Recitation 4: Multivariate OLS

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

October 6th, 2022

Gauss-Markov Theorem

Gauss-Markov Theorem: Big picture

- Assuming classical linear regression assumptions, OLS is the unbiased, linear estimator with the smallest variance (OLS = BLUE)

Gauss-Markov Theorem: Statement and condition

Assuming the following conditions

- Conditional expectation is zero: $E[u_i|X_i] = 0$
- Homoskedasticity: $var(u_i|X_i) = \sigma_u^2$
- No autocorrelation: $E[u_i u_j|X_i] = 0$ for $i \neq j$

OLS is best, linear, unbiased estimator (BLUE)

OLS is linear and unbiased

- Linearity: We can write the numerator from $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ as

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} Y_i + n\bar{X}\bar{Y}$$

  We use the fact that $\sum_{i=1}^{n} X_i = n\bar{X}$ to reduce the above to

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i - \bar{X}\sum_{i=1}^{n} Y_i + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - \bar{X}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$$

  Thus, OLS estimator is linear in Y_i

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} c_i Y_i$$

- Unbiasedness: We have shown this when we talked about the sampling distribution of $\hat{\beta}_1$. The same logic holds here

OLS has the smallest variance out of unbiased, linear estimators

- Let $\tilde{\beta}_1$ be linear in Y_i and unbiased ($\tilde{\beta}_1 = \sum_{i=1}^n a_i Y_i$). We can write

$$\tilde{\beta}_1 = \sum_{i=1}^n a_i(\beta_0 + \beta_1 X_i + u_i) = \sum_{i=1}^n a_i \beta_0 + \beta_1 \sum_{i=1}^n a_i X_i + \sum_{i=1}^n a_i u_i$$

- Since this is also unbiased, we can infer that (note that this applies to c_i too)

$$E\left[\sum_{i=1}^n a_i u_i | X_i\right] = 0 \sum_{i=1}^n a_i = 0 \quad \sum_{i=1}^n a_i X_i = 1$$

- We have

$$var(\tilde{\beta}_1|X_i) = var(\beta_1 + \sum_{i=1}^n a_i u_i|X_i) = \sum_{i=1}^n a_i^2 var(u_i|X_i) = \sum_{i=1}^n a_i^2 \sigma_u^2$$

$$var(\hat{\beta}_1|X_i) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n c_i^2 \sigma_u^2 (\text{Verify on your own})$$

OLS has the smallest variance out of unbiased, linear estimators

- So the difference in variances are

$$var(\tilde{\beta}_1|X_i) - var(\hat{\beta}_1|X_i) = \sigma_u^2 \sum_{i=1}^{n}(a_i^2 - c_i^2)$$

- Let $a_i = c_i + d_i$. Then we have

$$\begin{aligned}
\sum_{i=1}^{n} a_i^2 &= \sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} c_i d_i \\
&= \sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2 + 2\frac{\sum_{i=1}^{n}(X_i - \bar{X})d_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2 + 2\frac{\sum_{i=1}^{n} X_i a_i - \sum_{i=1}^{n} X_i c_i - \bar{X}\sum_{i=1}^{n} a_i + \bar{X}\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} d_i^2 (\because \text{ Conditions we set for } a_i)
\end{aligned}$$

OLS has the smallest variance out of unbiased, linear estimators

- Therefore,

$$var(\tilde{\beta}_1|X_i) - var(\hat{\beta}_1|X_i) = \sigma_u^2 \sum_{i=1}^{n} d_i^2 \geq 0$$

- Takeaway: Among the class of unbiased and linear estimators, OLS has the smallest variance of them all.
- Note that this also implies that there is an estimator with smaller variance but such estimator has a bias or is nonlinear (e.g. LASSO)

Multivariate OLS

Multivariate Regression: Why add more variables to the right?

- Suppose that there are more than one possible independent variable
- The set of models are

$$\begin{aligned}\text{True: } Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \\ \text{Mistake: } Y_i &= \beta_0 + \beta_1 X_i + u_i^* \\ \text{Sample: } Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i\end{aligned}$$

- Suppose you run an OLS regression without Z_i. $\hat{\beta}_1$ can be calculated as $\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Replacing this with the true model gives

$$\begin{aligned}\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_1 (X_i - \bar{X}) + \beta_2 (Z_i - \bar{Z}) + (u_i - \bar{u}))}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^n (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

Omitted variable bias: Bias from missing out needy independent variables

- If $\beta_2 \neq 0$ and $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \neq 0$, then the mean of $\hat{\beta}_1$ is not guaranteed to be β_1. This leads to the **omitted variable bias** problem
- This happens when both of the following cases hold
  - Z should explain Y: If the slope coefficient of Z (β_2) is nonzero, then the Z variable is part of the error term if we forget to include them
  - Z is correlated with X: If $cov(X, Z) \neq 0$ and the regression residual $\hat{u}$ is correlated with X, the independent variable is now correlated with $\hat{u}$, which leads to violation of the assumption that independent variable and the residual are not correlated.
- We can even determine the direction of the bias
  - $\hat{\beta}_1$ is overestimated if $\beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} > 0$
  - $\hat{\beta}_1$ is underestiated if $\beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} < 0$

Omitted variable bias: What to do about it?

- We can simply include the Z variable if we have the data for it.
- Another way is to conduct an ideal randomized controlled experiment (or randomized control trial) that randomly assigns value of X to all students.
- If none of the two are feasible, we should find another variable that can be a proxy to Z - they have to be related to the X variable and is uncorrelated with the errors - which is the Instrumental Variable method.

Multivariate regression: So now what does the coefficients really mean?

- The technicalities involved do not change drastically compared to the univariate regression.
- However, one should interpret the coefficients cautiously. Suppose that the regression is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- To see the impact of X_i and Y_i, one needs to take (partial) derivatives on Y_i with respect to X_i. This leads to

$$\beta_1 = \frac{\partial Y_i}{\partial X_i}$$

- In words, β_1 captures how much Y_i changes with respect to X_i *holding other variables constant* (ceteris paribus).
- If you do not hold other variables (Z_i in this case) fixed, the change will not exactly be β_1 (it could be more or less)

Sapling distribution of estimates: Just derive this once in your life

- The estimates for the $\hat{\beta}_j$, can be obtained in a similar way in which we have obtained the OLS estimates for the single variable version.

$$\min_{\{\beta_0,\beta_1,\beta_2\}} \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}]^2$$

- After some more amount of algebra (than the single variable case), the result we get is the following

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(Y_i-\bar{Y}) \sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 - \sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)}{\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 \sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 - [\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)]^2} \\
\hat{\beta}_2 &= \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 - \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(Y_i-\bar{Y}) \sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)}{\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 \sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 - [\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)]^2}
\end{aligned}$$

- What matters at this point is how we should **interpret** these coefficients.

What new problems do we have in multivariate regressions?

- We are quite likely to end up including independent variables that are highly correlated with each other. There are two
- We say two variables X_1 and X_2 are **perfectly multicollinear** if X_1 is in an exact linear relationship of some sort with X_2.
- Any multicollinearities that are not in exact linear relationship is referred to as **imperfect multicollinearity**.

When does perfect multicollinearity happen?

- **Assume that $X_2 = cX_1$ for some constant c**: Then we have $(X_{2i} - \bar{X}_2) = c(X_{1i} - \bar{X}_1)$. Then $\hat{\beta}_1$ changes to $\frac{0}{0}$
- **Dummy variable trap**: Say that you have the dummy variable for females and males. Let each of them be X_{1i} and X_{2i} with $X_{2i} = 1 - X_{1i}$. Then the regression can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \iff Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2(1 - X_{1i}) + u_i$$
$$\iff Y_i = \beta_0 + \beta_2 + (\beta_1 - \beta_2)X_{1i} + u_i$$

Therefore, by including both X_{1i} and X_{2i} in the same regression, the X_{2i} vanishes from the equation. This is why when you have dummy variables for all categories in the observation, **one of them must be left out.**

Adjusted R^2 and why we need this

- Additional complication rises from interpreting the goodness of fit. In addition to R^2, we now get the **adjusted** R^2 (or $\bar{R}^2$), which is defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{RSS}}{\text{TSS}} \quad (k : \text{number of independent variables})$$

- Since we are assuming that $k \geq 1$, adjusted R^2 is smaller than the R^2.
- As we include more variables, the $\frac{n-1}{n-k-1}$ increases, leading to further decrease in $\bar{R}^2$. (There is no such factor in R^2, so it always increases even with irrelevant variable)
- However, if the new variables are very relevent, $\frac{\text{RSS}}{\text{TSS}}$ decreases quickly.
- This reduces the gap between R^2 and the adjusted R^2. If the adjusted R^2 do not decrease drastically, it is a sign that we are adding a relevant variable.

Joint hypothesis tests (Why it is not straightforward)

- Suppose that you are running a two-sided test with 5 independent variables and significance level $\alpha = 5\%$ under the null hypothesis

$$H_0 : \beta_1 = ...\beta_5 = 0$$

- You reject the null hypothesis when $|t_i| \geq 1.96$ with probability 0.05.
- Now assume that each test statics are independent. Then the probability of incorrectly rejecting the null hypothesis using this approach is

$$\begin{aligned} \Pr(|t_1| > 1.96 \cup ... \cup |t_5| > 1.96) &= 1 - \Pr(|t_1| \leq 1.96 \cap .. \cap |t_1| \leq 1.96) \\ (\because \text{Independence of } t_i\text{'s}) &= 1 - \Pr(|t_1| \leq 1.96) \times .. \times \Pr(|t_5| \leq 1.96) \\ &= 1 - (0.95)^5 \\ &= 0.2262 \end{aligned}$$

- This means that the rejection rate under the null is not 5% but 22% percent - we end up rejecting the null hypothesis more than we have to.

F-test for joint significance

- This is a test where all parts of the joint hypothesis can be tested at once. It also has mechanism for correcting the correlation between the t-test statistics.
- It ultimately allows us to correctly set the significance level even for the multiple testing case.
- The usual joint hypothesis test for the regression with k variables (not including the constant term) is

$$H_0 : \beta_1 = ... = \beta_k = 0, \ H_1 : \neg H_0$$

  where H_1 refers to the case where there is a nonzero element in any one of β_1 to β_k.
- Note that the default F-test null hypothesis for STATA is as above

Tricks for deriving F-statistics

- Assuming $H_0 : \beta_1 = ... = \beta_q = 0$ hypothesis, we use R^2 from the 'unrestricted' and 'restricted' regressions.
  - Assume the following setup

$$\begin{aligned}\text{Restricted: } Y_i &= \beta_0 + 0X_{1,i} + ... + 0X_{q,i} + \beta_{q+1}X_{q+1,i} + ... + \beta_k X_{k,i} + u_i \\ \text{Unrestricted: } Y_i &= \beta_0 + \beta_1 X_{1,i} + ... + \beta_q X_{q,i} + \beta_{q+1}X_{q+1,i} + ... + \beta_k X_{k,i} + u_i\end{aligned}$$

  - Restricted regression assumes that H_0 is true and then only optimizes with respect to $\beta_{q+1}, ..., \beta_k$.
  - Unrestricted regression does not assume that H_0 is true and optimizes with respect to all slope coefficients.

Tricks for deriving The F-statistic

- We use R^2 from these two regressions.

$$\frac{(R^2_{\text{Unrestricted}} - R^2_{\text{Restricted}})/q}{(1 - R^2_{\text{Unrestricted}})/(n-k-1)}$$

  - k: number of independent variables (not counting intercept)
  - q is the number of restrictions.
  - Since unrestricted models allows roles for $X_1, ..., X_q$ variables, they have higher R^2 (Restricted: They should have no role)
- Another: Using $R^2_{\text{Restricted}} = 1 - \frac{RSS_{\text{Restricted}}}{TSS}$, we can write

$$\frac{(RSS_{\text{Restricted}} - RSS_{\text{Unrestricted}})/q}{(RSS_{\text{Unrestricted}})/(n-k-1)}$$

Other tests in multivariate regressions

- Suppose that instead of β_1 and β_2 being zero, we are just interested in whether they are equal.
- The F-test can also be used for testing this hypothesis. The setup of the hypothesis would be

$$H_0 : \beta_1 = \beta_2 \; H_1 : \beta_1 \neq \beta_2$$

- With this, you can answer various types of tests (e.g. is $\beta_1 + \beta_2 = 100$?)