ECONOMICS 6002 CLASS 15 - 16
MAXIMUM LIKELIHOOD ESTIMATION

1. Principles of Maximum Likelihood Estimation
    a. Maximum likelihood estimation is based on estimating parameter values for which the probability of drawing the data in the sample is greater than that for any other parameter values
    b. It is based on probabilistic principles rather than fitting criteria.
    c. Its main disadvantage is that the probability distribution function for the data must be fully specified.
    d. It is particularly useful in contexts (such as limited dependent variable models) in which fitting criteria produce inconsistent estimates.

2. Review of Concepts
    a. Likelihood function
        i. Generally, the log of L is easier to work with
    b. Gradient vector and likelihood equations
        i. L (and log L) is maximized where the gradient (first-order derivatives of logL) = 0

3. Examples
    a. Random sampling from unknown population
    b. Linear regression model
    c. Estimation of variance
    d. Generalized linear regression model

4. Other uses for ML estimator
    a. Models non-linear in the **dependent** variable $g(y,\theta) = h(x,\beta) + \varepsilon$
        i. e.g., limited dependent-variable models
    b. Models with non-normal distributions
        i. e.g., count data, duration models, stochastic frontier production models

5. Statistical properties
    a. Consistency
        i. Follows from *E* log*L* maximized at true value of parameters
        ii. Follows from *E* gradient = 0 at true value of parameters [$\mathbf{g}(\boldsymbol{\theta}^*) = 0$]
    b. **Not** generally unbiased in small samples
        i. If gradient vector is non-linear function of the random variables, mathematical expectation does not carry through (Example: ML estimate of the variance)
    c. **Not** generally normally distributed, but is asymptotically normal
        i. If gradient vector is non-linear function of the random variables, it is not normally distributed even if the random variables are
        ii. But Central Limit Theorem usually applies

d. Asymptotic efficiency
   i. Information number/matrix characterizes the information in the sample
      (1) Information matrix = variance of the gradient vector = - matrix of expected **second** order partial derivatives (Hessian) of log L (all evaluated at true value of the unknown parameters)
      (2) $\mathbf{I}(\boldsymbol{\theta}^*) = E\,[\mathbf{g}(\boldsymbol{\theta}^*)\,\mathbf{g}(\boldsymbol{\theta}^*)'] = -E\,\mathbf{H}(\boldsymbol{\theta}^*)$ – information matrix equality
   ii. Cramér-Rao lower bound is the variance of an estimator that is ascribed to the limited information in the sample, and is the smallest variance possible for that sample
      (1) Cramér-Rao lower bound is the inverse of the Information Matrix $\mathbf{I}(\boldsymbol{\theta}^*)^{-1}$
   iii. Asymptotic variance of ML estimator is the Cramér-Rao lower bound, and so the ML estimator is asymptotically efficient

6. Covariance matrix can be estimated by
   a. Expectation of Hessian $-E\,\mathbf{H}(\boldsymbol{\theta})^{-1}$
      i. Best, but requires taking expectations of all elements of the Hessian, which is not always possible
   b. Actual value of Hessian $-\,\mathbf{H}(\boldsymbol{\theta})^{-1}$
      i. Expectations not needed, but need to calculate all second order partial derivatives
   c. Outer Product of Gradient (BHHH estimator) $[\Sigma\,\mathbf{g}_i(\boldsymbol{\theta})\,\mathbf{g}_i(\boldsymbol{\theta})']^{-1}$
      i. Only need gradient (first-order partial derivatives)
      ii. But is frequently inaccurate in small and moderate sample sizes
      iii. Particularly useful in LM tests, for which only the restricted estimate is needed

7. Testing hypotheses
   a. Because ML estimates are generally biased and not normally distributed in small samples, t and F tests are not exact tests.
   b. Asymptotic tests are asymptotically distributed as χ^2
      i. Wald test (based on unrestricted estimate)
      ii. LM test (based on restricted estimate)
      iii. LR test $2\,[\ln L_U - \ln L_R]$ (based on both)
   c. In linear models, W > LR > LM in small samples