Multivariate Regression [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

September 9, 2021

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

Motivation

- Although school districts with lower student-teacher ratios tend to have higher test scores in the California data set,
- 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?

Omitted Variable Bias

- By focusing only on the student-teacher ratio, ignored some potentially important determinants of test scores.
  - school characteristics, such as teacher quality and computer usage,
  - student characteristics, such as family background.
- 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.

Control for percentage of English learners I

- By ignoring the **percentage of English learners** in the district, the OLS estimator could be biased;
  - not equal the true causal effect.
- Students who are still learning English might perform worse on standardized tests than native English speakers.
- Direction of the bias?
  - If positive correlation(districts with large classes also have many students still learning English).

Control for percentage of English learners II

- Wrong Implication: 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*.
- 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.**
- If the student–teacher ratio were unrelated to the percentage of English learners, safe to ignore.

Control for percentage of English learners III

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student-Teacher Ratios, by the Percentage of English Learners in the District

	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High Student-Teacher Ratio	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

Figure

Control for percentage of English learners IV

What can you do about omitted variable bias?

- select a subset of districts that have the same fraction of English learners but have different class sizes:
- class size cannot be picking up the English learner effect because the fraction of English learners is held constant.

Mathematical Expression for OVB

- **DGP**: $Y_i = X_i\beta + u_i$,
- $cor(u_i, x_i) = \rho_{ux} \neq 0$.

$$\hat{\beta}_1 \rightarrow_p \beta_1 + \frac{\sigma_u}{\sigma_x}\rho_{ux}$$

1. OVB cause problem for the consistency of $\hat{\beta}$.
2. The magnitude of the bias depends on $|\rho_{ux}|$.
3. The direction of the bias depends on the sign of ρ_{ux}.

Multiple Regression Model

- Multivariable OLS extends the single variable OLS, include additional variables as regressors.
- Causal inference, 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.
- e.g. **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}.
- the multiple regression model can improve predictions by using multiple variables as predictors.

The Population Regression Line

- Two independent variables, X_{1i} and X_{2i}.
- Interest in 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:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}.$$

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 $\beta_0, \beta_1, \ldots, \beta_k$.

The Least Square Assumption

Parameters of interest: estimate the coefficients $\beta_1, \ldots, \beta_k$.
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.
2. Assumption 2: $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are i.i.d
3. Assumption 3: Large outliers are unlikely.
4. Assumption 4: No perfect multicollinearity.
   other regressors.

Goodness of Fit I

1. The Standard Error of the Regression (SER).
   - ◇ The standard error of the regression (SER) estimates the standard deviation of the error term u_i.
   - ◇ and where SER is the sum of squared residuals, $SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}$ where $s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2$.
   - ◇ compare SER with the divisor is n - k - 1 rather than $n-2$.

Goodness of Fit II

2. The R^2.
   the fraction of the sample variance of Y explained by the regressors.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

   where the **explained sum of squares** is $ESS = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$ and the **total sum of squares** is $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
3. The Adjusted R^2, or $\bar{R}^2$.
   - ◇ More variables increase R^2.
   - ◇ Use adjusted $\bar{R}^2$, $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$.

Multi-collinearity I

Easy to attempt to calculate a regression with linearly dependent regressors.

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

Multi-collinearity II

Some Examples:

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 Variable I

Consider the multiple regression model with control variables

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

- k variables of interest, denoted by X,
- and r control variables, denoted by W

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

Control Variable II

Assumptions:

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

Conditional mean independence

Conditional mean independence requires that the conditional expectation of u_i given the variable of interest and the control variables does not depend on the variable of interest, although it can depend on control variables.

- once you control for the W's the X's can be treated as if they were randomly assigned.
- X's uncorrelated with the error term.
- OLS can estimate the causal effects on Y of a change in each of the X's.

Example of Control Variable I

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL, \quad (6.12)$$

$$\widehat{TestScore} = 700.2 - 1.00 \times STR - 0.122 \times PctEL - 0.547 \times LchPct. \quad (6.16)$$

- Consider the potential OVB arising from outside learning opportunities.
  - difficult to measure,
  - correlated with the students' economic background, (can be measured).

Example of Control Variable II

- regression of test scores on *STR* and *PctEL*
- with the percentage of students receiving a free or subsidized school lunch(*LchPct*).
- Students are eligible if their family income less than 150% of the poverty line
- the coefficient on the *STR* is the effect of the student–teacher ratio on test scores,
  - Including the control variable *LchPct* does not substantially change any conclusions about the class size effect: The coefficient on STR changes only slightly from its value of -1.10 in Equation (6.12) to -1.00 in Equation (6.16)

Choose Control Variables I

Multiple regression makes it possible to control for factors that could lead to omitted variable bias in the estimate of the effect of interest.

- how does one determine the "right" set of control variables?
- the conditional mean independence condition.
- to eliminate omitted variables bias, a set of control variables must satisfy $E[u_i|X_i, W_i] = E[u_i|W_i]$,
  - where X_i denotes the variable or variables of interest and W_i denotes one or more control variables.

Choose Control Variables II

- Use judgement to determining which control variables to include.
  - For example, economic conditions vary **across school districts** with the same percentage of English learners.
  - more affluent districts would be expected to have lower class sizes,
  - Affluent families tend to have more access to outside learning opportunities.
  - OVB even after controlling for the percentage of English learners.
- Logic leads to including one or more additional control variables in the test score regressions, where the additional control variables measure economic conditions of the district.
- Our approach to the challenge of choosing control variables is twofold.
  - First, a core or base set of regressors should be chosen using a combination of expert judgment, economic theory, and knowledge of how the data were collected; the regression using this base set of regressors is sometimes referred to as a **base specification**.

Choose Control Variables III

- This base specification should contain the variables of primary interest and the control variables suggested by expert judgment and economic theory.
- develop a list of candidate alternative specifications—that is, alternative sets of regressors.

▷ If the estimates of the coefficients of interest change substantially across specifications–OVB.

Hypothesis test for single coefficient I

- Want to test the hypothesis that a change in STR has no effect on test scores, holding constant $PctEl$.
- $H_0 : \beta_1 = 0$.
- Moregenerally: can test $H_0 : \beta_j = \beta_{j,0}$.
- If the alternative hypothesis is two-sided, then the two hypotheses can be written mathematically as

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

Hypothesis test for single coefficient II

- 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 $\beta_1 = 0$ (so $\beta_{1,0} = 0$).
- Task: test the null hypothesis H_0 against the alternative H_1 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

- Same as in the single-regressor model.
- Note: only work in large samples.
- 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 I

- test the null hypothesis that a change in the student-teacher ratio has no effect on test scores, control for the percentage of English learners in the district?
- The regression of test scores against STR and $PctEL$,

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} * STR - \underset{(0.031)}{0.650} * PctEL(7.5)$$

- Construct the t-statistic $t = (-1.10 - 0)/0.43 = -2.54$.
- Associated p-value is $2\Phi(-2.54) = 0.011$.
- Reject at H_0.

Application II

- A 95% confidence interval for the population coefficient on STR is $-1.10 \pm 1.96 * 0.43 = (-1.95, -0.26)$.
- If 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 $(-0.26 * -2, -1.95 * -2) = (0.52, 3.902)$.

Example with expenditures per student I

Adding expenditures per pupil to the equation.

- What is the effect on test scores of reducing the student-teacher ratio, holding expenditures per student constst?
-

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29} * STR + \underset{(1.59)}{3.87} * Expn - \underset{(0.032)}{0.656} * PctEL, \quad (7.6)$$

  - coefficient on STR is -**1.10**, but after adding Expn as a regressor in Equation (7 .6), it is only -**0.29**.
  - $t = (-0.29 - 0)/0.48 = -0.60$,
  - Cannot reject H_0 at the 10% significance level.
  - Interpretation of (7.6) is that: in these California data, school administrators allocate their budgets efficiently.
  - Counterfactual, the coefficient on STR in (7.6) were negative and large.
    - 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.
    - this transfer would have little effect on test scores.

Example with expenditures per student II

- **Remark**: the standard error on *STR* increased when *Expn* was added, from 0.43 to 0.48. 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

- Consider the model
- Consider the test

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

- Example of joint hypothesis on the coefficients in the multiple regression model.
- The null hypotheses is $H_0 : \beta_1 = 0$ and $\beta_2 = 0$.

A joint hypothesis is a hypothesis that imposes two or more restrictions on the regression coefficients.
Consider the alternative form: $H_0 : \beta_j = \beta_{j,0}, \ldots \beta_m = \beta_{m,0},$, for a total of q restrictions, vs. H_1 : 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?

- Seemingly possible to test a joint hypothesis by using 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 that $\beta_1 = 0$ and $\beta_2 = 0$.
- Let t_1 and t_2 be the t-statistic for testing $\beta_1 = 0$ and $\beta_2 = 0$
- What happens when you use the "one-at-a-time" testing procedure: Reject the joint null hypothesis if either t_1 or t_2 exceeds 1.96 in absolute value?

We derive the joint distribution of t_1 and t_2.

F-test I

- The **homoskedasticity-only** F-statistic is computed using a simple formula based on the sum of squared residuals from two regressions.
- **Restricted regression**: the null hypothesis is forced to be true.
- **Unrestricted regression**: the alternative hypothesis is allowed to be true.
- The homoskedasticity-only F-statistic is given by the formula

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

where SSR is the sum of squared residuals.

F-test II

We can write the F-test as

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}.$$

- If the errors are homoskedastic, then the difference between the homoskedasticityonly F-statistic and the heteroskedasticityrobust F-statistic vanishes as the sample size n increases.
- If the errors are homoskedastic, the sampling distribution of the homoskedasticity-only F-statistic under the null hypothesis is, in large samples, $F_{q,\infty}$.
- Advantages: easy to compute, clear to interpret. Disadvantage: only apply to homoskedastic errors.
- 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 has an $F_{q,n-k_{unrestricted}-1}$ distribution under the null hypothesis.
- Critical values depend on both q and $k_{unrestricted}$, check SW Appendix Table 5 or use `qf(p, df1, df2)` in R.
- $F_{q,n-k_{unrestricted}-1} \rightarrow_d F_{q,\infty}$ as n increases;
- For small samples, however, the two sets of critical values differ. However, it is easy to read more into them than they deserve.

Application to test scores and the student-teacher ratio I

- H_0 : coefficients on STR and and $Expn$ are 0, controlling for $PctEL$.
- The unrestricted regression has the regressors STR, $Expn$, and $PctEL$:

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

$$R^2 = 0.4366$$

- The restricted model:

$$\widehat{TestScore} = \underset{(1.0)}{664.7} - \underset{(0.032)}{0.671} * PctEL,$$

$$R^2 = 0.4149$$

Application to test scores and the student-teacher ratio II

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

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

- For $F_{2,\infty}$, 1% critical value of 4.61
- Rejected at the 1% level using the homoskedasticity-only test.
- Remark: The heteroskedasticity-robust F-statistic 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

- Economic theory may suggest a single restriction that involves two or more regression coefficients. For example, theory might suggest a null hypothesis of the form $\beta_1 = \beta_2$;
- $$H_0 : \beta_1 = \beta_2, \text{ v.s. } H_1 : \beta_1 \neq \beta_2.$$
- single restriction, $q = 1$, 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

Testing Single Restrictions Involving Multiple Coefficients II

1. Approach 1: Test the restriction directly.
   - Test with an F-statistic that, because $q = 1$, has an F_1-distribution
   - Under the null hypothesis, the 95% percentile of the F_1-distribution is $1.96^2 = 3.84$.

Testing Single Restrictions Involving Multiple Coefficients III

2. Approach 2: Transform the regression.
   - ◇ Rewrite the equation and turn the restriction on on a single regression coefficient.

Extension to q > 1 I

- In general, q restrictions under the null hypothesis in which some or all of these restrictions involve multiple coefficients.
- User the F-statistic to test this type of joint hypothesis.
- The F-statistic can be computed by either of the two methods just discussed for $q = 1$.

Confidence Sets for Multiple Coefficients I

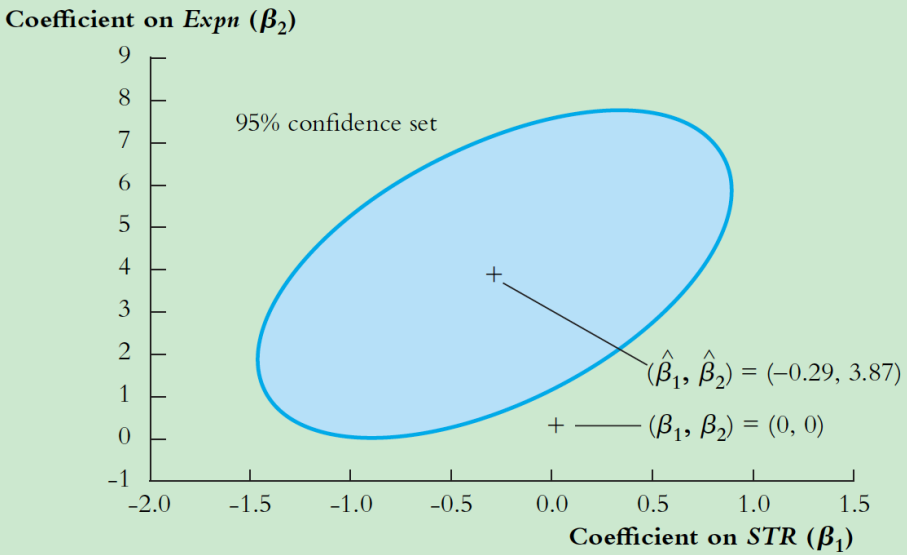
- The method is conceptually similar to that of a single coefficient using the t-statistic except that the confidence set for multiple coefficients is based on the F-statistic.
- A 95% confidence set for two or more coefficients is a set that contains the true population values of these coefficients 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.

Confidence Sets for Multiple Coefficients II

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

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 III

- 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 oriented in the lower-left/upper-right direction.
- Estimated correlation.

Some Remarks I

1. An increase in the R^2 or R^2 does not necessarily mean that an added variable is statistically significant.
2. A high R^2 or R^2 does not mean that the regressors are a true cause of the dependent variable.
3. A high R^2 or R^2 does not mean that there is no 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.

Tabulate Results I

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28	−1.10	−1.00	−1.31	−1.01
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
	[−3.30, −1.26]	[−1.95, −0.25]	[−1.53, −0.47]	[−1.97, −0.64]	[−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650	−0.122	−0.488	−0.130
		(0.031)	(0.033)	(0.030)	(0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547		−0.529
			(0.024)		(0.038)
Percentage qualifying for income assistance (X_4)				−0.790	0.048
				(0.068)	(0.059)
Intercept	698.9	686.0	700.2	698.0	700.4
	(10.4)	(8.7)	(5.6)	(6.9)	(5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

Tabulate Results II

- Way compare several possible sets of regressors, tabulate the results.
- Each column presents a separate regression.
  - same dependent variable: test score.
  - statistics
  - standard error in brackets is a 95% confidence interval for the population coefficient
- The final three rows contain **summary statistics**:
  - SER: the standard error of the regression, SER
  - R^2
  - the sample size (same for all of the regressions, 420 observations).
- Use the results to test single regressor.
- Although the table does not report t-statistics, they can be computed from the information provided.

Tabulate Results III

These results suggest three conclusions:

1. Controlling for these student characteristics cuts the estimated effect of the student–teacher ratio on test scores approximately in half.
   - ◇ This estimated effect is not very sensitive to which specific control variables are included in the regression.
   - ◇ In all cases, the hypothesis that the coefficient on the student–teacher ratio is 0 can be rejected at the 5% level. In the four specifications with control variables, regressions (2) through (5), reducing the student–teacher ratio by one student per teacher is estimated to increase average test scores by approximately 1 point, holding constant student characteristics.

Tabulate Results IV

2. The student characteristic variables are potent predictors of test scores. The student–teacher ratio alone explains only a small fraction of the variation in test scores: The R^2 in column (1) is 0.049. The R^2 jumps when the student characteristic variables are added. For example, the R2 in the base specification, regression (3), is 0.773. Districts with many English learners and districts with many poor children have lower test scores.
3. In contrast to the other two control variables, the percentage qualifying for income assistance appears to be redundant. As reported in regression (5), adding it to regression (3) has a negligible effect on the estimated coefficient on the student-teacher ratio or its standard error.

References I

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