Statistical methods and data analysis in developmental psychology

Antonio Calcagnì

DPSS, University of Padova

A.Y 2021-2022

Copyright © 2021 Antonio Calcagnì. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation A copy of the license is available at: https://www.gnu.org/licenses/fdl-1.3.html.

Outline

Introduction

The purpose of this course is to provide statistical theory and methods to analyse data which are organized in the form of pairs

$$(y_1, x_1), \ldots, (y_i, x_i), \ldots, (y_n, x_n)$$

where, in general, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$ can be J-dimensional.

The aim is to infer a **stochastic function** which relates y_i to x_i through a **statistical model**

$$y_i = f(x_i) + \epsilon_i$$

where the noise term ϵ_i is assumed to be additive.

Introduction

Throughout the course, y_i will be called **response variable** whereas, depending on the context, $x_{i1}, \ldots, x_{iJ}$ will be called **independent variables** (e.g., experimental settings) or simply **covariates** (e.g., social studies). In general, $\mathbf{x}_{i1}, \ldots, x_{iJ}$ are used as **predictors** of the responses y_i (asymmetric relation).

In this context, $x_{i1}, \ldots, x_{iJ}$ are considered *non stochastic* whereas y_i are thought as being random realizations from a random variable Y_i (only the response variables embed the stochasticity of the data collection process).

Introduction

Depending on the context, the predictors $x_{i1}, \ldots, x_{iJ}$ can be **continuous** (i.e., $\mathbf{x}_i \in \mathbb{R}^J$) or **categorical** (i.e., $\mathbf{x}_i \in \mathcal{C} \subset \mathbb{N}^J$). In the last case, the elements of $\mathcal{C} = \{c_1, \ldots, c_k\}$ represent the **levels** that $\mathbf{x}_i$ may assume. For instance, if $k = 2$ the predictor is a dichotomous variable, otherwise when $k > 2$ the predictor is a polithomous variable.

The same applies for the response observations y_i. We may have **continuous** (i.e., $y_i \in \mathbb{R}$) or **positive continuous** responses (i.e., $y_i \in \mathbb{R}^+$) as well as categorical responses (i.e., $y_i \in \mathcal{C} \subset \mathbb{N}$) in the simplest form of **unordered categorical** responses or **ordered categorical** responses (i.e., $\ldots < c_{k-1} < c_k < c_{k+1} < \ldots$). In some circumstances, observations can be also collected in the form of **counts** (i.e., $y_i \in \mathbb{N}_0$) or **frequencies** (i.e., $y_i \in [0, 1]$).

Introduction

As the stochasticity is embedded into y_i, the type of observation (e.g., continuous, categorical, counts) implies different random variable model Y_i. For instance, *continuous* responses may be modeled using a Normal random variable $Y_i \sim \mathcal{N}(y; \mu_i, \sigma^2)$, *dichotomous* responses may be modeled using a Bernoulli random variable $Y_i \sim \mathcal{B}er(\pi_i)$, *counts* may be modeled using a Poisson random variable $Y_i \sim \mathcal{P}oi(y; \lambda_i)$.

In order to infer the proper statistical model for a given response variable, we use **generalized linear models** (GLMs) which is a class of statistical models including many probabilistic models (e.g., Normal, Poisson, Gamma) for different response variables (e.g., continuous, counts, response times).

Introduction

Response variable and covariates can be organized by means of a $n \times (J + 1)$ matrix representation where n is the number of collected/sampled statistical units.

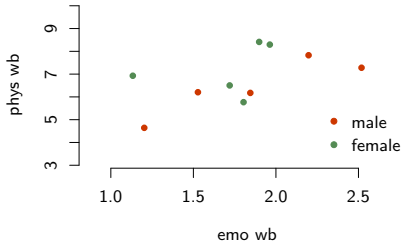
	Y	X_1	$\ldots$	X_J
1	y_1	x_{11}	$\ldots$	x_{1J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	x_{i1}	$\ldots$	x_{iJ}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	$\ldots$	x_{nJ}

Notes:

- We are interested in studying Y_i conditioned on $\mathbf{x}_i$ (asymmetric relation)
- For **non-grouped data** (most of the first part of the course), the number of statistical units n equals the number of observations

Introduction
Example 1

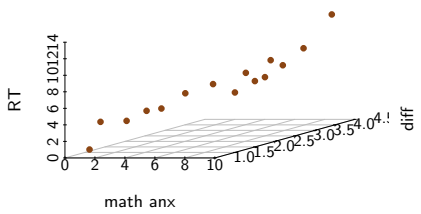
Subset of $n = 10$ data referring to the study of Physical well-being (`phys wb`) as a function of Emotional well-being (`emo wb`) and `gender`.



	phys wb	gender	emo wb
1	4.64	1	1.20
2	6.17	1	1.84
3	7.28	1	2.52
4	6.21	1	1.53
5	7.83	1	2.20
6	5.77	2	1.80
7	6.50	2	1.72
8	6.93	2	1.13
9	8.30	2	1.96
10	8.41	2	1.90

Introduction
Example 2

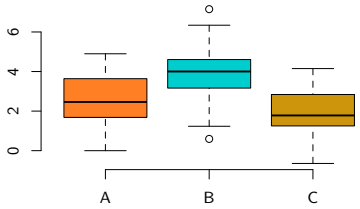
Subset of $n = 15$ data referring to the study of response times (in sec) to a math test (`RT`) as a function of math anxiety (`math anx`) and test difficulty `diff`.



	RT	math anx	diff
1	0.96	1.53	1.04
2	4.08	1.81	1.21
3	3.98	3.11	1.38
4	4.56	3.16	1.86
5	4.70	3.88	1.96
6	5.82	4.03	2.50
7	6.10	4.22	3.12
8	4.44	4.38	3.61
9	6.47	4.42	3.87
10	5.26	4.59	4.03
11	5.53	4.85	4.18
12	7.47	5.00	4.27
13	6.84	5.79	4.28
14	8.85	7.07	4.31
15	12.79	8.65	4.43

Introduction
Example 3

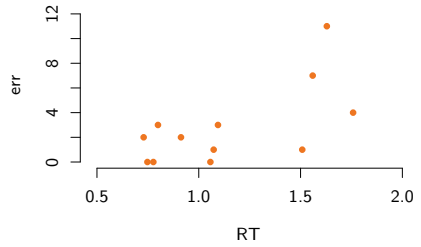
Subset of $n = 150$ data referring to the study of math anxiety to a math test (`math anx`) as a function of three types of test (`test`).



	math anx	test
1	2.38	2
2	5.00	2
3	4.47	2
4	1.95	3
5	5.61	2
6	3.44	2
7	4.01	2
8	5.64	2
9	0.56	1
10	1.35	1
11	2.88	1
12	2.56	1
13	3.27	2
14	2.30	1
15	2.91	2
⋮	⋮	⋮

Introduction
Example 4

Subset of $n = 12$ data referring to the number of errors in an experimental task (`err`) as a function of response times (`RT`).



	err	RT
1	11	1.63
2	0	1.06
3	0	0.75
4	4	1.76
5	3	1.09
6	7	1.56
7	1	1.07
8	2	0.91
9	0	0.78
10	2	0.73
11	3	0.80
12	1	1.51

Introduction
Notation

Throughout the course, **uppercase Roman letters** will denote both the *random variables* underlying the response variable $Y_1, \ldots, Y_i, \ldots, Y_n$ and the *explanatory variables* $X_{i1}, \ldots, X_{iJ}$. Instead, **lowercase Roman letters** will denote either the random realizations associated to the response variable $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)$ or the observed values for the predictors $\mathbf{x}_j = (x_1, \ldots, x_i, \ldots, x_n)$.

Boldface Roman letters will denote *matrices* (e.g., $\mathbf{X}$) as well as column *vectors* (e.g., $\mathbf{y}$), with dimensions in subscript (e.g., $\mathbf{X}_{n\times J}$, $\mathbf{y}_{n\times 1}$). Sometimes dimensions may be omitted to simplify notation.

Model parameters will be denoted by **Greek letters** (e.g, θ). As for the Roman case, boldface Greek symbols will indicate matrices or vectors of model parameters (e.g., $\boldsymbol{\theta}$, $\boldsymbol{\theta}_{q\times 1}$).

Introduction
Notation

The starting point of statistical modeling is the sample of observations $\mathbf{y} = (y_1, \ldots, y_n)$ which is a random realization of a set of random variables (*random vector*) $Y_1, \ldots, Y_n$. Most often, $Y_1, \ldots, Y_n$ are considered independent with identical distribution (iid) so that $\mathbf{y}$ is the outcome of a Bernoulli sampling schema.

The usual notation is then adopted to indicate the probabilistic model for a random variable $Y_i \sim \mathcal{F}(y; \boldsymbol{\theta})$ with $\mathcal{F}$ being a proper statistical distribution parameterized by $\boldsymbol{\theta}$. With a slight abuse of notation, the same will also be denoted by $\mathbf{y} \sim F(y; \boldsymbol{\theta})$.

Introduction
Defining a statistical model

For a correct statistical model specification we will need to evaluate:

1. the probabilistic model $\mathcal{F}$ (aka, probabilistic distribution) underlying the response variable Y
2. the characteristics of $\mathcal{F}$ (e.g., expected value, variance, covariance) to be associated with the explanatory variables $\mathbf{X}$ through a parametric specification

Introduction
Defining a statistical model

For instance, in **Example 1** the response `phys wb` may be modeled using the Normal distribution $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{N}$ with parameters $\boldsymbol{\theta} = \{\mu, \sigma^2\} \in \mathbb{R} \times \mathbb{R}^+$, i.e.:

$$y_i \sim \mathcal{N}(y; \mu_i, \sigma^2)$$

with the explanatory variables `emo wb` ($\mathbf{x}_1$) and `gender` ($\mathbf{x}_2$) being associated to the *mean* of the model (systematic variation):

$$\mu_i = \mathbb{E}[Y_i] = \beta_0 + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$$

The variance of the model σ^2 can be defined either as a function of the data (first part of the course) or as a function of the mean and, consequently, as a function of the explanatory variables (second part of the course).

Similarly, in **Example 4** the response `err` may be modeled using the Poisson distribution $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{P}oi$ with parameters $\boldsymbol{\theta} = \lambda \in \mathbb{R}^+$, i.e.:

$$y_i \sim \mathcal{P}oi(y; \lambda_i)$$

with the explanatory variable `RT` ($\mathbf{x}$) being associated to the *mean* of the model:

$$\lambda_i = \mathbb{E}[Y_i] = \exp(\beta_0 + \mathbf{x}\beta)$$

In this particular case, the variance equals the mean, which is in turn a function of `RT`:

$$\mathbb{V}\text{ar}[Y_i] = \lambda_i$$

This is common in GLMs.

..we will return to these topics (more in depth) at the beginning of module B.

In the **first part of the course** we will focus on the cases where:

- $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{N}$ with $y_i \sim \mathcal{N}(y; \mu_i, \sigma^2)$
- $\mu_i = \mathbf{x}^T_{J\times 1}\boldsymbol{\beta}_{J\times 1}$
- $\sigma^2 \perp\!\!\!\perp \mu_i$ for $i = 1, \ldots, n$ (mean and variance of the model are independent)

In the first place, we will review basics of probability theory and random variables.

For a review of linear algebra, see Appendix B.1 (Fox, 2016) in the supplementary materials of the course.

A **random experiment** is an experiment whose outcomes cannot be determined in advance. Whereas the set of all possible outcomes (**sample space** Ω) can distinctly be determined (there is no fuzziness in this step), what is affected by uncertainty is the occurrence of an **event** of the sample space.

The most typical example is the experiment where a (fair) coin is tossed a number of times (e.g., three times). In this case,

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

where THT means that the first toss is tail, the second is head, and the the third is tail.

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

Subsets $A_1, \ldots, A_K$ of Ω are called **events**. For instance, the event that the first toss is tail is $A = \{THH, THT, TTH, TTT\}$. An event *A is said to occur* if an element $a \in A$ (e.g., $a = \{THH\}$) is the outcome of the experiment.

Since events are sets, we can use the **standard set operations** to perform computation on random events.

Given two events A_k, A_h ($k \neq h$):

- $A_k \cup A_h$ (union: the event that A or B or both occur)
- $A_k \cap A_h$ (intersection: the event that A and B both occur)
- A^c (complement: the event that A does not occur)
- $A_k \cap A_h = \emptyset$ (disjoint event)

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

The **probability** $\mathbb{P}$ of an event A is a measure such that

P1 $\mathbb{P}(A_k) \in [0, 1]$

P2 $\mathbb{P}(\bigcup_{k=1}^{K} A_k) = \sum_{k=1}^{K} \mathbb{P}(A_k) = 1$

P1 states that $\mathbb{P}(A_k) = 0$ indicates that A_k does not occur certainly whereas P2 gives a calculus for the total probability of disjoint events.

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

There are two ways to assign probability at least:

- **classic**: the probability of A is given by computing the elementary events which have been occurred during an experiment

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

- **frequentist**: the probability of A is given as the limiting frequency after a sequence of n independent attempts

$$\mathbb{P}(A) = \lim_{n \to \infty} \frac{f_A}{n}$$

where f_A is the empirical frequency for the event A

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

It should be noted that the *classic* approach to computing probabilities is performed **before** the random experiment is done whereas the *frequentist* (or *empirical*) approach is performed **after** the experiment is done and it requires that the experiment can be repeated infinitely many times.

For example, the probability of the event $A = \{THH, THT, TTH, TTT\}$ is

$$\mathbb{P}(A) = |A| \ / \ |\Omega| = 4/2^3 = 1/2$$

according to the classic approach to probability.

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

Consider two events A_k and A_h ($k \neq h$). Then, by fixing one of the two terms (e.g., A_h) we may ask whether knowing A_h changes the probability of A_k:

$$\mathbb{P}(A_k|A_h) = \frac{\mathbb{P}(A_k \cap A_h)}{\mathbb{P}(A_h)}$$

This is known as **conditional probability**. When $\mathbb{P}(A_k \cap A_h) = \emptyset$ then $\mathbb{P}(A_k|A_h) = \mathbb{P}(A_k)$ and knowing $\mathbb{P}(A_h)$ does not affect the probability of $\mathbb{P}(A_k)$.

For instance, suppose we toss a fair coin three times. Let A_h be the event that the total number of heads is two and let A_k be the event that the first toss is heads. Then $\mathbb{P}(A_k|A_h) = (2/8)/(3/8) = 2/3 = 0.67$.

Random experiments

Source: Appendix D.1.1 (Fox, 2016) - supplementary materials

The conditional probability also provides a calculus for the joint probability of $A_k \cap A_h$:

$$\mathbb{P}(A_k \cap A_h) = \mathbb{P}(A_k|A_h)\mathbb{P}(A_h)$$

which can be generalized for a sequence of events:

$$\mathbb{P}(A_1 \cap \ldots \cap A_K) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1) \cdots \mathbb{P}(A_k|A_{k-1} \cap \ldots \cap A_1) \cdots \mathbb{P}(A_K|A_{K-1} \cap \ldots \cap A_1)$$

Two events A_k and A_h ($k \neq h$) are said to be **independent** when $\mathbb{P}(A_k|A_h) = \mathbb{P}(A_k)$ or $\mathbb{P}(A_k \cap A_h) = \mathbb{P}(A_k)\mathbb{P}(A_h)$. Independence models the *lack of information between events* and it is often a model assumption.

In the general case of independence:

$$\mathbb{P}(A_1 \cap \ldots \cap A_K) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_k) \cdots \mathbb{P}(A_K)$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Very often defining probability spaces for some interesting empirical phenomena is difficult. Sometimes it is also unnecessary as we are only interested in particular outcomes of the random experiment. To this end, random variables can circumvent these issues by introducing parametric classes of probabilistic models.

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Non-formal definition: A **random variable** X is a function that maps subsets of the sample space Ω (or subsets of the event space $\mathcal{A}$, the σ-algebra associated to Ω) to real numbers.

The **support** of X - i.e. $\sup(X)$ - is the set of values that X may assume. For discrete random variables, $\sup(X)$ is countable finite (e.g., discrete subset of real numbers). For real random variables, $\sup(X)$ is infinite.

Random variables can be **univariate** (e.g., $\sup(X) \subset \mathbb{R}$), **bivariate** (e.g., $\sup(X) \subset \mathbb{R} \times \mathbb{R}$), or more generally **multivariate** (e.g., $\sup(X) \subset \mathbb{R} \times \overset{J-1}{\ldots} \times \mathbb{R}$).

Note: the adjective *random* indicates that we are dealing with random experiments (the function X is not random per-sé).

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Example: Fair coin tossed $n = 3$ times

- $\Omega = \{ttt, ttc, tct, ctt, ccc, cct, ctc, tcc\}$, $|\Omega| = 2^n$
- $\mathbb{P}$ defined according to the classic assignment: $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
- $X \stackrel{\text{def}}{=}$ "number of heads", $\sup(X) = \{0, 1, 2, 3\}$

$$
\begin{array}{lcl}
X = 0 & \iff & \{ccc\} \\
X = 1 & \iff & \{ctc, tcc, cct\} \\
X = 2 & \iff & \{ttc, tct, ctt\} \\
X = 3 & \iff & \{ttt\}
\end{array}
$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Example: Fair coin tossed $n = 3$ times

- $\Omega = \{ttt, ttc, tct, ctt, ccc, cct, ctc, tcc\}$, $|\Omega| = 2^n$
- $\mathbb{P}$ defined according to the classic assignment: $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
- $X \stackrel{\text{def}}{=}$ "number of heads", $\sup(X) = \{0, 1, 2, 3\}$

$$
\begin{array}{lcll}
X = 0 & \iff & \{ccc\} & \mathbb{P}(X = 0) = 1/8 \\
X = 1 & \iff & \{ctc, tcc, cct\} & \mathbb{P}(X = 1) = 3/8 \\
X = 2 & \iff & \{ttc, tct, ctt\} & \mathbb{P}(X = 2) = 3/8 \\
X = 3 & \iff & \{ttt\} & \mathbb{P}(X = 3) = 1/8
\end{array}
$$

Note: $\sup(X)$ can be considered the new sample space over which $\mathbb{P}$ assigns probabilities.

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

The probabilities $\mathbb{P}(X = x)$ induced by a random variable give rise to the **distribution function** F_X. Depending on if X is discrete or continuous, probability distribution can be discrete or continuous too.

F_X defines the way probabilities can be computed by means of random variables:

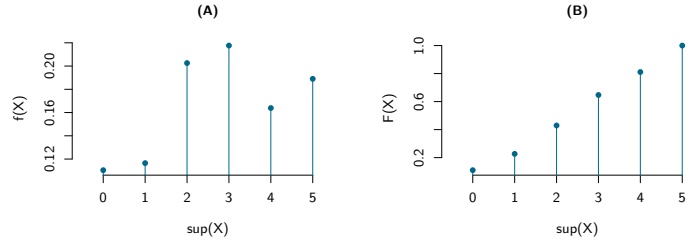
$$
\begin{aligned}
F_X(X = x) &= \mathbb{P}(X \leq x) \quad x \in \sup(X) \\
F_X(X \in [a, b]) &= \mathbb{P}(a \leq X \leq b) \quad x \in \sup(X)
\end{aligned}
$$

From F_X we can derive continuous or discrete **density functions** $f_X(X = x)$ or $f_X(X \in [a, b])$. In general, $f_X(x)$ satisfies the axioms of probabilities:

- $f_X(x \in [x_0, x_0 + \epsilon]) = \int_{x_0}^{x_0+\epsilon} f_X(x)\, dx \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Random variables

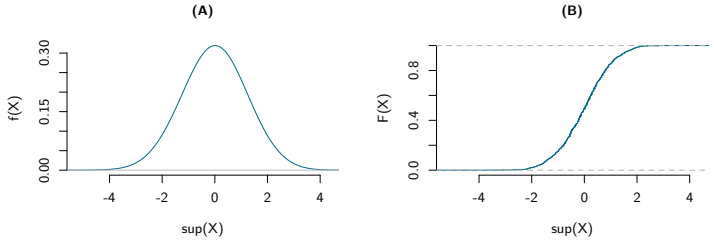
Source: Appendix D.1.2 (Fox, 2016) - supplementary materials



(A) Discrete density function (aka, *probability mass function*)
(B) Discrete distribution function (aka, *cumulative probability mass function*)

$f(X = x) \geq 0$

$\sum_{x \in \sup(X)} f(X = x) = 1$



(A) Continuous density function
(B) Continuous distribution function (aka, *cumulative density function*)

$$f(x_0 \leq X \leq x_0 + \epsilon) = \int_{x_0}^{x_0+\epsilon} f(x)dx \geq 0 \qquad (\epsilon > 0)$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$f(x_0) = 0$$

Random variables allows for using the same probabilistic models to represent different random experiments.

For instance, the Binomial random model can be used to formalize the experiments of drawing marbles from an urn or the experiment of purchasing a given product from a finite set of choices. Similarly, the Normal random model can be used to represent the measurement error of a physical as well as psychophysic experiments.

The distribution function F_X can be parameterized by some reals $\boldsymbol{\theta}$ called **parameters** that modify the way it assigns probabilities. The mathematical domain where the parameters lie is called **parameter space**.

In general $X \sim F_X(x;\boldsymbol{\theta})$ is used to signify that X has distribution function F_X parameterized by $\boldsymbol{\theta}$.

The class of parameterized distribution functions will be called **parametric probabilistic models**. Depending on the support of X we may then have discrete parametric models as well as continuous parametric models.

Some univariate **discrete probabilistic models**

Model	Notation	sup(X)	$\boldsymbol{\theta}$	f_X
Bernoulli	$\mathcal{B}er(x;\pi)$	$\{0,1\}$	$\pi \in [0,1]$	$\pi^x(1-\pi)^{1-x}$
Binomial	$\mathcal{B}in(x;n,\pi)$	$\mathbb{N}_0$	$n \in \mathbb{N}$, $\pi \in [0,1]$	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$
Poisson	$\mathcal{P}oi(x;\lambda)$	$\mathbb{N}_0$	$\lambda \in \mathbb{R}^+$	$\frac{\lambda^x}{x!}\exp(-\lambda)$
Geometric	$\mathcal{G}(x;\pi)$	$\mathbb{N}$	$\pi \in [0,1]$	$\pi(1-x)^{x-1}$
Multinomial	$\mathcal{M}ulti(\mathbf{x};n,\boldsymbol{\pi})$	$\mathbb{N}_0$	$n \in \mathbb{N}$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$, $\boldsymbol{\pi}^T\mathbf{1}_K = 1$	$\left(\frac{n!}{x_1!\cdots x_K!}\right)\pi_1^{x_1}\cdots\pi_K^{x_K}$

Random variables

Source: Appendix D.3 (Fox, 2016) - supplementary materials

Some univariate **continuous probabilistic models**

Model	Notation	sup(X)	$\boldsymbol{\theta}$	f_X
Normal	$\mathcal{N}(x;\mu,\sigma^2)$	$\mathbb{R}$	$\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$	$\left(\sigma\sqrt{2\pi}\right)^{-1} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$
Uniform	$\mathcal{U}(x;\alpha,\beta)$	$[\alpha,\beta] \subset \mathbb{R}$	$\alpha \in \mathbb{R}$ $\beta \in \mathbb{R}$ $\alpha < \beta$	$\frac{1}{\beta-\alpha}$
Exponential	$\mathcal{E}xp(x;\lambda)$	$\mathbb{R}^+$	$\lambda \in \mathbb{R}$	$\lambda \exp(-\lambda x)$
Beta	$\mathcal{B}eta(x;\alpha,\beta)$	$[0,1] \subset \mathbb{R}$	$\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}^+$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Chi-square	$\chi^2(x;\nu)$	$\mathbb{R}^+$	$\nu \in \mathbb{N}$	$\left(2^{\nu/2}\Gamma(\nu/2)\right)^{-1} x^{\nu/2-1} \exp(-x/2)$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

When using random variables it is useful to consider various characteristics (e.g., position, dispersion, shape) that can be summarized numerically.

Expectation. It is denoted by $\mathbb{E}[X]$ and quantifies the mean value to which a sequence of random experiments is expected to converge:

$$\mathbb{E}[X] = \sum_{x \in \text{sup}(X)} x f_X(x) \quad \text{(discrete case)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{(continuous case)}$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Variance. It is denoted by $\mathbb{V}\text{ar}[X]$ and quantifies the dispersion of the outcomes of a sequence of random experiments:

$$\mathbb{V}\text{ar}[X] = \sum_{x \in \text{sup}(X)} (x - \mathbb{E}[X])^2 f_X(x) \quad \text{(discrete case)}$$

$$\mathbb{V}\text{ar}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x)\, dx \quad \text{(continuous case)}$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

For two (or more) random variables $X_1, \ldots, X_J$ an important characteristic to be calculated is the **covariance**, which summarizes the joint variability of the involved r.vs.

Given a pair X_h, X_k ($h \neq k$), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[X_h, X_k] &= \mathbb{E}[(X_h - \mu_{X_h})(X_k - \mu_{X_k})] \\ &= \mathbb{E}[X_h X_k] - \mu_{X_h}\mu_{X_k}\end{aligned}$$

where in general $\mu_X = \mathbb{E}[X]$. The covariance offers a measure of linear association between X_h and X_k. In particular:

- $\mathbb{C}\text{ov}[X_h, X_k] > 0$ indicates that X_h and X_k are positively associated
- $\mathbb{C}\text{ov}[X_h, X_k] < 0$ indicates that X_h and X_k are negatively associated
- $\mathbb{C}\text{ov}[X_h, X_k] = 0$ indicates that X_h and X_k are not *linearly* associated

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Expectations for some important distributions

Model	Notation	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Bernoulli	$\mathcal{B}er(x;\pi)$	π	$\pi(1-\pi)$
Binomial	$\mathcal{B}in(x;n,\pi)$	$n\pi$	$n\pi(1-\pi)$
Poisson	$\mathcal{P}oi(x;\lambda)$	λ	λ
Geometric	$\mathcal{G}(x;\pi)$	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$
Multinomial	$\mathcal{M}ulti(\mathbf{x};n,\boldsymbol{\pi})$	$n\pi_1,\ldots,$ $n\pi_J$	$n\pi_1(1-\pi_1),\ldots,$ $n\pi_J(1-\pi_J)$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Expectations for some important distributions

Model	Notation	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Normal	$\mathcal{N}(x;\mu,\sigma^2)$	μ	σ^2
Uniform	$\mathcal{U}(x;\alpha,\beta)$	$\frac{1}{2}(\alpha+\beta)$	$\frac{1}{12}(\beta-\alpha)^2$
Exponential	$\mathcal{E}xp(x;\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Beta	$\mathcal{B}eta(x;\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Chi-square	$\chi^2(x;\nu)$	ν	2ν

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Some important properties for expectations

$$\mathbb{E}[\alpha] = \alpha$$

$$\mathbb{E}[X_h + X_k] = \mathbb{E}[X_h] + \mathbb{E}[X_k]$$

$$\mathbb{E}[\beta X_h] = \beta\mathbb{E}[X_h]$$

$$\mathbb{E}[\alpha + \beta X_h] = \alpha + \beta\mathbb{E}[X_h]$$

$$\mathbb{V}\text{ar}[\alpha] = 0$$

$$\mathbb{V}\text{ar}[X_h + X_k] = \mathbb{V}\text{ar}[X_h] + \mathbb{V}\text{ar}[X_k] + 2\mathbb{C}\text{ov}[X_h, X_k]$$

$$\mathbb{V}\text{ar}[\beta X_h] = \beta^2\mathbb{V}\text{ar}[X_h]$$

$$\mathbb{V}\text{ar}[\alpha + \beta X_h] = \beta^2\mathbb{V}\text{ar}[X_h]$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Often a random experiment is described by more than one random variable (**random vectors**).

Given a random vector $X = (X_1, \ldots, X_J)$ the **joint probability distribution** is defined as

$$F_{X_1,\ldots,X_J}(X_1 = x_1, \ldots, X_J = x_J) = \mathbb{P}(X_1 \leq x_1, \ldots, X_J \leq x_J)$$

The **marginal probability distribution** is obtained by integration (continuous case) or summation (discrete case). For example, in the continuous case ($J = 2$):

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

The **conditional probability distribution** is defined as follows ($J = 2$):

$$f_{X_1|X_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

If $X_1 \perp\!\!\!\perp X_2$ (**independence**), then:

$$f_{X_1|X_2}(x_1) = f_{X_1}$$

or alternatively

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

In the multivariate context, **conditional expectations** can be obtained as well:

$$\mathbb{E}[X_1|X_2 = x_2] = \int_{-\infty}^{\infty} x_1 \; f_{X_1|X_2}(x_1) \; dx_1$$

$$\mathbb{V}\text{ar}[X_1|X_2 = x_2] = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1|X_2])^2 \middle| X_2 = x_2\right]$$

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Similarly to the univariate case, there are several parametric probabilistic models for the multivariate case. For example, the most relevant model for the continuous case is the **multivariate Normal model** $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \vdots \\ \mu_j \\ \vdots \\ \mu_J\end{bmatrix}, \begin{bmatrix}\sigma^2_{11} & \dots & \sigma_{1j} & \dots & \sigma_{1J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{j1} & \dots & \sigma^2_{jj} & \dots & \sigma_{jJ} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{J1} & \dots & \sigma_{Jj} & \dots & \sigma^2_{JJ}\end{bmatrix}\right)$$

with $\boldsymbol{\mu}_{J\times 1}$ being the vector of the means and $\boldsymbol{\Sigma}_{J\times J}$ the **covariance matrix**.

For further details, see Appendix D.3.5 (Fox, 2016) in the supplementary materials of the course.

Random variables

Source: Appendix D.1.2 (Fox, 2016) - supplementary materials

Random variables $X_1, \dots, X_J$ are said to be **independent and identically distributed** (iid) iff:

$$f_{X_1,\dots,X_J}(x_1, \dots, x_J) = f_{X_1}(x_1) \cdots f_{X_J}(x_J)$$

$$f_{X_1} = f_{X_2} = \dots = f_{X_J}$$

Independent and identically distributed random variables constitute the building block of simple **random samples**. Moreover, they are at the base of **limit theorems**, which are important for statistical inference.

(Weak) Law of Large Numbers

Source: Appendix D.4.1 (Fox, 2016) - supplementary materials

Let $X_1, \ldots, X_n$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \sigma^2$ for each term of the sequence and let

$$\overline{X} = \frac{1}{n}\sum_i X_i$$

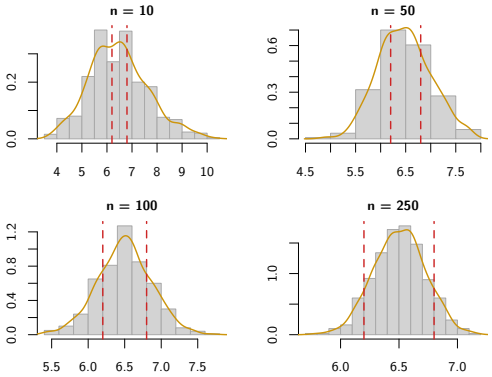
be the mean of the random sequence. Then given any positive number ϵ (no matter how small) we have:

$$\boxed{\lim_{n\to\infty} \mathbb{P}(\mu - \epsilon < \overline{X}_n < \mu + \epsilon) = 1}$$

In other words, the random variable $\overline{X}$ is close to μ for large n.

(Weak) law of large numbers

Source: Appendix D.4.1 (Fox, 2016) - supplementary materials



Notes:
$X_i \sim \chi^2(n, \lambda = 6.5)$, $n = (10, 50, 100, 250)$. Dotted red lines indicate the set $I_{\epsilon,\lambda} = [\lambda \pm \epsilon]$, $\epsilon = 0.3$.
As n increases, $I_{\epsilon,\lambda}$ gets larger:
$\mathbb{P}(\overline{X}_{n=10} \in I_{\epsilon,\lambda}) = 0.218$, $\mathbb{P}(\overline{X}_{n=50} \in I_{\epsilon,\lambda}) = 0.422$, $\mathbb{P}(\overline{X}_{n=100} \in I_{\epsilon,\lambda}) = 0.586$,
$\mathbb{P}(\overline{X}_{n=250} \in I_{\epsilon,\lambda}) = 0.814$

Central limit theorem

Source: Appendix D.4.3 (Fox, 2016) - supplementary materials

Let $X_1, \ldots, X_n$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \sigma^2$ for each term of the sequence and let

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

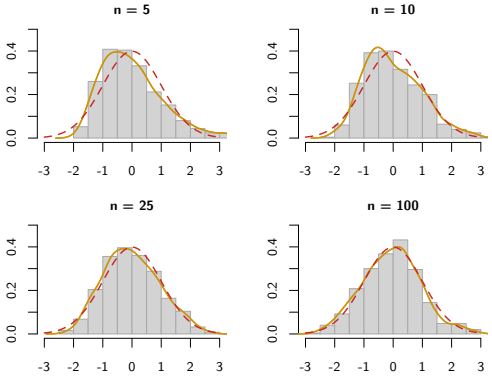
be the **standardized random variable** with $\mathbb{E}[Z] = 0$ e $\mathbb{V}\text{ar}[Z] = 1$. Then for $x \in \mathbb{R}$ we have:

$$\boxed{\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \mathbb{P}(Z \leq x)} \quad \text{with} \quad Z \sim \mathcal{N}(0,1)$$

In other words, the random variable Z_n has a distribution that is approximately standardized Normal (no matter how $X_1, \ldots, X_n$ are distributed).

Central limit theorem

Source: Appendix D.4.3 (Fox, 2016) - supplementary materials



Notes:
$X_i \sim \mathcal{E}xp(n, \lambda = 1)$, $n = (5, 10, 25, 100)$. Dotted red curves indicate the standardized Normal distribution.
As n increases, the distribution of Z_n approximates the standardized Normal distribution.

Statistical inference

A **statistical model** can generally be defined as a triplet

$$\mathcal{M} = \{F_Y(y;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p, y \in \mathcal{Y}\}$$

where

- $F_Y(y;\boldsymbol{\theta})$ is a parametric probabilistic model
- $\boldsymbol{\Theta}$ is the parametric space for $\boldsymbol{\theta}$
- $\mathcal{Y}$ is the sample space, i.e. the space where $\sup(Y)$ is defined

Examples:

- *Normal model*:
  $p = 2$, $\boldsymbol{\theta} = \{\mu, \sigma^2\} \in \boldsymbol{\Theta} \subset \mathbb{R} \times \mathbb{R}^+$, $\mathcal{Y} \subseteq \mathbb{R}$, and $F_Y(y;\boldsymbol{\theta}) = \mathcal{N}(y;\mu,\sigma^2)$
- *Bernoulli model*:
  $p = 1$, $\boldsymbol{\theta} = \pi \in [0,1]$, $\mathcal{Y} \subseteq \{0,1\}$, $F_Y(y;\boldsymbol{\theta}) = \mathcal{B}in(y;\pi)$

Statistical inference

In general, we have two ways for dealing with a statistical model $\mathcal{M}$:

(i) **Top-down approach**: the observer knows in advance the elements of the model - i.e. $\boldsymbol{\theta}$, $\mathcal{Y}$, and $F_Y(y;\boldsymbol{\theta})$ - with the purpose of simulating new instances/samples $\{y_1,\ldots,y_n\}$ from $\mathcal{M}$. For instance, this approach can be used to assess the inner-working mechanisms of $\mathcal{M}$.

(ii) **Bottom-up approach**: the observer has a set of instances/observations $\mathbf{y} = \{y_1,\ldots,y_n\}$ but nothing is known about $\mathcal{M}$ in advance. Then, the purpose here is to infer the most plausible model $\mathcal{M}^0 \in \{\mathcal{M}_1,\ldots,\mathcal{M}_K\}$ which has generated the observed sample $\mathbf{y}$.

Statistical inference

Example

With the goal of determining the level of a cognitive ability μ^0 in a non-clinical population, a sample of observations $\mathbf{y}_{n\times 1}$ has been collected by means of a cognitive test. From a statistical point-of-view, we need to determine which of the models $F(y;\mu_1),\ldots,F(y;\mu_k),\ldots,F(y;\mu_K)$ is the most plausible for μ^0 given $\mathbf{y}$.

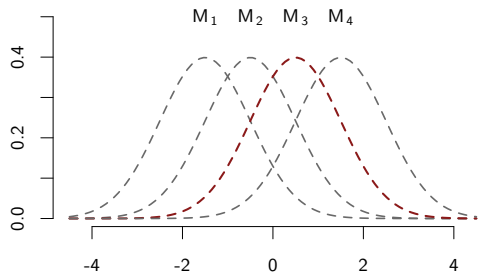
By previous knowledge about μ, we can set $\mathcal{Y} = \mathbb{R}$ and

$$F(\mathbf{y};\mu) = \mathcal{N}(\mathbf{y};\mu,\sigma^2 = 1)$$

Then, the goal becomes that of estimating $\mu^0 \in \mathbb{R}$ given $\mathbf{y}$, which implies selecting the most plausible model from the set $F(\mathbf{y};\mu_1),\ldots,F(\mathbf{y};\mu_k),\ldots,F(\mathbf{y};\mu_K)$.

Note: $\mathcal{N}(\mathbf{y};\mu,\sigma^2 = 1)$ is a *location model*.

Statistical inference



Notes:
$K = 4$ plausible location models for the cognitive ability estimation.
The most plausible model given the data $\mathbf{y}$ is M_3 (red dotted curve).

Statistical inference

Determining $\mathcal{M}^0$ means making inference about the true but unknown parameter $\mu^0 \in \mathbb{R}$ of the true model $F^0(y;\boldsymbol{\theta})$. The procedure requires a **theory of statistical inference** which establishes the *correctness*, the *bias*, and the *uncertainty* of the estimates $\hat{\boldsymbol{\theta}}$.

A couple of approaches are available to this end: frequentist, Bayesian, information-theoretic based. Within the frequentist framework, the **maximum likelihood theory** is the most studied and most commonly used approach to statistical inference.

For a brief review of ML theory, see Appendix D.6 (Fox, 2016) in the supplementary materials of the course.

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

A function of the data $t(\mathbf{y})$ is called **statistic** and, under some regularities, it summarizes the data information in the most optimal way. For example, the sample mean $\overline{y} = \frac{1}{n}\sum_i y_i$ is a statistic of the sample $\mathbf{y}$. As a statistic is computed over samples, which are in turn outcomes of r.vs., itself is a random variable $T(Y)$ with an own distribution as well. For example, the statistic $\overline{Y} = \sum_i Y_i$ is a random variable following the Normal distribution.

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

Estimators $\hat{\boldsymbol{\theta}}(Y)$ are statistics of the data and their outcomes $\hat{\boldsymbol{\theta}}(\mathbf{y})$ or simply $\hat{\boldsymbol{\theta}}$ are called **estimates**. For instance, in the location model $\mathcal{N}(y;\mu)$ the estimator for the parameter μ is $\hat{\mu} = \frac{1}{n}\sum_i y_i$. The probability distribution of $\hat{\boldsymbol{\theta}}(Y)$ is called **sampling distribution** and provides information about $\hat{\boldsymbol{\theta}}$. The variance of an estimator $\mathbb{V}\text{ar}\left[\hat{\boldsymbol{\theta}}\right]$ (or $\sigma^2_{\hat{\boldsymbol{\theta}}}$) provides important information about the uncertainty of the estimates $\hat{\boldsymbol{\theta}}$.

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

An estimator $\hat{\boldsymbol{\theta}}(Y)$ for the parameter θ^0 is:

- **ubiased** iff
$$\text{B}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] - \theta^0 = 0$$
This means that its average value over repeated samples is equal to the parameter being estimated

An estimator $\hat{\boldsymbol{\theta}}(Y)$ for the parameter θ^0 is:

- **efficient** iff its Mean Square Error (MSE)

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(Y) - \theta^0\right)^2\right] = \mathbb{V}\text{ar}\left[\hat{\boldsymbol{\theta}}\right] + \text{B}(\hat{\boldsymbol{\theta}})^2$$

is as lower as possible.
For unbiased estimators, the efficiency of an estimator increases, therefore, as its sampling variance $\mathbb{V}\text{ar}\left[\hat{\boldsymbol{\theta}}\right]$ declines.

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

An estimator $\hat{\boldsymbol{\theta}}(Y)$ for the parameter θ^0 is:

- **consistent** if the bias and the sampling variance approach zero as n increases.

Antonio Calcagnì University of Padova
PSQ1096299 - First Part (module A) Preliminaries 59/64

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

Example

Consider a random sample $Y_1, \ldots, Y_n$ from a probabilistic model with parameters $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}\text{ar}[Y_i] = \sigma^2$.

Then, two estimators for μ are the following

$$\hat{\boldsymbol{\theta}}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \hat{\boldsymbol{\theta}}_2 = Y_1$$

Antonio Calcagnì University of Padova
PSQ1096299 - First Part (module A) Preliminaries 60/64

Statistical inference
Source: Appendix D.5 (Fox, 2016) - supplementary materials

Example

Both estimators are unbiased:

$$\begin{aligned}\text{B}\left(\hat{\boldsymbol{\theta}}_1\right) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] - \mu \\ &= \frac{1}{n}n\mu - \mu = 0\end{aligned}$$

$$\begin{aligned}\text{B}\left(\hat{\boldsymbol{\theta}}_2\right) &= \mathbb{E}[Y_1] - \mu \\ &= \mu - \mu = 0\end{aligned}$$

Antonio Calcagnì University of Padova
PSQ1096299 - First Part (module A) Preliminaries 60/64

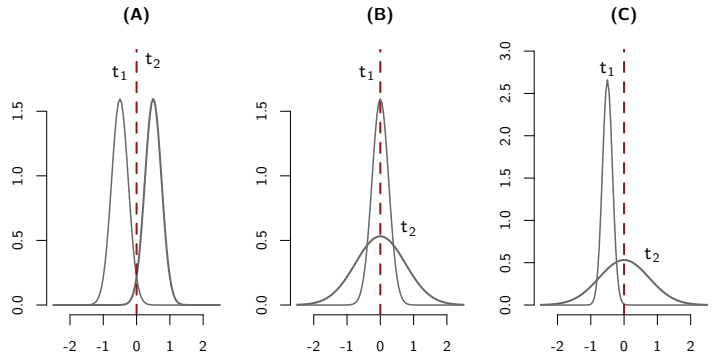
Example

However, the second estimator is not as good as the first one:

$$\begin{aligned}\text{MSE}\left(\hat{\boldsymbol{\theta}}_1\right) &= \mathbb{V}\text{ar}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \text{B}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \\ &= \frac{1}{n^2}\mathbb{V}\text{ar}\left[\sum_{i=1}^{n} Y_i\right] + 0 \\ &= \frac{1}{n^2} n\sigma^2 + 0 = \frac{\sigma^2}{n}\end{aligned}$$

$$\begin{aligned}\text{MSE}\left(\hat{\boldsymbol{\theta}}_2\right) &= \mathbb{V}\text{ar}\left[Y_1\right] + \text{B}\left(Y_1\right)^2 \\ &= \sigma^2 + 0 = \sigma^2\end{aligned}$$

As $\text{MSE}\left(\hat{\boldsymbol{\theta}}_1\right) < \text{MSE}\left(\hat{\boldsymbol{\theta}}_2\right)$, the estimator $\hat{\boldsymbol{\theta}}_1$ should be preferred over $\hat{\boldsymbol{\theta}}_2$.



Notes:
Sampling distributions for two estimators (dotted red line indicates the true parameter).
(A) - Biased estimators with same variance; **(B)** - Unbiased estimators with different variance;
(C) - Biased estimator (t_1) vs. unbiased estimator (t_2) with different variance.
Although t_2 is unbiased, t_1 would be preferred if bias could be removed.

There are several ways to build estimators for unknown parameters $\boldsymbol{\theta}$, e.g.:

- Method of moments
- Least squares method
- Maximum likelihood
- Bayesian approach
- Monte Carlo based methods
- Information-theoretic based methods

We will see some of them (more in depth) in Module B.

Example

Reconsider the location model used to estimate the cognitive ability μ^0 given the random sample $\mathbf{y}_{n\times 1}$:

$$y_i \sim \mathcal{N}(y_i; \mu, \sigma^2 = 1)$$

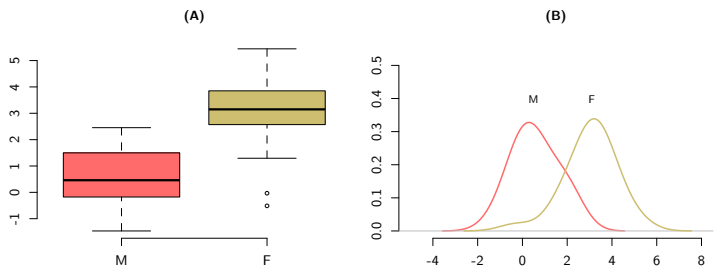
Then, we can extend the model to analyse whether the cognitive ability varies as a function of the categorical variable gender $\mathbf{z} \in \{0, 1\}^n$, which has the following levels $z_i = 0$ (male) and $z_i = 1$ (female).

This requires rewriting the mean of the model as a function of the new variable:

$$\mu_i = \beta_0 + z_i\beta_1$$

The result is still a location model but now it codifies two means, one for the male group when $z_i = 0$ ($\mu_i = \beta_0$) and the other one for the female group $z_i = 1$ ($\mu_i = \beta_0 + \beta_1$).

Statistical inference



Notes:
Linear model for the cognitive ability μ^0 as a function of gender.
In this example: $n = 100$ ($n_M = 50$), $\beta_0 = 0.5$, $\beta_1 = 2.6$, $\hat{\mu}_M = 0.5$, $\hat{\mu}_F = 3.1$.

(A) Observed data **y** plotted as a function of **z**
(B) Estimated densities $\hat{F}_Y(\mathbf{y}; \hat{\boldsymbol{\mu}})$ plotted as a function of **z**.

Statistical methods and data analysis in developmental psychology

Antonio Calcagnì

DPSS, University of Padova

A.Y 2021-2022

Copyright © 2021 Antonio Calcagnì. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation A copy of the license is available at: https://www.gnu.org/licenses/fdl-1.3.html.

Outline

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

Let $Y = (Y_1, \ldots, Y_i, \ldots, Y_n)$ be a collection of independent random variables. For each outcome y_i, a set of (non-random) variables is collected

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$$

so that the observed sample is represented in terms of pairs

$$\mathbf{y} = \{(y_1, \mathbf{x}_1), \ldots, (y_i, \mathbf{x}_i), \ldots, (y_n, \mathbf{x}_n)\}$$

From now on $\mathbf{x}_i$ is considered continuous without loss of generality. The statistical theory underlying the Normal linear model is the same for both continuous and categorical predictors.

We will discuss in more details the categorical case (e.g., *dummy coding*) later on.

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

The Normal linear model for the sample $\mathbf{y}$ is of the form:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{aligned}$$

where $\mathbf{x}_i$ is a $J \times 1$ vector of predictors, $\beta_0 \in \mathbb{R}$ is the parameter for the offset of the model, $\boldsymbol{\beta} \in \mathbb{R}^J$ is a $J \times 1$ vector of model coefficients, and $\sigma^2 \in \mathbb{R}^+$ is constant over the observations (homoscedasticity).

Unlike other models (e.g., Binomial, Poisson), this model codifies the mean of the random variables as a function of the predictors $\mathbb{E}[Y_i] = \mu_i$ whereas the variance is kept constant and independent from the mean $\mathbb{V}\text{ar}[Y_i] = \sigma^2$.

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{aligned}$$

The parameters of the model $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \sigma^2\}$ can be interpreted as follows:

- β_0 is the intercept term, i.e. $\mathbb{E}[Y_i]$ when $\boldsymbol{\beta} = \mathbf{0}_J$
- $\boldsymbol{\beta} > \mathbf{0}_J$: the regression line or (multidimensional) regression plane increases as a function of the predictors
- $\boldsymbol{\beta} < \mathbf{0}_J$: the regression line or (multidimensional) regression plane decreases as a function of the predictors

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{aligned}$$

The parameters of the model $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \sigma^2\}$ can be interpreted as follows:

- $\boldsymbol{\beta} = \mathbf{0}_J$: *location model*, i.e. $y_i \sim \mathcal{N}(\beta_0, \sigma^2)$
- σ^2: constant variance across observations and each level of the predictors (*homogeneity of the variance*)

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\
\mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\
\sigma_i^2 &= \sigma^2
\end{aligned}
$$

The parameters of the model $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \sigma^2\}$ can be interpreted as follows:

- the single coefficient β_j codifies the *partial effect* of the j-th predictor on the response variable Y_i when the remaining $\boldsymbol{\beta}_{-j}$ are fixed

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\
\mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\
\sigma_i^2 &= \sigma^2
\end{aligned}
$$

The following assumptions follow from the model definition:

- *linearity*: $\mathbb{E}[Y_i]$ is a linear function of $\mathbf{x}_i$ (i.e., $\mathbb{E}[Y_i] = g(\mathbf{x}_i^T \boldsymbol{\beta})$ with $g(.)$ identity function)
- *homoscedasticity*: $\sigma_i^2 = \sigma^2$, i.e. constant variance for all the observations
- *normality*: the conditional distribution of the response variable $Y_i|\mathbf{x}_i$ is Normal

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\
\mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\
\sigma_i^2 &= \sigma^2
\end{aligned}
$$

The following assumptions follow from the model definition:

- *independence*: the random variables underlying the responses are independent $Y_i \perp\!\!\!\perp Y_h$ ($i \neq h$). Moreover, they are also identically distributed
- *non-random predictors*: the explanatory variables $X_1, \ldots, X_J$ are fixed and measured without error
- *absence of collinearity*: no predictor is a perfect linear function of the others

Model definition

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

The (multidimensional) Normal linear model can also be represented more compactly using the matrix notation:

$$
\mathbf{y}_{n\times 1} \sim \mathcal{N}(\mathbf{X}_{n\times J}\boldsymbol{\beta}_{J\times 1}, \mathbf{I}_{n\times n}\sigma^2)
$$

$$
\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 1 & \ldots & x_{11} & \ldots & x_{1j} & \ldots & x_{1J} \\ \vdots & & \vdots & & \vdots & & \vdots \\ 1 & \ldots & x_{i1} & \ldots & x_{ij} & \ldots & x_{iJ} \\ \vdots & & \vdots & & \vdots & & \vdots \\ 1 & \ldots & x_{n1} & \ldots & x_{nj} & \ldots & x_{nJ} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_J \end{bmatrix}}_{\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}}, \underbrace{\begin{bmatrix} 1 & \ldots & \ldots \\ \vdots & & \vdots \\ \ldots & 1 & \ldots \\ \vdots & & \vdots \\ \ldots & \ldots & 1 \end{bmatrix} \sigma^2}_{\boldsymbol{\Sigma}=\mathbf{I}\sigma^2} \right)
$$

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

After the Normal linear model has been formulated, given an observed sample of data $\mathbf{y}$ the goal is to estimate the parameters $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \sigma^2\}$ which identify a linear model $\hat{\mathcal{M}}$ from the set of possible linear models indexed by $\boldsymbol{\theta}$.

Once $\hat{\mathcal{M}}$ has been found, the estimates $\hat{\boldsymbol{\theta}}$ can be used to evaluate the model and interpret the results.

The parameters can be estimated via the theory of Maximum Likelihood. For general details, see Appendix D.6 (Fox, 2016) in the supplementary materials of the course.

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

The log-Likelihood function of the Normal linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$ is as follows:

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) &= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 \\ &= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

The problem of estimate $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ is solved by maximizing $\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ w.r.t. the unknown quantities:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$$

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$$

With regards to $\boldsymbol{\beta}$ we get:

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}_J \\ &\frac{\partial}{\partial \boldsymbol{\beta}} \left(-\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) = \mathbf{0}_J \\ &\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}_J \\ &2\mathbf{y}^T\mathbf{X} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} = \mathbf{0}_J \\ &(\mathbf{X}^T\mathbf{X}) = \mathbf{X}^T\mathbf{y} \qquad \text{Normal equations} \\ &\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad \text{The solution exists providing that } (\mathbf{x}^T\mathbf{x})^{-1} \text{ exists.} \end{aligned}$$

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$$

With regards to σ^2 we get:

$$\begin{aligned} &\frac{\partial}{\partial \sigma^2} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = 0 \\ &\frac{\partial}{\partial \sigma^2} \left(-\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) = 0 \\ &-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ &-n + \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\ &\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$$

Finally, the (nested) solutions are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

Note: The point $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\}$ is a maximum for $\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ since $\frac{\partial^2}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) \geq \mathbf{0}_J$ and $\frac{\partial^2}{\partial \sigma^2} \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) < 0$.

Parameter estimation

Source: 5.2.1, 5.2.2, 9.3.3 (Fox, 2016); 2.4, 2.5 (Faraway, 2014)

In the simplest case of a single predictor ($J = 1$), the Normal linear model simplifies to

$$\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta, \mathbf{1}_n\sigma^2)$$

and the estimators simplify as well (see Table 9.1, Fox 2016):

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{x})}$$

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i\hat{\beta})^2$$

Parameter estimation

Source: 9.3.1 (Fox, 2016); 2.8 (Faraway, 2014)

The estimator $\hat{\boldsymbol{\beta}}$ has the following properties:

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \\
&= \cancel{(\mathbf{X}^T\mathbf{X})^{-1}}\cancel{(\mathbf{X}^T\mathbf{X})}\boldsymbol{\beta} \quad \text{(unbiasness)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\hat{\boldsymbol{\beta}}\right] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{V}\text{ar}[\mathbf{y}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}\sigma^2)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\cancel{\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

Parameter estimation

Source: 9.3.2 (Fox, 2016); 2.8 (Faraway, 2014)

From the **Gauss-Markov theorem** we know that $\hat{\boldsymbol{\beta}}$ is BLUE (*best linear unbiased estimator*), i.e. it is unbiased and show smaller variance among the set of all possible linear estimators for $\boldsymbol{\beta}$.

The **distribution** of $\hat{\boldsymbol{\beta}}$ can be used to make inference about the coefficients $\boldsymbol{\beta}$. It is obtained from the Normality assumption of the liner model:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\underbrace{\boldsymbol{\beta}}_{\mathbb{E}[\boldsymbol{\beta}]}, \underbrace{\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}}_{\mathbb{V}\text{ar}[\boldsymbol{\beta}]}\right)$$

The marginal distribution for a specific coefficient of the model is obtained by taking the j-th element of the estimator: $\beta_j \sim \mathcal{N}\left(\beta_r, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}\right)$.

Example
Math anxiety and test difficulty

Let consider the Example 2 (module A) again. The data refers to a subset of $n = 15$ data containing response times (in sec) to a math test (RT) as a function of math anxiety (`math anx`) and test difficulty `diff`.

	RT	math anx	diff
1	0.96	1.53	1.04
2	4.08	1.81	1.21
3	3.98	3.11	1.38
4	4.56	3.16	1.86
5	4.70	3.88	1.96
6	5.82	4.03	2.50
7	6.10	4.22	3.12
8	4.44	4.38	3.61
9	6.47	4.42	3.87
10	5.26	4.59	4.03
11	5.53	4.85	4.18
12	7.47	5.00	4.27
13	6.84	5.79	4.28
14	8.85	7.07	4.31
15	12.79	8.65	4.43

Example
Math anxiety and test difficulty

The aim is to study whether response times varies as a (linear) function of both math anxiety and test difficulty:

$$\texttt{RT}_i = \beta_0 + \texttt{math anx}\beta_1 + \texttt{diff}\beta_2 + \epsilon_i$$

Under the common Normality assumption for the error component $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{1}_n\sigma^2)$, we get a Normal linear model

$$\texttt{RT}_i \sim \mathcal{N}(\beta_0 + \texttt{math anx}\beta_1 + \texttt{diff}\beta_2, \sigma^2)$$

More compactly:

$$\underbrace{\mathbf{y}}_{\texttt{RT}} \sim \mathcal{N}(\underbrace{\mathbf{X}}_{[1\ \texttt{math\_anx}\ \texttt{diff}]} \underbrace{\boldsymbol{\beta}}_{[\beta_0\ \beta_1\ \beta_2]}, \mathbf{I}\sigma^2)$$

Example
Math anxiety and test difficulty

In matrix form:

$$\underbrace{\begin{bmatrix}0.96\\4.08\\3.98\\4.56\\4.70\\5.82\\6.10\\4.44\\6.47\\5.26\\5.53\\7.47\\6.84\\8.85\\12.79\end{bmatrix}}_{\mathbf{y}_{15\times 1}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix}1 & 1.53 & 1.04\\1 & 1.81 & 1.21\\1 & 3.11 & 1.38\\1 & 3.16 & 1.86\\1 & 3.88 & 1.96\\1 & 4.03 & 2.50\\1 & 4.22 & 3.12\\1 & 4.38 & 3.61\\1 & 4.42 & 3.87\\1 & 4.59 & 4.03\\1 & 4.85 & 4.18\\1 & 5.00 & 4.27\\1 & 5.79 & 4.28\\1 & 7.07 & 4.31\\1 & 8.65 & 4.43\end{bmatrix}}_{\mathbf{x}_{15\times(2+1)}} \underbrace{\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix}}_{\boldsymbol{\beta}_{3\times 1}}, \underbrace{\begin{bmatrix}1 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 1 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 0\\0 & 0 & \dots & 1\end{bmatrix}}_{\mathbf{I}_{15\times 15}} \sigma^2 \right)$$

Example
Math anxiety and test difficulty

To estimate the parameters of the model we apply the solutions provided before (see Slide 13).

For the linear coefficients $\boldsymbol{\beta}$ we have:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Then:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix}15.00 & 66.49 & 46.05\\66.49 & 341.02 & 231.41\\46.05 & 231.41 & 164.21\end{bmatrix} \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix}0.52209 & -0.05584 & -0.06772\\-0.05584 & 0.07300 & -0.08721\\-0.06772 & -0.08721 & 0.14798\end{bmatrix} \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix}87.85\\452.24\\304.06\end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix}0.52209 & -0.05584 & -0.06772\\-0.05584 & 0.07300 & -0.08721\\-0.06772 & -0.08721 & 0.14798\end{bmatrix}\begin{bmatrix}87.85\\452.24\\304.06\end{bmatrix} = \begin{bmatrix}0.0214\\1.5899\\-0.3949\end{bmatrix}$$

Example
Math anxiety and test difficulty

To estimate the parameters of the model we apply the solutions provided before (see Slide 13).

For the variance of the model σ^2 we have:

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\left(\frac{n}{n-J-1}\right)$$

Then:

$$\hat{\sigma}^2 = 0.06667 \left(\begin{bmatrix} 0.96 \\ 4.08 \\ 3.98 \\ 4.56 \\ 4.70 \\ 5.82 \\ 6.10 \\ 4.44 \\ 6.47 \\ 5.26 \\ 5.53 \\ 7.47 \\ 6.84 \\ 8.85 \\ 12.79 \end{bmatrix} - \begin{bmatrix} 2.043 \\ 2.421 \\ 4.421 \\ 4.311 \\ 5.416 \\ 5.442 \\ 5.499 \\ 5.560 \\ 5.521 \\ 5.728 \\ 6.082 \\ 6.285 \\ 7.537 \\ 9.560 \\ 12.025 \end{bmatrix} \right)^T \left(\begin{bmatrix} 0.96 \\ 4.08 \\ 3.98 \\ 4.56 \\ 4.70 \\ 5.82 \\ 6.10 \\ 4.44 \\ 6.47 \\ 5.26 \\ 5.53 \\ 7.47 \\ 6.84 \\ 8.85 \\ 12.79 \end{bmatrix} - \begin{bmatrix} 2.043 \\ 2.421 \\ 4.421 \\ 4.311 \\ 5.416 \\ 5.442 \\ 5.499 \\ 5.560 \\ 5.521 \\ 5.728 \\ 6.082 \\ 6.285 \\ 7.537 \\ 9.560 \\ 12.025 \end{bmatrix} \right) \cdot 1.25 = 0.9048$$

Example
Math anxiety and test difficulty

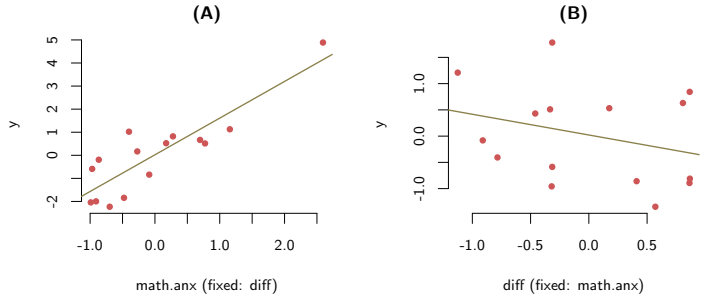
The estimated parameters are the following:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.0214 \\ 1.5899 \\ -0.3949 \end{bmatrix} \qquad \hat{\sigma}^2 = 0.9048$$

Parameter interpretation:

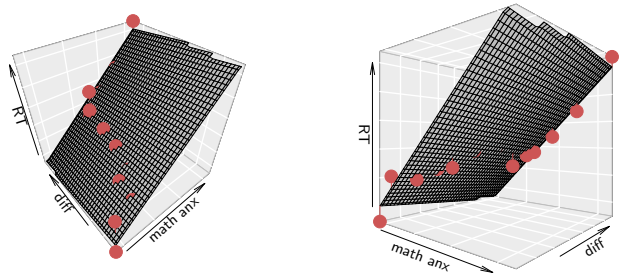
- $\hat{\beta}_0 = 0.0214$: the mean of `RT` when the predictors are fixed at zero
- $\hat{\beta}_1 = 1.5899$: as `math anx` increases by one unit, `RT` also increases by 1.59 sec (by controlling for `diff`)
- $\hat{\beta}_2 = -0.3949$: as `diff` increases by one unit, `RT` decreases by 0.39 sec (by controlling for `math anx`)

Example
Math anxiety and test difficulty



Partial regression plots: **(A)** `RT` as a function of `math anx` by fixing `diff` and **(B)** `RT` as a function of `diff` by fixing `math anx`. Note that variables are plotted on residual scale. The plots have been produced using the method described by:
Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4), 234-242.

Example
Math anxiety and test difficulty



Bivariate regression plot for two different perspectives.

Example

Math anxiety and test difficulty

To evaluate the accuracy of the estimates $\hat{\boldsymbol{\beta}}$, **standard errors** can be computed using the variance formula $\sigma^2_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ where $\mathbf{X}^T\mathbf{X}$ is the covariance matrix of the predictors:

$$\sigma_{\hat{\boldsymbol{\beta}}} = \sqrt{\text{diag}(\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2}$$

$$= \sqrt{\begin{bmatrix} 0.52209 & -0.05584 & -0.06772 \\ -0.05584 & 0.07300 & -0.08721 \\ -0.06772 & -0.08721 & 0.14798 \end{bmatrix}\begin{bmatrix}0.9048\end{bmatrix}} = \begin{bmatrix}0.6873 & 0.2570 & 0.3659\end{bmatrix}$$

Next, standard errors can be used to compute statistics to make inference about $\hat{\boldsymbol{\beta}}$ (e.g., t-statistic), to compute confidence intervals, and confidence bands for the regression line. We will see all of them soon.

Coefficient of determination R^2

Source: 5.2.1, 5.2.4 (Fox, 2016); 2.9 (Faraway, 2014)

Once the model parameters have been estimates, one can ask whether the estimated linear model is good enough in predicting the observations $\mathbf{y}$. More precisely, one can evaluate to what extend the **predicted values**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

resemble the observations $\mathbf{y}$.

To this end, the **coefficient of determination** R^2 can be used:

$$R^2 = 1 - \frac{(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}})}{(\mathbf{y}-\mathbf{1}\bar{y})^T(\mathbf{y}-\mathbf{1}\bar{y})} = \frac{\text{residual sum of squares}}{\text{total sum of squares}}$$

with $\bar{y}$ being the sample mean.

Note that $R^2 \in [0,1]$, with $R^2 = 1$ indicating a perfect fit for the model.

Coefficient of determination R^2

Source: 5.2.1, 5.2.4 (Fox, 2016); 2.9 (Faraway, 2014)

More generally, when $J > 1$ it can be useful to adjust the overall fit index R^2 to prevent the case where the index increases because of spurious predictors in the model.

The most common adjustment (i.e., *McNemar's* R^2) is as follows:

$$R^2_{\text{adj}} = 1 - (1-R^2)\cdot \underbrace{\left(\frac{n-1}{n-J-1}\right)}_{\text{adjustment factor}}$$

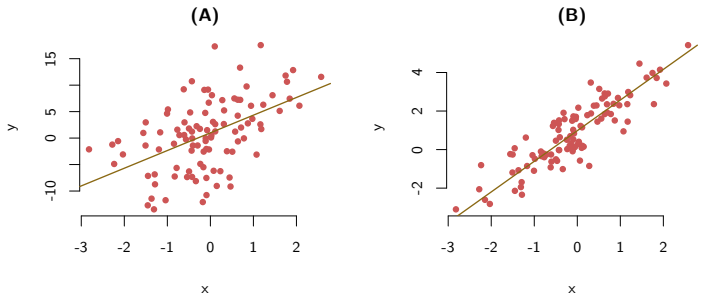
In general

$$R^2_{\text{adj}} \leq R^2$$

with $R^2_{\text{adj}} \in (-\infty, 1]$ (the adjusted index can sometimes be negative).

Coefficient of determination R^2

Source: 5.2.1, 5.2.4 (Fox, 2016); 2.9 (Faraway, 2014)



R^2 index for two Normal linear models: **(A)** Case with a lower overall fit index $R^2 = 0.25$ and **(B)** case with a higher overall fit index $R^2 = 0.76$. In the first case, the fitted model explains about the 25% of overall variance. By contrast, in the second case the fitted model explains a higher amount of overall variance (about 76%).

Example
Math anxiety and test difficulty

Consider the math anxiety example again.
Here, the *total sum of squares* is

$$(\mathbf{y} - \mathbf{1}\bar{y})^T(\mathbf{y} - \mathbf{1}\bar{y}) = 97.18$$

whereas the *residual sum of squares* is

$$(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = 10.86$$

The R^2 index is then

$$R^2 = 1 - \frac{10.86}{97.18} = 0.888$$

As $J = 2$, we can adjust for the number of predictors. This yields to

$$R^2_{adj} = 0.888 \left(\frac{15-1}{15-2-1}\right) = 0.869$$

Testing individual coefficients
Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

To test $\hat{\boldsymbol{\beta}}_{J+1\times 1}$ element-wise, we use the *statistic*:

$$\boxed{T_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta^0}{\sigma_{\hat{\beta}_j}}}$$

with β_0 being the value under the **null hypothesis** $H_0 = \beta_j = \beta^0$ against the test being performed.

From the distribution of the estimators $\{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\}$, we know that this statistic is distributed according to a t-Student distribution with $n - J$ degrees of freedom:

$$\boxed{T_{\hat{\beta}_j} \overset{H_0}{\sim} t(; n - J - 1)}$$

Testing individual coefficients
Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

To choose whether H_0 has to be *rejected* for a fixed α, the observed statistic $t_{\hat{\beta}_j}$ needs to be compared to the reference value $t^0_{n-J|1-\alpha/2}$ of the t-Student distribution for the test being computed.

For a **symmetric test**, H_0 is rejected iif

$$\boxed{|t_{\hat{\beta}_j}| > t^0_{n-J-1|1-\alpha/2}}$$

or alternatively if

$$\boxed{\alpha^{\text{obs}} = 2\min\left(\mathbb{P}(T^0_{n-J-1|1-\alpha/2} > t_{\hat{\beta}_j}), \mathbb{P}(T^0_{n-J-1|1-\alpha/2} < t_{\hat{\beta}_j})\right)}$$

is *lower then* a certain threshold (e.g., 0.05 or 0.001).

Note that α^{obs} is usually called p-value.

Example
Math anxiety and test difficulty

In the math anxiety example, we have

$$\hat{\boldsymbol{\beta}} = [0.0214, 1.5899, -0.3949]^T$$
$$\sigma_{\hat{\boldsymbol{\beta}}} = [0.6147, 0.2299, 0.3273]^T$$

For $\alpha = 0.05$ we can test whether $\boxed{H_0 : \hat{\beta}_{\text{math anx}} = 0}$ against $\boxed{H_1 : \hat{\beta}_{\text{math anx}} \neq 0}$.

Then, the reference quantile† for the current test is $t^0_{15-3|1-0.05/2} = 2.18$ and

$$t_{\hat{\beta}_{\text{math anx}}} = 1.5899/0.2299 = 6.187$$

As $6.187 > 2.18$ we can conclude that H_0 is rejected for the current parameter.

† Quantiles of t-Student distribution can be computed using the `R` function `qt(1-α/2,n-J)`.

Example
Math anxiety and test difficulty

In the math anxiety example, we have

$$\hat{\boldsymbol{\beta}} = [0.0214, 1.5899, -0.3949]^T$$
$$\sigma_{\hat{\boldsymbol{\beta}}} = [0.6147, 0.2299, 0.3273]^T$$

For $\alpha = 0.05$ we can test whether $\boxed{H_0 : \hat{\beta}_{\text{math anx}} = 0}$ against $\boxed{H_1 : \hat{\beta}_{\text{math anx}} \neq 0}$.

Alternatively, $\alpha^{\text{obs}} = 2 \min\left(1, 2.328e^{-05}\right)^{\dagger} = 4.656e^{-05}$

As $4.656e^{-05}$ is smaller then the usual predefined threshold ($\alpha^0 = 0.05$) then we can conclude that H_0 is rejected for the current parameter.

† Probabilities for the t-Student distribution can be computed using the R function `pt(t,n-J-1,1-α/2)`.

Testing individual coefficients
Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Based on the pivotal quantity $t_{\hat{\beta}_j}$ we can also compute $1-\alpha$ **confidence intervals** (CIs) for a certain parameter $\hat{\beta}_j$. In this case, the symmetric interval such that

$$\mathbb{P}(T_{\hat{\beta}_j} \in [lb_{\hat{\beta}_j}, ub_{\hat{\beta}_j}]) = 1 - \alpha$$

is given by

$$\boxed{\hat{\beta}_j \pm t^0_{n-J-1|1-\alpha/2} \cdot \sigma_{\hat{\beta}_j}}$$

where the bounds of the interval are as follows

$$lb_{\hat{\beta}_j} = \hat{\beta}_j - t^0_{n-J-1|1-\alpha/2} \cdot \sigma_{\hat{\beta}_j}$$
$$ub_{\hat{\beta}_j} = \hat{\beta}_j + t^0_{n-J-1|1-\alpha/2} \cdot \sigma_{\hat{\beta}_j}$$

Example
Math anxiety and test difficulty

In the math anxiety example, the $1-0.05$ CIs for $\beta_{\texttt{mathanx}} = 1.5899$ is as follows:

$$lb_{\beta_{\texttt{mathanx}}} = 1.5899 - t^0_{12|1-0.05/2} \cdot 0.2299 = 1.095$$
$$ub_{\beta_{\texttt{mathanx}}} = 1.5899 + t^0_{12|1-0.05/2} \cdot 0.2299 = 2.085$$

As the interval does not contain zero, the null hypothesis $H_0 : \hat{\beta}_{\texttt{mathanx}} = 0$ can be rejected at the 5% level.

Moreover, the CI is relatively wide in the sense that the upper bound is larger than the lower bound: We are not really confident about what the exact effect of `math anx` is, even though the p-value for the current parameter is smaller then the predefined threshold.

Testing all the coefficients
Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Testing the null hypothesis

$$\boxed{H_0 : \hat{\boldsymbol{\beta}}_{J+1\times 1} = \mathbf{0}_{J+1\times 1}}$$

where all the regression coefficients are all zero simultaneously (**omnibus test**) can be performed by computing the **analysis of variance table** for *nested models*.

Testing all the coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

In practice, two submodels are defined

$$\mathcal{M}_0 : \mathbf{y} \sim \mathcal{N}(\mathbf{1}_n\beta_0, \mathbf{I}\sigma_0^2) \quad \text{null model}$$
$$\mathcal{M}_1 : \mathbf{y} \sim \mathcal{N}(\mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta}_1, \mathbf{I}\sigma_1^2) \quad \text{full model}$$

with $\mathcal{M}_0$ containing the intercept coefficient only (*null model*).

Testing all the coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Then, under the constrain $\mathcal{M}_0 \subset \mathcal{M}_1$, the **likelihood-ratio statistic**

$$W_J = 2\left(\ln \mathcal{L}_1(\hat{\beta}_1, \hat{\sigma}_1^2; \mathbf{y}) - \ln \mathcal{L}_0(\hat{\beta}_0, \hat{\sigma}_0^2; \mathbf{y})\right)$$
$$= \ldots \text{ after a little algebra we get}$$
$$= \frac{R^2}{1-R^2}\left(\frac{n-J-1}{J}\right)$$

is distributed according to a $\mathcal{F}$-distribution with $\text{df}_1 = J$ and $\text{df}_2 = n - J - 1$ degrees of freedom:

$$W_J \overset{\text{H}_0}{\sim} \mathcal{F}(; J, n-J-1)$$

As usual, large values of W_J allows for rejecting H_0.

Testing all the coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Note that rejecting H_0 does not imply that all the predictors are still not needed to explain the response variable. For instance, by adding or removing one of them the result of the $\mathcal{F}$-test may change. In this sense, the omnibus test can be considered as a starting point for further improvements of the model being tested.

Example

Math anxiety and test difficulty

In the math anxiety example, we get:

$$\mathcal{M}_1 : \texttt{RT}_i \sim \mathcal{N}(\beta_0 + \texttt{math anx}\beta_1 + \texttt{diff}\beta_2, \sigma^2)$$
$$\mathcal{M}_0 : \texttt{RT}_i \sim \mathcal{N}(\beta_0, \sigma^2)$$

and

$$W = \frac{0.883}{1-0.883}\left(\frac{15-3}{2}\right) = 47.70$$

Since the observed significance level (p-value) for W is $\alpha^{\text{obs}} = 1.945e^{-06}$ is lower then the predefined threshold (0.05), we reject H_0 at the 5% level.

Testing subsets of coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Likewise for the omnibus test, we can use the analysis of variance to test sub-models with more then one predictor. In this case, the null model does not simply contains the intercept coefficient:

$$\mathcal{M}_0 : \mathbf{y} \sim \mathcal{N}(\mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta}^\circ, \mathbf{I}\sigma_0^2) \quad \text{null model}$$
$$\mathcal{M}_1 : \mathbf{y} \sim \mathcal{N}(\mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta}^\dagger, \mathbf{I}\sigma_1^2) \quad \text{full model}$$

The term $\boldsymbol{\beta}^\circ$ indicates a subset of coefficients (i.e., a subset of variables) different by $\boldsymbol{\beta}^\dagger$.

Testing subsets of coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

To test

$$\boxed{H_0 : \boldsymbol{\beta}^\circ_{Q\times 1} = \mathbf{0}_{Q\times 1}}$$

against

$$\boxed{H_1 : \exists\beta_k \in \boldsymbol{\beta}^\dagger_{K\times 1} \text{ such that } \beta_k \neq 0}$$

since $Q < K$ and $\mathcal{M}_0 \subset \mathcal{M}_1$, we can use the statistic

$$W_J = 2\left(\ln \mathcal{L}_1(\hat{\boldsymbol{\beta}}_1, \hat{\sigma}_1^2; \mathbf{y}) - \ln \mathcal{L}_0(\hat{\beta}_0, \hat{\sigma}_0^2; \mathbf{y})\right)$$
$$= \ldots \text{ after a little algebra we get}$$
$$= \frac{(\mathbf{e}^\circ)^T(\mathbf{e}^\circ) - (\mathbf{e}^\dagger)^T(\mathbf{e}^\dagger)}{(\mathbf{e}^\dagger)^T(\mathbf{e}^\dagger)}\left(\frac{n-K-1}{K-Q}\right)$$

which is distributed according to the $\mathcal{F}$-distribution:

$$W_{K-Q} \overset{H_0}{\sim} \mathcal{F}(; K-Q, n-K-1)$$

Testing subsets of coefficients

Source: 6.2.2 (Fox, 2016); 3.2, 3.5 (Faraway, 2014)

Again, H_0 is rejected for large values of W_{K-Q}.

Note that the term $\mathbf{e}$ in the definition of W_{K-Q} is the *residual sum of squares*, which is defined as usual:

$$\mathbf{e}^\circ = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^\circ$$
$$\mathbf{e}^\dagger = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^\dagger$$

The analysis of variance can also be used to test whether *adding* (or *removing*) a single coefficient from a null (or full model) increases the overall fit of the model being tested (**incremental test**). We will directly see this approach in the practical sessions of the course.

Example

Math anxiety and test difficulty

In the math anxiety example, we get as follows:

$$\mathcal{M}_1 : \texttt{RT}_i \sim \mathcal{N}(\beta_0 + \texttt{math anx}\beta_1 + \texttt{diff}\beta_2, \sigma^2)$$
$$\mathcal{M}_0 : \texttt{RT}_i \sim \mathcal{N}(\beta_0 + \texttt{math anx}\beta_1, \sigma^2)$$

and

$$W_{2-1} = \frac{11.91 - 10.86}{10.86}\left(\frac{15-2-1}{2-1}\right) = 1.16$$

Since the observed significance level (p-value) for W_{2-1} is $\alpha^{\text{obs}} = 0.3$ is higher then the predefined threshold (0.05), we canont reject H_0 at the 5% level. Therefore, adding the variable `diff` does not increase the fit of the model for `RT`.

Testing non-nested models

Source: 22.1.1 (Fox, 2016); 10.3 (Faraway, 2014)

More generally, in some cases (e.g., non-nested models) model selection can be performed without the use of inferential tests, which may otherwise be biased if applied. Two of the most common indices are the Akaike information criterion (**AIC**) and the Bayesian information criterion (**BIC**):

$$\text{AIC}(\hat{\boldsymbol{\beta}}) = -2 \ln \mathcal{L}(\hat{\boldsymbol{\beta}}; \mathbf{y}) + 2J$$
$$\text{BIC}(\hat{\boldsymbol{\beta}}) = -2 \ln \mathcal{L}(\hat{\boldsymbol{\beta}}; \mathbf{y}) + (J+2) \ln n$$

They both penalize the maximum likelihood of the estimated models by adding a constant depending on the number of variables being included in the model. Given a set of candidates, the best model is that minimizing AIC or BIC.

We will see how AIC and BIC may be used during the practical sessions of the course.

Confidence intervals for $\boldsymbol{\mu}$

Once $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\}$ has been estimated, **confidence bands** for the linear predictor $\hat{\boldsymbol{\mu}}$ can also be computed. They allow for assessing the uncertainty of the *predictions* based on the current explanatory variables $\mathbf{X}$.

As for confidence intervals for $\boldsymbol{\theta}$, confidence bands for a fixed $1-\alpha$ level are as follows:

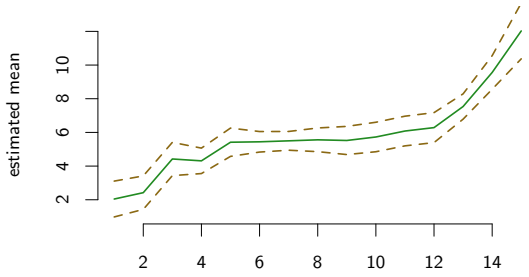
$$\boxed{\hat{\mathbf{y}} \pm t^0_{n-J-1|1-\alpha/2} \cdot \sqrt{\text{diag}\left(\mathbf{X}\left((\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2\right)\mathbf{X}^T\right)}}$$

where $\sqrt{\text{diag}(\mathbf{X}\left((\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2\right)\mathbf{X}^T)} = \sigma_{\hat{\boldsymbol{\mu}}}$ is the standard deviation for the mean $\hat{\boldsymbol{\mu}}$, $t^0_{n-J-1|1-\alpha/2}$ is the reference quantile for the t-distribution, and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Similarly, confidence bands can be computed for prediction of $\hat{\boldsymbol{\mu}}$ given a **new set of observations** $\mathbf{X}^{\text{new}}$ instead of using the current $\mathbf{X}$. We will see this topic in the practical sessions of the course.

Example

Math anxiety and test difficulty



Confidence bands for $\hat{\boldsymbol{\mu}}$ at the 95% level (dashed gold curves) and estimated means (straight green line).

Outline

Diagnostics and model evaluation

Source: 11-13 (Fox, 2016); 6 (Faraway, 2014)

Once $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\}$ has been estimated, diagnostics should be performed before using the estimated model for research purposes (e.g., testing experimental hypothesis, clinical interpretations, prediction).

Four main issues need to be checked before any use of the estimated model:

- Normality of the response variable $\mathbf{y}$ (or equivalently Normality of residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$)
- Homoscedasticity $\mathbb{V}\text{ar}[\mathbf{y}] = \sigma^2$
- The linear predictor $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ needs to be correctly specified
- Absence of influential observations or outliers

Residuals

Given $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\}$, the residuals $\mathbf{e}$ of a Normal linear model are defined as:

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \qquad \text{note that } \mathbf{H}_{n\times n} \text{ is the } \textit{hat matrix} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

With the following expectations:

$$\begin{aligned}\mathbb{E}[\mathbf{e}] &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= \mathbb{E}\left[\mathbf{y} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \mathbb{E}[\mathbf{y}] - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{y}] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\,\cancel{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}\boldsymbol{\beta} \\ &= 0\end{aligned}$$

$$\begin{aligned}\mathbb{V}\text{ar}[\mathbf{e}] &= \mathbb{V}\text{ar}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I} - \mathbf{H})\mathbb{V}\text{ar}[\mathbf{y}]\,(\mathbf{I} - \mathbf{H})^T \\ &= (\mathbf{I} - \mathbf{H})\mathbf{I}\sigma^2(\mathbf{I} - \mathbf{H})^T \\ &= (\mathbf{I} - \mathbf{H})\sigma^2\end{aligned}$$

Residuals

Three types of residuals can then be defined:

- **raw**: $e_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$
- **standardized**: $\tilde{e}_i = \frac{e_i}{\sqrt{1-h_{ii}}} \sim \mathcal{N}(0, \sigma^2)$
- **stundentized**: $r_i = \frac{e_i}{\hat{\sigma}^2\sqrt{1-h_{ii}}} \dot{\sim} \mathcal{N}(0, 1)$

The analysis of residuals consist in contrasting the residuals of the fitted model with that expected by their theoretical model.

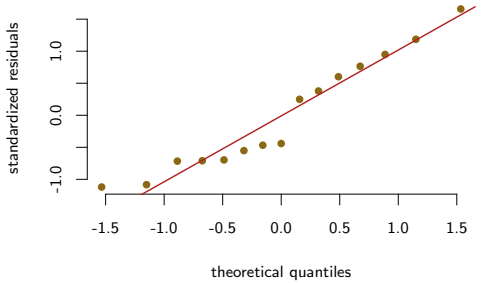
Checking Normality of residuals

Source: 12.1 (Fox, 2016); 6.1.2 (Faraway, 2014)

A quick graphical check for the Normality of residuals consist in plotting quantiles from standardized residuals and Normal distribution (**qq-plot**).

Checking Normality of residuals

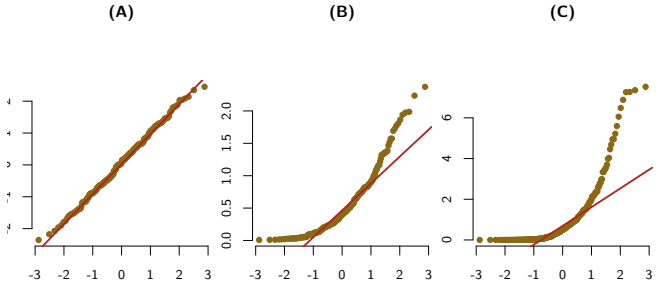
Source: 12.1 (Fox, 2016); 6.1.2 (Faraway, 2014)



qq-plot for the math anxiety example. In case of Normality, standardized residuals should follow the straight line approximately.

Checking Normality of residuals

Source: 12.1 (Fox, 2016); 6.1.2 (Faraway, 2014)



Examples of qq-plots: **(A)** Normal residuals and **(B)**-**(C)** residuals from skewed distributions.

Checking Normality of residuals

Source: 12.1 (Fox, 2016); 6.1.2 (Faraway, 2014)

Another way to check the Normality of residuals is the **Shapiro-Wilk test**. The null hypothesis H_0 is that standardized residuals follows the Normal distribution. Large values for the test statistic W indicate that null hypothesis should be rejected.

In case of non-Normal residuals, one may try transforming the response variable (e.g., Box-Cox transformation), using robust estimators for the model parameter, or changing the modeling approach from Normal linear models to Generalized linear models (second part of the course).

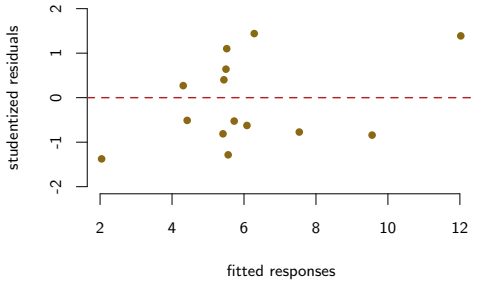
Checking Homoscedasticity

Source: 12.2 (Fox, 2016); 6.1.1 (Faraway, 2014)

Non constant variance for residuals can be checked graphically by inspecting the studentized residuals $\mathbf{r}$ as a function of the fitted values $\hat{\mathbf{y}}$. If the variance is constant, then no systematic pattern should be noted (residuals do not vary as a function of fitted values). By contrast, in case of heteroscedasticity residuals should varies as a function of fitted values and systematic patterns should be noted in the scatter plot.

Checking Homoscedasticity

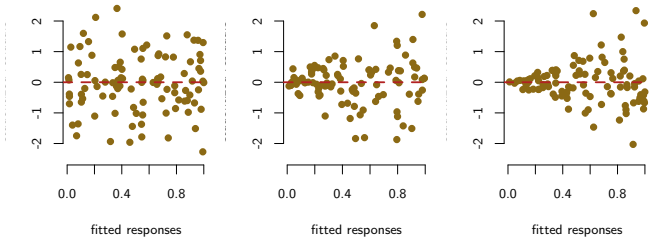
Source: 12.2 (Fox, 2016); 6.1.1 (Faraway, 2014)



Homoscedasticity check for the math anxiety example. In case of homoscedasticity, no systematic variation should be noted in the scatterplot (there should be approximately constant variation).

Checking Homoscedasticity

Source: 12.2 (Fox, 2016); 6.1.1 (Faraway, 2014)



Examples of **(A)** homoscedasticity, **(B)** mild heteroscedasticity, and **(C)** strong heteroscedasticity.

Checking Homoscedasticity

Source: 12.2 (Fox, 2016); 6.1.1 (Faraway, 2014)

Another way to check for the Homoscedasticity is the **Bartlett test** (for grouped data) or the **Breusch-Pagan test** (when single observations are available, i.e. ungrouped data). The null hypothesis H_0 is that residuals do not vary as a function of the explanatory variables. Large values for these test statistics indicate that null hypothesis should be rejected.

In case of heteroscedasticity, one may try transforming the response variable by using any **variance stabilizing transformation** (e.g., square root, logarithm, arcsin). Alternatively, in the case of non constant variance, maximum likelihood (or ordinary least squares) based estimators for $\boldsymbol{\beta}$ and σ^2 can be derived using the **weighted least squares method**.

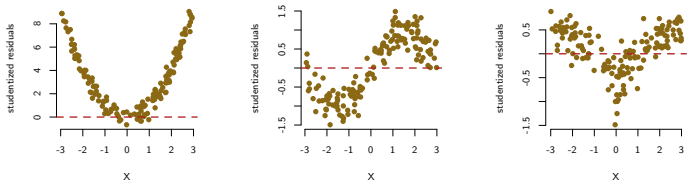
Checking the structural part of the model

Source: 12.3.1 (Fox, 2016); 6.3 (Faraway, 2014)

Plotting the studentized residuals **r** as a function of the explanatory variable X can help in identifying whether the linear part of the model $\mathbb{E}[Y]_i = \beta_0 + x_i\beta_1$ holds.

Checking the structural part of the model

Source: 12.3.1 (Fox, 2016); 6.3 (Faraway, 2014)



Examples of non-linearity w.r.t. an exploratory variable X.

Checking the structural part of the model

Source: 12.3.1 (Fox, 2016); 6.3 (Faraway, 2014)

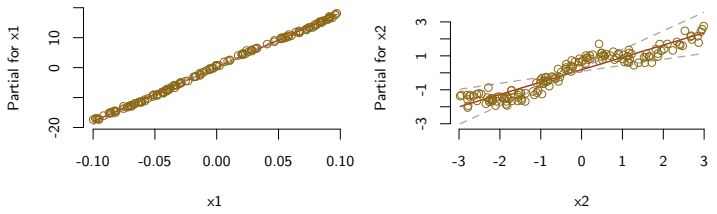
When $J \geq 2$ **partial regression plots** (Velleman & Welsch, 1981) should instead be preferred to find nonlinearities among Y_i and the predictors $X_1, \ldots, X_J$.

As an example, consider the simplest case where $J = 2$. Then, the partial plot for X_2 can be produced by applying the following procedure:

1. compute the residuals $\mathbf{e}_1 = \mathbf{y} - (\hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1)$ by taking the effect of $\mathbf{x}_2$ out
2. compute the residuals $\mathbf{e}_2 = \mathbf{x}_2 - (\hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1)$, which allows for computing the effect of $\mathbf{x}_2$ by taking out the effect of $\mathbf{x}_1$
3. plot $\mathbf{e}_1$ against $\mathbf{e}_2$ and look for nonlinearity

Checking the structural part of the model

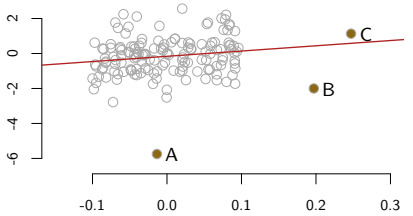
Source: 12.3.1 (Fox, 2016); 6.3 (Faraway, 2014)



Example of partial regression plots for a Normal linear model with $J = 2$. We can notice that the variable X_2 enters the model nonlinearly as opposed to X_1.

Looking for unusual observations

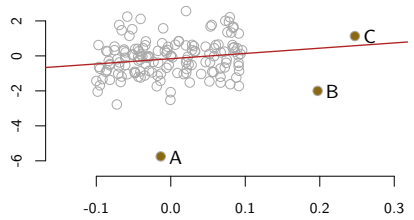
Source: 11 (Fox, 2016); 6.2 (Faraway, 2014)



Unusual observations consist in data that may change the overall fit of the model.

Looking for unusual observations

Source: 11 (Fox, 2016); 6.2 (Faraway, 2014)



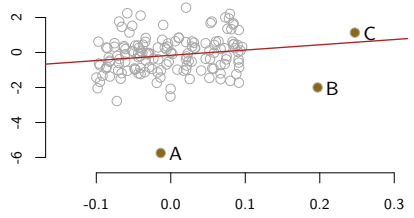
outliers: Observations that are far away from the estimated linear line ($J = 1$) or hyperplane ($J > 1$). They are points that do not fit the model very well (A, B).

leverage points: Observations that are far away from the estimated linear line or hyperplane conditionally on the exploratory variables (B, C).

influential points: Observations that change the fit of the model substantively (B).

Looking for unusual observations

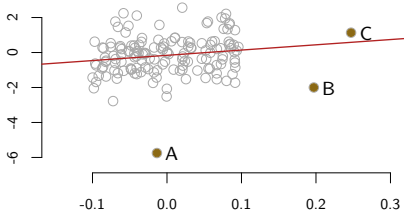
Source: 11 (Fox, 2016); 6.2 (Faraway, 2014)



By contrast, leverage points that cannot be classified as outliers (e.g., C) impact the model by increasing the overall fit only, for instance by increasing the R^2 index.

Looking for unusual observations

Source: 11 (Fox, 2016); 6.2 (Faraway, 2014)



Generally, observations that are outliers and leverage points simultaneously (e.g., B) can have a substantial influence on the regression coefficients. For instance, they can change the magnitude or the sign of $\hat{\beta}$.

Identifying leverage points

Source: 11.2 (Fox, 2016); 6.2.1 (Faraway, 2014)

To identity leverage points, *leverages* can be computed once the model has been estimated. The diagonal of the $n \times n$ **hat matrix H**

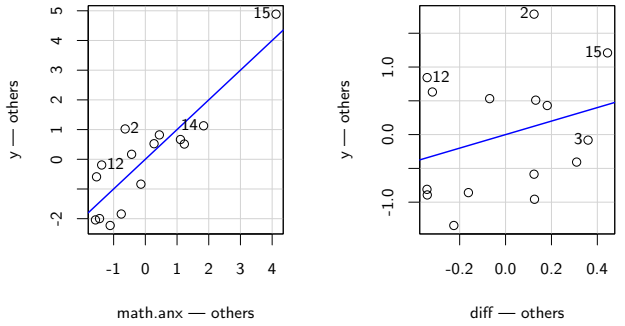
$$\begin{aligned} \mathbf{h} &= \text{diag}\,(\mathbf{H}) \\ &= \text{diag}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) \end{aligned}$$

is used to compute leverage values for each observation. As leverages are related to the variance of the residuals through the formula $\sigma^2_{\mathbf{e}} = \text{diag}\,(\mathbf{I} - \mathbf{H})\,\hat{\sigma}^2$, a large leverage h_i will make $\sigma^2_{e_i}$ very small and the fit $\hat{y}_i$ will be attracted toward the observation y_i.

Leverages larger three or more times the average hat-value $\overline{h}$ should be checked carefully.

Identifying leverage points

Source: 11.2 (Fox, 2016); 6.2.1 (Faraway, 2014)



Partial regression plots for the math anxiety example with *suspected* leverage points.

Identifying outliers

Source: 11.3 (Fox, 2016); 6.2.2 (Faraway, 2014)

To identify outliers from the set of observations, we can quickly inspect the (absolute value of the) **studentized residuals** $|\mathbf{r}|$ of the fitted model and look for those observations showing larger residuals.

As $r_i \sim t(; n-J-2)$, a two-sided Bonferroni adjusted p-value can be computed under the null hypothesis that r_i *is not* an outlier point:

$$\alpha_{\text{adj}}^{\text{obs}} = 1 - \mathbb{P}(T_{n-J-2|1-\alpha/2}^{0} > |r_i|)2n$$

If $\alpha_{\text{adj}}^{\text{obs}}$ is lower then a predefined threshold (e.g., 0.05), then H_0 is rejected and y_i has to be considered as outlier.

Identifying influential points

Source: 11.4 (Fox, 2016); 6.2.3 (Faraway, 2014)

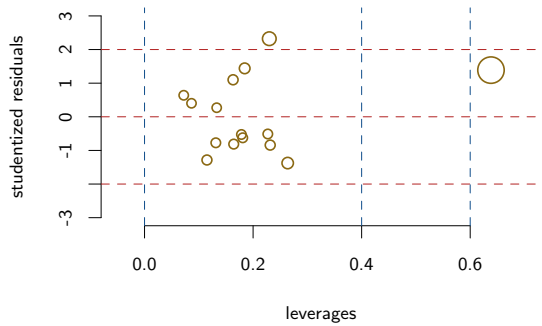
An influential point is one whose removal from the dataset would cause a large change in the fit. The influence of an observation can be summarized in terms of **Cook's distance**, which relates *studentized residuals* **r** to *leverages* **h**:

$$d_i = \frac{r_i^2}{J+1} \frac{h_i}{1-h_i}$$

The formula reflects the fact that observations with larger leverage are more likely to exert a substantive influence on the regression coefficients $\hat{\beta}$.

Identifying influential points

Source: 11.4 (Fox, 2016); 6.2.3 (Faraway, 2014)

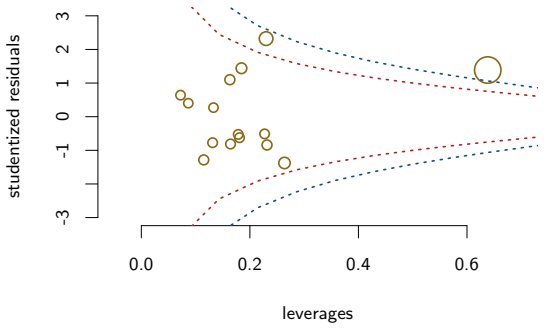


Influential plot for the math anxiety example. Note that the size of each point is proportional to Cook's distances, horizontal reference lines (in blue color) are drawn at studentized residuals of 0 and ± 2, vertical reference lines (in red color) are drawn at hat-values of $2\bar{h}$ and $3\bar{h}$.

Points with larger hat-value as well as leverage should carefully be checked.

Identifying influential points

Source: 11.4 (Fox, 2016); 6.2.3 (Faraway, 2014)



Influential plot for the math anxiety example. Note that the size of each point is proportional to Cook's distances whereas blue and red curves are the contour lines ($l_{d_0=0.5}$ and $l_{d_0=1}$, respectively) of the Cook's distance. Any point that lies beyond these contours might well be influential and require closer attention.

For a fixed d_0, they are obtained as follows: $l_{d_0} = \pm\sqrt{d_0(J+1)\frac{(1-h)}{h}}$.

Identifying influential points

Source: 11.4 (Fox, 2016); 6.2.3 (Faraway, 2014)

Influential observations can seriously affect the estimation of regression coefficients $\hat{\boldsymbol{\beta}}$.

Given an influential observation i, a way to quantify such an influence is evaluating the difference between regression coefficients computed including the point ($\hat{\boldsymbol{\beta}}$) and those obtained by excluding that point ($\hat{\boldsymbol{\beta}}_{-i}$):

$$\text{diff}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{1}_n\mathbf{x}_i)(y_i - \hat{y}_i)$$

where $\mathbf{x}_i$ is the $1 \times J$ vector containing the i-th influential observation.

The larger the quantity $\text{diff}_{\hat{\boldsymbol{\beta}}}$, the more attention should be paid to fitted coefficients.

Handling with unusual observations

Source: 11.7 (Fox, 2016)

Once unusual observations have been identified, one may ask what to do further.

- Unusual observations can be caused by errors during the sampling process. In this case, it may be legitimate to remove them from the data.
- It may be the case that the measurement process is incorrect for the unusual observations (e.g., different experimental conditions). In this case, removing those observations may not be legitimate. On the contrary, two analyses may be instead run by including and excluding the unusual observations. Then, appropriate comments may be made about that.
- It may also happen that unusual observations are extreme but still plausible realizations of the (random) sampling process. As before, removing those observations may be unfair. In this particular case, some solutions include using robust estimators for the regression coefficients or using linear models without the assumption of Gaussianity (second part of the course).

In general, removing unusual observations should be done carefully. Finally, in large samples, unusual data substantially alter the results only in extreme instances.

Outline

Categorical predictors
Source: 7.1, 7.2 (Fox, 2016)

Let the random realizations $\mathbf{y} \in \mathbb{R}^n$ be modeled as $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$. When $\mathbf{X}$ is categorical we need to appropriately codify the information of the independent variables in order to get meaningful regression coefficients.

Examples of categorical predictors in linear models include gender, occupation, experimental or treatment groups.

Dummy coding is a common way to represent categorical information in statistical models. Other representation include *Helmert* coding, *treatment* coding, and *sum* coding.

Categorical predictors: Dummy coding
Source: 7.1, 7.2 (Fox, 2016)

Let D be a categorical variable with K levels taking values in the set $\{d_1, \ldots, d_K\}$. The dummy coding transforms D into a set of $K-1$ Boolean variables $(Z_1, \ldots, Z_{K-1})$, with $Z_k \in \{0, 1\}$.

To dummify D we need as many boolean variables as the levels of D minus one.

Two examples are as follows:

Y	D	Z
y_1	M	0
y_2	M	0
y_3	M	0
y_4	F	1
y_5	F	1
y_6	F	1
$\vdots$	$\vdots$	$\vdots$

Y	D	Z_1	Z_2
y_1	G1	0	0
y_2	G1	0	0
y_3	G2	1	0
y_4	G2	1	0
y_5	G3	0	1
y_6	G3	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Categorical predictors: Dummy coding
Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

In this case, as $K = 3$ we need $K - 1 = 2$ dummy variables as follows:

$$d_i = \begin{cases} A, & z_i^{(1)} = 0 \wedge z_i^{(2)} = 0 \\ B, & z_i^{(1)} = 1 \wedge z_i^{(2)} = 0 \\ C, & z_i^{(1)} = 0 \wedge z_i^{(2)} = 1 \end{cases}$$

Categorical predictors: Dummy coding
Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

Rewriting the linear model with dummy coding leads to:

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(\mathbf{1}\boldsymbol{\beta}_0 + \mathbf{z}^{(1)}\boldsymbol{\beta}_1 + \mathbf{z}^{(2)}\boldsymbol{\beta}_2, \sigma^2\mathbf{I}) \\ &\sim \mathcal{N}(\mathbf{1}\boldsymbol{\beta}_0 + \mathbf{Z}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \end{aligned}$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

In matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 1 & \dots & z_1^{(1)} & \dots & z_1^{(2)} \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & z_i^{(1)} & \dots & z_i^{(2)} \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & z_n^{(1)} & \dots & z_n^{(2)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_1 \\ \vdots \\ \beta_2 \end{bmatrix}}_{\boldsymbol{\mu}=\mathbf{Z}\boldsymbol{\beta}}, \underbrace{\begin{bmatrix} 1 & \dots & \dots \\ \vdots & & \vdots \\ \dots & 1 & \dots \\ \vdots & & \vdots \\ \dots & \dots & 1 \end{bmatrix} \sigma^2}_{\boldsymbol{\Sigma}=\sigma^2\mathbf{I}} \right)$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

In this case, as $K = 3$ we need $K - 1 = 2$ dummy variables as follows:

$$d_i = \begin{cases} A, & z_i^{(1)} = 0 \wedge z_i^{(2)} = 0 \\ B, & z_i^{(1)} = 1 \wedge z_i^{(2)} = 0 \\ C, & z_i^{(1)} = 0 \wedge z_i^{(2)} = 1 \end{cases}$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

Rewriting the linear model with dummy coding leads to:

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{z}^{(1)}\beta_1 + \mathbf{z}^{(2)}\beta_2, \sigma^2\mathbf{I}) \\ &\sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{Z}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \end{aligned}$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

When used in Normal linear and generalized linear models, dummy coding generates a set of nested models.

To see this point, consider the case of a simple model $\mathbf{y} \sim \mathcal{N}(\beta_0 + \mathbf{x}\beta_1, \sigma^2)$, with $\mathbf{x} = \mathbf{d}_{n\times 1}$ being a categorical variables with three distinct levels, $d_i \in \{A, B, C\}$.

In matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 1 & \dots & z_1^{(1)} & \dots & z_1^{(2)} \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & z_i^{(1)} & \dots & z_i^{(2)} \\ \vdots & & \vdots & & \vdots \\ 1 & \dots & z_n^{(1)} & \dots & z_n^{(2)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_1 \\ \vdots \\ \beta_2 \end{bmatrix}}_{\boldsymbol{\mu}=\mathbf{Z}\boldsymbol{\beta}}, \underbrace{\begin{bmatrix} 1 & \dots & \dots \\ \vdots & & \vdots \\ \dots & 1 & \dots \\ \vdots & & \vdots \\ \dots & \dots & 1 \end{bmatrix} \sigma^2}_{\boldsymbol{\Sigma}=\sigma^2\mathbf{I}} \right)$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

As $z_i^{(1)} \in \{0, 1\}$ and $z_i^{(2)} \in \{0, 1\}$, we can note that

$$
\begin{aligned}
M_0 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\beta_0, \sigma^2) && \text{when} \quad z_i^{(1)} = 0 \wedge z_i^{(2)} = 0 \\
M_1 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2) && \text{when} \quad z_i^{(1)} = 1 \wedge z_i^{(2)} = 0 \\
M_2 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\beta_0 + \beta_2, \sigma^2) && \text{when} \quad z_i^{(1)} = 0 \wedge z_i^{(2)} = 1
\end{aligned}
$$

The models are nested w.r.t. the regression coefficients $\boldsymbol{\beta}$, i.e. $(M_0 \subset M_1 \subset M_2)$.

Note: as we are modeling the means of $\mathbf{y}$, there are no slopes in the models except intercepts.

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

Estimating and testing $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ in linear models with dummy coding is the same as for continuous predictors (just replace $\mathbf{X}$ with $\mathbf{Z}$).

Regression coefficients can be estimated in terms of mean.

In case of $K - 1 = 1$ (*one* dummy variable), we have:

- β_0 indicates the reference level which refers to the case $Z = 0$
- β_1 indicates the incremental/decremental quantity at $Z = 1$ from $Z = 0$

and

- $\mathbb{E}[\mathbf{y}|Z = 0] = \beta_0$
- $\mathbb{E}[\mathbf{y}|Z = 1] = \beta_0 + \beta_1$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

In case of $K - 1 = 2$ (*two* dummy variables), we instead have:

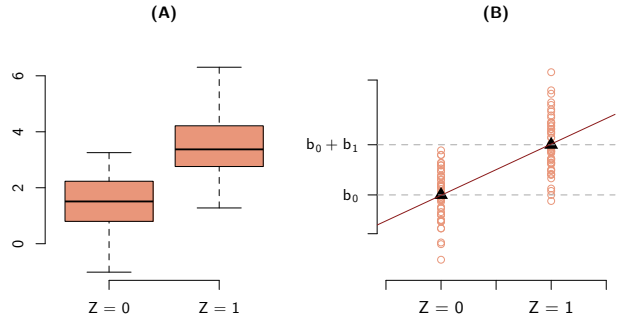
- β_0 indicates the reference level which refers to the case $Z_1 = 0$ and $Z_2 = 0$
- β_1 indicates the incremental/decremental quantity at $Z_1 = 1$ from $Z_1 = 0$ and $Z_2 = 0$
- β_2 indicates the incremental/decremental quantity at $Z_2 = 1$ from $Z_1 = 0$ and $Z_2 = 0$

and

- $\mathbb{E}[\mathbf{y}|Z_1 = 0, Z_2 = 0] = \beta_0$
- $\mathbb{E}[\mathbf{y}|Z_1 = 1, Z_2 = 0] = \beta_0 + \beta_1$
- $\mathbb{E}[\mathbf{y}|Z_1 = 0, Z_2 = 1] = \beta_0 + \beta_2$

Categorical predictors: Dummy coding

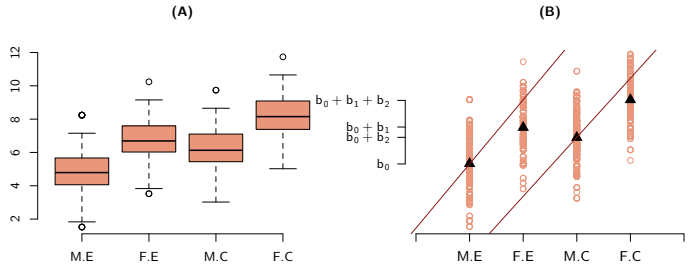
Source: 7.1, 7.2 (Fox, 2016)



Example of a Normal linear model with one dummy variable Z.
(A) Box-plot for the response variable as a function of the dummy levels.
(B) Estimated means (dotted gray lines with black triangles) and regression line (red straight line) as a function of the dummy levels.

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)



Example of a Normal linear model with two dummy variable $Z_1 \in \{E, C\}$ and $Z_2 \in \{M, F\}$.
(A) Box-plot for the response variable as a function of the dummy levels.
(B) Estimated means (dotted gray lines with black triangles) and regression lines (red straight lines) as a function of the dummy levels.

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

Dummy-coding can also be used when the matrix of data contains both categorical and continuous variables.

Let $\mathbf{X}_{n\times J}$ be a matrix of real data and $\mathbf{Z}_{n\times K-1}$ a matrix of dummy variables (with K being the number of categorical variables). Then,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta_X + \mathbf{Z}\beta_Z, \mathbf{I}\sigma^2)$$

is a Normal linear model containing both continuous and categorical variables.

With no loss of generality, we can rewrite the model as

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}^*\beta, \mathbf{I}\sigma^2)$$

where $\mathbf{X}^* = [\mathbf{X}|\mathbf{Z}]$ is the $n \times (J + K - 1)$ staked matrix obtained by juxtaposing $\mathbf{X}$ and $\mathbf{Z}$.

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

In matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 1 & x_{11} & \dots & x_{1J} & | & \dots & z_{1,K-1} \\ \vdots & & & \vdots & & & \vdots \\ 1 & x_{i1} & \dots & x_{iJ} & | & \dots & z_{i,K-1} \\ \vdots & & & \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{nJ} & | & \dots & z_{n,K-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_J \\ \vdots \\ \beta_{J+K-1} \end{bmatrix}}_{\mu = \mathbf{X}^*\ \beta}, \underbrace{\begin{bmatrix} 1 & \dots & \dots \\ \vdots & & \vdots \\ \dots & 1 & \dots \\ \vdots & & \vdots \\ \dots & \dots & 1 \end{bmatrix} \sigma^2}_{\Sigma = \sigma^2 \mathbf{I}} \right)$$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

Consider the simplest case where $J = 1$ and $K = 3$, then we have two dummy variables $z_i^{(1)} \in \{0, 1\}$ and $z_i^{(2)} \in \{0, 1\}$ and a single continuous independent variable $\mathbf{x}$. We then get:

$$\begin{aligned}
M_0 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}\beta_1, \sigma^2) && \text{when} \quad z_i^{(1)} = 0 \wedge z_i^{(2)} = 0 \\
M_1 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}\beta_1 + \mathbf{1}\beta_2, \sigma^2) && \text{when} \quad z_i^{(1)} = 1 \wedge z_i^{(2)} = 0 \\
M_2 &\stackrel{\text{def}}{=} y_i \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}\beta_1 + \mathbf{1}\beta_3, \sigma^2) && \text{when} \quad z_i^{(1)} = 0 \wedge z_i^{(2)} = 1
\end{aligned}$$

Also in this case, dummy coding generates a set of **nested linear equations** w.r.t. parameters ($M_0 \subset M_1 \subset M_2$).

Note: we now have slopes in the models (i.e., β_1). The parameters (β_2, β_3) of dummy variables can be aggregated with the intercept β_0.

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)

Estimating and testing $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ in linear models with dummy coding is the same as for continuous predictors (just replace $\mathbf{X}$ with $\mathbf{X}^*$).

In this case with *two* dummy variables and *one* continuous predictor, we have:

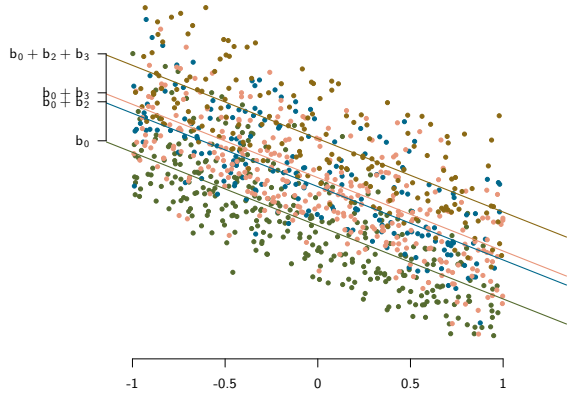
- β_0 indicates the intercept of the model
- β_1 indicates the slope of the model
- β_2 indicates the increment/decrement of the intercept when $Z_1 = 1$ and $Z_2 = 0$
- β_3 indicates the increment/decrement of the intercept when $Z_1 = 0$ and $Z_2 = 1$

and

- $\mathbb{E}[\mathbf{y}|Z_1 = 0, Z_2 = 0] = \beta_0 + \mathbf{x}\beta_1$
- $\mathbb{E}[\mathbf{y}|Z_1 = 1, Z_2 = 0] = (\beta_0 + \beta_2) + \mathbf{x}\beta_1$
- $\mathbb{E}[\mathbf{y}|Z_1 = 0, Z_2 = 1] = (\beta_0 + \beta_3) + \mathbf{x}\beta_1$

Categorical predictors: Dummy coding

Source: 7.1, 7.2 (Fox, 2016)



Example of a Normal linear model with two dummy variable $Z_1 \in \{E, C\}$ and $Z_2 \in \{M, F\}$ and one continuous predictor. Note that dummy levels are represented with different colors whereas the real variable is on the x-axis.

Modeling interactions

Source: 7.3 (Fox, 2016)

Consider a continuous variable X_1, X_2 along with a categorical variable with two levels $D = \{A, B\}$. Then, the product term

$$(X_1 \cdot X_2)$$

is the interaction between two continuous variables whereas

$$(D \cdot X_1) \quad \text{or} \quad (D \cdot X_2)$$

are the interactions between continuous and categorical variables.

Modeling interactions

Source: 7.3 (Fox, 2016)

When the terms $(X_1 \cdot X_2)$ and $(X_j \cdot D), j = 1, 2$ are used as **predictors** of a dependent variable Y, we say that

- X_1 predicts Y as a function of X_2 (continuous-continuous interaction)
- Y differs over the levels of D as a function of X_j (categorical-continuous interaction)

Modeling interactions
Source: 7.3 (Fox, 2016)

Consider the simplest case with $J = 1$ and $K = 2$. Then we have:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}_1\beta_1 + \mathbf{z}\beta_2 + \mathbf{x}_1 \circ \mathbf{z}\beta_3, \mathbf{I}\sigma^2)$$
$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}^\dagger\boldsymbol{\beta}, \mathbf{I}\sigma^2)$$

Note: $\circ$ is the element-wise product between two vectors.

Modeling interactions
Source: 7.3 (Fox, 2016)

In matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 1 & x_1^{(1)} & z_1 & z \cdot x_1^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_i^{(1)} & z_i & z \cdot x_i^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & z_n & z \cdot x_n^{(1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_1 \\ \vdots \\ \beta_2 \\ \vdots \\ \beta_3 \end{bmatrix}}_{\mu = \mathbf{X}^\dagger\boldsymbol{\beta}}, \underbrace{\begin{bmatrix} 1 & \dots & \dots \\ \vdots & & \vdots \\ \dots & 1 & \dots \\ \vdots & & \vdots \\ \dots & \dots & 1 \end{bmatrix} \sigma^2}_{\boldsymbol{\Sigma} = \sigma^2\mathbf{I}} \right)$$

Modeling interactions
Source: 7.3 (Fox, 2016)

When $J = 1$ and $K = 2$, then we have one dummy variable $z_i \in \{0, 1\}$ and two continuous independent variable $\mathbf{x}_1$, $\mathbf{x}_2$. We then get:

$$\begin{aligned} M_0 &\stackrel{\text{def}}{=} \mathbf{y} \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}_1\beta_1) && \text{when} \quad Z = 0 \\ M_1 &\stackrel{\text{def}}{=} \mathbf{y} \sim \mathcal{N}(\mathbf{1}\beta_0 + \mathbf{x}_1\beta_1 + \mathbf{1}\beta_2 + \mathbf{1}\beta_3 && \text{when} \quad Z = 1 \\ &\dots \mathbf{y} \sim \mathcal{N}(\mathbf{1}(\beta_0 + \beta_2) + \mathbf{x}_1(\beta_1 + \beta_3)) \end{aligned}$$

Also in this case, dummy coding generates a set of **nested linear equations** w.r.t. parameters ($M_0 \subset M_1 \subset M_2$).

Note:

- the effect of the categorical variable is included in the intercept $\mathbf{1}(\beta_0 + \beta_2)$
- the effect of the interaction variable is included in the slope $\mathbf{x}(\beta_1 + \beta_3)$

Modeling interactions
Source: 7.3 (Fox, 2016)

Estimating and testing $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ in linear models with dummy coding is the same as for continuous predictors (just replace $\mathbf{X}$ with $\mathbf{X}^\dagger$).

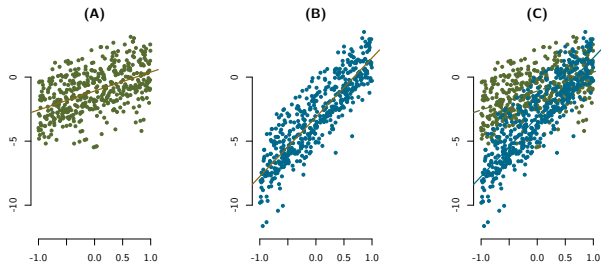
In this case with *one* dummy variable and *two* continuous predictors, we have:

- β_0 indicates the intercept of the model when $Z = 0$ (marginal effect of Z)
- β_2 indicates the slope when $Z = 0$ (marginal effect of X_1)
- $\beta_0 + \beta_1$ indicates the intercept of the model when $Z = 1$ (marginal effect of Z)
- $\beta_1 + \beta_3$ indicates the increment/decrement of the slope when $Z = 1$ (interaction effect)

and

- $\mathbb{E}[\mathbf{y}|Z = 0] = \beta_0 + \mathbf{x}\beta_1$
- $\mathbb{E}[\mathbf{y}|Z = 1] = (\beta_0 + \beta_2) + \mathbf{x}(\beta_1 + \beta_3)$

Modeling interactions
Source: 7.3 (Fox, 2016)



Example of a Normal linear model with an interaction between a dummy variable $Z \in \{0, 1\}$ and a continuous variable X. Note that dummy levels are represented with different colors whereas the real variable is on the x-axis. **(A)** Marginal plot of X for $Z = 0$; **(B)** Marginal plot of X for $Z = 1$; **(C)** Interaction plot for $X \cdot Z$.

Modeling interactions
Source: 7.3 (Fox, 2016)

Modeling **high-order interactions** ($J > 1$ and $K > 2$) is performed using the same rationale which applies for the simplest case. In this case, the model should also include all the low-order terms (**principle of marginality**) and the incremental F-test (slide 37, module B) is used to choose which of the terms should be retained in the final model.

However, with high-order interactions caution should be taken in interpreting the regression coefficients. Indeed, as interaction terms increase, the model complexity increases as well.

Outline

An illustrative example
Introduction

Background: Hear rate variability (HRV) is a measure regarding the modulation of the heart. Scientific findings have shown an important relation between sports performance and HRV. In general, athletes with great performance show increased HRV level and higher levels of pre-competitive anxiety can lead to impaired HRV during sport competitions.

Variables: HRV (as measured by the RRMSD), competitive anxiety (as measured by CSAI-2R), body fat percent (BF), international point score in 0-1000 (IPS: The closer the score is to 1000, the better the athlete's performance).

Goal: Define and fit a Normal linear model in order to predict HRV as a function of CSAI, BF, and IPS.

Source: Fortes, L. S., da Costa, B. D., Paes, P. P., do Nascimento Júnior, J. R., Fiorese, L., & Ferreira, M. E. (2017). Influence of competitive-anxiety on heart rate variability in swimmers. *Journal of sports science & medicine*, 16(4), 498.

An illustrative example

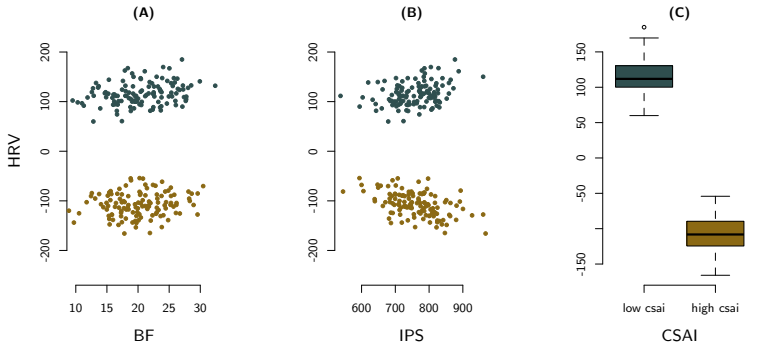
Data and descriptive analyses

	HRV	BF	IPS	CSAI (cat)
1	101.47	27.66	723.10	0
2	-153.45	19.65	821.96	1
3	-69.03	18.05	701.34	1
4	-88.95	21.19	723.41	1
5	-69.53	22.80	656.71	1
6	106.10	20.01	852.12	0
7	118.02	25.35	774.47	0
8	-96.01	17.09	699.36	1
9	139.44	21.87	648.08	0
10	-104.36	26.18	773.47	1
⋮	⋮	⋮	⋮	⋮

	n	mean	sd	median	min	max
hrv (low csai)	125	115.45	22.58	111.84	59.93	184.92
bf (low csai)	125	20.45	4.69	20.24	9.49	32.40
ips (low csai)	125	757.58	65.83	769.64	537.65	960.12
hrv (high csai)	125	-107.44	23.91	-108.20	-165.88	-54.19
bf (high csai)	125	20.03	4.58	20.04	6.57	30.47
ips (high csai)	125	760.94	73.11	758.61	545.08	967.37

An illustrative example

Data and descriptive analyses



(A) Scatter plot of HRV and BF as a function of CSAI
(B) Scatter plot of HRV and IPS as a function of CSAI
(C) Box plot of HRV as a function of CSAI

An illustrative example

First model (additive): definition

The goal here is to evaluate whether HRV varies as a linear function of BF, IPS, and CSAI:

$$\texttt{HRV}_i = \beta_0 + \texttt{BF}_i\beta_1 + \texttt{IPS}_i\beta_2 + \texttt{CSAI}\beta_3 + \epsilon_i$$

Under the assumption $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we get the Normal linear model:

$$\texttt{HRV}_i \sim \mathcal{N}\left(\beta_0 + \texttt{BF}_i\beta_1 + \texttt{IPS}_i\beta_2 + \texttt{CSAI}\beta_3, \sigma^2\right)$$

Note that CSAI is a categorical variable with two levels (0: low; 1: high). Using the (standard) dummy coding for this variable, we get as follows:

- β_0 codifies the level `CSAI=0` (baseline/intercept)
- β_3 quantifies the increment/decrement of `HRV` obtained when `CSAI=1`

An illustrative example

First model (additive): parameter estimation

Using the maximum-likelihood results for $\boldsymbol{\beta}$ (`estimate`) and σ^2 (`residual variance`), we get the following estimates along with the standard errors $\sigma_{\hat{\beta}}$ (`Std. Error`):

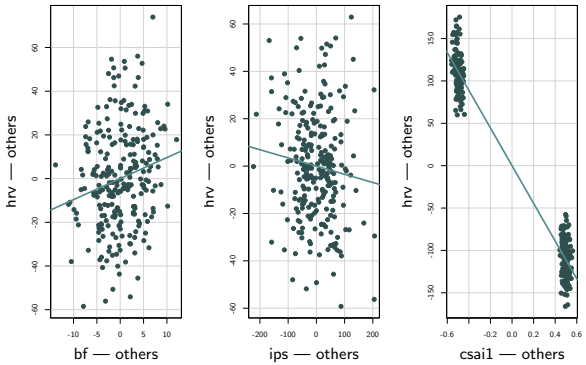
	Estimate	Std. Error
Intercept (CSAI:0)	121.935	17.403
BF	0.964	0.312
IPS	-0.035	0.021
CSAI:1	-222.376	2.882
Residual variance	518.002	
R2	0.961	

Note:

- The average level of `HRV` when `CSAI=0` is $\beta_0 = 121.935$
- When `CSAI=1`, the average level of `HRV` decreases by $\beta_3 = -222.376$ units
- `HRV` is positively associated to `BF` ($\beta_1 = 0.964$)
- `HRV` does not linearly depend on `IPS` ($\beta_2 = -0.035$)
- The overall fit of the model is satisfactory ($R^2 = 0.961$)

An illustrative example

First model (additive): parameter estimation



Partial regression plots for `HRV` as a function of the predictors `BF`, `IPS`, and `CSAI` respectively.

An illustrative example

First model (additive): inference

The t-statistics (`t-value`) along with their observed significance levels (`Pr(>|t|)`) can be computed to make inference about $\hat{\beta}$. Similarly, $1-\alpha$ CIs (`CI lb` and `CI ub`) can also be computed for the estimated regression coefficients ($\alpha = 0.05$).

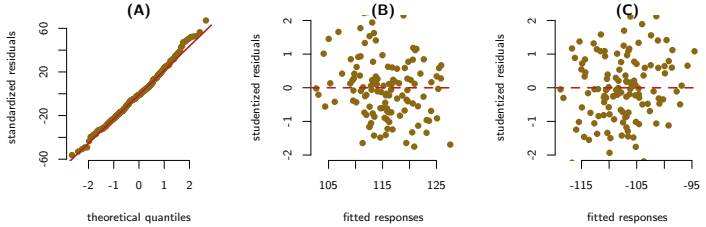
	Estimate	Std. Error	t-value	Pr(> |t|)	CI lb	CI ub
Intercept (CSAI:0)	121.935	17.403	7.007	0.000	87.658	156.212
BF	0.964	0.312	3.086	0.002	0.349	1.579
IPS	-0.035	0.021	-1.662	0.098	-0.076	0.006
CSAI:1	-222.376	2.882	-77.147	0.000	-228.054	-216.699
Residual variance	518.002					
R2	0.961					

Note:

- For the variable `IPS` the null hypothesis $H_0 : \beta_2 = 0$ cannot be rejected. Consistently, the 95% confidence interval for this coefficient contains zero.

An illustrative example

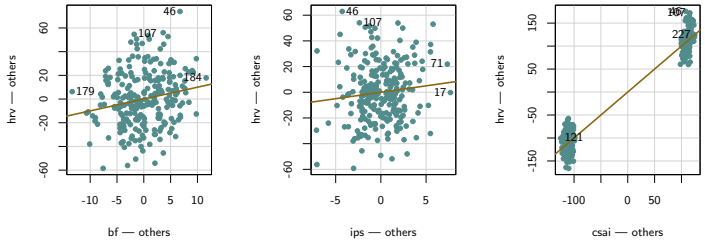
First model (additive): diagnostics



(A) Normality of the residuals
(B) Homoscedasticity conditioned on `CSAI=0`
(C) Homoscedasticity conditioned on `CSAI=1`

An illustrative example

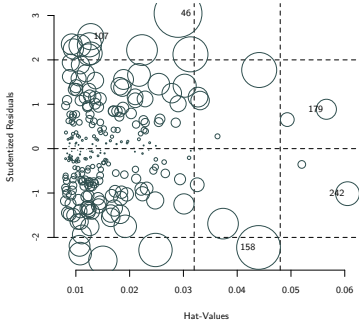
First model (additive): diagnostics



Partial residual plots with suspected leverage observations.

An illustrative example

First model (additive): diagnostics



Influential plot with suspected influential observations. Note that the most influential observation ($i = 46$) is not classified as influential since the Bonferroni-adjusted t-test is not significant ($r_{46} = 3.047, \alpha_{\text{obs}}^{\text{adj}} = 0.642$).

An illustrative example

Second model (interaction): definition

We can ask whether adding the term `CSAI`×`IPS` would generally improve the fit of the previous model:

$$\texttt{HRV}_i = \beta_0 + \texttt{BF}_i\beta_1 + \texttt{IPS}_i\beta_2 + \texttt{CSAI}\beta_3 + (\texttt{CSAI} \times \texttt{IPS})\beta_4 + \epsilon_i$$

Under the assumption $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we get the Normal linear model:

$$\texttt{HRV}_i \sim \mathcal{N}\left(\beta_0 + \texttt{BF}_i\beta_1 + \texttt{IPS}_i\beta_2 + \texttt{CSAI}\beta_3 + (\texttt{CSAI} \times \texttt{IPS})\beta_4, \sigma^2\right)$$

Note that CSAI is a categorical variable with two levels (0: low; 1: high). Using the (standard) dummy coding for this variable, we get as follows:

- β_0 codifies the level `CSAI=0` (baseline/intercept)
- β_3 quantifies the increment/decrement of `HRV` obtained when `CSAI=1`
- β_4 quantifies the interaction effect between `IPS` and `CSAI`

An illustrative example

Second model (interaction): parameter estimation and inference

	Estimate	Std. Error	t-value	Pr(> |t|)	CI lb	CI ub
Intercept (CSAI:0)	2.928	22.682	0.129	0.897	-41.748	47.604
BF	1.085	0.284	3.820	0.000	0.525	1.644
IPS	0.119	0.028	4.219	0.000	0.064	0.175
CSAI:1	-11.434	28.965	-0.395	0.693	-68.486	45.619
IPS x CSAI:1	-0.278	0.038	-7.313	0.000	-0.353	-0.203
Residual variance	426.935					
R2	0.968					

The incremental $\mathcal{F}$-test can be used to evaluate the new model:

	DF	RSS	SS	F stat	Pr(> F)	AIC
model 0	246.000	127428.516				2277.932
model 1	245.000	104599.000	22829.517	53.473	0.000	2230.576

Notes: `DF`: degrees of freedom calculated as $n - J - 1$; `RSS`: residual sum of squares; `SS`: sum of squares; `F stat`: $\mathcal{F}$-statistic calculated as ration of squares (see slide 37, module B) with associated p-value (or observed significance level).

An illustrative example

Second model (interaction): parameter estimation and inference

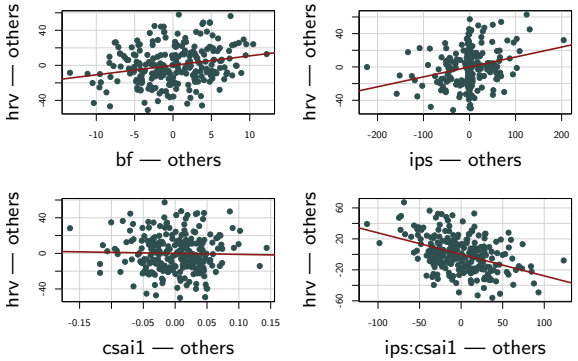
	DF	RSS	SS	F stat	Pr(> F)	AIC
model 0	246.000	127428.516				2277.932
model 1	245.000	104599.000	22829.517	53.473	0.000	2230.576

Notes:

- The t-statistic for the `IPS` variable is significant now
- The t-statistic for the `CSAI` variable is instead not significant after the interaction term has been added
- The R^2 index is large showing that the model still shows a very good fit
- The incremental $\mathcal{F}$-test (or Anova table) shows that the interaction model shows better performance as opposed to the simplest additive model

An illustrative example

Second model (interaction): parameter estimation and inference



Partial regression plots for `HRV` as a function of the predictors `BF`, `IPS`, `CSAI`, and `IPS`×`CSAI`, respectively.

An illustrative example

Likewise for the simplest additive model, diagnostics can be computed for the current model as well.

During the practical sessions of the course, we will learn further strategies to deal with Normal linear models and data analysis.

$(x_1, \dots, x_J), y$

$y = f(x_1, \dots, x_J)$

outcome / predictors / regressors

$J = 1$

$[(y_1, x_1)', \dots, (y_m, x_m)']$ first sample

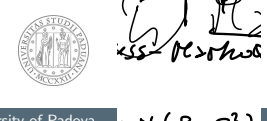
$[(y_1, x_1)'', \dots, (y_m, x_m)^4]$ second sample

Statistical methods and data analysis in developmental psychology

Antonio Calcagnì

DPSS, University of Padova

A.Y 2021-2022



Copyright © 2021 Antonio Calcagnì. Permis
this document under the terms of the GNU
later version published by the Free Software
https://www.gnu.org/licenses/fdl-1.3.h

Antonio Calcagnì
PSQ1096299 - First Part (module C)

Outline

Introduction

Source: 23.1 (Fox, 2016)

There are several situations where observations $y_1, \ldots, y_n$ are dependent such as when:

- Students are sampled from a random sample of schools (two levels of sampling: 1. schools; 2. students within each school).
- Patients are sampled within physicians which are in turn sampled within hospitals (three levels of sampling: 1. hospitals; 2. physicians; 3. patients).
- Subjects are measured over time in a controlled experiment (repeated measurements)

Introduction

Source: 23.1 (Fox, 2016)

In both cases, observations are no longer independent as they are **clustered** in $m = 1, \ldots, M$ clusters or subgroups (known in advance).

To deal with this issue, the standard Normal linear model needs to be extended properly. There are a number of ways to deal with non-independent observations including *marginal models* (where structured covariance matrices can be used to model the covariance matrix of the errors), *time-series*, and *linear mixed-effects models*.

In this course, we will briefly learn linear mixed-effects model (LMMs) along with their use in practical applications. Particularly, we will focus on *single-level* LMMs (or *random-intercept* Normal linear model).

Model specification

Source: 23.2 (Fox, 2016)

Let

$$Y_1 = (Y_{11}, \ldots, Y_{1m} \ldots, Y_{1M})$$
$$\vdots$$
$$Y_i = (Y_{i1}, \ldots, Y_{im} \ldots, Y_{iM})$$
$$\vdots$$
$$Y_i = (Y_{n1}, \ldots, Y_{nm} \ldots, Y_{nM})$$

be a collection of random variables independent over $i = 1, \ldots, n$. For each outcome y_i, a set of (non-random) variable is collected so that the observed sample can be represented in terms of pairs:

$$\mathbf{y} = \{(\mathbf{y}_1\mathbf{x}_1), \ldots, (\mathbf{y}_i\mathbf{x}_i), \ldots, (\mathbf{y}_n\mathbf{x}_n)\}$$

where $\mathbf{y}_i$ is a $M \times 1$ vector of dependent observations.

Model specification

Source: 23.2 (Fox, 2016)

i	m	Y	X_1	$\dots$	X_J
1	1	y_{11}	x_{11}	$\dots$	x_{1J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	m	y_{1m}	x_{11}	$\dots$	x_{1J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	M	y_{1M}	x_{11}	$\dots$	x_{1J}
2	1	y_{21}	x_{21}	$\dots$	x_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	m	y_{2m}	x_{21}	$\dots$	x_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	M	y_{2M}	x_{21}	$\dots$	x_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Structure of the data:

- $i = 1, \dots, n$ is the index set for the first level of sampling (subjects)
- $m = 1, \dots, M$ is the index set for the second level of sampling (groups or clusters)
- There are M clusters of dependent observations $\mathbf{y}_{M\times 1}^{(i)} = (y_{i1}, \dots, y_{iM})$
- The explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_J$ are repeated within each cluster
- The are $M \times n$ observations in the dataset

Model specification

Source: 23.2 (Fox, 2016)

The Normal linear model with random-intercept is of the form:

$$\begin{aligned}\boldsymbol{\eta}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\eta^2)\\ \mathbf{y}_i|\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}\sigma_y^2)\\ \boldsymbol{\mu}_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i\end{aligned}$$

where $\boldsymbol{\eta}_i$ is the $M \times 1$ vector containing the *random components* (also called *intercepts*) of the model, $\mathbf{X}_i = \mathbf{1}_{M\times 1}\mathbf{x}_i$ is a $M \times J$ matrix containing (row-wise) replicates of the *predictors* $\mathbf{x}_i$, $\{\beta_0, \boldsymbol{\beta}\}$ are the *regression coefficients*, whereas σ_η^2 and σ_y^2 are the *variances* of the random effects and errors, respectively.

Model specification

Source: 23.2 (Fox, 2016)

$$\begin{aligned}\boldsymbol{\eta}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\eta^2)\\ \mathbf{y}_i|\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}\sigma_y^2)\\ \boldsymbol{\mu}_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i\end{aligned}$$

The interpretation of the model parameter is the same as for the standard Normal linear model. In this case:

- $\beta_0, \boldsymbol{\beta}$ are called **fixed effects** (they are parameters)
- $\boldsymbol{\eta}_i$ are called **random effects** (they are not parameters but unobservable random realizations)

Model specification

Source: 23.2 (Fox, 2016)

$$\begin{aligned}\boldsymbol{\eta}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\eta^2)\\ \mathbf{y}_i|\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}\sigma_y^2)\\ \boldsymbol{\mu}_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i\end{aligned}$$

The following assumptions hold:

- $\eta_1, \dots, \eta_M$ are i.i.d. realizations as well as $\mathbf{y}_1, \dots, \mathbf{y}_n$
- y_{im} and y_{ih} $(m \neq h)$ for a fixed i are not independent
- σ_η^2 and σ_y^2 are independent

Model specification

Source: 23.2 (Fox, 2016)

$$\begin{aligned}\boldsymbol{\eta}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_\eta)\\ \mathbf{y}_i|\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}\sigma^2_y)\\ \boldsymbol{\mu}_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i\end{aligned}$$

In this case, the mean of the marginal model for $\mathbf{y}_i$ is

$$\mathbb{E}[Y_i] = \beta_0 + \mathbf{X}_i\boldsymbol{\beta}$$

whereas the variance is

$$\mathbb{V}\text{ar}[Y_i] = \sigma^2_y + \sigma^2_\eta$$

We can notice that the random-effect modifies the variance of the model (the mean is still the same).

Model specification

Source: 23.2 (Fox, 2016)

$$\begin{aligned}\boldsymbol{\eta}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_\eta)\\ \mathbf{y}_i|\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}\sigma^2_y)\\ \boldsymbol{\mu}_i &= \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i\end{aligned}$$

The correlation between two different observations Y_{im} and Y_{ih} for a fixed unit i is as follows:

$$\mathbb{C}\text{or}[Y_{im}, Y_{ih}] = \frac{\sigma^2_\eta}{\sigma^2_y + \sigma^2_\eta}$$

This term is also called *intraclass correlation* and it can be used to assess how much of the total variance is due to *within-subject variation* σ^2_η.

Parameter estimation

Source: 23.2, 23.9 (Fox, 2016)

The parameters of the single-level Normal linear model $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \sigma^2_y, \sigma^2_\eta\}$ can be estimated by maximizing the marginal likelihood of the model

$$\mathcal{L}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{\mathbb{R}} f_{Y_i||\eta_i}(\mathbf{y}; \boldsymbol{\mu}, \mathbf{I}\sigma^2_y) f_{\eta_i}(\boldsymbol{\eta}; \mathbf{I}\sigma^2_\eta)\ d\boldsymbol{\eta}$$

which is obtained by integrating out the random effects $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$.

To this purpose, several *numerical methods* are available such as Restricted Marginal Maximum Likelihood (REML) or Expectation Maximization (EM). In both cases, the procedure estimate the parameters only whereas random effects are recovered after parameter estimation.

Parameter estimation

Source: 23.2, 23.9 (Fox, 2016)

Once the parameters have been (numerically) estimated, inference on the fixed-effects can be performed by knowing that

$$\begin{aligned}\mathbb{E}[\boldsymbol{\beta}] &= \boldsymbol{\beta}\\ \mathbb{V}\text{ar}[\boldsymbol{\beta}] &= (\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X})^{-1}\end{aligned}$$

where

$$\mathbf{S} = \sigma^2_\eta \mathbf{Z}\mathbf{Z}^T + \mathbf{I}\sigma^2_y$$

is the variance-covariance matrix of the regression coefficients (fixed-effects) with

$$\mathbf{Z}_{nm\times n} = \mathbf{I}_{n\times n} \otimes \mathbf{1}_{m\times 1}$$

being the block matrix of the random-effects ($\otimes$ indicates the Kronecker product) and $\mathbf{I}$ an identity matrix of appropriate order.

Parameter estimation

Source: 23.8 (Fox, 2016)

ML-based methods allow for estimating $\boldsymbol{\theta}$ by marginalizing over the random effects. In doing so, they are not recovered together with the model parameters.

In order to recover $\hat{\boldsymbol{\eta}}$ conditioned on the current estimates, one may use the following linear estimator:

$$\hat{\boldsymbol{\eta}} = \mathbb{E}\left[\boldsymbol{\eta}|\hat{\boldsymbol{\theta}}, \mathbf{y}\right] = \sigma^2_\eta \mathbf{Z}^T \widehat{\mathbf{S}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

where the estimated variances $\hat{\sigma}^2_\eta$ and $\hat{\sigma}^2_y$ are used in the variance-covariance quantity $\widehat{\mathbf{S}}$. The estimator is the **best linear unbiased estimator** (BLUP) for the random-effect quantities.

The recovered $\hat{\boldsymbol{\eta}}$ can be used as new data in subsequent analyses (e.g., cluster analysis).

Inference

Source: 23.2, 23.9 (Fox, 2016)

Testing individual coefficients

Testing individual coefficients (fixed-effects) can be performed using t-statistics (see slides 30-31, Module B) with the **standard errors** of the estimates being computed as follows:

$$\sigma_{\hat{\beta}} = \sqrt{\text{diag}\left((\mathbf{X}^T \widehat{\mathbf{S}}^{-1}\mathbf{X})^{-1}\right)}$$

where the estimated variances $\hat{\sigma}^2_\eta$ and $\hat{\sigma}^2_y$ are used in the variance-covariance quantity $\widehat{\mathbf{S}}$.

Observed significance levels (p-values) for the t-statistics under the null hypotheses can be computed using the Sattherwite approximation or via parametric bootstrap.

Inference

Source: 23.2, 23.9 (Fox, 2016)

Testing subset of coefficients

Likewise for the standard Normal linear model (see slides 35-38, module B), testing subset of coefficients (fixed-effects) can be performed by comparing the baseline model $\mathcal{M}_0$ (e.g., null model) against the full or target model $\mathcal{M}_1$.

Unlike the standard case, the **Likelihood Ratio Test** (LRT) has to be used here:

$$W_{1|0} = 2\left(\ln \mathcal{L}_1(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ln \mathcal{L}_0(\hat{\boldsymbol{\theta}}; \mathbf{y})\right)$$

which under the null hypothesis is distributed according to a χ^2 distribution:

$$W_{1|0} \dot{\sim} \chi^2(W; \text{df}_1 - \text{df}_0)$$

where for the single-level model $\text{df} = J + 1 + 2$.

As usual, large values of $W_{1|0}$ allows for rejecting H_0.

Inference

Source: 23.2, 23.9 (Fox, 2016)

Testing subset of coefficients

Likewise for the standard Normal linear model (see slides 35-38, module B), testing subset of coefficients (fixed-effects) can be performed by comparing the baseline model $\mathcal{M}_0$ (e.g., null model) against the full or target model $\mathcal{M}_1$.

Unlike the standard case, the **Likelihood Ratio Test** (LRT) has to be used here:

$$W_{1|0} = 2\left(\ln \mathcal{L}_1(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ln \mathcal{L}_0(\hat{\boldsymbol{\theta}}; \mathbf{y})\right)$$

Note that in this case the log-likelihood of the model $\ln \mathcal{L}(\hat{\boldsymbol{\theta}}; \mathbf{y})$ is obtained by refitting the model via standard ML approach (e.g., no REML).

Inference

Source: 23.2, 23.9 (Fox, 2016)

Testing non-nested models

Non-nested random-effect Normal linear models can be compared by assessing AIC or BIC indices as for the standard case (see slide 39, module B). As usual, the minimum-AIC (or minimum-BIC) criterion has to be used in order to choose the best model between two competing models.

Diagnostics

Diagnostics for estimated random-effect Normal linear models can be performed similarly to the standard Normal linear model.

We will see more on this topic during the practical sessions of the course.

Outline

An illustrative example

Introduction

Background: Reaction times (RTs) is a well-known measure of cognitive processing. In this experiment, it is used to measure the cognitive loading of a math task in two experimental scenario: (a) standard homework (e.g., students do it alone); (b) innovative homework (e.g., students work partially in group).

Variables: RTs (in sec.) and type of homework (H) with two levels.

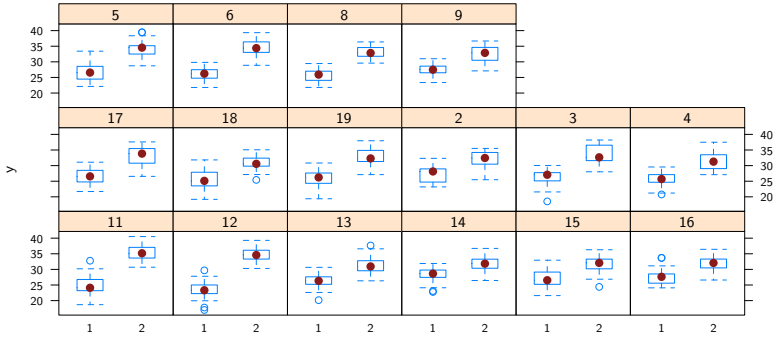
Goal: Define and fit a Normal linear model in order to predict RTs as a function of H. The model should take into account the within-subject variability measured during the task as it is expected that individuals have different response styles.

An illustrative example
Data and descriptive analyses

sbj	H	RT
1	1	17.16
1	1	16.75
1	1	19.42
1	1	23.20
1	1	17.34
2	1	19.70
2	1	23.54
2	1	18.33
2	1	21.83
2	1	21.83
3	1	21.76
3	1	23.39
3	1	19.72
3	1	16.87
3	1	21.95
⋮	⋮	⋮

	n	mean	sd	median	min	max
RT (standard homew)	500	19.87	2.72	19.94	10.60	27.30
RT (innovative homew)	500	26.34	2.88	26.21	18.00	34.11

An illustrative example
Data and descriptive analyses



RT as a function of the two-level variable H (1: standard homework, 2: innovative homework).
Note that each panel represents a single participant of the experiment.

An illustrative example
Model: definition

The goal here is to evaluate whether RT varies as a linear function of H by taking into account the within-subject variability (codified as sbj):

$$\texttt{RT}_i = \beta_0 + \texttt{H}_i\beta_1 + \texttt{sbj}_i + \epsilon_i$$

Under the assumptions $\epsilon_i \sim \mathcal{N}(0, \sigma^2_y)$ and $\texttt{sbj}_i \sim \mathcal{N}(0, \sigma^2_\eta)$, we get the random-effect Normal linear model:

$$\texttt{RT}_i \sim \mathcal{N}\left(\beta_0 + \texttt{H}_i\beta_1 + \texttt{sbj}_i, \sigma^2\right)$$

Note that H is a categorical variable with two levels (1: standard homew; 2: innovative homew). Using the (standard) dummy coding for this variable, we get as follows:

- β_0 codifies the level H=1 (baseline/intercept)
- β_2 quantifies the increment/decrement of H when H=2

An illustrative example
Model: parameter estimation and inference

	Estimate	Std. Error	df (Satterw approx)	t-value	Pr(> |t|)	CI lb	CI ub
Intercept (H:1)	19.846	0.174	34.962	113.971	0.000	19.501	20.188
H:2	6.512	0.187	827.853	34.853	0.000	6.145	6.877
σ^2_y	7.597						
σ^2_η	0.280						
Intraclass corr	0.161						

The LRT test can be used to evaluate the new model:

	DF	loglikel	Chisq	Pr(> *Chisq*)	AIC
model 0	3.000	-2834.718			5675.436
model 1	4.000	-2442.237	784.963	0.000	4892.473

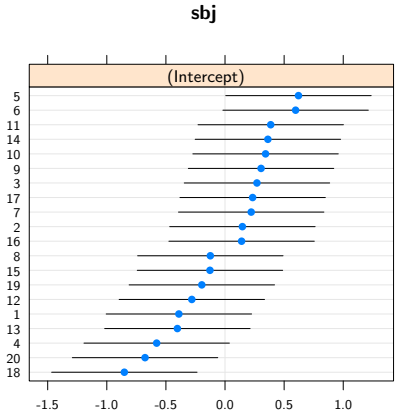
where

$$\text{model 0: } \texttt{RT}_i = \beta_0 + \texttt{sbj}_i + \epsilon_i$$

whereas `model 1` is the current model.

An illustrative example

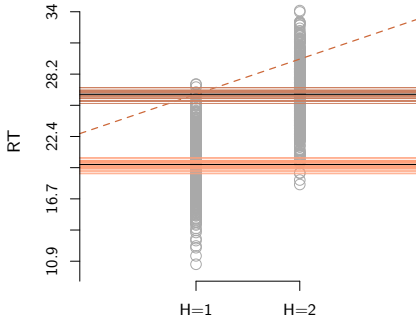
Model: parameter estimation and inference



Dotplot for the estimated random-effect quantities of the model.

An illustrative example

Model: parameter estimation and inference



Observed data (in gray) and estimated means (horizontal dark straight lines) along with the random-effect quantities (horizontal colored straight lines). Note that in the single-level model the random-effect quantities increment or decrement the intercept $\hat{\beta}_0$ only (the slope is unaffected by random deviations).

An illustrative example

Likewise for the standard Normal linear model, diagnostics and other graphical explorations can be computed in order to assess the fitted model.

During the practical sessions of the course, we will learn further strategies to deal with mixed-effect Normal linear models.

Statistical methods and data analysis in developmental psychology

Paolo Girardi

DPSS, University of Padova

A.Y 2021-2022

Copyright © 2021 Paolo Girardi Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation A copy of the license is available at: https://www.gnu.org/licenses/fdl-1.3.html.

Outline

Introduction to GLM

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

In the part A of the course the classical linear model can be summarized by:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{aligned}$$

The following assumptions **follow** from the model definition:

- *linearity*: $\mathbb{E}[Y_i]$ is a linear function of $\mathbf{x}_i$ (i.e., $\mathbb{E}[Y_i] = g(\mathbf{x}_i^T\beta)$ with $g(.)$ identity function);
- *homoscedasticity*: $\sigma_i^2 = \sigma^2$, i.e. constant variance for all the observations;
- *normality*: the conditional distribution of the response variable $Y_i|\mathbf{x}_i$ is Normal.

Introduction to GLM

Source: 5.2.1, 5.2.2, 6.1.1, 6.2.1, 9.1.0 (Fox, 2016); 2.1, 2.2 (Faraway, 2014)

In the part A of the course the classical linear model can be summarized by:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{aligned}$$

But these assumptions imply some constraints because:

- *linearity*: we imposed that the relationship between $\mathbb{E}[Y_i]$ and $\mathbf{x}_i$ is **linear**;
- *homoscedasticity*: σ^2 is **constant** for each observation.
- *normality*: $Y_i \sim \mathcal{N}()$, but many measurements are **not normally** distributed.

In order to deal with these constraints we can adopt, for i.e., some trasformation. Many measurements are log-normally distributed. This assumption implies that

$$\mathcal{Y} \sim \log \mathcal{N}(\mu, \sigma^2)$$

where $\log \mathcal{N}$ indicated that $\mathcal{Y}$ is a log-normal random variabile. With easy transformation we can obtain that

$$\mathcal{Y}^* = \log(\mathcal{Y}) \sim \mathcal{N}(\mu, \sigma^2)$$

But this transformation does not preserve, for example, the scale.

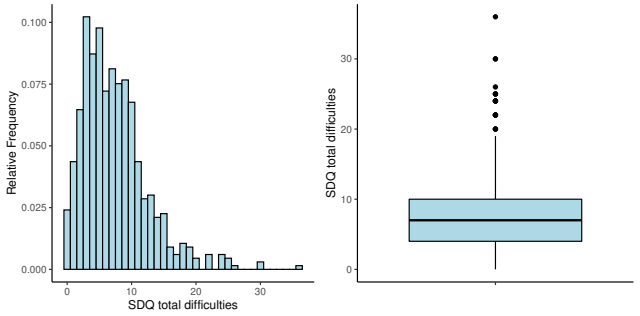
Other measurements cannot easily be led to follow a normal distribution after a transformation. In these settings, we need to introduce the framework of Generalized Linear Models (GLMs).

In order to infer the proper statistical model for a given response variable $\mathcal{Y}$, we use **Generalized Linear Models** (GLMs) which is a class of statistical models including many probabilistic models (e.g., Normal, Poisson, Gamma) for different response variables (e.g., continuous, counts, response times, ...).

A Generalized Linear Model (GLM) is a flexible generalization of linear regression that allows

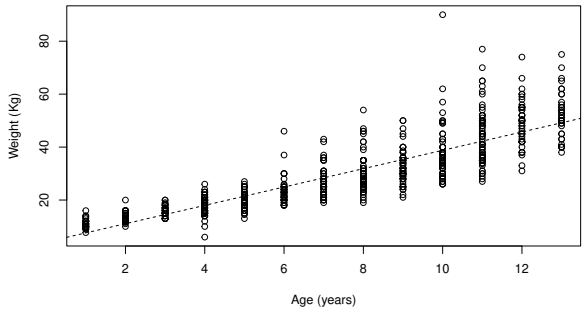
- *non normality*: $Y_i \sim \mathcal{D}(\theta)$ where $\mathcal{D}(\theta)$ is an appropriate probability distribution which depends on the parameters vector θ;
- *non linearity*: the relation between Y_i and $\mathbf{x}_i$ can be not linear (i.e., $\mathbb{E}[Y_i] = g(\mathbf{x}_i^T \beta)$ with $g(.)$ that is a function called **link function**);
- *heteroscedasticity*: magnitude of the variance of each measurement can be a function of its predicted value (i.e. $\mathbb{V}\text{ar}[Y_i] = \phi \mathbb{E}[Y_i]$).

We considered a Study conducted by the University of Padua (TEDDY Child Study, 2020) in which a sample of $n = 675$ children was assessed about the amount of total difficulties measured among children aged 3-13 years old with the SDQ questionnaire[1] **Question.** SDQ scores normally distributed?



[1]The Strengths and Difficulties Questionnaire (SDQ) is a brief behavioral screening questionnaire about 3-16-year-olds. The score is composed of the sum of behavioral indicators.

The assumption of homoscedasticity can be violated. In this example there is a linear relationship between y and x, but the variability increases with the values of y. This is a clear example of **heteroscedasticity**.

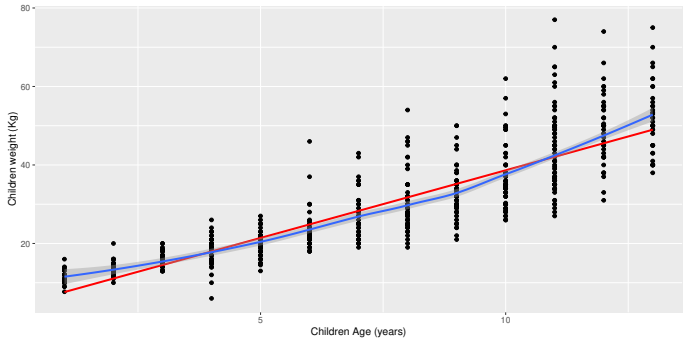


Question. Is the assumption of equal variance[2] supported by this figure?

[2]This assumpiton implies that $Var(\mathcal{Y}_i) = \sigma^2, \forall i$

Introduction
Example 3: TEDDY Child Study

In the previous example (675 children, TEDDY Child), the relationship can be assumed to be linear with the children's age?
This is a clear example of **non linear** relationship (in red the estimated linear, and in blue the "non linear" trend).



Question: Is linearity (and homoscedasticity) supported by this figure?

GLM in a nut-shell (or almost)
Source: 15.1 (Fox, 2016)

A Generalized Linear Model (or GLM) consists of three components:

- a **random component**, specifying the conditional distribution of the response variable, $\mathcal{Y}_i$ (for the i-th of n independently sampled observations), given the values of the explanatory variables X in the model;
- a **linear predictor** η_i that is, a linear function of regressors

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik};$$

- a smooth and invertible linearizing **link function** $g()$ which transforms the expectation of the response variable, $\mu_i = \mathbb{E}[Y_i]$ to the linear predictor in the following way

$$\mathbb{E}[Y_i] = g(\mu_i) = \eta_i.$$

GLM in a nut-shell (or almost): a random component
Source: 15.1 (Fox, 2016)

In Nelder and Wedderburn's (1972) original formulation, the distribution of the random variable $\mathcal{Y}_i$ is a member of an exponential family, such as the Gaussian (normal), Binomial, Poisson, gamma, or inverse-Gaussian families of distributions. The **Exponential Distribution** (ED) family can be express in the following form:

$$p(y_i; \theta_i, \phi) = \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

with y_i is the observation, while θ is called natural parameter and ϕ is the dispersion parameter. Specifying the expression of the functions $a()$, $b()$, and $c()$ obtains a particular parametric model.

This part will not be covered in depth. We focus our attention on two particular EDs: the **Binomial** and **Poisson** distribution.

GLM in a nut-shell (or almost): a random component
Example 1

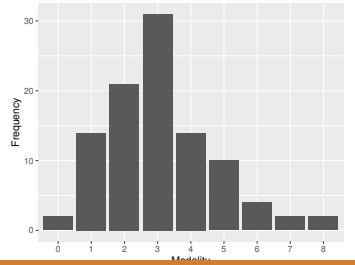
We simulated N=100 observations by a Poisson with mean/average parameter $\theta = \mu = 3$.

$$\mathcal{Y}_i \sim Poisson(\mu = 3)$$

for i=(1,2,...,100). Here the results

Number	0	1	2	3	4	5	6	7	8
Frequency	2	14	21	31	14	10	4	2	2

and a barplot of the frequencies



GLM in a nut-shell (or almost): a random component

Example 2

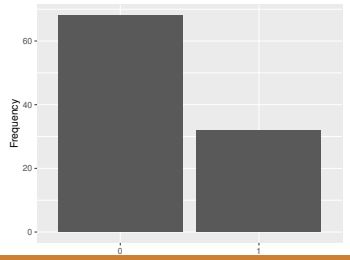
We simulated N=100 observations by a Bernoulli variable with parameter $\theta = \pi = 0.3$.

$$\mathcal{Y}_i \sim Bernoulli(\pi = 0.3)$$

for i=(1,2,...,100). Below the results .

Number	0	1
Frequency	68	32

The estimated $\hat{\pi} = \frac{32}{100} = 0.32$ and a barplot of the frequencies



GLM in a nut-shell (or almost): a random component

Example 3

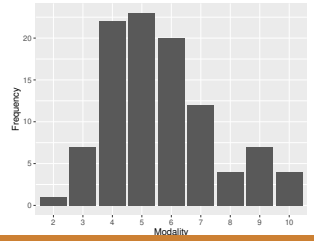
We simulated N=100 observations by a Binomial variable with parameter $\theta = \pi = 0.3$ and $n = 20$.

$$\mathcal{Y}_i \sim Binomial(\pi = 0.3, n = 20)$$

for i=(1,2,...,100). Below the results .

2	3	4	5	6	7	8	9	10
1	7	22	23	20	12	4	7	4

The estimated $\bar{y} = \frac{\sum_{i=1}^{20} y_i}{n} = \frac{565}{100} = 5.65$, $\hat{\pi} = 5.65/20 = 0.2825$ and a barplot of the frequencies



GLM in a nut-shell (or almost): a linear predictor

A **linear predictor** is a linear function of regressors

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

that are summed together in an additive form.

As in linear models, the regressors X_{ik} are prespecified functions of the explanatory variables and therefore may include quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial or regression-spline regressors, dummy regressors, interactions, and so on.

The novelty is that η_i is not directly connected to the $\mathbb{E}[Y_i]$, but...

GLM in a nut-shell (or almost): a link function

... the linear predictor is connected to the expectation of the response variable, $\mu_i = \mathbb{E}[Y_i]$ employing the **link function** $g()$ with respect to η_i as follows

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

Because the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}).$$

The inverse link $g^{-1}()$ is also called the **mean function**.

GLM in a nut-shell (or almost): a link function

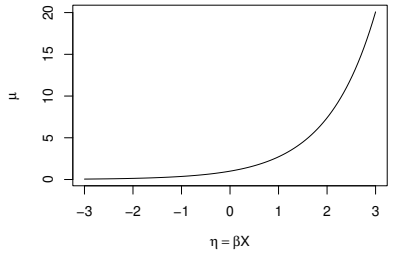
The **link function** permits to trasform the values of the linear predictor η_i in to the range of the response variables.

Example

If the response $\mathcal{Y}_i$ is a count, taking on only non-negative integer values, 0, 1, 2,., and consequently μ_i is an expected count, which (though not necessarily an integer) is also nonnegative, the log function is a link function because maps μ_i to the whole real line as follows:

$$log(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

for example if we considered $\mu = g^{-1}(\eta) = exp(\eta) = exp(\beta X)$



GLM in a nut-shell (or almost): the canonical link

The range of variation of the response variable (in the Bernoulli or Binomial variable: (0, 1), in the Poisson distribution: 0,1,2, etc...) identified in each family a so-called **canonical** (or "natural") **link function** associated with each family.

The canonical link simplifies the GLM, but other link functions may be used as well.

Family	**Canonical Link** $g()$	**Range of** $\mathcal{Y}_i$	**Var**$(\mathcal{Y}_i\|\eta)$
Gaussian Identity	Identity	$(-\infty, +\infty)$	$\phi = \sigma^2$
Binomial	Logit()	(0,1)	$\mu_i(1-\mu_i)$
Poisson	Log()	0,1,2, ...	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse	$(0,\infty)$	$\phi\mu_i^3$

Outline

Models for Dichotomous Data

Source: 14.1 (Fox, 2016)

This part introduces the generalized linear model for binary response variables. Dichotomous data treated the presence or the absence of a characteristic in our statistical units (a disease, the gender, the presence of child, ...).

However the probability distribution for dichotomous data can derived by a Bernoulli distribution

$$\mathcal{Y} \sim Be(\pi)$$

where $\mathcal{Y}$ can assume values of 0 or 1, the probability to observe 1 is π (in fact $\mathbb{E}[Y] = \Pr(\mathcal{Y} = 1) = \pi$)

We use to thinking of regression as a conditional average. Does this interpretation make sense when the response variable is dichotomous?

Models for Dichotomous Data

Source: 14.1 (Fox, 2016)

After all, an average between 0 and 1 represents a "score" for the dummy response variable that cannot be realized by any individual. In particular we are interested on a conditionated probability as follows

$$\mathbb{E}[\mathcal{Y}|x_i] = Pr(\mathcal{Y} = 1|X = x_i)$$

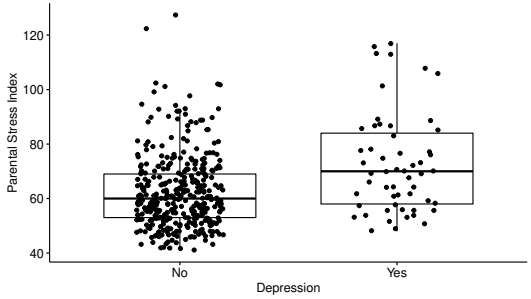
with

- $\mathbb{E}[\mathcal{Y}|x_i] = \pi_i * 1 + (1 - \pi_i) * 0 = \pi_i$
- π_i is the probability to have $\mathcal{Y} = 1$ given $X = x_i$ for the observation i.

Models for Dichotomous Data: a linear probability model

To understand why these models are required, let us begin by examining a representative problem, attempting to apply linear least-squares regression to it.

In the TEDDY Child Study, we asked the participants (mothers of a young child) about the presence of post-partum depression and we measured the parental stress[3].



[3] PSI-Parenting Stress Index, 4th Edition

Models for Dichotomous Data: a linear probability model

The Linear-Probability Model
As a first effort, let us try linear regression with the usual assumptions:

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

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and incorrelated ($cor(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$).
Since Y_i can assume a value of 0 or 1, its expection $\mathbb{E}[Y]$=$\mu = \pi$ and

$$\pi_i = \beta_0 + \beta X_i$$

For this reason, the linear-regression model applied to a dummy response variable is called the linear probability model. But in this formulation, the errors can assume a series of values that depends on π_i:

$$\varepsilon_i|Y_i = \begin{cases} \text{if } Y_i = 1 \text{ we have } \varepsilon_i = 1 - \mathbb{E}[Y_i] = 1 - \pi_i \\ \text{if } Y_i = 0 \text{ we have } \varepsilon_i = 0 - \mathbb{E}[Y_i] = -\pi_i \end{cases}$$

The assumption of $\varepsilon_i \sim N()$ is clearly **violated** since $\pi_i \in (0, 1)$ and the error can't take any values in the real line.

Models for Dichotomous Data: a linear probability model

The Linear-Probability Model
The variance of ε_i is

$$Var(\varepsilon_i) = \pi_i(1 - \pi_i)$$

which depends on the value of π_i leading to a clear constraint againt the potentiral presence of **heteroskedasticity**.
In addition, in the linear form of the model

$$\pi_i = \beta_0 + \beta X_i$$

$\beta_0 + \beta X_i$ is not limited to take values between (0, 1) and values outside the range are **permitted**. One solution to the problems of the linear-probability model —though not a good general solution— is simply to constrain π to the unit interval while retaining the linear relationship between π and X within this interval:

$$\pi|X_i = \begin{cases} 0 & \text{if } \beta_0 + \beta X_i < 0 \\ \beta_0 + \beta X_i & \text{if } \beta_0 + \beta X_i \in [0, 1] \\ 1 & \text{if } \beta_0 + \beta X_i > 1 \end{cases}$$

Models for Dichotomous Data: a linear probability model

The Linear-Probability Model
For the previous data, the estimated model is the following[4].

Table: Estimated linear probability model

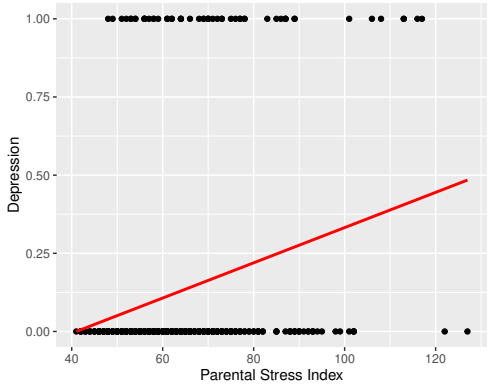
	Dependent variable:
	depression_n
parent_stress	0.006***
	(0.001)
Constant	−0.231***
	(0.071)
Observations	429
R^2	0.059
Adjusted R^2	0.056
Residual Std. Error	0.325 (df = 427)
F Statistic	26.601*** (df = 1; 427)
Note:	*p<0.1; **p<0.05; ***p<0.01

For each 1-point increase in its score, an increase in the probability of depression was estimated at 0.006 points (s.e. 0.001, p-value<0.01).

[4] estimated with the comand *lm()* in **R** software

Models for Dichotomous Data: a linear probability model

The Linear-Probability Model



A low probability of depression with low parental stress, while moving to high value of parentals stress index the probability increased up to 50%.
Is the model correct?

Models for Dichotomous Data: a logistic model

Transformations of π: Probit and Logit Models
A central difficulty of the unconstrained linear-probability model is its inability to ensure that π stays between 0 and 1. We need a positive monotone (i.e., nondecreasing) function that maps the linear predictor η to the interval [0,1].

Any Cumulative Probability Distribution function (**CDF**) meets this requirement, and we define P() as a selected **CDF**.

$$\pi_i = P(\eta_i) = \beta_0 + \beta X_i$$

where $P()$ is the so called mean function $g^{-1}()$ previously described.
If $P()$ is the CDF of a normal standard distribution $\mathcal{N}(0,1)$, $\Phi()$, the model is called **linear probit model**:

$$\pi_i = \Phi(\eta_i) = \beta_0 + \beta X_i$$

.

Models for Dichotomous Data: a logistic model

Another function $g^{-1}()$ which ensures that π stays between 0 and 1 is called logistic function

$$\Lambda(z) = \frac{1}{1+e^{-z}}$$

where e is the Euler's number= 2.718.
The function $\Lambda(z)$ is applied to the linear prediction η obtaining

$$\pi_i = \Lambda(\eta_i) = \beta_0 + \beta X_i.$$

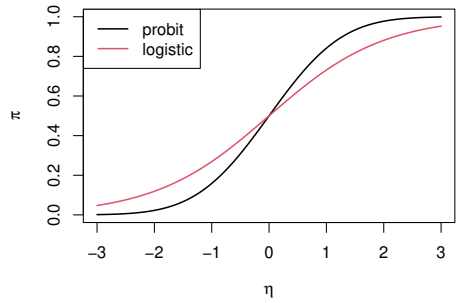
The model is called **linear logistic model or linear logit model:**

$$\pi_i = \frac{1}{1+e^{-(\beta_0+\beta X_i)}}.$$

The logistic function or logit is the canonical link function for the binomial regression.

Models for Dichotomous Data: a logistic model

The probit and the logistic function are very similar. It is also clear from this graph that both functions are nearly linear over much of their range, say between about π=0.2 and π=0.8.



Why is convenient to use the logistic (or logit) specification?

Models for Dichotomous Data: a logistic model

Despite their essential similarity, there are two practical advantages of the logit model compared to the probit model:

- The equation of the logistic CDF is very simple, while the normal CDF involves an unevaluated integral.
- More important, the inverse linearizing transformation for the logit model, $\Lambda^{-1}()$, is directly interpretable as a log-odds (log $\frac{\pi}{1-\pi}$):

$$\Lambda^{-1}(\pi_i) = \frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta X_i)$$

To make it clear here a simple table with values of π, *odds* and Log Odds:

π	**odds=$\frac{\pi}{1-\pi}$**	**Log(odds)**
0.01	0.0101010	-4.595120
0.05	0.0526316	-2.944439
0.20	0.2500000	-1.386294
0.50	1.0000000	0.000000
0.80	4.0000000	1.386294
0.95	19.0000000	2.944439
0.99	99.0000000	4.595120

Models for Dichotomous Data: a logistic model

The probit link does not provide the same good features of the logit specification. In particular, the logit model is a linear, additive model for the log odds, but it is also a multiplicative model for the odds:

$$\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta X_i) = exp(\beta_0)exp(\beta X_i) = e^{\beta_0}(e^{\beta})^{X_i}$$

So, increasing X by 1 changes the logit by β and multiplies the odds by e^{β}.

This trasformation of the coefficient β, (e^{β}), is called **Odds Ratio** and it is closely related to a variation of the probabilty of π by 1-points increase of X.

$$[\pi|X = (x+1) - \pi|X = x] = \Delta\pi \approx (e^{\beta})$$

To better understand this meaning, we compare the results of the linear probability model with the linear logit model on the data about parental stress and post-partum depression.

Models for Dichotomous Data: a logistic model

From the previous application...
The **linear probability model** estimats these coefficients between Y and X

$$\pi_i = -0.231 + 0.006X$$

For each increase of 1-point of parental stress index (X) the absolute probability of postpartum depression increases by 0.006, in the percentage of 0.6%.

If we use the **linear logistic model** we impose this equation to the data

$$\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \beta X_i)$$

Where the values of β and the related Odds ratio $=e^{\beta}$. The Odds Ratio is the increase in the relative probabilty due the increase of 1 points of x.

Models for Dichotomous Data: a logistic model

The estimates of the parameter $\boldsymbol{\beta} = (\beta_0, \beta)$[5] is commonly perfomed by a Maximum Likelihood Estimator (MLE) as following:

$$(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, y, x)$$

where the likelihood $\mathcal{L}(\boldsymbol{\beta}, Y, X)$ is formed by the product of each conditional probability

$$L(\boldsymbol{\beta}, y, x) = \prod_{i=1}^{n} \Pr(Y|X = x)$$

and, using the probability density of a Bernoulli variable, we obtain that

$$L(\boldsymbol{\beta}, y, x) = \prod_{i=1}^{n} (\pi_i|X = x)^{y_i} (1 - \pi_i|X = x)^{1-y_i}$$

where with a logit link we obtain

$$(\pi_i|X = x_i) = \frac{1}{1 - e^{-(\beta_0 + \beta x_i)}}$$

[5] This formula can easily be extended to k regressors $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$

Models for Dichotomous Data: a logistic model

At the end, jumping some intermediate steps, the estimation of the parameters was perfomed by the maximization of the log-likelihood $\ell(\boldsymbol{\beta}, y, x)$

$$\begin{aligned} (\hat{\boldsymbol{\beta}}) &= \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, y, x) \\ &= \sum_{i=1}^{n} [y_i \log(\pi_i|X = x) + (1 - y_i) \log(1 - \pi_i|X = x)] \end{aligned}$$

The solution is a system not linear in the parameter

$$\begin{cases} \sum_{i=1}^{n} (y_i - \pi_i|X = x_i) = 0 \text{ (for the solution of } \beta_0) \\ \sum_{i=1}^{n} (y_i - \pi_i|X = x_i) x_i = 0 \text{ (for the solution of } \beta) \end{cases}$$

The solution can be derived using iterative methods.
(Newton-Raphson like methods (see. paragraph 14.1.5 for details))

Models for Dichotomous Data: a logistic model

Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation

- The asyntotic distribution of $\mathcal{B} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{I}^{-1}(\boldsymbol{\beta})$, where $\mathbf{I}^{-1}(\beta) = \mathbb{E}\left[\frac{\partial^2 \ell(\beta_0, \beta)}{\partial \beta_0 \partial \beta}\right]$ is the Fischer information matrix.
- $\mathcal{B}$ is the minimum variance unbiased estimator (MVUE)
- Given this distribution, the associated hypothesis test (Wald test)

$$\begin{cases} H_0 : \beta = 0; \\ H_1 : \beta \neq 0; \end{cases}$$

Can be performed to test the association between Y and X
The test statistic $W = \frac{\mathcal{B} - 0}{s.e.(\mathcal{B})} \sim N(0, 1)$

And if we use the fitted values $|\frac{\hat{\beta}}{s.e.(\hat{\beta})}| \geq z_{1-\alpha/2}$, define the usual test of significance against the null hyphotesis

Models for Dichotomous Data: a logistic model

It is also possible to formulate a Likelihood-Ratio Test (LRT) for the hypothesis that several coefficients are simultaneously equal to 0. If we consider the following model with 2 regressors X_1 and X_2:

$$\text{- model 1: } logit(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The system of hypothesis

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0; \\ H_1 : \beta_1 \text{ or } \beta_2 \neq 0; \end{cases}$$

Can be solved by fitting a NULL model (model 0) with the only intercept

$$\text{- model 0: } logit(\pi) = \beta_0 + 0 * X_1 + 0 * X_2$$

$$\text{- model 0: } logit(\pi) = \beta_0$$

.

Models for Dichotomous Data: a logistic model

The likelihood ratio test (LRT) statistic used the likelihood of the model 1 ($\mathcal{L}_1$) and of the null model ($\mathcal{L}_0$) for the null hypothesis. The LRT can be conducted with a simple difference between the log-likelihood of the two models:

$$G_0^2 = 2(log\,\mathcal{L}_1 - log\,\mathcal{L}_0)$$

Under the null hypothesis, this difference G_0^2 follows a χ^2 distribution (in this case 2 degrees of freedom).

$$G_0^2(q) \sim \chi_q^2$$

As derived before, the degree of freedom was defined by the number of null coefficients between the two models.

Models for Dichotomous Data: a logistic model

By comparing $\log \mathcal{L}_0$ for the model containing only the constant (β_0) to the $\log \mathcal{L}_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y. The quantity pseudo-R^2 is a generalization of the residual sum of squares for a linear model (called also Nagelkerke R^2).
Thus,

$$\text{pseudo-}R^2 = \frac{G_1^2}{G_0^2} = 1 - \frac{\log \mathcal{L}_1}{\log \mathcal{L}_0}$$

is analogous to R^2 for a linear model and it ranges between 0 and 1.

Models for Dichotomous Data: a logistic model

Considering a logistic model for assessing the relationship between depression and parental stress index

Table: Estimated linear logistic model

	Dependent variable:
	depression_n
parent_stress	0.042^{***}
	(0.009)
Constant	-4.699^{***}
	(0.639)
Observations	429
Log Likelihood	−153.496
Akaike Inf. Crit.	310.992
Note:	*p<0.1; **p<0.05; ***p<0.01

In this model the effect was statistically significant

$$W_{oss} = \frac{\hat{\beta}}{s.e.(\hat{\beta})} == \frac{0.042}{0.009} = 4.7$$

which implies a p-value $\approxeq 0 << 0.05$.

Models for Dichotomous Data: a logistic model

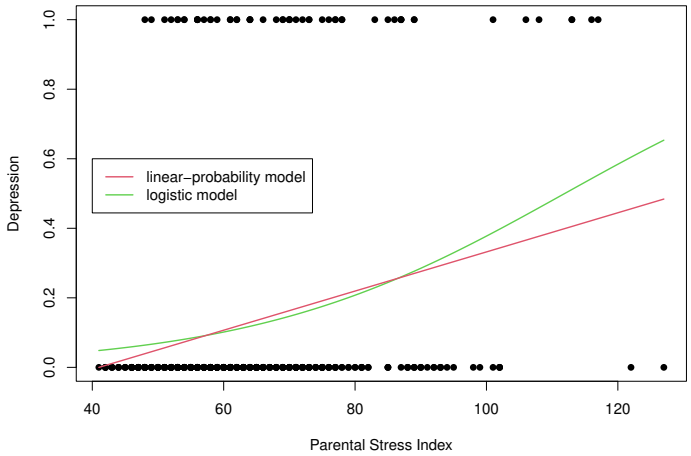
The estimated model is the following

$$logit(\pi_i, x_i) = \frac{\pi_i}{1-\pi_i} = exp(-4.7 + 0.042x_i)$$

The estimated slope coefficient $\hat{\beta}$=0.042.

The Odds Ratio is exp(0.042)=1.043 which gives us the information that each 1-point increase in parental stress index implies an increase in the relative probability of the post-partum depression of 4.3%.

Models for Dichotomous Data: a logistic model

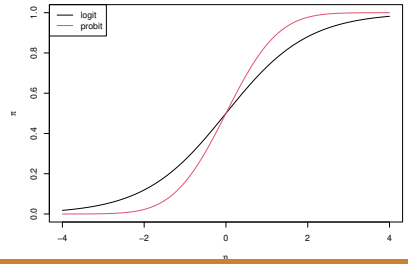


Models for Dichotomous Data: a probit model

The **Logit** specification for the binomial model is a common way to ensure that π belongs to the values [0,1].
As previously presented an alternative link is provided by any other CDF of a random variable (**probit** link).
The relationship between the probability of Y to assume the value 1 ($\Pr(Y_i = 1) = \pi_i$) and the regressor X_i si the following:

$$\Pr(Y_i = 1) = \pi_i = \Phi(\beta_0 + \beta X_i),$$

where $\Phi()$ is the CDF of a standard normal distribution ($\sim \mathcal{N}(0, 1)$).



Models for Dichotomous Data: a probit model

The results resulted by two different link in the binay regression (between logit and probit) are similar!

Table: Estimates of logistic and probit model

	Dependent variable:	
	depression_n	
	logistic	*probit*
	(1)	(2)
parent_stress	0.042***	0.023***
	(0.009)	(0.005)
Constant	−4.699***	−2.683***
	(0.639)	(0.349)
Observations	429	429
Log Likelihood	−153.496	−153.415
Akaike Inf. Crit.	310.992	310.831
Note:	*p<0.1; **p<0.05; ***p<0.01	

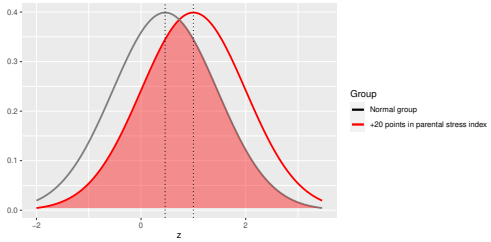
The estimated value for $\hat{\beta} = 0.023$ indicates that each 1-point increase in the X implies an increase of 0.023 point in the z scale value.

Models for Dichotomous Data: a probit model

The direction indicates a positive effect of the parenting stress on depression, but the interpretation can be only evaluated in terms of **effect size***.

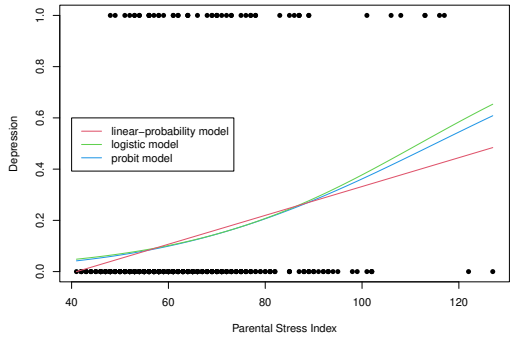
Example
(+10 pt increase: $\hat{\beta} * 10 = 0.023 * 10 = 0.23$ which is a low effect size).
(+20 pt increase: $\hat{\beta} * 20 = 0.023 * 20 = 0.46$ which is a moderate effect size).

*In statistics, an **effect size** is a number measuring the strength of the relationship between two variables in a population. In this case, measure the strength of the relationship between parental stress and depression.



Models for Dichotomous Data: to summarize

- probit and logit models lead to the same results in terms of goodness of fit
- the logit model allows to write π_i in a closed-form
- the logit model can be easily interpreted in terms of Odds Ratio
- the probit model is more difficult to estimate
- the linear probability model is the most simple, but not correctly specified



Models for Dichotomous Data: to summarize

- The dichotomous logit model can be fit to data by the method of maximum likelihood.
- Wald tests and likelihood-ratio tests for the coefficients of the model parallel t-tests and
- Incremental F-tests for the general linear model.
- The deviance for the model, defined as $G^2 = 2 \log \mathcal{L}(\hat{\beta})$ connected to the maximized log-likelihood can be used to calculate a pseudo R^2 which is the analogous of the residual sum of squares for a linear model.

Models for Dichotomous Data: a latent variable

An alternative derivation of the logit or probit model posits an underlying regression for a the continuous but unobservable response variable $\mathcal{E}_i$ (representing, e.g., the "propensity" to vote to get vaccinated (as an example)), scaled so that

$$Y_i = \begin{cases} 0 \text{ when } \mathcal{E}_i \leq k \\ 1 \text{ when } \mathcal{E}_i > k \end{cases}$$

with k being an unknown threshold value. The model becomes the following

$$\mathcal{E}_i = \beta_0 + \beta X_i - \varepsilon_i$$

where ϵ_i is the traditional regression model. Since $\mathcal{E}_i$ is not observed (we have the values of only Y_i) the equation can be expressed as

$$Pr(Y_i = 1) = Pr(\mathcal{E}_i > k) = Pr(\beta_0 + \beta X_i - \varepsilon_i > k)$$

Or better

$$Pr(\beta_0 + \beta X_i > k + \varepsilon_i)$$

Models for Dichotomous Data: a latent variable

Fixing as an example k=0 and imposing $\varepsilon_i \sim \mathcal{N}(0, 1)$

$$Pr(\beta_0 + \beta X_i > \varepsilon_i) = \Phi(\beta_0 + \beta X_i)$$

which is the probit model.

Alternatively, if the ε_i follows the logistic distribution, then we get the logit model

$$Pr(\beta_0 + \beta X_i > \varepsilon_i) = \Lambda(\beta_0 + \beta X_i)$$

We will have occasion to return to the unobserved-variable formulation of logit and probit models when we consider models for ordinal categorical data (but not in this course).

Models for Dichotomous Data: a guided analysis

The **Space Shuttle Challenger** disaster was a fatal accident in the United States' space program that occurred on January 28, 1986,

The failure was caused by the failure of the two redundant **O-ring** seals used in the joint, in part because of the unusually cold temperatures at the time of launch.

Models for Dichotomous Data: a guided analysis

However, engineers and scientists had conducted n=23 tests as reported measuring the temperature and the fail.

	temp	fail
4/12/81	66	0
11/12/81	70	1
3/22/82	69	0
11/11/82	68	0
4/4/83	67	0
6/18/83	72	0
8/30/83	73	0
11/28/83	70	0
2/3/84	57	1
4/6/84	63	1
8/30/84	70	1
10/5/84	78	0

	temp	fail
11/8/84	67	0
1/24/85	53	2
4/12/85	67	0
4/29/85	75	0
6/17/85	70	0
7/29/85	81	0
8/27/85	76	0
10/3/85	79	0
10/30/85	75	2
11/26/85	76	0
1/12/86	58	1

The space shuttle Challenger was cleared to launch at 11:38 a.m. EST, with an air temperature of 36 Fahrenheit degrees (2 Celsius degrees).
To establish if the temperature influenced the probability of an O-ring fail, the NASA scientist took only fails data because they said that data of working tests are "not informative".

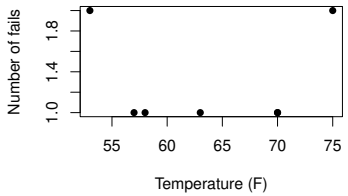
Models for Dichotomous Data: a guided analysis

The analysed dataset becomes formed by 7 observations

	temp	fail
11/12/81	70	1
2/3/84	57	1
4/6/84	63	1
8/30/84	70	1
1/24/85	53	2
10/30/85	75	2
1/12/86	58	1

and the analysis was conducted by means of a simple linear models

$$\text{number of fails}(Y_i) = \beta_0 + \beta\ \text{temperature}(X) + \varepsilon$$



Models for Dichotomous Data: a guided analysis

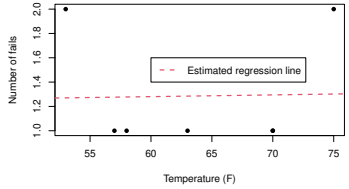
Questions

- Is this model appropriate? (is $\mathcal{Y}$ normallly distributed?)
- Do they have discarded useful information? (is "no fails" info irrelevant?)

The results were the following

Table: Estimated simple linear model

	Dependent variable:
	fail
temp	0.001
	(0.027)
Constant	1.195
	(1.715)
Observations	7
R^2	0.001
Adjusted R^2	−0.199
Residual Std. Error	0.534 (df = 5)
F Statistic	0.003 (df = 1; 5)
Note:	*p<0.1; **p<0.05; ***p<0.01



The conclusion led that temperature did not influence the number of fails.

Models for Dichotomous Data: a guided analysis

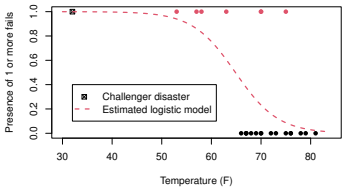
Considering the following logistic model we have

$$\Pr(\text{fails a the test i } > 0) = \pi_i = \Lambda(\beta_0 + \beta * \text{Temperature}(X))$$

allowing us to consider the probability to have a fail in the regression and the entire dataset. The results of the estimated linear logistic model are the following

Table: Estimated linear logistic model

	Dependent variable:
	I(fail >0)
temp	-0.232^{**}
	(0.108)
Constant	15.043^{**}
	(7.379)
Observations	23
Log Likelihood	-10.158
Akaike Inf. Crit.	24.315
Note:	*p<0.1; **p<0.05; ***p<0.01



The temperature has a negative influence on the probability to have a fail (low temperature, high probability of fail).

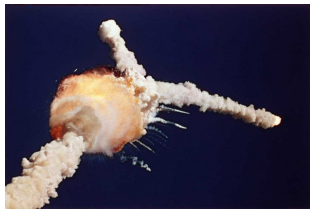
Models for Dichotomous Data: a guided analysis

The results of the logistic regresison reported that the estimated value for $\hat{\beta} = -0.23$ with a p-value <0.05 (0.032).

The Odds Ratio is exp(0.23)=0.79 that reported a protective effect of the temperature on the probability of failure. In particular, each 1-degree increase decreases the probability of fails by 21% (1-0.79).

The model estimated with a temperature of 36 Fahrenheit degrees a probability of failure of:

$$\Lambda(15.00 - 0.23 * 36) = 0.99871 = 99.871\%$$



Outline

Models for Counts: a gentle introduction

Source: 15.2 (Fox, 2016)

In this section we are interested to model data that are based on counts. Counts is evereything that has a support among integer numbers as:

$$\mathcal{S}_y = (0, 1, 2, 3, 4, \ldots)$$

What kind of characteristics has this kind of distribution?
Example: number of events, goals, errors, waiting time, etc...

However, these models have many applications, not only to the analysis of counts of events but also in the context of models for contingency tables and the analysis of survival data.

Models for Counts: a gentle introduction

Source: 15.2 (Fox, 2016)

The basic GLM for count data is the Poisson regression model with the log link. However, the Poisson regression is not the unique type of regression used to model counts. Alternative approaches:

- Negative binomial regression;
- Ordinal logistic regression (or proportional odds regression);
- Zero Inflated models;
- Truncated models;
- Bounded regression;
- ...

Models for Counts: a gentle introduction

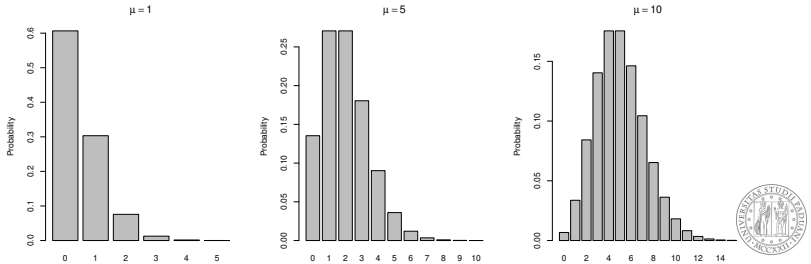
Source: 5.1 (Faraway, 2014)

The Poisson random variable distribution assigns a probability to each integer number in this way:

$$Y \sim Poi(\mu)$$

$$\Pr(Y = y; \mu) = \frac{e^{-\mu}\mu^{y}}{y!}$$

With $y = 0, 1, 2, 3, \ldots$. The parameter $\mu \in \Re^{+}$ s the expected number of events $\mathbb{E}[Y] = \mu$ that is also equal to the variance $\mathbb{V}\text{ar}[Y] = \mu$. Below probability value for $\mathcal{Y}(\mu)$ for μ equal to 1, 5 and 10.



The Poisson regression model

We are aimed to assess if a regressor (X) has an influence on the number of counts (Y).
We suppose that y_i is a realization of a Poisson random variable $\mathcal{Y}_i$ with parameter μ_i which may vary according to the values of X.

$$\mathcal{Y}_i \sim Pois(\mu_i)$$

We need some way to **link** the μ_i to the x_i.
As previously done in the binomial regression, a linear combination of the x_i form the **linear predictor** $\eta_i = X_i\beta^T$ and in order to ensure $\mu_i > 0$ we apply a **log** link function.
The Poisson regression model can be defined as:

$$\log(\mu_i) = \eta_i = \beta_0 + \beta X_i$$

where μ_i is the parameter of the Poisson random variable $\mathcal{Y}_i$.

The Poisson regression model

In this setting the previous formula is equlivalent to this

$$\mu_i = exp(\eta_i) = exp(\beta_0 + \beta X_i) = exp(\beta_0)exp(\beta)^{X_i}$$

denoting that the variation due to a unit change on x_i is proportional to $exp(\beta)$.

Replacing μ_i in the probability formula we obtain

$$\Pr(Y = y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \frac{e^{-(exp(\beta_0+\beta X_i)}(exp(\beta_0 + \beta X_i))^{y_i}}{y_i!}$$

In this way the mean parameter μ is replaced by a function of the regression parameters β_0 and β. We have to estimate a value for β_0 and β, how to do that?

The Poisson regression model

We use the Maximum Likelihood Estimator (MLE) to estimate the coefficient parameters. In particular, after some intermediate steps, we obtain that the log-likelihood for the Poisson regression model is the following:

$$\ell(\hat{\boldsymbol{\beta}}, \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} (y_i(\beta_0 + \beta x_i) - exp(\beta_0 + \beta x_i) - log(y_i!))$$

The estimates for $\hat{\beta}_0$ and $\hat{\beta}$ can be derived differentiating with respect to β_0 and β resorting to numerical methods to find a solution (as previously seen for the binomial regression).

The Poisson regression model

Interpretation of the coefficients β_0 and β.

- β_0 is the intercept: $exp(\beta_0)$ is the expected mean μ in absence of the effect of the variable X or however when X=0;
- β is the slope coefficient. If X is continuous, each 1-point increase is related to an increase of μ equal to $exp(\beta)$. In the Poisson Regression model is common to express the influence of the coefficients by means of Incidente Rate Ratio (IRR) that is the $exp(\beta)$. (IRR=$exp(\beta)$).

$$IRR(\beta) = \begin{cases} \text{if } > 1 \text{ we have a positive influence of X on Y} \\ \text{if } \approxeq 1 \text{ we do not have influence of X on Y} \\ \text{if } < 1 \text{ we have a negative influence of X on Y} \end{cases}$$

The Poisson regression model

An example: AIDS dataset

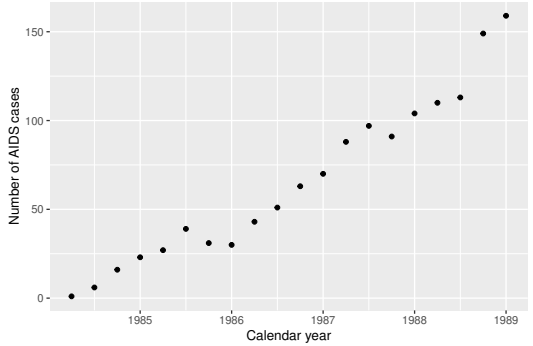
Whyte, et al 1987 (Dobson, 1990) reported the number of deaths due to AIDS in Australia per 3 month period from January 1983 – June 1986 with a total of 20 observations. (the dataset was inside the R packages *dobson*).

year	1984	1984	1984	1984	1985	1985	1985	1985	1986	1986
quarter	1	2	3	4	1	2	3	4	1	2
cases	1	6	16	23	27	39	31	30	43	51
year	1986	1986	1987	1987	1987	1987	1988	1988	1988	1988
quarter	3	4	1	2	3	4	1	2	3	4
cases	63	70	88	97	91	104	110	113	149	159

Question research: Is there any relationship between the number of AIDS cases and time?

The Poisson regression model

An example: AIDS dataset



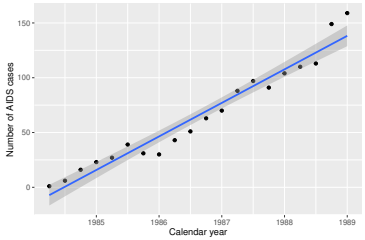
The number of AIDS cases is increasing. Can I apply the linear regression model in this case? Yes, but

The Poisson regression model

An example: AIDS dataset

Table: Estimated linear model

	Dependent variable:
	cases
time	30.614***
	(1.621)
Constant	−60,752.060***
	(3,219.706)
Observations	20
R^2	0.952
Adjusted R^2	0.949
Residual Std. Error	10.448 (df = 18)
F Statistic	356.801*** (df = 1; 18)
Note:	*p<0.1; **p<0.05; ***p<0.01

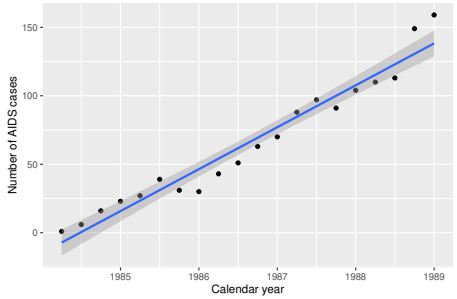


The model estimates an increase of 30.6 cases per year.

The model appears good, but what's wrong? Almost three things...

The Poisson regression model

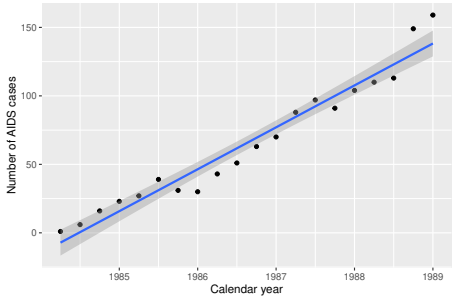
An example: AIDS dataset



What's wrong? Almost three things...
- 1) The expected number of AIDS cases in the year 1984 is below 0 (-7.15)

The Poisson regression model

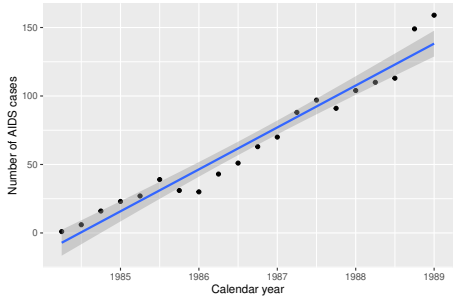
An example: AIDS dataset



What's wrong? Almost three things...
- 1) The expected number of AIDS cases in the year 1984 is below 0 (-7.15)
- 2) The increase is not so linear with the time

The Poisson regression model

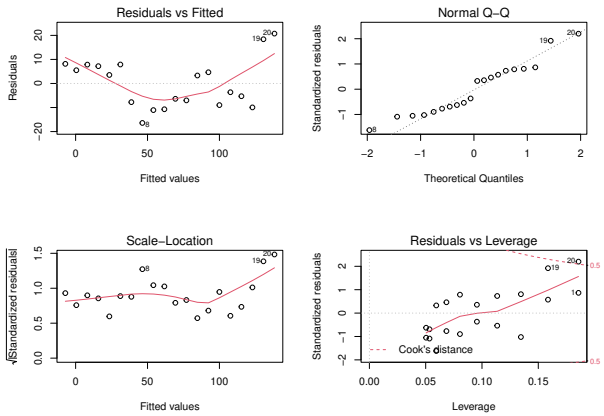
An example: AIDS dataset



What's wrong? Almost three things...
- 1) The expected number of AIDS cases in the year 1984 is below 0 (-7.15)
- 2) The increase is not so linear with the time
- 3) We are modelling counts, but linear models implies $\mathcal{Y} \sim \mathcal{N}(\mu, \sigma^2)$

The Poisson regression model
An example: AIDS dataset

Some indications are provided by the residual analysis



The Poisson regression model
An example: AIDS dataset

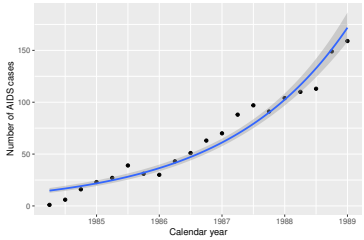
For this data we estimate a Poisson regression model for the variabile Y_t=AIDS cases at the time t as follows

$$log(\mu_t) = \beta_0 + \beta t$$

where t is the time and $Y_t \sim Poi(\mu_t)$.

Table: Estimated Poisson model

	Dependent variable:
	cases
time	0.517^{***}
	(0.022)
Constant	$-1{,}023.000^{***}$
	(44.400)
Observations	20
Log Likelihood	−82.700
Akaike Inf. Crit.	169.000
Note:	*p<0.1; **p<0.05; ***p<0.01

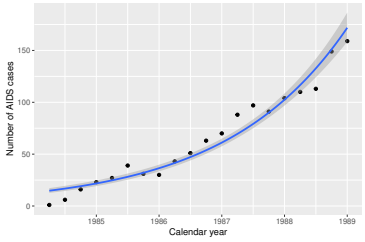


The model estimates an increasing trend (but not linear).

The Poisson regression model
An example: AIDS dataset

Table: Estimated Poisson model

	Dependent variable:
	cases
time	0.517^{***}
	(0.022)
Constant	$-1{,}023.000^{***}$
	(44.400)
Observations	20
Log Likelihood	−82.700
Akaike Inf. Crit.	169.000
Note:	*p<0.1; **p<0.05; ***p<0.01



The model estimates an increasing trend (but not linear).

The estimated coefficient $\hat{\beta}$=0.517 (>0) reported a increasing trend.
The IRR index = exp(0.517)=1.677 (>1) gives us the information that we registered an increase of 67.7% of cases per year.

Poisson regression model for contigency tables

When the explanatory variables—as well as the response—are discrete, the joint sample distribution of the variables defines a contingency table of counts: Each cell of the table records the number of observations possessing a particular combination of characteristics.
Let us consider the following dataset (dataset **drugpsy**, R package *faraway*)

y	diagnosis	drug
105	Schizophrenia	yes
12	Affective.Disorder	yes
18	Neurosis	yes
47	Personality.Disorder	yes
0	Special.Symptoms	yes
8	Schizophrenia	no
2	Affective.Disorder	no
19	Neurosis	no
52	Personality.Disorder	no
13	Special.Symptoms	no

The data contains a sample of 276 psychiatry patients classified by their diagnosis and whether drug treatment was prescribed. In this case the frequency (the number of patients who meet the two characteristics (diagnosis and drug).

Poisson regression model for contigency tables

This dataset can be obtained by the following original one:

ID	Diagnosis	drug presence
1	Neurosis	yes
2	Personality.Disorder	yes
3	Schizophrenia	no
4	Special.Symptoms	no
5	Neurosis	yes
6	Personality.Disorder	no

where Y is the number of patients in each combination of diagnosis and drug. The combination of the modality of diagnosis (5) and drug (2) defines the length of my dataset derived from the contingency tables (5x2=10).

Poisson regression model for contigency tables

The application of the Poisson regression model is quite simple. In this special case it is called also "log-linear" model. The related contingency 2x5 tables is

	Schizophrenia	Affective.Disorder	Neurosis	Personality.Disorder	Special.Symptoms
yes	105	12	18	47	0
no	8	2	19	52	13

The expected counts μ_{ij} depends on the row (drug) (i) and on the column (diagnosis)(j).

$$log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j$$

where μ is the overall mean, α_i is the effect of the drug and β_j is the effect of the diagnosis.

In this formulation is, under independence, the log expected frequencies η_{ij} depend additively on the logs of the row marginal expected frequencies, the column marginal expected frequencies, and the sample size.

Poisson regression model for contigency tables

We can add parameters to extend the loglinear model to data for which the row and column classifications are not independent in the population but rather are related in an arbitrary manner:

$$log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

leading to a *saturated model*, in fact, the number of independent parameters is equal to the number of cells in the table

Poisson regression model for contigency tables

In the log-linear model, we are not interested in the find if a regressor is statistically relevant (or significant). We are interested in the simplest model that influenced the observed counts To choose the best model we adopted an ANOVA test based on the Likelihood Ratio Test (**LRT**) on nested models.

Analysis of Deviance Table

Model 1: $\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$; In **R**: y $\sim$ drug * diagnosis (saturated model)
Model 2: $\eta_{ij} = \mu + \alpha_i + \beta_j$; In **R**: y $\sim$ drug + diagnosis
Model 3: $\eta_{ij} = \mu + \alpha_i$; In **R**: y $\sim$ drug
Model 4: $\eta_{ij} = \mu + \beta_j$; In **R**: y $\sim$ diagnosis
Model 5: $\eta_{ij} = \mu$; In **R**: y $\sim$ 1 (null model)

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Model 1	0	0.000	-	-	-
Model 2	4	96.537	-4	-96.537	<0.001
Model 3	8	268.499	-4	-171.962	<0.001
Model 4	5	125.091	3	143.408	<0.001
Model 5	9	297.053	-4	-171.962	<0.001

We conclude that both drug and diagnosis (and their interaction) reported an influence on the number of patients.

Outline

Overdispersion

The GLM regression is more flexible than the standard LM regression. However, depending on the underlying distribution probability there is a connection between the expected mean $\mathbb{E}[\mathcal{Y}|X]$ and the variance $Var(\mathcal{Y}|X)$.
In fact, for example:

- Poisson regression: $\mathbb{E}[\mathcal{Y}_i|X_i] = Var(\mathcal{Y}_i|X_i) = \mu_i$.
- Logistic regression: $\mathbb{E}[\mathcal{Y}_i|X_i] = \pi_i$ while $Var(\mathcal{Y}_i|X_i) = \pi_i(1-\pi_i)$ implying that $Var(\mathcal{Y}_i|X_i) = \mathbb{E}[\mathcal{Y}_i|X_i](1-\mathbb{E}[\mathcal{Y}_i|X_i])$.

If the GLM regression fits the data reasonably, we would expect the residual deviance to be roughly equal to the residual degrees of freedom. In fact, under the Null Hypothesis (no other regressors influenced the relationship between Y and X or the null-model is prefered)

$$D_{res} = 2(\log \mathcal{L}_k - \log \mathcal{L}_0) \sim \chi^2_{n-k}$$

where $\mathcal{L}_k$ is the likelihood of the model with k regressors, while $\mathcal{L}_0$ is the likelihood with only the intercept
It follows that $\mathbb{E}\left[\chi^2_{n-k}\right] = n-k$ and n-k is the residual degrees of freedom.

Overdispersion - Quasi Poisson

If the residual deviance is so large suggests that the conditional variance of the expected number of interlocks exceeds the variation of a Poisson-distributed variable, for which the variance equals the mean. This common occurrence in the analysis of count data is termed overdispersion.

A simple remedy for overdispersed count data is to introduce a dispersion parameter into the Poisson model, so that the conditional variance of the response is now

$$Var(\mathcal{Y}_i|X_i) = \phi\mu_i$$

with $(\phi > 1)$[6]
Nevertheless, the usual procedure for maximum-likelihood estimation of a GLM yields the so-called quasi-likelihood estimators of the regression coefficients.

[6] Although it is much less common, it is also possible for count data to be under dispersed—that is, for the conditional variance of the response to be less than the mean.

Overdispersion - Quasi Poisson

Important feautures:

- The Quasi-Poisson model yields to the same estimates ($\hat{\beta}$) of the Poisson model;
- The unique variation is in the standard error which are increased of $\phi^{1/2}$ ($s.e(\beta) * \phi^{1/2}$);
- The inflation of the standard errors implies a lower statistical significance of the parameters.

In the quasi-Poisson model, the dispersion estimator takes the form

$$\hat{\phi} = \frac{1}{n-k-1}\frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where $\hat{\mu}_i = g(\hat{\eta}_i)^{-1}$

Overdispersion - Quasi Poisson

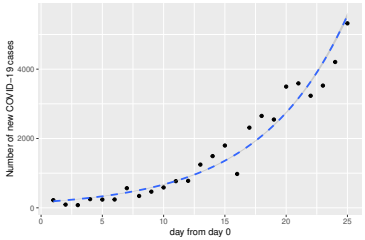
An example is provided by the recent COVID-19 epidemic. We analyse the first 25 days of the epidemic in Italy with the following model

$$\mathbb{E}[Y_i] = \mu_i = g(\eta)^{-1} = exp(\beta_0 + \beta X_i)$$

with Y_i the n. of new COVID-19 cases; X_i the day after the epidemic outbreak.

Table: Estimated Poisson model

	Dependent variable:
	new_positive
day	0.141***
	(0.001)
Constant	5.103***
	(0.018)
Observations	25
Log Likelihood	−1,004.689
Akaike Inf. Crit.	2,013.377
Note:	*p<0.1; **p<0.05; ***p<0.01

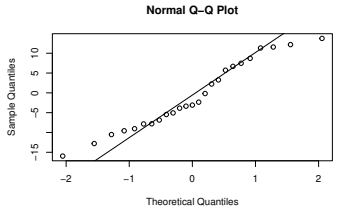


The model estimates an increasing trend.
IRR=exp($\hat{\beta}$)=exp(0.141)=1.15.

Overdispersion - Quasi Poisson

We evalute the presence of overdispersion:

- The model reported a Residual deviance of 1792.9 on 23 degrees of freedom
- The Deviance residuals (we will see them in the next part) appears to have tails too heavy respect to the normal distribution

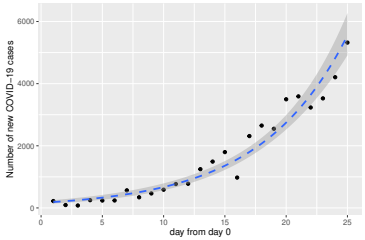


We estimated a Quasi-Poisson model, allowing the estimation of an over-dispersion parameter ϕ.

Overdispersion - Quasi Poisson

Table: Estimated Quasi-Poisson models

	Dependent variable:
	new_positive
day	0.141***
	(0.008)
Constant	5.103***
	(0.154)
ϕ	75.291***
Observations	25
Note:	*p<0.1; **p<0.05; ***p<0.01



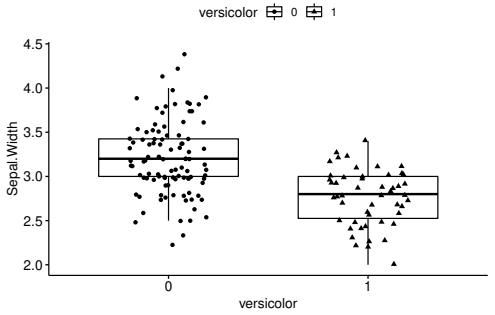
The dispersion parameter ϕ was estimated to be equal to $\hat{\phi}$ =75.291 (too different from 1). The standard errors of the estimates were increased by $\sqrt{75.291} = 8.68$ times.

Overdispersion - Quasi Binomial

The same method can be applied to the logistic model for binary regression. Considering the famous IRIS dataset, we want to explore if the length of the sepals discriminates between the species versicolor and the others. The model is

$$\mathbb{E}[Y_i] = \pi_i = g(\eta)^{-1} = exp(\beta_0 + \beta X_i)$$

with Y=1 if versicolor and 0 otherwise; X is the sepal width.



Overdispersion - Quasi Binomial

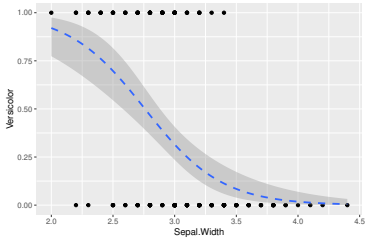
The same method can be applied to the logistic model for binary regression. Considering the famous IRIS dataset, we want to explore if the length of the sepals discriminates between the species Versicolor and the others. The model is

$$\mathbb{E}[Y_i] = \pi_i = g(\eta)^{-1} = exp(\beta_0 + \beta X_i)$$

with Y=1 if Versicolor and 0 otherwise; X is the sepal width.

Table: Estimated logistic model

	Dependent variable:
	versicolor
Sepal.Width	-3.220^{***}
	(0.637)
Constant	8.890^{***}
	(1.870)
Observations	150
Log Likelihood	−76.000
Akaike Inf. Crit.	156.000
Note:	*p<0.1; **p<0.05; ***p<0.01



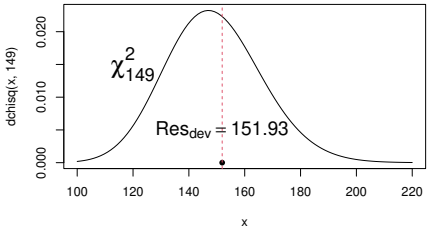
The model a inverse relationship **OR**=exp($\hat{\beta}$)=exp(-3.222)=0.04.

Overdispersion - Quasi Binomial

The residual deviance is equal to

$$Dev_{res} = 2(\log \mathcal{L}(\hat{\beta}) - \log \mathcal{L}_0) = 151.93$$

The number of residual degrees of freedom is 149.
To test if the presence of overdispersion is justifiable we can test if the observed residual deviance is a plausible value for a χ^2_{149}.



P-value=$Pr(\chi^2_{149} > 151.93)$=0.417 >0.05
In this case the presence of overdispersion is not supported.

Outline

Variable Selection and Hypothesis Test

In real settings, we want to evaluate influence of a set of regressors $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ on a dependent variable Y

$$log(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

However, not all the variables may have a significant effect on the Y.

How I can select variables most (statistically) influential?

This can be solved by Hypothesis Tests (Faraway: 15.3.3): Analysis of Deviance and correlated tests.

Variable Selection and Hypothesis Test

We define the log-likelihood for a GLM as a function $\log \mathcal{L}(\mu; y)$ where μ is the parameter of the related probability function $\mathcal{Y} \sim D(\mu)$. We can rewrite the log-likelihood in function of the parameters connected to $g(\mu) = X^T\boldsymbol{\beta}$

$$\log \mathcal{L}(\boldsymbol{\beta}; y) = \ell(\boldsymbol{\beta}; y)$$

where $\boldsymbol{\beta}$ is the vector of parameters of length k of k relative regressors (intercept included).

Following a backward approach (but in lab sessions both backward and forward approaches will be considered) we want to test if, for example, the last β_k is equal to 0.

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}$$

Variable Selection and Hypothesis Test

The following quantity

$$\Delta Dev(\boldsymbol{\beta}, y) = 2(\ell(\beta_0, \beta_1, \ldots, \beta_k; y) - \ell(\beta_0, \beta_1, \ldots, \beta_{k-1}, \beta_k = 0; y))$$

Is the difference between the log-likelihood between 2 nested models (the first model is the "full" model; the second model is the model with $\beta_k = 0$) multiplied by 2.

This difference is the quote of the deviance explained by β_k. Under the null hypothesis, this difference is approximately distributed following a chi-squared random variable with 1 degree of freedom

$$\Delta Dev(\boldsymbol{\beta}, y)_{|H_0} \sim^a \chi^2_1$$

This part provides us with the inferential part for the hypothesis test.
The degree of freedom derived from the number of coefficients tested to be 0 (in this case the degree of freedom is equal to 1).

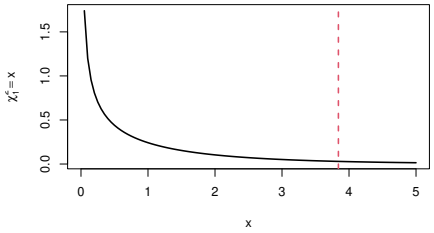
Variable Selection and Hypothesis Test

We have in this case two nested models

$$\begin{cases} M_0 : \hat{\eta}_0 = \hat{\beta}_{0,1} + \hat{\beta}_{0,1}X_1 + \ldots + \hat{\beta}_{0,k-1}X_{k-1} + \hat{\beta}_{0,k}X_k \\ M_1 : \hat{\eta}_1 = \hat{\beta}_{1,1} + \hat{\beta}_{1,1}X_1 + \ldots + \hat{\beta}_{1,k-1}X_{k-1} + 0\ X_k \end{cases}$$

Given the two models estimates, the value of the observed statistic test is

$$T_{oss} = \Delta Dev(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, y) = 2(\ell_{M_0}(\hat{\boldsymbol{\beta}}_0) - \ell_{M_1}(\hat{\boldsymbol{\beta}}_1))$$



If the observed values of the statistic test are greater than the quantile $1 - \alpha$ of a χ^2_1 (3.84) we will reject the null hypothesis.

Variable Selection and Hypothesis Test

An example

During the Challenger disaster (example of the previous chapter), the engineering tested if also other factors influenced the probability of o-rings damage. Another investigated factor was the pressure used to check the presence of leaks.

	temp	pres	fail
4/12/81	66	50	0
11/12/81	70	50	1
3/22/82	69	50	0
11/11/82	68	50	0
4/4/83	67	50	0
6/18/83	72	50	0
8/30/83	73	50	0
11/28/83	70	100	0
2/3/84	57	100	1
4/6/84	63	200	1
8/30/84	70	200	1
10/5/84	78	200	0

	temp	pres	fail
11/8/84	67	200	0
1/24/85	53	200	2
4/12/85	67	200	0
4/29/85	75	200	0
6/17/85	70	200	0
7/29/85	81	200	0
8/27/85	76	200	0
10/3/85	79	200	0
10/30/85	75	200	2
11/26/85	76	200	0
1/12/86	58	200	1

The increasing pressure during the tests may have generated additional damage to the o-rings.

Variable Selection and Hypothesis Test
An example

The linear logistic model becomes the following:

$$logit(Pr(Y_i = 1)) = logit(\pi_i) = \eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

where $X_{1\cdot}$ is the temperature and $X_{2\cdot}$ is the pressure.

Question. Had the pressure also influenced the probability to fail?

We can have a response with this Hypothesis Test:

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

Variable Selection and Hypothesis Test
An example

We have estimated two nested models. A Full model (1, M_0) and the model with β_2 setted to 0 (2, M_1).

Table: Estimated logistic models

	Dependent variable:	
	I(fail >0)	
	(1)	(2)
temp	-0.242^{**}	-0.232^{**}
	(0.110)	(0.108)
pres	0.010	
	(0.009)	
Constant	14.360^{*}	15.043^{**}
	(7.443)	(7.379)
Observations	23	23
Log Likelihood	−9.486	−10.158
Akaike Inf. Crit.	24.972	24.315
Note:	*p<0.1; **p<0.05; ***p<0.01	

The Wald test indicated the coefficient for the pressure is not statistically significant (p-value >0.05) and it can be removed.

Variable Selection and Hypothesis Test
An example

But in case of dummy variables (a dymmy variable has more than two coefficients) the use of a single case p-value is not a praticable solution.
Using a Deviance testing approach, the 2 times the difference between the two log-likelihoods is equal to

$$2(\ell_{M1} - \ell_{M0}) = 2(-9.486 - (-10.158)) = 1.34$$

The 0.95 quantile of a χ^2_1 (with 1 degree of freedom) at level 0.05[7] is 3.84.

$$Pr(\chi^2_1 < 3.84) = 0.95$$

Since 1.34<3.84, we accept the null Hypothesis. The pressure did not influence the probability of fail[8].
Other potential variable selection can follow computational indices (Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), ...).

[7] α is set to 0.05, that is a commonly choice, and 1-α=0.95

[8] For a complete statistical analysis of the Challenger disaster see this manuscript. (https://www.jstor.org/stable/2290069).

Diagnostics for GLMs
Source: 15.4 (Faraway, 2014)

Most of the diagnostics for linear models can be extended relatively straightforwardly to GLMs. Several types of residuals can then be defined. We considered:

- **Raw or response residuals**: it is simply the difference between the observed values and the estimated ones $Y_i - \hat{\mu}_i = Y_i - g(\hat{\eta}_i)^{-1}$.
- **Pearson residuals**: they are the equivalent of the standardized residuals. They can obtained by:

$$\frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(Y_i|\eta_i)}}$$

  where $\hat{Var}(Y_i|\eta_i)$ is the expected conditional variance.
- **Standardized Pearson residuals**:

$$R_{pi} \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(Y_i|\eta_i)(1 - h_i)}}$$

  where h_i is the i-th value of the diagonal of the hat matrix $\mathbb{H}$.

Diagnostics for GLMs
Source: 15.4 (Faraway, 2014)

- **Deviance residuals**: G_i, are the square roots of the casewise components of the residual deviance attaching the sign of the corresponding response residual.

$$G_i = sgn(y_i - \hat{\mu}_i)\sqrt{2(\frac{y_i(g(y_i) - g(\hat{\mu}_i)) - b(g(y_i) - g(\hat{\mu}_i))}{a_i})}$$

Family	*Residual Deviance*
Gaussian	$\sum (Y_i - \widehat{\mu}_i)^2$
Binomial	$2\sum\left[n_i Y_i \log_e \frac{Y_i}{\widehat{\mu}_i} + n_i(1-Y_i)\log_e \frac{1-Y_i}{1-\widehat{\mu}_i}\right]$
Poisson	$2\sum\left[Y_i \log_e \frac{Y_i}{\widehat{\mu}_i} - (Y_i - \widehat{\mu}_i)\right]$
Gamma	$2\sum\left[-\log_e \frac{Y_i}{\widehat{\mu}_i} + \frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right]$
Inverse-Gaussian	$\sum \frac{(Y_i - \widehat{\mu}_i)^2}{Y_i \widehat{\mu}_i^2}$

Diagnostics for GLMs
Source: 15.4 (Faraway, 2014)

- **Standardized Deviance residuals** are

$$R_{pi}\frac{G_i}{\sqrt{1-h_i}}$$

The analysis of residuals consists in contrasting the residuals of the fitted model with that expected by their theoretical model.

Among the others, the deviance residuals are often used since they approximately follow a normal distribution (better than Pearson deviance residuals).

Diagnostics for GLMs

Outliers or Influential points
Outliers can be observed by the quantile-quantile diagram. Residuals very far from the remaining points imply potential outliers.
As in LMs, Influential points can be obtained by this approximation of the Cook distance

$$d_i = \frac{R_{Pi}^2}{k+1}\frac{h_i}{1-h_i}$$

where k is the number of parameters, and h_i is the diagonal value of the Hat matrix $\mathbb{H}$. Values higher than 1 can be pointed as an outlier or influential observation.

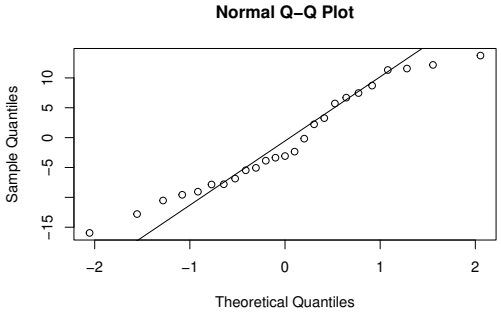
Diagnostics for GLMs

The Diagnostics for GLMs is similar to the diagnostic for LMs:

- Approximate Normality distribution of the Deviance residuals (i.e. **qqplot** or Shapiro-Wilk test):
- Presence of eteroschedasticiy or overdispersion;
- Cheking the structural part of the model (i.e. relationship of residuals with regressiors);
- Looking for unusual observations or Influential points;
- Collinearity Diagnostics.

Diagnostics for GLMs

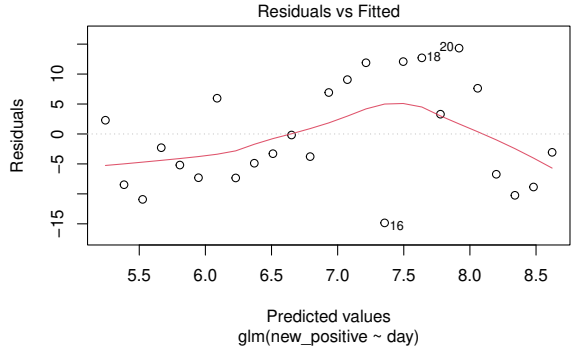
QQ-plot: we assess the presence of a normal distribution for deviance residuals. Before we analysed the trend of the first 25 days of the epidemic in Italy with a Poisson regression model. Here is the Quantile-quantile diagram of the Deviance residuals.



The normality of the deviance residuals is not so far.

Diagnostics for GLMs

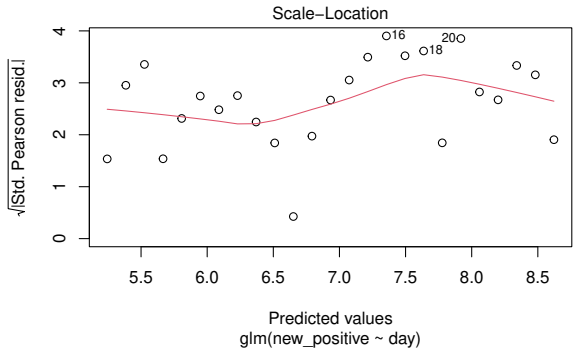
Raw Residuals vs Predicted value: the presence of a flat trend is good.



The parabolic trend can be meant that a quadratic term can be added ($+x^2$) in the model.

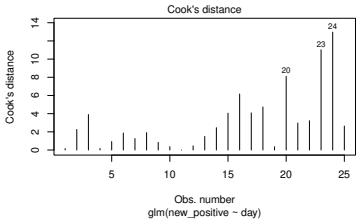
Diagnostics for GLMs

Scale-location: the presence of a flat trend is good.

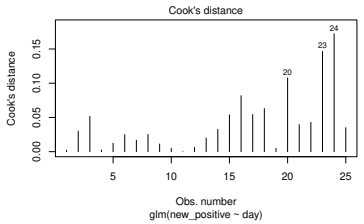


Diagnostics for GLMs

Cook distance: values higher than 4/(N-k-1) indicate outliers/influential points (N is the number of observation and k is the number of predictors)



All the observations, the Cook distance is higher than 1/(25-2-1)=0.0455. This may suggest the presence of overdispersion. The same plot with an over dispersion parameter estimated is reported below:



Collinearity Diagnostics for GLMs

For detecting collinearity several methods can be employed. We focus our attention to the Generalized Variance Inflation Factor (GVIF) index. Starting with a simple model

$$\eta = \alpha + \beta_1 X_1 + \beta_2 X_2$$

The VIF (Variance Inflation Factor) is defined as:

$$\text{VIF} = \frac{1}{1 - R^2}$$

where R^2 is quote of variance explained by X_1 to X_2 (or viceversa).
Higher values of VIF indicate the presence of collinearity.
An example:
If I have a R^2 equal to 0.8 in a linear regression between X_1 and X_2, the two variables should be highly (linearly) correlated. The resulting VIF is below

$$\text{VIF} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.8} = \frac{1}{0.2} = 5$$

Collinearity Diagnostics for GLMs

In a more complex settings formed by

$$\eta = \alpha + \boldsymbol{\beta} X + \boldsymbol{\beta} Z$$

where X is a matrix (p_1 x n) matrix and Z is a (p_2 x n) matrix, the GVIF can be obtained by

$$GVIF = \frac{\det R^1 \det R^2}{\det R}$$

there R^1, R^2 and R are the correlation matrix of X, Z and (X,Z) togheter.
To make generalized variance-inflation factors comparable across dimensions, Fox and Monette suggest reporting $GVIF^{\frac{p}{2}}$.
The threshold is equal to $10^{\frac{p}{2}}$ to assess collinearity. In presence of only a regressor with a dimension equal to 1, the threshold is equal to 3.16.
Larger values indicate the presence of collinearity.

Collinearity Diagnostics for GLMs

We consider the IRIS dataset. We want to assess, in a full model settings, the presence of collinearity
The model is the following:

$$\mathbb{E}[Y_i] = \pi_i = g(\eta)^{-1} = exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i})$$

where

- Y = 1 if the specie is versicolor (0 otherwise)
- X_{1i} is the Sepal Length
- X_{2i} is the Sepal Width
- X_{3i} is the Petal Length
- X_{4i} is the Petal Width

Collinearity Diagnostics for GLMs

	Estimate	**Std. Error**	Z value	Pr(>|z|)
(Intercept)	7.378	2.499	2.952	0.003
Sepal.Length	-0.245	0.650	-0.378	0.706
Sepal.Width	-2.797	0.784	-3.569	0.000
Petal.Length	1.314	0.684	1.921	0.055
Petal.Width	-2.778	1.173	-2.368	0.018

Petal and sepal width are statistically significant (p<0.05). We can for example drop the first variable Sepal Length that is not influent.

Collinearity Diagnostics for GLMs

However the GVIF calculated for each regressor is the following

Variable	GVIF
Sepal.Length	6.86
Sepal.Width	1.51
Petal.Length	27.93
Petal.Width	14.80

Three variables report a large GVIF value. After removing Sepal length (non significant) we obtain

Variables	GVIF
Sepal.Width	1.09
Petal.Length	12.76
Petal.Width	12.60

Petal Length and Petal Width are collinear (cor=0.963). We can drop from the model one of them.
In the end, the final model will contain **only** Sepal Width.

To be continued...

Frequent idioms on statistics:

- If you torture the numbers long enough, they will confess everything.
- A single death is a tragedy, a million deaths are a statistic.
- If you eat two chickens and I don't, statistically it turns out that we ate one each.
- If you want to inspire confidence, provide a lot of statistics.

Thanks for your attention.