Estimating the Regression Model by Least Squares

Henrique Veras

PIMES/UFPE

Today

We now examine the Least Squares as an **estimator** of the parameters of the linear regression model.

We start by analysing the question "*why should we use least squares*"?

We will compare the LS estimator to other candidates based on their **statistical properties**:

1. Unbiasedness
2. Efficiency
3. Consistency

Population orthogonality

Recall assumption A3: $E[\varepsilon_i|\mathbf{X}] = 0$

By iterated expectations, $E[\varepsilon] = E_x E[\varepsilon_i|\mathbf{X}] = E_x[0] = 0$.

Also, $cov(\mathbf{x}, \varepsilon) = cov[\mathbf{x}, E[\varepsilon_i|\mathbf{X}]] = cov(\mathbf{x}, 0) = 0$, so $\mathbf{x}$ and ε are uncorrelated.

From these results we can find that

$$E[\mathbf{X}\mathbf{y}] = E[\mathbf{X}'\mathbf{X}]\beta$$

Population orthogonality

Now recall the FOC of the LS problem: $\mathbf{X'y} = \mathbf{X'Xb}$. Dividing both sides by n and writing it as a summation:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i y_i} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i' x_i}\right)\mathbf{b}$$

Notice that this is the sample counterpart of the population condition $E[\mathbf{Xy}] = E[\mathbf{X'X}]\beta$.

Statistical Properties of the LS Estimator

An **estimator** is a strategy for using the sample data that are drawn from a population.

The **properties** of that estimator are descriptions of how it can be expected to behave when it is appied to a sample of data.

Unbiasedness

The least squares estimator is **unbiased** in every sample:

$$E[\mathbf{b}|\mathbf{X}] = \beta$$

Moreover,

$$E[\mathbf{b}] = E_x[E[\mathbf{b}|\mathbf{X}]] = E_x[\beta] = \beta$$

This is to say that the Least Squares estimator has expectation β.

Moreover, when we average this over the possible values of $\mathbf{X}$, the unconditional mean is also β.

Omitted Variable Bias (OVB)

Suppose the true population model is given by

$$\mathbf{y} = \mathbf{X}\beta + \gamma z + \varepsilon$$

If we estimate $\mathbf{y}$ on $\mathbf{X}$ only, without the *relevant* variable z, the estimator is

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \gamma z + \varepsilon) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\gamma z + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\end{aligned}$$

Omitted Variable Bias (OVB)

The expected value is given by

$$
\begin{aligned}
E[\mathbf{b}|\mathbf{X}, z] &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\gamma z \\
&= \beta + \mathbf{p}_{X.z}\gamma,
\end{aligned}
$$

where $\mathbf{p}_{X.z} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'z$. What does it represent? What happens if $\mathbf{X}$ and z are orthogonal?

Based on the FWL theorem and corollary 3.2.1, we can write

$$
E[b_k|\mathbf{X}, z] = \beta_k + \gamma \left(\frac{cov(z, x_k|\text{all other x's})}{var(x_k|\text{all other x's})} \right)
$$

An Example

Suppose we are interested in estimating the returns to education regression model below:

$$Income = \beta_0 + \beta_1 Educ + \beta_2 age + \beta_3 age^2 + \beta_4 Abil + \varepsilon$$

What is the sign of the bias if we estimate the model above without the $(unobserved)$ $Abil$?

An Example

The sign of the bias will depend on the signs of γ and $cov(z, x_k|\text{all other x's})$:

$$E[b_1|\mathbf{X}, z] = \beta_1 + \gamma \left(\frac{cov(Abil, Educ|age, age^2)}{var(Educ|age, age^2)} \right)$$

Thus, if $\gamma > 0$ and $cov(Abil, Educ|age, age^2) > 0$, b_1 will be biased upward:

$$E[b_1|\mathbf{X}, z] > \beta_1$$

Notice, however, that in some circumstances, the sign of the conditional covariance might not be obvious!

What happens if we include irrelevant variables instead?

Variance of the Least Squares Estimator

If Assumption A4 holds, the variance of the Least Squares estimator is given by

$$Var(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

If we wish to find a sample estimate of $Var(\mathbf{b}|\mathbf{X})$, we need to estimate the (unknown) population parameter σ^2.

Recall:

1. σ^2 is the variance of the error term: $\sigma^2 = E[\varepsilon_i^2|\mathbf{X}]$
2. e_i is the estimate of ε_i

Variance of the Least Squares Estimator

A natural estimator for σ^2 would then be $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$.

However, we would also need to estimate K parameters β, which would distort σ^2.

An **unbiased** estimator for σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

Like $\mathbf{b}$, s^2 is unbiased unconditionally because

$$E[s^2] = E_X[E[s^2|\mathbf{X}]] = E_X[\sigma^2] = \sigma^2$$

Variance of the Least Squares Estimator

The **standard error of the regression** is $s = \sqrt{s^2}$.

The variance of the Least Squares Estimator can thus be estimated by

$$V\hat{a}r(\mathbf{b}|X) = s^2(\mathbf{X'X})^{-1}$$

$V\hat{a}r(\mathbf{b}|X)$ is the sample estimate of the *sampling variance* of the LS estimator.

Notice that the k-th diagonal element of this matrix is $[s^2(\mathbf{X'X}_{kk})^{-1}]^{1/2}$, the standard error of the estimator b_k.

The Gauss-Markov Theorem

THEOREM 4.2 Gauss–Markov Theorem
In the linear regression model with given regressor matrix $\mathbf{X}$, (1) *the least squares estimator,* $\mathbf{b}$, *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$ and (2) *for any vector of constants* $\mathbf{w}$, *the minimum variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *is* $\mathbf{w}'\mathbf{b}$.

The Normality Assumption

We have not used assumption A6 until now. Recall:

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

If **Assumption A6** is satisfied: $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Thus, $\mathbf{b}|\mathbf{X} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

Asymptotic Properties of the Least Squares Estimator

The properties described above can be applied to any sample size (*finite sample* properties)

Issues:

1. The list of settings in which exact finite sample results can be obtained is extremely small.
2. The assumption of normality likewise narrows the range of the applications.

We now extend our analysis to large samples. We might be able to obtain more general approximate results.

Asymptotic Properties of the Least Squares Estimator

Unbiasedness has two important limitations:

1. It is rare for an econometric estimator to be unbiased (with the clear exception of the LS)
2. Does not imply that more data is better.

Throughout we'll rely on two crucial assumptions:

1. **A5a**: $(\mathbf{x}_i \varepsilon_i), i = 1, \cdots, n$, is a sequence of independent, identically distributed *observations*.
2. **A2'**: plim $\frac{(\mathbf{X}'\mathbf{X})}{n} = \mathbf{Q}$, a positive definite matrix.

Recall:

> **DEFINITION D.1 Convergence in Probability**
> *The random variable* x_n **converges in probability** *to a constant* c *if* $\lim_{n\to\infty} \text{Prob}(|x_n - c| > \varepsilon) = 0$ *for any positive* ε.

If x_n converges in probability to c, then, we say that plim $x_n = c$.

Asymptotic Assumptions about $\mathbf{X}$

At many points from here forward, we will make an assumption that the data are *well behaved* so that an estimator or statistic will converge to a result.

TABLE 4.2 Grenander Conditions for Well-Behaved Data

G1. For each column of $\mathbf{X}$, $\mathbf{x}_k$, if $d_{nk}^2 = \mathbf{x}_k'\mathbf{x}_k$, then $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. *Sums of squares will continue to grow as the sample size increases.*

G2. $\operatorname{Lim}_{n\to\infty} x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. No single observation will ever dominate $\mathbf{x}_k'\mathbf{x}_k$. *As $n \to \infty$, individual observations will become less important.*

G3. Let $\mathbf{C}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty}\mathbf{C}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample. *This rank condition will not be violated as the sample size increases.*

Consistency of the LS estimator of β

Recall the general definition of consistency of an estimator:

DEFINITION D.2 Consistent Estimator
An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if

$$\text{plim}\ \hat{\theta}_n = \theta. \qquad \textbf{(D-4)}$$

We will show that the LS estimator $\mathbf{b}$ is consistent, that is, plim $\mathbf{b} = \beta$.

We can write $\mathbf{b}$ as

$$\mathbf{b} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

Some useful results

Before proceeding to the consistency proof, recall some important results:

THEOREM D.1 Convergence in Quadratic Mean
If x_n has mean μ_n and variance σ_n^2 such that the ordinary limits of μ_n and σ_n^2 are c and 0, respectively, then x_n converges in mean square to c , and

$$\text{plim } x_n = c.$$

THEOREM D.4 Consistency of the Sample Mean
The mean of a random sample from any population with finite mean μ and finite variance σ^2 is a consistent estimator of μ.
Proof: $E[\bar{x}_n] = \mu$ *and* $\text{Var}[\bar{x}_n] = \sigma^2/n$. *Therefore, $\bar{x}_n$ converges in mean square to μ, or plim $\bar{x}_n = \mu$.*

Some useful results

THEOREM D.14 Rules for Probability Limits

If x_n *and* y_n *are random variables with* $\operatorname{plim} x_n = c$ *and* $\operatorname{plim} y_n = d$, *then*

$$\operatorname{plim}(x_n + y_n) = c + d, \quad \textbf{(sum rule)} \tag{D-7}$$

$$\operatorname{plim} x_n y_n = cd, \quad \textbf{(product rule)} \tag{D-8}$$

$$\operatorname{plim} x_n/y_n = c/d \quad \text{if} \quad d \neq 0. \quad \textbf{(ratio rule)} \tag{D-9}$$

If $\mathbf{W}_n$ *is a matrix whose elements are random variables and if* $\operatorname{plim} \mathbf{W}_n = \boldsymbol{\Omega}$, *then*

$$\operatorname{plim} \mathbf{W}_n^{-1} = \boldsymbol{\Omega}^{-1}. \quad \textbf{(matrix inverse rule)} \tag{D-10}$$

If $\mathbf{X}_n$ *and* $\mathbf{Y}_n$ *are random matrices with* $\operatorname{plim} \mathbf{X}_n = \mathbf{A}$ *and* $\operatorname{plim} \mathbf{Y}_n = \mathbf{B}$, *then*

$$\operatorname{plim} \mathbf{X}_n \mathbf{Y}_n = \mathbf{AB}. \quad \textbf{(matrix product rule)} \tag{D-11}$$

Consistency of the LS estimator of β

Now, the probability limit of $\mathbf{b}$ is

$$\text{plim } \mathbf{b} = \beta + \mathbf{Q}^{-1}\text{plim } \left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

To find plim $\mathbf{b}$, we need to check plim $\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$:

According to Theorem D.4, the sample mean is a consistent estimator of the population mean. therefore

$$\frac{\mathbf{X}'\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\varepsilon_i$$

converges in probability to its expected value, $E[\varepsilon\mathbf{X}]$, which is zero (why?).

Consistency of the LS estimator of β

It follows that plim $\left(\frac{\mathbf{X}'\varepsilon}{n}\right) = 0$

Thus, plim $\mathbf{b} = \beta$.

For any individual coefficient of the vector $\mathbf{b}$, we have

$$\lim_{n\to\infty} Prob[|b_k - \beta_k| > \delta] = 0$$

for any $\delta > 0$

The Estimator of the $Asy.Var[\mathbf{b}]$

To complete the derivation of the asymptotic properties of **b**, we will require an estimator of $Asy.Var[\mathbf{b}] = \sigma^2/n\mathbf{Q}^{-1}$, provided that **A2'** is satisfied.

It can be shown that s^2 is a **consistent** estimator of σ^2. (you should be able to show this!)

Thus, by the product rule of the probability limits, we have

$$\text{plim } s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2\mathbf{Q}^{-1}$$

The appropriate estimator of the asymptotic covariance matrix of **b** is

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

Asymptotic Normality of the Least Squares Estimator

Let us now derive the asymptotic distribution of the Least Squares estimator.

Notice that we do not require that **A6** is satisfied here. If it is, then the sampling distribution of **b** is *exact* normal for every sample, which also holds asymptotically.

We now rewrite **b** as

$$\mathbf{b} - \beta = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$$

A multivariate version of the following theorem can be useful:

THEOREM D.18 Lindeberg–Levy Central Limit Theorem (Univariate)
If $x_1, \ldots, x_n$ are a random sample from a probability distribution with finite mean μ and finite variance σ^2 and $\bar{x}_n = (1/n)\sum_{i=1}^{n} x_i$, then $\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[0, \sigma^2]$. A proof appears in Rao (1973, p. 127*).*

Asymptotic Normality of the Least Squares Estimator

Multiplying both sides by $\sqrt{n}$:

$$\sqrt{n}(\mathbf{b} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\varepsilon$$

As we know that plim $\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = \mathbf{Q}^{-1}$, the *limiting* distribution of $\sqrt{n}(\mathbf{b} - \beta)$ is the same as the limiting distribution of

$$\mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \mathbf{X}'\varepsilon$$

Asymptotic Normality of the Least Squares Estimator

We can show that $\frac{1}{\sqrt{n}}\mathbf{X'}\varepsilon$ can be written as

$$\sqrt{n}[\bar{\mathbf{w}} - E(\bar{\mathbf{w}})],$$

where $\mathbf{w}_i = \sum \mathbf{x}_i\varepsilon_i$ and $\bar{\mathbf{w}} = (1/n)\sum \mathbf{w}_i$.

The vector $\bar{\mathbf{w}}$ is the average of n $i.i.d.$ random vectors with mean 0 and variance $Var[\varepsilon_i\mathbf{x}_i] = \sigma^2 E[\mathbf{x_i x_i'}] = \sigma^2\mathbf{Q}$.

Asymptotic Normality of the Least Squares Estimator

We can show that the variance of $\sqrt{n}\bar{\mathbf{w}} = \sigma^2\mathbf{Q}$.

Thus, applying the Lindeberg–Levy central limit theorem, if $[\mathbf{x}_i\varepsilon_i]$, $i = 1, \cdots, n$, are independent vectors, each distributed with mean 0 and variance $\sigma^2\mathbf{Q} < \infty$, and **A2'** holds, then

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon \xrightarrow{d} N[0, \sigma^2\mathbf{Q}]$$

Asymptotic Normality of the Least Squares Estimator

From the previous result, we can find the following theorem:

THEOREM 4.3 Asymptotic Distribution of b with IID Observations
If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2 and x_{ik} is such that the Grenander conditions are met, then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right]. \qquad \textbf{(4-33)}$$

In practice, we need to estimate $\frac{1}{n}\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$ and σ^2 with $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$.

Asymptotic Efficiency

The Gauss-Markov Theorem establishes finite sample conditions under which the LS estimator is optimal.

The requirements that the estimator be *linear* and *unbiased* limit the theorem's generality, however.

In asymptotic theory, we are interested in expanding the scope of analysis to estimators that might be biased but consistent.

Moreover, we might also be interested in comparing non-linear estimators as well.

Asymptotic Efficiency

These cases extend beynd the reach of the Gauss-Marvok theorem.

Below we present an alternative criterion:

DEFINITION 4.1 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

Some comments on Asymptotic Efficiency

We can compare estimators based on their asymptotic variances.

The complication in comparing two consistent estimators is that both converge to the true parameter as the sample size increases.

Moreover, it usually happens that they converge at the same rate—that is, in both cases, the asymptotic variances of the two estimators are of the same order.

In such a situation, we can sometimes compare the asymptotic variances for the same n to resolve the ranking.

Asymptotic Distribution of a Function of $\mathbf{b}$: The Delta Method

Let $\mathbf{b}$ be a set of J continuous, linear or non-linear, and continuously differentiable functions of the LS estimators and let

$$\mathbf{C}(\mathbf{b}) = \frac{\partial f(\mathbf{b})}{\partial \mathbf{b}'}$$

a $J \times K$ matrix whose jth row is the vector of derivatives of the jth function with respect to $\mathbf{b}'$.

We'll use the Slutsky theorem:

THEOREM D.12 Slutsky Theorem
For a continuous function $g(x_n)$ that is not a function of n,

$$\operatorname{plim} g(x_n) = g(\operatorname{plim} x_n). \qquad \textbf{(D-6)}$$

Asymptotic Distribution of a Function of **b**: The Delta Method

According to the Slutsky Theorem,

$$\text{plim } f(\mathbf{b}) = f(\text{plim } \mathbf{b}) = f(\beta)$$

and

$$\text{plim } \mathbf{C}(\mathbf{b}) = \frac{\partial f(\beta)}{\partial \beta'} = \boldsymbol{\Gamma}$$

Asymptotic Distribution of a Function of $\mathbf{b}$: The Delta Method

Using a linear Taylor series approach, we expand this set of functions in the approximation

$$f(\mathbf{b}) = f(\beta) + \mathbf{\Gamma} \times (\mathbf{b} - \beta) + \text{higher order terms}$$

The higher order terms become negligible in large samples if plim $\mathbf{b} = \beta$.

Thus the asymptotic distribution of the LHS of the above expression is the same as the asymptotic distribution of the RHS.

Asymptotic Distribution of a Function of $\mathbf{b}$: The Delta Method

The mean of the asymptotic distribution is plim $f(\mathbf{b}) = f(\beta)$ and the asymptotic covariance matrix is

$$\mathbf{\Gamma}[Asy.Var(\mathbf{b} - \beta)]\mathbf{\Gamma}'$$

.

These results are displayed in the theorem below.

THEOREM 4.4 Asymptotic Distribution of a Function of b

If $\mathbf{f}(\mathbf{b})$ *is a set of continuous and continuously differentiable functions of* $\mathbf{b}$ *such that* $\mathbf{f}(\text{plim } \mathbf{b})$ exists and $\mathbf{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ *and if Theorem 4.4 holds, then*

$$\mathbf{f}(\mathbf{b}) \overset{a}{\sim} N\left[\mathbf{f}(\boldsymbol{\beta}), \mathbf{\Gamma}\left(\text{Asy.Var}[\mathbf{b}]\right)\mathbf{\Gamma}'\right]. \tag{4-43}$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}\{\text{Est. Asy.Var}[\mathbf{b}]\}\mathbf{C}'.$$

An Example: The US gasoline market

Suppose we are interested in estimating the short- and long-run impacts of changes in price and income on the demand for gasoline.

We can estimate the following model:

$$\begin{aligned} ln(G/Pop)_t = & \beta_1 + \beta_2 lnP_{G,t} + \beta_3 ln(Income/Pop)_t + \beta_4 lnP_{nc,t} \\ & + \beta_5 lnP_{uc,t} + \gamma ln(G/Pop)_{t-1} + \varepsilon_t \end{aligned}$$

In the short-run, price and income elasticities are β_2 and β_3, respectively.

In the long-run, equilibrium would require $ln(G/Pop)_t = ln(G/Pop)_{t-1}$.

Therefore, long-run elasticities are $\phi_2 = \frac{\beta_2}{1-\gamma}$ and $\phi_3 = \frac{\beta_3}{1-\gamma}$

How do we estimate these elasticities?

An Example: The US gasoline market

The LS estimates of the long-run elasticities are $f_2 = b_2/(1-c) = -0.411$ and $f_3 = b_3/(1-c) = 0.9705$.

The estimated results are shown in the table below.

TABLE 4.6 Regression Results for a Demand Equation

Sum of squared residuals:			0.0127352
Standard error of the regression:			0.0168227
R^2 based on 51 observations			0.9951081
Variable	***Coefficient***	***Standard Error***	***t Ratio***
Constant	−3.123195	0.99583	−3.136
$\ln P_G$	−0.069532	0.01473	−4.720
ln *Income / Pop*	0.164047	0.05503	2.981
$\ln P_{nc}$	−0.178395	0.05517	−3.233
$\ln P_{uc}$	0.127009	0.03577	3.551
last period ln *G / Pop*	0.830971	0.04576	18.158

An Example: The US gasoline market

The next step is to estimate the standard errors of ϕ_2 and ϕ_3.

Recall the estimated asymptotic variance matrix obtained in theorem 4.4:

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

This yields

$$Es.Asy.Va[b_2/(1-c)] = 0.0231$$

$$Es.Asy.Va[b_3/(1-c)] = 0.0264$$

The (asymptotic) standard errors are the square roots 0.1522 and 0.1623, respectively.

Interval Estimation

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate.

A general approach would be

$$\hat{\theta} \pm \text{sampling variability}$$

Consider two (uninformative) extreme cases:

1. 100% of confidence that the true population parameter lies in the range $\hat{\theta} \pm \infty$
2. 0% of confidence that the true population parameter lies in the range $\hat{\theta} \pm 0$

The objective is to choose a value α such that we can attach the confidence $100(1 - \alpha)\%$ to the interval.

Forming a Confidence Interval for a Coefficient

If $\varepsilon|\mathbf{X} \sim N(0, \sigma^2\mathbf{I})$, then any particular element of $\mathbf{b}$ is distributed as

$$\mathbf{b}_k|\mathbf{X} \sim N(\beta_k, \sigma^2 S^{kk})$$

where S^{kk} denotes the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

By standardizing the variable, we find

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}}$$

has a standard normal distribution.

Using $\alpha = 0.05$, we know that $Prob[-1.96 \leq z_k \leq 1.96] = 0.95$.

Forming a Confidence Interval for a Coefficient

By simple manipulation, we have

$$Prob[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2 S^{kk}}] = 0.95$$

But recall that σ^2 is unknown. Thus, we use s^2 as its estimator. Then, the ratio

$$t_k = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}}$$

has a t distribution with $n - K$ degrees of freedom (why?).

Forming a Confidence Interval for a Coefficient

If the disturbances do not follow a normal distribution, then the theory for the t distribution does not apply.

However, we can use the large sample results to find the *limiting* distribution of the statistic

$$z_k = \frac{\sqrt{n}(b_k - \beta_k)}{\sqrt{\sigma^2 Q^{kk}}}$$

is standard normal, where $\mathbf{Q} = \text{plim } (\mathbf{X}'\mathbf{X}/n)$ and Q^{kk} is the kth diagonal element of $\mathbf{Q}^{-1}$.

We can use s^2 as a consistent estimator for σ^2 and estimate $\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X}/n)^{-1}$.

Forming a Confidence Interval for a Coefficient

This gives us precisely the t statistic for the exact distribution case.

To compute the confidence interval, however, we use the standard normal distribution table rather than the t distribution.

In practice, however, if the degrees of freedom are moderately large, say greater than 100, the t distribution converges to the standard normal one.

The confidence interval would then be given by

$$Prob[b_k - 1.96\sqrt{Est.Asy.Var(b_k)} \leq \beta_k \leq b_k + 1.96\sqrt{Est.Asy.Var(b_k)}] = 1 - \alpha$$

Confidence Interval for a Linear Combination of Coefficients: The Oaxaca's Decomposition

Oaxaca (1973) and Blinder (1973) provide an application on how to form a confidence interval for a linear function of the parameters.

Let $\mathbf{w}$ denote a $K \times 1$ vector of known constants.

Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is asymptotically normally distributed with mean $\gamma = \mathbf{w}'\beta$ and variance

$$\sigma_c^2 = \mathbf{w}'[Asy.Var(\mathbf{b})]\mathbf{w}$$

which we estimate with $s_c^2 = \mathbf{w}'[Est.Asy.Var(\mathbf{b})]\mathbf{w}$.

The confidence interval for γ is thus

$$Prob[c - z_{(1-\alpha/2)}s_c \leq \gamma \leq c - z_{(1-\alpha/2)}s_c] = 1 - \alpha$$

The Oaxaca-Blinder Decomposition

Consider Oaxaca's (1973) application, in the context of labor supply.

The underlying regression model for men and women, separately, are

$$\ln wage_{m,i} = \mathbf{x}_{m,i}\beta_m + \varepsilon_{m,i}$$

$$\ln wage_{f,i} = \mathbf{x}_{f,i}\beta_f + \varepsilon_{f,i}$$

where $\mathbf{x}_i$ includes sociodemographic variables, such as age, education, and experience.

The purpose is to compute both equations by decomposing the estimated difference in wages in two components:

1. Differences in the levels of each observable variable of the model;
2. Differences in the (unexplained) "effects".

The Oaxaca-Blinder Decomposition

From the population regression equations, we have

$$\begin{aligned}
E[\ln wage_{m,i}|\mathbf{x}_{m,i}] - E[\ln wage_{f,i}|\mathbf{x}_{f,i}] &= \mathbf{x}_{m,i}\beta_m - \mathbf{x}_{f,i}\beta_f \\
&= \mathbf{x}_{m,i}\beta_m - \mathbf{x}_{m,i}\beta_f + \mathbf{x}_{m,i}\beta_f - \mathbf{x}_{f,i}\beta_f \\
&= \mathbf{x}_{m,i}(\beta_m - \beta_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,i})\beta_f
\end{aligned}$$

Assuming labor markets respond to differences in human capital properly, the second term captures differences in human capital *levels* across both groups.

The first term shows the differential in log wages that is attributed to differences unexplainable by human capital.

The Oaxaca-Blinder Decomposition

We are interested in forming a confidence interval for the first term. For this, we assume that both $\mathbf{x}_m$ and $\mathbf{x}_f$ are known and we have two independent set of observations.

Evaluating the model at the mean of the regression vectors, $\bar{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_f$, we find that $\mathbf{b}_m$ and $\mathbf{b}_f$ are independent with means β_m and β_f and estimated asymptotic covariance matrices $Est.Asy.Var[\mathbf{b}_m]$ and $Est.Asy.Var[\mathbf{b}_f]$.

We are, then, forming a confidence interval for $\bar{\mathbf{x}}_m\mathbf{d}$, where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

$$Est.Asy.Var[\mathbf{d}] = Est.Asy.Var[\mathbf{b}_m] + Est.Asy.Var[\mathbf{b}_f]$$

.

The CI will be constructed as before.

Prediction $\times$ Forecasting

Prediction: Using the regression model to compute fitted (predicted) values of the dependent variables (used in cross sections, panel data, time series)

Forecasting Same exercise, but explicitly giving role to time and the purpose of the model building is to forecast future outcomes (used in time series only).

Prediction Intervals

Suppose we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$.

The actual value would be

$$y^0 = \mathbf{x}^{0'}\beta + \varepsilon^0$$

The prediction error is

$$e^0 = \hat{y}^0 - y^0 = (\mathbf{b} - \beta)'\mathbf{x}^0 - \varepsilon^0$$

Prediction Intervals

The prediction variance is

$$Var[e^0|\mathbf{X}, \mathbf{x^0}] = Var[(\mathbf{b} - \beta) - \varepsilon^0|\mathbf{X}, \mathbf{x^0}] = \sigma^2 + \mathbf{x^0}'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x^0}$$

The prediction variance can be estimated by using s^2 in place of σ^2.

A confidence (prediction) interval for y^0 would then be formed using

$$y^0 \pm t_{(1-\alpha/2),[n-K]} se(e^0)$$

Table of Contents