This version: September 28, 2016

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 $\beta = \left(E\left[x_i x_i'\right]\right)^{-1} E\left[x_i y_i\right]$.

While population is something imaginary, in reality we possess a sample of n observations. 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

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

By the first-order condition

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

the optimality condition gives exactly the same $\widehat{\beta}$. Moreover, the secondorder condition

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

shows that $Q(\beta)$ is convex in β. ($Q(\beta)$ is strictly convex in β if $X'X$ is positive definite.)

Here we introduce some definitions and properties in OLS estimation.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projector: $P_X = X(X'X)^{-1}X$; 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: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = M_X Y = M_X(X\beta + e) = M_X e$.
- $X'\widehat{e} = XM_X e = 0$.
- $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$ 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 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 $M_\iota(z_1, \ldots, z_n)' = (z_1 - \overline{z}, \ldots, z_n - \overline{z})'$, 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-square is defined as $\widehat{Y}'M_\iota\widehat{Y}/Y'M_\iota Y$.

1.3 Frish-Waugh-Lovell Theorem

This theorem is the sample version of the subvector regression.

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

2 Statistical Properties of Least Squares

To talk about the statistical properties in finite sample, 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. $e_i | x_i \sim N\left(0, \sigma^2\right)$.

2.1 Normal Regression

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

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

and the (conditional) log-likelihood function is

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

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

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

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

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

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

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

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

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

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

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

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 independent. As a result, $T_k \sim t(n-K)$.

2.2 Mean and Variance

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

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E[e_i|x_i] &= 0 \qquad (1) \\ E\left[e_i^2|x_i\right] &= \sigma^2. \qquad (2) \end{aligned}$$

(1) is the *mean independence* assumption, and (2) is the *homoskedasticity* assumption. These assumptions are about the first and second moment of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the entire 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:

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

2.3 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, $(Z'X)^{-1}Z'y$ 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

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

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