Longitudinal and Panel data | (NMST 422)

Summer Term 2024 | Department of Probability and Mathematical Statistics

Doc. Matúš Maciak
www.karlin.mff.cuni.cz/~maciak

Faculty of Mathematics and Physics
Charles University, Prague

Topics to cover

- Linear regression overview and multivariate/multiple linear regression
- Longitudinal/panel data and their representation
- Linear mixed effect models (marginal vs. hierarchical)
- GLM overview and generalized estimating equations (GEE)
- GLMM for binary and count data
- Missing data concepts
- Bayesian approaches
- Futher extensions & generalizations

Bibliography

- **Diggle, P.J, Heagerty, P. Liang, K.Y., and Zeger, S. (2022)**
  Analysis of Longitudinal Data. Oxford University Press
- **Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2012)**
  Applied Longitudinal Analysis. John Wiley & Sons
- **Hardin, J.W. and Hilbe, J.M. (2007)**
  Generalized Linear Model and Extensions. StataPress.
- **Kulich, M. (2022)**
  NMST432 Advaced Regression Models: Extended Course Notes
  `www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/index.html` (18.02.2022)
- **Pinheiro, J. and Bates, D. (2006)**
  Mixed-effects models in S and S-PLUS. Springer Science & Business

Bibliography

- **Diggle, P.J, Heagerty, P. Liang, K.Y., and Zeger, S. (2022)**
  Analysis of Longitudinal Data. Oxford University Press
- **Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2012)**
  Applied Longitudinal Analysis. John Wiley & Sons
- **Hardin, J.W. and Hilbe, J.M. (2007)**
  Generalized Linear Model and Extensions. StataPress.
- **Kulich, M. (2022)**
  NMST432 Advaced Regression Models: Extended Course Notes
  `www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/index.html` (18.02.2022)
- **Pinheiro, J. and Bates, D. (2006)**
  Mixed-effects models in S and S-PLUS. Springer Science & Business

- **Additional studying material (NMST 422 web site)**
  Details will be given when/if needed

General course information

General conditions

- enrollment into the corresponding SIS group
- pre-requisite: Linear regression course (NMSA 407)

Credit requirements

- in-person lab session attendance
- active participation
- individual project assignment

Final Course Exam

- final exams at the end of the term (course credit required)
- the exam is composed of two parts – written and oral
- written part contains theory and examples from the lectures
- oral part includes a discussion of the written part and the project solution

Lecture organization

- **In-person teaching**
  - PDF slides (available apriori on the course web page)
  - hand written notes (on the board in the class)
  - additional literature to read/study

- **Individual work**
  - some lectures (lab sessions respectively) not taking place in person
  - individual assignment for styding/working given instead
  - all necessary information will be given when needed

↪ The PDF slides primarily serve as an extended (detailed) sylabus for the lecture. Additional material and specific pieces of information (such as calculations, various proofs, or derivations) will be given by hand.

The PDF slides do not comprehend all necessary information required for the exam!

Lecture 1 | 17.02.2025

Linear regression overview (i.i.d. and/vs. correlated data)

What is the linear regression in general?

- **historically** (Francis Galton)
- **mathematically** (functional relationship)
- **geometrically** (orthogonal projection)
- **numerically** (least squares/normal equations)
- **probabilistically** (conditioanl expectation)
- **statistically** (estimation of the expectation)
- **computationaly** (matrix QR decomposition)

Theoretical perspective: probabilistic model $\Longrightarrow$ model interpretation

Empirical perspective data $\Longrightarrow$ model $\Longrightarrow$ inference$\Longrightarrow$ interpretation

Linear regression model(s)

- ordinary linear regression (theoretical/generic/random model)

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

- ordinary linear regression (theoretical/probabilistic/deterministic model)

$$E[Y|X] = \alpha + \beta X$$

- ordinary linear regression (empirical/statistical/data model)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$\hookrightarrow$ recall the common notation, alternative model definitions, formulations for iid errors (random sample respectively), typical assumptions, and the consequent theoretical properties of the estimates $\widehat{\alpha}$ and $\widehat{\beta}$ (respectively $\widehat{\boldsymbol{\beta}}$).

Generalization for correlated data

- In practice: correlated observations (e.g., multiple observations)
  (paired t-test and further generalizations, repeated measures in general)

- Example: $X_1, \dots, X_n$ (random sample?) – estimate of the mean: $\overline{X}_n$
  (what is the mean and the variance of the corresponding estimate?)
  - $Var \overline{X}_n$ if $cor(X_i, X_j) = 0$ (e.g.., independence, random sample)
  - $Var \overline{X}_n$ if $cor(X_i, X_j) = 1$ (i.e., $cov(X_i, X_j) = \sigma^2$)
  - $Var \overline{X}_n$ if $cor(X_i, X_j) = \gamma \in (0, 1)$ (i.e., $cov(X_i, X_j) = \sigma^2$)

Generalization for correlated data

- In practice: correlated observations (e.g., multiple observations) *(paired t-test and further generalizations, repeated measures in general)*

- Example: $X_1, \dots, X_n$ (random sample?) – estimate of the mean: $\overline{X}_n$ *(what is the mean and the variance of the corresponding estimate?)*
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = 0$ (e.g.., independence, random sample)
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = 1$ (i.e., $cov(X_i, X_j) = \sigma^2$)
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = \gamma \in (0, 1)$ (i.e., $cov(X_i, X_j) = \sigma^2$)
  - Now, what if $\gamma < 0$?

Generalization for correlated data

- In practice: correlated observations (e.g., multiple observations)
  (paired t-test and further generalizations, repeated measures in general)

- Example: $X_1, \dots, X_n$ (random sample?) – estimate of the mean: $\overline{X}_n$
  (what is the mean and the variance of the corresponding estimate?)
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = 0$ (e.g.., independence, random sample)
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = 1$ (i.e., $cov(X_i, X_j) = \sigma^2$)
  - $Var\overline{X}_n$ if $cor(X_i, X_j) = \gamma \in (0, 1)$ (i.e., $cov(X_i, X_j) = \sigma^2$)
  - Now, what if $\gamma < 0$?

$\hookrightarrow$ the variance of a random variable $X \in \mathbb{R}$ is supposed to be always positive...
However, for the random vector $\boldsymbol{X} \in \mathbb{R}^p$ the condition becomes more strict...

$\Rightarrow$ the variance-covariance matrix must be positive definite!
What kind of consequences does it imply? *(curse of dimension problem)*

Data structures beyond random samples...

- **random sample (i.i.d. data)**
  - typical for many simple (but very practical) problems
  - simple theory behind, straightforward proofs, easy implementation
  - however, not always realistic ...

- **correlated (i.e., dependent) data**
  - different forms of dependence (time/spatial)
  - group dependent data (clustered/repeated/longitudinal/panel data)
  - however, still i.i.d. in some (well-formulated) sense

- **n.i.n.i.d. data**
  - generally not independent and not identically distributed observations
  - complex and sophisticated data structures (occuring in practical situations)
  - typical for panel data with nonstationry & dependent panels for instance

Data structures beyond random samples...

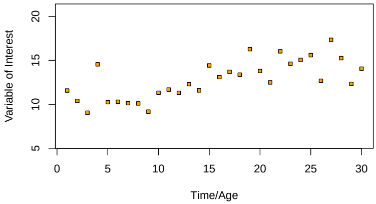
- **random sample (i.i.d. data)**
  - typical for many simple (but very practical) problems
  - simple theory behind, straightforward proofs, easy implementation
  - however, not always realistic ...

- **correlated (i.e., dependent) data**
  - different forms of dependence (time/spatial)
  - group dependent data (clustered/repeated/longitudinal/panel data)
  - however, still i.i.d. in some (well-formulated) sense

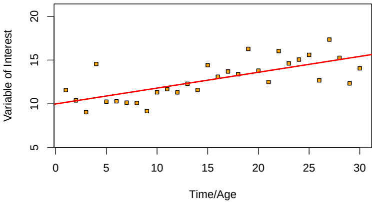
- **n.i.n.i.d. data**
  - generally not independent and not identically distributed observations
  - complex and sophisticated data structures (occuring in practical situations)
  - typical for panel data with nonstationry & dependent panels for instance

Question: In which category would you expect the time series to appear?

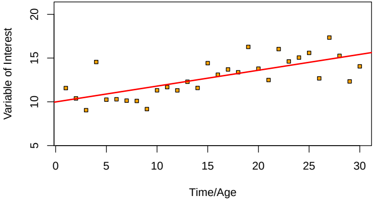
Example: independent/paired/panel data

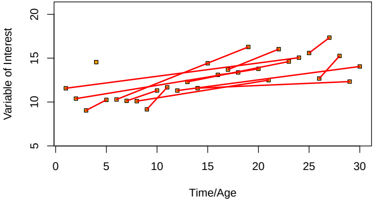


Example: independent/paired/panel data

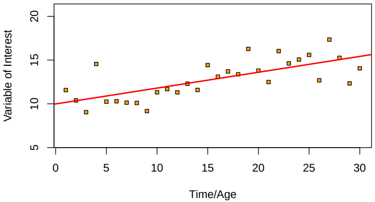


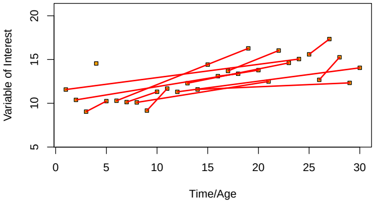
Example: independent/paired/panel data

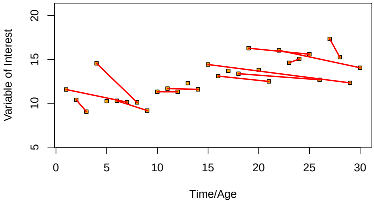




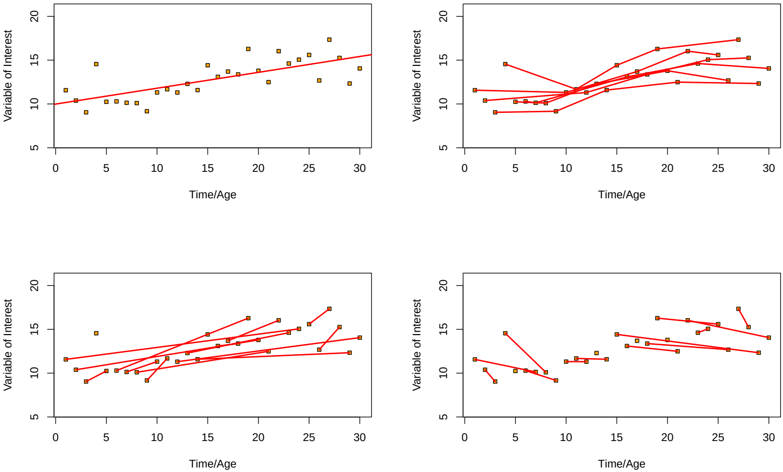
Example: independent/paired/panel data







Example: independent/paired/panel data



Some useful terminology

- **Cross sectional data**
  - typically $n \in \mathbb{N}$ independent subjects measured once but under different conditions
  - different conditions are reflected by the set of explanatory variables/covariates
  - typically a random sample from a joint distribution (of Y and $\boldsymbol{X}$) and $n \to \infty$
- **Time series data**
  - typically a subject (just one) measured/followed repeatedly over time ($T \in \mathbb{T}$)
  - measurements over time are mutually dependent and (theoretically) $T \to \infty$
  - multivariate time series are also assumed, but the dimension $p \in \mathbb{N}$ is fixed
- **Longitudinal data**
  - collection of observations with independent subjects ($n \in \mathbb{N}$) measured over time T
  - independent measurements between subjects, but dependence within each subject
  - typically a limited follow-up period is used (T is fixed) but (theoretically) $n \to \infty$
- **Panel data**
  - in some literature, the panel data and logitudinal data are equivalent/interchangeble
  - in general, more flexibility can be used within the panel data framework
  - typical scenarios involves $n, T \to \infty$, or $n \to \infty$ and T fixed, or vise-versa

Cross-sectional vs. longitudinal model

- Observations $(Y_{ij}, X_{ij1}, \dots, X_{ijp})^\top$, for $i = 1, \dots, N \in \mathbb{N}$ and $n_i, p \in \mathbb{N}$
- Cross-sectional model ($n_i = 1$)

$$Y_{i1} = \beta_{CS} X_{i1} + \varepsilon_{i1} \tag{1}$$

- Longitudinal model ($n_i \in \mathbb{N}$)

$$Y_{ij} = \beta_{CS} X_{i1} + \beta_L (X_{ij} - X_{i1}) + \varepsilon_{ij} \tag{2}$$

$\rightarrow$ for $j = 1$ the later model reduces to the former model, thus β_{CS} has the same interpretation in both models;

$\rightarrow$ in addition, there is also β_L (a longitudinal dependence structure) parameter – its interpretation is quite straightforward when substracting (1) from (2):

$$(Y_{ij} - Y_{i1}) = \beta_L (X_{ij} - X_{i1}) + (\varepsilon_{ij} - \varepsilon_{i1})$$

Cross-sectional vs. longitudinal interpretation

- **Cross-sectional interpretation of β_{CS}**
  (averaging over subpopulations with the same values of X)

  To estimate how individuals change over time with the cross-sectional data it needs to be assumed that the effects coincide $\Rightarrow \beta_{CS} = \beta_L$

- **Longitudinal interpretation of β_L**
  (change within a specific subject per unit change of X within the subject)

  No restriction in the form $\beta_{CS} = \beta_L$ is needed and longitudinal approaches are usually more powerfull even in situations when $\beta_{CS} = \beta_L$

- Population-specific interpretation vs. subject-specific interpretation
  (two different sources of variability that can be properly distinguished)

- Associative vs. causal interpretation of the model
  (however, this is not the causal inference)

Benefits and drawbacks of longitudinal data

There are several imporant advantages of the longitudinal data and longitudinal data models compared with purely cross-sections studies or time series data. Longitudinal data are, therefore, more complex, more powerful, and, also, more useful.
On the other hand, there is also a price to pay when working with them.

Benefits and drawbacks of longitudinal data

There are several imporant advantages of the longitudinal data and longitudinal data models compared with purely cross-sections studies or time series data. Longitudinal data are, therefore, more complex, more powerful, and, also, more useful.
On the other hand, there is also a price to pay when working with them.

Benefits

- ability to study dynamic relationships
- modeling heterogeneity among subjects
- ability to deal with relatively complex data structures

Benefits and drawbacks of longitudinal data

There are several imporant advantages of the longitudinal data and longitudinal data models compared with purely cross-sections studies or time series data. Longitudinal data are, therefore, more complex, more powerful, and, also, more useful.
On the other hand, there is also a price to pay when working with them.

Benefits

- ability to study dynamic relationships
- modeling heterogeneity among subjects
- ability to deal with relatively complex data structures

Drawbacks

- study design (subjects leaving the study before its completion)
- selection bias due to nonrandom (not stratified) effects
- more complex (explanatory and confirmatory) analysis

Borrowing power

- **Inference on β_{CS}** (marginal model)
  - *averaging individuals with one value of X and comparing with averaged individuals with another value of X – the estimated parameter $\widehat{\beta}_{CS}$ stands for the expected/estimated change between subpopulations which corresponds to the unit change of X*

- **Inference on β_L** (hierarchical model)
  - *comparing a specific person's response at two distinct time points while X changes over time within the given subject – the estimated parameter $\widehat{\beta}_L$ stands for the expected/estimated change (time development) within the subject which corresponds to the unit change of X over time (within the given subject)*

Borrowing power across subjects (sometimes possible, sometimes not)

Example: LM interpretation

```
lm(formula = mpg ~ wt + as.factor(cyl), data = mtcars)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       33.9908     1.8878  18.006  < 2e-16 ***
wt                -3.2056     0.7539  -4.252 0.000213 ***
as.factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
as.factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
```

- What is the interpretation of the intercept paramter?
- What is the interpretation of the parameter associated with wt?
- What is the interpretation of the parameter related to as.factor(cyl)6?

Exploration of the longitudinal data

- The first step (most important?) when analyzing (any) data is to perform a proper exploratory analysis... **Why?**
- In case of longitudinal data structures, the exploratory analysis becomes even more important (and also more complex)... **Why?**
  - exploratory of the mean structure
  - exploratory of the variance-covariance structure
  - exploratory of the between-subject dependence structure
  - exploratory of the subject-specific dependence structure

Question: What are common empirical/graphical tools to perform an exploratory analysis on longitudinal (time or spacial dependent) data?
(knowing or not knowing that the data are group-dependent)

Individual work for the next week

- Recall the theory of the maximum likelihood estimation and the proporties of the constructed estimates. Focus, in particular, on the following:
  - Normal linear regression model $Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$, for $i = 1, \dots, n$
  - Maximum likelihood in a multivariate normal model $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$
  - Maximul likelihood estimates for $\boldsymbol{\mu} \in \mathbb{R}^p$ and the variance-covariance Σ

- What would be the corresponding likelihood in a normal regression model of the form

$$\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_{n\times n})$$

  where $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\mathbb{X} \in \mathbb{R}^{n\times p}$ is the model (regression) matrix, and $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$ are the unknown parameters?