Multivariate Regression [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

August 11, 2021

[1] This section is based on Stock and Watson (2020), Chapter 6-7.

- Although school districts with lower student–teacher ratios tend to have higher test scores in the California data set, perhaps students from districts with small classes have other advantages that help them perform well on standardized tests.
- Could this have produced a misleading estimate of the causal effect of class size on test scores, and, if so, what can be done?

- By focusing only on the student–teacher ratio, the empirical analysis the bivariate regression model ignored some potentially important determinants of test scores by collecting their influences in the regression error term.
- These omitted factors include school characteristics, such as teacher quality and computer usage, and student characteristics, such as family background.
- We begin by considering an omitted student characteristic that is particularly relevant in California because of its large immigrant population: the prevalence in the school district of students who are still learning English.

- By ignoring the **percentage of English learners** in the district, the OLS estimator of the effect on test scores of the student–teacher ratio could be biased;
- that is, the mean of the sampling distribution of the OLS estimator might not equal the true causal effect on test scores of a unit change in the student–teacher ratio.
- Students who are still learning English might perform worse on standardized tests than native English speakers.
- If districts with large classes also have many students still learning English, then the OLS regression of test scores on the student–teacher ratio could erroneously find a correlation and produce a large estimated coefficient, when in fact the true causal effect of cutting class sizes on test scores is small, even zero.

- Accordingly, based on the analysis of bivariate regression, the superintendent might hire enough new teachers to reduce the student–teacher ratio by 2, but her hoped-for improvement in test scores will fail to materialize if *the true coefficient is small or zero*.
- A look at the California data lends credence to this concern. The correlation between the student–teacher ratio and the percentage of English learners (students who are not native English speakers and who have not yet mastered English) in the district is **0.19.** This small but positive correlation suggests that districts with more English learners tend to have a higher student–teacher ratio (larger classes).
- If the student–teacher ratio were unrelated to the percentage of English learners, then it would be safe to ignore English proficiency in the regression of test scores against the student–teacher ratio. But because the student–teacher ratio and the percentage of English learners are correlated, it is possible that the OLS coefficient in the regression of test scores on the student–teacher ratio reflects that influence.

Multiple Regression Model

The multiple regression model extends the single variable regression model to include additional variables as regressors. When used for causal inference, this model permits estimating the effect on Y_i of changing one variable X_{1i} while holding the other regressors (X_{2i} and X_{3i}, and so forth) constant.

In the class size problem, the multiple regression model provides a way to isolate the effect on test scores Y_i of the student–teacher ratio X_{1i} while holding constant the percentage of students in the district who are English learners X_{2i}.

When used for prediction, the multiple regression model can improve predictions by using multiple variables as predictors.

The Population Regression Line

Suppose for the moment that there are two independent variables, X_{1i} and X_{2i}. In the linear multiple regression model, the average relationship between these two independent variables and the dependent variable:

$$E[Y_i|X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where $E[Y_i|X_{i1}, X_{i2}]$ is the conditional expectation of Y_i given X.
The equation is referred as the **population regression line**.
The interpretation of the coefficient β_1 is the predicted change in Y between two observations with a unit difference in X_1 controlling for X_2

Distribution of Error

- The error term u_i in the multiple regression model is **homoskedastic** if the variance of the conditional distribution of u_i given $X_{1i}, \ldots, X_{ki}$ is constant for $i = 1, \ldots, n$,
- and thus does not depend on the values of $X_{1i}, \ldots, X_{ki}$. Otherwise, the error term is **heteroskedastic**.

OLS Estimator

- We estimate the unknown population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ using a sample of data.
- The estimators of the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize the sum of squared mistakes are called the **ordinary least squares (OLS) estimators** of $\beta_0, \beta_1, \ldots, \beta_k$.

In this section, we make precise the requirements for OLS to provide valid inferences about causal effects.
Consider the case in which we are interested in knowing the causal effects of all k regressors in the multiple regression model; that is, all the coefficients $\beta_1, \ldots, \beta_k$ are causal effects of interest.
There are **four least squares assumptions** for causal inference in the multiple regression model.
The first three are those of Section 4.3 for the single-regressor model extended to allow for multiple regressors.

1. Assumption 1: The conditional distribution of u_i given $X_1, \ldots, X_k$ has mean of 0.
   This assumption extends the first least squares assumption with a single regressor to multiple regressors. This assumption is implied if $X_1, \ldots, X_k$ are randomly assigned or are as-if randomly assigned; if so, for any value of the regressors, the expected value of u_i is 0. As is the case for regression with a single regressor, this is the key assumption that makes the OLS estimators unbiased.

2. Assumption 2: $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are i.i.d
   The second assumption is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are independently and identically distributed (i.i.d.) random variables. This assumption holds automatically if the data are collected by simple random sampling.
3. Assumption 3: Large outliers are unlikely.
   The third least squares assumption is that large outliers—that is, observations with values far outside the usual range of the data—are unlikely. This assumption serves as a reminder that, as in the single-regressor case, the OLS estimator of the coefficients in the multiple regression model can be sensitive to large outliers.
4. Assumption 4: No perfect multicollinearity.
   The fourth assumption is new to the multiple regression model. It rules out an inconvenient situation called perfect multicollinearity, in which it is impossible to compute the OLS estimator. The regressors are said to exhibit perfect multicollinearity (or to be perfectly multicollinear) if one of the regressors is a perfect linear function of the other regressors.

Data generating process:
$Y = X^\top \beta + u$, where $X = (X_1, \dots, X_k)^\top$ and $\beta = (\beta_1, \dots, \beta_k)^\top$.
The least square estimator is defined by $\hat{\beta} = \arg\min E((Y - X^\top \beta)^2)$.
Suppose we have n observations, we have n linear equations:

$$Y_1 = X_1^\top \beta + u_1$$
$$\vdots$$
$$Y_n = X_n^\top \beta + u_n$$

Define $\mathrm{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\mathrm{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, $\mathrm{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$. Observe that Y and u Are $n \times 1$ vectors, and X is $n \times k$ matrix.
The system of n equations can be written as $\mathrm{Y} = \mathrm{X}\beta + \mathrm{u}$.
Sample sums can be written as $\sum_{i=1}^n X_i X_i^\top = \mathrm{X}^\top \mathrm{X}$, $\sum_{i=1}^n X_i Y_i = \mathrm{X}^\top \mathrm{Y}$.
The least square estimator is $\hat{\beta} = (\mathrm{X}^\top \mathrm{X})^{-1} (\mathrm{X}^\top \mathrm{Y})$.
The matrix version of the estimator is $\mathrm{Y} = \mathrm{X}\hat{\beta} + \hat{\mathrm{u}}$.

Using the residual vector we can write that $\mathrm{X}^\top \hat{\mathrm{u}} = 0$.
The sum of squared errors is $SSE(\beta) = (\mathrm{Y} - \mathrm{X}\beta)^\top(\mathrm{Y} - \mathrm{X}\beta)$.

The OLS estimator is unbiased estimator: $E(Y_i|X_1,\ldots,X_n) = E(Y_i|X_i) = X_i\beta$.

$$
\begin{aligned}
E(\beta|X_1,\ldots,X_n) &= E(\sum_{i=1}^{n}(X_iX_i^\top)^{-1}\sum_{i=1}^{n}(X_iY_i)|X_1,\ldots,X_n) \\
&= \sum_{i=1}^{n}(X_iX_i^\top)^{-1}E(\sum_{i=1}^{n}(X_iY_i)|X_1,\ldots,X_n) \\
&= \sum_{i=1}^{n}(X_iX_i^\top)^{-1}\sum_{i=1}^{n}E((X_iY_i)|X_1,\ldots,X_n) \\
&= \sum_{i=1}^{n}(X_iX_i^\top)^{-1}\sum_{i=1}^{n}X_iE(Y_i|X_i) \\
&= \sum_{i=1}^{n}(X_iX_i^\top)^{-1}\sum_{i=1}^{n}X_iX_i^\top\beta \\
&= \beta
\end{aligned}
\tag{1}
$$

$var(\mathrm{u}|\mathrm{X}) = E(\mathrm{u}\mathrm{u}^\top|\mathrm{X}) = \mathrm{D}$.
The i-th diagonal element of D is $E(u_i^2|\mathrm{X}) = E(u_i^2|X_i) = \sigma_i^2$ and the off-diagonal (i,j)-th element is $E(u_iu_j|\mathrm{X}) = E(u_i|X_i)E(u_j|X_j) = 0$.

$$\mathrm{D} = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}.$$

Homoskedastic regression: $\mathrm{D} = \sigma \mathrm{I}_n$.
The variance of the least square regression
$V_{\hat{\beta}} = var(\hat{\beta}|\mathrm{X}) = (\mathrm{X}^\top\mathrm{X})^{-1}(\mathrm{X}^\top\mathrm{D}\mathrm{X})(\mathrm{X}^\top\mathrm{X})^{-1}$.
If the error is homoskedastic, $V_{\hat{\beta}} = \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}$.
Apply the law of iterated expectation:
$E(\hat{\beta}) = E(E(\hat{\beta}|\mathrm{X}))$ and $var(\hat{\beta}) = E(V_{\hat{\beta}}) = \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}$.

Multi-collinearity I

For the least squares estimator to be uniquely defined the regressors cannot be linearly dependent.
However, it is quite easy to attempt to calculate a regression with linearly dependent regressors. This can
occur for many reasons, including the following.

1. Including the same regressor twice.
2. Including regressors which are a linear combination of one another, such as education, experience

and age in the CPS data set example (recall, experience is defined as age-education-6).

1. Including a dummy variable and its square.
2. Estimating a regression on a sub-sample for which a dummy variable is either all zeros or all ones.
3. Including a dummy variable interaction which yields all zeros.
4. Including more regressors than observations.

Control Variables and Conditional Mean Independence I

To distinguish between variables of interest and control variables, we modify the notation of the linear regression model to include k variables of interest, denoted by X, and r control variables, denoted by W.
Accordingly, the multiple regression model with control variables is

$$Y_i = X_i^\top \beta_{(1)} + W_i^\top \beta_{(2)},$$

where $\beta_{(1)}$ is the causal effect, W_i is the control variable,

1. $E(u_i|X_i, W_i) = E(u_i|W_i) = 0$, conditional independence.
2. (X_i, W_i, Y_i) are i.i.d.
3. X_i and W_i have nonzero finite fourth moment.
4. No perfect collinearity.

Control Variables and Conditional Mean Independence II

- **Conditional mean independence** requires that the conditional expectation of u_i given the variable of interest and the control variables does not depend on (is independent of) the variable of interest, although it can depend on control variables.
- The idea of conditional mean independence is that once you control for the W's the X's can be treated as if they were randomly assigned. Controlling for W makes the X's uncorrelated with the error term, so that OLS can estimate the causal effects on Y of a change in each of the X's.
- In the class size example, $LchPct$ can be correlated with factors, such as learning opportunities outside school, that enter the error term; indeed, it is because of this correlation that $LchPct$ is a useful control variable. This correlation between $LchPct$ and the error term means that the estimated coefficient on $LchPct$ does not have a causal interpretation. What the conditional mean independence assumption requires is that, given the

Control Variables and Conditional Mean Independence III

control variables in the regression (*PctEL* and *LchPct*), the mean of the error term does not depend on the student–teacher ratio. Said differently, conditional mean independence says that among schools with the same values of *PctEL* and *LchPct*, class size is “as-if” randomly assigned: Including *PctEL* and *LchPct* in the regression controls for omitted factors so that *STR* is uncorrelated with the error term. If so, the coefficient on the student–teacher ratio has a causal interpretation even though the coefficient on *LchPct* does not.

Hypothesis test for single coefficient I

Suppose that you want to test the hypothesis that a change in the student–teacher ratio has no effect on test scores, holding constant the percentage of English learners in the district. This corresponds to hypothesizing that the true coefficient β_1 on the student–teacher ratio is 0 in the population regression of test scores on STR and $PctEL$.

More generally, we might want to test the hypothesis that the true coefficient β_j on the j th regressor takes on some specific value, $\beta_{j,0}$. The null value comes either from economic theory or, as in the student–teacher ratio example, from the decision-making context of the application. If the alternative hypothesis is two-sided, then the two hypotheses can be written mathematically as

$$H_0 : \beta_j = \beta_{j,0} \; v.s. \; H_1 : \beta_j \neq \beta_{j,0}$$

For example, if the first regressor is STR, then the null hypothesis that changing the student–teacher ratio has no effect on test scores corresponds to the null hypothesis that b1 = 0 (so b1,0 = 0). Our

Hypothesis test for single coefficient II

task is to test the null hypothesis H0 against the alternative H1 using a sample of data.
Steps:

1. compute the standard error
2. compute the t-statistics
3. compute the p-value

Confidence Intervals for a Single Coefficient I

The method for constructing a confidence interval in the multiple regression model is also the same as in the single-regressor model. Accordingly, it should be kept in mind that these methods for quantifying the sampling uncertainty are only guaranteed to work in large samples.

- A 95% two-sided confidence interval for the coefficient β_j is an interval that contains the true value of bj with a 95% probability; that is, it contains the true value of bj in 95% of all possible randomly drawn samples.
- Equivalently, it is the set of values of β_j that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, the 95% confidence interval is

$$CI_{0.95}(\beta_j) = [\hat{\beta}_j - 1.96se(\hat{\beta}_j), \hat{\beta}_j + 1.96se(\hat{\beta}_j)].$$

Application to Test Scores and the Student–Teacher Ratio I

Can we reject the null hypothesis that a change in the student–teacher ratio has no effect on test scores, once we control for the percentage of English learners in the district? What is a 95% confidence interval for the effect on test scores of a change in the student–teacher ratio, controlling for the percentage of English learners? We are now able to find out. The regression of test scores against *STR* and *PctEL*,

$$\widehat{TestScore} = \underset{8.7}{686.0} - \underset{0.43}{1.10} * STR - \underset{0.0031}{0.650} * PctEL$$

To test the hypothesis that the true coefficient on *STR* is 0, we first need to compute the t-statistic in Equation (7 .2). Because the null hypothe sis says that the true value of this coefficient is 0, the t-statistic is t = (-1.10 - 0) / 0.43 = -2.54. The associated p-value is $2\Phi(-2.54) = 0.011$; that is, the smallest significance level at which we can reject the null hypothesis is 1.1%. Because the p-value is less

Application to Test Scores and the Student–Teacher Ratio II

than 5%, the null hypothesis can be rejected at the 5% significance level (but not quite at the 1% significance level).
A 95% confidence interval for the population coefficient on STR is $-1.10 \pm 1.96 * 0.43 = (-1.95, -0.26)$; that is, we can be 95% confident that the true value of the coefficient is between -1.95 and -0.26. Interpreted in the context of the superintendent's interest in decreasing the student–teacher ratio by 2, the 95% confidence interval for the effect on test scores of this reduction is $(-0.26 * -2, -1.95 * -2) = (0.52, 3.902)$.
Adding expenditures per pupil to the equation. Your analysis of the multiple regression in Equation (7 .5) has persuaded the superintendent that, based on the evidence so far, reducing class size will improve test scores in her district. Now, however, she moves on to a more nuanced question. If she is to hire more teachers, she can pay for those teachers either by making cuts elsewhere in the budget (no new computers, reduced maintenance, and so on) or by asking for an increase in her budget, which taxpayers do not favor. What,

Application to Test Scores and the Student–Teacher Ratio III

she asks, is the effect on test scores of reducing the student–teacher ratio, holding expenditures per pupil (and the percentage of English learners) constant?

$$\widehat{TestScore} = \underset{15.5}{649.6} - \underset{0.48}{0.29} * STR + \underset{1.59}{3.87} * Expn - \underset{0.032}{0.656} * PctEL,$$

- The result is striking. Holding expenditures per pupil and the percentage of English learners constant, changing the student–teacher ratio is estimated to have a very small effect on test scores: The estimated coefficient on *STR* is -**1.10**, but after adding Expn as a regressor in Equation (7 .6), it is only -**0.29**.
- Moreover, the t-statistic for testing that the true value of the coefficient is 0 is now t = (-0.29 - 0) >0.48 = -0.60, so the hypothesis that the population value of this coefficient is indeed 0 cannot be rejected even at the 10% significance level.

Application to Test Scores and the Student–Teacher Ratio IV

- One interpretation of the regression in Equation (7 .6) is that, in these California data, school administrators allocate their budgets efficiently.
- Suppose, counterfactually, that the coefficient on *STR* in Equation (7 .6) were negative and large. If so, school districts could raise their test scores simply by decreasing funding for other purposes (textbooks, technology, sports, and so on) and using those funds to hire more teachers, thereby reducing class sizes while holding expenditures constant. However, the small and statistically insignificant coefficient on STR in Equation (7 .6) indicates that this transfer would have little effect on test scores. Put differently, districts are already allocating their funds efficiently.

Application to Test Scores and the Student–Teacher Ratio V

- Note that the standard error on STR increased when Expn was added, from 0.43 in Equation (7 .5) to 0.48 in Equation (7 .6). This illustrates the general point, introduced in Section 6.7 in the context of imperfect multicollinearity, that correlation between regressors (the correlation between STR and Expn is -0.62) can make the OLS estimators less precise.

Testing Hypotheses on Two or More Coefficients I

Joint null hypotheses. Consider the regression in Equation (7 .6) of the test score against the student–teacher ratio, expenditures per pupil, and the percentage of English learners.
Our angry taxpayer hypothesizes that neither the student–teacher ratio nor expenditures per pupil have an effect on test scores, once we control for the percentage of English learners.
Because STR is the first regressor in Equation (7 .6) and Expn is the second, we can write this hypothesis mathematically as

$$H0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \; vs \; H1 : \beta_1 \neq 0 \text{ and / or } \beta_2 \neq 0.$$

The hypothesis that both the coefficient on the student–teacher ratio (β_1) and the coefficient on expenditures per pupil (β_2) are 0 is an example of a joint hypothesis on the coefficients in the multiple regression model. In this case, the null hypothesis restricts the value of two of the coefficients, so as a matter of terminology we can say that the null hypothesis in Equation (7 .7) imposes two restrictions on the multiple regression model: $\beta_1 = 0$ and $\beta_2 = 0$.

Testing Hypotheses on Two or More Coefficients II

In general, a joint hypothesis is a hypothesis that imposes two or more restrictions on the regression coefficients. We consider joint null and alternative hypotheses of the form
$H0 : \beta_j = \beta_{j,0}, \ldots \beta_m = \beta_{m,0},$, for a total of q restrictions, vs. $H1$: one or more of the q restrictions under $H0$ does not hold,
Why can't I just test the individual coefficients one at a time?
Although it seems it should be possible to test a joint hypothesis by using the usual t-statistics to test the restrictions one at a time, the following calculation shows that this approach is unreliable.
Specifically, suppose you are interested in testing the joint null hypothesis in Equation (7 .6) that b1 = 0 and b2 = 0. Let t1 be the t-statistic for testing the null hypothesis that b1 = 0, and let t2 be the t-statistic for testing the null hypothesis that b2 = 0. What happens when you use the "one-at-a-time" testing procedure: Reject the joint null hypothesis if either t1 or t2 exceeds 1.96 in absolute value?

Testing Hypotheses on Two or More Coefficients III

Because this question involves the two random variables t1 and t2, answering it requires characterizing the joint sampling distribution of t1 and t2, so under the joint null hypothesis the t-statistics t1 and t2 have a bivariate normal distribution, where each t-statistic has a mean equal to 0 and variance equal to 1.

The homoskedasticity-only F-statistic is computed using a simple formula based on the sum of squared residuals from two regressions. In the first regression, called the restricted regression, the null hypothesis is forced to be true. When the null hypothesis is of the type in Equation (7 .8), where all the hypothesized values are 0, the restricted regression is the regression in which those coefficients are set to 0; that is, the relevant regressors are excluded from the regression. In the second regression, called the unrestricted regression, the alternative hypothesis is allowed to be true. If the sum of squared residuals is sufficiently smaller in the unrestricted than in the restricted regression, then the test rejects the null hypothesis. The homoskedasticity-only F-statistic is given by the formula

Testing Hypotheses on Two or More Coefficients IV

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{restricted} - 1)}.$$

- If the errors are homoskedastic, then the difference between the homoskedasticityonly F-statistic computed using Equation (7 .13) or (7 .14) and the heteroskedasticityrobust F-statistic vanishes as the sample size n increases. Thus, if the errors are homoskedastic, the sampling distribution of the homoskedasticity-only F-statistic under the null hypothesis is, in large samples, $F_{q,\infty}$.

Testing Hypotheses on Two or More Coefficients V

- These formulas are easy to compute and have an intuitive interpretation in terms of how well the unrestricted and restricted regressions fit the data. Unfortunately, the formulas apply only if the errors are homoskedastic. Because homoskedasticity is a special case that cannot be counted on in applications with economic data—or more generally with data sets typically found in the social sciences—in practice the homoskedasticity-only F-statistic is not a satisfactory substitute for the heteroskedasticityrobust F-statistic.
- Using the homoskedasticity-only F-statistic when n is small. If the errors are i.i.d., homoskedastic, and normally distributed, then the homoskedasticity-only F-statistic defined in Equations (7 .13) and (7 .14) has an $F_{q,n-k_{unrestricted}-1}$ distribution under the null hypothesis (see Section 19.4). Critical values for this distribution, which depend on both q and $k_{unrestricted}$, are given in Appendix Table 5. As discussed in Section 2.4, the $F_{q,n-k_{unrestricted}-1}$ distribution converges to the F_q, distribution as

Testing Hypotheses on Two or More Coefficients VI

n increases; for large sample sizes, the differences between the two distributions are negligible. For small samples, however, the two sets of critical values differ. However, it is easy to read more into them than they deserve. There are four potential pitfalls to guard against when using the R2 or R2.

Application to test scores and the student–teacher ratio. To test the null hypothesis that the population coefficients on *STR* and Expn are 0, controlling for *PctEL*, we need to compute the R2 (or SSR) for the restricted and unrestricted regressions. The unrestricted regression has the regressors *STR*, *Expn*, and *PctEL* and is given in Equation (7 .6). Its R^2 is 0.4366; that is, R^2 unrestricted $= 0.4366$. The restricted regression imposes the joint null hypothesis that the true coefficients on *STR* and Expn are 0; that is, under the null hypothesis STR and *Expn* do not enter the population regression, although *PctEL* does (the null hypothesis does not restrict the coefficient on *PctEL*). The restricted regression, estimated by OLS, is

Testing Hypotheses on Two or More Coefficients VII

$$\widehat{TestScore} = \underset{1.0}{664.7} - \underset{0.032}{0.671} * PctEL, R^2 = 0.4149$$

The number of restrictions is $q = 2$, the number of observations is $n = 420$, and the number of regressors in the unrestricted regression is $k = 3$. The homoskedasticity-only F-statistic, computed using Equation

$$F = \frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)/(420 - 3 - 1)} = 8.01.$$

Because 8.01 exceeds the 1% critical value of 4.61, the hypothesis is rejected at the 1% level using the homoskedasticity-only test. This example illustrates the advantages and disadvantages of the homoskedasticityonly F-statistic. An advantage is that it can be computed using a calculator. Its main disadvantage is that the values of the homoskedasticity-only and heteroskedasticityrobust F-statistics can be very different: The heteroskedasticity-robust F-statistic

Testing Hypotheses on Two or More Coefficients VIII

testing this joint hypothesis is 5.43, quite different from the less reliable homoskedasticityonly value of 8.01.

Testing Single Restrictions Involving Multiple Coefficients I

Sometimes economic theory suggests a single restriction that involves two or more regression coefficients. For example, theory might suggest a null hypothesis of the form b1 $=$ b2; that is, the effects of the first and second regressors are the same. In this case, the task is to test this null hypothesis against the alternative that the two coefficients differ

$$H_0 : \beta_1 = \beta_2, \text{ v.s. } H_1 : \beta_1 \neq \beta_2.$$

This null hypothesis has a single restriction, so $q = 1$, but that restriction involves multiple coefficients (β_1 and β_2). We need to modify the methods presented so far to test this hypothesis. There are two approaches; which is easier depends on your software

Approach 1: Test the restriction directly. Some statistical packages have a specialized I

command designed to test restrictions like Equation (7 .16), and the result is an F-statistic that, because q = 1, has an F1, distribution under the null hypothesis. (Recall from Section 2.4 that the square of a standard normal random variable has an F1, distribution, so the 95% percentile of the F1, distribution is 1.962 = 3.84

Approach 2: Transform the regression I

If your statistical package cannot test the restriction directly, the hypothesis in Equation (7 .16) can be tested using a trick in which the original regression equation is rewritten to turn the restriction in Equation (7 .16) into a restriction on a single regression coefficient. To be concrete, suppose there are only two regressors, X_{1i} and X_{2i}, in the regression, so the population regression has the form.

Extension to q $>$ 1 I

In general, it is possible to have q restrictions under the null hypothesis in which some or all of these restrictions involve multiple coefficients. The F-statistic of Section 7.2 extends to this type of joint hypothesis.

The F-statistic can be computed by either of the two methods just discussed for $q = 1$. Precisely how best to do this in practice depends on the specific regression software being used.

Confidence Sets for Multiple Coefficients I

- The method is conceptually similar to that of a single coefficient using the t-statistic except that the confidence set for multiple coefficients is based on the F-statistic.
- A 95% confidence set for two or more coefficients is a set that contains the true population values of these coefficients in 95% of randomly drawn samples.
- Recall that a 95% confidence interval is computed by finding the set of values of the coefficients that are not rejected using a t-statistic at the 5% significance level.

This approach can be extended to the case of multiple coefficients.

- Suppose you are interested in constructing a confidence set for two coefficients, β_1 and β_2.
- Similar to use the F-statistic to test a joint null hypothesis that $\beta_1 = \beta_1, 0$ and $\beta_2 = \beta_{2,0}$.

Confidence Sets for Multiple Coefficients II

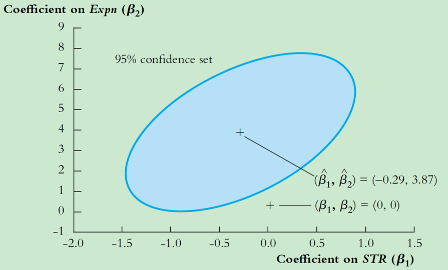
- Suppose you were to test every possible value of $\beta_{1,0}$ and $\beta_{2,0}$ at the 5% level. For each pair of candidates $(\beta_{1,0}, \beta_{2,0})$, you compute the F-statistic and reject it if it exceeds the 5% critical value of 3.00.
- Because the test has a 5% significance level, the true population values of β_1 and β_2 will not be rejected in 95% of all samples. Thus the set of values not rejected at the 5% level by this F-statistic constitutes a 95% confidence set for b1 and β_2.
- When there are two coefficients, the resulting confidence sets are ellipses.

Confidence Sets for Multiple Coefficients III

In practice it is much simpler to use an explicit formula for the confidence set. This formula for the confidence set for an arbitrary number of coefficients is obtained using the formula for the F-statistic. [2].

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level. The point $(\beta_1, \beta_2) = (0, 0)$ is not contained in the confidence set, so the null hypothesis H_0: $\beta_1 = 0$ and $\beta_2 = 0$ is rejected at the 5% significance level.



- ▷ This ellipse does not include the point $(0, 0)$.

Confidence Sets for Multiple Coefficients IV

- This means that the null hypothesis that these two coefficients are both 0 is rejected using the F-statistic at the 5% significance level.
- The confidence ellipse is a fat sausage with the long part of the sausage oriented in the lower-left/upper-right direction.
- The reason for this orientation is that the estimated correlation.

[2] See SW Section 19.3

Some Remarks I

1. An increase in the R^2 or R^2 does not necessarily mean that an added variable is statistically significant. The R^2 increases whenever you add a regressor, whether or not it is statistically significant. The R^2 does not always increase, but if it does, this does not necessarily mean that the coefficient on that added regressor is statistically significant. To ascertain whether an added variable is statistically significant, you need to perform a hypothesis test using the t-statistic.
2. A high R^2 or R^2 does not mean that the regressors are a true cause of the dependent variable. Imagine regressing test scores against parking lot area per pupil. Parking lot area is correlated with the student–teacher ratio, with whether the school is in a suburb or a city, and possibly with district income—all things that are correlated with test scores. Thus the regression of test scores on parking lot area per pupil could have a high R^2 and R^2, but the relationship is not causal (try telling the superintendent that the way to increase test scores is to increase parking space!).

Some Remarks II

3. A high R^2 or R^2 does not mean that there is no omitted variable bias. Recall the discussion of Section 6.1, which concerned omitted variable bias in the regression of test scores on the student–teacher ratio. The R^2 of the regression was not mentioned because it played no logical role in this discussion. Omitted variable bias can occur in regressions with a low R^2, a moderate R^2, or a high R^2. Conversely, a low R^2 does not imply that there necessarily is omitted variable bias.
4. A high R^2 or R^2 does not necessarily mean that you have the most appropriate set of regressors, nor does a low R^2 or R^2 necessarily mean that you have an inappropriate set of regressors. The question of what constitutes the right set of regressors in multiple regression is difficult, and we return to it throughout this textbook. Decisions about the regressors must weigh issues of omitted variable bias, data availability, data quality, and, most importantly, economic theory and the nature of the substantive questions being addressed. None of these questions can be answered simply by having a high (or low) regression R2 or R2.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.