TABLE 2.2 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

Algebraic Properties of OLS Statistics

There are several useful algebraic properties of OLS estimates and their associated statistics. We now cover the three most important of these.

(1) The sum, and therefore the sample average of the OLS residuals, is zero. Mathematically,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \qquad [2.30]$$

This property needs no proof; it follows immediately from the OLS first order condition (2.14), when we remember that the residuals are defined by $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. In other words, the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are *chosen* to make the residuals add up to zero (for any data set). This says nothing about the residual for any particular observation i.

(2) The sample covariance between the regressors and the OLS residuals is zero. This follows from the first order condition (2.15), which can be written in terms of the residuals as

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \qquad [2.31]$$

The sample average of the OLS residuals is zero, so the left-hand side of (2.31) is proportional to the sample covariance between x_i and $\hat{u}_i$.

(3) The point $(\bar{x},\bar{y})$ is always on the OLS regression line. In other words, if we take equation (2.23) and plug in $\bar{x}$ for x, then the predicted value is $\bar{y}$. This is exactly what equation (2.16) showed us.

EXAMPLE 2.7 WAGE AND EDUCATION

For the data in WAGE1.RAW, the average hourly wage in the sample is 5.90, rounded to two decimal places, and the average education is 12.56. If we plug $educ = 12.56$ into the OLS regression line (2.27), we get $\widehat{wage} = -0.90 + 0.54(12.56) = 5.8824$, which equals 5.9 when rounded to the first decimal place. These figures do not exactly agree because we have rounded the average wage and education, as well as the intercept and slope estimates. If we did not initially round any of the values, we would get the answers to agree more closely, but to little useful effect.

Writing each y_i as its fitted value, plus its residual, provides another way to interpret an OLS regression. For each i, write

$$y_i = \hat{y}_i + \hat{u}_i. \qquad [2.32]$$

From property (1), the average of the residuals is zero; equivalently, the sample average of the fitted values, $\hat{y}_i$, is the same as the sample average of the y_i, or $\bar{\hat{y}} = \bar{y}$. Further, properties (1) and (2) can be used to show that the sample covariance between $\hat{y}_i$ and $\hat{u}_i$ is zero. Thus, we can view OLS as decomposing each y_i into two parts, a fitted value and a residual. The fitted values and residuals are uncorrelated in the sample.

Define the **total sum of squares** (**SST**), the **explained sum of squares** (**SSE**), and the **residual sum of squares** (**SSR**) (also known as the sum of squared residuals), as follows:

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2. \qquad [2.33]$$

$$\text{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \qquad [2.34]$$

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \qquad [2.35]$$

SST is a measure of the total sample variation in the y_i; that is, it measures how spread out the y_i are in the sample. If we divide SST by $n - 1$, we obtain the sample variance of y, as discussed in Appendix C. Similarly, SSE measures the sample variation in the $\hat{y}_i$ (where we use the fact that $\bar{\hat{y}} = \bar{y}$), and SSR measures the sample variation in the $\hat{u}_i$. The total variation in y can always be expressed as the sum of the explained variation and the unexplained variation SSR. Thus,

$$\text{SST} = \text{SSE} + \text{SSR}. \qquad [2.36]$$

Proving (2.36) is not difficult, but it requires us to use all of the properties of the summation operator covered in Appendix A. Write

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \text{SSR} + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \text{SSE}.
\end{aligned}$$

Now, (2.36) holds if we show that

$$\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) = 0. \qquad [2.37]$$

But we have already claimed that the sample covariance between the residuals and the fitted values is zero, and this covariance is just (2.37) divided by $n-1$. Thus, we have established (2.36).

Some words of caution about SST, SSE, and SSR are in order. There is no uniform agreement on the names or abbreviations for the three quantities defined in equations (2.33), (2.34), and (2.35). The total sum of squares is called either SST or TSS, so there is little confusion here. Unfortunately, the explained sum of squares is sometimes called the "regression sum of squares." If this term is given its natural abbreviation, it can easily be confused with the term "residual sum of squares." Some regression packages refer to the explained sum of squares as the "model sum of squares."

To make matters even worse, the residual sum of squares is often called the "error sum of squares." This is especially unfortunate because, as we will see in Section 2.5, the errors and the residuals are different quantities. Thus, we will always call (2.35) the residual sum of squares or the sum of squared residuals. We prefer to use the abbreviation SSR to denote the sum of squared residuals, because it is more common in econometric packages.

Goodness-of-Fit

So far, we have no way of measuring how well the explanatory or independent variable, x, explains the dependent variable, y. It is often useful to compute a number that summarizes how well the OLS regression line fits the data. In the following discussion, be sure to remember that we assume that an intercept is estimated along with the slope.

Assuming that the total sum of squares, SST, is not equal to zero—which is true except in the very unlikely event that all the y_i equal the same value—we can divide (2.36) by SST to get $1 = \text{SSE/SST} + \text{SSR/SST}$. The ***R*-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 \equiv \text{SSE/SST} = 1 - \text{SSR/SST}. \qquad [2.38]$$

R^2 is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the *fraction of the sample variation in y that is explained by x*. The second equality in (2.38) provides another way for computing R^2.

From (2.36), the value of R^2 is always between zero and one, because SSE can be no greater than SST. When interpreting R^2, we usually multiply it by 100 to change it into a percent: $100 \cdot R^2$ is the *percentage of the sample variation in y that is explained by x*.

If the data points all lie on the same line, OLS provides a perfect fit to the data. In this case, $R^2 = 1$. A value of R^2 that is nearly equal to zero indicates a poor fit of the OLS line: very little of the variation in the y_i is captured by the variation in the $\hat{y}_i$ (which all lie on the OLS regression line). In fact, it can be shown that R^2 is equal to the *square* of the sample correlation coefficient between y_i and $\hat{y}_i$. This is where the term "R-squared" came from. (The letter R was traditionally used to denote an estimate of a population correlation coefficient, and its usage has survived in regression analysis.)

EXAMPLE 2.8 CEO SALARY AND RETURN ON EQUITY

In the CEO salary regression, we obtain the following:

$$\widehat{salary} = 963.191 + 18.501\, roe \qquad [2.39]$$

$$n = 209, R^2 = 0.0132.$$

We have reproduced the OLS regression line and the number of observations for clarity. Using the *R*-squared (rounded to four decimal places) reported for this equation, we can see how much of the variation in salary is actually explained by the return on equity. The answer is: not much. The firm's return on equity explains only about 1.3% of the variation in salaries for this sample of 209 CEOs. That means that 98.7% of the salary variations for these CEOs is left unexplained! This lack of explanatory power may not be too surprising because many other characteristics of both the firm and the individual CEO should influence salary; these factors are necessarily included in the errors in a simple regression analysis.

In the social sciences, low *R*-squareds in regression equations are not uncommon, especially for cross-sectional analysis. We will discuss this issue more generally under multiple regression analysis, but it is worth emphasizing now that a seemingly low *R*-squared does not necessarily mean that an OLS regression equation is useless. It is still possible that (2.39) is a good estimate of the ceteris paribus relationship between *salary* and *roe*; whether or not this is true does *not* depend directly on the size of *R*-squared. Students who are first learning econometrics tend to put too much weight on the size of the *R*-squared in evaluating regression equations. For now, be aware that using *R*-squared as the main gauge of success for an econometric analysis can lead to trouble.

Sometimes, the explanatory variable explains a substantial part of the sample variation in the dependent variable.

EXAMPLE 2.9 VOTING OUTCOMES AND CAMPAIGN EXPENDITURES

In the voting outcome equation in (2.28), $R^2 = 0.856$. Thus, the share of campaign expenditures explains over 85% of the variation in the election outcomes for this sample. This is a sizable portion.

2.4 Units of Measurement and Functional Form

Two important issues in applied economics are (1) understanding how changing the units of measurement of the dependent and/or independent variables affects OLS estimates and (2) knowing how to incorporate popular functional forms used in economics into regression analysis. The mathematics needed for a full understanding of functional form issues is reviewed in Appendix A.