

Notation: y_i is a scalar, and x_i is a $K \times 1$ vector. Y is an $n \times 1$ vector, and X is an $n \times K$ matrix.

1 Algebra of Least Squares

1.1 OLS estimator

As we have learned from the linear project model, the parameter β

$$\begin{aligned} y_i &= x_i' \beta + e_i \\ E[x_i e_i] &= 0 \end{aligned}$$

can be written as $\beta = (E[x_i x_i'])^{-1} E[x_i y_i]$.

While population is something imaginary, in reality we possess a sample of n observations. We thus replace the population mean $E[\cdot]$ by the sample mean, and the resulting estimator is

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i y_i = (X'X)^{-1} X'y.$$

This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals $\sum_{i=1}^n (y_i - x_i' \beta)^2$. By a routine optimization, we obtain exactly the same $\hat{\beta}$.

Here we introduce some definitions and properties from OLS estimation.

- Fitted value: $\hat{Y} = X\hat{\beta}$.
- Residual: $\hat{e} = Y - \hat{Y}$.
- Projector: $P_X = X(X'X)^{-1}X'$; Annihilator: $M_X = I_n - P_X$.
- $P_X M_X = M_X P_X = 0$.
- Idempotent matrix: $P_X P_X = P_X$, $M_X M_X = M_X$.
- $\hat{e} = Y - \hat{Y} = Y - X\hat{\beta} = M_X Y = M_X (X\beta + e) = M_X e$.

- $X'\hat{e} = XM_Xe = 0$.
- $\frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0$ if x_i contains a constant.

1.2 Goodness of Fit

The so-called R-square is the most popular measure of goodness-of-fit in the linear regression. R-square is well defined only when a constant is included in the regressors. Let $M_\iota = I_n - \frac{1}{n}\iota\iota'$, where ι is an $n \times 1$ vector of 1's. M_ι is the *demeaner*, in the sense that $M_\iota(z_1, \dots, z_n)' = (z_1 - \bar{z}, \dots, z_n - \bar{z})'$, where $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$. For any X , we can decompose $Y = P_X Y + M_X Y = \hat{Y} + \hat{e}$. The total variation is

$$Y'M_\iota Y = (\hat{Y} + \hat{e})' M_\iota (\hat{Y} + \hat{e}) = \hat{Y}' M_\iota \hat{Y} + 2\hat{Y}' M_\iota \hat{e} + \hat{e}' M_\iota \hat{e} = \hat{Y}' M_\iota \hat{Y} + \hat{e}' \hat{e}$$

where the last equality follows by $M_\iota \hat{e} = \hat{e}$ as $\frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0$, and $\hat{Y}' \hat{e} = Y' P_X M_X e = 0$. R-square is defined as $\hat{Y}' M_\iota \hat{Y} / Y' M_\iota Y$.

1.3 Frish-Waugh-Lovell Theorem

If $Y = X_1\beta_1 + X_2\beta_2 + e$, then $\hat{\beta}_1 = (X_1' M_{X_2} X_1)^{-1} X_1' M_{X_2} Y$.

2 Statistical Properties of Least Squares

To talk about the statistical properties, we impose the following assumptions.

1. The data $(y_i, x_i)_{i=1}^n$ is a random sample from the same data generating process $y_i = x_i' \beta + e_i$.
2. e_i and x_i are independent.
3. $e_i \sim N(0, \sigma^2)$.

2.1 Normal Regression

Under the normality assumption, $y_i|x_i \sim N(x_i'\beta, \gamma)$, where $\gamma = \sigma^2$. The *conditional* likelihood of observing a sample $(y_i, x_i)_{i=1}^n$ is

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

and the (conditional) log-likelihood function is

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

Therefore, the maximum likelihood estimator (MLE) coincides with the OLS estimator, and $\hat{\gamma}_{\text{MLE}} = \hat{e}'\hat{e}/n$.

We can show the finite-sample exact distribution of $\hat{\beta}$. Since

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

we have the estimator $\hat{\beta}|X \sim N(\beta, \sigma^2 (X'X)^{-1})$, and

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

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

Consider the T -statistic

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 (X'X)^{-1}_{kk}}} = \frac{(\hat{\beta}_k - \beta_k) / \sqrt{\sigma^2 (X'X)^{-1}_{kk}}}{\sqrt{s^2 / \sigma^2}}.$$

The numerator follows a standard normal, and the denominator follows $\chi^2(n - K)$. Moreover, the numerator and the denominator are independent. As a result, $T_k \sim t(n - K)$.

2.2 Gauss-Markov Theorem

Now we relax the normality assumption and statistical independence. Instead, we assume a random sample and

$$y_i = x_i' \beta + e_i$$

$$E[e_i | x_i] = 0 \quad (1)$$

$$E[e_i^2 | x_i] = \sigma^2. \quad (2)$$

(1) is called the mean independence assumption, and (2) is the homoskedasticity assumption.

- Unbiasedness: $E[\hat{\beta} | X] = E[(X'X)^{-1} X'Y | X] = E[(X'X)^{-1} X'(X'\beta + e) | X] = \beta$. Unbiasedness does not rely on the homoskedasticity assumption.
- Variance:

$$\begin{aligned} \text{var}(\hat{\beta} | X) &= E\left[(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' | X\right] \\ &= E\left[(X'X)^{-1} X'ee'X (X'X)^{-1} | X\right] \\ &= (X'X)^{-1} X' E[ee' | X] X (X'X)^{-1} \\ &= (X'X)^{-1} X' (\sigma^2 I_n) X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

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

There are numerous linearly unbiased estimators. For example, $(Z'X)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E[(Z'X)^{-1} Z'y] = E[(Z'X)^{-1} Z'(X\beta + e)] = \beta$.

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

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

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

variance

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

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

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

where the last equality follows as

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

Therefore $A'A - (X'X)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\hat{\beta}$.

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