1 The Algebra of Least Squares

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

1.1 OLS estimator

As we have learned from the previous lecture, the parameter β in the linear projection model

$$\begin{aligned} y &= x'\beta + e \\ E[xe] &= 0 \end{aligned}$$

can be written as $\beta = (E[xx'])^{-1} E[xy]$. In reality we possess a sample of n observations, not the population. We thus replace the population mean $E[\cdot]$ by the sample mean, and the resulting estimator is

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

This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals

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

which is the sample counterpart of the MSE. By a similar optimization procedure, we derive exactly the same $\widehat{\beta}$.

Some definitions.

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

Some properties of the residual

- $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = M_X Y = M_X (X\beta + e) = M_X e$.
- $X'\widehat{e} = X M_X e = 0$.
- $\frac{1}{n} \sum_{i=1}^{n} \widehat{e}_i = 0$ 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 defined 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, that is, $M_\iota (z_1, \ldots, z_n)' = (z_1 - \bar{z}, \ldots, z_n - \bar{z})'$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$. For any X, we can decompose $Y = P_X Y + M_X Y = \widehat{Y} + \widehat{e}$. The squared sum is

$$\begin{aligned} Y' M_\iota Y &= \left(\widehat{Y} + \widehat{e}\right)' M_\iota \left(\widehat{Y} + \widehat{e}\right) \\ &= \widehat{Y}' M_\iota \widehat{Y} + 2\widehat{Y}' M_\iota \widehat{e} + \widehat{e}' M_\iota \widehat{e} \\ &= \widehat{Y}' M_\iota \widehat{Y} + \widehat{e}'\widehat{e}. \end{aligned}$$

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

1.3 Frish-Waugh-Lovell Theorem

If $y_i = x_1'\beta_1 + x_2'\beta_2 + e$, then $\widehat{\beta}_1 = \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 Y$.

2 Statistical Properties of Least Squares

2.1 Bias and Variance

When we discuss the statistical properties of the least squares in finite sample, we assume

$$\begin{aligned} y &= x'\beta + e \\ E[e|x] &= 0, \end{aligned}$$

which is equivalent to assume $E\left[y|x\right] = x'\beta$ so the linear projection model and the conditional mean model coincide. In other words, the following statistical properties hold only when the conditional mean is a linear function.

- Unbiasedness: $E\left[\widehat{\beta}|X\right] = E\left[\left(X'X\right)^{-1} XY|X\right] = E\left[\left(X'X\right)^{-1} X\left(X'\beta + e\right)|X\right] = \beta$.
- Variance $\mathrm{var}\left(\widehat{\beta}|X\right) = E\left[\left(\widehat{\beta} - E\widehat{\beta}\right)\left(\widehat{\beta} - E\widehat{\beta}\right)'|X\right] = E\left[\left(X'X\right)^{-1} X' e e' X \left(X'X\right)^{-1}|X\right] = \left(X'X\right)^{-1} X' E\left[ee'|X\right] X \left(X'X\right)^{-1}$. Under the *homoskedasticity* assumption,[1] we simplify it as

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

2.2 Gauss-Markov Theorem

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.

[1] The error term is homoskedastic if $E\left[e_i^2|X\right] = \sigma^2$ for all i. Homoskedasticity is a restrictive assumption. An example of *heteroskedasticity*: consumption = income + e. A rich family has more variation in consumption than a poor family.

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

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

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

So the unbiasedness of $\tilde{\beta}$ implies $AX = I_n$. Moreover, the variance

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

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

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

where the last equality follows as

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

The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$.

However, notice that Gauss-Markov theorem is derived only in the conditional mean model under homoskedasticity. These conditions are restrictive.

2.3 Variance Estimation

Under homoskedasticity, $\mathrm{var}\left(\widehat{\beta}\right) = \sigma^2 (X'X)^{-1}$. 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,

$$\mathrm{var}\left(\widehat{\beta}\right) = \left(X'X\right)^{-1} X'DX\left(X'X\right)^{-1}$$

where $D = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$. We propose to estimate $\widehat{\sigma}_i^2 = \widehat{e}_i^2$ so that $X'DX = \sum_{i=1}^n x_i x_i' \widehat{e}_i^2$.

2.4 Normal Regression

If we assume $e_i \sim N\left(0, \sigma^2\right)$ is independent of x_i, then $y_i \sim N\left(x_i'\beta, \sigma^2\right)$. The likelihood of observing a sample $(y_i, x_i)_{i=1}^n$ is

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

and the log-likelihood function is

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

Therefore, the maximum likelihood estimator coincides with the OLS estimator.

Moreover, consider the statistic

$$T_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2 \left(X'X\right)^{-1}_{kk}}} = \frac{\eta_k' \left(X'X\right)^{-1} X'e/\sigma}{\sqrt{\frac{s^2}{\sigma^2}\eta_k' \left(X'X\right)^{-1}\eta_k}} = \frac{a_k' \frac{e}{\sigma} / \sqrt{a_k' a_k}}{\sqrt{s^2/\sigma^2}}$$

where $\eta_k = \left(1\left\{l = k\right\}\right)_{l=1,\ldots,K}$ is the selector, and $a_k = \eta_k'\left(X'X\right)^{-1}X'$. It is easy to see that the numerator follows a standard normal, and the denominator follows a χ^2 with the degree of freedom $n-K$. Therefore $T_k \sim t\left(n-K\right)$.