LECTURE 6

PROPERTIES OF $\overline{R}^2$, MODEL MISSPECIFICATION, TEST OF STRUCTURAL CHANGE, DUMMY VARIABLES, FORECASTS

Properties of $\overline{R}^2$

In this section, we will show that, when adding new regressors, $\overline{R}^2$ will rise/fall if the F-statistic associated with the added regressors is greater/less than 1, regardless of the number of the regressors added. Consider the unrestricted $(k+q)$ regression model and restricted model with only k regressors:

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

where $q \geq 1$. Let RSS and RSS_r denote unrestricted and restricted Residual Sum-of-Squares respectively. The corresponding adjusted coefficients of determination are given by

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

Next,

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

The desired results follows because

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

is the F-statistic associated with the null

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

For comparison, the critical values of F-distribution exceed 1. Consequently, model selection based on the adjusted coefficient determination can lead to inclusion of irrelevant regressors.

Model misspecification

Exclusion of relevant regressors

Suppose that the true model is given by

$$Y = X_1\beta_1 + X_2\beta_2 + U, \tag{1}$$

where X_1 is $n \times k_1$, X_2 is $n \times k_2$, $\beta_2 \neq 0$, and the Assumptions (A1)-(A5) are satisfied with $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$. Suppose that the econometrician runs a regression of Y on X_1 alone, either because X_2 is unavailable, or because he does not know that it should be included.

First, we study properties of the LS estimator of β_1:

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

By Assumption (A2),

$$E\left(\widetilde{\beta}_1 | X\right) = \beta_1 + (X_1'X_1)^{-1} X_1'X_2\beta_2. \tag{2}$$

In this case, the LS estimator of β_1 is *biased*, with the bias given by $(X_1'X_1)^{-1} X_1'X_2\beta_2$. The bias term disappears if X_1 and X_2 are orthogonal with probability 1, i.e.

$$P(X_1'X_2 = 0) = 1.$$

Next, consider the conditional variance of $\widetilde{\beta}_1$. Due to (2) and Assumption (A3),

$$\begin{aligned} Var\left(\widetilde{\beta}_1 | X\right) &= (X_1'X_1)^{-1} X_1' E(UU'|X) X_1 (X_1'X_1)^{-1} \\ &= \sigma^2 (X_1'X_1)^{-1}. \end{aligned}$$

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

$$Var\left(\widehat{\beta}_1 | X\right) = \sigma^2 (X_1'M_2X_1)^{-1},$$

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

$$\begin{aligned} X_1'X_1 - X_1'M_2X_1 &= X_1'P_2X_1 \\ &\geq 0. \end{aligned}$$

The inequality follows because P_2 is symmetric and idempotent and therefore positive semi-definite. Consequently,

$$(X_1'X_1)^{-1} - (X_1'M_2X_1)^{-1} \leq 0,$$

and

$$Var\left(\widetilde{\beta}_1 | X\right) - Var\left(\widehat{\beta}_1 | X\right) \leq 0.$$

Thus, the variance increases with the number of regressors.

Under Assumption (A5), we obtain that

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

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

$$s^2 = \frac{Y'M_1Y}{n - k_1},$$

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

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

and

$$\begin{aligned} E\left(s^2|X\right) &= E\left(\frac{U'M_1U}{n-k_1}|X\right) + 2E\left(\frac{U'M_1X_2\beta_2}{n-k_1}|X\right) + \frac{\beta_2'X_2'M_1X_2\beta_2}{n-k_1} \\ &= \sigma^2 + \frac{\beta_2'X_2'M_1X_2\beta_2}{n-k_1} \\ &\geq \sigma^2. \end{aligned}$$

The bias remains, even if X_1 and X_2 are orthogonal, since in this case $M_1X_2 = X_2$.

One of the consequences of exclusion of relevant variables is that tests and confidence intervals are invalid.

Inclusion of irrelevant variables

Suppose that the true model is given by

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

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

$$\widetilde{\beta}_1 = \left(X_1'M_2X_1\right)^{-1}X_1'M_2Y.$$

In this case, the LS estimator is unbiased under Assumption (A2):

$$\begin{aligned} E\left(\widetilde{\beta}_1|X\right) &= E\left(\left(X_1'M_2X_1\right)^{-1}X_1'M_2\left(X_1\beta_1 + U\right)|X\right) \\ &= \beta_1 + \left(X_1'M_2X_1\right)^{-1}X_1'M_2E\left(U|X\right) \\ &= \beta_1. \end{aligned}$$

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

$$Var\left(\widetilde{\beta}_1|X\right) = \sigma^2\left(X_1'M_2X_1\right)^{-1}.$$

However, $\widetilde{\beta}_1$ is inefficient, due to the Gauss-Markov Theorem. As we have seen in the previous section, the variance increases with the number of regressors. Under Assumption (A5), we have

$$\widetilde{\beta}_1|X \sim N\left(\beta_1, \sigma^2\left(X_1'M_2X_1\right)^{-1}\right).$$

Next, consider s^2. In this case,

$$s^2 = \frac{Y'M_XY}{n-k_1-k_2},$$

where $M_X = I_n - X\left(X'X\right)^{-1}X'$. Since $M_XX_1 = 0$, it follows that

$$s^2 = \frac{U'M_XU}{n-k_1-k_2},$$

and

$$E\left(s^2|X\right) = \sigma^2.$$

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 comparing to the correctly specified equation. The discussion in the last two section that for model selection purpose, one should start with the most general model, and eliminate irrelevant regressors by applying F-tests.

Test of structural change

Suppose that there are two regression models representing, for example, observations in two countries or in 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 $n_1 + n_2 = n$. In this case, (Y_1, X_1) and (Y_2, X_2) composed of observations for the same dependent variable and regressors for *two different sub-samples*. One can ask whether the response of the dependent variable to changes in the regressor differs in the two sub-samples by testing

$$H_0 : \beta_1 = \beta_2. \tag{3}$$

In order to combine two equations into a single equation, it is convenient to define

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

Using the above definitions, the unrestricted model can be written as

$$Y = X\beta + U, \tag{4}$$

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

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

Note that in this case,

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

Consequently,

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

Now, the unrestricted RSS are given by the sum of RSS's from two separate regressions:

$$\begin{aligned}\widehat{U}'\widehat{U} &= Y'M_XY \\ &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}' \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= Y_1'M_1Y_1 + Y_2'M_2Y_2 \\ &= RSS_1 + RSS_2.\end{aligned}$$

Note that there are $2k$ regression slope coefficients in the unrestricted model.

Next, the restricted model can be written as

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

Therefore, the restricted RSS must be obtained by pulling together the two sub-samples. Define

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

and

$$M_r = I_n - X_r \left(X_r'X_r\right)^{-1} X_r'.$$

The restricted RSS are given by

$$RSS_r = Y'M_rY.$$

Therefore, the test of no structural changed is based on the following statistic:

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

One rejects the null of no structural change when

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

Dummy variables

Frequently in regression analysis the econometrician is interested in the effect of the variables that are qualitive and cannot be quantified in a usual way. For example, one may be interested in studying the effects of sex, marital status, race, religion on other economic variables such as income or education. A common approach to quantifying such variables is to introduce artificial variables that indicate if a particular quality is present. Supposed that a qualitive has m categories. For observations $i = 1, \ldots, n$, define the *dummy variables* d_{ij}, $j = 1, \ldots, m$ such that

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

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

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

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

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

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 $\alpha_1 = 0$.

Note that in the above example one cannot include the intercept and both dummy variables, since for all i's

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

which violates Assumption (A4). The general rule is that if a categorical variable has m categories, include m dummy and no intercept, or $m-1$ variables 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 modelled by including the *interaction term* of X_i with the dummy variable d_{i1}:

$$Y_i = \alpha_0 + \alpha_1 d_{i1} + X_i'\beta + (d_{i1}X)'\,\delta + U_i.$$

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

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

or equivalently,

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

In this case, $\beta_2 = \beta_1 + \delta$, and the test of no structural change is equivalent to testing $H_0 : \delta = 0$. In order to show that the two approaches, with and without dummy variables, are equivalent, it is sufficient 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):

$$Y_i = X_i'\beta + U_i.$$

In this section, we discuss *forecasting* the dependent variable Y_i given some *fixed* k-vector of values for the regressors x_f, and construction of confidence intervals for such forecasts. Let $\widehat{\beta}$ be the LS estimator of β based on the data $\{(Y_i, X_i) : i = 1, \dots, n\}$. Note that x_f may or may not be one of the realized values for the regressors in the observed sample. Since

$$E\left(Y_i | X_i = x_f\right) = x_f'\beta,$$

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

$$\widehat{Y}_f = x_f'\widehat{\beta}. \tag{7}$$

Note that $\widehat{Y}_f$ is the predicted value of a point *on the regression line*. Since x_f is fixed, $\widehat{Y}_f$ random only due to the randomness in $\widehat{\beta}$. Using the results for $\widehat{\beta}$, we obtain

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

The a-level confidence interval for the point on the regression line that corresponds to x_f is given by

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

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

$$Y_f = x_f'\beta + U_f,$$

where $(n+1)$-vector $(U', U_f)'$ satisfies

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

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

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

The result in (8) implies that

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

Therefore, the α-level confidence interval for the predicted value of Y_f is given by

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