256: Linear Statistical Models

Textbooks:

- Monahan J.F. (2008) A primer on linear regression models. Chapman & Hall/CRC Texts in Statistical Science. (M).
- Christensen, R. (2001) Plane Answers to Complex Questions: Theory of Linear Models. (C)
- Faraway, J.J. (2000). Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science. (F)

Course Topics:

- Basic notions of linear algebra, e.g., vector spaces, column and null spaces of a matrix, inverse and generalized inverse, solutions to systems of linear equations, bases, orthogonal matrices, idempotent matrices, eigenvalues and eigenvectors. (M and C)
- Definition and examples of the general linear model, including simple and multiple linear regression, analysis of variance, and analysis of covariance models. (M and C)
- Ordinary and generalized Least Squares Estimation. Estimable functions. Best linear unbiased estimators and the Gauss-Markov Theorem. (M and C)
- Distribution Theory. Class notes but available in many books. Covariances. Properties of covariances. Quadratic forms. Expectations of quadratic forms. Multivariate Normal distribution and its properties. Orthogonal transformations of MVN vectors. Partitions and conditional distributions. Quadratic forms in Multivariate Normal Variables and its distributions. Cochran's theorem. Non-central F distribution. (M and C)
- Maximum likelihood estimation, interval estimation and hypothesis testing under the Gaussian Gauss-Markov model. (M and C)
- You should also be familiar with fitting linear models using `R`. Class notes for examples (F)

General linear regression

Notations :

1. Vector : boldface lowercase

$\underline{a} = [a_1, \ldots, a_n]^T \in R^n$,

$\underline{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \Big\}$ repeated n times

$\underline{0}_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \Big\}$ repeated n times.

2. Matrices : boldface uppercase

$\underline{A}_{n \times m} \longrightarrow$ matrix of dimension n by m.

$\underline{A}_{\cdot j}$: jth column of the matrix $\underline{A}$.

$\underline{I}_n = \begin{bmatrix} 1 & & \underline{0} \\ & \ddots & \\ \underline{0} & & 1 \end{bmatrix}$ $n \times n$ identity matrix.

$\underline{J}_{n \times m} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ $n \times m$ matrix with each entry equal to 1.

3. transpose of a matrix $\underline{A}$ by $\underline{A}^T$ (I can also use $\underline{A}'$)

similarly transpose of a vector $\underline{a}$ by $\underline{a}^T$.

The General Linear Model

Focus on general linear model of the form

$\underline{y} = \underline{X}\,\underline{\beta} + \underline{e}$,

where $\underline{y}$ is an $n \times 1$ vector of responses

$\underline{X}$ is an $n \times p$ matrix of predictors (covariates)

$\underline{\beta}$ is a $p \times 1$ vector of regression coefficients

$\underline{e}$ is an $n \times 1$ vector of unobserved errors.

Linear model is linear in terms of parameter $\underline{\beta}$.

Some of the topics of this class will include.

Estimation:

1. Least square estimation of the model parameter $\underline{\beta}$.
2. BLUE (Best linear unbiased estimator) of the model parameter.
   BLUE is based upon a distribution assumption on the error $\underline{e}$.
   Typically $\underline{e} \sim N(\underline{0}, \sigma^2 I_n)$
   But one can extend the theory to the case where $\underline{e} \sim N(\underline{0}, \underline{\Sigma})$.
3. We will look at estimability of functions of parameters, i.e. whether $\underline{c}'\underline{\beta}$ can be unbiasedly estimated.
4. Hypothesis testing.

To discuss them

① linear algebra — vector spaces, orthogonality, and projections.

② Distribution theory.

<u>usefulness</u>

① Univariate and multivariate regression analysis.

② Analysis of variance (ANOVA)

③ Analysis of covariance (ANCOVA)

④ Random effect modeling.

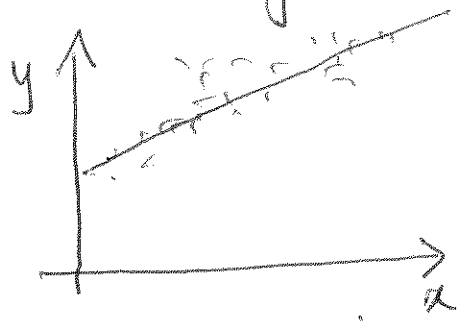
<u>Two main application</u>

① Regression analysis : A response y and a number of predictors x_j, $j=1,\dots,k$ (x_j can be both cont. and categorical)

② Analysis of variance (ANOVA).

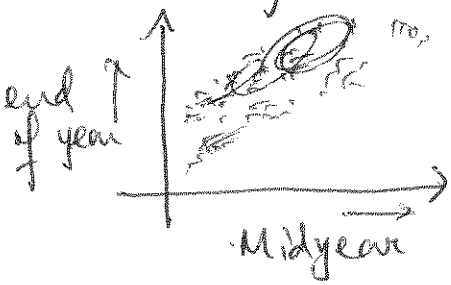
<u>Regression Models</u>

If we have a response y and only one predictor x then that is called a simple linear regression.



When the number of predictors is more than one, that is called multiple linear regression

Example: Performance evaluation for 10 employees were obtained both middle of the year and end of the year.



$$y = \beta_0 + \beta_1 x + e$$

~~β_0 and β_1 are~~

β_0 = intercept and β_1 = slope.

$y_1, \ldots, y_{10}$, $x_1, \ldots, x_{10}$

y = end year evaluation , x = midyear evaluation

$$y_1 = \beta_0 + \beta_1 x_1 + e_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + e_2$$

$$y_{10} = \beta_0 + \beta_1 x_{10} + e_{10}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{10} \end{pmatrix} = \beta_0 \underline{1}_{10} + \beta_1 \begin{pmatrix} x_1 \\ \vdots \\ x_{10} \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_{10} \end{pmatrix} \quad \ldots \quad (1)$$

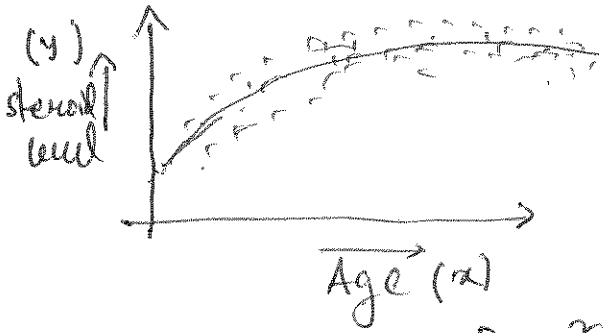
$$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{10} \end{pmatrix} , \quad \underline{X} = \left[\underline{1}_{10} \vdots \underline{x} \right] , \quad \underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{10} \end{pmatrix}$$

and $\underline{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_{10} \end{pmatrix}$

I can write (1) as

$$\underline{Y} = X \underline{\beta} + \underline{e} , \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

eg: the age and level of steroid in plasma for 27 healthy females between 8 and 15 years of age has the following type of relationship.



$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$ (Linear model)

$y = e^{\beta_0 + \beta_1 x} + e$ (Not a linear model)

Multiple linear regression model

$y \rightarrow$ response

$x_1, \dots, x_p \longrightarrow$ predictors

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e$

when we have observations $y_1, \dots, y_n$

$x_{11}, \dots, x_{1n}$

$\vdots$

$x_{p1}, \dots, x_{pn}$

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i$

$$\left.\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{p1} + e_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_p x_{pn} + e_n \end{aligned}\right\} \quad \cdots (2)$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \underline{1}_n + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \cdots + \beta_p \underline{x}_p + \underline{e}$$

$$X = \left[\underline{1}_n : \underline{x}_1 : \underline{x}_2 : \cdots : \underline{x}_p \right]$$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \Rightarrow (2) \text{ can be written as}$$

$$\underset{n\times 1}{\underline{y}} = \underset{n\times(p+1)}{X} \underset{(p+1)\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{e}}; \qquad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\underline{x}_p = \begin{pmatrix} x_{p1} \\ \vdots \\ x_{pn} \end{pmatrix}$$

Regression with AR errors

$$y_t = \beta_0 + \beta_1 t + e_t$$
$$e_t = \nu + \rho e_{t-1} + a_t$$

you can show that

$$y_t = \beta_0(1-\rho) + \nu + \rho\beta_1 + \beta_1(1-\rho)t + \rho y_{t-1} + a_t$$

this is strictly speaking not a linear regression model.

Design of experiment

Definitions:

Experiments: An experiment deliberately imposes a treatment on a group of objects on subjects in the interest of observing the response.

Experimental units: A unit is a person, animal, plant or thing which is actually studied by the researcher; the basic objects upon which the study of experiment is carried out. For example, in the blood pressure example an experimental unit is a person. In the agricultural example it can be a plot.

Factor: A factor of an experiment is a controlled independent variable, whose levels are set by the experimenter.

Treatment: Treatment is something that the researcher administer to experimental units.

Design of experiment is based upon three principles ① replication ② randomization ③ blocking.

① replication: repetition of the basic experiment

② randomization: Both the allocation of the experimental material and the order in which the individual runs or trials of the experiment are to be performed are randomly determined.

blocking: Grouping experimental units into the units that will act similarly into blocks (homogeneous clusters) and then randomly applying the treatments to the units in each block.

Complete vs. incomplete design:

Complete means that each block contains all treatments. Incomplete means every treatment is not present in every block.

Balanced vs. unbalanced design:

A balanced design has an equal number of observations for each treatment. An unbalanced design has an unequal number of observations.

1. completely randomized design (One way Anova model)
2. Randomized block design (Two way Anova model)
3. Latin square design.

~~Completely~~ One way Anova model

One way Anova is used to determine whether there are any significant differences between means of two or more independent (unrelated) groups.

For example, we want to compare K treatments and we assign n_i experimental units to each treatment and measure the response y_{ij}.

<u>Eg:</u> y_{ij} is the jth measurement of nitrogen concentration in the soil that received treatment i.

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad i = 1, \dots, K$$
$$j = 1, \dots, n_i$$
$$= (\mu + c) + (\alpha_i - c) + e_{ij}$$

To avoid this, one either works with

$$y_{ij} = \theta_i + e_{ij}$$

or, they add one extra constraint

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