R+PAS, lec 2

RStudio and Git

- Today we will practice how to keep R code under version control using Git.
- The Git software can run locally on your computer, but if you need to share work with others you need a central repository. We will use GitHub as our central repository, but other options are available. E.g. GitLab, BitBucket, or one of the AAU servers (no graphical interface).
- We use RStudio as the client to use the underlying Git software and GitHub.
- Whenever you work with a project with R code it is convenient to make an "project" in RStudio (independent of whether you use Git or not).
- How you create the project depends on whether you are going to use Git or not.

Getting started

- Go to your home page on GitHub and create a new repository (select a sensible name that you will also use as the main directory holding the files on your computer).
- Initialize it with a README.md file.
- Make a new project in RStudio where you select “check out from version control” and enter the repository url from GitHub.
- Our reference for working with R code in Git will be http://happygitwithr.com/

Working with Git in RStudio

- Stage file for version control and commit it (happens locally on computer)
- Push changes to repository (sends to GitHub.com)

PART II

Population vs. sample

- **Population:** The entire set of individuals we want to investigate. Typically assumed infinite.
- **Sample:** A randomly chosen subset of the population. Should be representative.

The parametric model $f_Y(y; \theta)$ and its (unknown) parameter θ describes the population, while the corresponding estimate $\hat{\theta}$ describes the sample.

The aim of a statistical analysis is to derive conclusions about the population based on the sample.

Steps in a statistical analysis

Typically we have (some of) the following steps in a statistical analysis (not necessarily in this order):

- **Preliminary investigations:** The first step is to get to know your data. Make descriptive statistics and plots.
- **Model specification:** Specify the model $f_Y(y;\theta)$, e.g. a multiple regression model or an autoregressive model.
- **Parameter estimation:** Estimate the parameters in the model, e.g. using maximum likelihood estimation or method of least squares.
- **Model checking:** All models contain a number of assumptions about the data, e.g. independence or normal distribution. Here we check whether these are (approximately) fulfilled by the data. If not, we need to modify our model.
- **Hypothesis testing:** Test relevant hypotheses of the model, e.g. can the model be simplified by removing some terms, or is it reasonable that the true unknown have specific relevant values.

Steps in a statistical analysis (cont.)

- **Model selection:** If we have different potential models, we need to compare them to see which ones fit best, e.g. the model may contain multiple variables, and we do not know which ones to include.
- **Prediction:** Can we use the model to make predictions of the future (or other missing parts of the data).
- **Conclusions:** At the end we gather our findings to make conclusions about the population in question. Can we conclude something that we can explain to non-statisticians?

We will look at practical issues regarding all of these steps in the case of linear models, including how to perform the steps in R.

Linear models

- A linear model is a model of the form

$$y_i = \beta_0 x_{i,0} + \cdots + \beta_k x_{i,k} + \epsilon_i, \qquad i = 1, \dots, n$$

  where β_j are parameters, y_i and $x_{i,j}$ are (functions of) observed data, $\epsilon_i \sim N(0, \sigma^2)$ are i.i.d. error terms.
- On vector-form

$$y = X\beta + \epsilon$$

  with

  - design matrix: X is $n \times (k+1)$ matrix with (i,j)th element $x_{i,j}$,
  - parameter vector: $\beta = (\beta_1, \dots, \beta_k)^\top$,
  - error vector: $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$,
  - observation vector: $y = (y_1, \dots, y_n)^\top$
- $x_{\cdot,j}$ are called explanatory or independent variables, y is called a response or dependent variable

Linear models - a simple example of regression

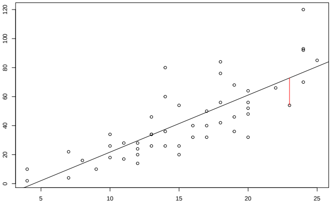
- Simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Vector form $y = X\beta + \epsilon$ with

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

  If we pretend we know the true line $y = \beta_0 + \beta_1 x$ (we don't!), one of the ϵ_i is illustrated in red.



Linear models - other examples of regression

- Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{i,0} + \cdots + \beta_k x_{i,k} + \epsilon_i$$

- Polynomial regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$$

Estimation in linear models

- The parameter vector can be estimated by method of least squares (same for maximum likelihood estimation) by solving the **normal equations**

  $$(X^\top X)\beta = X^\top y$$

  yielding

  $$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

  Note that the matrix inverse should only be used for math formulas etc. The numerical solution should always be done by solving the normal equations.
- The variance estimated by method of least squares

  $$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^\top (y - X\hat{\beta})}{n - (k+1)}$$

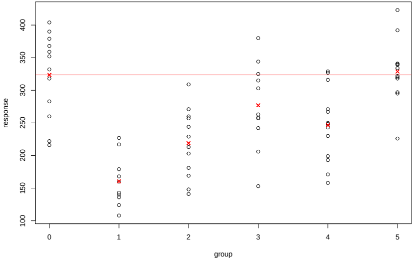
PART 3

Linear models - ANOVA types of models

- So far x_i are real (continuous!) variables - now assume x_i is a discrete variable with n categories, say $x_i \in \{1, \dots, k\}$
- One-way analysis of variance

$$y_i = \beta_0 + \beta_1 \mathbf{1}[x_i = 1] + \cdots + \beta_{k-1} \mathbf{1}[x_i = k-1] + \epsilon_i$$

- The indicator functions $\mathbf{1}[x_i = j]$ are called dummy variables.
- Group k is a reference group with mean β_0 (marked by a red line), while for $j < k$ group j have mean $\beta_0 + \beta_j$ (marked by red crosses).


response
group