We can use MATLAB or other software packages to do regression analysis. For example, the following MATLAB code can be used to obtain the estimated regression line in Example 8.31.

```
x=[1;2;3;4];
x0=ones(size(x));
y=[3;4;8;9];
beta = regress(y,[x0,x]);
```

Coefficient of Determination (R-Squared):

Let's look again at the above model for regression. We wrote

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is a $N(0, \sigma^2)$ random variable independent of X. Note that, here, X is the only variable that we observe, so we estimate Y using X. That is, we can write

$$\hat{Y} = \beta_0 + \beta_1 X.$$

The error in our estimate is

$$Y - \hat{Y} = \epsilon.$$

Note that the randomness in Y comes from two sources: X and ϵ. More specifically, if we look at $\mathrm{Var}(Y)$, we can write

$$\mathrm{Var}(Y) = \beta_1^2 \mathrm{Var}(X) + \mathrm{Var}(\epsilon) \quad \text{(since } X \text{ and } \epsilon \text{ are assumed to be independent).}$$

The above equation can be interpreted as follows. The total variation in Y can be divided into two parts. The first part, $\beta_1^2 \mathrm{Var}(X)$, is due to variation in X. The second part, $\mathrm{Var}(\epsilon)$, is the variance of error. In other words, $\mathrm{Var}(\epsilon)$ is the variance left in Y after we know X. If the variance of error, $\mathrm{Var}(\epsilon)$, is small, then Y is close to $\hat{Y}$, so our regression model will be successful in estimating Y. From the above discussion, we can define

$$\rho^2 = \frac{\beta_1^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)}$$

as the *portion* of variance of Y that is explained by variation in X. From the above discussion, we can also conclude that $0 \leq \rho^2 \leq 1$. More specifically, if ρ^2 is close to 1, Y can be estimated very well as a linear function of X. On the other hand if ρ^2 is small, then the variance of error is large and Y cannot be accurately estimated as a linear function of X. Since $\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$, we can write

$$\rho^2 = \frac{\beta_1^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)} = \frac{[\mathrm{Cov}(X,Y)]^2}{\mathrm{Var}(X)\mathrm{Var}(Y)} \qquad (8.6)$$

The above equation should look familiar to you. Here, ρ is the correlation coefficient that we have seen before. Here, we are basically saying that if X and Y are highly correlated (i.e., $\rho(X,Y)$ is large), then Y can be well approximated by a linear function of X, i.e., $Y \approx \hat{Y} = \beta_0 + \beta_1 X$.

We conclude that ρ^2 is an indicator showing the strength of our regression model in estimating (predicting) Y from X. In practice, we often do not have ρ but we have the observed pairs (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). We can estimate ρ^2 from the observed data. We show it by r^2 and call it *R-squared* or *coefficient of determination*.

Coefficient of Determination

For the observed data pairs, (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), we define **coefficient of determination**, r^2 as

$$r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}},$$

where

$$s_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad s_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2, \quad s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

We have $0 \leq r^2 \leq 1$. Larger values of r^2 generally suggest that our linear model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is a good fit for the data.

Two sets of data pairs are shown in Figure 8.12. In both data sets, the values of the y_i's (the heights of the data points) have considerable variation. The data points shown in (a) are very close to the regression line. Therefore, most of the variation in y is explained by the regression formula. That is, here, the $\hat{y}_i$'s are relatively close to the y_i's, so r^2 is close to 1. On the other hand, for the data shown in (b), a lot of variation in y is left unexplained by the regression model. Therefore, r^2 for this data set is much smaller than r^2 for the data set in (a).

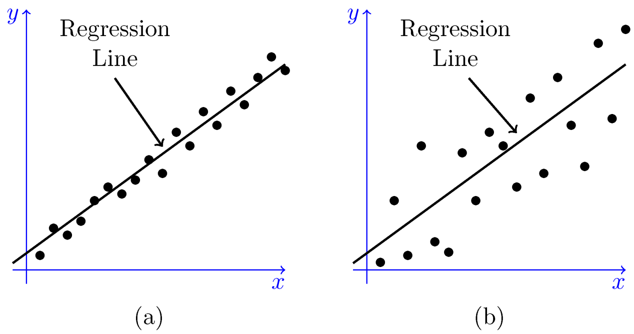

Figure 8.12 - The data in (a) results in a high value of r^2, while the data shown in (b) results in a low value of r^2.

Example 8.32

For the data in Example 8.31, find the coefficient of determination.

Solution

In Example Example 8.31, we found

$$s_{xx} = 5, \quad s_{xy} = 11.$$

We also have

$$s_{yy} = (3-6)^2 + (4-6)^2 + (8-6)^2 + (9-6)^2 = 26.$$

We conclude

$$r^2 = \frac{11^2}{5 \times 26} \approx 0.93$$
