Chapter 2

Least Squares Estimation

Reference: Verbeek (2012) 2 and 4; Greene (2018) 2-4.
More advanced material is denoted by a star (*). It is not required reading.

2.1 Least Squares: The Optimization Problem and Its Solution

2.1.1 Simple Regression

The simplest regression model is

$$y_t = \beta_0 + \beta_1 x_t + u_t \text{, where } \mathrm{E}\, u_t = 0 \text{ and } \mathrm{Cov}(x_t, u_t) = 0, \tag{2.1}$$

where we can observe (have data on) the dependent variable y_t and the regressor x_t but not the residual u_t. In principle, the residual should account for all the movements in y_t that we cannot explain by x_t. The subscript t refers to observation t, which could represent period t (when data is a time series) or investor t (when data is a cross-section). In the latter case, it is common to instead use i as subscript.

Remark 2.1 *(On notation) These notes sometimes use alternative notations for the regression equation, for instance, $y_t = \alpha + \beta x_t + u_t$ (as is typical in CAPM regressions) or $y_i = a + bx_i + u_i$.*

Notice the two very important assumptions: (*i*) the mean of the residual is zero; and (*ii*) the residual is not correlated with the regressor, x_t. This basically says that the residual is pure noise. In contrast, if the average of u_t was non-zero, then $\beta_0 + \beta_1 x_t$ would get the general level of y_t wrong. Also, if x_t and u_t were correlated, then the best guess of y_t based on x_t would not be $\beta_0 + \beta_1 x_t$.

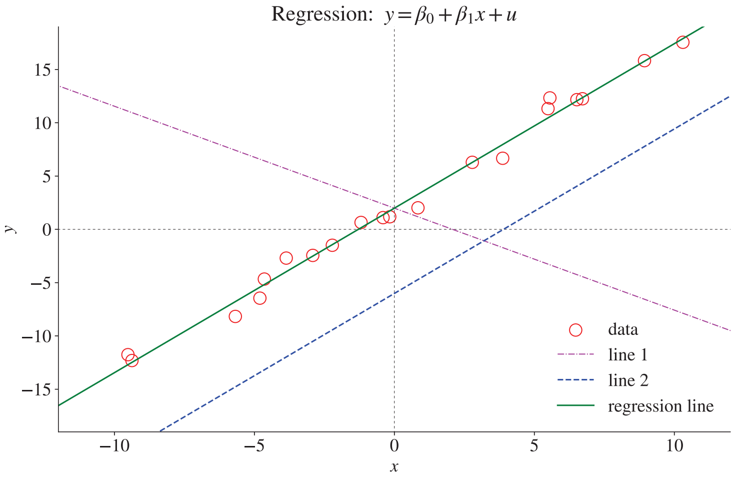


Figure 2.1: Example of OLS

Suppose you do not know β_0 or β_1, and that you have a sample of data: y_t and x_t for $t = 1, ..., T$. The LS estimator of β_0 and β_1 minimizes the loss function

$$\textstyle\sum_{t=1}^{T}(y_t - b_0 - b_1 x_t)^2 = (y_1 - b_0 - b_1 x_1)^2 + (y_2 - b_0 - b_1 x_2)^2 + ... \tag{2.2}$$

by choosing b_0 and b_1 to make the loss function value as small as possible. The objective is thus to pick values of b_0 and b_1 in order to make the model fit the data as closely as possible—where close is taken to be a small variance of the unexplained part (the residual). See Figures 2.1–2.2 for illustrations. (Least squares is only one of many possible ways to estimate regression coefficients. We will discuss other methods later on.)

Remark 2.2 *Note that β_i is the true (unobservable) value which we estimate to be $\hat{\beta}_i$. Whereas β_i is an unknown (deterministic) number, $\hat{\beta}_i$ is a random variable since it is calculated as a function of the random sample of y_t and x_t. We use b_i as an argument in the loss function (so we contemplate different values of b_i) —and the optimal value is clearly $\hat{\beta}_i$.*

Remark 2.3 *(First order condition for minimizing a differentiable function). We want to find the value of b in the interval $b_{low} \leq b \leq b_{high}$, which makes the value of the differentiable function $f(b)$ as small as possible. The answer is b_{low}, b_{high}, or a value of b where $df(b)/db = 0$. See Figure 2.3.*

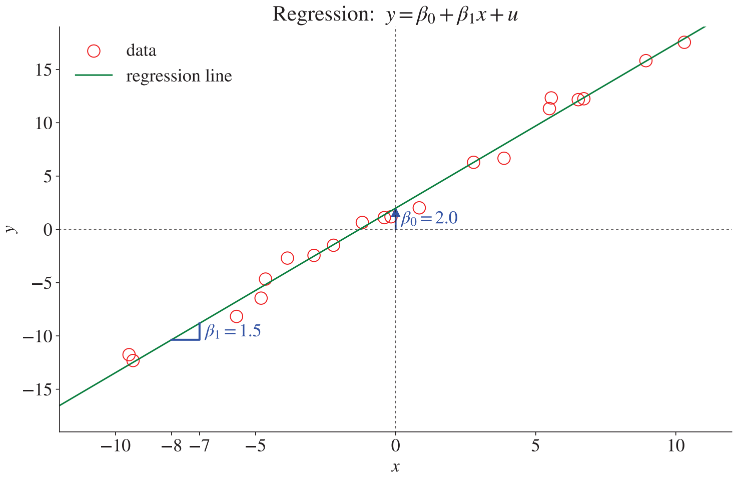


Figure 2.2: Example of OLS

The first order conditions for a minimum are that the derivatives of this loss function with respect to b_0 and b_1 should be zero. Notice that

$$\frac{\partial}{\partial b_0}(y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)1 \tag{2.3}$$

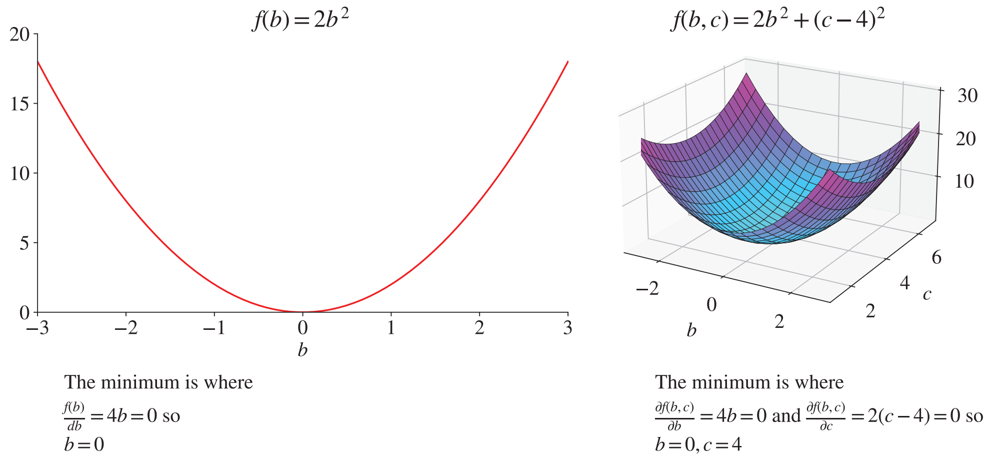
$$\frac{\partial}{\partial b_1}(y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)x_t. \tag{2.4}$$

Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the values of (b_0, b_1) where the derivatives are zero (that is, $(\hat{\beta}_0, \hat{\beta}_1)$ are the optimal values)

$$\frac{\partial}{\partial \beta_0}\textstyle\sum_{t=1}^{T}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2\sum_{t=1}^{T}1(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0 \tag{2.5}$$

$$\frac{\partial}{\partial \beta_1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2\sum_{t=1}^{T}x_t(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0, \tag{2.6}$$

which are two equations in two unknowns ($\hat{\beta}_0$ and $\hat{\beta}_1$), which must be solved simultaneously. These equations show that both the constant and x_t should be *orthogonal* to the fitted residuals, $\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t$. This is indeed a defining feature of LS and can be seen as the sample analogues of the assumptions in (2.1) that $\mathrm{E}\, u_t = 0$ and $\mathrm{Cov}(x_t, u_t) = 0$. To see this, note that (2.5) says that the sample average of $\hat{u}_t$ should be zero. Similarly, (2.6) says that the sample cross moment of $\hat{u}_t$ and x_t (that is, $\sum_{t=1}^{T}\hat{u}_t x_t / T$) should also be zero, which implies that the sample covariance is zero as well since $\hat{u}_t$ has a zero



The minimum is where
$\frac{f(b)}{db} = 4b = 0$ so
$b = 0$

The minimum is where
$\frac{\partial f(b,c)}{\partial b} = 4b = 0$ and $\frac{\partial f(b,c)}{\partial c} = 2(c-4) = 0$ so
$b = 0, c = 4$

Figure 2.3: Quadratic loss function. Subfigure a: 1 coefficient; Subfigure b: 2 coefficients

sample mean (see Remark 2.4).

Remark 2.4 *(Cross moments and covariance) A covariance is defined as*

$$\begin{aligned}\operatorname{Cov}(x, y) &= \mathrm{E}[(x - \mathrm{E}\, x)(y - \mathrm{E}\, y)] \\ &= \mathrm{E}(xy - x \,\mathrm{E}\, y - y \,\mathrm{E}\, x + \mathrm{E}\, x \,\mathrm{E}\, y) \\ &= \mathrm{E}\, xy - \mathrm{E}\, x \,\mathrm{E}\, y - \mathrm{E}\, y \,\mathrm{E}\, x + \mathrm{E}\, x \,\mathrm{E}\, y \\ &= \mathrm{E}\, xy - \mathrm{E}\, x \,\mathrm{E}\, y.\end{aligned}$$

If $\mathrm{E}\, x = 0$ *or* $\mathrm{E}\, y = 0$*, then* $\operatorname{Cov}(x, y) = \mathrm{E}\, xy$*. When* $x = y$*, then we get* $\operatorname{Var}(x) = \mathrm{E}\, x^2 - (\mathrm{E}\, x)^2$*. These results hold for sample moments too.*

When the means of y and x are zero, then we know that intercept is zero ($\beta_0 = 0$). In this case, (2.6) with $\hat{\beta}_0 = 0$ immediately gives

$$\begin{aligned}\textstyle\sum_{t=1}^{T} x_t y_t &= \hat{\beta}_1 \textstyle\sum_{t=1}^{T} x_t x_t \text{ or} \\ \hat{\beta}_1 &= \frac{\sum_{t=1}^{T} x_t y_t / T}{\sum_{t=1}^{T} x_t x_t / T}. \end{aligned} \tag{2.7}$$

In this case, the coefficient estimator is the sample covariance (recall: means are zero) of y_t and x_t, divided by the sample variance of the regressor x_t (this statement is actually true even if the means are not zero and a constant is included on the right hand side—just more tedious to show it).

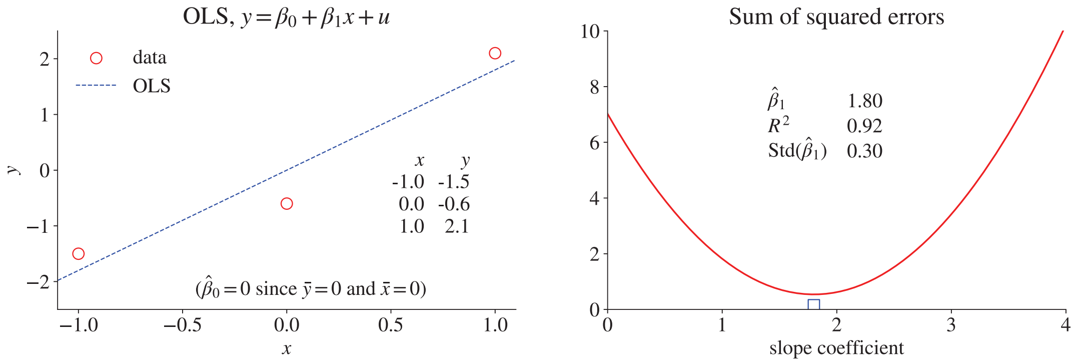


Figure 2.4: Example of OLS estimation

Empirical Example 2.5 *(CAPM regressions) See Table 2.1 and Figure 2.5 for CAPM regressions for two industry portfolios. The betas clearly differ.*

	HiTec	Utils
constant	−0.08	0.23
	(−0.57)	(1.64)
market return	1.25	0.51
	(36.53)	(13.36)
R^2	0.75	0.32
obs	612	612

Table 2.1: CAPM regressions, monthly returns, %, US data 1970:01-2020:12. Numbers in parentheses are t-stats.

Example 2.6 *(Simple regression) Consider the simple regression model (PSLS1). Suppose we have the following data*

t	x	y
1	−1	−1.5
2	0	−0.6
3	1	2.1

To calculate the LS estimate according to (2.7) we note that

$$\sum_{t=1}^{T} x_t x_t = (-1)^2 + 0^2 + 1^1 = 2 \textit{ and}$$
$$\sum_{t=1}^{T} x_t y_t = (-1)(-1.5) + 0(-0.6) + 1 \times 2.1 = 3.6$$

This gives

$$\hat{\beta}_1 = \frac{3.6}{2} = 1.8.$$

The fitted residuals are

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix} - 1.8 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ -0.6 \\ 0.3 \end{bmatrix}.$$

The fitted residuals indeed obey the first order condition (2.6) since

$$\sum_{t=1}^{T} x_t \hat{u}_t = (-1) \times 0.3 + 0(-0.6) + 1 \times 0.3 = 0.$$

See Figure 2.4 for an illustration.

Example 2.7 *Using the same data as in Example 2.6 we can also calculate the sums of squared residuals for different values of the slope coefficient. With* $\beta_1 = 1.6$ *we get*

t	u_t	u_1^2
1	$-1.5 - \mathbf{1.6} \times (-1) = 0.1$	0.01
2	$-0.6 - \mathbf{1.6} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{1.6} \times 1 = 0.5$	0.25
sum	0	0.62

With $\beta = 1.8$ *and* $\beta = 2.0$ *we instead get*

t	u_t	u_1^2
1	$-1.5 - \mathbf{1.8} \times (-1) = 0.3$	0.09
2	$-0.6 - \mathbf{1.8} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{1.8} \times 1 = 0.3$	0.09
sum	0	0.54

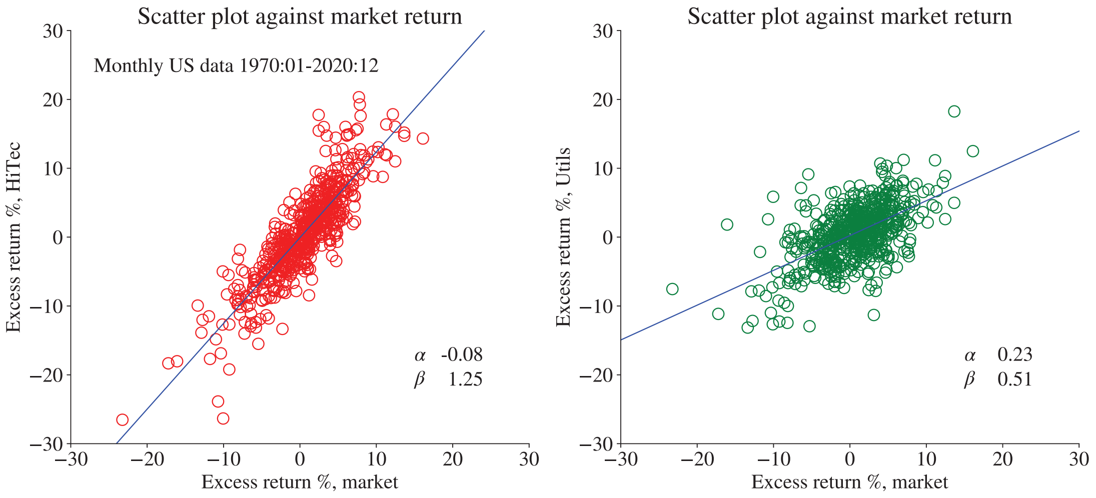


Figure 2.5: Scatter plot against market return

t	u_t	u_1^2
1	$-1.5 - \mathbf{2.0} \times (-1) = 0.5$	0.25
2	$-0.6 - \mathbf{2.0} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{2.0} \times 1 = 0.1$	0.01
sum	0	0.62

Among these alternatives, $\beta = 1.8$ *has the lowest sum of squared residuals (it is actually the optimum). See Figure 2.4.*

2.1.2 Multiple Regression

All the previous results still hold in a multiple regression—with suitable reinterpretations of the notation.

Consider the linear model

$$\begin{aligned} y_t &= x_{1t}\beta_1 + x_{2t}\beta_2 + \cdots + x_{kt}\beta_k + u_t \\ &= x_t'\beta + u_t, \end{aligned} \tag{2.8}$$

where y_t and u_t are scalars, x_t a $k \times 1$ vector, and β is a $k \times 1$ vector of the true coefficients. In this expression, one of the elements of x_t is typically a constant equal to one (and the intercept is its coefficient).

Remark 2.8 *(On notation) These notes typically denote a vector of regression coefficients by β. The distinction from the $y_t = \alpha + \beta x_t + u_t$ notation sometimes used for simple regressions should be clear from the context.*

Least squares minimizes the sum of the squared fitted residuals

$$\sum_{t=1}^{T} (y_t - x_t' b)^2, \tag{2.9}$$

by choosing the vector b. The first order conditions (zero derivatives) hold at the (optimal) values $\hat{\beta}$, and can then be written

$$\mathbf{0}_{kx1} = \sum_{t=1}^{T} x_t (y_t - x_t' \hat{\beta}) \text{ or } \sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{T} x_t x_t' \hat{\beta}. \tag{2.10}$$

Solve this as

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t. \tag{2.11}$$

If the regressors are orthogonal (for instance, $\Sigma x_{1t} x_{2t} = 0$) then the results from the multiple regression (2.31) are the same as those from a series of simple regressions: y_t regressed on x_{1t}, y_t regressed on x_{2t}, etc. (This is easy to see since in this case $\Sigma x_t x_t'$ is a diagonal matrix which carries over to the inverse.) This is an unlikely case, unless the regressors have been pre-processed to indeed be orthogonal.

Remark 2.9 *(Matrix notation*) Let X be a $T \times k$ matrix where row t is filled with the elements of x_t and let Y be a $T \times 1$ where element t is y_t. Then, $X'X = \Sigma_{t=1}^{T} x_t x_t'$ and $X'Y = \Sigma_{t=1}^{T} x_t y_t$, so (2.11) can also be written $\hat{\beta} = (X'X)^{-1} X'Y$.*

Example 2.10 *(OLS with 2 regressors) With 2 regressors ($k = 2$) denoted x_{1t} and x_{2t},*

$$x_t y_t = \begin{bmatrix} x_{1t} y_t \\ x_{2t} y_t \end{bmatrix} \text{ and } x_t x_t' = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \begin{bmatrix} x_{1t} & x_{2t} \end{bmatrix} = \begin{bmatrix} x_{1t} x_{1t} & x_{1t} x_{2t} \\ x_{2t} x_{1t} & x_{2t} x_{2t} \end{bmatrix}.$$

This means that (2.10) is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} x_{1t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \\ x_{2t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \end{bmatrix}$$

and (2.11) is

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left(\sum_{t=1}^{T} \begin{bmatrix} x_{1t} x_{1t} & x_{1t} x_{2t} \\ x_{2t} x_{1t} & x_{2t} x_{2t} \end{bmatrix} \right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} x_{1t} y_t \\ x_{2t} y_t \end{bmatrix}.$$

Example 2.11 *(OLS with constant and one more regressor) In Example 2.10, let $x_{1t} = 1$. The first order conditions are then*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \textstyle\sum_{t=1}^{T} \begin{bmatrix} y_t - \hat{\beta}_1 - x_{2t}\hat{\beta}_2 \\ x_{2t}(y_t - \hat{\beta}_1 - x_{2t}\hat{\beta}_2) \end{bmatrix}.$$

The first line mean that $\hat{\beta}_1 = \bar{y}_t - \bar{x}_{2t}\hat{\beta}_2$ (since dividing 0 by T is still 0). Using this in the second line to replace $\hat{\beta}_1$ and noticing that it does not matter if the term outside the parenthesis is x_{2t} or $x_{2t} - \bar{x}_{2t}$ (since the term in parenthesis is zero on average) gives $\Sigma(x_{2t} - \bar{x}_{2t})[(y_t - \bar{y}_t) - (x_{2t} - \bar{x}_{2t})\hat{\beta}_2] = 0$. We can then solve as $\hat{\beta}_2 = \Sigma(x_{2t} - \bar{x}_{2t})(y_t - \bar{y}_t)/\Sigma(x_{2t} - \bar{x}_{2t})^2$, which is the sample covariance of x_{2t} and y_t divided by the sample variance of x_{2t} (divide both numerator and denominator by T to see this).

Example 2.12 *(Regression with an intercept and slope) Suppose we have the following data:*

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix}, x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{and } x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This is clearly the same as in Example 2.6, except that we allow for an intercept (which turns out to be zero in this particular example). The notation we need to solve this problem is the same as for a general multiple regression. Therefore, calculate the following:

$$\begin{aligned} \textstyle\sum_{t=1}^{T} x_t x_t' &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \textstyle\sum_{t=1}^{T} x_t y_t &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-1.5) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-0.6) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2.1 \\ &= \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix} + \begin{bmatrix} -0.6 \\ 0 \end{bmatrix} + \begin{bmatrix} 2.1 \\ 2.1 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 3.6 \end{bmatrix} \end{aligned}$$

To calculate the LS estimate, notice that the inverse of the $\sum_{t=1}^{T} x_t x_t'$ is

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix},$$

which can be verified by

$$\begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The LS estimate is therefore

$$\begin{aligned} \hat{\beta} &= \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t \\ &= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 3.6 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 1.8 \end{bmatrix}. \end{aligned}$$

Remark 2.13 *(The Frisch-Waugh-Lovell theorem*) Let split up x_t into the* ***vectors*** *x_{1t} and x_{2t} and write (2.8) as $y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + u_t$. First, regress y_t on x_{1t} and get the residuals $\tilde{e}_1$. Second, regress x_{2t} on x_{1t} and get the residuals $\tilde{x}_{2t}$. Third, regress $\tilde{e}_1$ on $\tilde{x}_{2t}$. This gives the same estimates as β_2 from the multiple regression of y_t on both x_{1t} and x_{2t}. (The proof is a straightforward reshuffling of the first order conditions, see, for instance, Greene (2018) 3.) The perhaps most common application of this is when x_{1t} contains various dummy variables (for instance, for different cross-sectional units) and x_{2t} are the variables of key interest. It can then be convenient to apply this 3-step approach. This is used in the fixed effects estimator for panel data.*

2.1.3 Least Squares: Goodness of Fit

The quality of a regression model is often measured in terms of its ability to explain the movements of the dependent variable.

Let $\hat{y}_t$ be the fitted (predicted) value of y_t. For instance, with (2.1) it would be $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$. If a constant is included in the regression (or the means of y and x are zero), then a check of the *goodness of fit* of the model is given by the fraction of the variation in

y_t that is explained by the model

$$R^2 = \frac{\text{Var}(\hat{y}_t)}{\text{Var}(y_t)} = 1 - \frac{\text{Var}(\hat{u}_t)}{\text{Var}(y_t)}, \tag{2.12}$$

which can also be rewritten as the squared correlation of the actual and fitted values

$$R^2 = \text{Corr}(y_t, \hat{y}_t)^2. \tag{2.13}$$

Notice that we must have constant in regression (unless both y_t and x_t have zero means) for R^2 to make sense.

Example 2.14 *(R^2) From Example 2.6 we have* $\text{Var}(\hat{u}_t) = 0.18$ *and* $\text{Var}(y_t) = 2.34$*, so*

$$R^2 = 1 - 0.18/2.34 \approx 0.92.$$

See Figure 2.4.

Proof. (of (2.12)–(2.13)) Write the regression equation as

$$y_t = \hat{y}_t + \hat{u}_t,$$

where hats denote fitted values. Since $\hat{y}_t$ and $\hat{u}_t$ are uncorrelated (always true in OLS—provided the regression includes a constant), we have

$$\text{Var}(y_t) = \text{Var}(\hat{y}_t) + \text{Var}(\hat{u}_t).$$

R^2 is defined as the fraction of $\text{Var}(y_t)$ that is explained by the model

$$R^2 = \frac{\text{Var}(\hat{y}_t)}{\text{Var}(y_t)} = \frac{\text{Var}(y_t) - \text{Var}(\hat{u}_t)}{\text{Var}(y_t)} = 1 - \frac{\text{Var}(\hat{u}_t)}{\text{Var}(y_t)}.$$

Equivalently, we can rewrite R^2 by noting that

$$\text{Cov}\left(y_t, \hat{y}_t\right) = \text{Cov}\left(\hat{y}_t + \hat{u}_t, \hat{y}_t\right) = \text{Var}\left(\hat{y}_t\right).$$

Use this in the denominator of R^2 and multiply by $\text{Cov}\left(y_t, \hat{y}_t\right) / \text{Var}\left(\hat{y}_t\right) = 1$

$$R^2 = \frac{\text{Cov}\left(y_t, \hat{y}_t\right)^2}{\text{Var}(y_t)\,\text{Var}\left(\hat{y}_t\right)} = \text{Corr}\left(y_t, \hat{y}_t\right)^2.$$

■

To understand this result, suppose that x_t has no explanatory power, so R^2 should be zero. How does that happen? Well, if x_t is uncorrelated with y_t, then $\hat{\beta}_1 = 0$. As a consequence $\hat{y}_t = \hat{\beta}_0$, which is a constant. This means that R^2 in (2.12) is zero, since the fitted residual has the same variance as the dependent variable ($\hat{y}_t$ captures nothing of the movements in y_t). Similarly, R^2 in (2.13) is also zero, since a constant is always uncorrelated with anything else (as correlations measure comovements around the means).

Remark 2.15 *(R^2 from simple regression*) Suppose $\hat{y}_t = \beta_0 + \beta_1 x_t$, so (2.13) becomes*

$$R^2 = \frac{\text{Cov}\,(y_t, \beta_0 + \beta_1 x_t)^2}{\text{Var}(y_t)\,\text{Var}(\beta_0 + \beta_1 x_t)} = \frac{\text{Cov}\,(y_t, x_t)^2}{\text{Var}(y_t)\,\text{Var}(x_t)} = \text{Corr}(y_t, x_t)^2.$$

The R^2 can never decrease as we add more regressors, which might make it attractive to add more and more regressors. To avoid that, some researchers advocate using an ad hoc punishment for many regressors, $\bar{R}^2 = 1 - (1 - R^2)(T - 1)/(T - k)$, where k is the number of regressors (including the constant). This measure can be negative.

Empirical Example 2.16 *(CAPM regressions) See Table 2.1 for CAPM regressions for two industry portfolios where the R^2 values clearly differ. This is seen also from the dispersion around the regression line in Figure 2.5.*

2.2 Missing Data

It is often the case that some data is missing For instance, we may not have data on regressor 3 for observation $t = 7$. If data is *missing in a random way*, then we can simply exclude (y_t, x_t) for the t with some missing data. In contrast, if data is missing in a non-random way (for instance, depending on the value of y_{it}), then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

Remark 2.17 *(Replacing missing values with 0*) Instead of excluding (y_t, x_t) for the t with some missing data, we could set $(y_t, x_t) = (0, \mathbf{0}_K)$. This would not change the estimates, but it could lead to the wrong standard errors unless we are careful (see below for details).*