

Overview of Linear Regression

Part II contains the statistical theory of the OLS estimation. This theory rests on three basic assumptions about the sampling distribution from which one observes the data in the LHS variable \mathbf{y} and the RHS variables \mathbf{X} . As we accumulate the assumptions, we build an increasingly detailed model of the population and develop more sophisticated properties for OLS. Our primary goal is to provide order to the classical statistical theory by emphasizing the progressive character of these assumptions and their associated results.

This part of the book also extends the application of projection to random variables. The geometry of the OLS fit also appears in the conditional mean and in the relative efficiency of estimators. This geometry and the mathematics of the multivariate normal distribution comprise the probability distribution theory that undergirds the OLS statistical theory.

12.1 STATISTICAL THEORY

Part I explains the fit of a linear relationship by OLS. For the OLS fitted coefficients $\hat{\boldsymbol{\beta}}$ to be well defined, we assume that the matrix \mathbf{X} of explanatory variables is full-column rank. In *Part II*, we add the distributional assumptions listed in Table 12.1. The assumptions map to results in three categories: first moment, second moment, and distribution.

That the first moment of \mathbf{y} conditional on \mathbf{X} is $\mathbf{X}\boldsymbol{\beta}_0$ implies two first-moment results. In particular, the OLS fitted coefficients $\hat{\boldsymbol{\beta}}$ are unbiased estimators of the population coefficients in $\boldsymbol{\beta}_0$. The linearity of $\hat{\boldsymbol{\beta}}$ in \mathbf{y} is the key property of the OLS fit that supports these results: a linear combination of expected values equals the expected value of the linear combination.

That the second moment of \mathbf{y} conditional on \mathbf{X} is $\sigma_0^2 \cdot \mathbf{I}_N$ implies three second-moment results. First, because $\hat{\boldsymbol{\beta}}$ is linear in \mathbf{y} , the conditional variance matrix of $\hat{\boldsymbol{\beta}}$ follows easily from the conditional variance matrix of \mathbf{y} . Second, the variance parameter σ_0^2 possesses an unbiased estimator s^2 , which is the sample variance of the OLS fitted residuals adjusted for overfitting relative to $\mathbf{X}\boldsymbol{\beta}_0$. Third, the variance matrix of the OLS fitted coefficients yields the smallest variances for linear unbiased estimators of linear combinations of the elements of $\boldsymbol{\beta}_0$.

Finally, that the conditional distribution of \mathbf{y} is multivariate normal implies the conditional distribution of the OLS estimators. This stronger assumption also strengthens the relative efficiency of these estimators. The distributional properties lead to pivotal statistics that make interval estimators and hypothesis tests feasible.

12.2 PROBABILITY DISTRIBUTION THEORY

1. The conditional mean $E[\mathbf{y}_n|\mathbf{x}_n]$ and the MMSE linear predictor $E^*[\mathbf{y}_n|\mathbf{x}_n]$ are orthogonal projections, analogous to the OLS fitted vector $\hat{\boldsymbol{\mu}}$. By construction $E^*[\mathbf{y}_n|\mathbf{x}_n]$ is also linear in \mathbf{x}_n , like the elements of $\hat{\boldsymbol{\mu}}$.¹

If $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$, then $E[\hat{\boldsymbol{\mu}}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$. If \mathbf{X} is full-column rank also, then $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}_0$.

2. The conditional variance matrix $\text{Var}[\mathbf{y}|\mathbf{X}] = \sigma_0^2 \cdot \mathbf{I}_N$ has a geometric representation as an ellipse. The variance ellipse of a projection of \mathbf{y} , $\mathbf{P}\mathbf{y}$, equals the projection of the variance ellipse of \mathbf{y} .

The variance ellipse of a scalar variance matrix is a sphere. Thus, if $\text{Var}[\mathbf{y}|\mathbf{X}] = \sigma_0^2 \cdot \mathbf{I}_N$, then \mathbf{y} , $\hat{\boldsymbol{\mu}} \equiv \mathbf{P}_X\mathbf{y}$, and $\mathbf{y} - \hat{\boldsymbol{\mu}}$ are also spherically distributed, despite the apparent heteroskedasticity and covariance among their elements. Furthermore, the random variables in $\hat{\boldsymbol{\mu}}$ are orthogonal (uncorrelated) to those in $\mathbf{y} - \hat{\boldsymbol{\mu}}$.

On the other hand, $\hat{\boldsymbol{\beta}}$ is not spherically distributed. Its elliptical character depends on \mathbf{X} .

3. Covariance is the source of predictive power (in MSE) in one random variable for another:

$$\text{Cov}[z_1, z_2] \neq 0 \quad \Leftrightarrow \quad \min_{\alpha, \beta} E[(z_1 - \alpha - \beta z_2)^2] < E[(z_1 - E[z_1])^2]$$

Table 12.1.
Summary of Assumptions and Results for the Classical Regression Model

Assumptions	Results
<i>First moment:</i> $E[\mathbf{y} \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}_0$	<ul style="list-style-type: none"> • $E(\hat{\boldsymbol{\mu}} \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_0$, • $E[\hat{\boldsymbol{\beta}} \mathbf{X}] = \boldsymbol{\beta}_0$, where $\hat{\boldsymbol{\mu}} = \mathbf{P}_X\mathbf{y}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
<i>Second moment:</i> $\text{Var}[\mathbf{y} \mathbf{X}] = \sigma_0^2 \cdot \mathbf{I}_N$	<ul style="list-style-type: none"> • $\text{Var}[\hat{\boldsymbol{\beta}} \mathbf{X}] = \sigma_0^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ • $E(s^2 \mathbf{X}) = \sigma_0^2$, where $s^2 = (\mathbf{y} - \hat{\boldsymbol{\mu}})'(\mathbf{y} - \hat{\boldsymbol{\mu}})/(N - K)$ • $\hat{\boldsymbol{\beta}}$ is efficient relative to other linear unbiased estimators
<i>Distribution:</i> $\mathbf{y} \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_0, \sigma_0^2 \cdot \mathbf{I}_N)$	<ul style="list-style-type: none"> • $\hat{\boldsymbol{\beta}} \mathbf{X} \sim \mathcal{N}[\boldsymbol{\beta}_0, \sigma_0^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}]$ and independent of s^2 • $s^2 \sim \sigma_0^2 \chi_{N-K}^2/(N - K)$ • $\hat{\boldsymbol{\beta}}$ and s^2 are efficient relative to other unbiased estimators

¹ Thus, $E^*[\mathbf{y}_n|\mathbf{x}_n]$ is a restricted projection relative to $E[\mathbf{y}_n|\mathbf{x}_n]$ and its MMSE linear predictor as well.

This is one of several examples of the projection theorem at work. In addition,

- (a) the Gram–Schmidt orthonormalization of a set of random variables through a sequence of orthogonal projections provides the Choleski factorization $\mathbf{C}\mathbf{C}'$ of a variance matrix $\mathbf{\Omega}$ and
 - (b) the orthogonality condition $E[(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}'] = \mathbf{0}$ coincides with the efficiency of an unbiased estimator $\hat{\boldsymbol{\theta}}$ relative to another unbiased estimator $\tilde{\boldsymbol{\theta}}$ and the set of unbiased estimators $\hat{\boldsymbol{\theta}} + \mathbf{A}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ indexed by a matrix \mathbf{A} .
4. The multivariate normal distribution has the following properties:
 - (a) the distribution is characterized by its first two moments;
 - (b) linear combinations of multivariate normal random variables also possess a multivariate normal distribution;
 - (c) if they are uncorrelated, then multivariate normal random variables are also independently distributed;
 - (d) the conditional mean and MMSE linear predictor are identical; and
 - (e) variance ellipses coincide with isodensity contours.
 5. The chi-square and F distributions arise as the distributions of transformations of multivariate normal random variables. These are distributions for pivotal test statistics and corresponding interval estimators. In their noncentral generalizations, they determine the power functions of the test statistics.