14.32 Recitation 4

Nina Wang

MIT Department of Economics

Lectures 8-9

Table of Contents

Table of Contents

Multivariate Regressions

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i$$

- We can interpret β_1 as the expected effect on Y of a one-unit change in X_1 holding X_2 constant.
- In theory, we are still solving for the coefficients using the same objective of minimizing the sum of the squared residuals.

$$\underset{\beta_0,\beta_1,...,\beta_k}{\text{argmin}} \ \Sigma_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2$$

Multivariate Regressions

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i$$

- In theory, we are still solving for the coefficients using the same objective of minimizing the sum of the squared residuals.

$$\underset{\beta_0, \beta_1, ..., \beta_k}{\text{argmin}} \ \Sigma_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2$$

- However, this gets unwieldy pretty quickly (aka after you have more than 2 regressors). Thus, it's easier to solve for these coefficients using matrix notation.

Multivariate Regressions

$$\underset{\beta_0,\beta_1,...,\beta_k}{\text{argmin}} \ \Sigma_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2$$

- For a regression with 2 covariates (another way of saying regressors), solving this minimization gives us

Multivariate Regressions

$$\underset{\beta_0,\beta_1,\ldots,\beta_k}{\text{argmin}} \ \Sigma_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_k X_{k,i})^2$$

- For a regression with 2 covariates (another way of saying regressors), solving this minimization gives us

$$\alpha = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 \qquad\qquad \beta_1 = \frac{Cov(Y_i, \tilde{x}_{1,i})}{Var(\tilde{x}_{1,i})}$$

- where $X_{1,i} = \gamma_0 + \gamma_1 X_{2,i} + \tilde{x}_{1,i}$ is used to partial out (remove) the influence of $X_{2,i}$ on $X_{1,i}$

Multivariate Regressions

$$\underset{\beta_0,\beta_1,\ldots,\beta_k}{\text{argmin}} \ \Sigma_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_k X_{k,i})^2$$

- For a regression with 2 covariates (another way of saying regressors), solving this minimization gives us

$$\alpha = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 \qquad\qquad \beta_1 = \frac{Cov(Y_i, \tilde{x}_{1,i})}{Var(\tilde{x}_{1,i})}$$

- where $X_{1,i} = \gamma_0 + \gamma_1 X_{2,i} + \tilde{x}_{1,i}$ is used to partial out (remove) the influence of $X_{2,i}$ on $X_{1,i}$
- However, this gets unwieldy pretty quickly (aka after you have more than 2 regressors). Thus, it's easier to solve for these coefficients using matrix notation.

Multivariate Regressions

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{k,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{k,2} \\ & & \dots & & \\ 1 & X_{1,n} & X_{2,n} & \dots & X_{k,n} \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}.$$

- Our new model is $Y = X\beta + e$.

Multivariate Regressions

- Example with 3 regressors and 4 observations, giving us a total of 12 datapoints:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}, X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & X_{3,1} \\ 1 & X_{1,2} & X_{2,2} & X_{3,2} \\ 1 & X_{1,3} & X_{2,3} & X_{3,3} \\ 1 & X_{1,4} & X_{2,4} & X_{3,4} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Multivariate Regressions

- Example with 3 regressors and 4 observations, giving us a total of 12 datapoints:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}, X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & X_{3,1} \\ 1 & X_{1,2} & X_{2,2} & X_{3,2} \\ 1 & X_{1,3} & X_{2,3} & X_{3,3} \\ 1 & X_{1,4} & X_{2,4} & X_{3,4} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

- Multiplying out $X\beta$ gives us $\begin{pmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{2,1} + \beta_3 X_{3,1} \\ \beta_0 + \beta_1 X_{1,2} + \beta_2 X_{2,2} + \beta_3 X_{3,2} \\ \beta_0 + \beta_1 X_{1,3} + \beta_2 X_{2,3} + \beta_3 X_{3,3} \\ \beta_0 + \beta_1 X_{1,4} + \beta_2 X_{2,4} + \beta_3 X_{3,4} \end{pmatrix}$

Multivariate Regressions

- We want to minimize the sum of the squared residuals $e'e$

$$
\begin{aligned}
e'e &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}
\end{aligned}
$$

Multivariate Regressions

- We want to minimize the sum of the squared residuals $e'e$

$$\begin{aligned} e'e &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

- Taking the derivative wrt $\hat{\beta}$ and setting to 0, we have

$$\begin{aligned} &X'Y - X'X\hat{\beta} = 0 \\ &X'Y = X'X\hat{\beta} \\ &(X'X)^{-1}X'Y = (X'X)^{-1}X'X\hat{\beta} \\ &\hat{\beta} = (X'X)^{-1}X'Y \end{aligned}$$

Multiple Regressors vs. Controls

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i$$

vs

$$Y_i = \beta_0 + \beta X_i + \gamma Z_i + e_i$$

- We distinguish our β_1, β_2, etc. as coefficients/parameter of interest vs. controls.

Multiple Regressors vs. Controls

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i$$

vs

$$Y_i = \beta_0 + \beta X_i + \gamma Z_i + e_i$$

- We distinguish our β_1, β_2, etc. as coefficients/parameter of interest vs. controls.
  - For coefficients of interest, we are interested in the causal effect between X and Y, must worry about OVB, etc.
  - For controls, we add to minimize OVB, but we are not actually interested in estimating that relationship.

Multiple Regressors vs. Controls

- You might see a regression model described as:
  $Y_i = \beta_0 + \beta X_i + \gamma Z_i + e_i$ where Z_i is a vector of controls.

Multiple Regressors vs. Controls

- You might see a regression model described as:
  $Y_i = \beta_0 + \beta X_i + \gamma Z_i + e_i$ where Z_i is a vector of controls.

- Essentially, here $\gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & \ldots & \gamma_n \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_{1,i} \\ Z_{2,i} \\ \ldots \\ Z_{n,i} \end{bmatrix}$

Multiple Regressors vs. Controls

- You might see a regression model described as:
  $Y_i = \beta_0 + \beta X_i + \gamma Z_i + e_i$ where Z_i is a vector of controls.

- Essentially, here $\gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & ... & \gamma_n \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_{1,i} \\ Z_{2,i} \\ ... \\ Z_{n,i} \end{bmatrix}$

- We can see that $\gamma Z_i = \begin{bmatrix} \gamma_1 Z_{1,i} + \gamma_2 Z_{2,i} + ... + \gamma_3 Z_{3,i} \end{bmatrix}$

A note on R^2

- Although R^2 can always be used as a measure of fit, we do not really look at R^2 when determining whether a model is "good" or not

A note on R^2

- Although R^2 can always be used as a measure of fit, we do not really look at R^2 when determining whether a model is "good" or not
- One reason for this is we don't really use regressions as a predictive model, more so to estimate the relationship between variables.

A note on R^2

- Although R^2 can always be used as a measure of fit, we do not really look at R^2 when determining whether a model is "good" or not
- One reason for this is we don't really use regressions as a predictive model, more so to estimate the relationship between variables.
- Also when dealing with multivariate regressions, be mindful that R^2 will **always increase when you add a regressor**, even if adding the regressor biases the coefficient of interest.

A guide to adding covariates/controls

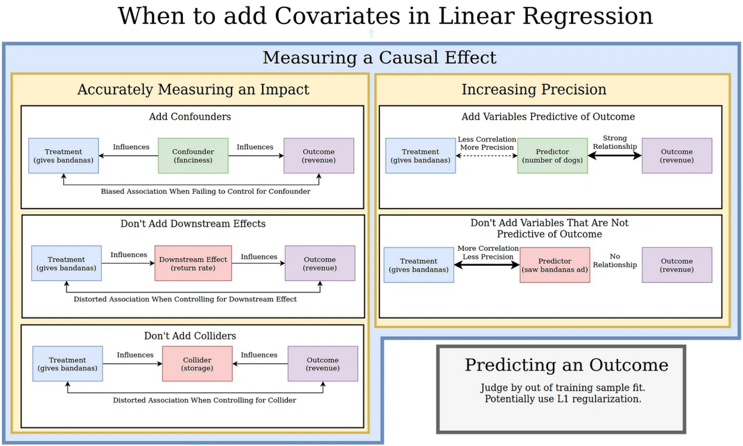

Table of Contents

F-test

- When testing the joint hypothesis that multiple coefficients are 0, we use an F-test.
- To test coefficients in Stata, use the `test` command followed by the variables whose coefficients you want to test.
- Generally, if $F > 2.5$, you reject

```
. test height sex educ

 ( 1)  height = 0
 ( 2)  sex = 0
 ( 3)  educ = 0

       F(  3, 17866) = 1085.81
            Prob > F =    0.0000
```

Table of Contents

Homework 2: 1) c

We are given regression $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

- How can we test the hypothesis $H_0 : 3\beta_1 = 4\beta_2$?

Stata Example: Running multivariate regressions

Table 1: Earnings vs. Height

	(1) Simple	(2) With Controls	(3) Male	(4) Female
height	707.7***	377.0***	93.52	683.5***
	(50.49)	(66.05)	(92.12)	(94.73)
Education of Individual		3836.0***	3908.8***	3732.2***
		(70.87)	(99.71)	(100.7)
Sex		552.3		
		(524.1)		
1:Northeast		0	0	0
		(.)	(.)	(.)
2:Midwest		-4177.7***	-3839.9***	-4714.4***
		(548.7)	(728.6)	(833.3)
3:South		-6001.0***	-6669.2***	-5209.0***
		(523.0)	(696.1)	(792.3)
4:West		-2238.6***	-2116.0**	-2371.2**
		(570.8)	(775.8)	(843.1)
Constant	-512.7	-27042.5***	-9636.9	-46648.5***
	(3386.9)	(4227.2)	(5949.4)	(6511.6)
Observations	17870	17870	9974	7896
R^2	0.011	0.161	0.149	0.174
Adjusted R^2	0.011	0.161	0.149	0.173
rmse	26777.2	24664.5	24762.2	24506.6

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Stata Example: Running multivariate regressions

- What is the estimated effect of height on earnings in the univariate regression (regression with no controls)? **707.7 dollars per inch**
- What is the estimated effect of height on earnings in regression 2? **377 dollars per inch**
- Use regression 2 to estimate the difference between the earnings of a male with 12 years of schooling and a woman with 11 years of schooling (who are the same height). $3836 + 552.3 =$ **4388.3 dollars**
- Test (at the 5% significance level), that the effect of height on earnings is the same for men as women.
  - t-statistic $= \frac{683.5 - 93.52}{\sqrt{94.73^2 + 92.12^2}} = \frac{589.98}{132.136} = 4.465$. **We reject the null hypothesis.**