Econometrics 1

Linear Regression and Ordinary Least Squares

Linear Regression and Ordinary Least Squares

Outline:

(1) Linear Regression,

(2) Ordinary Least Squares: Estimation,

(3) Ordinary Least Squares: Inference,

(4) Ordinary Least Squares: Extensions.

(1) Linear Regression

EconTutor.com

What Is Econometrics?

An econometric analysis typically (but not always) begins with a set of theoretical *economic* propositions that can be tested empirically.

Economic theory often helps to formalize the hypothesis of interest, e.g., returns to crime, returns to job training programs, etc.

Economic data give the possibility of *identifying* objects of interest, e.g., correlations, causal effects, channels of causation, etc.

Econometrics helps to navigate this misty space, but first ...

Linear Regression

Linear regression remains one of the most widely used econometric techniques in applied empirical research in economics.

Linear models are analytically and computationally tractable; they are well studied and their properties are well understood.

The linear regression model is nearly always the point of departure for further econometric analysis, and it will be ours as well.

Simple Linear Regression

The *simple linear regression model* posits a relationship between a *dependent variable* and one *independent variable*, *i.e.*,

$$\begin{aligned} y &= f(x) + u \\ &= \beta_0 + \beta_1 x + u, \end{aligned}$$

where y is the *dependent/explained/response/predicted/outcome variable* or *regressand*, and where x is the *independent/explanatory/control/predictor variable*, *regressor*, or *covariate*.

The underlying theory informs the specification of 'dependent' and 'independent' variables in this *population regression* model.

A *disturbance* or *error* term u is included additively to account for the combined effects of influences on y other than x; note, u is unobserved.

Other influences include, among many other things, omitted variables, nonlinear effects of included variables, and measurement error.

Simple Linear Regression

If $y' = \beta_0 + \beta_1 x' + u'$, and we define $\Delta y = y' - y$, $\Delta x = x' - x$, and $\Delta u = u' - u$, then $\Delta y = \beta_1 \Delta x + \Delta u$; if $\Delta u = 0$, then $\beta_1 = \Delta y / \Delta x$.

In essence, $\Delta u = 0$ embodies the *ceteris paribus* assumption, i.e., we are holding all other (unobserved) factors fixed, a strong assumption.

A weaker assumption is that $\mathrm{E}(u) = 0$, i.e., the combined effects of influences on y other than x have a mean of zero; this is innocuous so long as the regression has an intercept.

The *critical* identifying assumption is that $\mathrm{E}(u|x) = \mathrm{E}(u)$, i.e., u is said to be *mean independent* of x; therefore, $\mathrm{E}(u|x) = \mathrm{E}(u) = 0$.

The *population regression function* (PRF) is then given by

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x.$$

Multiple Linear Regression

The mean independence assumption is critical for identification, but it often remains problematic; we would like to further *control for* other factors which might systematically affect y.

The *multiple linear regression model* posits a relationship between a dependent variable and two or more independent variables, i.e.,

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_K) + u \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u, \end{aligned}$$

where y is the dependent variable, $(x_1, x_2, \ldots, x_K)$ are independent variables, and u is the error term; $(\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$ are parameters.

Mean independence now states that $\mathrm{E}(u|x_1, x_2, \ldots, x_K) = \mathrm{E}(u)$; with an intercept, $\mathrm{E}(u) = 0$, and so $\mathrm{E}(u|x_1, x_2, \ldots, x_K) = \mathrm{E}(u) = 0$.

Multiple Linear Regression

Given the sample $\{y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^{N}$, each observation satisfies

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i,$$

where the observed y_i is composed of two types of *heterogeneity*:

(1) *Observed/systematic*, i.e., $\mathrm{E}(y_i | x_{1i}, x_{2i}, \ldots, x_{Ki})$,

(2) *Unobserved/idiosyncratic*, i.e., u_i.

Ultimately, our empirical objective is going to be to *estimate* the set of population parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$.

Consumption and Income

Let C denote expenditure on consumption and X denote income.

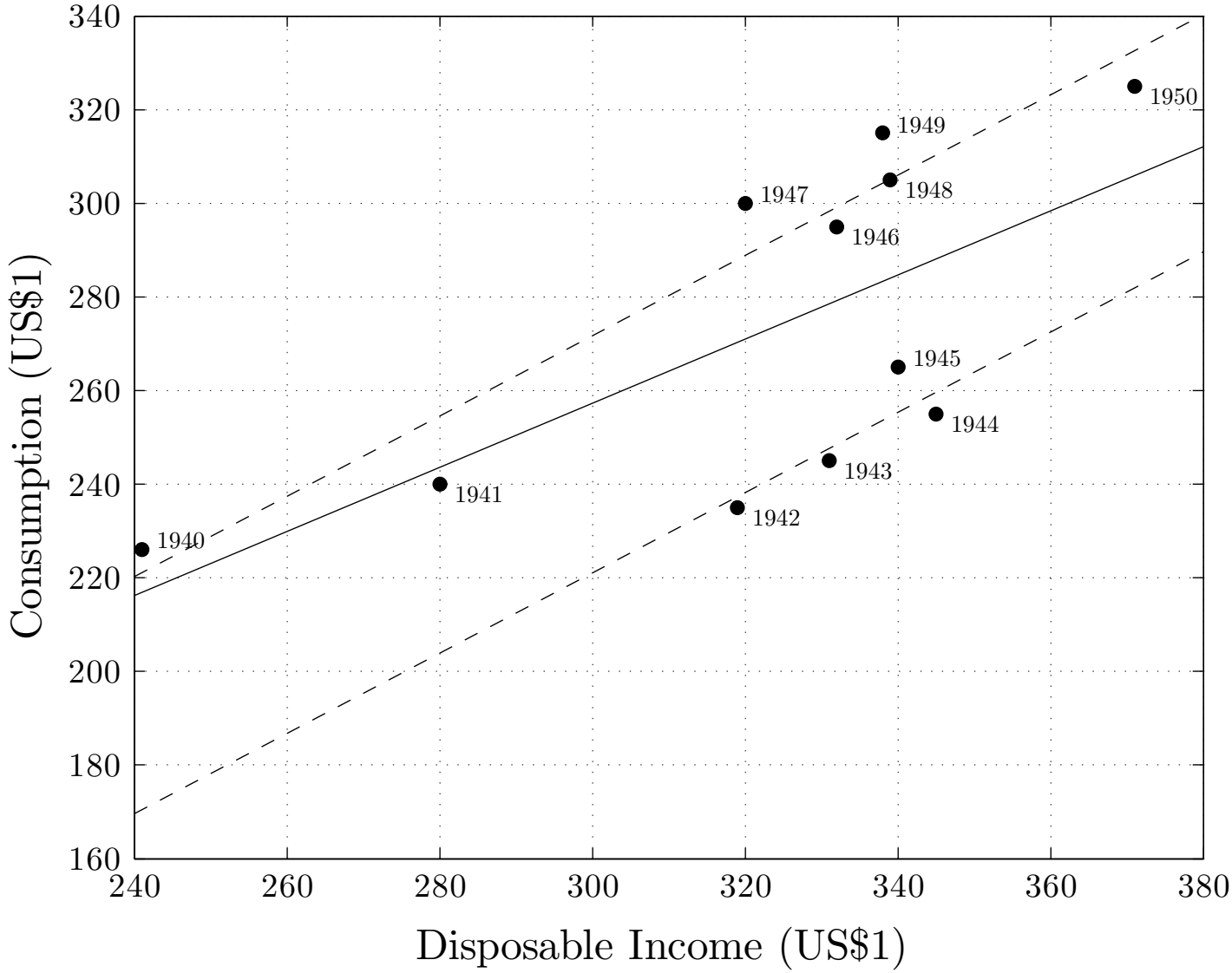
Let $C = f(X)$, where f denotes the propensity to consume.

H_0: $0 < dC/dX < 1$, and $d(C/X)/dX < 0$.

H_0^*: if $C = \alpha + \beta X$, then $\alpha > 0$ and $0 < \beta < 1$.

H_0^{**}: if $C = \alpha + \beta X + u$, then $\alpha + u > 0$ and $0 < \beta < 1$.

Consumption and Income, 1940–1950



Consumption and Income

Let $d_{waryears} = 1$ in 1942–1945, and 0 otherwise, i.e., the variable $d_{waryears}$ is said to be a *dummy variable*.

H_0^{***}: if $C = \alpha + \beta X + \delta_w d_{waryears} + u$, then $\alpha + \delta_w + u > 0$, $0 < \beta < 1$, and $\delta_w < 0$.

Multiple regression analysis allows us to identify the independent effects of X and $d_{waryears}$ on C.

Multiple Linear Regression

The linear specification is not as restrictive as it might seem.

Variables can enter as nonlinear transformations of other variables, e.g., quadratics, cubics, roots, logs, etc.

E.g., if $y = \log(earnings)$, $x_1 = education$, $x_2 = age$, and $x_3 = age^2$, we could specify any of the following relationships:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + u \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u. \end{aligned}$$

Linear models typically include an unrestricted intercept term.

The linear regression model is restrictive because it is linear in the parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$ and the error term u.

Multiple Linear Regression (Matrix Form)

Given the sample $\{y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$, each observation satisfies

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i.$$

Denote $\mathbf{x}_i \in \mathbb{R}^{K+1}$ and $\boldsymbol{\beta} \in \mathbb{R}^{K+1}$ according to

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}.$$

We can now write the model as

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

Multiple Linear Regression (Matrix Form)

Given the more compactly written model

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

we can stack observations for a sample of size N, i.e.,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{K1} \\ 1 & x_{12} & \cdots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{KN} \end{pmatrix}, \text{ and } \ \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}.$$

Notice that $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{X} \in \mathbb{R}^{N\times(K+1)}$, $\boldsymbol{\beta} \in \mathbb{R}^{(K+1)}$, and $\mathbf{u} \in \mathbb{R}^N$, so that the model can be written even more compactly as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Classical Linear Regression

(A0) Random sampling.

(A1) Linearity in parameters.

(A2) There are at least as many observations as parameters, and there is no linear dependence among the explanatory variables.

(A3) Exogeneity: $\mathrm{E}(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = 0$, i.e., the error terms are independent of and uncorrelated with the regressors.

(A4) Homoskedasticity: $\mathrm{Var}(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma^2$, i.e., the error terms have identical conditional variances.

(A5) Normality: $u_i|x_{1i}, x_{2i}, \ldots, x_{Ki} \sim \mathrm{Normal}\,(0, \sigma^2)$, i.e., the error terms are normally distributed with mean 0 and variance σ^2.

Classical Linear Regression (Matrix Form)

(A1) Linearity in parameters.

(A2) There are at least as many observations as parameters, and there is no linear dependence among the explanatory variables.

(A3) Exogeneity: $\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$, i.e., the error terms are independent of and uncorrelated with the regressors.

(A4) Homoskedasticity: $\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}$, i.e., the error terms have identical variances and are uncorrelated.

(A5) Normality: $\mathbf{u}|\mathbf{X} \sim \mathrm{Normal}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, i.e., the error terms are normally distributed with mean 0 and variance σ^2.

(2) Ordinary Least Squares: Estimation

EconTutor.com

Linear Models

Linear models are linear in their parameters.

Given the sample $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^{N}$, each observations satisfies

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i,$$

where y_i is the dependent variable, $(x_{1i}, \ldots, x_{Ki})$ are the independent variables, $(\beta_0, \beta_1, \ldots, \beta_K)$ are parameters, and u_i is the error term.

There are a multitude of approaches to estimating the population parameters in linear models; the method of *ordinary least squares* (OLS) is a natural and well established benchmark.

More importantly, other 'more sophisticated' estimators are often mild modifications or extensions to the least squares approach.

Ordinary Least Squares

The OLS estimator chooses a value of $(\beta_0, \beta_1, \ldots, \beta_K)$ to minimize the *sum of squared deviations* (SSD), i.e., which minimizes the expression

$$\sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_K x_{Ki})^2.$$

Because of the squared loss function, the OLS estimator chooses the parameters to avoid large deviations and to tolerate small ones.

The OLS estimator has many desirable properties in linear models.

Since the OLS estimator is sensitive to outliers, robustness checks are often recommended in empirical practice.

Deriving OLS

Implementing the least squares routine requires $K+1$ derivatives

$$\frac{\partial \text{SSD}}{\partial \beta_0} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_K x_{Ki}),$$

$$\frac{\partial \text{SSD}}{\partial \beta_k} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_K x_{Ki})x_{ki}.$$

Since $u_i = y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_K x_{Ki}$, we can write

$$\frac{\partial \text{SSD}}{\partial \beta_0} = -2\sum_{i=1}^{N} u_i,$$

$$\frac{\partial \text{SSD}}{\partial \beta_k} = -2\sum_{i=1}^{N} x_{ki}u_i.$$

Deriving OLS

The OLS estimator then satisfies the $K+1$ first-order conditions

$$-2\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_K x_{Ki}) = 0,$$

$$-2\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_K x_{Ki})x_{ki} = 0.$$

Since $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_K x_{Ki}$, we can write

$$-2\sum_{i=1}^{N}\hat{u}_i = 0,$$

$$-2\sum_{i=1}^{N}x_{ki}\hat{u}_i = 0.$$

Deriving OLS

Given (A2), which states that there are at least as many observations as parameters, and that no regressors are perfectly collinear,

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)$$

exists and achieves a *unique* minimum.

Notice that other than (A2), we have only assumed linearity (A1).

The remaining assumptions (A3)–(A5) will allow us to derive other (distributional) properties of the OLS estimator; we will come to this.

Deriving OLS (Matrix Form)

The OLS estimator chooses the value of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$ to minimize the *sum of squared deviations* (SSD), i.e.,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \\ &= \arg\min_{\boldsymbol{\beta}} \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

Since $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta}$$

$$\implies 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'y = \mathbf{0} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Algebraic Properties of OLS

The *fitted/predicted values* are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_K x_{Ki},$$

and the *residuals* are given by

$$\begin{aligned}\hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_K x_{Ki}.\end{aligned}$$

The following algebraic properties of OLS must hold:

(1) $\sum_{i=1}^{N} \hat{u}_i = 0$, i.e., the sum of the residuals is zero,

(2) $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_K \bar{x}_K$, i.e., the regression passes through the sample means of the observed values,

(3) $\bar{\hat{y}} = \bar{y}$, i.e., the means of the fitted and observed values are equal.

Algebraic Properties of OLS

Define the total sum of squares (TSS), the explained sum of squares (ESS), and the residual sum of squares (RSS) as follows:

$$\text{TSS} = \sum_{i=1}^{N} (y_i - \bar{y})^2,$$

$$\text{ESS} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2,$$

$$\text{RSS} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \hat{u}_i^2.$$

Notice that TSS = ESS + RSS, i.e., the total sample variation in the dependent variable can be additively decomposed into explained and residual components; $R^2 = \text{ESS}/\text{TSS}$ is a measure of *fit*.

Simple Linear Regression

As a special case, consider the simple linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i,$$

where the OLS estimator for β_0 is given according to

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1,$$

and for β_1 according to

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)^2}.$$

The fitted relationship passes through sample means (generalizable), and $\hat{\beta}_1$ is the (scaled) sample correlation (non-generalizable).

Partitioned Regression

Consider the linear regression model given by

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,$$

with fitted/predicted values and residuals given by

$$\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}.$$

$$\hat{u}_i = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}.$$

Now consider the linear regression models given by

$$y_i = \gamma x_{1i} + v_i,$$

$$x_{2i} = \delta x_{1i} + w_i.$$

The OLS estimators for γ and δ are given by

$$\hat{\gamma} = \frac{\sum_{i=1}^N x_{1i} y_i}{\sum_{i=1}^N x_{1i}^2}, \quad \hat{\delta} = \frac{\sum_{i=1}^N x_{1i} x_{2i}}{\sum_{i=1}^N x_{1i}^2}.$$

Partitioned Regression

We can write the OLS estimator $\hat{\gamma}$ as follows:

$$\begin{aligned}
\hat{\gamma} &= \frac{\sum_{i=1}^{N} x_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2} \\
&= \frac{\sum_{i=1}^{N} x_{1i} (\hat{y}_i + \hat{u}_i)}{\sum_{i=1}^{N} x_{1i}^2} \\
&= \frac{\sum_{i=1}^{N} x_{1i} (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i)}{\sum_{i=1}^{N} x_{1i}^2} \\
&= \hat{\beta}_1 + \hat{\delta} \hat{\beta}_2.
\end{aligned}$$

Thus, $\hat{\gamma} \neq \hat{\beta}_1$ unless $\hat{\beta}_2 = 0$, $\hat{\delta} = 0$, or both; in general, the estimated coefficient on x_1 differs between multiple and simple regressions.

Partitioned Regression

Finally, consider the linear regression model given by

$$\hat{v}_i = \eta \hat{w}_i + \epsilon_i,$$

where the residuals are given by

$$\hat{v}_i = y_i - \hat{\gamma} x_{1i},$$

$$\hat{w}_i = x_{2i} - \hat{\delta} x_{1i},$$

Therefore, we have that

$$\hat{\eta} = \frac{\sum_{i=1}^{N} \hat{v}_i \hat{w}_i}{\sum_{i=1}^{N} \hat{w}_i^2} = \hat{\beta}_2.$$

In words, we *partial out*, or *control for*, the effect of x_1 on both y and x_2 before considering the relationship between y and x_2.

OLS estimates are informative about *partial correlations* in the data.

(3) Ordinary Least Squares: Inference

Classical Linear Regression

The *classical linear regression* (CLR) model requires a set of strong assumptions, but yields convenient OLS properties in finite samples.

However, the OLS estimator will survive in large samples under much weaker assumptions, i.e., we can appeal to the Central Limit Theorem.

Our point of departure is to consider the most restrictive case, and then to relax unnecessarily restrictive assumptions.

CLR with Fixed Regressors

Given the model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i$, we let y_i and u_i be stochastic, and $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ be non-stochastic.

The regressors are said to be *fixed* in repeated samples.

This assumption might be acceptable in an experimental setting, with control over the controls, but rarely in an observational one.

However, the CLR model with fixed regressors is illustrative.

CLR with Fixed Regressors

More formally, given the (A0) random sample $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$, consider the (A1) linear model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i,$$

where (A3′) $\mathrm{E}(u_i) = 0$, (A4′) $\mathrm{Var}(u_i) = \sigma^2$, and $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ are non-stochastic and satisfy the restrictions in (A2).

Equivalently, we can say that given $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$,

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki},$$

$$\mathrm{Var}(y_i) = \sigma^2,$$

and $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ are non-stochastic and satisfy (A2).

Under (A0)–(A3′), we can show that $\mathrm{E}(\hat{\beta}_k) = \beta_k$, $k = 0, \ldots, K$, i.e., that the OLS estimator is *unbiased*.

CLR with Fixed Regressors

Under (A3′) and (A4′), $\text{Var}(u_i) = \text{E}(u_i^2) - (\text{E}(u_i))^2 = \text{E}(u_i^2) = \sigma^2$.

Under (A0)–(A4′), we can also derive $\text{Var}(\hat{\beta}_k)$, $k = 0, \ldots, K$, which in general depends upon σ^2 and the non-stochastic regressors.

Furthermore, $\text{Var}(\hat{\beta}_k) \leqslant \text{Var}(\tilde{\beta}_k)$, $k = 0, \ldots, K$, for any other linear unbiased estimator $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_K)$, i.e., $\hat{\boldsymbol{\beta}}$ is *efficient*.

We often say that $\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* (BLUE) in the classical linear regression model with fixed regressors.

The *Gauss-Markov Theorem* proves that the OLS estimator is BLUE, i.e., that $\hat{\boldsymbol{\beta}}$ has minimum variance.

CLR with Stochastic Regressors

Given the (A0) random sample $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$, we once again consider the (A1) linear model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i,$$

where $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ continue to satisfy the requirements in (A2) but are now *stochastic*, and where (A3) $\mathrm{E}(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = 0$ and (A4) $\mathrm{Var}(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma^2$.

Equivalently, we can say that given $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$,

$$\mathrm{E}(y_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki},$$

$$\mathrm{Var}(y_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma^2,$$

and $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ are stochastic and satisfy (A2).

CLR with Stochastic Regressors

Under (A0)–(A3), we can again show that $\mathrm{E}(\hat{\beta}_k) = \beta_k$, $k = 0, \ldots, K$, i.e., that the OLS estimator is *unbiased*.

Under (A0)–(A4), we can also derive $\mathrm{Var}(\hat{\beta}_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N)$, $k = 0, \ldots, K$, which depends upon σ^2 and the stochastic regressors.

Furthermore, $\hat{\boldsymbol{\beta}}$ is *efficient* conditional on the regressors.

The *Gauss-Markov Theorem* proves that $\hat{\boldsymbol{\beta}}$ is BLUE, i.e., it has the minimum conditional variance among unbiased linear estimators, in the classical linear regression model with stochastic regressors.

Point and Interval Estimation

We have already shown that ordinary least squares (OLS) estimator is *unbiased* and *efficient* in a classical linear regression framework.

Recall that the point estimates produced by an estimator vary across samples; we therefore want to know something about *precision*.

In order to do this, we can estimate *confidence intervals* around the true parameters; we can also perform a range of *hypothesis tests*.

Recall that we have already derived the (conditional) expectation and variance of the OLS estimator $\hat{\boldsymbol{\beta}}$; now we need to know something about the (conditional) *distribution* of $\hat{\boldsymbol{\beta}}$.

Finite Sample Inference

To derive *finite sample* distributional properties of $\hat{\boldsymbol{\beta}}$, which are *exact*, we assume *conditional normality* (A5).

Given the (A0) random sample $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$, we once again consider the (A1) linear model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i,$$

where $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ are stochastic and satisfy the requirements in (A2), and where (A5) $u_i | x_{1i}, x_{2i}, \ldots, x_{Ki} \sim \text{Normal}\,(0, \sigma^2)$.

Equivalently, we can say that given $\{(y_i, (x_{1i}, \ldots, x_{Ki})\}_{i=1}^N$,

$$y_i | x_{1i}, x_{2i}, \ldots, x_{Ki} \sim \text{Normal}\,(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki}, \sigma^2).$$

Note that the stronger conditional normality assumptions *imply* the weaker conditional mean and homoskedasticity assumptions.

Finite Sample Inference

Under (A0)–(A2) and (A5), we can once again show that $\mathrm{E}(\hat{\beta}_k) = \beta_k$, i.e., that the OLS estimator is *unbiased*.

Under (A0)–(A2) and (A5), we know $\mathrm{Var}(\hat{\beta}_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N)$, which in general depends upon σ^2 and the stochastic regressors.

Let $v_k = \mathrm{Var}(\hat{\beta}_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N)$; given properties particular to the multivariate normal distribution, we have that

$$\hat{\beta}_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N \sim \text{Normal}\left(\beta_k, v_k\right).$$

Finite Sample Inference

For the linear function of $\hat{\beta}_k$ given by

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_k}},$$

we can show that $z_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N \sim \text{Normal}\,(0, 1)$.

Since the distribution of $z_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ is the same for any realization of $\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$, we also have $z_k \sim \text{Normal}\,(0, 1)$.

Realizations of $\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ determine realizations of $\hat{\beta}_k$, v_k, and z_k; furthermore, the distribution of $\hat{\beta}_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ also depends upon realizations of $\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$.

Critically, the distribution of $z_k | \{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ does *not*.

Confidence Intervals

Assuming that σ^2 is *known*, the unconditional distributional result for the test statistic z_k allows us to perform exact finite sample inference.

For $z_k \sim \text{Normal}\,(0, 1)$, we have that

$$\text{P}(-1.96 < z_k < 1.96) = 0.95,$$

which implies that

$$\text{P}(\hat{\beta}_k - 1.96\sqrt{v_k} < \beta_k < \hat{\beta}_k + 1.96\sqrt{v_k}) = 0.95.$$

The 95% *confidence interval* (CI) for the true parameter β_k is written as $(\hat{\beta}_k - 1.96\sqrt{v_k},\ \hat{\beta}_k + 1.96\sqrt{v_k})$ or $\hat{\beta}_k \pm 1.96\sqrt{v_k}$.

Hypothesis Testing

To test $H_0 : \beta_k = \beta_k^0$, we construct $z_k = (\hat{\beta}_k - \beta_k^0)/\sqrt{v_k}$, which has a standard normal distribution under the null hypothesis H_0.

If z_k is an unlikely standard normal draw, i.e., if it falls outside of the interval $(-1.96, 1.96)$, then it is unlikely that H_0 is correct.

We either *reject* (z_k outside $(-1.96, 1.96)$) or *fail to reject* (z_k within $(-1.96, 1.96)$) the null hypothesis H_0 at the 0.05 significance level.

Equivalently, if β_k^0 falls outside of the 95% CI, then we reject the null hypothesis H_0 at the 0.05 significance level.

Standard Errors

Since σ^2 is typically *unknown*, we estimate it according to

$$\hat{\sigma}^2 = \frac{1}{N-K-1}\sum_{i=1}^{N}\hat{u}_i^2,$$

which then gives $\hat{v}_k = \widehat{\text{Var}}(\hat{\beta}_k|\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N)$.

We can now construct the standardized test statistic

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_k}} = \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)},$$

where $\text{se}(\hat{\beta}_k) = \sqrt{\hat{v}_k}$ is the *standard error* of the estimate $\hat{\beta}_k$.

Student t-Distribution

Since we have replaced $\sqrt{v_k}$ with $\text{se}(\hat{\beta}_k)$, the statistic t_k has different distributional properties than the statistic z_k.

We can show that $t_k|\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N \sim \text{t}(N-K)$, i.e., that $t_k|\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ is *t-distributed* ($N-K$ degrees of freedom).

Since the distribution of $t_k|\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$ is the same for any realization of $\{(x_{1i}, x_{2i}, \ldots, x_{Ki})\}_{i=1}^N$, we have that $t_k \sim \text{t}(N-K)$.

Confidence Intervals

This unconditional distributional result for the standardized test statistic t_k allows us to perform exact finite sample inference.

For $t_k \sim \mathrm{t}(N-K)$, we have that

$$\mathrm{P}(-c_{0.025}(N-K) < t_k < c_{0.025}(N-K)) = 0.95,$$

which implies that

$$\mathrm{P}(\hat{\beta}_k - c_{0.025}(N-K)\,\mathrm{se}(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + c_{0.025}(N-K)\,\mathrm{se}(\hat{\beta}_k)) = 0.95.$$

The 95% CI for the true parameter β_k is narrower when $\mathrm{se}(\hat{\beta}_k)$ is smaller, i.e., a lower standard error implies a more precise estimate.

Notice that as $N-K \to \infty$, we have that $\mathrm{t}(N-K)$ approaches the standard normal distribution, so that $c_{0.025}(N-K) \to 1.96$.

Hypothesis Testing

To test $H_0 : \beta_k = \beta_k^0$, we construct $t_k = (\hat{\beta}_k - \beta_k^0)/\text{se}(\hat{\beta}_k)$, which is t-distributed with $N - K$ degrees of freedom under the null H_0.

If t_k is an unlikely draw from a t-distribution with $N - K$ degrees of freedom, then it is unlikely that the null hypothesis H_0 is correct.

If β_k^0 falls outside of the 95% CI, then we reject the null hypothesis H_0 at the 0.05 significance level; otherwise, we fail to reject.

It is common to conduct hypothesis tests at the 0.05 significance level, but results are sometimes reported at the 0.01 and 0.10 levels as well.

Statistical Significance

One can also report the *p-value*, i.e., the significance level at which the null hypothesis is just rejected.

When $\beta_k^0 = 0$, one can report the *t-ratio* given by $\hat{\beta}_k / \text{se}(\hat{\beta}_k)$, which is t-distributed with $N - K$ degrees of freedom under the null.

Typical regression output includes standard errors, t-ratios, p-values, and 95% confidence intervals; see the worked example in *Stata*.

More Hypothesis Testing

We may also be interested in further hypothesis testing, e.g.,

- One-sided alternatives, e.g., $H_0 : \beta_k = \beta_k^0,\ H_a : \beta_k > \beta_k^0$,
- Linear combinations, e.g., $H_0 : \beta_2 + \beta_3 = 1$,
- Joint hypotheses, e.g., $H_0 : \beta_2 = 0,\ \beta_3 = 0$.

Large Sample Inference

Finite sample distributional properties, which are exact, will hold in samples of all sizes, but they also require strong assumptions.

Asymptotic distributional results are approximate in large samples, or more precisely, as the number of observations $N \to \infty$, and with the number of stochastic regressors K fixed.

Large sample distributional properties of an estimator require much weaker assumptions and provide useful approximations.

Consistency

Under the assumptions (A0)–(A3), $\text{plim}\,(\hat{\beta}_k) = \beta_k$, or $\hat{\beta}_k \xrightarrow{P} \beta_k$, i.e., $\hat{\beta}_k$ is a consistent estimator for β_k.

The critical assumption for consistency is that the error term u_i is *orthogonal* to, or does not covary with, the regressors.

We do not require conditional mean and/or variance assumptions, let alone the conditional normality assumption, to establish consistency.

Notice that consistency is a large sample property, while unbiasedness is a finite sample property; one does not necessarily imply the other.

An unbiased estimator is usually (but not necessarily) consistent, and a consistent estimator is usually (but not necessarily) unbiased.

Asymptotic Normality

Under the assumptions (A0)–(A4), we have the distributional result

$$\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{D} \text{Normal}\,(0, \sigma^2/m_k^2),$$

where $m_k^2 = \text{plim}\left(\frac{1}{N}\sum_{i=1}^N \hat{r}_{ki}^2\right)$, and where $\hat{r}_{ki}$ are the residuals from regressing the kth explanatory variable on the other regressors.

Equivalently, we also have that $\hat{\beta}_k \overset{a}{\sim} \text{Normal}\,(\beta_k, \sigma^2/Nm_k^2)$.

Let $\text{AVar}(\hat{\beta}_k) = \sigma^2/Nm_k^2$ denote the *asymptotic variance* of $\hat{\beta}_k$.

Since σ^2 and m_k^2 are unknown, we can estimate them according to $\hat{m}_k^2 = \frac{1}{N}\sum_{i=1}^N \hat{r}_{ki}^2$ and $\hat{\sigma}^2 = \frac{1}{N-K-1}\sum_{i=1}^N \hat{u}_i^2$.

Finally, we have that $\hat{\beta}_k \overset{a}{\sim} \text{Normal}\,(\beta_k, \hat{v}_k)$, which coincides with the finite sample distributional results, but under milder assumptions.

Asymptotic Inference

To construct confidence intervals and perform hypothesis tests, we use the estimate for the asymptotic variance to form standard errors.

Note that standard errors are, in fact, asymptotic standard errors.

We then compare test statistics of various kinds to critical values from the standard normal distributions.

(4) Ordinary Least Squares: Extensions

Heteroskedasticty

To obtain finite sample distributional results, we have assumed that $\mathrm{Var}(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \mathrm{E}(u_i^2|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma^2$ for every observation $i = 1, \ldots, N$, i.e., conditional homoskedasticity.

We often can relax this assumption to *conditional heteroskedasticity*, i.e., $\mathrm{E}(u_i^2|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma_i^2$ for all $i = 1, \ldots, N$.

Unbiasedness and consistency of the OLS estimator do not require the conditional homoskedasticity assumption.

Neither does asymptotic normality; we use the consistent estimator

$$\widehat{\mathrm{AVar}}(\hat{\beta}_k|\{x_{1i}, x_{2i}, \ldots, x_{Ki}\}_{i=1}^N) = \frac{\sum_{i=1}^N \hat{r}_{ki}^2 \hat{u}_i^2}{\mathrm{RSS}_k^2}.$$

Asymptotic Inference

To construct confidence intervals and perform hypothesis tests, we can use the estimate for the asymptotic variance to form standard errors.

Notice that the standard errors are again asymptotic, and this time *heteroskedasticity-consistent* or *heteroskedasticty-robust*.

We can then compare test statistics of various kinds to critical values from the standard normal distribution.

Homoskedasticty Versus Heteroskedasticty

Notice that $\mathrm{E}(u_i^2|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma^2$ is just a special case of $\mathrm{E}(u_i^2|x_{1i}, x_{2i}, \ldots, x_{Ki}) = \sigma_i^2$, with $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, N$.

The natural implication here is that asymptotic inference based on a robust estimator is also valid if conditional homoskedasticity obtains.

However, there is potentially an efficiency loss.

In empirical practice, many researchers are agnostic, but there are some helpful tests, e.g., White's test for heteroskedasticity.

(Feasible) Generalized Least Squares

A more *efficient* approach is to specify the form of heteroskedasticity.

This gives rise to a *generalized least squares* (GLS) approach, often operationalized via *feasible generalized least squares* (FGLS).

Under stronger (milder) assumptions, the FGLS estimator can be shown to be (asymptotically) normally distributed.

Inference using the standard normal distribution once again obtains.