

## The Basic Linear Model

The basic linear model is

$$y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \dots + X_{i,k}\beta_k + \varepsilon_i, \quad \text{for } i = 1, \dots, n \quad (1)$$

which is written compactly in matrix notation as

$$y = X\beta + \varepsilon \quad (2)$$

The 'data' is arranged in two objects,  $X$  and  $y$ . It is usual to think of the  $(n \times 1)$  vector  $y$  as the variable to be 'explained' and that the  $(n \times k)$  matrix  $X$  (representing each of the  $n$  values of the  $k$  variables) does the 'explaining'. The meaning of the statement ' $X$  explains  $y$ ' depends heavily on the context.

## The Basic Linear Model (2)

Some mechanics of  $y = X\beta + \varepsilon$  :

$$\begin{matrix} y & = & X & \beta & + & \varepsilon \\ (n \times 1) & & (n \times k) & (k \times 1) & & (n \times 1) \end{matrix}, \quad (3)$$

so we see that  $X\beta$  is an  $(n \times 1)$  object and so the right hand side (RHS) of (3) adds two  $(n \times 1)$  objects together to get the LHS  $y$  which is also an  $(n \times 1)$  object.

In most applications the first column of  $X$  is a column of '1' s.

## Interpretations of the Basic Linear Model

” Interpretations” here means almost ” attitudes towards” .

One common attitude is that the ‘model’ represents a data generating process (“DGP”). Little is lost in this interpretation if one thinks of (1) as

1. Nature or the experimenter chooses  $X$ .
2.  $X$  is multiplied by  $\beta$ . (More concretely, each column of  $X$  is multiplied by its corresponding component of  $\beta$ .)
3. Nature chooses  $\varepsilon$ .
4. “We” see  $y$ .

## Interpretations of the Basic Linear Model (2)

Another view would be that there is a joint probability density function  $f(y, X)$ . We see random draws from from this density. Then  $y = X\beta + \varepsilon$  (or  $y_i = X_i\beta + \varepsilon_i$ , if we want to draw attention to observation  $i$ ) is a description or approximation of what we see in our sample: in writing either of these expressions we make no claim about the *genesis* of  $y$ , that is we say nothing of how  $y$  was created.

The difference is that if we have the first attitude, that the model corresponds to a DGP, we tend to think in terms of counterfactuals: that if  $X$  had been different,  $y$  would have been different. Something akin to a causal mechanism is apparently (or ostensibly) implied.

In most contexts we do not have to make a choice about adopting either of these views, and we can shift (sometimes uncomfortably) between them.

## Common Assumptions about the Linear Model

Following Greene (Fifth Edition) Table 2.1 p.10.

**A1. Linearity.**

**A2. Full rank.** There is no exact linear relationship among any of the independent variables in the model. This assumption will be necessary for estimation of the parameters of the model.

**A3. Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jK}] = 0$ . This states that the expected value of the disturbance at observation  $i$  in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of  $\varepsilon_i$ .

**A4. Homoscedasticity and nonautocorrelation:** Each disturbance  $\varepsilon_i$  has the same finite variance,  $\sigma^2$  and is uncorrelated with every other disturbance,  $\varepsilon_j$ . This assumption limits the generality of the model, and we will want to examine how to relax it.

**A5. Exogenously generated data:** The data in  $(x_{j1}, x_{j2}, \dots, x_{jK})$  may be any mixture of constants and random variables. The process generating the data operates outside the assumptions of the model— that is, independently of the process that generates  $\varepsilon_i$ . Note that this extends **A3**. Analysis is done conditionally on the observed  $X$ .

**A6. Normal distribution:** The disturbances are normally distributed. (This is a convenience that we will dispense with after some analysis of its implications.)

## Discussion of Linearity and Functional Form.

Sometimes our first impulse is to think in terms of variables that are obviously NOT related to each other in the linear form envisaged by equations 1, 2, and 3. But in many such instances we can transform the relation to one that is linear. The most straightforward example of this is the Cobb–Douglas production function:

$$Y_i = \gamma K_i^{\beta_K} L_i^{\beta_L}, \quad (4)$$

which is often ‘stochasticized’ by

$$Y_i = \gamma K_i^{\beta_K} L_i^{\beta_L} e^{\varepsilon_i}, \quad (5)$$

which, if we take logarithms of both sides, becomes

$$\log(Y_i) = \log(\gamma) + \beta_K \log(K_i) + \beta_L \log(L_i) + \varepsilon_i \quad (6)$$

This is linear in  $\log(Y_i)$ ,  $\log(K_i)$ , and  $\log(L_i)$  with coefficients  $\beta = \{\log(\gamma), \beta_K, \beta_L\}$ .

## Another 'Model' in Log Form: The Wage Equation

It is no exaggeration to say that much of modern econometrics has been developed to deal with the problem of the explanation of individuals' wages. Here it is plausible to start from:

$$wage = f(education, ability, labor\ market\ history, \varepsilon), \quad (7)$$

which very rapidly becomes e.g.

$$\begin{aligned} \log(wage) = & \beta_0 + \beta_1(education) + \beta_2(experience) \\ & + \text{linear terms in age, other variables} + \varepsilon \end{aligned} \quad (8)$$

The wage equation is good vehicle for exploring differences between the extremes of the 'DGP attitude' and the ' $f(y, X)$  attitude' and their relations to issues in contemporary econometrics. Let's estimate one using least squares.

## A Wage Equation

Let's estimate a very simple (log) wage equation by ordinary least squares (OLS): (dataset description)

$$\log(wage) = 5.5027 + .0778 * educ + .0198 * exper$$

(0.1120) (.0066)      (.0033)

Other things equal, an extra year of education increases wages by 7.8% and another year of experience increases wages 2%. If we use the DGP interpretation of this regression, we would be tempted to say that *had* a person attended school for another year, then their wage would be 7.8% higher (or more likely  $7.8 - 2 = 5.8\%$  higher since they would probably have one less year of experience as a result). In the  $f(y, X)$  interpretation, we might say that for equal experience, those with an additional year of education did 'in fact' earn 7.8% more, on average. (A description.)

## Formal Properties of OLS Estimation of the Linear Model

OLS='ordinary least squares'=minimize the sum of the squared residuals. If the 'true model' is  $y = X\beta + \varepsilon$ , we write the estimated model as  $y = Xb + u$ ; the  $u$ 's are residuals, the  $\varepsilon$ 's disturbances;  $b$  is an estimate of  $\beta$ . Thus  $u = y - Xb$  and  $u$  is  $(n \times 1)$ ; so  $u'u$  is the sum of the squared residuals. Thus:

$$\begin{aligned}Q(b) &= u'u = (y - Xb)'(y - Xb) \\ \frac{\partial Q}{\partial b} &= -2X'(y - Xb) = 0 \text{ (FOC)} \\ X'y &= X'Xb \\ (X'X)^{-1}X'y &= b\end{aligned}$$

The last line tells how to compute  $b$  and that  $(X'X)$  must be nonsingular to do so (if we want a unique  $b$ ). A further interesting way to write the FOC is:

$$X'u = 0$$

## The Expectation and Variance of $b$

$$E(b) = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'(X\beta + \varepsilon)] = \beta + E[(X'X)^{-1}X'\varepsilon].$$

The value of the  $E[\cdot]$  term is a question about the relation between  $X$ 's and disturbances. If the  $X$ 's are independent of the disturbances, it is zero. Greene's Assumptions A3 or A5 also make this term zero. Now  $b - \beta = (X'X)^{-1}X'\varepsilon$  and by definition  $V(b) = E\{[b - E(b)][b - E(b)]'\}$ ; when  $E(b) = \beta$  we have

$$\begin{aligned} V(b) &= E\{[(X'X)^{-1}X'\varepsilon][\varepsilon'X(X'X)^{-1}]\} \\ &= E\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\} \end{aligned}$$

If we further assume A4 of Greene (homoscedasticity and no autocorrelation) then  $E(\varepsilon\varepsilon'|X) = \sigma^2I$  and the last line can be written as

$$\begin{aligned} V(b) &= (X'X)^{-1}X'(\sigma^2I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

## Normality of the Disturbances and the Normality of $b$

We have just demonstrated that—at least with enough assumptions—the OLS estimator of  $\beta$  in the linear model is unbiased and has variance  $\sigma^2(X'X)^{-1}$ . Under these conditions,  $b = \beta + (X'X)^{-1}X'\varepsilon$ . If  $\varepsilon$  is normal, then  $b$  is normal also, since it is a linear combination (  $(X'X)^{-1}X'$  ) of the normally distributed disturbances. Thus under these rather strict assumptions we have the exact finite sample distribution of  $b$ , since  $b$  is normal and we know its mean and variance. (Unconditionally if we take  $X$  to be fixed and conditioned on  $X$  if we take  $X$  to be stochastic.)

It should be understood that finite sample normality of  $b$  is based on normality of the disturbances—indeed on normal, homoscedastic, non-autocorrelated disturbances; asymptotic or ‘large-sample’ normality of  $b$  can be based on other arguments.

## Exploiting the normality of $b$ : hypothesis testing

If  $b$  is normal with mean  $\beta$  and variance  $\sigma^2(X'X)^{-1}$ , we can compute the distribution of  $b$  (and/or its components) for known  $\sigma^2$  when  $\beta$  is assumed known: the distribution of  $b[j]$  is  $N(\beta[j], \sigma^2(X'X)^{-1}_{j,j})$  and hypothesis tests can be computed. The difficulty with just proceeding on this basis is that  $\sigma^2$  is not known—it must be estimated. This can be done by taking the sum of the squared residuals and dividing, not by  $n$ , but by  $n - k$ , to get an unbiased estimate. The division by  $n - k$  is known as ‘adjusting for degrees of freedom.’ This adjustment for unknown  $\sigma^2$  results in a battery of exact finite sample tests for our ideal conditions:  $t$  and  $F$  tests are used in place of their ‘normal’ counterparts (i.e. tests that are based on the normal and  $\chi^2$  distributions.) Because these finite sample tests are valid only under ideal conditions and have closely related asymptotic counterparts, we will not study them in detail but give formulas for them subsequently.

## Ordinary Least Squares as an Estimator of the Conditional Mean

If  $y = X\beta + \varepsilon$  and  $E(\varepsilon|X) = 0$ , then  $X\beta$  is  $E(y|X)$ . We certainly know what this means under a DGP interpretation. Does OLS produce the conditional mean always?

Let  $E(y|X) = \mu(X)$ . This is typically (i.e. in wide but not complete generality) a well defined object;  $\mu(X)$  may or may not be linear in  $X$ .

## Ordinary Least Squares as an Estimator of the Conditional Mean (2)

We need to understand that  $\mu(X)$  minimizes the average of the squared deviations of  $y$  from  $\mu(X)$ .

$$\mu(X) = E(y|X) = \int y f(y|X) dy$$

$$Q(\mu(X)) = \int [y - \mu(X)]^2 f(y|X) dy = \int (y^2 - 2\mu(X)y + [\mu(X)]^2) f(y|X) dy$$

Picking  $\mu(X)$  to minimize the expected squared residual means minimizing this last quantity; differentiate it with respect to  $\mu(X)$  to get the FOC:

$$\begin{aligned} \frac{\partial Q}{\partial \mu(X)} &= \int (-2y + 2\mu(X)) f(y|X) dy = 0 \\ \int \mu(X) f(y|X) dy &= \int y f(y|X) dy \\ \mu(X) \int f(y|X) dy &= \int y f(y|X) dy \\ \mu(X) &= \frac{\int y f(y|X) dy}{\int f(y|X) dy} = E(y|X) \end{aligned}$$

## Ordinary Least Squares as an Estimator of the Conditional Mean: Discussion

We have just demonstrated that the conditional mean is the number that minimizes the expected (or average) squared deviation; it is an implication of  $f(y, X)$ . (The derivation of course was just a notational adornment of the demonstration that the 'simple' mean of a scalar random variable minimizes the average squared deviation.)

If  $\mu(X)$  is indeed  $X\beta$ , when we apply least squares, we 'get' the conditional mean. By applying least squares, we show that our intention is to estimate the conditional mean.

Is there something else we could do in the context of the simple linear model? Sure. We could minimize the sum of the absolute values of the deviations, in which case we would be aiming for the conditional median. (This is covered under 'quantile regression.')

## Practical Regression Analysis (under Ideal Conditions)

We now turn to applications of regression analysis under ideal conditions, postponing the discussion of the implications of the failure or relaxation of the ideal conditions. Many of the tricks we learn now have analogs under relaxed conditions. These ‘tricks’ are concerned with how to make the linear model and its relatives address ‘practical’ problems.

In our first regression we implicitly assumed that *each* year of education had the same marginal percentage effect on wages. Suppose we define the ‘dummy’ variables  $ed10$ ,  $ed11$ , ...,  $ed18$  which take the value 1 if the observation has 10, 11, ..., 18 years of education respectively, and zero otherwise.

If we do this we obtain:

Call:

```
lm(formula = lwage ~ ed10 + ed11 + ed12 + ed13 + ed14 + ed15 +  
    ed16 + ed17 + ed18 + exper)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.228176	0.138237	45.054	< 2e-16	***
ed10TRUE	-0.050020	0.141203	-0.354	0.723241	
ed11TRUE	0.115968	0.138660	0.836	0.403177	
ed12TRUE	0.178594	0.126922	1.407	0.159728	
ed13TRUE	0.310547	0.133813	2.321	0.020517	*
ed14TRUE	0.389017	0.135026	2.881	0.004055	**
ed15TRUE	0.478936	0.140341	3.413	0.000671	***
ed16TRUE	0.515503	0.131464	3.921	9.46e-05	***
ed17TRUE	0.600010	0.143027	4.195	2.99e-05	***
ed18TRUE	0.575781	0.138209	4.166	3.39e-05	***
exper	0.021231	0.003376	6.290	4.90e-10	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3932 on 924 degrees of freedom

Multiple R-Squared: 0.1378, Adjusted R-squared: 0.1284

F-statistic: 14.76 on 10 and 924 DF, p-value: < 2.2e-16

## Practical Regression Analysis (2)

We can see that different levels of education have different levels of wages, for a given level of experience. But it is difficult to calculate by sight whether the succeeding increments are equal to the previous increments. It is even harder (well, impossible) to compute the standard errors of these increments. If we redefine  $ed10$ ,  $ed11$ , ...,  $ed18$  to take the value 1 if the observation *at least* 10, 11, ..., 18 years of education respectively, we obtain:

Call:

```
lm(formula = lwage ~ ed10 + ed11 + ed12 + ed13 + ed14 + ed15 +  
    ed16 + ed17 + ed18 + exper)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.228176	0.138237	45.054	< 2e-16	***
ed10TRUE	-0.050020	0.141203	-0.354	0.72324	
ed11TRUE	0.165987	0.089659	1.851	0.06444	.
ed12TRUE	0.062626	0.063219	0.991	0.32213	
ed13TRUE	0.131953	0.047892	2.755	0.00598	**
ed14TRUE	0.078471	0.061911	1.267	0.20530	
ed15TRUE	0.089918	0.073778	1.219	0.22324	
ed16TRUE	0.036567	0.066826	0.547	0.58438	
ed17TRUE	0.084507	0.070181	1.204	0.22885	
ed18TRUE	-0.024229	0.081158	-0.299	0.76536	
exper	0.021231	0.003376	6.290	4.9e-10	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3932 on 924 degrees of freedom

Multiple R-Squared: 0.1378, Adjusted R-squared: 0.1284

F-statistic: 14.76 on 10 and 924 DF, p-value: < 2.2e-16

## The Lessons of the Education Example

There are several lessons of this example. The first is: make the dummy variables do the work for you. Many hypotheses can be addressed by defining dummies or interactions with dummies. A second lesson is that when the hypothesis of interest is directly addressed by an estimated coefficient, the standard error of the coefficient allows you to test the hypothesis of interest more easily—often at sight.

Generally speaking, if we take  $X$  and posmultiply it by a  $(k \times k)$  nonsingular matrix  $W$ , then we get an equivalent model with the coefficients ‘scrambled’ by  $W^{-1}$ , as is pretty evident by writing  $y = (XW)(W^{-1}\beta) + \varepsilon = Z\gamma + \varepsilon$ ; it is easy to confirm that OLS estimates transform equivalently. A little exercise: in our example, what was  $W$ ? Do you think I defined the second set of dummies by knowing  $W^{-1}$ ?