Econometrics I - Goodness of fit

Ryan T. Godwin

University of Manitoba

Goodness of fit

Can measure "quality" of fitted regression model by the extent that it "explains" the sample variation for $\boldsymbol{y}$. Our LS model tries to explain the sample variance of $\boldsymbol{y}$, and our measure of fit tells how good a job the model does. Recall that the sample variance of $\boldsymbol{y}$ is:

$$\frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Or, we could just use

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

One estimator for variance has an $(n-1)$ in the denominator (above) while another just has n (not shown). Since we will be comparing the sample variance of two vectors with the same dimension $(n \times 1)$, the denominator will cancel out.
Two measures of fit that are often used in conjunction with LS are the R-squared R^2 and adjusted-R-squared $\bar{R}^2$. We'll see why $\bar{R}^2$ is usually better.

Coefficient of Determination - R^2

R^2 is the ratio of variance in $\boldsymbol{y}$ that can be explained using the model (the X variables and LS estimates) over the total variance in $\boldsymbol{y}$. Start by writing our measure of sample variance in matrix form. Measures of variability use the squared *deviations-from-means*, so we can use the M_i matrix:

$$\boldsymbol{y}' M_i \boldsymbol{y} = \boldsymbol{y}' M_i' M_i \boldsymbol{y} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1}$$

Recall that LS "decomposes" $\boldsymbol{y}$ into two components, fitted values and residuals:

$$\boldsymbol{y} = P_X \boldsymbol{y} + M_X \boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$$

We can take the sample variance of *both* sides of $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$. Start by taking the deviations in means of both sides:

$$M_i \boldsymbol{y} = M_i \widehat{\boldsymbol{y}} + M_i \boldsymbol{e} = M_i \widehat{\boldsymbol{y}} + \boldsymbol{e} \tag{2}$$

We have converted the components of the model into deviations from means.

Question: Why does $M_i e = e$?

Now, pre-multiply both sides of equation 2 by its own transpose:

$$\begin{aligned} \boldsymbol{y}' M_i \boldsymbol{y} &= \boldsymbol{y}' M_i' M_i \boldsymbol{y} = \left(M_i \widehat{\boldsymbol{y}} + \boldsymbol{e}\right)' \left(M_i \widehat{\boldsymbol{y}} + \boldsymbol{e}\right) \\ &= \widehat{\boldsymbol{y}}' M_i \widehat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} + 2e' M_i \widehat{y} \end{aligned} .$$

however,

$$e' M_i \widehat{y} = e' M_i{}' \widehat{y} = \left(M_i e\right)' \widehat{y} = e' \widehat{y} = e' X \left(X'X\right)^{-1} X' y = 0$$

We have "decomposed" the sample variance of $\boldsymbol{y}$ into two parts: that which is explained by the estimated model ($\widehat{\boldsymbol{y}}$), and that which is unexplained ($\boldsymbol{e}$).

Question: Why does $e' M_i \widehat{y} = 0$? Recall the rule for taking the variance of a sum of two random variables.

So, we have (recall that $\overline{\hat{y}} = \bar{y}$):

$$\begin{array}{ccccc} \boldsymbol{y}'M_i\boldsymbol{y} & = & \widehat{\boldsymbol{y}}'M_i\widehat{\boldsymbol{y}} & + & \boldsymbol{e}'\boldsymbol{e} \\ \sum_{i=1}^{n}(y_i-\bar{y})^2 & = & \sum_{i=1}^{n}(\widehat{y_i}-\bar{y})^2 & + & \sum_{i=1}^{n}e_i^2 \\ \text{TSS} & = & \text{ESS} & + & \text{RSS} \end{array}$$

Where TSS = "total sum of squares", ESS = "explained sum of squares", and RSS = "residual sum of squares". This lets us define the R-squared:

$$R^2 = \left(\frac{ESS}{TSS}\right) = 1 - \left(\frac{RSS}{TSS}\right)$$

R^2 is the portion of variance in the dependent variable that can be explained by the estimated model.

- The second equality in the definition of R^2 holds only if model includes an intercept.
- $R^2 = \left(\frac{ESS}{TSS}\right) \geq 0$
- $R^2 = 1 - \left(\frac{RSS}{TSS}\right) \leq 1$
- So, $0 \leq R^2 \leq 1$
- R^2 is *unitless*.

Question: What is the interpretation of $R^2 = 0$ and $R^2 = 1$?

R^2 increases when a regressor is added to the model

What happens if we add any regressor(s) to the model? Consider the population model:

$$y = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \tag{3}$$

then consider adding regressors to it:

$$y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{u} \tag{4}$$

Optimization problem A - apply LS to 4:

$$\min\left(\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}\right) \quad ; \quad \widehat{\boldsymbol{u}} = \boldsymbol{y} - X_1\boldsymbol{b}_1 - X_2\boldsymbol{b}_2$$

Optimization problem B - apply LS to 3:

$$\min\left(\boldsymbol{e}'\boldsymbol{e}\right) \quad ; \quad \boldsymbol{e} = \boldsymbol{y} - X_1\hat{\boldsymbol{\beta}}_1$$

Problem B is just Problem A, subject to the restriction: $\boldsymbol{\beta}_2 = \mathbf{0}$.
Minimized value in A must be $\leq$ minimized value in B. So, $\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} \leq \boldsymbol{e}'\boldsymbol{e}$.

- Adding any regressor(s) to the model cannot increase (and typically will decrease) the sum of squared residuals.
- So, adding any regressor(s) to the model cannot decrease (and typically will increase) the value of R^2.
- "Junk" variables could be added to the model to get the R^2 arbitrarily high.
- Means that R^2 is not really a very interesting measure of the "quality" of the regression model, in terms of explaining sample variability of the dependent variable.

For these reasons, we usually "adjust" the Coefficient of Determination.

Adjusted R-square: $\bar{R}^2$

Modify R^2:

$$R^2 = \left[1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}' M_i \boldsymbol{y}}\right]$$

to become:

$$\bar{R}^2 = \left[1 - \frac{\boldsymbol{e}'\boldsymbol{e}/(n-k)}{\boldsymbol{y}' M_i \boldsymbol{y}/(n-1)}\right]$$

We're adjusting for "degrees of freedom" in the numerator and denominator. Now, when a regressor is added to the model, $\bar{R}^2$ increases due to the improved "fit" ($\boldsymbol{e}'\boldsymbol{e}$ decreases), and decreases due to the penalty imposed by k. $\bar{R}^2$ only increases if the improvement in model fit due to the additional regressor dominates the penalty for adding another regressor.

"Degrees of freedom" are the number of independent pieces of information.

When we calculate $\bar{y}$, we lose one degree of freedom. That is, the sample y, together with $\bar{y}$, only contains $(n-1)$ independent pieces of information.

For example, if $\boldsymbol{y} = \{1, 3, z\}$, and $\bar{y} = 3$, then you know that $z = 5$. Similarly, in order to calculate $\boldsymbol{e}$, we first need to calculate the $(k \times 1)$ vector $\boldsymbol{b}$. Once we have $\boldsymbol{b}$, there are only $(n-k)$ pieces of independent information left in $\boldsymbol{e}$.

$\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{b}$. We estimate k parameters from the n data-points, before we can calculate $\boldsymbol{e}$. We have $(n-k)$ "degrees of freedom" associated with the fitted model.

Some final points about R^2 and $\bar{R}^2$:

- Possible for $\bar{R}^2 \leq 0$ (even with an intercept in the model).
- $\bar{R}^2$ can *increase* or *decrease* when we add regressors.
- In multiple regression, $\bar{R}^2$ will increase (decrease) if a variable is deleted, if and only if the associated t-statistic has *absolute value less than* (greater than) unity.
- If the model doesn't include an intercept, then $\text{TSS} \neq \text{ESS} + \text{RSS}$, and in this case there is no longer any guarantee that $0 \leq R^2 \leq 1$.
- Must be careful comparing R^2 and $\bar{R}^2$ values across models. For example:

$$\begin{aligned} \widehat{C}_i &= 0.5 + 0.8Y_i \quad ; \quad R^2 = 0.90 \\ \log\left(\widehat{C}_i\right) &= 0.2 + 0.75Y_i \quad ; \quad R^2 = 0.80 \end{aligned}$$

The sample variation is in *different units*.