Advanced Econometrics

Chapter 3: Least Squares Methods

In Choi

Sogang University

Least squares methods for estimating coefficients

Reading: Chapter 3 of Greene

Methods for estimating β

Least squares estimation

Maximum likelihood estimation

Method of moments estimation

Least absolute deviation estimation

$\vdots$

Least squares estimation

- The objective function for the least squares estimation is

$$S(\beta_1, \ldots, \beta_K) = \sum_{i=1}^{n} \left(y_i - \beta_1 x_{i1} - \cdots - \beta_K x_{iK}\right)^2 .$$

  We need to minimize this function.

- The first–order conditions for the minimization is

$$\begin{aligned}
\frac{\partial S(\beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_{i1} \left(y_i - \beta_1 x_{i1} - \cdots - \beta_K x_{iK}\right) = 0 \\
&\vdots \\
\frac{\partial S(\beta_K)}{\partial \beta_K} &= -2 \sum_{i=1}^{n} x_{iK} \left(y_i - \beta_1 x_{i1} - \cdots - \beta_K x_{iK}\right) = 0.
\end{aligned}$$

Least squares estimation

- These equations can be written as

$$
\begin{aligned}
\sum_{i=1}^{n} x_{i1} y_i &= \beta_1 \sum_{i=1}^{n} x_{i1} x_{i1} + \cdots + \beta_K \sum_{i=1}^{n} x_{i1} x_{iK} \\
&\vdots \\
\sum_{i=1}^{n} x_{iK} y_i &= \beta_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \cdots + \beta_K \sum_{i=1}^{n} x_{ik} x_{iK}
\end{aligned}
$$

or

$$
\begin{bmatrix} \mathbf{x}_1' y \\ \vdots \\ \mathbf{x}_K' y \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \mathbf{x}_1 & \cdots & \mathbf{x}_1' \mathbf{x}_K \\ \vdots & \vdots & \vdots \\ \mathbf{x}_K' \mathbf{x}_1 & \cdots & \mathbf{x}_K' \mathbf{x}_K \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}
$$

or

$$
X' y = (X' X) \beta
$$

- The solution of these equations (b in vector notation) is

$$b = \left(X'X\right)^{-1} X'y.$$

- This is the least squares estimator of β, or the ordinary least squres estimator (OLS). If $rank\left(X\right) = K$, $rank\left(X'X\right) = K$. Thus, the inverse of $X'X$ exists.

Least squares estimation

- Residual vector

$$\begin{aligned} e &= y - Xb \\ &= y - X(X'X)^{-1}X'y \\ &= (I - X\left(X'X\right)^{-1}X')y \\ &= (I - P)y, \end{aligned} \tag{1}$$

where $P = X\left(X'X\right)^{-1}X'$. The matrix P is called the projection matrix. We also let $I - P = M$. Then, we may write (1) as

$$y = Xb + e = Py + My.$$

We often write $Py = \hat{y}$. This is the part of y that is explained by X.

Least squares estimation

- Properties of the matrices P and M are:

1. $P' = P$, $P^2 = P$ (idempotent matrix)
2. $M' = M$, $M^2 = M$
3. $PX = X$, $MX = 0$
4. $PM = 0$

Least squares estimation

- Using (1) and (iii), we have

$$X'e = X'My = 0.$$

If the first column of X is $\mathbf{x}_1 = (1, \cdots, 1)'$, this relation implies

$$\mathbf{x}_1'e = \sum_{i=1}^{n} e_i = 0.$$

- In addition, (iv) gives

$$y'y = y'P'Py + y'M'My = \hat{y}'\hat{y} + e'e$$

Partitioned regression

- Consider

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon. \;\; X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

The normal equations for b_1 and b_2 are

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}.$$

Partitioned regression

- The first part of these equations are

$$\left(X_1'X_1\right)b_1 + \left(X_1'X_2\right)b_2 = X_1'y$$

which gives

$$\begin{aligned} b_1 &= \left(X_1'X_1\right)^{-1}X_1'y - \left(X_1'X_1\right)^{-1}X_1'X_2b_2 \\ &= \left(X_1'X_1\right)^{-1}X_1'\left(y - X_2b_2\right). \end{aligned}$$

Partitioned regression

- Plug this into the second part of the normal equations. Then, we have

$$\begin{aligned}
& X_2'X_1 b_1 + X_2'X_2 b_2 \\
= \; & X_2'X_1 \left(X_1'X_1\right)^{-1} X_1'y - X_2'X_1 \left(X_1'X_1\right)^{-1} X_1'X_2 b_2 + X_2'X_2 b_2 \\
= \; & X_2'X_1 \left(X_1'X_1\right)^{-1} X_1'y + X_2' \left(I - P_{X_1}\right) X_2 b_2 \\
= \; & X_2'y.
\end{aligned}$$

Thus

$$b_2 = \left(X_2' \left(I - P_{X_1}\right) X_2\right)^{-1} X_2' \left(I - P_{X_1}\right) y.$$

In the same manner,

$$b_1 = \left(X_1' \left(I - P_{X_2}\right) X_1\right)^{-1} X_1' \left(I - P_{X_2}\right) y.$$

Partitioned regression

- Suppose that

$$X_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ and } X_2 = Z_{(n \times K_2)}.$$

Then

$$b_2 = \left(Z' \left(I - P_{\mathbf{1}}\right) Z\right)^{-1} Z' \left(I - P_{\mathbf{1}}\right) y.$$

Partitioned regression

- But

$$\left(I - P_{\mathbf{1}}\right) Z = Z - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'Z$$

and

$$\mathbf{1}'\mathbf{1} = n$$

$$\begin{aligned} \mathbf{1}'Z &= \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} z_{11} & \cdots & z_{1K_2} \\ \vdots & & \\ z_{n1} & \cdots & z_{nK_2} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} z_{i1} & \cdots & \sum_{i=1}^{n} z_{iK_2} \end{pmatrix}. \end{aligned}$$

Partitioned regression

- Thus,

$$
\begin{aligned}
(I - P_{\mathbf{1}})\,Z &= Z - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \bar{z}_1 & \cdots & \bar{z}_{K_2} \end{pmatrix} \\
&= \begin{pmatrix} z_{11} - \bar{z}_1 & \cdots & z_{1K_2} - \bar{z}_{K_2} \\ z_{21} - \bar{z}_1 & \cdots & z_{2K_2} - \bar{z}_{K_2} \\ \vdots & & \\ z_{n1} - \bar{z}_1 & \cdots & z_{nK_2} - \bar{z}_{K_2} \end{pmatrix}
\end{aligned}
$$

In the same way,

$$
(I - P_{\mathbf{1}})\,y = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.
$$

- These show that b_2 is equivalent to the OLS estimator of β in the demeaned regression equation

$$y_i - \bar{y} = \beta' \left(z_i - \bar{z} \right) + \varepsilon_i .$$

$$\left(\bar{z} = \left(\bar{z}_1, \cdots, \bar{z}_{K_2} \right)' \right)$$

Whether we demean the data and run regression or put a constant term in the model and run regression, we get the same results.

- Write

$$y = Xb + e = \hat{y} + e.$$

Let

$$M^0 = I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}' \text{ with } \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

M_0 transforms observations into deviations from sample means.

- Then

$$\begin{aligned} M^0 y &= M^0 Xb + M^0 e \\ &= M^0 Xb + e \end{aligned}$$

or

$$y - \mathbf{1}\bar{y} = \hat{y} - \mathbf{1}\bar{y} + e.$$

- The total sum of variation (TSS) of y_i is

$$\begin{array}{ccccc} y'M^0 y & = & b'X'M^0 Xb & + & e'e. \\ \| & & \| & & \| \\ \sum_{i=1}^n (y_i - \bar{y})^2 & & \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 & & \sum_{i=1}^n e_i^2 \end{array}$$

- Note that

$$
\begin{aligned}
b'X'M^0e &= b'X'M^0M\varepsilon \\
&= b'X'\left(I - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'\right)M\varepsilon \\
&= b'X'M\varepsilon - b'X'\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'M\varepsilon \\
&= 0
\end{aligned}
$$

because $X'M = 0$ and $\mathbf{1}'M = 0$. The term $b'X'M^0b$ is called the explained sum of squares (ESS), and $e'e$ the residual sum of squares (RSS).

- How well the regression line fits the data can be explained by

$$R^2 = \frac{ESS}{TSS} = \frac{b'XM^0Xb}{y'M^0y} = 1 - \frac{e'e}{y'M^0y}.$$

We call R^2 the coefficient of determination.

(i)

$$0 \leq R^2 \leq 1$$

0 : no fit
1 : perfect fit

(ii)

$R^2_{X,Z}$: R^2 for the regression of y on X and an additional variable Z.
R^2_X : R^2 for the regression of y on X.

Then

$$R^2_{X,Z} = R^2_X + \left(1 - R^2_X\right) r^{*2}_{yz} \tag{2}$$

where

$$r^{*2}_{yz} = \frac{\left(z_*' y_*\right)^2}{\left(z_*' z_*\right)\left(y_*' y_*\right)}, \;\; z_* = \left(I - P_X\right) z, \;\; y_* = \left(I - P_X\right) y.$$

The coefficient of determination R^2 increases as the number of regressors increases whatever quality the additional regressors have.

- Theil's $\bar{R}^2$ (adjusted R^2)

$$\bar{R}^2 = 1 - \frac{e'e/\left(n-K\right)}{y'M^0y/\left(n-1\right)} = 1 - \frac{n-1}{n-K}\left(1-R^2\right)$$

$\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the t–ratio associated with this variable is greater (less) than 1.

(i) AIC (Akaike Information Criterion)

- $AIC(K) = \ln \frac{e'e}{n} + \frac{2K}{n}$
- Select a set of regressors that minimize AIC.
- AIC was designed to be an approximately unbiased estimator of the expected Kullback-Leibler information of a fitted model.
- If the true model is finite dimensional, AIC does not provide consistent model order selections.
- AIC tends to overfit.

(ii) AIC_c (Corrected AIC)

- See Hurvich and Tsai (1989), "Regression and time series model selection in small samples," Biometrika, 76, 297–307.
- AIC_c is a bias-corrected version of AIC

$$AIC_c = AIC + \frac{2(K+1)(K+2)}{T-K-2}$$

- AIC_c is useful particularly in finite samples.

(iii) *BIC* (Bayesian information criterion)

- $BIC\left(K\right) = \ln \frac{e'e}{n} + \frac{K \ln n}{n}$
- *BIC* also tends to overfit as *AIC* does, but it appears that *BIC* is uniformly better than *AIC* at selecting the correct model (see Hurvich and Tsai (1990), "The impact of model selection on inference in linear regression," American Statistician, vol. 44, for some simulation results regarding linear regression).

An alternative way of deriving OLS

- Write the objective function for the least squares estimation as

$$S(\beta) = (y - X\beta)'(y - X\beta)$$

  and let the OLS be $b = \arg\min_\beta S(\beta)$. The residual vector is $\hat{u} = y - Xb$.

- Write

$$\begin{aligned} S(\beta) &= (y - Xb + Xb - X\beta)'(y - Xb + Xb - X\beta) \\ &= (\hat{u} + X(b - \beta))'(\hat{u} + X(b - \beta)) \\ &= \hat{u}'\hat{u} + 2(b - \beta)'X'\hat{u} + (b - \beta)'X'X(b - \beta). \end{aligned}$$

An alternative way of deriving OLS

Lemma Let $f(x) = a + b'x + x'Hx$, where b is an $n \times 1$ vector and H is an $n \times n$ symmetric matrix. Then, $f(\cdot)$ is minimized uniquely at $x = 0$ if and only if $b = 0$ and $H > 0$.

Proof Assume $b = 0$ and $H > 0$. Then, $f(0) = 0$ and $f(x) > a$ for all $x \neq 0$. Thus, $f(\cdot)$ is minimized at $x = 0$. This proves the sufficiency part of the lemma.

An alternative way of deriving OLS

Proof (continued) Assume $H \leq 0$ and b is an arbitrary vector. Choose x^o $(\neq 0)$ such that $b'x^o \leq 0$. Then, $f(x^o) = a + b'x^o + x^{o\prime}Hx^o \leq a$. Thus, $f(\cdot)$ is not minimized uniquely at $x = 0$. Assume $H > 0$ but $b \neq 0$. Put $y = -\frac{1}{2}H^{-1}b$. Then,

$$\begin{aligned} f(y) &= a - \frac{1}{2}b'H^{-1}b + \frac{1}{4}b'H^{-1}b \\ &= a - \frac{1}{4}b'H^{-1}b \leq a, \end{aligned}$$

since $H^{-1} > 0$. Thus, $f(\cdot)$ is not minimized uniquely at $x = 0$. This proves the necessity part of the lemma.

An alternative way of deriving OLS

This lemma shows that $S(\beta)$ is uniquely minimized at b if and only of $X'\hat{u} = 0$ and $X'X > 0$. Since $X'\hat{u} = X'(y - Xb) = 0$, $b = (X'X)^{-1}X'y$.

Exercises

1. Instead of estimating the coefficients β_1 and β_2 in model[1]

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{3}$$

it is decided to use OLS on the following equation

$$y = X_1^*\beta_1 + X_2\beta_2 + \varepsilon^*, \tag{4}$$

where X_1^* is the residual vector from the regression of X_1 on X_2.
a. Show that the OLS estimator of β_2 in model (4) is the same as the OLS coefficient estimator of y on X_2.
b. Prove that the OLS estimators of β_1 in models (3) and (4) are identical.

[1] Assume β_1 is a scalar.

Exercises

2. In the linear regression model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

under what condition $b_1 = (X_1'X_1)^{-1}X_1'y$?

3. Show that the OLS estimators of β_1 in the following regression equations are identical.

$$\begin{aligned} y_t &= X_t\beta_1 + t\beta_2 + e_t; \\ y_t^* &= X_t^*\beta_1 + u_t \end{aligned}$$

where y_t^* and X_t^* are detrended y_t and X_t, respectively, obtained by regressing y_t and X_t on t and setting y_t^* and X_t^* equal to the respective residuals.

4. Show that $(\hat{y} - \mathbf{1}\bar{y})' e = 0$.
5. Prove relation (2).
6. Prove the following statement.

   $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the t–ratio associated with this variable is greater (less) than 1.

Exercises

7. In the linear regression model

$$y = X\beta + \varepsilon,$$

there is a need for changing the unit of measurement for the dependent variable y. So $y^* = cy$ (c is a constant) is now used as a dependent variable.

a. Does this practice change R^2?

b. What happens to R^2 if the unit of measurement is changed only for the regressor?

8. Consider the linear regression model

$$y_i = \alpha + \beta' X_i + \varepsilon_i, \ \varepsilon_i \sim iid\left(\mu, \sigma^2\right), \ \mu \neq 0$$

a. Is the OLS estimator of β affected by the nonzero mean of ε_i?
b. Can the least squares estimator of α estimate it accurately?

9. Discuss the validity of the following statements.

a. Sum of residuals is always zero.

b. If a regression produces R^2 greater than 0.5, the regression is a reliable one.

c. In a regression model

$$y_i = \alpha x_i + \varepsilon_i,$$

switching the independent and dependent variables and running a least squares provide a valid estimator of $\frac{1}{\alpha}$.

d. $\bar{R}^2$ tends to favor larger models.

10. Data on wages from a group of women and a group of men are available. Denote them as $\{w_i\}_{i=1}^{N_W}$ and $\{m_i\}_{i=1}^{N_M}$, respectively. Note that N_W and N_M are the numbers of samples. In order to study gender difference in wage, a statistician considers using the difference in sample means, i.e., $\bar{w} - \bar{m}$ with $\bar{w} = \frac{1}{N_W}\sum_{i=1}^{N_W} w_i$ and $\bar{m} = \frac{1}{N_M}\sum_{i=1}^{N_M} m_i$. Another statistician intends to use the regression model

$$y = \beta_0 + \beta_1 D + \varepsilon,$$

where $y = [w_1, ..., w_{N_W}, m_1, ..., m_{N_M}]'$ and $D = [1, ..., 1, 0, ...0]$, where the number of 1's in D is equal to N_W. D is a collection of dummy variables.

Exercises

a. Show that the OLS estimator of β_1 is equal to $\bar{w} - \bar{m}$.
b. Assume that $\varepsilon_i \sim iid(0, \sigma^2)$ for all i. Is it equivalent to assuming common variance for w_i and m_i?

Exercises

c. If w_i and m_i have the common variance σ^2, the usual t-ratio using $\bar{w} - \bar{m}$ is defined by

$$\frac{\bar{w} - \bar{m}}{\sqrt{\frac{\hat{\sigma}^2}{N_W} + \frac{\hat{\sigma}^2}{N_M}}},$$

where $\hat{\sigma}^2 = \frac{1}{N_W + N_M - 2}\left(\sum_{i=1}^{N_W}(w_i - \bar{w})^2 + \sum_{i=1}^{N_M}(m_i - \bar{m})^2\right)$. Is this equivalent to the t-ratio for the null hypothesis $H_0 : \beta_1 = 0$ that uses the regression model?

Exercises

11. a. Using the partial regression result, show that

$$P_X = X_1(X_1' M_{X_2} X_1)^{-1} X_1' M_{X_2} + X_2(X_2' M_{X_1} X_2)^{-1} X_2' M_{X_1},$$

where $X = [X_1, X_2]$.
b. Show that matrix $X_1(X_1' M_{X_2} X_1)^{-1} X_1' M_{X_2}$ is idempotent.

Exercises

12. Show that Wald test statistics for β_1 in the following two regression equations

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + e; \\ y^* &= X_1^*\beta_1 + u \end{aligned}$$

are identical. Here $y^* = (I - P_{X_2})y$ and $X_1^* = (I - P_{X_2})X_1$. The divisor for the computation of the estimator of the error variance is set to be the sample size.
(Hint: $y'P_X y = y'P_{X_2}y + y^{*\prime}P_{X_1^*}y^*$, where $X = [X_1 \; X_2]$.)