

MATH 60604A
Statistical modelling
§ 5a - Correlated data

HEC Montréal
Department of Decision Sciences

Modifications of the ordinary linear regression model

- The goal of this chapter is to show how the linear regression model can be modified to account for the dependence between observations.
- We focus on modelling the covariance matrix to account for dependence between observations (for longitudinal data and clustered data) and heteroscedasticity (different variance per group).

When independence fails

- When observations are positively correlated, the estimated standard errors of the coefficients of the linear model are **too small**.
- We are overconfident and will reject the null hypothesis more often than we should if the null is true (inflated Type I error, false positives).

Generally, correlation between observations can come from

- time dependence, roughly categorized into
 - longitudinal data: repeated measurements are taken from the same subjects (few time points)
 - time series: observations observed at multiple time periods (many time points). Time series require dedicated models not covered in this course.
- clustered data: measurements are taken from subjects that are not independent from one another (family, groups, etc.)

Moments of random vectors

- Consider a **random vector** Y of dimension n .
 - Such a vector would usually comprise repeated measures on an individual, or even observations from a group of individuals.
- The expected value (theoretical mean) of this vector $E(Y)$ is taken componentwise, i.e., $E(Y) = (E(Y_1), \dots, E(Y_n))$.
- We denote the variance of the i th component by $\sigma_{ii} = \sigma_i^2 = \text{Var}(Y_i)$.
- Similarly, the covariance between observations Y_i and Y_j is $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$.

Covariance matrix

- For a random vector Y , we define the **covariance matrix** as the $n \times n$ symmetric matrix

$$\text{Cov}(Y) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \ddots & \sigma_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_n^2 \end{pmatrix}.$$

- The i th diagonal element of $\text{Cov}(Y)$ is the variance of Y_i .
- Since the matrix is symmetric, $\sigma_{ij} = \sigma_{ji}$.

Covariance and correlation matrix

- The correlation between Y_i and Y_j is

$$\rho_{ij} = \text{Corr}(Y_i, Y_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

- The **correlation matrix** of Y is an $n \times n$ symmetric matrix with ones on the diagonal and the pairwise correlations off the diagonal,

$$\text{Corr}(Y) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \ddots & \rho_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{pmatrix}.$$

One of the most important parts of modelling correlated (or longitudinal) data is the need to account for within-group correlations.

- This basically comes down to modelling a covariance matrix for observations within the same group (or within the same individual in the case of repeated measures).

Longitudinal studies on independent subjects

- In this kind of study, several measurements are taken from the same individuals, usually over time.
 - these data are termed **repeated measures** or **longitudinal data**, but econometricians use the vocable **panel data**.
- The individuals are **independent** from one another; however, measurements from the same subject are not independent.
- A data file might look like

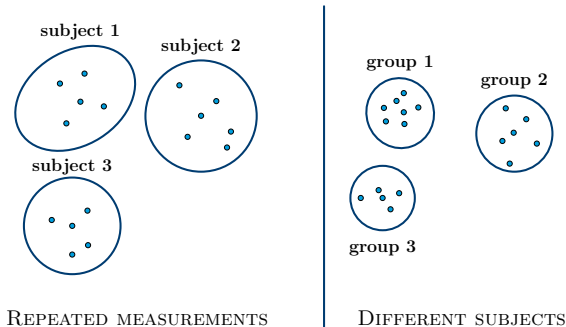
subject	time	score	sex
1	1	5	0
1	2	6	0
1	3	4	0
2	1	2	1
2	2	4	1
2	3	7	1

Studies on subjects that are not independent

- In this kind of study, the subjects are sampled within a **group**.
- Here are several examples:
 - subjects sampled from the same household,
 - subjects sampled from within several businesses,
 - subjects sampled within schools, hospitals, etc.
- In all these examples, the measurements between subjects in the same group (household, school, business) are correlated.

Correlated data is grouped data

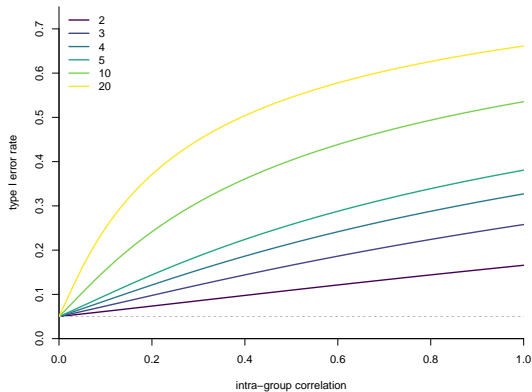
- We can always consider correlated data as grouped data, where there is within-group correlation.
- In longitudinal data, we have several records for each individual.
- In other examples, the groups could be households, schools, hospitals, businesses, etc.



One dot equals one line in the data file.

What happens if we ignore within-group correlation?

- Suppose that we have grouped data and we perform a one-sample t -test with level $\alpha = 5\%$.
- The following figure shows the true Type I error probability as a function of the within-group correlation for different values of the group size m .



Type I error inflation with correlated data

- It is alarming to see how quickly the probability of a Type I error increases with correlation, as well as with the number of samples within each group.
- The conclusions drawn from the t -tests are invalid, since they do not account for the within-group correlation.
- The size distortion illustrates the fact that **statistical inference is typically no longer valid** when we use a method that assumes independence between observations, when in truth the data are correlated.