Least Squares

Zhentao Shi

September 16, 2019

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 projection coefficient β in the regression

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

can be written as $\beta = \left(E\left[x_i x_i'\right]\right)^{-1} E\left[x_i y_i\right]$. While population is something imaginary, in reality we possess a sample of n observations $\left(y_i, x_i\right)_{i=1}^n$. We thus replace the population mean $E\left[\cdot\right]$ 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

$$Q\left(\beta\right) = \sum_{i=1}^n \left(y_i - x_i'\beta\right)^2 = \left(Y - X\beta\right)'\left(Y - X\beta\right).$$

Solve the first-order condition

$$\frac{\partial}{\partial\beta} Q\left(\beta\right) = -2X'\left(Y - X\beta\right) = 0.$$

This necessary condition for optimality gives exactly the same $\widehat{\beta}$. Moreover, the second-order condition

$$\frac{\partial^2}{\partial\beta\partial\beta'} Q\left(\beta\right) = 2X'X$$

shows that $Q\left(\beta\right)$ is convex in β due to the positive semi-definite matrix $X'X$. ($Q\left(\beta\right)$ is strictly convex in β if $X'X$ is positive definite.)

Here are some definitions and properties of the OLS estimator.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projector: $P_X = X\left(X'X\right)^{-1} X$; Annihilator: $M_X = I_n - P_X$.

- $P_X M_X = M_X P_X = 0$.
- If $AA = A$, we call it an idempotent matrix. Both P_X and M_X are idempotent.
- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - X(X'X)^{-1}X'Y = (I - P_X)Y = M_X Y = M_X (X\beta + e) = M_X e$. (Note: $M_X X = (I - P_X)X = X - X = 0 \Longrightarrow M_X X\beta = 0$)
- $X'\widehat{e} = X'M_X e = 0$. (Note again $X'M_X = 0$)
- $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$ if x_i contains a constant.

  (Justification: $X'\widehat{e} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ * & * & \cdots & * \\ \cdots & \cdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix} \begin{bmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the the first row implies $\sum_{i=1}^n \widehat{e}_i = 0$.)

1.2 Goodness of Fit

The so-called *R-squared* is a popular measure of goodness-of-fit in the linear regression. R-squared 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 $M_\iota (z_1, \ldots, z_n)' = (z_1 - \bar{z}, \ldots, z_n - \bar{z})'$, 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 total variation is

$$Y'M_\iota Y = \left(\widehat{Y} + \widehat{e}\right)' M_\iota \left(\widehat{Y} + \widehat{e}\right) = \widehat{Y}'M_\iota\widehat{Y} + 2\widehat{Y}'M_\iota\widehat{e} + \widehat{e}'M_\iota\widehat{e} = \widehat{Y}'M_\iota\widehat{Y} + \widehat{e}'\widehat{e}$$

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

1.3 Frish-Waugh-Lovell Theorem

The Frish-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem We need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}'\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & \left(A_{22} - A_{12}'A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

$$
\begin{aligned}
\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} &= (X'X)^{-1} X'Y \\
&= \left(\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \begin{pmatrix} X_1 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\
&= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\
&= \begin{pmatrix} \left(X_1' M_{X_2}' X_1\right)^{-1} & -\left(X_1' M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2' X_2\right)^{-1} \\ * & * \end{pmatrix} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}.
\end{aligned}
$$

The subvector

$$
\begin{aligned}
\widehat{\beta}_1 &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1'Y - \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2' X_2\right)^{-1} X_2'Y \\
&= \left(X_1' M_{X_2}' X_1\right)^{-1} \left(X_1'Y - X_1' P_{X_2} Y\right) \\
&= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' M_{X_2} Y.
\end{aligned}
$$

Notice that $\widehat{\beta}_1$ can be obtained by the following:

1. Regress y on X_2, obtain residuals $\tilde{e}_2$;
2. Regress X_1 on X_2, obtain residuals $\tilde{X}_2$;
3. Regress $\tilde{e}_2$ on $\tilde{X}_2$, obtain OLS estimates $\widehat{\beta}_1$.

Similar derivation can also be carried out in the population linear projection. See Hansen (2019) Chapter 2.22-23.

2 Statistical Properties of Least Squares

In this section we return to the classical statistical framework under restrictive distributional assumption $y_i | x_i \sim N\left(x_i'\beta, \gamma\right)$, where $\gamma = \sigma^2$.

2.1 Maximum Likelihood Estimation

The *conditional* likelihood of observing a *random sample* $(y_i, x_i)_{i=1}^n$ is

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

and the (conditional) log-likelihood function is

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

The maximum likelihood estimator $\widehat{\beta}_{MLE}$ can be found using the FOC:

$$\frac{\partial}{\partial \beta} L(\beta, \gamma) = \frac{1}{2\gamma} \sum_{i=1}^{n} 2x_i \left(y_i - x_i'\beta\right)^2 = 0.$$

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

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

The maximum likelihood estimator (MLE) coincides with the OLS estimator. Similarly, another FOC gives $\widehat{\gamma}_{\text{MLE}} = \widehat{e}'\widehat{e}/n$.

2.2 Classical 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. Let the "error term" $e_i = y_i - x_i'\beta$, we have $e_i|x_i = y_i|x_i - x_i'\beta \sim N(0, \gamma)$. Since the conditional distribution of e_i on x_i is invariant with x_i, the discrepancy e_i is statistical independent of x_i. Assume The 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

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

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

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

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

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

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

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2(n-K)$. Moreover, the numerator and the denominator are statistically independent (Basu's theorem). As a result, we conclude $T_k \sim t(n-K)$. This finite sample distribution is crucial when conducting statistical inference.

2.3 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we represent the regression model as $y_i = x_i'\beta + e_i$ and

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

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:

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

  Unbiasedness does not rely on homoskedasticity.

- Variance:

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

  Homoskedasticity is essential in this derivation.

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

2.4 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.

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

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 $\widehat{\beta}$.

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.