Chapter 4

Least Squares: Finite Sample Theory

4.1 Maximum Likelihood

There are very few *principles* in statistics, and maximum likelihood is one of them. In this chapter, we first give an introduction of the maximum likelihood estimation. Consider a random sample of $Z = (z_1, z_2, \ldots, z_n)$ drawn from a parametric distribution with density $f_z(z_i; \theta)$, where z_i is either a scalar random variable or a random vector. A parametric distribution is completely characterized by a finite-dimensional parameter θ. We know that θ belongs to a parameter space Θ. We use the data to estimate θ.

The log-likelihood of observing the entire sample Z is

$$L_n(\theta; Z) := \log\left(\prod_{i=1}^{n} f_z(z_i; \theta)\right) = \sum_{i=1}^{n} \log f_z(z_i; \theta). \tag{4.1}$$

In reality the sample Z is given and for each $\theta \in \Theta$ we can evaluate $L(\theta; Z)$. The maximum likelihood estimator is

$$\widehat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} L_n(\theta; Z).$$

Why maximizing the log-likelihood function is desirable? An intuitive explanation is that $\widehat{\theta}_{MLE}$ makes observing Z the "most likely" in the entire parametric space.

A more formal justification requires an explicitly defined distance. Suppose that the true parameter value that generates the data is θ_0, so that the true distribution is $f_z(z_i; \theta_0)$. Any generic point $\theta \in \Theta$ produce $f_z(z_i; \theta)$. To measure their difference, we introduce *Kullback-Leibler divergence*, or Kullback-Leibler distance, defined as the logarithms of the expected likelihood ratio

$$\begin{aligned} D_f(\theta_0 \| \theta) = D(f_z(z_i; \theta_0) \| f_z(z_i; \theta)) &:= E_{\theta_0}\left[\log \frac{f_z(z_i; \theta_0)}{f_z(z_i; \theta)}\right] \\ &= E_{\theta_0}[\log f_z(z_i; \theta_0)] - E_{\theta_0}[\log f_z(z_i; \theta)]. \end{aligned}$$

We call it a "distance" because it is non-negative, although it is not symmetric $D_f(\theta_1, \theta_2) \neq D_f(\theta_2, \theta_1)$ and it does not satisfy the triangle inequality. To see $D_f(\theta, \theta_0)$ is non-negative, notice that $-\log(\cdot)$ is strictly convex and then by Jensen's inequalityCover and Thomas (2006, p.19)

$$\begin{aligned} E_{\theta_0}\left[\log \frac{f_z(z_i;\theta_0)}{f_z(z_i;\theta)}\right] &= E_{\theta_0}\left[-\log \frac{f_z(z_i;\theta)}{f_z(z_i;\theta_0)}\right] \geq -\log\left(E_{\theta_0}\left[\frac{f_z(z_i;\theta)}{f_z(z_i;\theta_0)}\right]\right) \\ &= -\log\left(\int \frac{f_z(z_i;\theta)}{f_z(z_i;\theta_0)} f_z(z_i;\theta_0)\, dz_i\right) = -\log\left(\int f_z(z_i;\theta)\, dz_i\right) \\ &= -\log 1 = 0, \end{aligned}$$

where $\int f_z(z_i;\theta)\, dz_i = 1$ for any pdf. The equality holds if and only if $f_z(z_i;\theta) = f_z(z_i;\theta_0)$ almost everywhere. Furthermore, if there is a one-to-one mapping between θ and $f_z(z_i;\theta)$ on Θ (identification), then $\theta_0 = \arg\min_{\theta\in\Theta} D_f(\theta, \theta_0)$ is the unique solution.

In information theory, $-E_{\theta_0}[\log f_z(z_i;\theta_0)]$ is the *entropy* of the continuous distribution of $f_z(z_i;\theta_0)$. Entropy measures the uncertainty of a random variable; the larger is the value, the more chaotic is the random variable. The Kullback-Leibler is the *relative entropy* between the distribution $f_z(z_i;\theta_0)$ and $f_z(z_i;\theta)$. It measures the inefficiency of assuming that the distribution is $f_z(z_i;\theta)$ when the true distribution is indeed $f_z(z_i;\theta_0)$. (Cover and Thomas, 2006, p.19)

Example 4.1. Consider the Gaussian location model $z_i \sim N(\mu, 1)$, where μ is the unknown parameter to be estimated. The likelihood of observing z_i is $f_z(z_i;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_i - \mu)^2\right)$. The likelihood of observing the sample Z is

$$f_Z(\mu; Z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$$

and the log-likelihood is

$$L_n = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(z_i - \mu)^2.$$

The (averaged) log-likelihood function for n observations is

$$\ell_n(\mu) = -\frac{1}{2}\log(2\pi) - \frac{1}{2n}\sum_{i=1}^{n}(z_i - \mu)^2.$$

We work with the averaged log-likelihood ℓ_n, instead of the (raw) log-likelihood L_n, to make it directly comparable with the expected log density

$$\begin{aligned}
E_{\mu_0}[f_z(z;\mu)] &= E_{\mu_0}[\ell_n(\mu)] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[(z_i - \mu)^2\right] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[((z_i - \mu_0) + (\mu_0 - \mu))^2\right] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[(z_i - \mu_0)^2\right] - E[z_i - \mu_0](\mu_0 - \mu) - \frac{1}{2}(\mu_0 - \mu)^2 \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2} - \frac{1}{2}(\mu - \mu_0)^2.
\end{aligned}$$

where the first equality holds because of random sampling. Obviously, $\ell_n(\mu)$ is maximized at $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ while $E_{\mu_0}[\ell_n(\mu)]$ is maximized at

$\mu = \mu_0$. The Kullback-Leibler divergence in this example is

$$D(\mu, \mu_0) = E_{\mu_0}[\ell_n(\mu_0)] - E_{\mu_0}[\ell_n(\mu)] = \frac{1}{2}(\mu - \mu_0)^2,$$

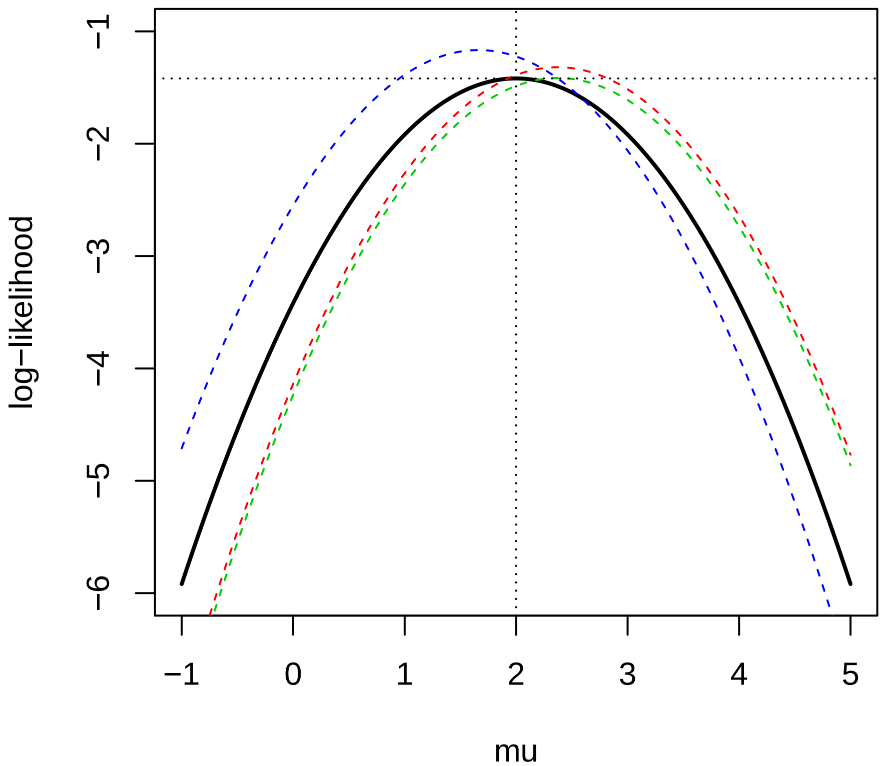
where $-E_{\mu_0}[\ell_n(\mu_0)] = \frac{1}{2}(\log(2\pi) + 1)$ is the entropy of the normal distribution with unit variance.

We use the following code to demonstrate the population log-likelihood $E[\ell_n(\mu)]$ when $\mu_0 = 2$ and the 3 sample realizations when $n = 4$.

```
set.seed(2020-10-7)
mu0 <- 2; gamma0 <- 1
# population likelihood function
L <- function(mu) {
  ell = -0.5 * log(2*pi*gamma0) - 0.5 / gamma0 *( 1 + (mu - mu0)^2 )
  return(ell) }
# sample likelihood function
Ln <- function(mu) {
  elln = -0.5 * log(2*pi*gamma0) - 0.5 / gamma0 * mean( (z - mu)^2 )
  return(elln) }

mu_base = mu0 + seq(-3, 3, by = 0.01)
# draw sample log-likelihood graph
n = 4
```

```r
lnz = matrix(0, length(mu_base), 3)
for (rr in 1:3){
  z <- rnorm(n, mu0, sqrt(gamma0) )
  lnz[,rr] <- plyr::laply(.data = mu_base, .fun = Ln)
}
# draw the graph
matplot(x = mu_base, y = cbind( L(mu_base), lnz),
        type = "l", lty = c(1, rep(2,3)),
        lwd =  c(2,rep(1,3)), col = 1:4, ylim = c(-6, -1),
        xlab = "mu", ylab = "log-likelihood")
abline(v = mu0, lty = 3)
abline(h = L(mu0), lty = 3)
```



4.2 Likelihood Estimation for Regression

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. Y is an $n \times 1$ vector, and X is an $n \times K$ matrix.

We continue with properties of OLS. Noticing that OLS coincides with the maximum likelihood estimator if the error term follows a normal distribution, we derive its finite-sample exact distribution which can be used for statistical inference. The Gauss-Markov theorem justifies the optimality of OLS under the classical assumptions.

In this chapter we employ the classical statistical framework under restrictive distributional assumption

$$y_i|x_i \sim N\left(x_i'\beta, \gamma\right), \tag{4.2}$$

where $\gamma = \sigma^2$ to ease the differentiation. This assumption is equivalent to $e_i|x_i = \left(y_i - x_i'\beta\right)|x_i \sim N(0,\gamma)$. Because the distribution of e_i is invariant to x_i, the error term $e_i \sim N(0,\gamma)$ and is statistically independent of x_i. This is a very strong assumption.

The likelihood of observing a pair (y_i, x_i) is

$$\begin{aligned} f_{yx}(y_i, x_i) &= f_{y|x}(y_i|x_i)\, f_x(x) \\ &= \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times f_x(x), \end{aligned}$$

where f_{yx} is the joint pdf, $f_{y|x}$ is the conditional pdf and f_x is the marginal pdf of x, and the second equality holds under the assumption (4.2). The likelihood a random sample $(y_i, x_i)_{i=1}^n$ is

$$\begin{aligned}\prod_{i=1}^{n} f_{yx}(y_i, x_i) &= \prod_{i=1}^{n} f_{y|x}(y_i|x_i)\, f_x(x) \\ &= \prod_{i=1}^{n} f_{y|x}(y_i|x_i) \times \prod_{i=1}^{n} f_x(x) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times \prod_{i=1}^{n} f_x(x).\end{aligned}$$

The parameters of interest (β, γ) are irrelevant to the second term $\prod_{i=1}^n f_x(x)$ for they appear only in the conditional likelihood

$$\prod_{i=1}^{n} f_{y|x}(y_i|x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right).$$

We focus on the conditional likelihood. To facilitate derivation, we work with the (averaged) conditional log-likelihood function

$$\ell_n(\beta, \gamma) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\gamma - \frac{1}{2n\gamma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2,$$

for $\log(\cdot)$ is a monotonic transformation that does not change the maximizer. The maximum likelihood estimator $\widehat{\beta}_{MLE}$ can be found using the FOC:

$$\frac{\partial}{\partial\beta}\ell_n(\beta, \gamma) = \frac{1}{n\gamma}\sum_{i=1}^{n} x_i\left(y_i - x_i'\beta\right) = 0.$$

Rearranging the above equation in matrix form $X'X\widehat{\beta}_{MLE} = X'Y$, we explicitly solve

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y$$

when $X'X$ is invertible. The maximum likelihood estimator (MLE) coincides with the OLS estimator. Similarly, the other FOC with respect to γ gives $\widehat{\gamma}_{\mathrm{MLE}} = \widehat{e}'\widehat{e}/n$.

4.3 Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$ assuming the error term follows a Gaussian distribution. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*, which is a large sample approximation to the finite sample distribution. We first review some properties of a generic jointly normal random vector.

Fact 4.1. *Let $z \sim N(\mu, \Omega)$ be an $l \times 1$ random vector with a positive definite variance-covariance matrix Ω. Let A be an $m \times l$ non-random matrix where $m \leq l$. Then $Az \sim N(A\mu, A\Omega A')$.*

Fact 4.2. *If $z \sim N(0,1)$, $w \sim \chi^2(d)$ and z and w are independent. Then $\frac{z}{\sqrt{w/d}} \sim t(d)$.*

The OLS estimator

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y = \left(X'X\right)^{-1} X' \left(X'\beta + e\right) = \beta + \left(X'X\right)^{-1} X'e,$$

and its conditional distribution can be written as

$$\begin{aligned}
\widehat{\beta}|X &= \beta + \left(X'X\right)^{-1} X'e|X \\
&\sim \beta + \left(X'X\right)^{-1} X' \cdot N\left(0_n, \sigma^2 \cdot I_n\right) \\
&\sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1} X'X \left(X'X\right)^{-1}\right) \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)
\end{aligned}$$

by Fact 4.1. The k-th element of the vector coefficient

$$\widehat{\beta}_k|X = \eta_k'\widehat{\beta}|X \sim N\left(\beta_k, \sigma^2\eta_k' \left(X'X\right)^{-1} \eta_k\right) \sim N\left(\beta_k, \sigma^2 \left(X'X\right)^{-1}_{kk}\right),$$

where $\eta_k = \left(1\left\{l = k\right\}\right)_{l=1,\ldots,K}$ is the selector of the k-th element.

In reality, σ^2 is an unknown parameter, and

$$s^2 = \widehat{e}'\widehat{e}/\left(n-K\right) = e'M_X e/\left(n-K\right)$$

is an unbiased estimator of σ^2. Consider the t-statistic

$$\begin{aligned}
T_k &= \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2\left[\left(X'X\right)^{-1}\right]_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}} \cdot \frac{\sqrt{\sigma^2}}{\sqrt{s^2}} \\
&= \frac{\left(\widehat{\beta}_k - \beta_k\right)/\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}}{\sqrt{\frac{e'}{\sigma}M_X\frac{e}{\sigma}/\left(n-K\right)}}.
\end{aligned}$$

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n-K\right)$. Moreover, the numerator and the denominator are statistically independent (See Section 4.7.2). As a result, we conclude $T_k \sim t(n-K)$ by Fact 4.2. This finite sample distribution allows us to conduct statistical inference.

4.4 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we represent the regression model as $Y = X\beta + e$ and

$$\begin{aligned} E[e|X] &= 0_n \\ \text{var}\left[e|X\right] &= E\left[ee'|X\right] = \sigma^2 I_n. \end{aligned}$$

where the first condition is the *mean independence* assumption, and the second condition is the *homoskedasticity* assumption. These assumptions are about the first and second *moments* of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the distribution of e_i.

- Unbiasedness:

$$\begin{aligned} E\left[\widehat{\beta}|X\right] &= E\left[\left(X'X\right)^{-1} XY|X\right] = E\left[\left(X'X\right)^{-1} X\left(X'\beta + e\right)|X\right] \\ &= \beta + \left(X'X\right)^{-1} XE\left[e|X\right] = \beta. \end{aligned}$$

  Unbiasedness does not rely on homoskedasticity.

- Variance:

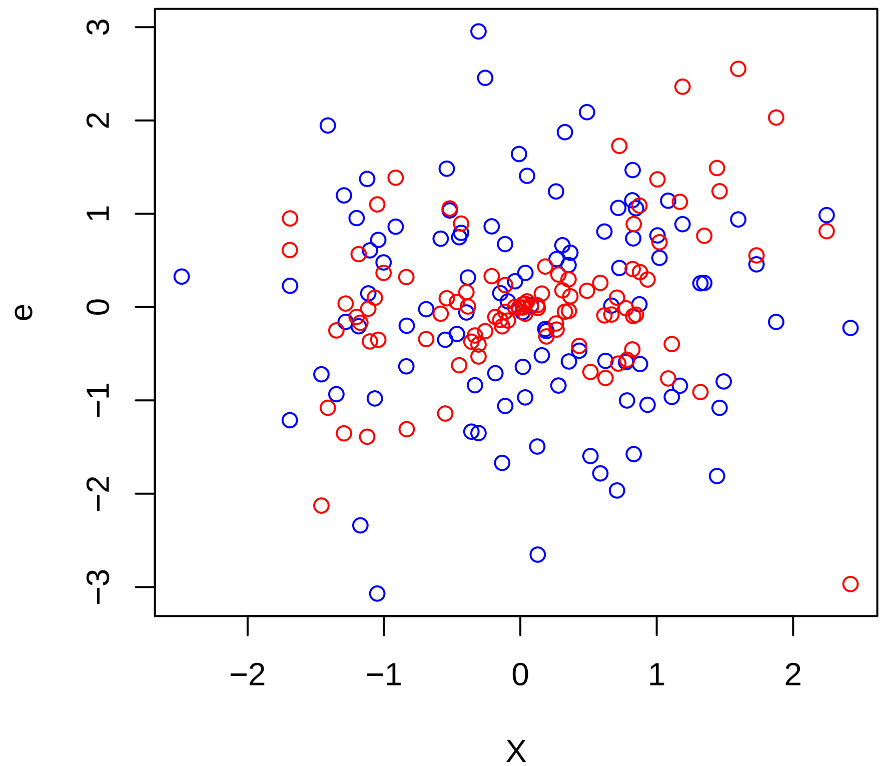
$$
\begin{aligned}
\operatorname{var}\left[\widehat{\beta}|X\right] &= E\left[\left(\widehat{\beta}-E\widehat{\beta}\right)\left(\widehat{\beta}-E\widehat{\beta}\right)'|X\right] \\
&= E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'|X\right] \\
&= E\left[\left(X'X\right)^{-1}X'ee'X\left(X'X\right)^{-1}|X\right] \\
&= \left(X'X\right)^{-1}X'E\left[ee'|X\right]X\left(X'X\right)^{-1}
\end{aligned}
$$

where the second equality holds as $E\left[\widehat{\beta}\right] = E\left[E\left[\widehat{\beta}|X\right]\right] = \beta$. Under the assumption of homoskedasticity, it can be simplified as

$$
\operatorname{var}\left[\widehat{\beta}|X\right] = \left(X'X\right)^{-1}X'\left(\sigma^2 I_n\right)X\left(X'X\right)^{-1} = \sigma^2\left(X'X\right)^{-1}.
$$

Example 4.2. (Heteroskedasticity) If $e_i = x_i u_i$, where x_i is a scalar random variable, u_i is statistically independent of x_i, $E\left[u_i\right] = 0$ and $E\left[u_i^2\right] = \sigma^2$. Then $E\left[e_i|x_i\right] = 0$ but $E\left[e_i^2|x_i\right] = \sigma^2 x_i^2$ is a function of x_i. We say e_i^2 is a heteroskedastic error.

```
n = 100; X = rnorm(n)
e1 = rnorm(n)
plot( y = e1, x = X, col = "blue", ylab = "e")
e2 = X * rnorm(n)
points( y = e2, x = X, col = "red")
```



It is important to notice that independently and identically distributed sample (iid) (y_i, x_i) does not imply homoskedasticity. Homoskedasticity or heterskedasticity is about the relationship between $(x_i, e_i = y_i - \beta x)$, whereas iid is about the relationship between (y_i, x_i) and (y_j, x_j) for $i \neq j$.

4.5 Gauss-Markov Theorem

Gauss-Markov theorem is concerned about the optimality of OLS. It justifies OLS as the efficient estimator among all linear unbiased ones. *Efficient* here means that it enjoys the smallest variance in a family of estimators.

We have shown that OLS is unbiased in that $E\left[\widehat{\beta}\right] = \beta$. There are numerous linearly unbiased estimators. For example, $\left(Z'X\right)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[\left(Z'X\right)^{-1} Z'y\right] = E\left[\left(Z'X\right)^{-1} Z'\left(X\beta + e\right)\right] = \beta$. We cannot say OLS is better than those other unbiased estimators because they are equally good in this aspect. Thus, we move to the second order property of variance: an estimator is better if its variance is smaller.

Fact 4.3. *For two generic random vectors X and Y of the same size, we say X's variance is smaller or equal to Y's variance if $(\Omega_Y - \Omega_X)$ is a positive semi-definite matrix. The comparison is defined this way because for any non-zero constant vector c, the variance of the linear combination of X*

$$\operatorname{var}\left(c'X\right) = c'\Omega_X c \leq c'\Omega_Y c = \operatorname{var}\left(c'Y\right)$$

is no bigger than the same linear combination of Y.

Let $\tilde{\beta} = A'y$ be a generic linear estimator, where A is any $n \times K$ functions of X. As

$$E\left[A'y|X\right] = E\left[A'\left(X\beta + e\right)|X\right] = A'X\beta.$$

So the linearity and unbiasedness of $\tilde{\beta}$ implies $A'X = I_n$. Moreover, the variance

$$\operatorname{var}\left(A'y|X\right) = E\left[\left(A'y-\beta\right)\left(A'y-\beta\right)'|X\right] = E\left[A'ee'A|X\right] = \sigma^2 A'A.$$

Let $C = A - X\left(X'X\right)^{-1}$.

$$\begin{aligned} A'A - \left(X'X\right)^{-1} &= \left(C + X\left(X'X\right)^{-1}\right)'\left(C + X\left(X'X\right)^{-1}\right) - \left(X'X\right)^{-1} \\ &= C'C + \left(X'X\right)^{-1}X'C + C'X\left(X'X\right)^{-1} \\ &= C'C, \end{aligned}$$

where the last equality follows as

$$\left(X'X\right)^{-1}X'C = \left(X'X\right)^{-1}X'\left(A - X\left(X'X\right)^{-1}\right) = \left(X'X\right)^{-1} - \left(X'X\right)^{-1} = 0.$$

Therefore $A'A - \left(X'X\right)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$. The above derivation shows OLS achieves the smallest variance among all linear unbiased estimators.

Homoskedasticity is a restrictive assumption. Under homoskedasticity, $\operatorname{var}\left[\widehat{\beta}\right] = \sigma^2\left(X'X\right)^{-1}$. Popular estimator of σ^2 is the sample mean of the residuals $\widehat{\sigma}^2 = \frac{1}{n}\widehat{e}'\widehat{e}$ or the unbiased one $s^2 = \frac{1}{n-K}\widehat{e}'\widehat{e}$. Under heteroskedasticity, Gauss-Markov theorem does not apply.

4.6 Summary

The linear algebraic properties holds in finite sample no matter the data are taken as fixed numbers or random variables. The exact distribution under the normality assumption of the error term is the classical statistical results. The Gauss Markov theorem holds under two crucial assumptions: linear CEF and homoskedasticity.

Historical notes: MLE was promulgated and popularized by Ronald Fisher (1890–1962). He was a major contributor of the frequentist approach which dominates mathematical statistics today, and he sharply criticized the Bayesian approach. Fisher collected the iris flower dataset of 150 observations in his biological study in 1936, which can be displayed in R by typing `iris`. Fisher invented the many concepts in classical mathematical statistics, such as sufficient statistic, ancillary statistic, completeness, and exponential family, etc.

Further reading: Phillips (1983) offers a comprehensive treatment of exact small sample theory in econometrics. After that, theoretical studies in econometrics swiftly shifted to large sample theory, which we will introduce in the next chapter.

4.7 Appendix

4.7.1 Joint Normal Distribution

It is arguable that normal distribution is the most frequently encountered distribution in statistical inference, as it is the asymptotic distribution of many popular estimators. Moreover, it boasts some unique features that facilitates the calculation of objects of interest. This note summaries a few of them.

An $n \times 1$ random vector Y follows a joint normal distribution $N(\mu, \Sigma)$, where μ is an $n \times 1$ vector and Σ is an $n \times n$ symmetric positive definite matrix. The probability density function is

$$f_y(y) = (2\pi)^{-n/2} (\det(\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)' \Sigma^{-1} (y-\mu)\right)$$

and the moment generating function is

$$M_y(t) = \exp\left(t'\mu + \frac{1}{2}t'\Sigma t\right).$$

We will discuss the relationship between two components of a random vector. To fix notation,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

where Y_1 is an $m \times 1$ vector, and Y_2 is an $(n-m) \times 1$ vector. μ_1 and μ_2 are the corresponding mean vectors, and Σ_{ij}, $j = 1,2$ are the corresponding variance and covariance matrices. From now on, we always maintain the assumption that $Y = (Y_1', Y_2')'$ is jointly normal.

Fact 4.1 immediately implies a convenient feature of the normal distribution. Generally speaking, if we are given a joint pdf of two random variables and intend to find the marginal distribution of one random variables, we need to integrate out the other variable from the joint pdf. However, if the variables are jointly normal, the information of the other random variable is irrelevant to the marginal distribution of the random variable of interest. We only need to know the partial information of the part of interest, say the mean μ_1 and the variance Σ_{11} to decide the marginal distribution of Y_1.

Fact 4.4. *The marginal distribution* $Y_1 \sim N(\mu_1, \Sigma_{11})$.

This result is very convenient if we are interested in some component if an estimator, but not the entire vector of the estimator. For example, the OLS estimator of the linear regression model $y_i = x_i'\beta + e_i$, under the classical assumption of (i) random sample; (ii) independence of z_i and e_i; (iii) $e_i \sim N\left(0, \sigma^2\right)$ is

$$\widehat{\beta} = \left(X'X\right)^{-1} X'y,$$

and the finite sample exact distribution of $\widehat{\beta}$ is

$$\left(\widehat{\beta}-\beta\right)|X \sim N\left(0,\sigma^2\left(X'X\right)^{-1}\right)$$

If we are interested in the inference of only the j-th component of $\beta_0^{(j)}$, then from Fact 4.4,

$$\left(\widehat{\beta}_k-\beta_k\right)/\left(X'X\right)_{kk}^{-1} \sim N\left(0,\sigma^2\right)$$

where $\left(X'X\right)_{kk}^{-1}$ is the k-th diagonal element of $\left(X'X\right)^{-1}$. The marginal distribution is independent of the other components. This saves us from integrating out the other components, which could be troublesome if the dimension of the vector is high.

Generally, zero covariance of two random variables only indicates that they are uncorrelated, whereas full statistical independence is a much stronger requirement. However, if Y_1 and Y_2 are jointly normal, then zero covariance is equivalent to full independence.

Fact 4.5. *If $\Sigma_{12}=0$, then Y_1 and Y_2 are independent.*

Fact 4.6. *If Σ is invertible, then $Y'\Sigma^{-1}Y \sim \chi^2\left(\text{rank}\left(\Sigma\right)\right)$.*

The last result, which is useful in linear regression, is that if Y_1 and Y_2 are jointly normal, the conditional distribution of Y_1 on Y_2 is still jointly normal, with the mean and variance specified as in the following fact.

Fact 4.7. $Y_1|Y_2 \sim N\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}\left(Y_2-\mu_2\right),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$.

4.7.2 Basu's Theorem

$Y = (y_1, \ldots, y_n)$ consists of n iid observations. We say $T(Y)$ is a sufficient statistic for a parameter θ if the conditional probability $f(Y|T(Y))$ does not depend on θ. For example, for $y_i \sim N(\mu, \sigma^2)$ with known σ^2 and unknown μ, We verify that the sample mean $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ is a sufficient statistic for μ. Notice that the joint density of Y is

$$
\begin{aligned}
f(Y) &= (2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \bar{y})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right).
\end{aligned}
$$

Because $\bar{y} \sim N(\mu, \sigma^2/n)$, the marginal density is

$$
f(\bar{y}) = \left(2\pi\sigma^2/n\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{y} - \mu)^2\right).
$$

The conditional density is

$$
f(Y|\bar{y}) = \frac{f(Y)}{f(\bar{y})} = \frac{\left(2\pi\sigma^2\right)^{-n/2}}{\left(2\pi\sigma^2/n\right)^{-1/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \bar{y})^2\right)
$$

is independent of μ, and thus $\bar{y}$ is a sufficient statistic for μ.

In the meantime, the sample standard deviation $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})$ is an *ancillary statistic* for μ , because the distribution of s^2 does not depend on μ.

Basu's theorem says that a *complete* sufficient statistic is statistically independent from any ancillary statistic. For a normal distribution with unknown mean and known variance, the sample mean $\bar{y}$ is the sufficient statistic and the sample standard deviation s^2 is an ancillary statistic.

A parametric distribution indexed by θ is a member of the *exponential family* is its PDF can be written as

$$f(Y|\theta) = h(Y)\, g(\theta) \exp\left(\eta(\theta)'\, T(Y)\right),$$

where $g(\theta)$ and $\eta(\theta)$ are functions depend, only on θ and $h(Y)$ and $T(Y)$ are functions depend only on Y. The normal distribution with known σ^2 and unknown μ belongs to the exponential family in view of the decomposition

$$\begin{aligned} f(Y) &= \left(\sqrt{2\pi}\sigma\right)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ &= \underbrace{\exp\left(-\sum_{i=1}^{n}\frac{y_i^2}{2\sigma^2}\right)}_{h(Y)} \cdot \underbrace{\left(\sqrt{2\pi}\sigma\right)^{-n}\exp\left(-\frac{n}{2\sigma^2}\mu^2\right)}_{g(\theta)} \cdot \underbrace{\exp\left(\frac{\mu n}{2\sigma^2}\bar{y}\right)}_{\exp\left(\eta(\theta)'T(Y)\right)}. \end{aligned}$$

The exponential family is a class of distributions with the special functional form which is convenient for deriving sufficient statistics as well as other desirable properties in classical mathematical statistics.

`Zhentao Shi. Oct 6.`

Bibliography

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory (2nd ed.)*. John Wiley & Sons.

Phillips, P. C. (1983). Exact small sample theory in the simultaneous equations model. *Handbook of Econometrics 1*, 449–516.