

LECTURE 6

PROPERTIES OF \bar{R}^2 , MODEL MISSPECIFICATION, TEST OF STRUCTURAL CHANGE, DUMMY VARIABLES, FORECASTSProperties of \bar{R}^2

In this section, we show that \bar{R}^2 rises or falls when new regressors are added, depending on whether the F -statistic associated with them exceeds 1. This result holds regardless of how many regressors are added. Consider the unrestricted $(k + q)$ regression model and the restricted model with only k regressors:

$$\begin{aligned} \text{Unrestricted:} \quad & Y_i = \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \beta_{k+1} X_{i,k+1} + \dots + \beta_{k+q} X_{i,k+q} + U_i, \\ \text{Restricted:} \quad & Y_i = \beta_1 X_{i1} + \dots + \beta_k X_{ik} + U_i, \end{aligned}$$

where $q \geq 1$. Let RSS and RSS_r denote the unrestricted and restricted residual sums of squares, respectively. The corresponding adjusted coefficients of determination are given by

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{n-1}{n-k-q} \frac{RSS}{TSS}, \\ \bar{R}_r^2 &= 1 - \frac{n-1}{n-k} \frac{RSS_r}{TSS}. \end{aligned}$$

Then

$$\begin{aligned} \bar{R}^2 - \bar{R}_r^2 &= \frac{n-1}{n-k} \frac{RSS_r}{TSS} - \frac{n-1}{n-k-q} \frac{RSS}{TSS} \\ &= \frac{n-1}{TSS} \left(\frac{RSS_r}{n-k} - \frac{RSS}{n-k-q} \right) \\ &= \frac{n-1}{TSS} \frac{RSS}{n-k} \left(\frac{RSS_r}{RSS} - \frac{n-k}{n-k-q} \right) \\ &= \frac{n-1}{TSS} \frac{RSS}{n-k} \left(\frac{RSS_r}{RSS} - 1 + 1 - \frac{n-k}{n-k-q} \right) \\ &= \frac{n-1}{TSS} \frac{RSS}{n-k} \left(\frac{RSS_r - RSS}{RSS} - \frac{q}{n-k-q} \right) \\ &= \frac{n-1}{TSS} \frac{RSS}{n-k} \frac{q}{n-k-q} \left(\frac{(RSS_r - RSS)/q}{RSS/(n-k-q)} - 1 \right). \end{aligned}$$

The desired result follows because

$$\frac{(RSS_r - RSS)/q}{RSS/(n-k-q)}$$

is the F -statistic associated with the null

$$H_0 : \beta_{k+1} = \dots = \beta_{k+q} = 0.$$

For comparison, the critical values of the F -distribution at conventional significance levels exceed 1. Consequently, model selection based on \bar{R}^2 can lead to the inclusion of irrelevant regressors. Specifically, \bar{R}^2 increases whenever $F > 1$, but $F > 1$ does not imply statistical significance at conventional levels. A variable that is statistically insignificant can still raise \bar{R}^2 .

Model misspecification

Exclusion of relevant regressors

Suppose the true model is

$$Y = X_1\beta_1 + X_2\beta_2 + U, \quad (1)$$

where X_1 is $n \times k_1$, X_2 is $n \times k_2$, $\beta_2 \neq 0$, and Assumptions (A1)–(A5) are satisfied with $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$. Suppose the econometrician runs a regression of Y on X_1 alone, either because X_2 is unavailable or because its relevance is not recognized.

Consider the properties of the LS estimator of β_1 :

$$\begin{aligned} \tilde{\beta}_1 &= (X_1^\top X_1)^{-1} X_1^\top Y \\ &= (X_1^\top X_1)^{-1} X_1^\top (X_1\beta_1 + X_2\beta_2 + U) \\ &= \beta_1 + (X_1^\top X_1)^{-1} X_1^\top X_2\beta_2 + (X_1^\top X_1)^{-1} X_1^\top U. \end{aligned}$$

By Assumption (A2),

$$\mathbb{E}[\tilde{\beta}_1 | X] = \beta_1 + (X_1^\top X_1)^{-1} X_1^\top X_2\beta_2. \quad (2)$$

In this case, the LS estimator of β_1 is *biased*, with bias $(X_1^\top X_1)^{-1} X_1^\top X_2\beta_2$. The bias vanishes if X_1 and X_2 are orthogonal with probability one, that is,

$$\Pr(X_1^\top X_2 = 0) = 1.$$

Consider next the conditional variance of $\tilde{\beta}_1$. By (2) and Assumption (A3),

$$\begin{aligned} \text{Var}(\tilde{\beta}_1 | X) &= (X_1^\top X_1)^{-1} X_1^\top \mathbb{E}[UU^\top | X] X_1 (X_1^\top X_1)^{-1} \\ &= \sigma^2 (X_1^\top X_1)^{-1}. \end{aligned}$$

Compare that to the variance of $\hat{\beta}_1$, the LS estimator of β_1 from the regression that includes both X_1 and X_2 :

$$\text{Var}(\hat{\beta}_1 | X) = \sigma^2 (X_1^\top M_2 X_1)^{-1},$$

where $M_2 = I_n - P_2$, and $P_2 = X_2 (X_2^\top X_2)^{-1} X_2^\top$. Consider first the difference

$$\begin{aligned} X_1^\top X_1 - X_1^\top M_2 X_1 &= X_1^\top P_2 X_1 \\ &\geq 0. \end{aligned}$$

The inequality follows because P_2 is symmetric and idempotent and therefore positive semidefinite. Consequently,

$$(X_1^\top X_1)^{-1} - (X_1^\top M_2 X_1)^{-1} \leq 0,$$

and

$$\text{Var}(\tilde{\beta}_1 | X) - \text{Var}(\hat{\beta}_1 | X) \leq 0.$$

Thus, $\tilde{\beta}_1$ has smaller variance than $\hat{\beta}_1$. The cost of including additional regressors in the correctly specified model is a larger variance for the coefficient estimates.

Under Assumption (A5), we obtain that

$$\tilde{\beta}_1 | X \sim N\left(\beta_1 + (X_1^\top X_1)^{-1} X_1^\top X_2\beta_2, \sigma^2 (X_1^\top X_1)^{-1}\right).$$

We now study the effect of misspecification on s^2 , the estimator of σ^2 . In this case,

$$s^2 = \frac{Y^\top M_1 Y}{n - k_1},$$

where k_1 is the number of columns in X_1 , and $M_1 = I_n - X_1 (X_1^\top X_1)^{-1} X_1^\top$. Since the true model is (1),

$$s^2 = \frac{(X_1\beta_1 + X_2\beta_2 + U)^\top M_1 (X_1\beta_1 + X_2\beta_2 + U)}{n - k_1},$$

and

$$\begin{aligned} \mathbb{E}[s^2 | X] &= \mathbb{E}\left[\frac{U^\top M_1 U}{n - k_1} | X\right] + 2\mathbb{E}\left[\frac{U^\top M_1 X_2 \beta_2}{n - k_1} | X\right] + \frac{\beta_2^\top X_2^\top M_1 X_2 \beta_2}{n - k_1} \\ &= \sigma^2 + \frac{\beta_2^\top X_2^\top M_1 X_2 \beta_2}{n - k_1} \\ &\geq \sigma^2. \end{aligned}$$

The bias persists even when X_1 and X_2 are orthogonal, since $M_1 X_2 = X_2$ in that case.

One consequence of excluding relevant variables is that tests and confidence intervals based on the misspecified model are invalid.

Inclusion of irrelevant variables

Suppose the true model is

$$Y = X_1\beta_1 + U.$$

However, the econometrician includes X_2 as well and estimates β_1 by

$$\tilde{\beta}_1 = (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 Y.$$

The LS estimator is unbiased under Assumption (A2):

$$\begin{aligned} \mathbb{E}[\tilde{\beta}_1 | X] &= \mathbb{E}\left[(X_1^\top M_2 X_1)^{-1} X_1^\top M_2 (X_1\beta_1 + U) | X\right] \\ &= \beta_1 + (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 \mathbb{E}[U | X] \\ &= \beta_1. \end{aligned}$$

Under Assumption (A3), its variance is given by the usual formula

$$\text{Var}(\tilde{\beta}_1 | X) = \sigma^2 (X_1^\top M_2 X_1)^{-1}.$$

However, $\tilde{\beta}_1$ is inefficient relative to the correctly specified estimator. By the Gauss–Markov theorem, the OLS estimator in the correctly specified model has the smallest variance among all linear unbiased estimators; as shown in the previous section, adding irrelevant regressors inflates the variance. Under Assumption (A5), we have

$$\tilde{\beta}_1 | X \sim N\left(\beta_1, \sigma^2 (X_1^\top M_2 X_1)^{-1}\right).$$

Consider now s^2 . Here,

$$s^2 = \frac{Y^\top M_X Y}{n - k_1 - k_2},$$

where $M_X = I_n - X (X^\top X)^{-1} X^\top$. Since $M_X X_1 = 0$, it follows that

$$s^2 = \frac{U^\top M_X U}{n - k_1 - k_2},$$

and

$$\mathbb{E}[s^2 | X] = \sigma^2.$$

Naturally, the usual tests and confidence intervals remain valid in the case of inclusion of irrelevant variables. However, the confidence regions for β_1 will be larger, and tests less powerful, compared to the correctly specified equation. The preceding two sections suggest a general-to-specific approach to model selection: start with the most general specification and eliminate irrelevant regressors using F -tests.

Test of structural change

Suppose that there are two regression models representing, for example, observations from two countries or two different time periods:

$$\begin{aligned} Y_1 &= X_1\beta_1 + U_1, \\ Y_2 &= X_2\beta_2 + U_2, \end{aligned}$$

where X_1 is $n_1 \times k$, X_2 is $n_2 \times k$, and $n_1 + n_2 = n$. Here (Y_1, X_1) and (Y_2, X_2) contain observations on the same variables but from *two different sub-samples*. The hypothesis of interest is whether the regression coefficients differ across sub-samples:

$$H_0 : \beta_1 = \beta_2. \quad (3)$$

To apply the F -test framework, we first combine the two equations into a single regression. Define

$$\begin{aligned} Y &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \\ U &= \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \\ X &= \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}, \\ \beta &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}. \end{aligned}$$

With these definitions, the unrestricted model takes the form

$$Y = X\beta + U, \quad (4)$$

which is the usual linear regression model. We assume that (4) satisfies Assumptions (A1)–(A5). In this framework, the restrictions given in (3) can be written as

$$R\beta = \begin{pmatrix} I_k & -I_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0.$$

We now derive the unrestricted RSS . The block-diagonal structure of X simplifies the computation:

$$\begin{aligned} (X^\top X)^{-1} &= \begin{pmatrix} X_1^\top X_1 & 0 \\ 0 & X_2^\top X_2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 \\ 0 & (X_2^\top X_2)^{-1} \end{pmatrix}. \end{aligned}$$

It follows that the projection annihilator M_X is also block-diagonal:

$$\begin{aligned} M_X &= I_n - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 \\ 0 & (X_2^\top X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1^\top & 0 \\ 0 & X_2^\top \end{pmatrix} \\ &= I_n - \begin{pmatrix} X_1 (X_1^\top X_1)^{-1} X_1^\top & 0 \\ 0 & X_2 (X_2^\top X_2)^{-1} X_2^\top \end{pmatrix} \\ &= \begin{pmatrix} I_{n_1} - X_1 (X_1^\top X_1)^{-1} X_1^\top & 0 \\ 0 & I_{n_2} - X_2 (X_2^\top X_2)^{-1} X_2^\top \end{pmatrix} \\ &= \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}. \end{aligned}$$

The unrestricted RSS equals the sum of the residual sums of squares from the two separate regressions:

$$\begin{aligned}\hat{U}^\top \hat{U} &= Y^\top M_X Y \\ &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}^\top \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= Y_1^\top M_1 Y_1 + Y_2^\top M_2 Y_2 \\ &= RSS_1 + RSS_2.\end{aligned}$$

The unrestricted model has $2k$ regression coefficients.

The restricted model imposes $\beta_1 = \beta_2$, so the two sub-samples share a common coefficient vector. It can be written as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta_1 + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}. \quad (5)$$

The restricted RSS is obtained by pooling the two sub-samples. Define

$$X_r = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

and

$$M_r = I_n - X_r (X_r^\top X_r)^{-1} X_r^\top.$$

The restricted RSS is given by

$$RSS_r = Y^\top M_r Y.$$

The test statistic follows from the general F -test formula, with $q = k$ restrictions and $n - 2k$ degrees of freedom in the denominator:

$$F = \frac{(RSS_r - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n - 2k)}.$$

Under H_0 , the statistic F follows an $F_{k, n-2k}$ distribution. One rejects the null of no structural change when

$$F > F_{k, n-2k, 1-\alpha}.$$

Dummy variables

Regression analysis often involves qualitative variables that cannot be quantified in the usual way. For example, one may be interested in studying the effects of gender, marital status, race, and religion on economic variables such as income or education. A common approach to quantifying such variables is to introduce artificial variables that indicate whether a particular quality is present. Suppose that a qualitative variable has m categories. For observations $i = 1, \dots, n$, define the *dummy variables* d_{ij} , $j = 1, \dots, m$, such that

$$d_{ij} = \begin{cases} 1, & \text{if observation } i \text{ belongs to category } j, \\ 0, & \text{otherwise.} \end{cases}$$

For example, let Y_i be the salary of individual i , and

$$d_{i1} = \begin{cases} 1, & \text{if male,} \\ 0, & \text{if female,} \end{cases}$$

$$d_{i2} = \begin{cases} 1, & \text{if female,} \\ 0, & \text{if male.} \end{cases}$$

Consider the regression

$$Y_i = \alpha_1 d_{i1} + \alpha_2 d_{i2} + X_i^\top \beta + U_i,$$

where X_i is the vector of other regressors, such as years of schooling and experience. In this case, α_1 and α_2 give the starting salary for men and women, respectively. Alternatively, one may consider the following specification:

$$Y_i = \alpha_0 + \alpha_1 d_{i1} + X_i^\top \beta + U_i.$$

In this case, the starting salary for women is α_0 , and the starting salary for men is $\alpha_0 + \alpha_1$. The coefficient α_1 gives the difference in starting salaries between male and female workers. One can test whether their starting salaries differ by testing the hypothesis $\alpha_1 = 0$.

One cannot include the intercept together with both dummy variables, since for all i

$$d_{i1} + d_{i2} = 1,$$

which violates Assumption (A4). The general rule is that if a categorical variable has m categories, one should include m dummies and no intercept, or $m - 1$ dummies with an intercept.

One may also allow for the effect of other regressors X_i to be different across categories. In the above example, this can be modeled by including the *interaction term* of X_i with the dummy variable d_{i1} :

$$Y_i = \alpha_0 + \alpha_1 d_{i1} + X_i^\top \beta + (d_{i1} X_i)^\top \delta + U_i.$$

Now the marginal effect of X_i is β for women and $\beta + \delta$ for men. One can test whether the model is different for men and women by testing $H_0 : \alpha_1 = 0, \delta = 0$.

Consider the test for structural change discussed in the previous section. Define

$$d_i = \begin{cases} 0, & \text{for } i = 1, \dots, n_1, \\ 1, & \text{for } i = n_1 + 1, \dots, n. \end{cases}$$

One can write the model for $i = 1, \dots, n$ as

$$Y_i = X_i^\top \beta_1 + (d_i X_i)^\top \delta + U_i,$$

or equivalently,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ X_2 \end{pmatrix} \delta + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}. \quad (6)$$

In this case, $\beta_2 = \beta_1 + \delta$, and the test of no structural change is equivalent to testing $H_0 : \delta = 0$. To show that the two approaches are equivalent, it suffices to show that the matrix of regressors in (4),

$$\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix},$$

spans the same linear space as that in (6):

$$\begin{pmatrix} X_1 & 0 \\ X_2 & X_2 \end{pmatrix}.$$

Forecasts

Consider again the classical normal linear regression model defined by Assumptions (A1)–(A5):

$$Y_i = X_i^\top \beta + U_i.$$

In this section, we discuss *forecasting* the dependent variable Y_i given a fixed k -vector of regressor values x_f , and constructing confidence intervals for such forecasts. There are two distinct forecasting problems: predicting the conditional mean $x_f^\top \beta$ (a point on the regression line), and predicting a new observation $Y_f = x_f^\top \beta + U_f$ (a point off the regression line). The latter involves an additional source of uncertainty from the unobserved error U_f , which widens the prediction interval. Let $\hat{\beta}$ be the LS estimator of β based on

the data $\{(Y_i, X_i) : i = 1, \dots, n\}$. The vector x_f need not coincide with any regressor value in the observed sample. Since

$$E[Y_i | X_i = x_f] = x_f^\top \beta,$$

it is natural to estimate the conditional expectation of the dependent variable, $E[Y_i | X_i = x_f]$, by

$$\widehat{Y}_f = x_f^\top \widehat{\beta}. \quad (7)$$

Here \widehat{Y}_f is the predicted value of a point *on the regression line*; since x_f is fixed, \widehat{Y}_f is random only through $\widehat{\beta}$. Using the results for $\widehat{\beta}$, we obtain

$$\widehat{Y}_f | X \sim N\left(x_f^\top \beta, \sigma^2 x_f^\top (X^\top X)^{-1} x_f\right).$$

The $(1 - \alpha)$ -level confidence interval for the conditional mean $x_f^\top \beta$ is

$$x_f^\top \widehat{\beta} \pm t_{n-k, 1-\alpha/2} \sqrt{s^2 x_f^\top (X^\top X)^{-1} x_f}.$$

Next, consider predicting a point *off the regression line*. Define

$$Y_f = x_f^\top \beta + U_f,$$

where the $(n + 1)$ -vector $(U^\top, U_f)^\top$ satisfies

$$\begin{pmatrix} U \\ U_f \end{pmatrix} | X \sim N(0, \sigma^2 I_{n+1}). \quad (8)$$

Since U_f is not predictable from X , the predicted value of Y_f is given by (7). Next, the *forecast error* is given by

$$\begin{aligned} \widehat{U}_f &= Y_f - x_f^\top \widehat{\beta} \\ &= U_f - x_f^\top (\widehat{\beta} - \beta). \end{aligned}$$

The result in (8) implies that

$$\widehat{U}_f | X \sim N\left(0, \sigma^2 + \sigma^2 x_f^\top (X^\top X)^{-1} x_f\right).$$

Therefore, the $(1 - \alpha)$ -level prediction interval for Y_f is

$$x_f^\top \widehat{\beta} \pm t_{n-k, 1-\alpha/2} \sqrt{s^2 \left(1 + x_f^\top (X^\top X)^{-1} x_f\right)}.$$