Chapter 1. Linear regression Model

January 4, 2007

1 Review of linear regression models

1.1 the linear regression model

Suppose that we have variables: predictors (also called independent variable, covariates) $\mathbf{x}_1, ..., \mathbf{x}_p$ and response (also called dependent variable) Y. Statisticians usually fit the relation by a linear regression model

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p + \varepsilon.$$

(why? do you agree with this model assumption)

Here $\mathbf{x}_i$ is usually assumed to be non-random.

In practice, we usually have n observations:

observation	constant	$\mathbf{x}_1$	...	$\mathbf{x}_p$	Y
1	1	$\mathbf{x}_{1,1}$	...	$\mathbf{x}_{1,p}$	Y_1
2	1	$\mathbf{x}_{2,1}$	...	$\mathbf{x}_{2,p}$	Y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	1	$\mathbf{x}_{n,1}$	...	$\mathbf{x}_{n,p}$	Y_n

People also like to formulate it as

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 \mathbf{x}_{1,1} + ... + \beta_p \mathbf{x}_{1,p} + \varepsilon_1, \\ Y_2 &= \beta_0 + \beta_1 \mathbf{x}_{2,1} + ... + \beta_p \mathbf{x}_{2,p} + \varepsilon_2, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 \mathbf{x}_{n,1} + ... + \beta_p \mathbf{x}_{n,p} + \varepsilon_n, \end{aligned}$$

and assume that

$$E\varepsilon_1 = E\varepsilon_2 = ... = E\varepsilon_n = 0$$

(this is also called linear regression model)

By writing $X_i = (1, \mathbf{x}_{i,1}, ..., \mathbf{x}_{i,p})$, and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_{1,1} & ... & \mathbf{x}_{1,p} \\ 1 & \mathbf{x}_{2,1} & ... & \mathbf{x}_{2,p} \\ \cdots & & & \\ 1 & \mathbf{x}_{n,1} & ... & \mathbf{x}_{n,p} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

the model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \mathcal{E} \tag{1.1}$$

with

$$E\mathcal{E} = 0$$

1.2 the Least Squares estimation

the Least Squares estimation is to estimate β by minimizing

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 \mathbf{x}_{i,1} + ... + \beta_p \mathbf{x}_{i,p})\}^2$$

with respect to β. Note that

$$\begin{aligned} &\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 \mathbf{x}_{i,1} + ... + \beta_p \mathbf{x}_{i,p})\}^2 \\ &= \sum_{i=1}^{n} \{Y_i - X_i^\top \beta\}^2 \\ &= (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta). \end{aligned}$$

The estimate (solution) is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{1.2}$$

proof: Write

$$\begin{aligned} & (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\beta) \\ = & \{(I - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X})\mathbf{Y} - \mathbf{X}(\hat{\beta} - \beta)\}^\top \{(I - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X})\mathbf{Y} - \mathbf{X}(\hat{\beta} - \beta)\} \\ = & \mathbf{Y}^\top (I - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X})\mathbf{Y} + (\hat{\beta} - \beta)^\top \mathbf{X}^\top \mathbf{X}(\hat{\beta} - \beta) \end{aligned}$$

Therefore, the minimum point achieved when the second term is 0, i.e. $\beta = \hat{\beta}$

1.3 inference of linear regression model

- $\hat{\beta}$ is unbiased:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\beta + \mathcal{E}) = \beta + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathcal{E}$$

  Thus

$$E\hat{\beta} = \beta + E\{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathcal{E}\} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top E\mathcal{E} = \beta$$

- σ can be estimated by $\hat{\sigma}^2 = RSS/(n-p-1)$ where

$$RSS = \sum_{i=1}^{n} \{Y_i - X_i^\top \hat{\beta}\}^2$$

  is the residual sum of squares.

- the distribution of $\hat{\beta}$: If

$$\mathcal{E} \sim N(0, \sigma^2 I) \tag{1.3}$$

  where I is a $(p+1) \times (p+1)$ identity matrix, then

$$\hat{\beta} - \beta \sim N(0, (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2) \tag{1.4}$$

- we need to check whether $H_0 : \beta_k = 0$. Because

$$\hat{\beta}_k - \beta_k \sim N(0, c_{kk}\sigma^2)$$

  where c_{kk} is the (k,k) entry of $(\mathbf{X}^\top \mathbf{X})^{-1}$. We call $\sqrt{c_{kk}\hat{\sigma}^2}$ the standard error of $\hat{\beta}_k$. At significant level 0.05,

  - If $|\hat{\beta}_k| \le 1.96\sqrt{c_{kk}\hat{\sigma}^2}$, we accept $H_0 : \beta_k = 0$;
  - If $|\hat{\beta}_k| > 1.96\sqrt{c_{kk}\hat{\sigma}^2}$, we reject $H_0 : \beta_k = 0$.

- Other inference about the model, such as testing $\beta_1 = \beta_2 = ... = \beta_p = 0$.

Example 1.1 **(data)** *There are 5 predictors* $\mathbf{x}_1, ..., \mathbf{x}_5$ *and one response* Y*, we fit model*

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_5 \mathbf{x}_p + \varepsilon$$

The estimated model is

$$\begin{array}{lcl} Y & = & 0.20068 - 0.35520\mathbf{x}_1 + 0.98745\mathbf{x}_2 - 0.22444\mathbf{x}_3 - 0.73906\mathbf{x}_4 + 0.06922\mathbf{x}_5 \\ SE & & 0.14461 \quad\quad 0.15538 \quad\quad 0.13356 \quad\quad 0.14737 \quad\quad 0.11848 \quad\quad 0.14619 \end{array}$$

R code for teh calculation **(code)**

2 Variable selection and Cross-validation

We can see from the above example that some of the predictors actually can be removed from the model. Removing of the irrelevant regressors (predictors) is helpful in the prediction, because it can reduce the variation of the prediction. The above inference procedure can be applied for the purpose, but it is based on very strong assumption of the model; see (1.3).

We need to select variables from $\mathbf{x}_1, ..., \mathbf{x}_p$ to be included in the model. There are many candidate variables. For example,

$$\begin{aligned} \mathit{model}\ 1: &\quad Y = a_0 + a_1\mathbf{x}_1 + \varepsilon \\ \mathit{model}\ 2: &\quad Y = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_4 + \varepsilon \\ \mathit{model}\ 3: &\quad Y = c_0 + c_1\mathbf{x}_2 + \varepsilon \\ \cdots & \end{aligned}$$

Suppose we have n samples.

We need to compare the performances of all the candidate models. But RSS can not be used for the comparison.

We need to validate these models by additional data/observations by checking their prediction capability. However, there is usually no additional data. One possible idea is to partition the n samples into 2 parts: one part (called training set) for model estimation, the other part for model validation. There are many partitions. Using all the partitions is the idea of cross-validation (CV). The idea was proposed by M. Stone (1974).

If we use 1 observation for validation and the other n-1 for model estimation, it is the leave-one-observation-out cross-validation

If we use m observations for validation and the other n-m for model estimation, it is the leave-m-observation-out cross-validation.

Here is an example for the procedure: consider the models above.

For each i = 1, ..., n, we use data $(Y_1, X_1), ..., (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), ..., (Y_n, X_n)$ to estimate the models. the estimated models are, say,

$$\begin{aligned} \mathit{model}\ 1: &\quad Y = \hat{a}_0^i + \hat{a}_1^i\mathbf{x}_1 \\ \mathit{model}\ 2: &\quad Y = \hat{b}_0^i + \hat{b}_1^i\mathbf{x}_1 + \hat{b}_2^i\mathbf{x}_4 \\ \mathit{model}\ 3: &\quad Y = \hat{c}_0^i + \hat{c}_1^i\mathbf{x}_2 \\ \cdots & \end{aligned}$$

The prediction errors for (Y_i, X_i) are respectively

$$\begin{aligned}
&model\ 1: \quad err_1(i) = \{Y_i - \hat{a}_0^i - \hat{a}_1^i \mathbf{x}_{i,1}\}^2 \\
&model\ 2: \quad err_2(i) = \{Y_i - \hat{b}_0^i - \hat{b}_1^i \mathbf{x}_{i,1} - \hat{b}_2^i \mathbf{x}_{i,4}\}^2 \\
&model\ 3: \quad err_3(i) = \{Y_i - \hat{c}_0^i - \hat{c}_1^i \mathbf{x}_{i,2}\}^2 \\
&\quad \cdots
\end{aligned}$$

The overall prediction errors (also called Cross-validation value) are respectively then

$$\begin{aligned}
&model\ 1: \quad CV_1 = n^{-1} \sum_{i=1}^n err_1(i) \\
&model\ 2: \quad CV_2 = n^{-1} \sum_{i=1}^n err_2(i) \\
&model\ 3: \quad CV_3 = n^{-1} \sum_{i=1}^n err_3(i) \\
&\quad \cdots
\end{aligned}$$

The model with the smallest CV value is the model we prefer.

Example 2.1 *For the same data above* **(data)** *Our candidate models are*

$$\begin{aligned}
&\textit{model 0} \quad Y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4 + \beta_5 \mathbf{x}_5 + \varepsilon \\
&\textit{model 1} \quad Y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4 + \varepsilon \\
&\textit{model 2} \quad Y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_5 \mathbf{x}_5 + \varepsilon \\
&\textit{model 3} \quad Y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_4 \mathbf{x}_4 + \beta_5 \mathbf{x}_5 + \varepsilon \\
&\textit{model 4} \quad Y = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4 + \beta_5 \mathbf{x}_5 + \varepsilon \\
&\textit{model 5} \quad Y = \beta_0 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4 + \beta_5 \mathbf{x}_5 + \varepsilon
\end{aligned}$$

The CV values for the above model are respectivly

$$cv0 = 0.3633548, cv1 = 0.333161, cv2 = 1.216745,$$

$$cv3 = 0.3922781, cv4 = 1.400237, cv5 = 0.4589498$$

Thus model 2 is selected (and variable $\mathbf{x}_5$ *is deleted)*

R code for the calculation **(code)**