General linear regression

Notations:

1. vector: boldface lowercase.

$\underline{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ it is an n-dimensional vector.

we customarily write $\underline{a} \in \mathbb{R}^n$.

$\underline{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ } repeated n times. $\underline{0}_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ } repeated n times.

$\underset{n\times 1}{\underline{a}} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$, $\underset{1\times n}{\underline{a}^T} = (a_1, \ldots, a_n)$

2. Matrices: boldface uppercase.

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

$\underline{A}_{\cdot j} \longrightarrow$ j th column of the matrix $\underline{A}$.

$\underline{A}_{j\cdot} \longrightarrow$ j th row of the matrix $\underline{A}$.

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

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

3. Transpose of a matrix $\underline{A}$ is denoted by $\underline{A}^T$.

$\underline{A}'$

The General Linear Model

We will mainly focus on the 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 coefficients.
$\underline{e}$ is an $n \times 1$ vector of unobserved errors.

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

Some of the important topics of this class will include

1. Least square estimation of $\underline{\beta}$.
2. BLUE (Best Linear Unbiased Estimator) of the model parameter $\underline{\beta}$.
3. BLUE is generally obtained based on a distributional assumption on the error vector $\underline{e}$. Typically $\underline{e} \sim N(0, \sigma^2 I_n)$

$$\underline{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \quad e_i \sim N(0, \sigma^2) \text{ and } e_i \text{ and } e_j \text{ are mutually uncorrelated.}$$

(4) We will look at which linear functions of $\underline{\beta}$ can be unbiasedly estimated, i.e. whether $\underline{c}'\underline{\beta}$, where $\underline{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_p \end{pmatrix}$ can be unbiasedly estimated,

$$\underline{c}'\underline{\beta} = \sum_{l=1}^{p} c_l \beta_l$$

(5) Hypothesis testing. We will test hypotheses like

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$.

Let $\underline{A}_{q \times p}$ is a matrix. More generally we will be interested to carry out the hypothesis testing $H_0: \underline{A}\,\underline{\beta} = \underline{0}$ vs. $H_1: \underline{A}\,\underline{\beta} \neq \underline{0}$.

$$\underline{A} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{q1} & \cdots & a_{qp} \end{pmatrix}$$

$$\underline{A}\,\underline{\beta} = \begin{pmatrix} a_{11}\beta_1 + \cdots + a_{1p}\beta_p \\ \vdots \\ a_{q1}\beta_1 + \cdots + a_{qp}\beta_p \end{pmatrix}$$

Thus the above null hypotheses is equivalent to

$$H_0: \begin{cases} a_{11}\beta_1 + \cdots + a_{1p}\beta_p = 0 \\ \vdots \\ a_{q1}\beta_1 + \cdots + a_{qp}\beta_p = 0 \end{cases}$$

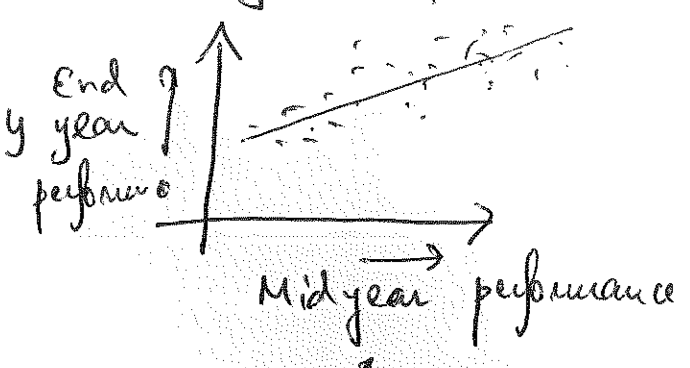
Testing these hypothesis would require distribution theory of quadratic forms.

① ~~univariate~~ simple and multiple regressions.
② Analysis of variance (ANOVA)
③ Analysis of covariance (ANCOVA)
④ Random effect modeling.

Regression models

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

Example: ~~Performance~~ 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 = intercept

and β_1 = slope.

$y_1, \ldots, y_{10} \longrightarrow$ End year performance for 10 indiv.

$x_1, \ldots, x_{10} \longrightarrow$ mid year performance for 10 indiv.

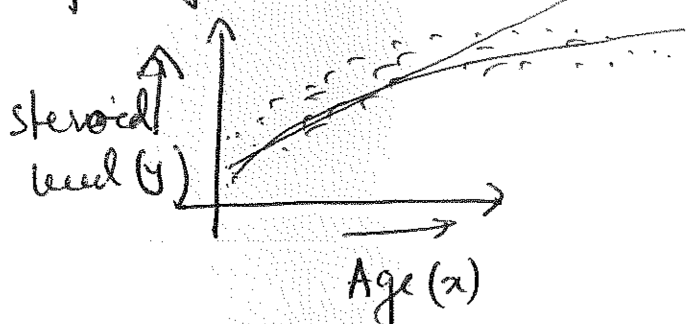
$$y_1 = \beta_0 + \beta_1 x_1 + e_1$$
$$\vdots \qquad\qquad\qquad \vdots$$
$$y_{10} = \beta_0 + \beta_1 x_{10} + e_{10}$$

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

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{10} \end{pmatrix}, \quad \underline{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{10} \end{bmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \underline{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_{10} \end{pmatrix}$$

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

Example: The age and level of steroids in plasma for 27 healthy females between 8 and 15 years of age has the following relationship.



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

(Linear model)

as the relationship is linear in terms of parameters.

$$y = e^{\beta_0 + \beta_1 x} + e$$ [Here the relationship is not linear in terms of the parameter]

Multiple linear regression model

$y \rightarrow$ response $\quad x_1, \ldots, x_p \rightarrow$ predictors.

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

the model that we have seen above

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e \leftarrow$ multiple linear regression model.

Regression with ~~Autog~~ Autoregressive errors

$y_t = \beta_0 + \beta_1 x_t + e_t$

$e_t = \nu + \rho e_{t-1} + a_t$, a_t's are uncorrelated over t.

Design of experiment

Definitions:

Experiments: An experiment deliberately imposes a treatment on a group of 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 experiments is carried out. For example in the blood pressure example, the experimental unit is a female.

Factor: A factor of an experiment are levels of the

Treatment: Treatment is something that the researcher administer to experimental units. In the first example the treatment is the seed. In the second example it is the medication.

Factors: A factor is the levels of the treatment set by the experimenter.

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 treatment and the order in which the treatment is administered to the experimental units are to be determined randomly.

③ Blocking: Sometimes it is important to group experimental units into blocks which will act similarly..

Complete and incomplete Designs:

Complete design means that each block or experimental unit receives all the treatments.

Balanced and unbalanced designs

A balance design administers treatments to equal number of experimental units.

1. Completely randomized design (basis of one way ANOVA model)
2. Randomized Block Design (Two-way ANOVA model)
3. Latin Square Design.