Chapter 2
Simple Linear Regression

2.1 Simple linear model

The simple linear regression model shows how one known dependent variable is determined by a single explanatory variable (regressor). Is is written as:

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{2.1}$$

The subscript i refers to the observation $i = 1, 2, \ldots n$, and Y_i is the dependent variable. We break down Y_i into two components, the deterministic (nonrandom) component $\beta_1 + \beta_2 X_i$ and the stochastic (random) component u_i. The explanatory variable is X_i and the population parameters we want to estimate are given by intercept β_1 and the slope β_2. The term u_i is the disturbance term. Figure 2.1 shows a graphical representation of the problem. The regression line $Y_i = \beta_1 + \beta_2 X_i + u_i$ is shown as the upward sloping blue line. Only a single observation point at (X_i, Y_i) is illustrated. We can see how for this observation i, we break down Y_i into the disturbance term u_i given by the vertical distance between Y_i and $\hat{Y}_i$ and the height of the regression line at point X_i, given by $\beta_1 + \beta_2 X_i$.

2.2 Least squares regression

The main idea in econometric analysis is to estimate the parameters β_1 and β_2. The most popular estimator for these population parameters is the Ordinary Least Squares (OLS) estimator. Let the OLS estimators of β_1 and β_2 be b_1 and b_2, respectively. Then, the fitter regression equation is:

$$Y_i = b_1 + b_2 X_i + e_i. \tag{2.2}$$

The difference between Equations 2.1 and 2.2 is that the first correspond to the population, while the second is the sample counterpart. The idea in the OLS es-

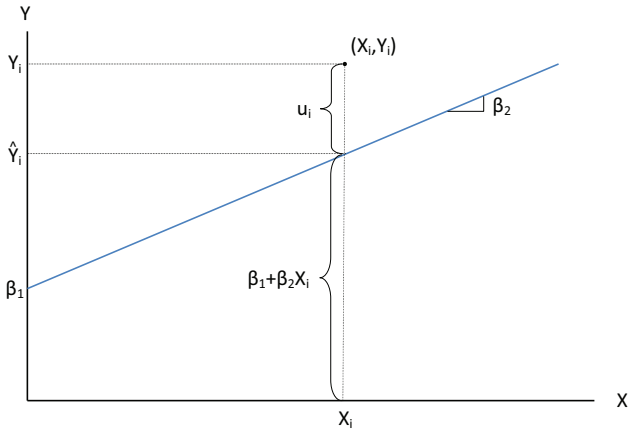

Fig. 2.1 Regression line $Y_i = \beta_1 + \beta_2 X_i + u_i$.

timator is simple, we want to pick values for the intercept b_1 and slope b_2 coefficients that are as close as possible to the actual data points. That is, we want to e_i ($e_i = Y_1 - b_1 - b_2 X_i$) to be small. Because some of the e_i are positive and some are negative, we will first square them to have all positive numbers. Then, to take into account all data points we will sum across all observations. That is how our objective is to pick b_1 and b_2 to minimize the following residual sum of squares:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2. \tag{2.3}$$

This minimization exercise yields the OLS estimators:

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{2.4}$$

for the slope coefficient, and

$$b_1 = \bar{Y} - b_2 \bar{X} \tag{2.5}$$

for the intercept. The derivation of the least squares coefficient estimators (Equations 2.4 and 2.5) has the following steps. We start with the regression equation:

$$Y_i = b_1 + b_2 X_i + e_i$$
$$\hat{Y}_i = b_1 + b_2 X_i$$

For observation i we obtain the residual, then square it and finally sum across all observations to obtain the residual sum of squares:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ e_i^2 &= (Y_i - \hat{Y}_i)^2 \\ \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \end{aligned} \tag{2.6}$$

The coefficients b_1 and b_2 are chosen to minimize the residuals sum of squares:

$$\begin{aligned} &\min_{b_1, b_2} \quad \textstyle\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &\min_{b_1, b_2} \textstyle\sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2 \end{aligned} \tag{2.7}$$

The first order necessary condition are:

$$-2 \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i) = 0 \qquad \text{w.r.t.} \qquad b_1 \tag{2.8}$$

$$-2 \sum_{i=1}^{n} X_i (Y_i - b_1 - b_2 X_i) = 0 \qquad \text{w.r.t.} \qquad b_2 \tag{2.9}$$

Dividing Equation 2.9 by n and working through some math we obtain the OLS estimators for the constant:

$$b_1 = \bar{Y} - b_2 \bar{X}.$$

Plugging this result into Equation 2.9 we obtain:

$$b_2 = \frac{\sum_{i=0}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^{n} (X_i - \bar{X})^2}.$$

2.3 Interpretation of the regression coefficients

If the estimated regression equation is given by:

$$\widehat{wage}_i = 4.64 + 0.09 exper_i, \tag{2.10}$$

where *wage* is the hourly wage measured in dollars, and *exper* is the number of years of experience, then the interpretation of the slope coefficient is the following:

$$\frac{\Delta wage}{\Delta exper} = 0.09.$$

Therefore, if the change in the number of years of experience is one, $\Delta exper$, then the change in the hourly wage in dollars is given by $\Delta wage = 0.09$. In words, an additional year of experience will increase your hourly wage by 0.09 dollars (or 9 cents). For the interpretation of the intercept, just consider the case where someone has not experience, $exper = 0$. Then, this person's predicted wage will be 4.64 dollars.

If the estimated regression equation takes the form:

$$\widehat{\log wage}_i = 1.38 + 0.02 exper_i, \tag{2.11}$$

where the $\log wage$ is the natural logarithm of *wage*, then the interpretation is different. Here, if the number of years of experience increases by one, the wage increases by 2% (0.02$\times$ 100 percent). Finally, for the folowing estimated equation:

$$\widehat{\log wage}_i = 0.98 + 0.26 \log exper_i. \tag{2.12}$$

A one percent increase in *exper* will increase *wage* by 0.25 percent. The 0.26 is interpreted as an elasticity.

2.4 Goodness of fit

How good is the regression equation in explaining the variation in variable Y? First we need a way to measure the total variation in Y. Let's try the sum of squared deviations about the sample mean of Y. That is,

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{2.13}$$

Now, let's start with a simple equality:

$$Y_i - \bar{Y} = Y_i - \bar{Y}.$$

If we add and subtract $\hat{Y}_i$ on the right hand side of the above equality, we have

$$\begin{aligned} Y_i - \bar{Y} &= Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i \\ Y_i - \bar{Y} &= (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \end{aligned}$$

Squaring both sides of the equation and then summing across all observations i we obtain:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2.14}$$

$$TSS = ESS + RSS. \tag{2.15}$$

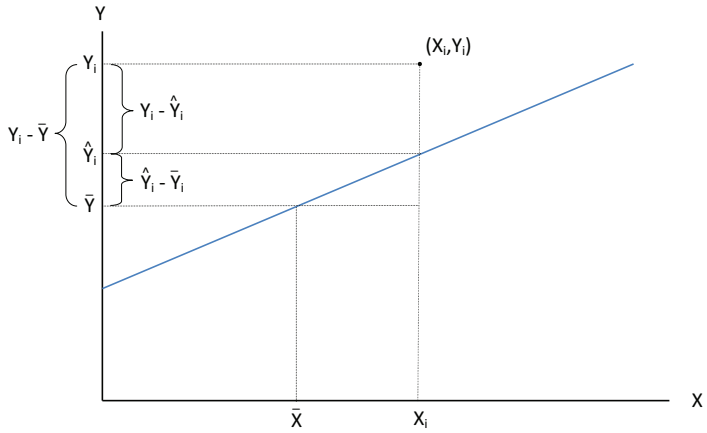

Fig. 2.2 Decomposition of $\hat{Y}_i - \bar{Y}$.

Notice that the sum of deviations from the mean is zero, that is why there are only two components on the right hand side. The TSS is the Total Sum of Squares, as presented in Equation 2.13. The first term on the right hand side is ESS, the Explained Sum of Squares, and the second term on the right hand side is the RRS, Residual Sum of Squares. This decomposition of the variable Y into two components can be appreciated in Figure 2.2. For every observation Y_i in the sample, the distance between Y_i and $\bar{Y}$ can be decomposed in two, the part that the regression equation can explain, $\hat{Y}_i - \bar{Y}$, and the part that the regression equation cannot explain, $Y_i - \hat{Y}_i$.

What is the proportion of the variation in Y that is explain by the regression equation? We just need to divide Equation 2.15 by TSS and define the ratio of ESS to TSS as the proportion of the explained variation in Y, the R^2:

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \tag{2.16}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.17}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2.18}$$

The R^2 is a number between zero and one, being higher when the model explains more of the variation in Y. Figures 2.3 and 2.4 illustrate how the regression line explain the variation in Y when the R^2 is low and high, respectively.

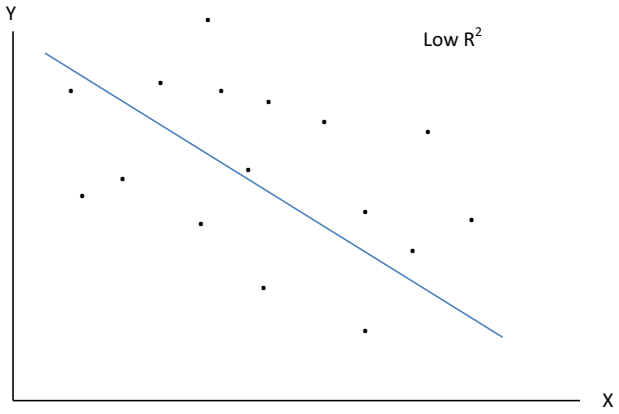

Fig. 2.3 Low R^2.

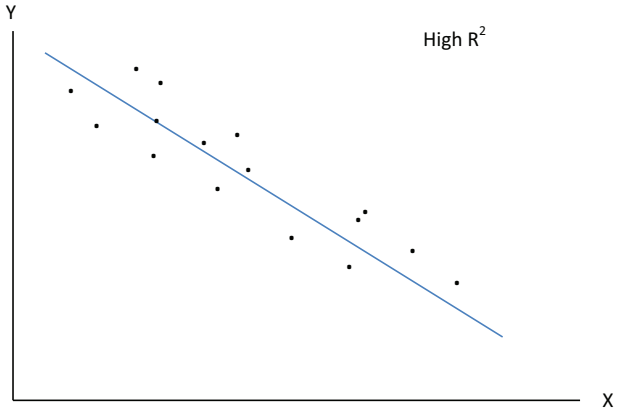

Fig. 2.4 High R^2.