Linear Models and Estimation by Least Squares

Jin-Lung Lin

1 Introduction

Causal relation investigation lies in the heart of economics.
Effect (Dependent variable) $\leftarrow$ cause (Independent variable)

Example:

Demand for money depends upon income, interest rate, etc.

Two most frequently estimation methods:

- Maximum likelihood estimate
- Least square estimate

2 Linear Statistical Models

Definition 11.1 A linear statistical model relating a random response Y to a set of independent variables $x_1, x_2, \ldots, x_k$ is of the form

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_{2i} x_2 + \cdots + \beta_k x_{ki} + \epsilon_i \quad i = 1, 2, \cdots, T \\ E(\epsilon_t) &= 0, \quad \forall t \\ V(\epsilon_t) &= \sigma^2, \quad \forall t \\ E(\epsilon_t \epsilon_s) &= 0, \quad \forall t \neq s \\ E(x_{it}\epsilon_t) &= i = 1, 2, \cdots, k0 \end{aligned}$$

Take expectation of both sides to obtain

$$\mu_y = \beta_0 + \beta_1 \mu_{x1} + \cdots + \beta_k \mu_{xk}$$

and subtract it from the regression model,

$$(y_t - \mu_y) = \beta_1(x_{1t} - \mu_{x1}) + \beta_2(x_{2i} - \mu_{x2}) + \cdots + \beta_k(x_{ki} - \mu_{xk}) + \epsilon_t, \quad t = 1, 2, \cdots, T$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ϵ is a random variable, and $x_1, x_2, \ldots, x_k$ are fixed variables.

3 The Method of Least Squares

Simple Linear Regression Model

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \epsilon \\ \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \end{aligned}$$

Minimize the sum of squared deviation:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Least Squares Estimators for Simple Linear Regression Model

1. $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, where $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, and $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

2. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

   Note that

$$\begin{aligned} (\hat{Y} - \bar{y}) &= \hat{\beta}_1 (x - \bar{x}) \\ \hat{\beta}_1 &= \sum_{i=1}^{n} w_i y_i \\ \hat{\beta}_0 &= \sum_{i=1}^{n} v_i y_i \\ w_i &= \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ v_i &= \frac{1}{n} - \bar{x} w_i \end{aligned}$$

$$\begin{aligned} \sum_{i}^{n} w_i &= 0 \\ \sum_{i=1}^{n} w_i^2 &= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

4 Properties of the Least Squares Estimators-Simple Linear Regression

1. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, that is, $E(\hat{\beta}_i) = \beta_i$, for i=0,1.

2. $V(\hat{\beta}_0) = c_{00}\sigma^2$, where $c_{00} = \frac{\sum x_i^2}{nS_{xx}}$.
3. 3. $V(\hat{\beta}_1) = c_{11}\sigma^2$, where $c_{00} = \frac{1}{S_{xx}}$.
4. $Cov(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$, where $c_{01} = \frac{-\bar{x}}{S_{xx}}$.
5. An unbiased estimator of σ^2 is $S^2 = SSE/(n-2)$, where $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ and $S_{yy} = \sum(y_i - \bar{y})^2$.
   If, in addition, the ϵ_i, for $i = 1, 2, \ldots, n$ are normally distributed:
6. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal distributed.
7. The random variable $\frac{(n-2)S^2}{\sigma^2}$ has a χ^2 distribution with n-2 degrees of freedom.
8. The statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

5 Inference Concerning the Parameters β_i

Test of Hypothesis for β_i
H_0 :

$$\beta_i = \beta_{i0}$$

H_a :

$$\begin{cases} \beta_i > & \beta_{i0} \quad (upper-tail \quad alternative) \\ \beta_i < & \beta_{i0} \quad (lower-tail \quad alternative) \\ \beta_i \neq & \beta_{i0} \quad (two-tailed \quad alternative) \end{cases}$$

Test statistic: $T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$
Rejection region:

$$\begin{cases} t > & t_\alpha \quad (upper-tail \quad alternative) \\ t < & -t_\alpha \quad (lower-tail \quad alternative) \\ |t| > & t_{\alpha/2} \quad (two-tailed \quad alternative) \end{cases}$$

where $c_{00} = \frac{\sum x_i^2}{nS_{xx}}$, and $c_{11} = \frac{1}{S_{xx}}$.
Notice that t_α is based on (n-2) degrees of freedom.
A $(1-\alpha)$ 100 % Confidence Interval for β_i

$$\hat{\beta} \pm t_{\alpha/2} S\sqrt{c_{ii}}$$

where

$$c_{00} = \frac{\sqrt{x}_i^2}{nS_{xx}}, c_{11} = \frac{1}{S_{xx}}$$

6 Inference Concerning Linear Functions of the Model Parameters: Simple Linear Regression

A Test for $\theta = \mathbf{a_0}\beta_0 + \mathbf{a_1}\beta_1$
H_0 :

$$\theta = \theta_0$$

H_a :

$$\begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

Test statistic: $T = \frac{\hat{\theta}-\theta_0}{S\sqrt{(\frac{a_0^2 \sum x_i^2/n + a_1^2 - 2a_0 a_1 \bar{x}}{S_{xx}})}}$
Rejection region:

$$\begin{cases} t > t_\alpha \\ t < -t_\alpha \\ |t| \neq t_{\alpha/2} \end{cases}$$

Here t_α is based upon n-2 degrees of freedom.
A $(1-\alpha)$ **100 % Confidence Interval for** $\theta = a_0\beta_0 + a_1\beta_1$

$$\hat{\theta} \pm t_{\alpha/2} S \sqrt{(\frac{a_0^2 \sum x_i^2/n + a_1^2 - 2a_0 a_1 \bar{x}}{S_{xx}})}$$

where the tabulated $t_{\alpha/2}$ is based upon n-2 degrees of freedom.
A $(1-\alpha)$ **100 % Confidence Interval for** $E(Y) = \beta_0 + \beta_1 x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1/n + (x^* - \bar{x})^2/S_{xx}}$$

where the tabulated $t_{\alpha/2}$ is based upon n-2 degrees of freedom.

7 Predicting a Particular Value of Y Using Simple Linear Regression

A $(1-\alpha)$ **100 % Prediction Interval for Y when** $x = x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + 1/n + (x^* - \bar{x})^2/S_{xx}}$$

8 Correlation

In $Y = \beta_0 + \beta_1 x + \epsilon$, x could be a fixed regressor controlled by researchers or can be observed values of a random variable. For the former, $E(Y) = \beta_0 + \beta_1 x$. For the latter, $E(Y|X = x) = \beta_0 + \beta_1 x$. Typically, we often assume that (X, Y) has a bivariate normal distribution with $E(X) = \mu_x, E(Y) = \mu_y, V(X) = \sigma_x^2, V(Y) = \sigma_y^2, Cov(X, Y) = \sigma_{xy} = \rho\sigma_x\sigma_y$. In this case,

$$E(Y|X = x) = \beta_0 + \ \beta_1 x, \text{ where } \beta_1 = \frac{\sigma_y}{\sigma_x}\rho$$

The MLE of correlation ρ is given by the sample correlation coefficient:

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{X_{xy}}{\sqrt{S_{xx} S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}} \end{aligned}$$

Testing $\beta_1 = 0$ is equivalent to testing $\rho = 0$.

$$\begin{aligned} t &= \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} \\ &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \end{aligned}$$

Z-test
$(1/2)\log[(1 + r)/(1 - r)] \sim N((1/2)\log[(1 + \rho)/(1 - \rho)], 1/(n - 3))$

$$Z = \frac{(\frac{1}{2})\log(\frac{1+r}{1-r}) - (\frac{1}{2})\log(\frac{1+\rho_0}{1-\rho_0})}{\frac{1}{\sqrt{n-3}}}$$

9 Fitting the Linear Model by Using Matrices

Least Squares Equations and Solutions for a General Linear Model
Equations : $(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{Y}$
Solutions : $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
$\mathbf{SSE} = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$

10 Linear Functions of the Model Parameters: Multiple Linear Regression

Multiple Linear Regression

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, \ldots, n. \\ \hat{\beta} &= (X'X)^{-1}X'Y \end{aligned}$$

Properties of the Least Squares Estimators- Multiple Linear Regression

1. $E(\hat{\beta}_i) = \beta_i, i = 0, 1, \ldots, k.$
2. $V(\hat{\beta}_i) = c_{ii}\sigma^2$, where c_{ij} (or, in this special case, c_{ii}) is the element in row i and column j (here, column i) of $(X'X)^{-1}$.
3. $Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$
4. An unbiased estimator of σ^2 is $S^2 = SSE/[n - (k+1)]$, where $SSE = YY' - \hat{\beta}'X'Y$.
5. Each $\hat{\beta}_i$ is normally distributed.
6. The random variable
$$\frac{[n - (k+1)]S^2}{\sigma^2}$$
has a χ^2 distribution with n-(k+1) degrees of freedom.
7. The statistics S^2 and $\hat{\beta}_i$, for $i = 0, 1, 2, \ldots, k$ are independent.

11 Inference Concerning Linear Functions of the Model Parameters: Multiple Linear Regression

A Test for a$'\beta$
H_0 :

$$a'\beta = (a'\beta)_0$$

H_a :

$$\begin{cases} a'\beta > (a'\beta)_0 \\ a'\beta < (a'\beta)_0 \\ a'\beta \neq (a'\beta)_0 \end{cases}$$

Test statistic:

$$T = \frac{a'\hat{\beta} - (a'\beta)_0}{S\sqrt{a'(X'X)^{-1}a}}$$

Rejection region:

$$\begin{cases} t > t_\alpha \\ t < -t_\alpha \\ |t| \neq t_{\alpha/2} \end{cases}$$

Here t_α is based upon [n-(k+1)] degrees of freedom.
A $(1-\alpha)$ 100 % Confidence Interval for $a'\beta$

$$a'\hat{\beta} \pm t_{\alpha/2} S\sqrt{a'(X'X)^{-1}a}$$

12 Predicting a Particular Value of Y Using Multiple Regression

A $(1-\alpha)$ 100 % Prediction Interval for Y when $x_1 = x_1^*, x_2 = x_2^*, \ldots, x_k = x_k^*$

$$a'\hat{\beta} \pm t_{\alpha/2} S\sqrt{1 + a'(X'X)^{-1}a}$$

where $a' = [1, x_1^*, x_2^*, \ldots, x_k^*]$

13 A test for $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

Reduced model (R):

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \epsilon$$

Complete model (C):

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k \epsilon$$

$$SSE_R = SSE_C + (SSE_R - SSE_C)$$

$$\begin{aligned} \chi_3^2 &= \frac{SSE_R}{\sigma^2} \\ \chi_2^2 &= \frac{SSE_C}{\sigma^2} \\ \chi_1^2 &= \frac{SSE_R - SSE_C}{\sigma^2} \end{aligned}$$

$$\begin{aligned} F &= \frac{\chi_1^2/(k-g)}{\chi_2^2/(n-[k+1])} = \frac{SSE_R - SSE_C/(k-g)}{SSE_C/(n-[k+1])} \\ F &> F_\alpha \end{aligned}$$