MATH3091 Statistical Modelling II
Lecture 5: Revisit linear models

Chao Zheng

14 Feb 2022

Recap

In previous lecture, we have

- asymptotics of MLE
- construct the confidence interval
- revised the log-likelihood test for testing a nested model pair
- introduced the AIC and BIC for comparing different models

All the contents so far are about just observations of (a set of) random variables. This week we are going to revise the linear model, where we have a response variable and a group of explanatory variables.

3.1.1 The linear model

Linear regression

We denote the n observations of the response variable by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. These are assumed to be observations of *random variables* $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Associated with each y_i is a vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables.

In a linear model, we assume that

$$\begin{aligned} Y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \\ &= \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \qquad i = 1, \ldots, n \end{aligned}$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently.

Matrix form

We can write our observations of explanatory variables in matrix form:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

The $n \times p$ matrix $\boldsymbol{X}$ consists of known (observed) constants and is called the *design matrix*.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$. Then we can wiret down the most economical expression of linear model in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Distribution of response variable

Instead directly assume a joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y}, \theta)$ for the observational response $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^T$, in linear model we relate it to our explanatory variables $\boldsymbol{X}$.

Since $\text{Var}(\epsilon_i) = \sigma^2$, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, as $\epsilon_1, \dots, \epsilon_n$ are independent of one another, the error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

Then the distribution of $\boldsymbol{Y}$ is multivariate normal with mean vector $\boldsymbol{X}\beta$ and variance covariance matrix $\sigma^2 \boldsymbol{I}$, *i.e.*

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}).$$

3.1.2 Examples of linear model structure

Example: the null model

If we do not include any variables x_i in the model, we have

$$Y_i = \beta_0 + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = (\beta_0).$$

This is one (dummy) explanatory variable. In practice, this variable is present in all models.

Example: simple linear regression

If we include a single variable x_i in the model, we might have

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \qquad i = 1, \ldots, n$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

There are two explanatory variables: the dummy variable and one 'real' variable.

Example: multiple regression

To include multiple explanatory variables, we might model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i\,p-1} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

for $i = 1, \ldots, n$. So

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n\,p-1} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

There are p explanatory variables: the dummy variable and $p-1$ 'real' variables.

Example: categorical explanatory variable

Suppose x_i is a categorical variable, taking values in a set of k possible categories. For simplicity we write $x_i \in \{1, \ldots, k\}$. We wish to model

$$Y_i = \mu_{x_i} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so that the mean of Y_i is the same for all observations in the same category, but differs for different categories.

We could rewrite this model to include an intercept, as

$$Y_i = \beta_0 + \beta_{x_i} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so that $\mu_j = \beta_0 + \beta_j$, for $j = 1, \ldots, k$.

Example: categorical explanatory variable (continued)

It is not possible to estimate all of the $\boldsymbol{\beta}$ parameters separately, as they only affect the distribution through the combination $\beta_0 + \beta_j$. Instead, we choose a **reference category** l, and set $\beta_l = 0$.

The intercept term β_0 then gives the mean for the reference category, with β_j giving the difference in mean between category j and the reference category.

We can rewrite the model as a form of multiple regression by first defining a new explanatory variable $\boldsymbol{z}_i$

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})^T,$$

where

$$z_{ij} = \begin{cases} 1 & \text{if } x_i = j \\ 0 & \text{otherwise.} \end{cases}$$

categorical explanatory variable (continued)

$\boldsymbol{z}_i$ is sometimes called the **one-hot encoding** of x_i, as it contains precisely one 1 (corresponding to the category x_i), and is 0 everywhere else. We then have

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} + \epsilon_i,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1k} \\ 1 & z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nk} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

where each row of $\boldsymbol{X}$ will have two ones, and the remaining entries will be zero.

We can also do linear models with more than categorical explanatory variables, and even allow an interaction between them.

3.1.3 Maximum likelihood estimation

MLE for $\boldsymbol{\beta}$ and σ^2

We use the observed data $y_1, \ldots, y_n$ to *estimate* the regression coefficients $\beta_1, \ldots, \beta_p$.

The likelihood for a linear model is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right). \qquad (1)$$

This is maximised with respect to $(\boldsymbol{\beta}, \sigma^2)$ at

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)^2.$$

Residuals

The corresponding fitted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

or

$$\hat{y}_i = \boldsymbol{x}_i^T\hat{\beta}, \quad i = 1, \ldots, n.$$

The residuals $\boldsymbol{r} = (r_1, \ldots, r_n)$ are $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ or $r_i = y_i - \boldsymbol{x}_i^T\hat{\beta}$ for $i = 1, \ldots, n$.. These residuals describe the variability in the observed responses $y_1, \ldots, y_n$ which has not been explained by the linear model. We call

$$D = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\beta}\right)^2$$

the *residual sum of squares* or *deviance* for the linear model.

3.1.4 Properties of the MLE

Properties of the MLE

As $\boldsymbol{Y}$ is normally distributed, and $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is a linear function of $\boldsymbol{Y}$, then $\hat{\beta}$ must also be normally distributed. We have

$$E(\hat{\beta}) = \beta \quad \text{and} \quad \text{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1},$$

so

$$\hat{\beta} \sim N(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

It is possible to prove that

$$\frac{D}{\sigma^2} \sim \chi^2_{n-p}$$

which implies that

$$E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2,$$

so the maximum likelihood estimator is biased for σ^2. We often use the unbiased estimator of σ^2

$$\tilde{\sigma}^2 = \frac{D}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2.$$

3.1.5 Comparing linear models

Hypothesis testing

As described previously, we proceed by comparing models pairwise using a likelihood ratio test.

We will assume that model H_1 contains p linear parameters and model H_0 a subset of $q < p$ of these. Without loss of generality, we can think of H_1 as the model

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n$$

and H_0 being the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Likelihood ratio test

Now, a likelihood ratio test of H_0 against H_1 has a critical region of the form

$$C = \left\{ \boldsymbol{y} : \frac{\max_{(\beta,\sigma^2)\in\Theta^{(1)}} L(\boldsymbol{\beta},\sigma^2)}{\max_{(\beta,\sigma^2)\in\Theta^{(0)}} L(\boldsymbol{\beta},\sigma^2)} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\boldsymbol{y} \in C; \boldsymbol{\beta}, \sigma^2) = \alpha.$$

For a linear model,

$$L(\boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right).$$

This is maximised with respect to $(\boldsymbol{\beta}, \sigma^2)$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\sigma^2 = \hat{\sigma}^2 = D/n$.

Critical region

Therefore we have

$$\max_{\beta,\sigma^2} L(\beta,\sigma^2) = (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right).$$

Let the deviances under models H_0 and H_1 be denoted by D_0 and D_1 respectively. Then the critical region is of the form

$$\frac{(2\pi D_1/n)^{-\frac{n}{2}}}{(2\pi D_0/n)^{-\frac{n}{2}}} > k.$$

Rearranging,

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k',$$

for some k'.

We refer to the left hand side of this inequality as the F-statistic. We reject the simpler model H_0 in favour of the more complex model H_1 if F is 'too large'.

Distribution of F test under H_0

As we have required H_0 to be nested in H_1, $F \sim F_{p-q,\,n-p}$ when H_0 is true.

Therefore, the precise critical region can be evaluated given the size, α, of the test. We reject H_0 in favour of H_1 when

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k'$$

where k is the $100(1-\alpha)\%$ quantile of the $F_{p-q,\,n-p}$ distribution.

Conclusion

- We have revised the linear models in different forms
- We have revisited the MLE and its properties for the unknown parameters in linear model
- We have looked at the theory of likelihood ratio test for comparing linear models.