Chapter 2. Least Square Regression (II)

Multiple Linear Regression Models

Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

1. Y: dependent variable.
2. x: independent (or, explanatory) variable.
3. ε: random error, with $E[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$.
4. $\beta_0, \beta_1, \ldots, \beta_k$: population parameter, unknown.

$$E[Y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Matrix notation

Data of size n: The i-th data point $(x_{i1}, x_{i2}, \ldots, x_{ik}; Y_i)$, $i = 1, 2, \ldots, n$. And

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}_{n \times (k+1)}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

Method of least square

Find $\hat{\beta}$ so as to minimize Sum of squares for error

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

Solution:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Proof and the special case of $k = 1$.

Properties of the regression

1. $\hat{\beta}$ is an unbiased estimate for β.
2. The variance matrix
$$\mathrm{Var}[\hat{\beta}] = \sigma^2 (X'X)^{-1}.$$
3. $\mathrm{SSE} = Y'(Y - X\hat{\beta}) = Y'Y - \hat{\beta}'X'Y$.
4. An unbiased estimate for σ^2 is
$$S^2 = \frac{1}{n-k-1}\mathrm{SSE}$$

If we further assume that $\varepsilon_i \sim N(0, \sigma^2)$, then

1. $\hat{\beta}$ is normally distributed.
2. $$\frac{n-k-1}{\sigma^2} S^2 \sim \chi^2(n-k-1)$$
3. S^2 and $\hat{\beta}$ are independent.

Inference of regression coefficient

Assumption: $\varepsilon_i \sim N(0, \sigma^2)$

Variance matrix:

$$\mathrm{Var}[\hat{\beta}] = \sigma^2 (X'X)^{-1}$$

Quantity of interest: $\theta = a'\beta$, where $a = (a_0, a_1, \ldots, a_k)'$.

An unbiased estimate for θ is $a'\hat{\beta}$.

Proposition.

$$T = \frac{a'\hat{\beta} - \theta}{S\sqrt{a'(X'X)^{-1}a}} \sim t(n-k-1).$$

1. Confidence interval for θ.

2. Hypothesis testing for θ. (Two-sided and one-sided).

3. The special case of $\theta = \beta_i$.

Estimation and Prediction

Two problems:

1. Estimate the population mean of Y given $x = (x_1^*, x_2^*, \ldots, x_k^*)$.
2. Predict an individual Y given $x = (x_1^*, x_2^*, \ldots, x_k^*)$.

Answer: Let

$$a = \begin{bmatrix} 1 \\ x_1^* \\ x_2^* \\ \vdots \\ x_k^* \end{bmatrix}$$

1.

$$\hat{\mu}_Y = a'\hat{\beta}.$$

$$T = \frac{\hat{\mu}_Y - \mu_Y}{S\sqrt{a'(X'X)^{-1}a}} \sim t(n-k-1).$$

Confidence interval and hypothesis testing.

2.

$$\hat{Y} = a'\hat{\beta}.$$

$$\text{error} = \hat{Y} - Y = a'\hat{\beta} - (a'\beta + \varepsilon)$$

$$\text{error} \sim N\left(0, \sigma^2[1 + a'(X'X)^{-1}a]\right)$$

$$T = \frac{\hat{Y} - Y}{S\sqrt{1 + a'(X'X)^{-1}a}} \sim t(n - k - 1)$$

Analysis of Variance for Multiple Regression

Testing the hypothesis that none of the explanatory variables contribute to the regression?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

Total variation of Y.

$$\text{SST} \doteq \sum_i (Y_i - \bar{Y})^2.$$

- Variation accounted for by linear regression.

$$\text{SSR} \doteq \sum_i (\hat{Y}_i - \bar{Y})^2.$$

- Variation NOT accounted for by linear regression.

$$\text{SSE} \doteq \sum_i (Y_i - \hat{Y}_i)^2.$$

$$\text{SST} = \text{SSR} + \text{SSE},$$

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = Y'Y - n\bar{Y}^2, \qquad \text{SSE} = Y'(Y - X\hat{\beta})$$

Source of Variation	d.f.	Sum of Squares	Mean Square
Regression	k	SSR	$\text{MSR} = \text{SSR}/k$
Error	$n-k-1$	SSE	$\text{MSE} = S^2$
Total	$n-1$	SST	

SSR and SSE are independent.

F-test

$$H_0 : \beta_1 = \cdots = \beta_k = 0, \quad H_a : \text{one of } \beta_i \text{ is non-zero.}$$

Under H_0,

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2(k), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-k-1)$$

and

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(k, n-k-1).$$

Rejection Region with significance level $\alpha = \{F > F_\alpha\}$.

Extension of Analysis of Variance

Testing the hypothesis that only a subset of the explanatory variables, say $(x_1, x_2, \ldots, x_g)$, contribute to the regression.

$$H_0 : \beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0.$$

In general,

g = number of explanatory variables in the reduced model.

1. Reduced model: Fit linear regression with explanatory variables $(x_1, x_2, \ldots, x_g)$.

   SSE_R = sum of squares for errors in the reduced model
   = variance NOT accounted for by the reduced model.

2. Complete model: Fit linear regression with explanatory variables $(x_1, x_2, \ldots, x_g, x_{g+1}, \ldots, x_k)$.

   SSE_C = sum of squares for errors in the complete model
   = variance NOT accounted for by the complete model.

$\text{SSE}_R - \text{SSE}_C$ = Variance accounted for by variables $(x_{g+1}, \ldots, x_k)$, after variables $(x_1, \ldots, x_g)$.

Source of Variation	d.f.	Sum of Squares	Mean Square
by $(x_{g+1}, \ldots, x_k)$ after $(x_1, \ldots, x_g)$	$k-g$	SSE_R - SSE_C	$\text{SSE}_R - \text{SSE}_C/(k-g)$
Complete model	$n-k-1$	SSE_C	$\text{SSE}_C/(n-k-1)$
Reduced model	$n-g-1$	SSE_R	

$\text{SSE}_R - \text{SSE}_C$ is independent of SSE_C.

Disscusion on the special case where $g = 0$.

F-test

$$H_0 : \beta_{g+1} = \cdots = \beta_k = 0.$$

Under H_0,

$$\frac{\text{SSE}_R - \text{SSE}_C}{\sigma^2} \sim \chi^2(k-g), \quad \frac{\text{SSE}_C}{\sigma^2} \sim \chi^2(n-k-1)$$

and

$$F = \frac{(\text{SSE}_R - \text{SSE}_C)/(k-g)}{\text{SSE}_C/(n-k-1)} \sim F(k-g, n-k-1).$$

Rejection Region with significance level $\alpha = \{F > F_\alpha\}$.

Examples

1. Data from the study of metabolic rate on a set of dogs.

Log-body mass (kg)	3.44	3.18	2.99	2.90	2.26	1.87	1.16
Log-body surface (cm^2)	9.28	9.08	8.92	8.94	8.57	8.22	7.79
Log-Metabolic rate (kcal/day)	7.01	6.89	6.81	6.74	6.44	6.06	5.64

(a) Write down the vector Y and the matrix X. Compute $X'X$, $(X'X)^{-1}$, $X'Y$, and $Y'Y$.

(b) Calculate $\hat{\beta}$, and find an unbiased estimate for σ^2.

(c) What is you estime for the average metabolic rate for a dog with weight 12 kg and surface area 6000 cm^2. Find a 95% confidence interval.

(d) Test the following three hypotheses: (i) $H_0 : \beta_1 = \beta_2 = 0$; (ii) $H_0 : \beta_1 = 0$; (iii) $H_0 : \beta_2 = 0$.

Solution: (a) & (b) $n = 7$, $k = 2$.

$$Y = \begin{bmatrix} 7.01 \\ 6.89 \\ 6.81 \\ 6.74 \\ 6.44 \\ 6.06 \\ 5.64 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 3.44 & 9.28 \\ 1 & 3.18 & 9.08 \\ 1 & 2.99 & 8.92 \\ 1 & 2.90 & 8.94 \\ 1 & 2.26 & 8.57 \\ 1 & 1.87 & 8.22 \\ 1 & 1.16 & 7.79 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 7.0 & 17.8 & 60.8 \\ 17.8 & 49.25 & 157.17 \\ 60.8 & 157.17 & 529.75 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 45.59 \\ 118.36 \\ 397.55 \end{bmatrix}.$$

$$(X'X)^{-1} = \begin{bmatrix} 5235.87 & 477.48 & -742.59 \\ 477.48 & 43.93 & -67.83 \\ -742.59 & -67.83 & 105.35 \end{bmatrix}.$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 0.191 \\ 0.175 \\ 0.677 \end{bmatrix},$$

$$\text{SSE} = Y'Y - Y'X\hat{\beta} = Y'Y - \hat{\beta}'X'Y = 0.01157$$

$$\text{An unbiased estimate for } \sigma^2 \text{ is } S^2 = \frac{\text{SSE}}{n-k-1} = 0.00289.$$

(c). $x_1^* = \ln(12) = 2.48$, $x_2^* = \ln(6000) = 8.70$. Let

$$a = \begin{bmatrix} 1 \\ x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 1 \\ 2.48 \\ 8.70 \end{bmatrix}$$

The estimate is

$$a'\hat{\beta} = 6.512.$$

A 95% confidence interval is

$$a'\hat{\beta} \pm t_{0.025}(n-k-1)S\sqrt{a'(X'X)^{-1}a}$$

or

$$6.512 \pm 2.776 * \sqrt{0.00289} * \sqrt{0.418} = [6.415, 6.609]$$

The estimate of average metabolic rate is $e^{6.512} = 673$, and a 95% confidence interval is $[e^{6.415}, e^{6.609}] = [611, 742]$.

(d). (i) Testing the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. $g = 0$.
The complete model use both explanatory variables x_1, x_2, and

$$\text{SSE}_C = 0.01157.$$

The reduced model uses none of the explanatory variables, that is, the reduced model is

$$Y = \beta_0 + \varepsilon.$$

and

$$\text{SSE}_R = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 1.50154.$$

Source of Variation	d.f.	Sum of Squares	Mean Square	F	P-value
by (x_1, x_2)	2	1.48997	0.745		
Complete model	4	0.01157	0.00289	$\frac{0.745}{0.00289} = 257$	< 0.005
Reduced model	6	1.50154			

(ii) Testing the hypothesis $H_0 : \beta_1 = 0$. $g = 1$

The complete model use both explanatory variables x_1, x_2, and

$$\text{SSE}_C = 0.01157.$$

The reduced model only uses the explanatory variables x_1, that is, the reduced model is

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

and (Discussion)

$$\text{SSE}_R = 0.01227$$

Source of Variation	d.f.	Sum of Squares	Mean Square	F	P-value
by (x_2)	1	0.00070	0.00070		
Complete model	4	0.01157	0.00289	$\frac{0.00070}{0.0289} = 0.25$	0.64
Reduced model	5	0.01227			

Remark: Another way to test $H_0 : \beta_1 = 0$ is to use the t-test. This is equivalent to the F-test.

Under H_0,

$$T = \frac{\hat{\beta}_1 - 0}{S\sqrt{a'(X'X)^{-1}a}} \sim t(n-k-1)$$

where

$$\hat{\beta}_1 = a'\hat{\beta}, \text{ or } a = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$T = \frac{0.175}{\sqrt{0.00289 * 43.93}} = 0.5$$

$$\text{P-value} = 2P(t(4) > |T|) = 0.64.$$

Note $T^2 = F$.

(iii) Testing the hypothesis $H_0 : \beta_2 = 0$. $g = 1$

The complete model use both explanatory variables x_1, x_2, and

$$\text{SSE}_C = 0.01157.$$

The reduced model only uses the explanatory variables x_2, that is, the reduced model is

$$Y = \beta_0 + \beta_2 x_2 + \varepsilon.$$

and

$$\text{SSE}_R = 0.01592$$

Source of Variation	d.f.	Sum of Squares	Mean Square	F	P-value
by (x_1)	1	0.00435	0.00435		
Complete model	4	0.01157	0.00289	$\frac{0.00435}{0.0289} = 1.50$	0.29
Reduced model	5	0.01592			

DISCUSSIONS: The interpretation of coefficients in multiple linear regression.

Graphs.

$$
\begin{aligned}
Y &= 0.19 + 0.175x_1 + 0.677x_2 + \varepsilon \\
Y &= 4.96 + 0.611x_1 + \qquad\quad + \varepsilon \\
Y &= -1.71 + \qquad\quad + 0.95x_2 + \varepsilon
\end{aligned}
$$

2. Data on wheat yield (x) and protein content (Y). $n = 19$.

Graph

Model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

or

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

$$x_1 = x, \quad x_2 = x^2.$$

$$X'X = \begin{bmatrix} 19 & 421 & 11469 \\ 421 & 11469 & 355621 \\ 11469 & 355621 & 11966805 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 1.1448 & -0.1019 & 0.001930 \\ -0.1019 & 0.01018 & -0.0002048 \\ 0.001930 & -0.0002048 & 4.319 \times 10^{-6} \end{bmatrix}$$

$$Y = 18.677 - 0.4367x + 0.005869x^2 + \varepsilon$$

$$\text{SSE} = 26.3, \quad S^2 = \frac{1}{19-2-1}\text{SSE} = 1.6445$$

Graph of the fitted model.

Remark: The test of departure from linearity. That is H_0 : $\hat{\beta}_2 = 0$. We use analysis of variance (one can also use t-test.)
The reduced model is

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Source of Variation	d.f.	Sum of Squares	Mean Square	F	P-value
by $(x_2 = x^2)$	1	7.975	7.975		
Complete model	16	26.311	1.644	$\frac{7.975}{1.644} = 4.85$	0.0427
Reduced model	17	34.286			