Examples 1 and 2

Let $X \sim \mathit{Binomial}(17, 0.4)$. Then

$$\text{cov}(X) = var(X) = n(\theta(1-\theta)) = 17 \cdot 0.4(0.6)$$

$var(Y_i) = \sigma_i^2 \quad , i = 1, \ldots, n$

If $Y_1, \ldots, Y_n$ are independent then

$$\text{cov}(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}) = \begin{pmatrix} cov(Y_1, Y_1) & \cdots & cov(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ cov(Y_n, Y_1) & \cdots & cov(Y_n, Y_n) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

Example 3

Let X, Y be independent r.v. with $X \sim N(5,2)$ and $Y \sim \text{Binomial}(10, 0.5)$. Then

$$\text{cov}\left(\begin{pmatrix} X \\ -X \end{pmatrix}\right) = \text{cov}\left(\begin{pmatrix} X \\ -X \end{pmatrix}, \begin{pmatrix} X \\ -X \end{pmatrix}\right) = \begin{pmatrix} \text{Var}\,X & \text{Cov}(X,-X) \\ \text{Cov}(-X,X) & \text{Var}(-X) \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

$$\text{cov}\left(\begin{pmatrix} X \\ X+Y \end{pmatrix}\right) = \begin{pmatrix} \text{Var}\,X & \text{Cov}(X, X+Y) \\ \text{Cov}(X+Y, X) & \text{Var}(X+Y) \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2+10(0.5)(0.5) \end{pmatrix}$$

$$\text{cov}\left(X, \begin{pmatrix} 2X \\ X-Y \end{pmatrix}\right) = \begin{pmatrix} \text{Cov}(X,2X) & \text{Cov}(X,X-Y) \end{pmatrix} = \begin{pmatrix} 2\,\text{Var}\,X & \text{Var}\,X \end{pmatrix} = \begin{pmatrix} 4 & 2 \end{pmatrix}$$

Looking ahead

In the next lecture we discuss how to use these concepts to specify and work with general linear models (with multiple predictors)

Imperial College London

Lecture 12: Linear Models with Second Order Assumptions

Statistical Modelling I

Dr. Riccardo Passeggeri

Outline

Linear Model

Definition: The General Linear Model

In a **linear model**

$$Y = X\beta + \epsilon,$$

where

Y is an n-dimensional random vector (observable),

$X \in \mathbb{R}^{n \times p}$ is a known matrix (often called "design matrix"),

$\beta \in \mathbb{R}^p$ is an *unknown parameter* and

ϵ is an n-variate random vector (not observable) with $E(\epsilon) = 0$.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$p = 2$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} (\beta_1 \ \beta_2) + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$Y = X\beta + \varepsilon$$

$$E[Y] = X\beta$$

Example: clinical study

20 patients, 2 drugs, A and B
10 given A, 10 given B
Y_{Aj} = response of jth patient to receive A, $j = 1, \ldots, 10$
Y_{Bj} = response of jth patient to receive B, $j = 1, \ldots, 10$
The simplest model is $E(Y_{Aj}) = \mu_A$, $E(Y_{Bj}) = \mu_B$. In matrix form:

$$E \begin{pmatrix} Y_{A1} \\ \vdots \\ Y_{A,10} \\ Y_{B1} \\ \vdots \\ Y_{B,10} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

$E[Y] = X\beta$

Scientific reasons to add variables to a simple linear model

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad vs \qquad Y = \beta_0 + \beta_1 x + \beta_2 w + \epsilon$$

Suppose we are interested in the relationship between y (e.g. 100m dash time) and a single covariate x (e.g. single-rep max for squat), but we also have information on $w = 0, 1$ (e.g. $0 =$professional sprinter or $1 =$distance runner)

Should we add w to the model?

Linear models $E(Y) = X\beta$ allow for $X \in \mathbb{R}^{n \times p}$.
We will assume a suitable choice of X is known a priori (we will return to this point later)

Assumptions and Identifiability

Assumptions

Second Order Assumption (SOA): $Cov(\epsilon) = (Cov(\epsilon_i, \epsilon_j))_{i,j=1,\dots,n} = \sigma^2 I_n$ for some $\sigma^2 > 0$.
So, (SOA) consists of two parts: First, the errors of two different observations, ϵ_i and ϵ_j for $i \neq j$ are uncorrelated. Second, the variance of all errors is identical (recall: $Var(\epsilon_i) = Cov(\epsilon_i, \epsilon_i)$).

Normal theory assumptions (NTA): $\epsilon \sim N(0, \sigma^2 I_n)$ for some $\sigma^2 > 0$.
N denotes the n-dimensional multivariate normal distribution. Equivalently, NTA ca be written as: $\epsilon_1, \dots, \epsilon_n \sim N(0, \sigma^2)$ independently.
(NTA) implies (SOA). We will use (NTA) to construct tests and confidence intervals.

Full rank (FR) The matrix X has full rank.
We say that a matrix has "full rank" if it has the highest possible rank for its dimensions, i.e. if $rank(X) = \min(n, p)$. As we are mostly working with the situation $n > p$, (FR) reduces to $rank(X) = p$. We will denote the rank of X always by $r = rank(X)$.

Identifiability

In statistical models, one of the main aims is to determine the unknown parameter. If two parameter values lead to the same distribution for the observed data we cannot distinguish between these parameter values.

[redacted]

Suppose we have a statistical model with unknown parameter θ. We call θ *identifiable* if no two different value of θ lead to the same distribution of the observed data.

For a linear model: the main parameter we are interested is β and the observation is Y. It turns out that if $r < p$, then the parameter vector β is not identifiable. The following example shows this.

r = rank(X)

p = numbers of parameters

$$X\tilde{\beta} = E[Y] = X\beta$$

Example: Twin Study

10 pairs of twins, 2 drugs: A and B
one twin in each pair receives A, the other one receives B
twins are alike - we need to modify our previous model:
$E(Y_{Aj}) = \mu_A + \tau_j$, $E(Y_{Bj}) = \mu_B + \tau_j$, where τ_j=effect of twin pair j.

$$E\begin{pmatrix} Y_{A1} \\ \vdots \\ Y_{A,10} \\ Y_{B1} \\ \vdots \\ Y_{B,10} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 1 \end{pmatrix}}_{=X} \underbrace{\begin{pmatrix} \mu_A \\ \mu_B \\ \tau_1 \\ \vdots \\ \tau_{10} \end{pmatrix}}_{\beta}$$

12 x 1

20 x 12

Example: Twin Study

Observe that $r = 11$ and $p = 12$.
Let $\delta > 0$ let

$$\tilde{\beta} = \begin{pmatrix} \mu_A - \delta \\ \mu_B - \delta \\ \tau_1 + \delta \\ \vdots \\ \tau_{10} + \delta \end{pmatrix}. \text{ Then } X\beta = \begin{pmatrix} \mu_a + \tau_1 \\ \vdots \\ \mu_a + \tau_{10} \\ \mu_b + \tau_1 \\ \vdots \\ \mu_b + \tau_{10} \end{pmatrix} = \begin{pmatrix} \mu_a - \delta + \tau_1 + \delta \\ \vdots \\ \mu_a - \delta + \tau_{10} + \delta \\ \mu_b - \delta + \tau_1 + \delta \\ \vdots \\ \mu_b - \delta + \tau_{10} + \delta \end{pmatrix} = X\tilde{\beta}.$$

Thus β and $\tilde{\beta}$ lead to the same expected value of Y. Thus β is not identifiable.

Least Squares Estimation

The Least Squares Problem

$$S(\beta) = \sum_{i=1}^{n} \Bigg(\overbrace{Y_i - \sum_{j=1}^{p} X_{ij}\beta_j}^{\varepsilon_i} \Bigg)^2 = (Y - X\beta)^T(Y - X\beta)$$

x_{ij}

$$\frac{\partial S(\beta)}{\partial \beta_j} = -2 \sum_{i=1}^{n} (Y_i - X_{ij}\beta_j)\, x_{ij} \qquad j = 1, \dots, p$$

$$\left(\frac{\partial S(\beta)}{\partial \beta_j} \right)_{j=1,\dots,p} = -2X^TY + 2X^TX\beta = 0 \implies X^TX\hat{\beta} = X^TY$$

BY THE FR ASSUMPTION $\operatorname{RANK}(X^TX) = \operatorname{RANK}(X) = p \implies (X^TX)^{-1}$ EXISTS

$$\implies \hat{\beta} = (X^TX)^{-1}X^TY$$

$\hat{\beta} = (X^T X)^{-1} X^T Y$ minimises $S(\beta)$

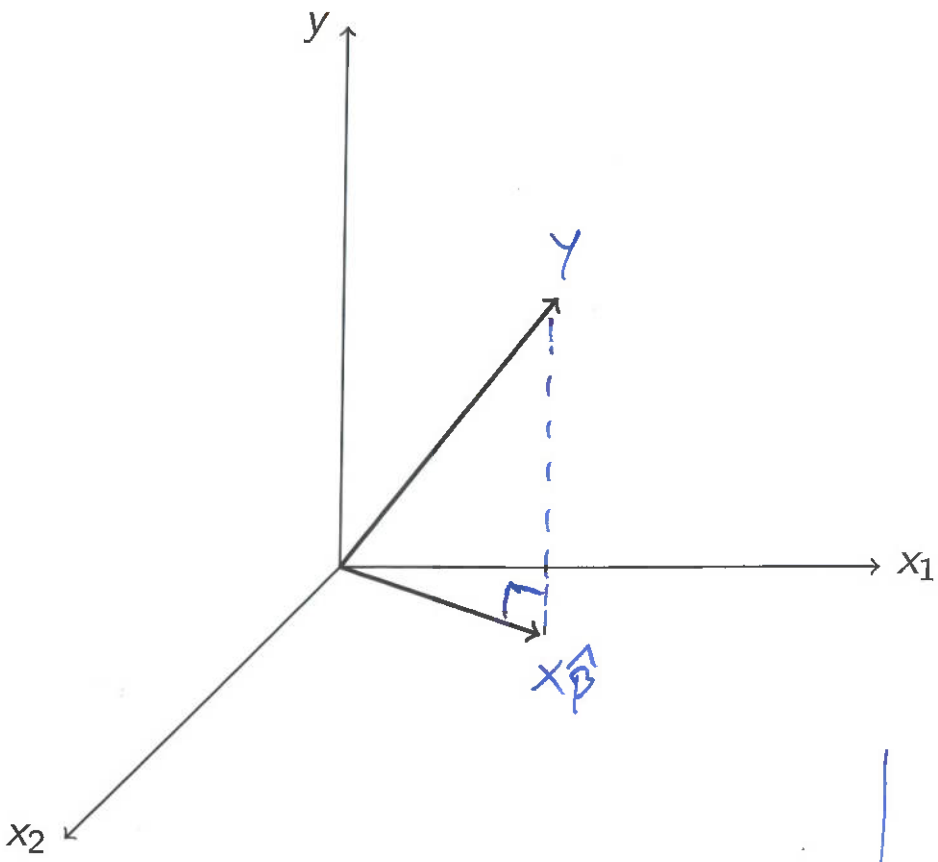
$S(\beta) = (Y - X\beta)^T (Y - X\beta) = (Y - X\hat{\beta} + X\hat{\beta} - X\beta)^T (Y - X\hat{\beta} + X\hat{\beta} - X\beta)$

$= S(\hat{\beta}) + (X\hat{\beta} - X\beta)^T (X\hat{\beta} - X\beta) + 2(X\hat{\beta} - X\beta)^T (Y - X\hat{\beta})$

$= S(\hat{\beta}) + \underbrace{\| X\hat{\beta} - X\beta \|^2}_{\geq 0} + 2(\hat{\beta} - \beta)^T \underbrace{(X^T Y - X^T X\hat{\beta})}_{= 0}$ BY DEFINITION OF $\hat{\beta}$

$\geq S(\hat{\beta})$

Geometry of Least Squares



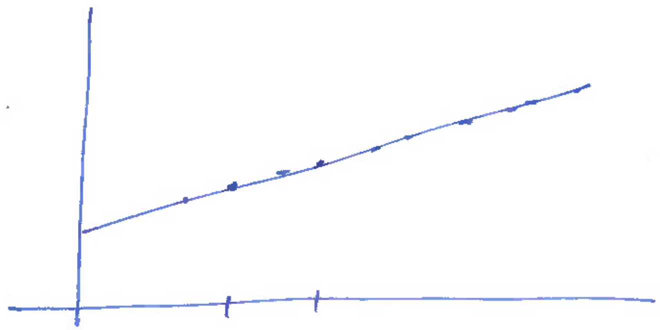
$$Y - X\hat\beta \perp X\hat\beta$$

$$(X\hat\beta)^T(Y - X\hat\beta) = \hat\beta^T X^T (Y - X\hat\beta)$$

$$= \hat\beta^T(\underbrace{X^TY - X^TX\hat\beta}_{=0 \text{ BY DEF. OF } \hat\beta}) = 0$$

$X\hat\beta$ IS THE PROJECTION OF Y ONTO X

$$\text{SPAN}(X) = \{ \cancel{X\beta}\ X\beta : \beta \in \mathbb{R}^p \}$$



Remarks

- Under Full Rank assumptions on X, the least squares estimator is

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

- We will find in the next lecture that $\hat{\beta}$ is optimal in a certain sense