

## The Simple Regression: Measures of Fit

- The technique of Ordinary Least Squares guarantees that the estimated regression line is the best-fitting line that can be drawn through the data. In the sense that it has the smallest possible sum of squared residuals.
- Although it is the best-fitting line, whether it fits well or not so well depends on the data. If the data points in the scatter-plot seem to lie close to some line, the best-fitting line will fit very well; but if the points are widely dispersed, the best-fitting line will not fit very well.
- We will go over two statistics that allow us to quantify how well the regression fits the data.
- Remember that we start from a set of data on  $Y$  and  $X$ , the estimated regression line yields a set of fitted values, or predictions, for the actual  $Y_i$  values. These are given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \quad (1)$$

The associated errors of fit are given by the residuals

$$e_i = Y_i - \hat{Y}_i, \quad (2)$$

these residuals will serve as the basis for the two measures of fit.

## The Standard Error of the Regression: SER

- The  $n$  residuals we just computed in (2) constitute a data variable in itself,  $e$ . And it can be described with the tools we have learned regarding any other variable. We are going to use a statistic that uses  $e$  to compute the typical error of fit of the estimated regression.
- The **standard error of the regression** (SER) is defined by

$$SER = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} \quad (3)$$

and its value is interpreted as a measure of the typical error of fit.

- Notice that to compute the SER we first have to find the estimates of the coefficients  $\beta_0$  and  $\beta_1$ , and then compute the fitted or predicted values of our dependent variable and then compute the  $n$  residuals.
- The units of measurement of the SER are always the same as those of the dependent variable  $Y$ , because each residual is equal to the actual value  $Y_i$  minus the fitted value,  $\hat{Y}_i$ .
- In general the smaller the SER the better is the fit of the estimated regression to the scatter of data. We do need, however, a value to compare it to. We usually compare it with the fitted value,  $\hat{Y}_i$ , or the mean of  $Y$ .

## The Coefficient of Determination: $R^2$ .

- This measure yields a pure, dimensionless number. It varies between zero and 1, with a higher value indicating a better fit.
- We are going to go through several steps of a formal development because these are useful for understanding the interpretation we make. Write

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) = (\hat{Y}_i - \bar{Y}) + e_i \quad (4)$$

which says that for any observation the total deviation of the observed  $Y_i$  from the mean is equal to the deviation of the fitted value from the mean plus the error of fit.

- Given the data, it is the regression that determines or ‘explains,’ the fitted value, while the error of fit is left ‘unexplained’. Giving new names to the items in the decomposition in (4), we can say that the total deviation of  $Y_i$  from  $\bar{Y}$  is equal to the **explained deviation** plus the **unexplained deviation**.

- If we square the leftmost and the rightmost portions of the equation (4), we get

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + e_i^2 + 2e_i(\hat{Y}_i - \bar{Y}) \quad (5)$$

if we add up the  $n$  equations like this one that hold for each observation  $i$ , then

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum e_i(\hat{Y}_i - \bar{Y}), \quad (6)$$

it can be shown that the last term on the right-hand side of this last equation is equal to zero, so it remains

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2. \quad (7)$$

- All the terms in this equation are either positive or zero. The expression in the left hand side is called the **total variation** in  $Y$ , because it is the sum of squares of the total deviations identified above.
- The first expression on the right-hand side is called the **explained variation** and the second is called the **unexplained variation**. Thus we can say that the total variation in  $Y$  is equal to the variation that is explained by the regression plus the unexplained variation.

- Rearranging the last equation,(7), and dividing by the total variation, we get

$$1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \quad (8)$$

in words

$$1 - \frac{\text{unexplained variation}}{\text{total variation in } Y} = \frac{\text{explained variation}}{\text{total variation in } Y} \quad (9)$$

the right hand side of this equation is the ratio of the explained variation to the total variation. Both the numerator and denominator are positive, and since the numerator is a component of the denominator the ratio can take on values only between zero and 1. This ratio can serve as a measure of goodness of fit.

- The left-hand side of the expression is often easier to compute, this is why we will use the left-hand side of the definition but give the interpretation of the right-hand side.
- The **coefficient of determination**, which is usually denoted by  $R^2$  and read as ‘R-squared,’ is defined by

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{explained variation}}{\text{total variation in } Y}. \quad (10)$$

- Our interpretation of this ratio is that it measures ‘the proportion of the total variation in  $Y$  that is explained by the regression.’ Since the regression model explains the values of  $Y$  on the basis of the given values of  $X$ , we might say instead that  $R^2$  measures ‘the proportion of the total variation in  $Y$  that is explained by  $X$ ’.
- In the data I will provide in P.S.4. on saving and income, the  $R^2$  is around 0.443. That means that in the regression of family saving on family income, family income explains about 44% of the variation in family saving. This might not seem as much but it is quite good.
- In general in time series data the  $R^2$  are quite large, mainly because  $Y$  and  $X$  have common trends. By contrast in cross-section data it is usually low since there is no trend, and individual behavior has quite a lot of variation difficult to explain.
- Notice that in general computer packages directly report  $R^2$ , but they also usually report the total sum of squares (total variation, or TSS), the explained sum of squares (ESS), and the residual sum of squares (RSS), and you can compute  $R^2$  yourself.