{is}/\sqrt{\check{v}_{is}}\,) \\ &\equiv \text{Sample Correlation}(\check{e}_{it}, \check{e}_{is}) \end{aligned} \tag{46}$$

where $\check{e}_{it} \equiv \check{u}_{it}/\sqrt{\check{v}_{it}}$ are the standardized (Pearson) residuals.

• Again, it is important to remember that we are not assuming the true conditional correlation matrix is constant. It virtually never is in LEF applications (except for linear models).

• Nevertheless, under general conditions, $\hat{\rho}_{st}$ converges in probability to, say, ρ^*_{st}, the population correlation between $e_{it} = u_{it}/\sqrt{v_t(\mathbf{x}_i, \boldsymbol{\theta}_o)}$ and $e_{is} = u_{is}/\sqrt{v_s(\mathbf{x}_i, \boldsymbol{\theta}_o)}$.

• A common form of $\mathbf{R}(\boldsymbol{\rho})$ in panel data settings is an **exchangeable working correlation matrix**, which introduces a single correlation parameter, ρ, for all pairs: $\rho_{st} = \rho$.

- Given $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\rho}}$, the **working variance matrix** estimates are

$$\hat{\mathbf{W}}_i = \mathbf{V}(\mathbf{x}_i, \check{\boldsymbol{\theta}})^{1/2} \mathbf{R}(\hat{\boldsymbol{\rho}}) \mathbf{V}(\mathbf{x}_i, \check{\boldsymbol{\theta}})^{1/2}. \tag{47}$$

- GEE is then essentially multivariate WNLS: $\hat{\boldsymbol{\theta}}$ solves

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} [\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})]' \hat{\mathbf{W}}_i^{-1} [\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})]. \tag{48}$$

- If the mean function is correctly specified, $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\rho}}$ do not affect the limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$. (This is typically assumed in GEE. The standard errors are labeled “semi-robust” in Stata.)

• The expression for the estimated asymptotic variance is the sandwich form,

$$\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i'\hat{\mathbf{W}}_i^{-1}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i\right)^{-1}\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i'\hat{\mathbf{W}}_i^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\hat{\mathbf{W}}_i^{-1}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i\right)$$
$$\cdot\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i'\hat{\mathbf{W}}_i^{-1}\nabla_{\boldsymbol{\theta}}\hat{\mathbf{m}}_i\right)^{-1} \tag{49}$$

• This allows both the variances from the LEF density and the constant correlation structure to be misspecified.

- Wald tests for joint restrictions are easily computed.
- In State, the command is “xtgee.” We will see specific examples later with count data and fractional responses.
- GEE also applies to binary responses as a substitute for, say, pooled probit and RE probit. GEE can be more efficient than pooled probit while being more robust to serial correlation than RE probit.

• Why use MWNLS if we are admitting that the variance-covariance matrix is misspecified? Experience (simulations) show that accounting for serial correlation in a misspecified way is often better than ignoring it entirely.

• Of course, MLE using the joint distribution $D(\mathbf{y}_i|\mathbf{x}_i)$ is best, but usually not robust and often computationally demanding. (An example is RE probit.)

• GEE tries to get back some of the efficiency lost by not using full MLE, but it maintains the robustness of pooled methods (assuming strict exogeneity). Remember, GEE is consistent whether or not the model for $Var(\mathbf{y}_i|\mathbf{x}_i)$ is correct.

• For example, suppose we start with

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\theta}_o + c_i)$$

and assume c_i is independent of $\mathbf{x}_i$ and normally distributed. For average partial effects, we can used pooled probit, but this is likely to be inefficient. We can use RE probit, but this uses conditional independence, $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i, c)$. As a middle ground, we can use GEE. If we maintain strict exogeneity, GEE uses the same assumptions for consistency as pooled probit.

- If we want to use GEE and allow $\mathbf{m}(\mathbf{x}_i, \boldsymbol{\theta})$ to be misspecified for $E(\mathbf{y}_i|\mathbf{x}_i)$, it is tricky to obtain valid standard errors via the delta method (but not impossible). These are sometimes called "fully robust" standard errors in the GEE literature.
- An alternative way to obtain valid inference under misspecification of the mean is to implement the panel bootstrap incorporating the estimation in all three stages (namely, $\check{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\rho}}$, and $\hat{\boldsymbol{\theta}}$).

• Pooled QMLE and GEE methods are attractive for conditional mean models with unobserved heterogeneity, say c_i, and strictly exogenous regressors conditional on c_i: $E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},c_i) = E(y_{it}|\mathbf{x}_{it},c_i)$. When we combine strict exogeneity with a specific correlated random effects assumptions, such as $c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$, where $\bar{\mathbf{x}}_i$ is the vector of time averages and a_i is independent of $(\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$, then we can often find $E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$ as a function of $(\mathbf{x}_{it},\bar{\mathbf{x}}_i)$.

• Or, we can just start with a model for $E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i)$.

• We will cover exponential and fractional regression later, where we will derive

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = \exp(\alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}) \tag{50}$$

and, if we add normality of a_i,

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = E(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = \Phi(\alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}). \tag{51}$$

Empirical Application: **Labor Force Participation**

- For binary responses, can apply GEE as an alternative to pooled methods or random effects estimation. GEE is computationally much simpler than, say, RE probit.
- GEE is also more robust than RE probit. Like pooled probit, we only need the response probability at time *t* to be correctly specified. We do need to maintain strict exogeneity of the covariates (as with RE probit).

• If we start with the usual unobserved effects model and impose the Chamberlain-Mundlak device,

$$\begin{aligned} y_{it} &= 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + r_{it} > 0] \\ &= 1[\mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + r_{it} > 0] \end{aligned}$$

then both pooled probit and GEE estimate scaled coefficients if $(a_i + r_{it})$ is independent of $\mathbf{x}_i$ with the $Normal(0, 1 + \sigma_a^2)$ distribution.

• RE probit, which imposes independence of $\{r_{it} : t = 1, \ldots, T\}$, estimates the unscaled coefficients and σ_a^2.

• If we define

$$e_{it} = \frac{[y_{it} - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{w}}_i\boldsymbol{\xi}_a)]}{\{\Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{w}}_i\boldsymbol{\xi}_a)[1 - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{w}}_i\boldsymbol{\xi}_a)]\}^{1/2}}$$

it is very difficult to derive $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ even when $\{r_{it} : t = 1, \ldots, T\}$ is serially independent, and it is not constant unless $\boldsymbol{\beta}_a = \mathbf{0}$, $\boldsymbol{\xi}_a = \mathbf{0}$.

• GEE uses the "working" assumption that $Corr(e_{it}, e_{is}|\mathbf{x}_i) = Corr(e_{it}, e_{is})$, and may further restrict the working correlations to be the same for all t and s.

```
. use lfp

. tab period

  1 through |
  5, each 4 |
months long |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      5,663       20.00       20.00
          2 |      5,663       20.00       40.00
          3 |      5,663       20.00       60.00
          4 |      5,663       20.00       80.00
          5 |      5,663       20.00      100.00
------------+-----------------------------------
      Total |     28,315      100.00

. tab lfp if per5

   =1 if in |
labor force |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,850       32.67       32.67
          1 |      3,813       67.33      100.00
------------+-----------------------------------
      Total |      5,663      100.00

. egen kidsbar = mean(kids), by(id)

. egen lhincbar = mean(lhinc), by(id)
```

```
. * First compute standard errors assuming probit model is correct and
. * no serial correlation

. probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5

Probit regression                                 Number of obs   =      28315
                                                  LR chi2(12)     =    2385.17
                                                  Prob > chi2     =     0.0000
Log likelihood = -16516.436                       Pseudo R2       =     0.0673

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1173749   .0372874    -3.15   0.002    -.1904569    -.044293
       lhinc |  -.0288098   .0248077    -1.16   0.246     -.077432    .0198125
     kidsbar |  -.0856913   .0380322    -2.25   0.024     -.160233   -.0111495
    lhincbar |  -.2501781   .0290625    -8.61   0.000    -.3071396   -.1932167
        educ |   .0841338   .0032539    25.86   0.000     .0777562    .0905114
       black |   .2030668   .0335069     6.06   0.000     .1373945     .268739
         age |   .1516424   .0062081    24.43   0.000     .1394748    .1638101
       agesq |  -.0020672   .0000762   -27.13   0.000    -.0022166   -.0019179
        per2 |  -.0135701   .0253864    -0.53   0.593    -.0633265    .0361862
        per3 |  -.0331991   .0253348    -1.31   0.190    -.0828544    .0164562
        per4 |  -.0390317   .0253325    -1.54   0.123    -.0886825     .010619
        per5 |  -.0552425   .0252773    -2.19   0.029    -.1047851   -.0056999
       _cons |  -.7260562     .14268    -5.09   0.000    -1.005704   -.4464086
------------------------------------------------------------------------------
```

```
. * Now allow probit model to be wrong but (for some reason) do not allow
. * serial correlation

. probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5, robust

Probit regression                                     Number of obs   =      28315
                                                      Wald chi2(12)   =    2103.73
                                                      Prob > chi2     =     0.0000
Log pseudolikelihood = -16516.436                     Pseudo R2       =     0.0673

------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1173749   .0393598    -2.98   0.003    -.1945187   -.0402311
       lhinc |  -.0288098   .0263352    -1.09   0.274    -.0804259    .0228063
     kidsbar |  -.0856913   .0400968    -2.14   0.033    -.1642796    -.007103
    lhincbar |  -.2501781   .0306941    -8.15   0.000    -.3103375   -.1900187
        educ |   .0841338   .0033147    25.38   0.000     .0776371    .0906305
       black |   .2030668     .03372     6.02   0.000     .1369768    .2691568
         age |   .1516424   .0062405    24.30   0.000     .1394114    .1638735
       agesq |  -.0020672   .0000771   -26.82   0.000    -.0022183   -.0019162
        per2 |  -.0135701   .0253565    -0.54   0.593     -.063268    .0361277
        per3 |  -.0331991   .0252992    -1.31   0.189    -.0827847    .0163864
        per4 |  -.0390317   .0253413    -1.54   0.124    -.0886998    .0106364
        per5 |  -.0552425   .0252942    -2.18   0.029    -.1048182   -.0056668
       _cons |  -.7260562    .143069    -5.07   0.000    -1.006466   -.4456461
------------------------------------------------------------------------------
```

```
. * Fully robust (misspecified model and serial correlation) inference below.
. * Standard errors are substantially smaller!

. probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5,
        cluster(id)

Probit regression                                 Number of obs   =      28315
                                                  Wald chi2(12)   =     538.09
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -16516.436                 Pseudo R2       =     0.0673

                                  (Std. Err. adjusted for 5663 clusters in id)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1173749   .0269743    -4.35   0.000    -.1702435   -.0645064
       lhinc |  -.0288098    .014344    -2.01   0.045    -.0569234   -.0006961
     kidsbar |  -.0856913   .0311857    -2.75   0.006     -.146814   -.0245685
    lhincbar |  -.2501781   .0352907    -7.09   0.000    -.3193466   -.1810097
        educ |   .0841338   .0067302    12.50   0.000     .0709428    .0973248
       black |   .2030668   .0663945     3.06   0.002     .0729359    .3331976
         age |   .1516424   .0124831    12.15   0.000      .127176    .1761089
       agesq |  -.0020672   .0001553   -13.31   0.000    -.0023717   -.0017628
        per2 |  -.0135701   .0103752    -1.31   0.191    -.0339051    .0067648
        per3 |  -.0331991   .0127197    -2.61   0.009    -.0581293    -.008269
        per4 |  -.0390317   .0136244    -2.86   0.004    -.0657351   -.0123284
        per5 |  -.0552425   .0146067    -3.78   0.000    -.0838711   -.0266139
       _cons |  -.7260562   .2836985    -2.56   0.010    -1.282095   -.1700173
------------------------------------------------------------------------------
```

```
. xtgee lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5,
        fam(binomial) link(probit) corr(exch) robust

GEE population-averaged model                   Number of obs      =     28315
Group variable:                         id      Number of groups   =      5663
Link:                                probit     Obs per group: min =         5
Family:                            binomial                    avg =       5.0
Correlation:                   exchangeable                    max =         5
                                                Wald chi2(12)      =    536.66
Scale parameter:                          1     Prob > chi2        =    0.0000

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semirobust
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.1125361   .0281366    -4.00   0.000    -.1676828   -.0573894
       lhinc |  -.0276543    .014799    -1.87   0.062    -.0566598    .0013511
     kidsbar |  -.0892543   .0323884    -2.76   0.006    -.1527344   -.0257742
    lhincbar |   -.252001   .0360377    -6.99   0.000    -.3226337   -.1813684
        educ |   .0841304   .0066834    12.59   0.000     .0710312    .0972296
       black |    .205611   .0668779     3.07   0.002     .0745328    .3366893
         age |    .152809   .0125434    12.18   0.000     .1282245    .1773936
       agesq |  -.0020781   .0001565   -13.28   0.000    -.0023847   -.0017714
...
       _cons |  -.7532503    .285216    -2.64   0.008    -1.312263   -.1942373
------------------------------------------------------------------------------
```

• Surprisingly, the GEE approach does not improve the precision of the pooled probit estimators. The robust standard errors for GEE are slightly above those for pooled probit. This finding is particularly puzzling because there is substantial serial correlation in the standardized residuals, written generally after pooled probit estimation as

$$\hat{e}_{it} \equiv \frac{[y_{it} - \Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}}_a)]}{\{\Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}}_a)[1 - \Phi(\mathbf{x}_{it}\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{w}}_i\hat{\boldsymbol{\xi}}_a)]^{1/2}},$$

where $\bar{\mathbf{w}}_i$ is the time average of variables that change across i and t ($kids_{it}$ and $lhinc_{it}$ in this application). The first-order correlation in the $\{\hat{e}_{it} : t = 2, \ldots, T; i = 1, \ldots, N\}$ is about .83.

```
. qui probit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5

. predict phat
(option pr assumed; Pr(lfp))

. gen eh = (lfp - phat)/sqrt(phat*(1 - phat))

. gen eh_1 = l.eh
(5663 missing values generated)

. corr eh eh_1
(obs=22652)

             |       eh     eh_1
-------------+------------------
          eh |   1.0000
        eh_1 |   0.8315   1.0000
```

```
. * Now CRE probit estimated by joint MLE. Stata allows no robust inference.
. * Remember that the pooled probit coefficients are scaled versions of the
. * RE coefficients.

. xtprobit lfp kids lhinc kidsbar lhincbar educ black age agesq per2-per5, re

Random-effects probit regression              Number of obs      =     28315
Group variable: id                            Number of groups   =      5663

Random effects u_i ~Gaussian                  Obs per group: min =         5
                                                             avg =       5.0
                                                             max =         5

                                              Wald chi2(12)      =    623.40
Log likelihood  = -8609.9002                  Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.3970102   .0701298    -5.66   0.000     -.534462   -.2595584
       lhinc |  -.1003399   .0469979    -2.13   0.033    -.1924541   -.0082258
     kidsbar |  -.4085664   .0898875    -4.55   0.000    -.5847428   -.2323901
    lhincbar |  -.8941069   .1199703    -7.45   0.000    -1.129244   -.6589695
        educ |   .3189079    .024327    13.11   0.000     .2712279     .366588
       black |   .6388784   .1903525     3.36   0.001     .2657945    1.011962
         age |   .7282057   .0445623    16.34   0.000     .6408651    .8155462
       agesq |  -.0098358   .0005747   -17.11   0.000    -.0109623   -.0087094
...
       _cons |  -5.359375   1.000514    -5.36   0.000    -7.320346   -3.398404
```

```
-------------+----------------------------------------------------------------
    /lnsig2u |   2.947234   .0435842                      2.861811    3.032657
-------------+----------------------------------------------------------------
     sigma_u |   4.364995   .0951224                      4.182484     4.55547
         rho |   .9501326    .002065                       .945926    .9540279
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  1.6e+04 Prob >= chibar2 = 0.000

. * This version of Stata (10 or 11) gives different estimates from those
. * in Table 15.3. The log likelihood is now substantially higher, too.
```