Wang Yanan Institute for Studies in Economics

Microeconometrics Course Report#2

Notes on Multivariate Linear Regression and Gauss-Markov Theorem

Student
Baohao Wei
15220172202657

Supervisor
Jiaming Mao

Summary

In this report, we will study some important properties for multivariate linear regression. First, we discuss how to derive the ordinary least squares estimator for our regression model. Then, we will study the asymptotic distribution of our estimator and how to do joint hypothesis testing. At last, we will show that our OLS estimator is the Best Linear Unbiased Estimator(BLUE) of the true parameter under some conditions and introduce Gauss-Markov theorem.

Keywords: linear regression; asymptotic distribution; joint hypothesis testing; Gauss-Markov theorem

1 Multivariate linear regression model

Here, we formally introduce the multivariate linear regression model in matrix form:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = X\beta + U = \begin{pmatrix} x_{11} \cdots x_{1p} \\ \vdots \\ x_{n1} \cdots x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \tag{1.1}$$

where y is $n \times 1$ response vector, X is $n \times p$ predictor matrix, β is $p \times 1$ parameter vector and U is $n \times 1$ error term vector[1]. Here, we have n sample points and the dimension of parameters is p. Also, we use X_i to denote the ith sample point: $X_i = (x_{i1}, ..., x_{ip})$ and our model can be written as $Y_i = X_i\beta + u_i$. Now, we introduce the least squares assumptions for ordinary least squares estimation.

Least Squares Assumptions. Consider the multivariate linear regression model: $Y_i = X_i\beta + u_i$, the least squares assumptions are: (1) $\mathbb{E}(u_i|X_i) = 0$; (2) $\{(X_i, y_i)\}_{i=1}^n$ are independently and identically distributed draws from their joint distribution; (3) X_i and u_i have finite fourth moments; (4) X has full column rank; (5) $Var(u_i|X_i) = \sigma_u^2$(homoskedasticity).

1.1 The Ordinary Least Squares estimator

Now, we consider estimating the true parameter vector β. The OLS estimation use l_2 loss of the distance as loss function. The estimator for β is:

$$\hat{\beta} = \underset{\beta}{argmin}\, L(\beta) = \underset{\beta}{argmin}(y - X\beta)^T(y - X\beta). \tag{1.2}$$

To derive the solution for OLS estimator, first we expand our loss function as:

$$\begin{aligned} L(\beta) &= \left(y^T - \beta^T X^T\right)(y - X\beta) \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta. \end{aligned} \tag{1.3}$$

Here we can define the matrix derivatives. Define the derivative of a vector function $f(x) = (f_1(x), ..., f_m(x))$ with respective to a multivariate n-dimensional random vectors x as following:

$$\frac{\partial f(x)}{\partial x} = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ & \vdots & \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}.$$

Then, we take the derivative[2] with respective to β and let it be 0:

$$\begin{aligned} \frac{\partial L(\beta)}{\partial \beta} &= -X^T y - X^T y + 2X^T X\beta = 0 \\ &\Rightarrow 2X^T X\beta = 2X^T y \\ &\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y. \end{aligned} \tag{1.4}$$

Note that the second derivative of $L(\beta)$ with respective to β is $\frac{\partial^2 L(\beta)}{\partial \beta^2} = X^T X$, which is semi-positive definite. This proves $\hat{\beta}$ is the minimizer of the loss function.

[1] We assume that both X and y have been standardized such that they have mean 0 and variance 1. So there is no intercept term in our regression model.

[2] If A is a $n \times n$ matrix and x is a $n \times 1$ vector, then

$$\frac{\partial(Ax)}{\partial x} = A \quad \text{and} \quad \frac{\partial(x^T Ax)}{\partial x} = 2Ax,$$

the first result still holds if A is a general $m \times n$ matrix.

1.2 Asymptotic distribution of $\hat{\beta}$

This section we study the asymptotic distribution of our OLS estimator. Note that $\hat{\beta} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta + U) = \beta + (X^TX)^{-1}X^TU$ under the true regression model that $y = X\beta + U$, which gives us:

$$\sqrt{n}(\hat{\beta} - \beta) = (\frac{X^TX}{n})^{-1}\frac{X^TU}{\sqrt{n}}. \tag{1.5}$$

First we consider the asymptotic distribution of $(X^TX/n)^{-1}$. We aim to show that $(X^TX/n)^{-1}$ converges to Q_X in probability, where $Q_X = \mathbb{E}(X_iX_i^T)$. Note that we have $(X^TX/n)^{-1} = n^{-1}\sum_{i=1}^n X_iX_i^T$.

Here, we use $(A)_{kl}$ to denote the (k,l)th element in matrix A. To show our conclusion, it is equivalent to show that $n^{-1}\sum_{i=1}^n (X_iX_i^T)_{kl}$ converges to $(Q_X)_{kl}$, which is equivalent to show that $n^{-1}\sum_{i=1}^n x_{ik}x_{il}$ converges to $\mathbb{E}(x_{ik}x_{il})$. We try to use the law of large number[1] to prove this. From our second least squares assumption, we can know that $\{x_{ik}x_{il}\}_{i=1}^n$ are indepedent with each other for any $k,l \in \{1,...,p\}$. And from the third least squares assumption, it's obvious that $\mathbb{E}(x_{ik}x_{il})$ is finite. And by Cauchy-Schwarz inequality[2], we have$\left|\mathbb{E}\left(x_{ik}^2x_{il}^2\right)\right| < \sqrt{\mathbb{E}(x_{ik}^4)\cdot\mathbb{E}(x_{il}^4)} < \infty$. So now we have $Var(x_{ik}x_{il})$ is finite. Then, by the law of large number,

$$(\frac{X^TX}{n})^{-1} \xrightarrow{p.} \mathbb{E}(X_iX_i^T). \tag{1.6}$$

Next, we show that $X^TX/\sqrt{n}$ converges to some normal distribution. We try to prove this using central limit theorem. Note that $X^TU = \sum_{i=1}^n X_iu_i = \sum_{i=1}^n V_i$, where $V_i = X_iu_i$. Obviously, v_i's are indepedent with each other under least squares assumptions. From the first least squares assumption, it's obvious that $\mathbb{E}(V_i) = \mathbb{E}\{\mathbb{E}(X_iu_i|X_i)\} = 0$. Then, we show that Σ_{V_i} is positive definite and finite. Note $(\Sigma_{V_i})_{kl} = Cov(u_ix_{ik}, u_ix_{il}) = u_i^2Cov(x_{ik}, x_{il})$. Then, $\Sigma_{V_i} = u_i^2\Sigma_{X_i}$. Also, for any vector y,

$$\begin{aligned} y^T\Sigma_Xy &= y^T(X_i - \mu_X)(X_i - \mu_X)^Ty \\ &= ((X_i - \mu_X)^Ty)^T(X_i - \mu_X)^Ty \geq 0. \end{aligned} \tag{1.7}$$

Then, Σ_{V_i} is positive semidefinite. And since $|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}$, we know that Σ_{V_i} is finite. Now we've shown that V_i's are independently distributed with mean 0 and variance Σ_{V_i}. By central limit theorem[3], we have

$$\frac{X^TU}{\sqrt{n}} \xrightarrow{d.} \mathcal{N}(0, \Sigma_V) \tag{1.8}$$

Finally, by Slutsky's theorem[4], we have the asymptotic distribution for our OLS estimator:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d.} \mathcal{N}(0, {Q_X}^{-1}\Sigma_V{Q_X}^{-1}) \tag{1.9}$$

[1] **Law of large number:** suppose $\{X_i\}_{i=1}^n$ is a series of independently and indentically distributed random variables with mean $\mathbb{E}(X)$ and finite variance, then $\overline{X} = \frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p.} \mathbb{E}(X)$.

[2] **Cauchy-Schwarz inequality:** for random variables X and Y, we have $|\mathbb{E}(XY)| < \sqrt{\mathbb{E}(X^2)\cdot\mathbb{E}(Y^2)}$

[3] **Multivariate central limit theorem:** suppose $\{X_i\}_{i=1}^n$ is a series of independently and indentically distributed m-dimensional random variables with mean $\mathbb{E}(X_i) = \mu_X$ and finite covariance matrix Σ_X which is also positive definite. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^n x_i$. Then,

$$\sqrt{n}(\overline{X} - \mu_X) \xrightarrow{d.} \mathcal{N}(\mu_X, \Sigma_X).$$

[4] **Slutsky's theorem:** if $\{X_i\}_{i=1}^n$ be a i.i.d sequence that converges to a random variable X with a distribution function $F(x)$ and if $\{Y_i\}_{i=1}^n$ is a i.i.d sequence that converges in probability to constant c. Then, X_iY_i is distributed asymptotically as Xc.

1.3 Test of joint hypothesis

In this section, we consider how to do hypothesis testing for our estimator. Consider joint hypothesis which is linear in our coefficients and impose q restrictions. We can write it in a matrix form:

$$R\beta = r, \tag{1.10}$$

where R is a $q \times n$ full rank matrix and r is $q \times 1$ restriction vector. Now we derive the F-statistic testing this joint hypothesis. Note that under null hypothesis, $\sqrt{n}(R\hat{\beta} - r) = \sqrt{n}R(\hat{\beta} - \beta) \sim \mathcal{N}(0, R\Sigma_{\sqrt{n}(\hat{\beta}-\beta)}R^T)^1$.

Futhermore, under null hypothesis we have

$$\begin{aligned} &\sqrt{n}(R\hat{\beta} - r)^T (R\Sigma_{\sqrt{n}(\hat{\beta}-\beta)}R^T)^{-1}\sqrt{n}(R\hat{\beta} - r) \\ &= (R\hat{\beta} - r)^T (R\hat{\Sigma}_{\hat{\beta}}R^T)^{-1}(R\hat{\beta} - r) \xrightarrow{p.} \mathcal{X}_q^2. \end{aligned} \tag{1.11}$$

Also note that $\hat{\Sigma}_{\hat{\beta}} \to \Sigma_{\beta}$. By Slutsky's theorem, the F-statistic testing this joint hypothesis is given by:

$$F = (R\hat{\beta} - r)^T (R\Sigma_{\beta}R^T)^{-1}(R\hat{\beta} - r)/q \xrightarrow{d.} F_{q,\infty} \tag{1.12}$$

2 Gauss-Markov theorem

2.1 Gauss-Markov conditions for multivariate regression

The Gauss-Markov conditions for multivariate regression are (1) $\mathbb{E}(U|X) = 0$; (2) $\mathbb{E}(UU^T|X) = \sigma_u^2 I$; (3) X has full rank, that is, $(X^TX)^{-1}$ exists.

2.2 Linear conditionally unbiased estimator

An estimator β^{est} is said to be linear if it's a linear combination of $\{y_i\}_{i=1}^n$, which can be written as $\beta^{est} = A^Ty$ where A is a $n \times p$ weight matrix. Note that A can depend on X but not depend on y. And we say that a linear estimator is conditionally unbiased if $\mathbb{E}(\beta^{est}|X) = \beta$. Note that our ordinary least squares estimator is also a linear conditionally unbiased estimator since $\hat{\beta} = (X^TX)^{-1}X^Ty = \beta + (X^TX)^{-1}X^TU$ where $A^T = (X^TX)^{-1}X^T$ and under our Gauss-Markov conditions, $\mathbb{E}[(X^TX)^{-1}X^TU|X] = 0$, which means $\hat{\beta}$ is conditionally unbiased.

2.3 Gauss-Markov theorem for multivariate regression

Typically, Gauss-Markov theorem says that our ordinary least squares estimator has the smallest variance among all linear conditionally unbiased estimator, which means that our OLS estimator is the Best Linear Unbiased Estimator (BLUE). The theorem compares the variance between two vector estimators using the linear combination of the elements of the estimators, that's, we compare $Var(c^T\beta_1)$ and $Var(c^T\beta_2)$ for any vector c.

Now, we start to prove this theorem. Let β_1 be any linear conditionally unbiased estimator, that's, $\beta_1 = A^Ty$ and $\mathbb{E}(\beta_1|X) = \beta$. Under our regression model, we can know that $\beta_1 =$

[1] If V follows multivariate normal distribution with mean vetor μ_V and covariance matrix Σ_V, then for any general $p \times 1$ vector d and $p \times n$ matrix A, we have

$$d + AV \sim \mathcal{N}(d + A\mu_V, A\Sigma_V A^T) \quad \text{and} \quad (V - \mu_V)^T {\Sigma_V}^{-1}(V - \mu_V) \sim \mathcal{X}_n^2.$$

$A^T(X\beta + U) = A^T X\beta + A^T U$. Combine the Gauss-Markov condition and the unbiasness of β_1 we can know that

$$\begin{aligned}\mathbb{E}(\beta_1|X) &= \mathbb{E}(A^T X\beta|X) + \mathbb{E}(A^T U|X) \\ &= A^T X\mathbb{E}(\beta|X) + A^T\mathbb{E}(U|X) \\ &= A^T X\mathbb{E}(\beta|X) = \beta.\end{aligned} \tag{2.1}$$

This indicates that $A^T X = I$. Also, we have the following conclusion:

$$\begin{aligned}Var(\beta_1|X) &= Var(A^T X\beta + A^T U|X) \\ &= Var(\beta + A^T U|X) = Var(A^T U|X) \\ &= \mathbb{E}(A^T U U^T A|X) - \mathbb{E}(A^T U|X)\mathbb{E}(A^T U|X) \\ &= A^T\mathbb{E}(UU^T|X)A = \sigma_u^2 A^T A.\end{aligned} \tag{2.2}$$

Note that for our least squares estimator, which is also a linear conditionally unbiased estimator, we have $\hat{A} = X(X^T X)^{-1}$. Now let $A = \hat{A} + D$, where D is the difference between weight matrix of $\hat{\beta}$ and β. Note that

$$\begin{aligned}\hat{A}^T D &= \hat{A}^T(A - \hat{A}) = \hat{A}^T A - \hat{A}^T\hat{A} \\ &= (X^T X)^{-1}X^T A - (X^T X)^{-1}X^T X(X^T X)^{-1} \\ &= (X^T X)^{-1} - (X^T X)^{-1} = 0 \quad (\text{Since } A^T X = I)\end{aligned} \tag{2.3}$$

Now we consider the variance of our estimator β_1,

$$\begin{aligned}Var(\beta_1|X) &= \sigma_u^2(\hat{A} + D)^T\hat{A} + D \\ &= \sigma_u^2(\hat{A}^T\hat{A} + \hat{A}^T D + D^T\hat{A} + D^T D) \\ &= \sigma_u^2(X^T X)^{-1} + \sigma_u^2 D^T D. \quad (\text{Since } \hat{A}^T D = 0)\end{aligned} \tag{2.4}$$

Finally, we consider the linear combination of elements in β_1 and $\hat{\beta}$, for any $p \times 1$ vector c,

$$Var(c^T\beta_1|X) - Var(c^T\hat{\beta}|X) = \sigma_u^2 c^T D^T Dc \geq 0 \tag{2.5}$$

Now we've shown $\hat{\beta}$ has the smallest variance among linear conditionally unbiased estimators, that is, $\hat{\beta}$ is BLUE.

References

[1] C. Radhakrishna Rao, Shalabh, Helge Toutenburg, Christian Heumann. ***Linear Models and Generalizations***[M]. Springer, 2010.

[2] James H. Stock, Mark W.Watson. ***Introduction to Econometrics***[M]. Pearson, 2011.