Lecture 6: Linear models and Gauss-Markov theorem

Linear model setting

Results in simple linear regression can be extended to the following general linear model with independently observed response variables $Y_1, ..., Y_n$:

$$Y_i = \beta' x_i + \varepsilon_i, \qquad i = 1, ..., n,$$

where x_i is the ith value of a p-dimensional vector of explanatory variables (multiple covariates) and $\beta \in \mathscr{R}^p$ is an unknown parameter vector, $p < n$ is a positive integer, ε_i's are random errors with mean 0 (more assumptions on the distributions of ε_i's will be added later).

x_i's are deterministic or a given set of observed values of covariates, in which case conditional analysis given $x_1, ..., x_n$ is considered.

A matrix form of the general linear model is

$$Y = X\beta + \mathscr{E}$$

where X is the $n \times p$ matrix whose i raw is x_i, Y is the n-dimensional vector whose ith component is Y_i, and $\mathscr{E}$ is the n-dimensional vector whose ith component is ε_i.

Obviously the simple linear regression model is a special case of the general linear model with $p = 2$.

Another very useful model is described as follows.

One way analysis of variance (ANOVA)

Consider $i = 1, ..., k$ populations with unknown means $\mu_1, ..., \mu_k$, respectively, and a common unknown variance σ^2.

Suppose that a random sample is taken from each population so that the total sample consists of

$$Y_{i1}, ..., Y_{in_i} \text{ iid} \sim \text{population } i \quad i = 1, ..., k$$

where n_i is the sample size of the ith sample and samples from different populations are independent.

Thus Y_{ij}'s are independent but may not be iid.

The ANOVA is a special case of linear model with

$$\beta = (\mu_1, ..., \mu_k)', \quad k = p,$$

$$Y = (Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{k1}, ..., Y_{kn_k})'$$

and

$$X = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1_{n_k} \end{pmatrix}$$

where 1_t is the t-dimensional vector of ones.

The least squares estimator (LSE)

The least squares estimator (LSE) has been defined as an estimator that minimizes the sum of squared distances between points (Y_i, x_i) and a straightline in the simple linear regression.

As an extension, the LSE of β under a gneral linear model is obtained by minimizing

$$\|Y - Xb\|^2 = \sum_{i=1}^{n} (Y_i - x_i' b)^2 \qquad \text{over } b \in \mathscr{R}^p$$

Besides the motivation of minimizing the sum of squared distances, the LSE can also be motivated by the fact that the true parameter vector β is a minimizer of $E(\|Y - Xb\|^2)$ over $b \in \mathscr{R}^p$.

Let $\tilde{\beta} \in \mathscr{R}^p$ satisfying

$$X'X\tilde{\beta} = X'Y \qquad \text{(the so-called normal equation)}$$

Then, for any $b \in \mathscr{R}^p$,

$$\begin{aligned}
\|Y - Xb\|^2 &= \|Y - X\tilde{\beta} + X(\tilde{\beta} - b)\|^2 \\
&= \|Y - X\tilde{\beta}\|^2 + \|X(\tilde{\beta} - b)\|^2 + 2(\tilde{\beta} - b)'X'(Y - X\tilde{\beta}) \\
&= \|Y - X\tilde{\beta}\|^2 + \|X(\tilde{\beta} - b)\|^2 \\
&\geq \|Y - X\tilde{\beta}\|^2
\end{aligned}$$

That means any solution of the normal equation is an LSE of β.

If $X'X$ is nonsingular (of rank p), then the LSE is unique and equal to

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

If $X'X$ is singular, then there are infinitely many solutions to the normal equation; in fact, the model parameter β is not identifiable in the sense that there exist $\gamma \neq \beta$ but $X\gamma = X\beta$ so that with the observed data we cannot estimate β (cannot tell the difference between γ and β).

Regression parameter β

The parameter vector β is called regression parameter.

Suppose that the rank of X is $q < p$.

Then there exists an $n \times q$ submatrix X_* of X such that X_* is of rank q and

$$X = X_* C$$

where C is a $q \times p$ matrix.

Then our linear model becomes

$$Y = X\beta + \mathscr{E} = X_* C\beta + \mathscr{E}$$

Note that $X_*'X_*$ is of rank q and nonsingular.

This means that after re-formulate the model with re-parameterization $\beta_* = C\beta$, we can estimate β_* by the LSE $\widehat{\beta}_* = (X_*'X_*)^{-1}X_*'Y$.

Since the dimension of β_* is $q < p$, the singularity of X is caused by having too many regression parameters in the linear model.

If the rank of X is q, then at most we can estimate q free regression parameters in the linear model.

In simple linear regression, X has two columns with one column 1_n and the other column whose ith element is the unvariate covariate x_i, and $X'X$ is nonsingular iff x_i's are not all the same.

In the case of one-way ANOVA, the matrix X is of the full rank $k = p$ and the LSE of μ_i is $\bar{Y}_i$.

Assumptions on $\mathscr{E}$.

One of the following assumptions is typically assumed:

A1 $\mathscr{E} \sim N(0, \sigma^2 I_n)$ with an unknown $\sigma > 0$.

A2 $E(\mathscr{E}) = 0$ and $\mathrm{Var}(\mathscr{E}) = \sigma^2 I_n$ with an unknown $\sigma > 0$.

A3 $E(\mathscr{E}) = 0$ and $\mathrm{Var}(\mathscr{E})$ exists.

Obviously A1 implies A2 and A2 implies A3.

Typically, some condition has to be added to $V = \mathrm{Var}(\mathscr{E})$ in A3, otherwise we may also have too many parameters.

Either assumption A1 or A2 is assumed in one-way ANOVA or simple linear regression.

Next, we consider a two-way ANOVA model.

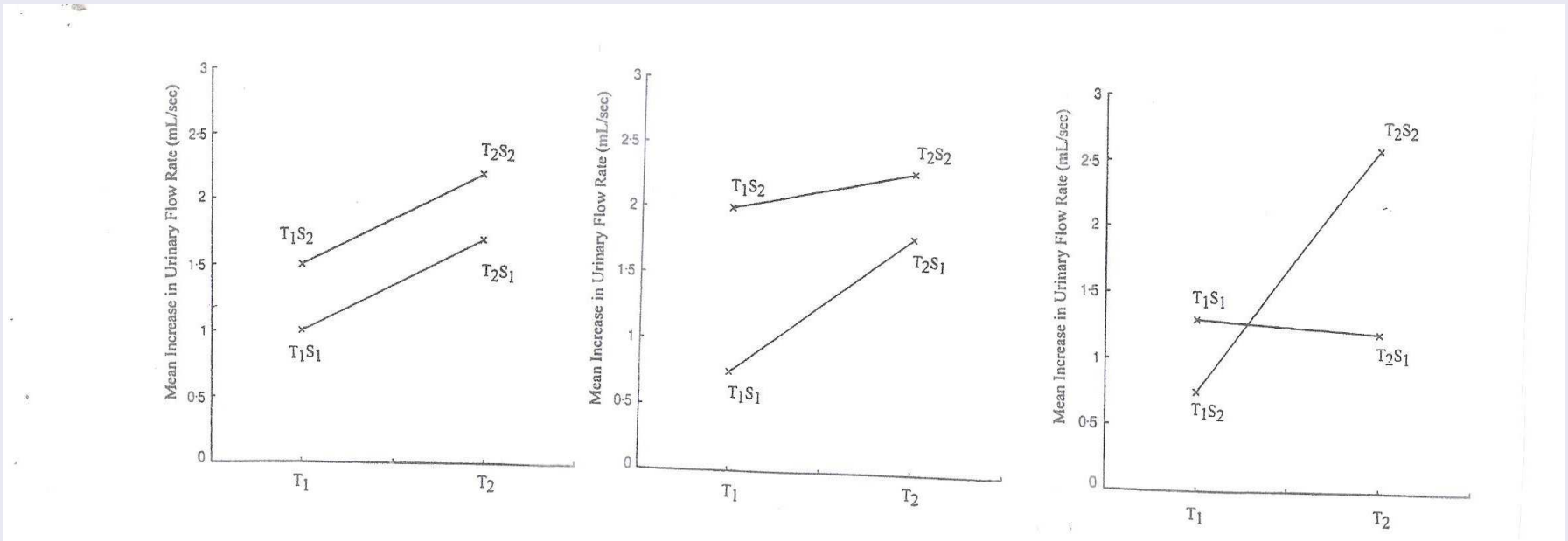
Example (two-way balanced ANOVA)

The one-way ANOVA model studies the effect of p groups constructed using one group factor (variable).

In some applications we have p groups constructed by two factors with one having a groups and the other having b groups so that $p = ab$.

In addition to studying the effects of the two factors, we also want to study whether the two factors have the so called interaction effect.

The following figure illustrates the interaction effect with one factor having two groups T_1 and T_2 and another fact having two groups S_1 and S_2, $a = b = 2$, $p = ab = 4$.



The effects of two factors, A and B, and the interaction effect AB can be modeled as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \qquad i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., c,$$

where a is the number of groups for factor A, b is the number of groups for factor B, c is the number of observations in each combination group of A and B, α_i is the effect of the ith group for factor A, β_j is the effect of the jth group for factor B, γ_{ij} is the interaction effect of the ith group for factor A and the jth group for factor B, μ is an overall effect, Y_{ijk}'s are observations, ε_{ijk}'s are random errors, the total number of groups is ab, and the total number of observations is $n = abc$.

This model is called a two-way balanced ANOVA model, because the number of observations in each combination group is a constant c.

We now show that this model is a special case of the general linear model with the regression parameter vector (denoted as θ instead of β)

$$\theta = (\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, \gamma_{11}, ..., \gamma_{1b}, ..., \gamma_{a1}, ..., \gamma_{ab})'$$

We need to find the form of the matrix X in the general linear model.

Define

$$X_1 = \begin{pmatrix} 1_{ab} & A & B & I_{ab} \end{pmatrix}_{ab\times(a+b+ab+1)}$$

where 1_t denotes the t-dimensional vector of ones, I_m denotes the identity matrix of order m,

$$A = \begin{pmatrix} 1_b & 0 & \cdots & 0 \\ 0 & 1_b & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1_b \end{pmatrix}_{ab\times a} \quad \text{and} \quad B = \begin{pmatrix} I_b \\ I_b \\ \cdots \\ I_b \end{pmatrix}_{ab\times b}$$

Then X can be obtained by replicating X_1 c times, i.e.,

$$X = \begin{pmatrix} 1_{ab} & A & B & I_{ab} \\ 1_{ab} & A & B & I_{ab} \\ \vdots & \vdots & \vdots & \vdots \\ 1_{ab} & A & B & I_{ab} \end{pmatrix}_{abc\times(a+b+ab+1)}$$

This is an $n\times p$ matrix, where $n = abc$ and $p = a+b+ab+1$.

Let $Y_k = (Y_{11k}, ..., Y_{1bk}, Y_{21k}, ..., Y_{2bk}, ..., Y_{a1k}, ..., Y_{abk})'$, $Y = (Y_1', ..., Y_k')'$.

Then $Y = X\theta + \mathscr{E}$ with $\mathscr{E}$ defined similarly to Y.

The rank of X is $ab < p$, even if $n > p$.

This means that there are too many regression parameters: we can only estimate ab regression parameters so the number of extra parameters is $p - ab = a + b + ab + 1 - ab = a + b + 1$.

Before we consider how to reduce the number of parameters, we first show that the following p-dimensional vector is an LSE although LSE's are not unique:

$$\widehat{\theta} = (\widehat{\mu}, \widehat{\alpha}_1, ..., \widehat{\alpha}_a, \widehat{\beta}_1, ..., \widehat{\beta}_b, \widehat{\gamma}_{11}, ..., \widehat{\gamma}_{1b}, ..., \widehat{\gamma}_{a1}, ..., \widehat{\gamma}_{ab})'$$

where

$$\widehat{\mu} = \bar{Y}_{...} \quad \widehat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \quad \widehat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...} \quad \widehat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

and a dot indicates averaging over the indicated subscript, e.g.,

$$\bar{Y}_{...} = \frac{1}{abc} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} Y_{ijk} \qquad \bar{Y}_{.j.} = \frac{1}{ac} \sum_{i=1}^{a} \sum_{k=1}^{c} Y_{ijk} \qquad \bar{Y}_{ij.} = \frac{1}{c} \sum_{k=1}^{c} Y_{ijk}$$

To prove this, we just need to verify that the normal equation holds:

$$X'X\widehat{\theta} = X'Y$$

From the construction of X,

$$X'X = cX_1'X_1 \qquad X'Y = X_1 \sum_{k=1}^{c} Y_k$$

Hence, we need to show that

$$X_1'X_1\widehat{\theta} = X_1'\bar{Y}_{\cdot}$$

where $\bar{Y}_{\cdot} = (\bar{Y}_{11\cdot}, ..., \bar{Y}_{1b\cdot}, \bar{Y}_{21\cdot}, ..., \bar{Y}_{2b\cdot}, ..., \bar{Y}_{a1\cdot}, ..., \bar{Y}_{ab\cdot})'$.

By the construction of X and X_1, the result follows from

$$X_1\widehat{\theta} = \begin{pmatrix} \widehat{\mu} + \widehat{\alpha}_1 + \widehat{\beta}_1 + \widehat{\gamma}_{11} \\ \vdots \\ \widehat{\mu} + \widehat{\alpha}_1 + \widehat{\beta}_b + \widehat{\gamma}_{1b} \\ \vdots \\ \widehat{\mu} + \widehat{\alpha}_a + \widehat{\beta}_1 + \widehat{\gamma}_{a1} \\ \vdots \\ \widehat{\mu} + \widehat{\alpha}_a + \widehat{\beta}_1 + \widehat{\gamma}_{ab} \end{pmatrix} = \begin{pmatrix} \bar{Y}_{11\cdot} \\ \vdots \\ \bar{Y}_{1b\cdot} \\ \vdots \\ \bar{Y}_{a1\cdot} \\ \vdots \\ \bar{Y}_{ab\cdot} \end{pmatrix} = \bar{Y}_{\cdot}$$

where the second equality follows from the definition of $\widehat{\mu}$, $\widehat{\alpha}_i$, $\widehat{\beta}_j$, $\widehat{\gamma}_{ij}$.

We now consider re-parameterization to reduce the number of regression parameters.

If factor A has a different effects and μ is the overall mean, we don't need a parameters $\alpha_1, ..., \alpha_a$ to describe the effects of A.

One of α_i can be absorbed into μ so we can impose a constraint such as

$$\sum_{i=1}^{a} \alpha_i = 0$$

Similarly, we can impose a constraint on the effects of factor B:

$$\sum_{j=1}^{b} \beta_j = 0$$

and constraints on the effects of interactions:

$$\sum_{j=1}^{b} \gamma_{ij} = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0, \qquad i = 1, ..., a, \quad i = 1, ..., b$$

But one of the above constraint on γ_{ij} is redundant so that the total number of constraints is $2 + a + b - 1 = a + b + 1$.

With $a+b+1$ constraints, the total number of free regression parameters is ab and we can estimate all of them by the LSE.

Under the re-parameterization according to the constraints, we may try to find the full rank $n \times ab$ matrix X_* and then estimate the new parameters by the unique LSE.

But it is more convenient to equivalently solve the following minimization problem with constraints:

$$\min_{b \in \mathscr{R}^p} \| Y - Xb \|^2 \qquad \text{subject to the } a+b+1 \text{ constraints}$$

where $p = 1 + a + b + ab$.

It is easy to check that the components of the previously defined $\widehat{\theta}$ satisfies all the $a+b+1$ constraints.

Hence, $\widehat{\theta}$ is the unique LSE under the re-parameterization.

Another advantage of using constraints instead of the new parameters in re-parameterization is the nice interpretation of main effects and interactions under the model with $1+a+b+ab$ parameters.

The following result, without the normality assumption on $\mathscr{E}$, explains why the LSE is popular.

Gauss-Markov Theorem

Assume a general linear model previously described: $Y = X\beta + \mathscr{E}$ with assumption A2, i.e., $\mathrm{Var}(\mathscr{E}) = \sigma^2 I_n$ and X is of full rank $p < n$. Let $\widehat{\beta}$ be the LSE and $l \in \mathscr{R}^p$ be a fixed vector. Then the $l'\widehat{\beta}$ is the *best linear unbiased estimator* (BLUE) of $l'\beta$ in the sense that it has the minimum variance in the class of unbiased estimators of $l'\beta$ that are linear functions of Y.

Proof.

Since $\widehat{\beta} = (X'X)^{-1}X'Y$, it is a linear function of Y and

$$E(\widehat{\beta}) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$$

Thus, $l'\widehat{\beta}$ is unbiased for $l'\beta$.

Let $c'Y$ be any linear unbiased estimator of $l'\beta$, where $c \in \mathscr{R}^p$ is a fixed vector.

Since $c'Y$ is unbiased, $E(c'Y) = c'E(Y) = c'X\beta = l'\beta$ for all β, which implies that $c'X = l'$, i.e., $l = X'c$.

Then

$$\begin{aligned}
\mathrm{Var}(c'Y) &= \mathrm{Var}(c'Y - l'\widehat{\beta} + l'\widehat{\beta}) \\
&= \mathrm{Var}(c'Y - l'\widehat{\beta}) + \mathrm{Var}(l'\widehat{\beta}) \\
&\quad + 2\mathrm{Cov}(c'Y - l'\widehat{\beta}, l'\widehat{\beta}) \\
&= \mathrm{Var}(c'Y - l'\widehat{\beta}) + \mathrm{Var}(l'\widehat{\beta}) \\
&\geq \mathrm{Var}(l'\widehat{\beta})
\end{aligned}$$

where the third equality follows from

$$\begin{aligned}
\mathrm{Cov}(c'Y - l'\widehat{\beta}, l'\widehat{\beta}) &= \mathrm{Cov}(c'Y - l'(X'X)^{-1}X'Y, l'(X'X)^{-1}X'Y) \\
&= \mathrm{Cov}(c'Y, l'(X'X)^{-1}X'Y) - \mathrm{Var}(l'(X'X)^{-1}X'Y) \\
&= c'\mathrm{Var}(Y)X(X'X)^{-1}l - l'(X'X)^{-1}X'\mathrm{Var}(Y)X(X'X)^{-1}l \\
&= \sigma^2 c'X(X'X)^{-1}l - \sigma^2 l'(X'X)^{-1}X'X(X'X)^{-1}l \\
&= \sigma^2 l'(X'X)^{-1}l - \sigma^2 l'(X'X)^{-1}l \\
&= 0.
\end{aligned}$$