Reading Assignment 6

Andrew Shaw

09-20-2025

Reading: Stock & Watson 6.1-6.4

1. What are the two requirements for omitted variable bias to be a problem?

1. The omitted variable is correlated with the regressor.
2. If the omitted variable is a determinant of the dependent variable.

2. What is the formula for omitted variable bias?

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{xu}\frac{\sigma_u}{\sigma_X}$$

3. How can we lower omitted variable bias without running a multiple regression model?

The data could be reexeamined to find an independent variable that is more exclusive and provide the same

4. Write a linear equation that relates wages to gender, age and education. Take a guess for the likely signs of each of these variables.

$$Wages = \beta_0 + \beta_{sex}X_{sex} + \beta_{edu}X_{edu} + \beta_{age}X_{age}$$

If we assume that X_{sex} as a dummy variable equal to 1 for females, the expected sign of this coefficient is negative while the others positive as wages typically increase with more education and age.

5. What variables might be omitted from your equation that could be biasing the effect of gender on wages?

Occupation or sector, job levels, full-time or part-time could also bias such a general model.

6. Write down the term for bias created by what you left out.

This would be the covariance of the dummy variable when female and the error term times the ratio of standard deviations of the error term and the sex dummy variable.

7. How can we remedy this situation and find a better estimate of the effect of gender on wages?

Adding in additional regressors that explicitly account for other factors that influence wages so they are captured in their own coefficient and not captured in the coefficient for the dummy variable for gender or in the error term.

8. How can we minimize errors in this multiple variable context?

Adding regressors that are relevant can make explicit more information otherwise captured in the error term or miscaptured in other coefficients.

9. What is the difference between Rˆ2 and adjusted Rˆ2?

The difference is largely how either handles the presence of additional regressors. R^2 has no penalty for more regressors thereby inflating the ratio as with more regressors, SSR decreases which makes it look like the model fits better. $\bar{R}^2$ accounts for the additional regressors making adjusted R^2 smaller

10. How can you interpret Rˆ2 in the multiple linear regression model.

A larger magnitude can mean that the model is doing a good job of fitting to the data or it could mean the model has too many regressors. It could also mean that Y is easily predictable.

Reading: Stock & Watson 6.5-6.7

11. Can we ever be sure that the first assumption of multiple linear regression holds without randomization?

No, not without randomization and even then it's never 100% sure.

12. The second and third assumption are essentially the same as in the single variable case, what is the fourth assumption?

The fourth is that there can be no perfect multicollinearity. If present, the model couldn't be calculated as there would be two variables that are linear functions of each other and you can't both try to find the effect of X on Y while also holding X constant.

13. Why is it not so surprising that the central limit theorem applies in the multiple linear regression case?

The coefficients are all partial averages of the whole set of sample data so if CLT were to hold if did single regression of the same data, it must be true that if the same data were broken down into smaller parts, CLT should hold as well.

14. The most common cause of perfect multicollinearity is the dummy variable trap. What is the dummy variable trap?

If data is all perfectly captured in a single category and all categories are all represented in the model then there is perfect multicollinearity because information would be redundant because it would be known if it's not in this group then that information is also captured in the other variable that does represent that group.

15. Why is imperfect multicollinearity a problem? Is it a problem if we have a large enough sample size? Use the variance formula in your answer.

Imperfect can still cause estimation problems for at least one regressor because there will exist a linear function that can describe the other which would make identifying which variable has an effect on Y difficult. The problem is that with multicollinearity the more data doesn't help decipher signal and noise if the covariance of the two variables is near 1.

16. Look up the Frisch Waugh lovell theorem on google. What is this theorem?

It's separating out the residuals of a regressor on a control variable and the dependent variable on the same control variable and essentially taking the difference so you have X_1 stripped of any residuals resulting from the control and Y stripped of any noise coming from the control so there's a clean comparison.

17. How does this demonstrate the held constant principle of multiple linear regression?

Because this allows the data to be examined without the noise coming from other variables

Readings: Stock & Watson Chapter 7.1-7.4

1. Is there anything especially different about testing a single estimate hypothesis in a multivariate model?

No.

2. Is there anything especially different about creating a single estimate confidence interval in a multivariate model?

No. The mechanics are the same unless discussing joint hypotheses.

3. What would be the correct interpretation of an f-test on the set of regressors in a model?

When more hypotheses are jointly made, it restricts the solution space for the model. The F-stat is the ratio of the restricted SSR to the unrestricted which reflects the models fit compared to the unrestricted model's fit.

4. Why is it not appropriate to jointly test that every estimate is equal to zero but individually testing each estimate with a t-test?

Multiple t-tests would compound the significance level required to make determinations about hypotheses which in turn leads to higher rejections rates.

5. How many sets of degrees of freedom are involved in a f-statistic?

There are two sets of df in F-stats.

6. Why might you want to test the hypothesis that two regression coefficients are equal?

It can test if the two coefficients have the same effects on the population.

7. Why might you want to construct a confidence set?

The confidence set for joint hypotheses would give the range of values for which the relationship of two coefficients hold.