

L5: Multiple regression



Feng Li

`feng.li@cufe.edu.cn`

**School of Statistics and Mathematics
Central University of Finance and Economics**

What we have learned last time...

- ① Two-variable linear model without intercept.
- ② Scaling and units, standardizing...
- ③ Variations of two-variable model.

Today we are going to learn...

- 1 Three-variable model
- 2 OLS estimation of regression coefficients
- 3 The multiple coefficient of determination
- 4 The matrix form
- 5 Method of Moments
- 6 Multiple Regression Inference
- 7 Likelihood Ratio, Wald and Lagrange Multiplier Tests

Three-variable model

↳ Model and assumptions

- 1 The three-variable population regression function

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where β_2 and β_3 are called **partial regression coefficients**

- 2 The assumptions

- 1 Linear model in terms of parameters.
- 2 X_{2i} and X_{3i} are fixed and independent of error term
 $\text{cov}(X_{2i}, u_i) = \text{cov}(X_{3i}, u_i) = 0$
- 3 Zero expectation: $E(u_i | X_{2i}, X_{3i}) = 0$
- 4 Homoscedasticity: $\text{var}(u_i) = \sigma^2$
- 5 Error terms are not correlated: $\text{cov}(u_i, u_j) = 0$ for $i \neq j$
- 6 $n > p$ where $p = 3$ in this case.
- 7 $\text{var}(X_{2i}) \neq 0$ and $\text{var}(X_{3i}) \neq 0$.
- 8 No exactly linear relationship between X_{2i} and X_{3i} — **no multicollinearity**.
- 9 The model is correctly specified.

- 3 Assumptions **1-7** are the same as in two-variable model.

- 4 Why do we need two more assumptions **8-9** ?

Three-variable model

↳ Why multicollinearity is evil?

- 1 **No multicollinearity** means non of the regressors can be written as exact linear combinations of the remaining regressors. That means you should not be able to find $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$ such that

$$\lambda_1 X_{2i} + \lambda_2 X_{3i} = 0$$

- 2 But if you happen to have $\lambda_1 X_{2i} + \lambda_2 X_{3i} = 0$, what will happen to your model then?

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ &= \beta_1 + \beta_2 \left(-\frac{\lambda_3}{\lambda_2} X_{3i} \right) + \beta_3 X_{3i} + u_i \\ &= \beta_1 + \left(\beta_3 - \beta_2 \frac{\lambda_3}{\lambda_2} \right) X_{3i} + u_i \end{aligned}$$

- 3 You will in fact have a two-variable regression model.
- 4 This perfect collinearity will not likely to happen in real data analysis.
- 5 Multicollinearity only applies to **linear relationships** between regressors. Other situations like $X_{2i} = X_{3i}^2$ will not violate our assumptions.

Three-variable model

↪ How do you interpret the model?

- 1 We always use the conditional mean

$$E(Y_i | X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

- 2 Example p. 191

Three-variable model

↪ OLS estimation of regression coefficients

- ① The sample regression function is

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

- ② The OLS is aiming to minimize

$$RSS = \sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2$$

- ③ Differentiating with respect to β_i , $i = 1, 2, 3$ and set to zero yields

$$\begin{aligned}\bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_{2i} + \hat{\beta}_3 \bar{X}_{3i} \\ \sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} \\ \sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2\end{aligned}$$

- ④ Solving the preceding formulas show that

$$\hat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum y_i x_{3i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}, \hat{\beta}_3 = \frac{\sum y_i x_{3i} \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}.$$

Then $\hat{\beta}_1$ is easily obtained.

Three-variable model

↪ Variance of regression coefficients

- ① Let r_{23} be the correlation coefficient between X_2 and X_3 , $r_{23}^2 = \frac{(\sum x_{2i} x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2}$.
The variance for β_i are

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2} \right] \sigma^2$$

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2} \sigma^2 = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2} \sigma^2 = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

- ② And the covariance between $\hat{\beta}_2$ and $\hat{\beta}_3$ is

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23} \sigma^2}{(1 - r_{23}^2) \sqrt{\sum x_{2i}^2} \sqrt{\sum x_{3i}^2}}$$

- ③ σ^2 is not known and estimated via $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$

Three-variable model

↳ Properties of OLS

- 1 The regression line passes through the mean \bar{Y} , \bar{X}_2 and \bar{X}_3 .
- 2 The mean value of the estimated Y_i is equal to the mean of the actual Y_i (**Why?**)
- 3 $\sum \hat{u}_i = \tilde{u} = 0$, **why?**
- 4 $\sum \hat{u}_i X_{2i} = \sum \hat{u}_i X_{3i} = \sum \hat{u}_i Y_i = 0$
- 5 $r_{23} \rightarrow 1$, $\hat{\beta}_2 \rightarrow ?$, $\text{var}(\hat{\beta}_2) \rightarrow ?$. $r_{23} \rightarrow 0$, $\hat{\beta}_2 \rightarrow ?$, $\text{var}(\hat{\beta}_2) \rightarrow ?$
- 6 The OLS estimator is the **best linear unbiased estimator** (BLUE).

Three-variable model

↪ R^2 and adjusted R^2

- ① Define the **multiple coefficient of determination** R^2 as

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} = 1 - \frac{(n-3)\hat{\sigma}^2}{(n-1)S_y^2}$$

- ② $R^2 \rightarrow 1$ means ?
- ③ What will happen if you increase the regressors? R^2 is increasing **which is bad (why?, see 4)**.
- ④ R^2 is not comparable for different models.
- ⑤ The adjusted R^2 ,

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i / (n-k)}{\sum y_i^2 / (n-1)} = 1 - (1 - R^2) \frac{n-1}{n-k} = 1 - \frac{\hat{\sigma}^2}{S_y^2}$$

where k = number of parameters. The adjusted R^2 can be used for comparing two models.

- ⑥ Think about r^2 in two-variable regression: which is both goodness of fit coefficient and correlation coefficient. (**Read p. 213**)

The matrix form I

- ① For a model with more than two regressors,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

we write the matrix form as

$$y = X\beta + u$$

where y is the $n \times 1$ response vector, X is the $n \times k$ covariate matrix (each column corresponds to a single covariate, the first column is just a vector of ones if the intercept is included), β is $k \times 1$ coefficient vector, and u is the $n \times 1$ error term vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

where model with intercept can be viewed the first column of X contains only ones.

The matrix form II

② $E(\mathbf{u}) = \mathbf{0}$

$$E \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

③ $E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}$ where \mathbf{I} is an $n \times n$ identity matrix.

$$\begin{aligned} E(\mathbf{u}\mathbf{u}') &= E \left[\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} (u_1 \quad u_2 \quad \dots \quad u_n) \right] \\ &= \begin{pmatrix} E(u_1^2) & E(u_1u_2) & \dots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \dots & E(u_2u_n) \\ \vdots & & & \\ E(u_nu_1) & E(u_nu_2) & \dots & E(u_n^2) \end{pmatrix} = \sigma^2\mathbf{I} \end{aligned}$$

The matrix form III

4 The OLS estimation

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$$\hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

$$= y'y - \hat{\beta}'X'y = \sum \hat{u}_i^2$$

$$\hat{\sigma}^2 = \hat{u}'\hat{u}/(n - k) = (y'y - \hat{\beta}'X'y)/(n - k) \text{ **Verify this!**}$$

Details can be found in **Appendix C: Matrix approach**)

5 The hat matrix

In the matrix form, we have the fitted value $\hat{y} = X\hat{\beta}$. We then have

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'y \\ &= [X(X'X)^{-1}X']y \\ &= Hy\end{aligned}$$

The matrix form IV

where H is the so-called **hat matrix**.

⑥ Some properties of the hat matrix

- The hat matrix is also called projection matrix— it maps the observed vector (y) to the fitted value (\hat{y}).
- The hat matrix is symmetric ($H' = H$) and idempotent ($H^2 = H$) in the linear regression (**verify this!**).
- The trace of the hat matrix equals the number of independent parameters (k) of the linear model which is the rank of covariate matrix (X).

⑦ Predictions

Estimation with method of moments I

↳ A two-variable example

- In the **population regression function**, we see that the error term u has zero expected value $E(u) = 0$ and that the covariance between X and u is zero $Cov(X, u) = 0$ which implies

$$E(u) = 0$$

$$E(Xu) = 0$$

- In terms of the observable variables x and y and the unknown parameters β_0 and β_1 , the above equation can be written as

$$E(Y - \beta_0 - \beta_1 X) = 0$$

$$E[X(Y - \beta_0 - \beta_1 X)] = 0$$

Estimation with method of moments II

↳ A two-variable example

- The above equations can be used to obtain good estimators of β_0 and β_1 if we change the expectation with its sample mean by **given a sample of data**

$$\frac{1}{n} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{1}{n} \sum [X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)] = 0$$

- These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.
 - From the first equation we have $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. And plugging it into the second equation yields

$$\frac{1}{n} \sum [X_i (Y_i - \bar{Y} - \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)] = 0$$

Estimation with method of moments III

↳ A two-variable example

which gives

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

- ② Now we go back to the first equation and solve the intercept.
- This is the **method of moments** approach to parameter estimation.
 - ① It relies on the sample average as an unbiased estimator of the population average and the sample variance as an unbiased estimator of the population variance.
 - ② The only assumption needed to compute the estimates for a particular sample X should not be a constant, which is obvious.
- Generally, **method of moments** estimation is to replace the population moment with its sample counterpart as follows.
- The parameter θ is shown to be related to some expected value in the distribution of Y , usually $E(Y)$ or $E(Y^2)$.

Estimation with method of moments IV

↳ A two-variable example

- Suppose, for example, that the parameter of interest θ , is related to the population mean as $\theta = g(\mu)$ for some function g .
- Because the sample average \bar{Y} is an unbiased and **consistent** estimator of μ , it is natural to replace μ with \bar{Y} , which gives us the estimator $g(\bar{Y})$ of θ .
- The estimator $g(\bar{Y})$ is consistent for θ , and if $g(\cdot)$ is a linear function of θ , then $g(\bar{Y})$ is unbiased as well.
- The **matrix form** of method of moments in linear regression:
<http://russell.vcharite.univ-mrs.fr/GMMboot/chp09big.pdf>

Multiple Regression Inference

↳ Hypothesis testing

- 1 For a model with more than two regressors,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

- 2 Testing individual regression coefficients: **the t test**

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0 \text{ or } H_a : \beta_i > 0, \text{ or } H_a : \beta_i < 0$$

- 3 Testing overall significance: **the t test**

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_a : \text{otherwise}$$

- 1 Method A: Do a lot of t test
- 2 Method B: ANOVA table (**the F test**)

Multiple Regression Inference

↪ Hypothesis testing: Testing overall significance

Source of Variation	SS	df	MSS
ESS	$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$	$k - 1$	$\sum \hat{y}_i^2 / (k - 1)$
RSS	$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$	$n - k$	$\sum \hat{u}_i^2 / (n - k)$
TSS	$\sum y_i^2 = \sum (Y_i - \bar{Y})^2 = \sum y_i^2 + \sum \hat{u}_i^2$	$n - 1$	

where k is number of parameters in the unrestricted model.

$$F = \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = ?$$
$$= \frac{n - k}{k - 1} \frac{R^2}{1 - R^2}$$

Multiple Regression Inference

↪ Testing the equality of two regression coefficients

- 1 For a model with more than two regressors,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

- 2 We want to test e.g.

$$H_0 : \beta_3 = \beta_5 \text{ or } \beta_3 - \beta_5 = 0$$

- 3 Under the classical assumption, we have

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_5) - (\beta_3 - \beta_5)}{se(\hat{\beta}_3 - \hat{\beta}_5)}$$

notice that

$$\text{var}(\hat{\beta}_3 - \hat{\beta}_5) = \text{var}(\hat{\beta}_3) + \text{var}(\hat{\beta}_5) - 2\text{cov}(\hat{\beta}_3, \hat{\beta}_5).$$

- 4 Then just do the usual t test.
- 5 **Question:** What are the degrees of freedom in your t statistic?

Multiple Regression Inference

↪ The general F test

- 1 For a model with more than two regressors,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$$

- 2 We want to test e.g.

$$H_0 : \beta_2 = \beta_3 \text{ or}$$

$$H_0 : \beta_3 + \beta_4 + \beta_5 = 3$$

- 3 If we assume the big model as unrestricted model (**UR**) and the restricted model (**R**) where H_0 satisfied.
- 4 For the two models,

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n - k)} = \frac{(\sum \hat{u}_R^2 - \sum \hat{u}_{UR}^2)/m}{\sum \hat{u}_{UR}^2/(n - k)} = \frac{(R_{UR}^2 - R_R^2)/m}{(1 - R_{UR}^2)/(n - k)}$$
$$\sim F(m, n - k)$$

where m = number of linear restrictions.

Likelihood Ratio, Wald and Lagrange Multiplier Tests I

- The **likelihood ratio (LR) test** is based on the maximum likelihood (ML) principle.
- Under the assumption that the disturbances u_i are normally distributed, we showed that, for the two-variable regression model, the OLS and ML estimators of the regression coefficients are identical, but the estimated error variances are different. The same is true in the multiple regression case.
- To illustrate the LR test, consider the three-variable regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

You will be able to write down the likelihood function as

$$\log \mathcal{L} = -n \log(\sigma) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i})^2$$

Likelihood Ratio, Wald and Lagrange Multiplier Tests II

- The null hypothesis: $\beta_3 = 0$, which gives the log likelihood function will then be

$$\log \mathcal{L} = -n \log(\sigma) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_{2i} - \beta_3 X_{3i})^2$$

which is known as the **restricted log-likelihood function** (RLLF) because it is estimated with the restriction that a priori β_3 is zero, whereas the previous is known as the **unrestricted log likelihood function** (ULLF).

- The LR test obtains the following test statistic

$$\lambda = 2(\text{ULLF} - \text{RLLF})$$

that follows the chi-square distribution with r degrees of freedom equal to the number of restrictions imposed by the null hypothesis.

Likelihood Ratio, Wald and Lagrange Multiplier Tests III

- Letting $RRSS$ and $URSS$ denote the restricted and unrestricted residual sums of squares. The LR test statistic can also be expressed as

$$-2\log(\lambda) = n(\log(RRSS) - \log(URSS))$$

which is distributed as χ^2 with r degrees of freedom where r is the number of coefficients omitted from the original model.

- The basic idea behind the LR test is simple:
 - If the a priori restriction(s) are valid, the restricted and unrestricted (log) LF should not be different, in which case λ will be zero.
 - But if that is not the case, the two LFs will diverge.
 - And since in a large sample we know that λ follows the chi-square distribution, we can find out if the divergence is statistically significant, say, at a 1 or 5 percent level of significance. Or else, we can find out the p value of the estimated λ .

Likelihood Ratio, Wald and Lagrange Multiplier Tests IV

- Letting $RRSS$ and $URSS$ denote the restricted and unrestricted residual sums of squares. We can have the **Wald statistic**

$$\frac{(n - k)(RRSS - URSS)}{URSS}$$

which is distributed as χ^2 with r degrees of freedom.

- Furthermore,

$$\frac{(n - k + r)(RRSS - URSS)}{RRSS}$$

where k is the number of regressors in the unrestricted model is known as the **Lagrange Multiplier statistic** which also follows χ^2 distribution with r degrees of freedom.

- Comparison of the three methods
 - All three are **asymptotically equivalent** (they give the similar answers in large samples).

Likelihood Ratio, Wald and Lagrange Multiplier Tests V

- But in small samples, the relationship among three test statistics are

$$\text{Wald} > \text{Likelihood Ratio} > \text{Lagrange Multiplier}$$

That means in small samples, a hypothesis can be rejected by the Wald but not rejected by Lagrange Multiplier.

- The three test statistics can be applied to test nonlinear hypothesis in linear models.
- They can be used for testing restrictions on variance-covariance matrices.
- They can also be applied to the models where the error term is not normally distributed.
- The choice of the three test statistics depends on the computational convenience

Take home questions

- 1 Read the partial correlation coefficients **p. 213.**
- 2 Read the **Chow test p. 254.**
- 3 Verify the BLUE property of OLS estimator with matrix form (**Appendix CA.4**)
- 4 Compare the three approaches in parameter estimation: OLS, MLE and method of moments.
- 5 Exercises (Set 3): **7.10, 7.14, 7.20, 8.2, 8.3, 8.6, 8.7, 8.11, 8.19, 8.20, C.10(p.863)**
- 6 Redo **Example 8.3** with maximum likelihood estimation and carry out likelihood ratio test, Wald test, and Lagrange Multiplier test.