# Introduction to Ordinary Least Squares

Krishna Pendakur

January 25, 2016

## 1 Introduction

1. Let $Y_i, X_i, \varepsilon_i$ be an observed scalar *dependent* variable, a $K-$vector of observed *independent* variables and an *unobserved* scalar *error* term, respectively, for observations $i = 1, ..., N$.

   (a) Saying that $X$ is *independent* and that $Y$ is *dependent* means that we think $X$ causes Y but not vice versa.

   (b) We can think of $\varepsilon_i$ as an *unobserved* independent variable. In the linear model (below), we can think of it as a linear index in a vector of unobserved independent variables.

   (c) Let $X$ be a fixed (nonstochastic) variable unless stated otherwise. Let $\varepsilon$ be a random variable. We can make $X$ a random variable if we wish to. It makes some things, like consideration of endogeneity and weak instruments, much more natural. But fixed $X$ makes the math much easier and doesn't usually change any results.

2. Suppose there is a relationship which holds for each observation (say each year)

$$Y_i = f(X_i, \beta, \varepsilon_i)$$

   where $\beta$ is an unobserved vector of *parameters* (could be a $K-$vector)

   (a) The parameters $\beta$ together with the function $f$ govern how $X$ affects $Y$. We call such a model 'parametric' because we have *a priori* information on its functional structure (parametric form). Notice that the function $f$ which depends on the parameters $\beta$ does *not* depend on $i$.

   (b) For every value of $X_i, \beta, \varepsilon_i$ there is exactly one value of $Y_i$. Since $\varepsilon_i$ is random, $Y_i$ must be random.

3. This may be written in terms of matrices and vectors, too. Let $Y, X$ be the $N-$vector of observations $Y_i$, the $NxK$ matrix comprised of the stacked observations of $X_i$. Let $\varepsilon$ be the vector (of unspecified length) of *error* terms. Here, rows are observations and columns (of $X$) are variables. We may then write

$$Y = f(X, \beta, \varepsilon).$$

4. We typically cannot make much progress on a model this general, so we assume that the unobserved error term comes in additively:

$$Y = f(X, \beta) + \varepsilon$$

   (a) The error term $N-$vector $\varepsilon$ has many interpretations. It could be:
      i. Unmodeled factors and left-out variables
      ii. Measurement error in $Y$
      iii. Unobserved heterogeneity
      iv. It could be the sum of random stuff which is independent of $X$.
      v. You could think of this as random *shifters* (like business cycles or the weather), or
      vi. As left-out variables (like the age distribution of the population)
      vii. It could be mean-zero measurement error on $Y$
      viii. It could be true indeterminacy in the relationship

5. We often have trouble estimating a model so general, so we further specify that $Y$ is linear in $X$

$$Y = X\beta + \varepsilon,$$

   where $\beta$ is an unobserved $K-$vector.

   (a) Given that we know $X$, if we knew $\beta$, then we would know everything predictable about $Y$. That is, our ignorance would only be about the error term, which given its definition above, is perfectly acceptable.
   (b) Equivalently, we may say that $Y$ is linear in $\beta$.
   (c) A simple function of $X$ is desirable, but why linear? Linear functions with additive errors have the property that the derivative of $Y$ with respect to $X$ is $\beta$. So, if we learn $\beta$, we immediately have something useful that we can interpret.

6. Let the word *estimator* denote a function of observed variables $Y, X$ that provides an estimate of $\beta$. If we had an estimator $\beta^*$ which yielded an estimate $\hat{\beta}$, then we could write

$$Y = X\hat{\beta} + e$$

   where $e$ is the $N-$vector of *residuals*.

   (a) Note that the word *estimator* describes a tool that gives an *estimate*. The estimator $\beta^*$ is a function of the data, and the estimate $\hat{\beta}$ is the value of that function.
   (b) A residual is an empirical quantity; an error is a feature of the true model.
   (c) Here, $X\hat{\beta}$ is the *prediction* of $Y$ given $X$ (more on this later), and $e$ is the *prediction error*. (If $e$ were 0 then, the prediction is exactly right.) Given an estimate of $\beta$, we can define the prediction error in reverse:

$$e = Y - X\hat{\beta}$$

7. How do you estimate $\beta$? What would characterise a good estimator $\beta^*$? (Kennedy)

# 2 Nice Things About Estimators

## 2.1 Computational ease

1. Used to be much more important when computers were tiny (actually really large, but weak).

2. Linear estimators can be computed in one (or maybe two) steps. Nonlinear estimators typically require more steps. Nonlinear parametric models used to be time-consuming to estimate, but now they aren't.

## 2.2 Motivation 1 for OLS: Minimising Prediction Error

1. If prediction error is bad, then one might wish to minimise its (weighted) sum, which is the (weighted) sum of residuals: obtain $\beta^*$ via

$$\min_{\widehat{\beta}} \sum_{i=1}^{N} \omega_i |e_i|,$$

where $\omega_i$ is a weight for each residual.

   (a) We minimize a function of the *within sample* residuals.

   (b) Focusing on prediction error/residuals has some features:

      i. This is not a statistical criterion.

      ii. It minimises a function of the absolute value of residuals. This implies that the investigator feels the same about negative and positive errors—she doesn't like either kind.

      iii. It should yield a good predictor.

      iv. But, it is a predictor of something you already have: the sample values of $Y_i$.

   (c) weighting by the size of the residual gives *ordinary least squares*

$$\begin{aligned}\min_{\widehat{\beta}} \sum_{i=1}^{N} |e_i| \cdot |e_i| &= \min_{\widehat{\beta}} \sum_{i=1}^{N} |Y_i - X_i\widehat{\beta}| \cdot |Y_i - X_i\widehat{\beta}| \\ &= \min_{\widehat{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i\widehat{\beta}\right)^2.\end{aligned}$$

      i. Here, big residuals are very important. If one residual is twice as large as another, it gets twice the weight. This means that *outliers*, defined as big residuals in $Y$, can drag the $\widehat{\beta}$ around.

      ii. Consider a model where $X_i = 1$, so that we minimize

$$= \min_{\widehat{Y}} \sum_{i=1}^{N} \left(Y_i - \widehat{Y}\right)^2.$$

The first-order condition is

$$2\sum_{i=1}^{N}\left(Y_i-\widehat{Y}\right)=0$$

$$2\sum_{i=1}^{N}Y_i=2N\widehat{Y}$$

$$\widehat{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$$

Here, $\widehat{Y}$ is the mean of $Y_i$. For this reason, OLS is often called *mean regression*. If we threw in conditioning variables, it would be conditional mean regression.

(d) Weighting with unitary weights gives *least absolute deviations* (LAD):

$$\min_{\widehat{\beta}}\sum_{i=1}^{N}1\cdot|e_i|=\min_{\widehat{\beta}}\sum_{i=1}^{N}|Y_i-X_i\widehat{\beta}|,$$

and if we consider the case where $X_i=1$, we have

$$\min_{\widehat{Y}}\sum_{i=1}^{N}|Y_i-\widehat{Y}|$$

with a first-order condition of

$$\sum_{i=1}^{N}\left(2I(Y_i-\widehat{Y}>0)-1\right)=0$$

$$\sum_{i=1}^{N}\left(I(Y_i-\widehat{Y}>0)\right)=\frac{N}{2}$$

where $I$ is the indicator function. Here, we select $\widehat{Y}$ so that half $(N/2)$ of the observations are above $\widehat{Y}$ (so that $I(Y_i-\widehat{Y}>0)=1$). For this reason, least absolute deviations is often called *median regression*. If we threw in conditioning variables, it would be conditional median regression.

(e) Weighting with asymmetric weights will push the errors to one side or the other of the regression curve. Quantile regression is analogous to median regression, except that it picks a quantile different from the 0.5 quantile (that is, the median). So, for the $\tau$'th quantile regression, we define the weight function to give a first-order condition that puts $\tau$ of the data below the regression line and $1-\tau$ of it above:

$$\min_{\widehat{\beta}}\sum_{i=1}^{N}\omega\left(\tau,e_i\right)\cdot|e_i|,$$

where

$$\begin{aligned}\omega(\tau, e_i) &= \tau \qquad \text{if } e_i \geq 0, \\ &= 1-\tau \text{ if } e_i < 0.\end{aligned}$$

or, equivalently,

$$\omega(\tau, e_i) = \tau - (2\tau - 1)\, I(Y_i - \widehat{Y} < 0)$$

Here $\omega(\tau, e_i) \cdot |e_i| = \rho_\tau(e_i)$ from Koenker and Hallock. For $\tau = 0.1$, for example, we put 9 times as much weight on negative errors as on positive ones. This puts the regression line way down in the $Y$ direction in the data cloud.

i. if we consider the case where $X_i = 1$, we have

$$\min_{\widehat{\beta}} \sum_{i=1}^{N} \omega\left(\tau, Y_i - \hat{Y}\right) \cdot |Y_i - \hat{Y}|,$$

with a first-order condition of

$$\sum_{i=1}^{N} \left(I(Y_i - \widehat{Y} < 0)\right) = \tau N$$

where $I$ is the indicator function. Here, we select $\widehat{Y}$ so that a fraction $\tau$ of the observations are below $\widehat{Y}$ (so that $I(Y_i - \widehat{Y} > 0) = 1$).

ii. *expectile* regression is a cute mix of quantile and mean regression. In expectile regression, we use the same weight function from, e.g. Koenker and Hallock, and $\min_{\widehat{\beta}} \sum_{i=1}^{N} \omega(\tau, e_i) \cdot e_i^2$,

(f) Computation

i. OLS computation is one-step:

$$\min_{\widehat{\beta}} \sum_{i=1}^{N} \left(Y_i - X_i\widehat{\beta}\right)^2,$$

or, in matrix notation,

$$\min_{\widehat{\beta}} \left(Y - X\widehat{\beta}\right)' \left(Y - X\widehat{\beta}\right),$$

leading to a FOC:

$$2X'\left(Y - X\widehat{\beta}\right) = 0,$$

leading to a solution

$$\begin{aligned}X'Y - X'X\widehat{\beta} &= 0 \\ X'X\widehat{\beta} &= X'Y \\ \widehat{\beta} &= (X'X)^{-1} X'Y.\end{aligned}$$

ii. For both LAD and QR, we solve the model via linear programming rather than by using the one-step OLS solution (because the indictor function uses inequalities). However, with modern computers, this is no big deal. QR computation is a linear programming problem, that is a problem of satisfying many inequalities. With a finite number of inequalities, such problems often can be solved in a finite number of computations.

(g) For all these estimators based on prediction error/residuals:

i. draw the weight function;

ii. draw the first-order condition;

iii. show how the estimator treats residuals and outliers.

(h) None of these estimators use a model for $Y$. That is, none requires us to know the true relationship between $Y, X, \varepsilon$. One therefore think about $X_i\widehat{\beta}$ from a linear regression as the *best linear unbiased predictor* of $Y$ given $X$, even when $Y$ is in fact a nonlinear function of $X$. Here, best-ness is in a mean-squared error sense.

2. Interpretation of the regression function as a conditional mean function works as follows (see Angrist and Pischke Ch3):

(a) The conditional mean function always exists for the random variables $X, Y$, it is

$$m(X_i) = E\left[Y_i|_{X_i}\right].$$

Causility is irrelevant here: this just requires the existence of a joint distribution of $X, Y$, and such a distribution exists regardless of the support and discreteness of either $Y$ or $X$.

(b) We can try to estimate this thing. If $m(X_i)$ is linear in $X$, then the linear regression function estimated on the population of $X, Y$ will pick it up exactly. That is,

$$X_i\widehat{\beta} = E\left[Y_i\right]|_{X_i},$$

subject to the restriction that the conditional expectation function $E\left[Y_i\right]|_{X_i}$ is linear in $X$.

(c) This interpretation carries through even when $E\left[Y_i\right]|_{X_i}$ is not a linear function of $X$, that is, when the linear regression is *misspecified*. In this case, the estimated linear regression function will give the smallest mean-squared error,

$$E\left[\left(m(X_i) - X_i\widehat{\beta}\right)^2\right],$$

of any linear function of $X$.

3. Saturation and Nonparametrics

(a) A model is called *saturated* if the regressors are discrete, and the model contains a parameter for every combination of regressors observed in the data.

i. For example, if sex=1, 2 and province=1, ..., 10, and if persons of both sex are observed in every province and nothing else is observed, the saturated model has 20 parameters.

ii. A saturated regression model estimated on the population of data will pick up the conditional mean function $m$ exactly.

(b) Nonparametric models are the analogue to saturated models when the regressors are continuous. They pick up the conditional mean function $m$ as the population size gets infinite.

i. The idea of nonparametric estimation is to let the data determine the shape of the function, rather than priors. However, the way we do this is to make the model incredibly highly parametric (a funny use of the prefix non!). In particular, nonparametric estimators let the number of parameters grow with the sample size, which means that asymptotically, you have an infinite number of parameters, so that asymptotically, you can pick up any shape at all.

(c) With samples from the population, both saturated and nonparametric models provide estimates of the conditional mean function, which can in principle be computed exactly.

## 2.3 Unbiasedness and Efficiency

1. Mean, Variance, Unbiasedness and Efficiency are concepts related to how the estimated values of an estimator behave across repeated samples. Consider the sample mean

$$\widehat{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Here, it is computed off of a sample $i = 1, ..., N$. But, what if it had been computed off of a different sample, eg, $i = N+1, ..., 2N$? What can we say about the distribution of $\widehat{Y}$ over repeated samples of size $N$?

2. *unbiasedness*: Bias is defined as

$$bias\left[\widehat{\beta}\right] = E\left[\widehat{\beta}\right] - \beta$$

if in repeated samples of a given size the mean value of the estimated parameters $\widehat{\beta}$ is equal to $\beta$, so that $E\left[\widehat{\beta}\right] = \beta$, then we say that the estimator $\beta^*$ is *unbiased*.

3. *efficiency*: if in repeated samples of a given size the variance of the estimated parameters $\widehat{\beta}$ is the lowest of any estimator, then we say that the estimator $\beta^*$ is *efficient*. If it is the lowest of any estimator in a particular class (for example, the class of estimators linear in $X$), then we say that it is efficient in that class.

(a) for any random variable $z$, the variance of $z$ is

$$V(z) = E[z]^2 - E[z^2],$$

or, equivalently,

$$V(z) = E[z - E[z]]^2.$$

(b) If $z$ is a random $T-$vector, then we may express it in matrix notation as the $TxT$ symmetric matrix

$$V(z) = E[z - E[z]]E[z - E[z]]'$$

the covariance between any two elements of $z$ is given by the appropriate element of $V(z)$, and the covariance of any two subvectors of $z$ is the appropriate submatrix of $V(z)$.

4. *mean squared error*: mean squared error is defined over the difference between an estimate, $\hat{\beta}$, and its target, $\beta$. Mean squared error is a $KxK$ matrix given by

$$MSE = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = bias[\hat{\beta}]bias[\hat{\beta}]' + V[\hat{\beta}]$$

5. You don't always want both unbiasedness and low variance. If an estimator is efficient and biased, but only a little biased, then one might prefer it to an inefficient unbiased estimator because it would be closer to the target ($\beta$) in most of the repeated samples.

   (a) minimising mean squared error, MSE=squared bias+variance, provides another measure of goodness of the estimator $\beta$*.

## 2.4 Asymptotic characteristics (very large samples)

1. although you work with data of a given finite size, we often can't say much about the behaviour of estimators in small samples.

2. the analogue to unbiasedness in small samples is *consistency* as the sample size grows to infinity.

3. we call an estimator *consistent* if it is unbiased in the limit as the sample size goes to infinity.

the analogue to efficiency is *asymptotic variance*. For any sample size $N$, the asymptotic variance is usually defined as $AV = NV(\hat{\beta})$ as $N$ goes to infinity. Thus, although $V(\hat{\beta})$ goes to zero as $N$ goes to infinity, the asymptotic variance may converge to something nonzero.

## 2.5 Motivation 2 for OLS: Exogeneity and the Method of Moments

1. *Method of moments* is a strategy for making estimators. The strategy is:

   (a) Take an expectation from theory.

   (b) Insert sample values and remove expectation operator

   (c) create the estimator by replacing parameters with estimates.

2. Consider a model

$$Y = X\beta + \varepsilon$$

(Kennedy Assumption 1: linear model) where

$$E\left[X'\varepsilon\right] = 0_K$$

(exogeneity)

3. Implement the Method of Moments Estimator

   (a) Expectation from theory is $E\left[X'\varepsilon\right] = 0_K$

   (b) Sample values to insert are: $X_i$ for $X$ and $e_i$ for the error term $\varepsilon_i$. Here, the residual is the sample analog of the error term

$$E\left[X'\varepsilon\right] = 0_K$$

$$X'e = 0_K$$

$$X'\left[Y - X\beta\right] = 0$$

   (c) Solve

$$X\left[Y - X\hat{\beta}\right] = 0$$

$$X'Y = X'X\hat{\beta}$$

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y$$

4. Ordinary Least squares. Again. Here, we did not invoke any conditions on the bias or variance of the estimator. All we did was take the exogeneity restriction as seriously as possible, and substitute residuals for errors in that restriction.

## 2.6 Motivation 3 for OLS: Efficiency and a Familiar Result

Ordinary Least Squares is the efficient estimator in the class of linear unbiased estimators (Best Linear Unbiased Estimator) for estimation of $\beta$ in models whose data generating process (DGP) is

1.

$$Y = X\beta + \varepsilon$$

(Kennedy Assumption 1: linear model) where

$$E\left[X'\varepsilon\right] = 0_K$$

(exogeneity), and $X$ includes a column of ones, and

$$E\left[\varepsilon\varepsilon'\right] = \sigma^2 I_N$$

where $I$ is the identity matrix.

2.
$$E\left[X'\varepsilon\right] = 0_K$$
implies $E\left[1'\varepsilon\right] = E\left[\varepsilon_i\right] = 0$: the errors are mean zero (Kennedy Assumption 2)

3. $X$ and $\varepsilon$ are uncorrelated implying $X$ is exogenous (Kennedy Assumption 4)

4. $E\left[\varepsilon\varepsilon'\right] = \sigma^2 I_N$ implies (Kennedy Assumption 3)

   (a) errors are homoskedastic (diagonals of $E\left[\varepsilon\varepsilon'\right]$ have the same value)

   (b) errors are uncorrelated (off-diagonals of $E\left[\varepsilon\varepsilon'\right]$ are zero)

5. A unique $\hat{\beta}$ exists as a least squares solution if and only if $X$ is of full rank and *N>rank(X)*. (Kennedy Assumption 5)

6. The OLS estimator has the lowest variance of all unbiased estimators for this model. This is a statistical assessment of a statistic (the estimated parameter vector $\hat{\beta}$).

7. It is a statistical assessment based on what would happen in *repeated samples* from the data generating function process above. The idea of repeated samples is weird: we only have one world, and yet we imagine redrawing the world and estimating the parameters using the OLS estimator. Then, we consider the statistical properties (bias, variance) of the estimates over these repeated samples.

## 2.7 Motivation 4 for OLS: Improbability and the Method of Maximum Likelihood

1. If improbability is bad, then one might wish to maximise by choice of parameters the probability that the observed data could come from the model.

2. You need to specify parametric model as well as the (conditional) distribution of error terms

3. Let subscripts $i = 1, ..., N$ index the observations (rows).

4. Consider
$$Y_i = X_i\beta + \varepsilon_i,\ \varepsilon_i \tilde{} f(\varepsilon)$$
which states the model for each observation as well as the distribution of the error term. The function $f$ gives the density of the error at any particular value. Usually, we assume that this function is high for error values near zero, and low for error values far from zero.

5. The function $f$ is the *probability density function* (pdf) for any random variable $x$. Probability density functions are related to probabilities. In particular,
$$P\left[a < x < b\right] = \int_a^b f(x)\partial x.$$

The integral of the pdf between two values gives the probability of observing the random variable between those two values. The pdf is related to a primitive, the *cumulative density function* (cdf) which gives the integral of the pdf from $-\infty$ to a value. For a pdf $f$, we write the cdf $F$ as

$$F(u) = \int_{-\infty}^{u} f(x)\partial x.$$

implying

$$\partial F(u)/\partial x = f(x).$$

Here, the cdf is the primitive because the cdf is defined even if is zero, or discontinuous. In contrast, the pdf is only defined everywhere if the cdf has derivatives everywhere.

6. Since probabilities are defined by integrals of densities, we can write probabilities in terms of cdfs, too:
$$P\left[a < x < b\right] = F(b) - F(a).$$

7. pdf's have the useful feature that, like with probabilities, the pdf for two independent random variables is the product of their pdf's.

8. pdf's can be interpreted similarly to probabilities in the following way: as the span $[a, b]$ gets very small, the density is approximately equal to the probability of observed the random variable in the range $[a, b]$. Of course, when the span is exactly zero (for a continuous random variable), the probability is exactly zero.

9. The trick with maximum likelihood estimation is to use densities instead of probabilities for continuous random variables, and to maximize the probability density (by choice of parameters) of observing the sample we actually saw. (With discrete random variables, you can actually maximize the probability.)

10. We call the pdf for the entire sample of data, given the parameters, as the *likelihood function*:
$$L\left(f, \beta, Y, X\right) = \prod_{i=1}^{N} f\left(\varepsilon(Y_i, X_i)\right) = \prod_{i=1}^{N} f\left(Y_i - X_i\beta\right).$$
The first equality follows from the fact that the only random variable in the model is $\varepsilon_i$, and since each of these is an independent draw from $f$, the likelihood is the product of the value of $f$ for each $Y_i$,$X_i$ pair. The second equality follows from the fact that the value of $\varepsilon_i$ corresponding to each pair is given by the model.

11. Clearly, we cannot easily proceed without knowledge of $f$. Let $f$ be the *normal* density function:
$$f(\varepsilon) = \frac{\exp\left(-\left(\varepsilon - \mu\right)^2/2\sigma^2\right)}{\sqrt{2\pi\sigma^2}}.$$
where $\mu$ and $\sigma$ are parameters of the distribution. It turns out that the mean of this distribution is $\mu$ (one of the parameters). Since $\varepsilon$ are presumed mean-zero, we can drop this parameter:
$$f(\varepsilon) = \exp\left(-\varepsilon^2/\sigma^2 2\right)/\sigma\sqrt{2\pi}.$$

You will see in a moment why an investigator might use this seemingly ugly density function.

12. Nobody likes products, so we will maximize the log of the likelihood function rather than its level:

$$\begin{aligned}\ln L\left(f,\beta,Y,X\right) &= \sum_{i=1}^{N}\ln f\left(Y_i - X_i\beta\right)\\ &= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{Y_i - X_i\beta}{\sigma}\right)^2 - N\ln\left(\sigma\sqrt{2\pi}\right)\end{aligned}$$

13. The method of maximum likelihood tells us that this equation holds in the model. Now, we maximize the (log) likelihood by choice of the parameter vector $\hat{\beta}$:

$$\begin{aligned}&\max_{\hat{\beta},\sigma} -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{Y_i - X_i\hat{\beta}}{\sigma}\right)^2 - N\ln\left(\sigma\sqrt{2\pi}\right)\\ &\Leftrightarrow \min_{\hat{\beta},\sigma}\frac{1}{2\sigma}\sum_{i=1}^{N}\left(Y_i - X_i\hat{\beta}\right)^2 + N\ln\left(\sigma\sqrt{2\pi}\right).\end{aligned}$$

If we had $\sigma$ in our pocket already, the right-hand term would drop out of the minimisation, leaving

$$\min_{\hat{\beta}}\sum_{i=1}^{N}\left(Y_i - X_i\hat{\beta}\right)^2,$$

that is, ordinary least squares.

(a) If we don't have $\sigma$ in our pocket, we still get OLS because $\sigma$ does not interact with $\hat{\beta}$–we can *concentrate* it out.

14. In many of these contexts, there exists a true parameter vector $\beta$, and $\hat{\beta}$ may be taken as an estimator of it. In our MoM, BLUE and ML approaches we assumed the linear model

$$Y = X\beta + \varepsilon$$

and exogeneity of fixed (nonstochastic) $X$

$$E\left[X'\varepsilon\right] = 0_K.$$

Can we say much about the OLS estimator's relationship to $\beta$ with just this?

(a) The bias of the estimator is the deviation between its expected value and the target: $E\left[\hat{\beta}_{OLS}\right] - \beta$ .

$$E\left[\hat{\beta}_{OLS}\right] - \beta = E\left[\left(X'X\right)^{-1}X'Y\right] - \beta$$

$$
\begin{aligned}
&= E\left[\beta + (X'X)^{-1} X'\varepsilon\right] - \beta \\
&= \beta + E\left[(X'X)^{-1} X'\varepsilon\right] - \beta \\
&= E\left[(X'X)^{-1} X'\varepsilon\right] \\
&= (X'X)^{-1} E\left[X'\varepsilon\right] = (X'X)^{-1} 0_K = 0_K
\end{aligned}
$$

(b) The estimator is unbiased. Its unbiasedness is due entirely to the assumed exogeneity of $X$. There's no distance between the assumption of exogeneity and the implication of unbiasedness. Assuming exogeneity is equivalent to assuming unbiasedness.

(c) Another way to see this is to start by assuming that $X$ is not exogenous (aka: is endogenous), and so assume

$$Y = X\beta + \varepsilon$$

and

$$E\left[X'\varepsilon\right] = \Gamma$$

where $\Gamma$ is a nonzero $K-$vector.

(d) By the same reasoning as above, the estimator is now biased, and its bias equals

$$E\left[\hat{\beta}_{OLS}\right] - \beta = (X'X)^{-1} E\left[X'\varepsilon\right] = (X'X)^{-1} \Gamma$$

(e) Suppose we had the population of data, and that we knew $\varepsilon$, and that we regressed them on $X$. The coefficients from this regression would be $(X'X)^{-1} X'\varepsilon$. Since it is for the population, we would get exactly the bias term $(X'X)^{-1} \Gamma$. For that reason, I like to think of the bias of OLS in a linear model as the *load* of $\varepsilon$ of $X$ that is misattributed to $X$ instead of to $\varepsilon$. Much more on this later.

15. Fixed Versus Stochastic $X$

(a) I often use fixed $X$ to describe stuff. It is easier—less stuff to carry through expectations.

(b) Consider the model above

$$Y = X\beta + \varepsilon$$

and

$$E\left[X'\varepsilon\right] = 0_K,$$

with associated estimator

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y.$$

(c) Here, the fixed $X$ assumption is a bit clunky in the exogeneity condition. It says that the random variable $\varepsilon$ has the property that, for the given data $X$, $E\left[X'\varepsilon\right] = 0_K$. This is strange. What if we had a different set of $X$ in our data? That would be a different exogeneity restriction. Nonetheless, we can write down this restriction and say it to ourselves: $\varepsilon$ is orthogonal to $X$ in expectation.

(d) The fixed $X$ assumption has a payoff: When we calculate the expection of $\hat{\beta}_{OLS}$, it is very straighforward:

$$E[\hat{\beta}_{OLS}] = E\left[(X'X)^{-1}X'Y\right] = \beta + E\left[(X'X)^{-1}X'\varepsilon\right] = \beta + (X'X)^{-1}E[X'\varepsilon] = \beta + (X'X)^{-1}0_K =$$

(e) You get to just pull the $X$'s out of the expectation, and only worry about the random variable $\varepsilon$.

(f) Suppose $X$ were random. Then, you'd have to consider the covariance of $(X'X)^{-1}$and $X'Y$ and you'd have to know something about the distribution of $X$.

(g) When assuming random $X$, which is more natural than fixed $X$, we often invoke two strong restrictions: 1) $X_i \sim iid$ and 2) a stricter exogeneity condition $E[\varepsilon_i]|_{X_i} = 0_K$ for all $i = 1, ..., N$.

(h) Then, we have

$$E[\hat{\beta}_{OLS}] = E\left[(X'X)^{-1}X'Y\right] = \beta + E\left[(X'X)^{-1}X'\varepsilon\right].$$

We can condition and integrate the right-hand expectation (the matrix expressions in the left-hand expression are sums over the observations $i$, so the expectation of the matrix sums is equal to the sum over the $X$'s of the elements in the matrix sums):

$$E\left[(X'X)^{-1}X'\varepsilon\right] = \int_x E\left[(X_i'X_i)^{-1}X_i'\varepsilon_i\right]|_{X_i=x}\partial x$$

$$= \int_x (X_i'X_i)^{-1}E[X_i'\varepsilon_i]|_{X_i=x}\partial x = \int_x (X_i'X_i)^{-1}0_K|_{X_i=x}\partial x = 0_K$$

yielding

$$E[\hat{\beta}_{OLS}] = E\left[(X_i'X_i)^{-1}X_i'Y_i\right] = \beta + 0_K.$$

So, we typically get to the same place, but with more effort.

(i) Now, consider estimating the same model by ML with random $X$:

$$Y = X\beta + \varepsilon,$$

$$\varepsilon_i \sim N(0, \sigma^2),$$

and

$$X_i \sim N(\mu, \Sigma).$$

These assumptions imply that $\varepsilon_i \perp X_i$ and all $\varepsilon_i, X_i$ are iid. The likelihood function for independent random variables that are iid is the product of each density producted over the observations:

$$L(f, g, \beta, Y, X) = \sum_{i=1}^{N} f(Y_i - X_i\beta)\, g(X_i - \mu)$$

where $f$ is the density of the $N(0, \sigma^2)$ and $g$ is the density of the $N(0_K, \Sigma)$.

(j) If you take the log of this likelihood function, and take a derivative with respect to $\beta$, you get the same first order condition as if $X_i$ were fixed, because the parameters in the distribution of $X_i$ ($\mu$ and $\Sigma$) do not show up in the derivative of the log-likelihood with respect to $\beta$.

(k) Consequently, I will mostly use the fixed $X$ approach to get you through this material, and will alert you to when random $X$ leads you to substantively different places (as in the consideration of weak instruments, for example). For your information, my approach is a bit old-timey, and has been abandoned in textbooks for roughly a decade.

16. Compare these approaches:

| | Min. Pred. Error | MoM estimator | BLUE Estimator | ML Estimator |
|---|---|---|---|---|
| Full DGP known | no | no | no | yes |
| error mean known | no | yes | yes | yes |
| error variance known | no | no | yes | yes |
| error pdf known | no | no | no | yes |
| linear model only | no | no | yes | no |