Asymptotics for Least Squares

James L. Powell
Department of Economics
University of California, Berkeley

Least Squares and Linear Predictors

Given the standard assumptions for the classical linear regression model – which include the strong assumptions of nonrandom regressors $\mathbf{X}$, with dependent variable $\mathbf{y}$ having linear expectation (in the regressors) and scalar covariance matrix – it might seem natural to replace the additional assumption of multinormality of $\mathbf{y}$ with weaker sufficient conditions that ensure that the classical least squares estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}} &\equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i\right]\end{aligned}$$

is approximately (asymptotically) normal. While this is certainly feasible, it is simpler to apply asymptotic theory under a different, and in some ways weaker, set of assumptions on the process that generates the data. Instead of imposing linearity of the mean of $\mathbf{y}$ in $\mathbf{X}$, etc., we instead assume just that the observations $\{\mathbf{z}_i \equiv (y_i, \mathbf{x}_i')'\}_{i=1}^{N}$ are i.i.d with bounded fourth moments, i.e.,

$$E\left[||\mathbf{z}_i||^4\right] < \infty.$$

In this setting, which treats the regressors $\mathbf{x}_i$ more symmetrically with the dependent variable y_i, the classical LS estimator $\hat{\boldsymbol{\beta}}$ does not estimate the coefficients for the mean of $\mathbf{y}$ as a function of $\mathbf{X}$, but rather the coefficients of the best linear predictor

$$\boldsymbol{\beta} \equiv \arg\min_{\mathbf{c}} E\left[(y_i - \mathbf{x}_i'\mathbf{c})^2\right].$$

Assuming, in addition, that the matrix

$$\mathbf{D} \equiv E\left[\mathbf{x}_i\mathbf{x}_i'\right]$$

is nonsingular – a population version of the usual assumption that $\mathbf{X}'\mathbf{X}$ is invertible – the parameter vector $\boldsymbol{\beta}$ is uniquely defined as

$$\boldsymbol{\beta} = \mathbf{D}^{-1}\boldsymbol{\delta},$$

where

$$\boldsymbol{\delta} \equiv E\left[\mathbf{x}_i y_i\right].$$

In contrast to the standard assumptions, the usual finite-sample results for the classical LS estimator $\hat{\boldsymbol{\beta}}$ do not apply. For instance, the law of iterated expectations implies that the expectation of $\hat{\boldsymbol{\beta}}$, if it exists, satisfies

$$E\left[\hat{\boldsymbol{\beta}}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left[\mathbf{y}|\mathbf{X}\right]\right],$$

but without the assumption of linearity of $E\left[\mathbf{y}|\mathbf{X}\right]$ in $\mathbf{X}$, in general

$$E\left[\hat{\boldsymbol{\beta}}\right] \neq \boldsymbol{\beta},$$

so classical LS is not generally unbiased for the best linear prediction coefficients $\boldsymbol{\beta}$. Nevertheless, $\hat{\boldsymbol{\beta}}$ is clearly a smooth function of sample averages,

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\delta}},$$

where

$$\begin{aligned}
\hat{\mathbf{D}} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i', \\
\hat{\boldsymbol{\delta}} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i.
\end{aligned}$$

Thus demonstration of the *consistency* of $\hat{\boldsymbol{\beta}}$ – i.e.,

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

– and its asymptotic normality are straightforward applications of the asymptotic theory discussed previously.

Consistency of Least Squares

Consistency of $\hat{\boldsymbol{\beta}}$ follows from a straightforward application of the Weak Law of Large Numbers and the continuity theorem. Each component of $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{D}}$ is a sample average of products of elements of the vector of regressors $\mathbf{x}_i$ with another component of $\mathbf{x}_i$ or with the dependent variable y_i; since fourth moments are assumed to exist for y_i and $\mathbf{x}_i$, the variances of these products are finite, and their means and variances are identical and covariances are zero because the data are assumed i.i.d. Thus

$$\hat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta}$$

and

$$\hat{\mathbf{D}} \xrightarrow{p} \mathbf{D}$$

by the WLLN. Furthermore,

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{D}^{-1}\boldsymbol{\delta} \\ &\equiv g(\mathbf{D}, \boldsymbol{\delta}) \end{aligned}$$

is a continuous function of $\mathbf{D}$ and $\boldsymbol{\delta}$ at all arguments with $|\mathbf{D}| \neq \mathbf{0}$. Thus, the continuity theorem implies that

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\delta}} \xrightarrow{p} \mathbf{D}^{-1}\boldsymbol{\delta} = \boldsymbol{\beta}.$$

Note that, if the stronger conditions $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and $E(||\hat{\boldsymbol{\beta}}||^2) < \infty$, were imposed, it could be possible to alternatively demonstrate consistency of $\hat{\boldsymbol{\beta}}$ by showing $\mathbf{V}(\hat{\boldsymbol{\beta}}|\mathbf{X}) \to \mathbf{0}$ as $N \to \infty$, which would imply quadratic mean convergence of any linear combination $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$.

Asymptotic Normality of LS

Now we can write the normalized difference $\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ between the LS estimator and its probability limit as

$$\begin{aligned} \sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= \sqrt{N}\left(\hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}\right) \\ &= \sqrt{N}\left(\hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\delta}} - \hat{\mathbf{D}}^{-1}\hat{\mathbf{D}}\boldsymbol{\beta}\right) \\ &= \hat{\mathbf{D}}^{-1}\sqrt{N}\left(\hat{\boldsymbol{\delta}} - \hat{\mathbf{D}}\boldsymbol{\beta}\right) \\ &= \hat{\mathbf{D}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mathbf{x}_i y_i - \mathbf{x}_i\mathbf{x}_i'\boldsymbol{\beta}\right) \\ &= \hat{\mathbf{D}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) \\ &\equiv \hat{\mathbf{D}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i, \end{aligned}$$

where $\varepsilon_i \equiv y_i - \mathbf{x}_i'\boldsymbol{\beta}$ has

$$\begin{aligned} E\left(\mathbf{x}_i\varepsilon_i\right) &= E\left(\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right) \\ &= \boldsymbol{\delta} - \mathbf{D}\boldsymbol{\beta} \\ &= \mathbf{0} \end{aligned}$$

by the definition of $\boldsymbol{\beta}$. Thus, application of the multivariate version of the Lindeberg-Levy CLT implies that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{x}_i \varepsilon_i \xrightarrow{d} N(\mathbf{0}, \mathbf{C}),$$

where

$$\begin{aligned} \mathbf{C} &\equiv \mathbf{V}(\mathbf{x}_i \varepsilon_i) \\ &= E\left[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'\right], \end{aligned}$$

which exists because of the assumed finite fourth moments of the observations.

Since

$$\hat{\mathbf{D}}^{-1} \xrightarrow{p} \mathbf{D}^{-1}$$

by consistency of $\hat{\mathbf{D}}$ and the continuity theorem, the Slutsky theorem implies that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathbf{D}^{-1} \cdot N(\mathbf{0}, \mathbf{C}) = N\left(\mathbf{0}, \mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}\right).$$

Consistent Estimation of the Asymptotic Covariance Matrix

In order to conduct large-sample inference on $\boldsymbol{\beta}$, a consistent estimator of the asymptotic covariance matrix $\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}$ needs to be constructed. (This matrix is sometimes called a "sandwich form," with $\mathbf{C}$ analogous the filling and $\mathbf{D}^{-1}$ to the bread.) Since $\hat{\mathbf{D}}$ has already been shown to be consistent for $\mathbf{D}$, only a consistent estimator of the middle matrix $\mathbf{C}$ is needed; such an estimator is

$$\begin{aligned} \hat{\mathbf{C}} &\equiv \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}\right)^2 \mathbf{x}_i \mathbf{x}_i' \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)^2 \mathbf{x}_i \mathbf{x}_i' + \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2 \mathbf{x}_i \mathbf{x}_i' + \frac{2}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right) \left(\mathbf{x}_i' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right) \mathbf{x}_i \mathbf{x}_i'. \end{aligned}$$

The first of these last three terms can be shown to converge in probability to $\mathbf{C}$ by application of another law of large numbers (which assumes only i.i.d. data and existence of first moments), and the last two terms can be shown to converge to zero using that LLN and consistency of $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

The covariance estimator $\hat{\mathbf{D}}^{-1}\hat{\mathbf{C}}\hat{\mathbf{D}}^{-1}$ was proposed or implied independently by a number of authors, and can be called the *Huber-Eicker-White heteroskedasticity-robust asymptotic covariance matrix estimator*, a daunting title often replaced by the simpler *robust covariance matrix estimator.*

Given this estimator and the results above, a large-sample test for a nonlinear hypothesis

$$H_0 : \mathbf{g}(\boldsymbol{\beta}) = \mathbf{0},$$

where $g(\boldsymbol{\beta})$ is a differentiable function with

$$\mathbf{G} \equiv \frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$$

assumed to be continuous and have full row rank. (An example would be $\mathbf{g}(\boldsymbol{\beta}) = \mathbf{G}\boldsymbol{\beta} - \boldsymbol{\gamma}_0$, with $\mathbf{G}$ fixed and of full row rank, which is typically called the "general linear hypothesis.")

Application of the so-called "delta method" implies that, under the null hypothesis,

$$\sqrt{N} g(\hat{\boldsymbol{\beta}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{G}'\right).$$

Since

$$\hat{\mathbf{G}} \equiv \frac{\partial \mathbf{g}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'} \xrightarrow{p} \mathbf{G}$$

by the continuity theorem, it follows from Slutsky's theorem and the continuous mapping theorem that the "Generalized Wald statistic"

$$\begin{aligned} GW_N &\equiv N\left(\mathbf{g}(\hat{\boldsymbol{\beta}})\right)' \left[\hat{\mathbf{G}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{C}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{G}}'\right]^{-1} \mathbf{g}(\hat{\boldsymbol{\beta}}) \\ &\xrightarrow{d} \chi^2_r \end{aligned}$$

under the null hypothesis, where $r = \dim\{\mathbf{g}(\hat{\boldsymbol{\beta}})\}$. (Wald proposed this statistic in the context of maximum likelihood (ML) estimation, but the approach applies to any asymptotically normal estimator for which a consistent estimator of its asymptotic covariance matrix is available.) The Generalized Wald test of the null hypothesis would reject when W_N exceeds the upper α critical value for a chi-squared random variable with r degrees of freedom.

A Special Case - I.I.D. Linear Regression Model

We can specialize the results above to the case where each dependent variable y_i is the sum of a linear function of $\mathbf{x}_i$ and an independent error term ε_i :

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

with ε_i assumed to be independent of $\mathbf{x}_i$ with

$$\begin{aligned} E(\varepsilon_i) &= 0, \\ Var(\varepsilon_i) &= \sigma^2. \end{aligned}$$

For this special case, it follows that

$$\begin{aligned}\mathbf{C} &= E\left[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'\right] \\ &= E\left[\varepsilon_i^2\right] E\left[\mathbf{x}_i \mathbf{x}_i'\right] \\ &= \sigma^2 \mathbf{D}\end{aligned}$$

by the assumed independence of ε_i and $\mathbf{D}$. Then $\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1} = \sigma^2\mathbf{D}^{-1}$ and

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \mathbf{D}^{-1}\right) = N\left(\mathbf{0}, \sigma^2 \left[p\lim \frac{1}{N}\mathbf{X}'\mathbf{X}\right]^{-1}\right).$$

Similarly to the demonstration of consistency of $\hat{\mathbf{C}}$, it is possible to show the consistency of the usual estimator s^2 of σ^2 under these conditions, so in this case we would approximate the distribution of the LS estimator $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, s^2(\mathbf{X}'\mathbf{X})^{-1}),$$

which is the large-sample version of the finite-sample results for the LS estimator when the dependent variables are multinormal.