Lecture 6: The Coefficient of Determination

1. Evaluating Regression Quality

It's easy to accept the OLS coefficients without question, but it's important to think carefully about the quality of a regression when interpreting the results. Empirical studies typically spend quite a lot of time considering such issues. Some things to consider are:

1. Is there a good theoretical foundation for the regression equation?
2. How well does the regression fit?
3. Is the sample large enough and representative?
4. Is OLS the best method?
5. Do the regression results support your hypothesis?
6. Are any important variables omitted?
7. Is the functional form appropriate?
8. Are there econometric problems with the regression?

We will discuss model fit in the remainder, but it is important to remember that fit is only one of many considerations.

2. The Coefficient of Determination

The most common measure of the overall fit of a regression is the \textit{coefficient of determination}, denoted $R^2$. This measure summarizes the fit of a single regression independently, but it is also useful to compare the fit of a collection of regressions with different combinations of included independent variables.

First of all, we need some notation. Let TSS denote the total variation in the data, the \textit{total sum of squares}, defined as the variance of $Y_i$:

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

This variance can be factored into two components: one capturing deviations in $Y_i$ from the fitted values $\hat{Y}_i$ and another capturing deviations of the fitted values from the mean $\bar{Y}$. In
Econ 444: Elementary Econometrics    Lecture 6: The Coefficient of Determination

We have seen RSS before, as OLS is defined by the values of \( \hat{\beta}_k \) which minimize the sum of squared residuals, or RSS. The ESS term is the *explained sum of squares*, which is the part of TSS which is “explained” by the fitted regression model. The part that is “unexplained” is captured by the residuals and falls in the RSS component. Usually we prefer to write this *decomposition of variance* as

\[
\text{TSS} = \text{ESS} + \text{RSS},
\]

which is more like our regression equation, with the dependent variable on the left hand side, followed by the deterministic or explained portion of the regression and the residuals on the right hand side.

Now, the coefficient of determination is defined as

\[
R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.
\]

It measures the fraction of the total variation in \( Y \) that is explained by the included regressors, as measured by the ESS. Measures such as \( R^2 \) are called *goodness of fit* measures.

A higher value of \( R^2 \) indicates that the regression fits the data better. In other words, \( X \) can explain more of the variation in \( Y \).

Note that since the OLS coefficients minimize RSS, OLS will always have the highest \( R^2 \) of any linear model with the same regressors.

Since TSS, RSS, and ESS are all nonnegative, and since \( \text{ESS} \leq \text{TSS} \), we know that \( R^2 \) must be between 0 and 1. Values of \( R^2 \) close to 1 indicate a good overall fit, while values close to zero indicate a poor overall fit. If \( R^2 = 0 \), it means that the \( X_i \)'s fail to explain the \( Y_i \)'s better than a line at \( \bar{Y} \) could. At the other extreme, if \( R^2 = 1 \), then each and every point falls exactly on the regression line, meaning that all of the residuals are equal to zero, along with the RSS.

If, for example, \( X \) and \( Y \) are uncorrelated, then \( R^2 = 0 \). In that case, \( \text{ESS} = 0 \) and \( \text{RSS} = \text{TSS} \). All of the variation in \( Y \) is explained by the residuals. In this case, the regression line would be \( \hat{Y} = \bar{Y} \), which would be the case if \( X \) was omitted from the regression.

Graphical examples:

- \( R^2 = 0 \),
- \( R^2 = 0.95 \),
- \( R^2 = 1 \).
Example 1. Recall our simple example dataset on stock price, $Y_i$, and trade volume, $X_i$, given in the table below.

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

The OLS coefficients we calculated before were $\hat{\beta}_0 = 8/3$ and $\hat{\beta}_1 = -1/6$. Using these coefficients yields the following predicted values:

$\hat{Y}_1 = \frac{8}{3} - \frac{1}{6} \cdot 1 = 2.5,$
$\hat{Y}_2 = \frac{8}{3} - \frac{1}{6} \cdot 4 = 2.0,$
$\hat{Y}_3 = \frac{8}{3} - \frac{1}{6} \cdot 1 = 2.5.$

The corresponding residuals are:

$e_1 = Y_1 - \hat{Y}_1 = -0.5,$
$e_2 = Y_2 - \hat{Y}_2 = 0,$
$e_3 = Y_3 - \hat{Y}_3 = 0.5.$

Now, we can calculate the TSS, ESS, and RSS. First, of all $\bar{Y} = \frac{7}{3}$. The total sum of squares is

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \left(2 - \frac{7}{3}\right)^2 + \left(2 - \frac{7}{3}\right)^2 + \left(3 - \frac{7}{3}\right)^2 = \frac{2}{3}.$$

The sum of squared residuals is

$$RSS = (-0.5)^2 + 0^2 + 0.5^2 = 0.5.$$

Finally, we have

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{1/2}{2/3} = 1 - \frac{3}{4} = 0.25.$$

3. Adjusted $R^2$

One fundamental problem with the definition of $R^2$ is that adding an additional independent variable to the regression, even a completely unrelated one, will always weakly increase the $R^2$ (meaning that it will either stay the same or increase but never decrease). To see this, recall that

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Adding an additional $X$ variable to the regression will not change the TSS but will (again, weakly) increase the ESS. Therefore, looking at $R^2$ alone can encourage the addition of too many explanatory variables.
Example 2. Recall our height-weight example and suppose we are considering adding another independent variable, the post office box number of each individual. Clearly this is irrelevant, but it could have the effect of increasing $R^2$! Suppose the estimated regression equation is

$$\text{WEIGHT}_i = 103.40 + 6.38 \cdot \text{HEIGHT}_i$$

with $R^2 = 0.74$. After including $BOX_i$, the post office box number, we have

$$\text{WEIGHT}_i = 102.35 + 6.36 \cdot \text{HEIGHT}_i + 0.02 \cdot BOX_i$$

with $R^2 = 0.75$. Adding $BOX_i$ increased the value of $R^2$ because in the sample, $BOX_i$ happened to be slightly correlated with $\text{WEIGHT}_i$, which in turn slightly increased the fit and decreased the RSS. Since $R^2 = 1 - \text{RSS}/\text{TSS}$, this increases $R^2$.

So, an econometrician using $R^2$ alone might settle on the second regression specification, but an econometrician using common sense would select the first.

As the previous example illustrates, $R^2$ helps us summarize the fit of a regression in a single number, but it cannot help us determine whether the addition of a regressor improves our regression in a meaningful sense. As a result, an alternative measure known as adjusted $R^2$ was developed, denoted $\bar{R}^2$.

Before defining the $\bar{R}^2$, we need to introduce some more notation. Recall that we use $n$ as the number of observations and $K$ as the number of regressors. Including the intercept, that means there are $K + 1$ coefficients to be estimated. The difference, $n - K - 1$, is known as the number of degrees of freedom.

Now, the adjusted $R^2$ is defined as

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n - K - 1)}{\text{TSS}/(n - 1)}.$$

This measure includes a “penalty” in the sense that increasing the number of regressors, $K$, decreases $\bar{R}^2$ unless the RSS decreases (ESS increases) enough to compensate. Intuitively, the adjusted $R^2$ measures the percentage of the variation in $Y$ around $\bar{Y}$ that is explained by the regressors, with an adjustment for the degrees of freedom.

As opposed to the standard $R^2$, $\bar{R}^2$ may either increase, decrease, or stay the same when an additional regressor is added. The direction of the change will depend on whether the fit improves enough to justify the decrease in the degrees of freedom. As with $R^2$, $\bar{R}^2$ is bounded above by 1.0. However, it may also be negative, while the minimum possible $R^2$ is 0.0.

Example 3. Returning to our stock price, trade volume example dataset, the adjusted $R^2$ is

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n - K - 1)}{\text{TSS}/(n - 1)} = 1 - \frac{\text{RSS}}{\text{TSS}} \cdot \frac{n - 1}{n - K - 1} = 1 - \frac{0.5}{2/3} \cdot \frac{3 - 1}{3 - 1 - 1} = 1 - \frac{3}{4} \cdot \frac{2}{1} = -0.5.$$
\[ \hat{\beta}_0 = -\frac{2}{3} \quad \text{and} \quad \hat{\beta}_1 = \frac{3}{20} \]

<table>
<thead>
<tr>
<th>( X_i )</th>
<th>( Y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
</tr>
</tbody>
</table>