Statistics for Data Science

Lecture Notes

Victor M. Panaretos

Chair of Mathematical Statistics
Institute of Mathematics
Swiss Federal Institute of Technology, Lausanne

MATH - 413

September 2018

Logistics

- Instructor: yours truly.
- TAs: Laya Ghodrati, Tomas Masak, Matthieu Simeoni, Kartik Waghmare
- Schedule:
  - Lectures: Mondays 12:00–14:00 (CE 1 3) + Tuesdays 14:00–16:00 (CM 1 5)
  - Exercises: Wednesdays 13:00-15:00 (CE 1 100 and CE 1101)
- Midterm: Date TBA
- Final Exam: January Exam Period, Written
- Course website: `http://smat.epfl.ch/courses/datasci.php`
- Lecture notes: =slides (more references on website).

What is Statistics?

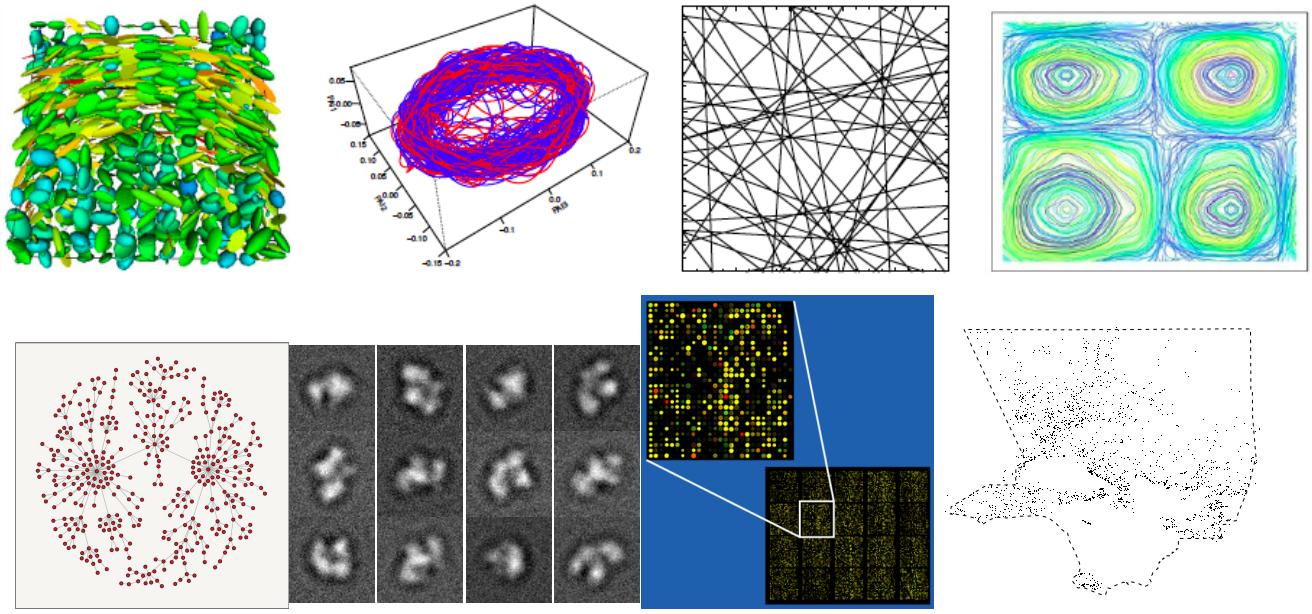
Learning from Data under Uncertainty

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

Ronald A. Fisher

(take note of: particular, general, uncertainty, rigour)

Anything[1]



[1] That we can mathematically represent

My own opinion (others might object):

$$\text{Machine Learning} \subset \text{Statistics}$$

(this does *not* mean that all machine learning was invented by statisticians!)

In some ways, machine learning corresponds one of two cultures in statistics[2]:

1. Inference/Modeling/Uncertainty
   - ↪ Focus on interpretability, reduction, and statistical efficiency.
   - ↪ Traditionally linked with science.
2. Prediction/Algorithms/Optimisation
   - ↪ Focus on emulation, automation, and computational efficiency.
   - ↪ Linked more with technology.

The two need not be mutually exclusive – depends on the problem and client!

Cultural differences run deep – philosophy of mind:

what does it mean to know/learn?

[2] Breiman (2001), "Statistical Modeling: The Two Cultures" – make sure to read discussion.

Uncertainty may stem from many sources:

1. Measurement error.
2. Chaos.
3. Intrinsic stochasticity.
4. Sampled data (observe the *particular* but not the *general*).
5. Fundamental limitations to precision.

⋮

Will use Kolmogorov's axiomatic system of probability theory to mathematically encapsulate uncertainty.

Probability:

1. Process of interest conceptualised as a probability model
2. Use model to learn about probability of potential outcomes.

Statistics:

1. Process of interest conceptualised as a probability model
2. Data viewed as observed outcomes from model
3. Use outcomes to learn about the model.

The Job of the Probabilist

Given a probability model $\mathbb{P}$ on a space Ω find the probability $\mathbb{P}[A]$ that the outcome of the experiment is $A \subset \Omega$.

The Job of the Statistician

Given an outcome of $A \subset \Omega$ (the data) of a probability experiment on Ω, tell me something *interesting** about the (uknown) probability model $\mathbb{P}$ that generated it.

(*something in addition to what I knew before observing the outcome A)

Such *interesting* questions can be:

1. Are the data more more consistent with one or another model?
2. Given a family of models, can we determine which model generated the data?
3. What range of models are consistent with a given set of data?
4. How to best answer these questions? (is there even a best way?)

Example (A Probabilist and a Statistician Flip a Coin)

Let $Y_1, ..., Y_{10}$ denote the results of flipping a coin ten times, with

$$Y_i = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}, \quad i = 1, ..., 10.$$

A plausible model is $Y_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$. We record the outcome

$$(0, 0, 0, 1, 0, 1, 1, 1, 1, 1).$$

Probabilist Asks:

- Probability of outcome as function of θ?
- Probability of k-long run?
- If keep tossing, how many k-long runs? How long until k-long run?
- What about the sum of observations? How does it behave? How does it scale?

Example (A Probabilist and a Statistician Flip a Coin (cont'd))

Statistician Asks:

- Is the coin fair?
- What is a good guess of the value of θ on the basis of the observations?
- What range of θ is plausible on the basis of the observations?
- How much error do we make when trying to decide the above from the observations?
- How does our answer change if the observations are perturbed?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from the model?
- How do our "answers" behave as # tosses $\longrightarrow \infty$?
- How many tosses would we need until we can get "accurate answers"?

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

Seems too simple?

$\rightarrow$ Spans essence of most of the ideas used in the most complex of problems!

$\rightarrow$ In principle, $\mathcal{Y}^n$ and Θ can be quite complicated, though:

- Almost always $\mathcal{Y}^n \subseteq \mathbb{R}^n$.
- Typically $\Theta \in \mathbb{R}^p$, some fixed p. Sometimes Θ is a function space. These are the parametric vs nonparametric regimes.

Prototypical Statistical Inference Tasks

1. **Estimation**. Given realisation $(Y_1, \ldots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

2. **Hypothesis Testing**. Given two disjoint regions Θ_0 and Θ_1, which is more plausible to contain the true θ that generated our observation $(Y_1, \ldots, Y_n)^\top$?

3. **Confidence Intervals**. Instead of estimating a unique θ that may have generated $(Y_1, \ldots, Y_n)^\top$, how can we give a whole range of θ that are plausible on the basis of $(Y_1, \ldots, Y_n)^\top$?

Additional tasks that can be formulated as versions/extensions of the above are:

- **Prediction**. Given data $(Y_1, \ldots, Y_n)^\top$ from a distribution $F(y_1, ..., y_n; \theta)$ where θ is unknown, predict a future outcome from the same distribution.
- **Classification**. Given observations $(Y_1^{(i)}, \ldots, Y_n^{(i)})^\top$ from various distributions $F(y_1, ..., y_n; \theta_i)$ (where $i = 1, ..., k$) depending on unknown parameters, and given a new observation Y, declare which of these distribution generated the observation Y.

Absence or Presence of Covariates

Can distinguish between two broad types of inference settings:

1. **Marginal Inference.** Here $(Y_1, ..., Y_n)^\top$ has i.i.d. entries each from the same distribution $F(y; \theta)$ with the same parameter θ.
   - In other words, all observations were obtained under identical experimental conditions, and thus depend in the same way on the same unknown θ.

2. **Regression.** Here $(Y_1, ..., Y_n)^\top$ has independent entries, each with distribution $F(y; \theta_i)$ of the same family but with different parameters.
   - Each observation was generated under slightly different experimental conditions. They depend in a similar way on different θ_i.
   - These θ_i correspond to different experimental conditions, say x_i.
   - Each x_i is called a covariate/feature, and is an input that the experimenter can vary. They are known. The index i reminds us that it corresponds to the ith observation Y_i.
   - Usually θ_i is postulated to have a special relationship to x_i, for example $\theta_i = \exp\{\alpha + \beta x_i\}$, for (α, β) uknown parameters.
   - The point here is to understand the effect of varying the covariate/feature on the distribution of the observable.

Example

1. **Marginal Inference.** Here we have independent realisations $(Y_1, ..., Y_n)$ each from the same distribution F_θ.
   - For instance, Y_i represents the outcome of flipping the same coin independently, and we wish to understand the probability of success.
   - Then we can model $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$.

2. **Regression.** Here we have independent realisations $(Y_1, ..., Y_n)$ each from a distribution F_{θ_i} of the same family but with different parameters.
   - For instance, Y_i represents the voting intention of the ith voter on a referendum. The corresponding feature may be his/her income level x_i.
   - We may model the probability of the ith voter voting 1 as $\text{Bernoulli}(\theta_i)$ with
   $$\theta_i = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$
   for (α, β) a pair of unknown parameters.
   - Note the parsimony: even though each θ_i is different, there are not n unknown parameters but only 2 unknown parameters.

Course Structure

Probabilistic background

1. Probability.
2. Reminder on Basic Probability Distributions.
3. Entropy and Exponential Families.

Sampling Theory

1. Sufficient Statistics.
2. Sampling distributions.
3. Stochastic convergence

Marginal Inference

1. Point Estimation and Likelihood Theory.
2. Hypothesis Testing and Confidence Intervals.
3. Nonparametric marginal inference and smoothing.

Inference with covariates (Regression)

1. Gaussian Linear Models.
2. Generalised Linear Models
3. Nonparametric regression and regularisation.

- This is intended to be (and will be) a mathematical course.
- There will be some proofs.
  - ↪ Proofs marked with an $*$ will not be examined, though.
- We will start from first principles and build up our theory.
- There will be the inevitable "elegantification"
- Reality is messy, complicated, and does not easily submit to narrative.

Still, mathematics is the best tool we've got.

The object of rigor is to sanction and legitimize the the conquests of intuition, and there was never any other object for it.

Jacques Hadamard

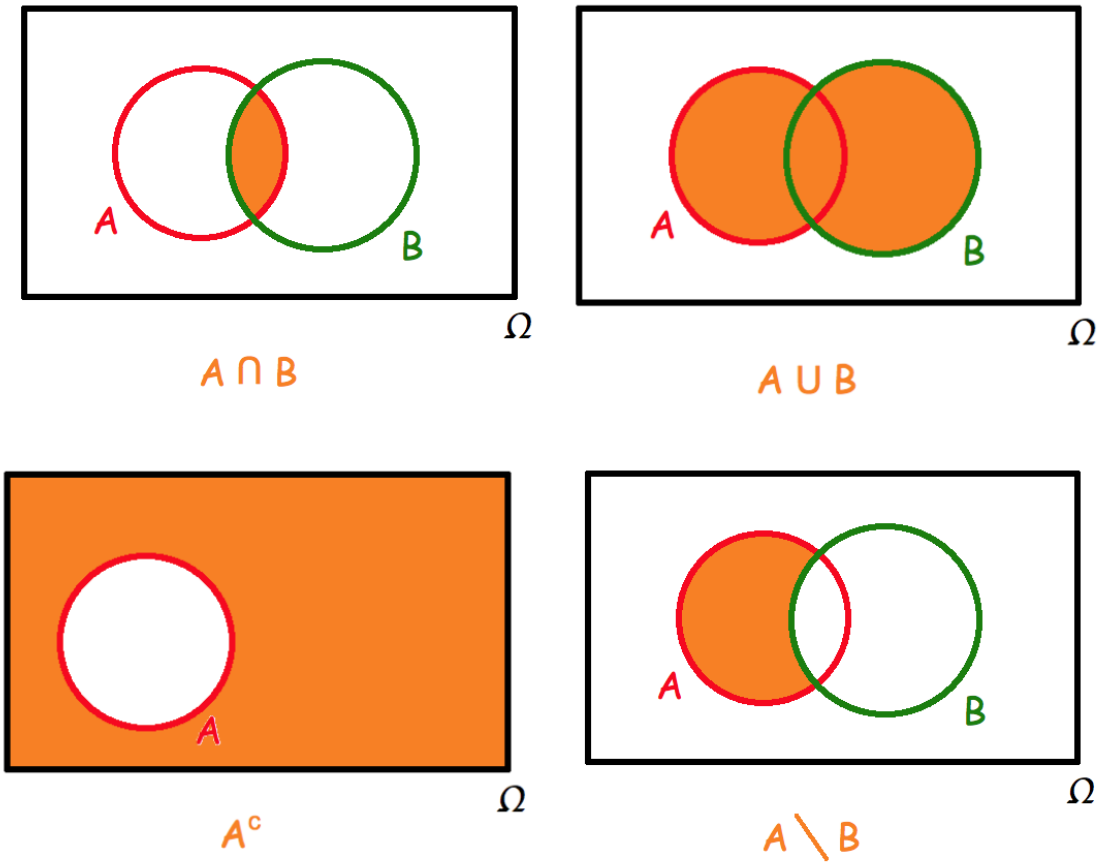
Probability Snapshot

Random experiment: process whose outcome is uncertain.
Outcomes and any statement involving them must be expressed via set theory.

- A possible outcome ω of a random experiment is called an elementary event.
- The set of all possible outcomes, say Ω is assumed non-empty, $\Omega \neq \emptyset$.
- An event is a subset $F \subset \Omega$ of Ω. An event F "is realised" (or "occurs") whenever the outcome of the experiment is an element of F.
- The union of two events F_1 and F_2, written $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs. Equivalently, $\omega \in F_1 \cup F_2$ if and only if $\omega \in F_1$ or $\omega \in F_2$,
$$F_1 \cup F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ or } \omega \in F_2\}$$
- The intersection of two events F_1 and F_2, written $F_1 \cap F_2$ occurs if and only both F_1 and F_2 occur. Equivalently, $\omega \in F_1 \cap F_2$ if and only if $\omega \in F_1$ and $\omega \in F_2$,
$$F_1 \cap F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ and } \omega \in F_2\}$$
- Unions and intersections of several events, $F_1 \cup \ldots \cup F_n$ and $F_1 \cap \ldots \cap F_n$ are defined iteratively from the definition for unions and intersections of pairs.

- The complement of an event F, denoted F^c, contains all the elements of Ω that are not contained in F,
$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$
- Two events F_1 and F_2 are called disjoint if the contain no common elements, that is $F_1 \cap F_2 = \emptyset$.
- A partition $\{F_n\}_{n \geq 1}$ of Ω is a collection of events such that $F_i \cap F_j = \emptyset$ for all $i \neq j$, and $\cup_{n \geq 1} F_n = \Omega$.
- The difference of two events F_1 and F_2 is defined as $F_1 \setminus F_2 = F_1 \cap F_2^c$. It contains all the elements of F_1 that are not contained in F_2. Notice that the difference is not symmetric: $F_1 \setminus F_2 \neq F_2 \setminus F_1$.
- It can be checked that the following properties hold true
  (i) $(F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3$
  (ii) $(F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3$
  (iii) $F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)$
  (iv) $F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)$
  (v) $(F_1 \cup F_2)^c = F_1^c \cap F_2^c$ and $(F_1 \cap F_2)^c = F_1^c \cup F_2^c$

Venn Diagrams



Probability Measures

Probability measure $\mathbb{P}$: real function defined over the events of Ω, assigning a probability to any event.

Interpreted as a measure of how certain we are that the event will occur.

Postulated to satisfy the following properties:

1. $\mathbb{P}(F) \geq 0$, for all events F.
2. $\mathbb{P}(\Omega) = 1$.
3. If an event F is a countable union $F = \cup_{n \geq 1} F_n$ of disjoint events $\{F_n\}_{n \geq 1}$,

$$\mathbb{P}(F) = \sum_{n \geq 1} \mathbb{P}(F_n).$$

The following properties are immediate consequences of the probability axioms:

- $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$.
- $\mathbb{P}(F_1 \cap F_2) \leq \min\{\mathbb{P}(F_1), \mathbb{P}(F_2)\}$.
- $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$.
- Continuity from below: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \subseteq F_{j+1}$ for all j, and let F be an event given by $F = \cup_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- Continuity from above: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \supseteq F_{j+1}$ for all j, and let F be an event given by $F = \cap_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- If $\Omega = \{\omega_1, ..., \omega_K\}$, $K < \infty$, is a finite set, then for any event $F \subseteq \Omega$, we have $\mathbb{P}(F) = \sum_{j:\omega_j \in F} \mathbb{P}(\omega_j)$.

Conditional Probability and Independence

Suppose we don't know the precise outcome $\omega \in \Omega$ that has occurred, but we are told that $\omega \in F_2$ for some event F_2, and are asked to now calculate the probability that $\omega \in F_1$ also, for some other event F_1.

- For any pair of events F_1, F_2 such that $\mathbb{P}(F_2) > 0$, we define the conditional probability of F_1 given F_2 to be

$$\mathbb{P}(F_1|F_2) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_2)}.$$

- Let G be an event and $\{F_n\}_{n\geq 1}$ be a partition of Ω such that $\mathbb{P}(F_n) > 0$ for all n. We then have:
  - Law of total probability: $\mathbb{P}(G) = \sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)$
  - Bayes' theorem: $\mathbb{P}(F_j|G) = \frac{\mathbb{P}(F_j \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(G|F_j)\mathbb{P}(F_j)}{\sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)}$
- The events $\{G_n\}_{n\geq 1}$ are called independent if and only if for any finite sub-collection $\{G_{i_1}, \ldots, G_{i_K}\}$, $K < \infty$, we have:

$$\mathbb{P}(G_{i_1} \cap \cdots \cap G_{i_K}) = \mathbb{P}(G_{i_1}) \times \mathbb{P}(G_{i_2}) \times \ldots \times \mathbb{P}(G_{i_K})$$

Random variables: numerical summaries of the outcome of a random experiment.

They allow us to not worry too much about precise structure of outcome $\omega \in \Omega$

We can concentrate on range of a random variable, rather than consider Ω.

- A random variable is a real function $X : \Omega \to \mathbb{R}$.
- We write $\{a \leq X \leq b\}$ to denote the event

$$\{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

  More generally, if $A \subset \mathbb{R}$ is a more general subset, we write $\{X \in A\}$ to denote the event

$$\{\omega \in \Omega : X(\omega) \in A\}.$$

- If we have a probability measure defined on the events of Ω, then X induces a new probability measure on subsets of the real line. This is described by the distribution function (or cumulative distribution function) $F_X : \mathbb{R} \to [0, 1]$ of a random variable X (or the law of X),

$$F_X(x) = \mathbb{P}(X \leq x).$$

- By its definition, a distribution function satisfies the following properties:

  (i) $x \leq y \;\Rightarrow\; F_X(x) \leq F_X(y)$

  (ii) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$

  (iii) $\lim_{y\downarrow x} F_X(y) = F_X(x)$, that is, F_X is right-continuous.

  (iv) $\lim_{y\uparrow x} F_X(y)$ exists, that is, F_X is left-limited.

  (v) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

  (vi) $\mathbb{P}(X > a) = 1 - F(a)$.

  (vii) Let $D_X := \{x \in \mathbb{R} : F_X(x) - \lim_{y\uparrow x} F_X(y) > 0\}$ be the set of points where F_X is not continuous.
  - D_X is a countable set.
  - If $\mathbb{P}(\{X \in D_F\}) = 1$ then X is called a *discrete* random variable (equivalently, X has a finite or countable range, with probability 1).
  - If $D_X = \emptyset$ then X is called a *continuous* random variable (the distribution function F_X is continuous).
  - It may very well happen that a random variable may be neither discrete nor continuous.

Given a probability $\alpha \in (0,1)$, which is the (smallest) real number x such that $\mathbb{P}[X \leq x] = \alpha$?

Let X be a random variable and F_X be its distribution function. We define the quantile function of X (or equivalently of F_X) to be the function

$$F_X^- : (0,1) \to \mathbb{R}$$

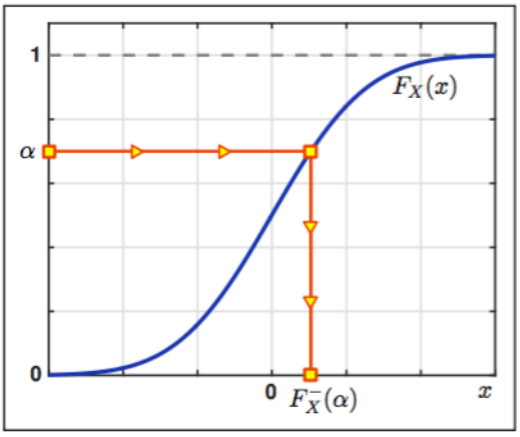
$$F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}.$$

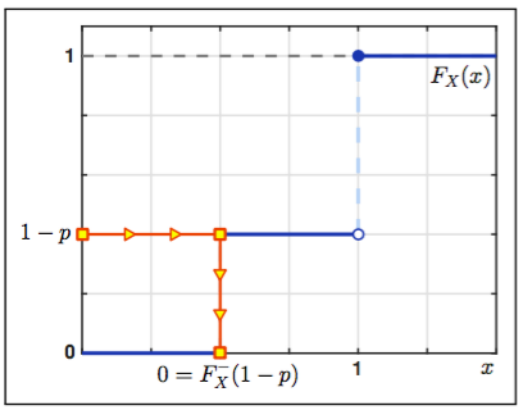
- If F_X is strictly increasing and continuous, then $F_X^- = F_X^{-1}$

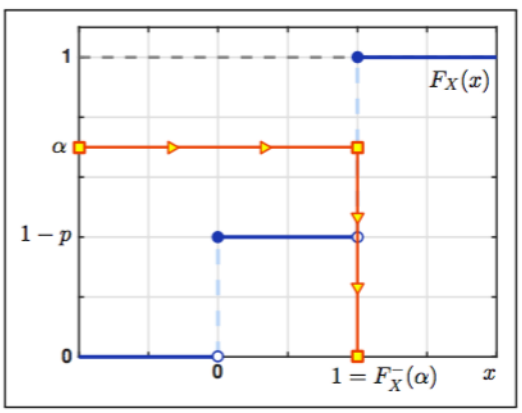
Given an $\alpha \in (0,1)$, the α-quantile of X (or equivalently of F_X) is the real number

$$q_\alpha = F_X^-(\alpha).$$

The pictures that say it all:







Lemma

Let $Y \sim Unif(0,1)$ and let F be a distribution function. Then, the distribution function of the random variable $X = F^{-}(Y)$ is given precisely by F.

- Can be used to generate realisations from any distribution
  - *Provided we can generate realsations from uniform on $[0,1]$.*
  - *Can do this with binary expansions and Bernoulli draws.*
  - *Reduces problem to infinite coin flipping.*

Partial Converse

Let X be a random variable with strictly increasing and continuous distribution function F_X. Then, $F_X(X) \sim Unif(0,1)$

The probability mas function (or frequency function) $f_X : \mathbb{R} \to [0,1]$ of a discrete random variable X is defined as

$$f_X(x) = \mathbb{P}(X = x).$$

By its definition, a probability mass function satisfies

(i) $\mathbb{P}(X \in A) = \sum_{t \in A \cap \mathcal{X}} f_X(t)$, for $A \subseteq \mathbb{R}$ and $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

(ii) $F_X(x) = \sum_{t \in (-\infty, x] \cap \mathcal{X}} f_X(t)$, for all $x \in \mathbb{R}$ and $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

(iii) An immediate corollary is that $F_X(x)$ is piecewise constant with jumps at the points in $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

A continuous random variable X has probability density function $f_X : \mathbb{R} \to [0, +\infty)$ if

$$F_X(b) - F_X(a) = \int_a^b f_X(t)\,dt.$$

for all real numbers $a < b$. By its definition, a probability density satisfies

(i) $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$

(ii) $f_X(x) = F_X'(x)$, whenever f_X is continuous at x.

(iii) Note that $f_X(x) \neq \mathbb{P}(X = x) = 0$. In fact, it can be $f(x) > 1$ for some x. It can even happen that f is unbounded.

Let X be discrete, taking values in $\mathcal{X}$, and define $Y = g(X)$. Then, Y takes values in $\mathcal{Y} = g(\mathcal{X})$ and

$$\begin{aligned} F_Y(y) = \mathbb{P}[g(X) \leq y] &= \sum_{x \in \mathcal{X}} f_X(x) \mathbf{1}\{g(x) \leq y\}, \qquad \forall y \in \mathcal{Y} \\ f_Y(y) = \mathbb{P}[g(X) = y] &= \sum_{x \in \mathcal{X}} f_X(x) \mathbf{1}\{g(x) = y\}, \qquad \forall y \in \mathcal{Y}. \end{aligned}$$

Let X be continuous, taking values in $\mathcal{X} \subseteq \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ a transformation that is

1. monotone,
2. continuously differentiable,
3. with non-vanishing derivative.

If $Y = g(X)$, then Y takes values in $\mathcal{Y} = g(\mathcal{X})$ and

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \qquad y \in \mathcal{Y}.$$

A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is a finite collection of random variables (arranged as the coordinates of a vector)

The point is that we may want to make probabilistic statements on the joint behaviour of all these random variables.

- The joint distribution function of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is defined as:
$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$
- Correspondingly, one defines the
  - joint frequency function, if the $\{X_i\}_{i=1}^d$ are all discrete,
  $$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$
  - the joint density function, if there exists $f_{\boldsymbol{X}} : \mathbb{R}^d \to [0, +\infty)$ such that:
  $$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{\boldsymbol{X}}(u_1, \ldots, u_d)\, du_1 \ldots du_d$$
  In this case, when $f_{\boldsymbol{X}}$ is continuous at the point $\boldsymbol{x}$,
  $$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_{\boldsymbol{X}}(x_1, \ldots, x_d)$$

Marginal Distributions

Given the joint distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, we can isolate the distribution of a single coordinate, say X_i.

- discrete case, the marginal frequency function of X_i is given by

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$$

- In the continuous case, the marginal density function of X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d)\, dy_1 \ldots dy_{i-1}\, dy_{i+1}\, dy_d.$$

- More generally, we can define the joint frequency/density of a random vector formed by a subset of the coordinates of $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, say the first k
  - Discrete case:
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)$.
  - Continuous case
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)\, dx_{k+1} \ldots dx_d$.
- I.e. to marginalise we integrate/sum out the remaining random variables from the overall joint density/frequency.
- Marginals do not uniquely determine the joint distribution.

Conditional Distributions

We may wish to make probabilistic statements about the potential outcomes of one random variable, if we already know the outcome of another.

For this we need the notion of a conditional density/frequency function.

If $(X_1, \ldots, X_d)$ is a continuous/discrete random vector, we define the conditional probability density/frequency function of $(X_1, \ldots, X_k)$ given $\{X_{k+1} = x_{k+1}, \ldots, X_d = x_d\}$ as

$$f_{X_1,\ldots,X_k | X_{k+1},\ldots,X_d}(x_1, \ldots, x_k | x_{k+1}, \ldots, x_d) = \frac{f_{X_1,\ldots,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d)}$$

provided that $f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d) > 0$.

Independent Random Variables

The random variables $X_1, \ldots, X_d$ are called independent if and only if for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1) \times \ldots \times F_{X_d}(x_d).$$

Equivalently, $X_1, \ldots, X_d$ are independent if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d).$$

For two random variables X and Y, we denote their independence as $X \perp\!\!\!\perp Y$.

Note that when random variables are independent, conditional distributions reduce to the corresponding marginal distributions.

Knowing the value of one of the random variables gives us no information about the distribution of the rest.

Conditionally Independent Random Variables

The random vector X in $\mathbb{R}^d$ is called conditionally independent of the random vector Y given the random vector Z, written

$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y \mid Z,$$

if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d|Y,Z}(x_1, \ldots, x_d) = F_{X_1,\ldots,X_d|Z}(x_1, \ldots, x_d).$$

Equivalently, if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d|Y,Z}(x_1, \ldots, x_d) = f_{X_1,\ldots,X_d|Z}(x_1, \ldots, x_d).$$

Knowing Y in addition to knowing Z gives us no more information about X.
Consequence: if X is conditionally independent of Y given Z, then

$$F_{X,Y|Z} = F_{X|Y,Z}F_{Y|Z} = F_{X|Z}F_{Y|Z}$$

Consequence: $X \perp\!\!\!\perp_Z Y \iff Y \perp\!\!\!\perp_Z X$

Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable bijection,

$$g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x})), \qquad \boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Let $X = (X_1, \ldots, X_n)^\top$ have joint density $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, and define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top = g(\boldsymbol{X})$. Then, $\boldsymbol{Y}$ takes values in $\mathcal{Y}^n = g(\mathcal{X}^n)$, and

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y})) \left| \det \left[J_{g^{-1}}(\boldsymbol{y}) \right] \right|, \qquad \text{for } \boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathcal{Y}^n,$$

and zero otherwise, whenever $J_{g^{-1}}(\boldsymbol{y})$ is well-defined. Here, $J_{g^{-1}}(\boldsymbol{y})$ is the Jacobian of g^{-1}, i.e. the $n \times n$ matrix-valued function,

$$J_{g^{-1}}(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial y_n} g_1^{-1}(\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_1} g_n^{-1}(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial y_n} g_n^{-1}(\boldsymbol{y}) \end{bmatrix}.$$

Example (Convolution of densities)

Let X and Y be independent, continuous random variables with densities f_X and f_Y. The density of $X + Y$ is the *convolution* of f_X with f_Y:

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u - v) f_Y(v)\, dv.$$

Dfine $g : \mathbb{R}^2 \to \mathbb{R}^2$, $\quad (x, y) \overset{g}{\mapsto} (x + y, y) \quad (u, v) \overset{g^{-1}}{\mapsto} (u - v, v)$.
The Jacobian of the inverse is

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and its determinant is 1. It follows that

$$f_{X+Y,Y}(u, v) = f_{X,Y}(u - v, v) = f_X(u - v) f_Y(v),$$

and we integrate out v to find the marginal f_{X+Y}:

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u - v) f_Y(v)\, dv.$$

Moments

Expectation

The expectation (or expected value) of a random variable X formalises the notion of the "average" value taken by that random variable.

- For continuous variables:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx.$$

- For discrete variables:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

The expectation satisfies the following properties:

- Linearity: $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha\, \mathbb{E}[X_2]$.
- $\mathbb{E}[h(X)] = \sum_{x \in \mathcal{X}} h(x) f_X(x)$ (discrete case)
  or
  $\mathbb{E}[h(x)] = \int_{-\infty}^{+\infty} h(x) f(x)\, dx$ (continuous case).

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ be a random vector in $\mathbb{R}^d$ with joint density function $f_{\boldsymbol{X}}(x_1, \ldots, x_d)$. For any $g : \mathbb{R}^d \to \mathbb{R}$, we define

$$\mathbb{E}\{g(X_1, \ldots, X_d)\} = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_d) f_{\boldsymbol{X}}(x_1, \ldots, x_d)\, dx_1 \ldots dx_d.$$

Similarly, in the discrete case,

$$\mathbb{E}\{g(X_1, \ldots, X_d)\} = \sum_{x_1 \in \mathcal{X}_1} \ldots \sum_{x_d \in \mathcal{X}_d} g(x_1, \ldots, x_d) f_{\boldsymbol{X}}(x_1, \ldots, x_d).$$

The mean vector or a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is defined as

$$\mathbb{E}[\boldsymbol{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

i.e. it is the vector of means.

The variance of a random variable X expresses how disperse the realisations of X are around its expectation.

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{if } \mathbb{E}[X^2] < \infty).$$

Furthermore, the covariance of a random variable X_1 with another random variable X_2 expresses the degree of linear dependency between the two.

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{if } \mathbb{E}[X_i^2] < \infty).$$

The correlation between X_1 and X_2 is defined as

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}}.$$

Conveys equivalent dependence information to covariance. Advantages: (1) it is invariant to changes of scale, (2) can be be understood in absolute terms (ranges in $[-1, 1]$), as a result of the correlation inequality (itself a consequence of the Cauchy-Schwarz inequality):

$$|\mathrm{Corr}(X_1, X_2)| \leq \sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}.$$

Some useful formulae relating expectations, variance, and covariances are:

- $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{cov}(X, X)$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$
- $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j)$
- $\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$
- $\text{cov}(aX_1 + bX_2, Y) = a\text{cov}(X_1, Y) + b\text{cov}(X_2, Y)$
- if $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, then the following are equivalent:
  (i) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\,\mathbb{E}[X_2]$
  (ii) $\text{cov}(X_1, X_2) = 0$
  (iii) $\text{Var}(X_1 \pm X_2) = \text{Var}(X_1) + \text{Var}(X_2)$

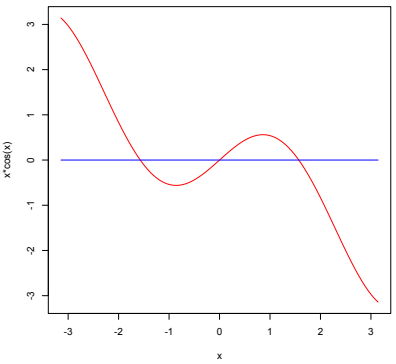
  Independence will imply these three last properties, but none of these properties imply independence.

Example ($\text{Corr}(X, Y) = 0 \not\Rightarrow$ Independence)

Let $X \sim \text{Unif}[-\pi, \pi]$ and define

$$Y = \cos(X).$$

- Clearly X and Y are not independent.
- To the contrary, they are perfectly dependent.
- Their covariance is, nevertheless, zero!



The function $x\,cos(x)$

Concretely, we calculate

$$\mathbb{P}[Y > 0] = 1/2 \qquad \text{mais} \qquad \mathbb{P}[Y > 0 | X \in (-\pi, -2)] = 1.$$

Despite this, we have

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \int_{-\pi}^{+\pi} x \cos(x) \frac{1}{2\pi}\, dx - 0 = 0.$$

Why: Because some non-linear dependencies cannot be detected by covariance...

Example ($\mathrm{Corr}(X, Y) = 0 \not\Rightarrow$ Independence)

Let X and Y have joint density

$$f_{XY}(x, y) = \begin{cases} \pi & \text{si } x^2 + y^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ by symmetry. Hence, $\mathrm{Cov}\{X, Y\} = \mathbb{E}[XY]$. But

$$\mathbb{E}[XY] = \iint\limits_{x^2+y^2=1} xy\pi \, dxdy = \iint\limits_{x^2+y^2=1, y\geq 0} xy\pi \, dxdy + \iint\limits_{x^2+y^2=1, y<0} xy\pi \, dxdy$$

The two terms are equal, by symmetry. Moreover,

$$\iint\limits_{x^2+y^2=1, y\geq 0} xy\pi \, dxdy = \pi \int_{-1}^{1} x \int_{0}^{1-x^2} y dy dx = \pi \int_{-1}^{1} x \frac{(1-x^2)^2}{2} dx = 0$$

and so the correlation is zero. But X and Y are clearly dependent, since knowing X restricts the possible values of Y.

Conditional Moments

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X \,|\, Y = y] = \begin{cases} \sum_{x \in \mathcal{X}} x \, \mathbb{P}[X = x \,|\, Y = y], & \text{if } X, Y \text{ are discrete,} \\ \int_{-\infty}^{+\infty} x \, f_{X|Y}(x|y) \, dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

- Simply the expectation of the conditional distribution.
- Note that calculation of $\mathbb{E}[X \,|\, Y = y] = q(y)$ results in a function of y.
- One can plug Y into $q(\cdot)$ and consider $Z = q(Y)$ as a random variable itself.
- Denoted by $\mathbb{E}[X \,|\, Y]$, this is the formal definition of conditional expectation.
- Important property/interpretation:

$$\mathbb{E}[X \,|\, Y] = \arg\min_{g} \mathbb{E} \, \|X - g(Y)\|^2$$

Among all functions[3] of Y, $\mathbb{E}[X \,|\, Y]$ best approximates X in mean square.

[3] measurable

Important properties of $\mathbb{E}[X \,|\, Y]$:

1. Unbiasedness: $\mathbb{E}\Big\{\mathbb{E}[X \,|\, Y]\Big\} = \mathbb{E}[X]$
2. If X independent of Y, then $\mathbb{E}[X \,|\, Y] = \mathbb{E}[X]$.
3. "Taking out known factors": $\mathbb{E}[g(Y)X \,|\, Y] = g(Y)\mathbb{E}[X \,|\, Y]$
4. "Tower property": $\mathbb{E}\Big[\mathbb{E}[X \,|\, Y]\Big|g(Y)\Big] = \mathbb{E}[X \,|\, g(Y)]$
5. Linearity: $\mathbb{E}[aX_1 + X_2 \,|\, Y] = a\mathbb{E}[X_1 \,|\, Y] + \mathbb{E}[X_2 \,|\, Y]$.
6. Monotonicity: $X_1 \leq X_2 \implies \mathbb{E}[X_1 \,|\, Y] \leq \mathbb{E}[X_2 \,|\, Y]$

The conditional variance of X given Y is defined as

$$\operatorname{var}[X \,|\, Y] = \mathbb{E}\Big[\,(X - \mathbb{E}[X \,|\, Y])^2 \,\Big|\, Y\Big] = \mathbb{E}[X^2 \,|\, Y] = \big(\mathbb{E}[X \,|\, Y]\big)^2$$

The law of total variance states that

$$\operatorname{var}(X) = \mathbb{E}\left[\operatorname{var}[X \,|\, Y]\right] + \operatorname{var}\left(\mathbb{E}[X \,|\, Y]\right)$$

Proof:

$$\begin{aligned}
\operatorname{var}(X) &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \\
&= \mathbb{E}\big[\mathbb{E}[X^2 \,|\, Y]\big] - \mathbb{E}^2\big[\mathbb{E}[X \,|\, Y]\big] \\
&= \mathbb{E}\Big[\operatorname{var}[X \,|\, Y] + \mathbb{E}^2[X \,|\, Y]\Big] - \mathbb{E}^2\big[\mathbb{E}[X \,|\, Y]\big] \\
&= \mathbb{E}\big[\operatorname{var}[X \,|\, Y]\big] + \mathbb{E}\big[\mathbb{E}^2[X \,|\, Y]\big] - \mathbb{E}^2\big[\mathbb{E}[X \,|\, Y]\big] \\
&= \mathbb{E}\left[\operatorname{var}[X \,|\, Y]\right] + \operatorname{var}\left(\mathbb{E}[X \,|\, Y]\right).
\end{aligned}$$

The covariance matrix or a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$, say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$, is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \text{cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \leq i \leq j \leq d.$$

That is, the covariance matrix encodes the variances of the coordinates of Y (on the diagonal) and the pairwise covariances between any two coordinates of $\boldsymbol{Y}$ (off the diagonal). If we write

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}] = (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])^\top$$

for the mean vector of $\boldsymbol{Y}$, then

$$\mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^\top] = \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Similarly to the vector case, the expectation of a matrix with random entries is the matrix of expectations of the random entries.

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of $\boldsymbol{A}\boldsymbol{Y}$ are $\boldsymbol{A}\boldsymbol{\mu}$ and $\boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}^\top$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, the variance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ is $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.
- If $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ are deterministic vectors, the covariance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ with $\boldsymbol{\gamma}^\top \boldsymbol{Y}$ is $\boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.

Inequalities Involving Moments

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\text{var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

For any convex[4] function $\varphi : \mathbb{R} \to \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then one has

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen]}$$

Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non decreasing function such that $\mathbb{E}[g^2(X)] < \infty$. Then,

$$\text{cov}[X, g(X)] \geq 0 \qquad \text{[Monotonicity and Covariance]}$$

[4] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.

Moment Generating Functions

Let X be a random variable taking values in $\mathbb{R}$. The moment generating function (MGF) of X is defined as

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$

$$M_X(t) = \mathbb{E}\Big[e^{tX}\Big], \qquad t \in \mathbb{R}.$$

When $M_X(t), M_Y(t)$ exist (are finite) for $t \in I \ni 0$, then:

- $\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \in \mathbb{N}$.
- $M_X = M_Y$ on I if and only if $F_X = F_Y$
- $M_{X+Y} = M_X M_Y$ when X and Y are independent

Similarly, for a random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, the MGF is

$$M_{\boldsymbol{X}}(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$$

$$M_{\boldsymbol{X}}(\boldsymbol{u}) = \mathbb{E}\Big[e^{\boldsymbol{u}^\top \boldsymbol{X}}\Big], \qquad \boldsymbol{u} \in \mathbb{R}^d.$$

and has analogous properties.

Elementary Distributions Factsheet

Bernoulli Distribution

A random variable X is said to follow the Bernoulli distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Bern}(p)$, if

1. $\mathcal{X} = \{0, 1\}$,
2. $f(x; p) = p\mathbf{1}\{x = 1\} + (1 - p)\mathbf{1}\{x = 0\}$.

The mean, variance and moment generating function of $X \sim \text{Bern}(p)$ are given by

$$\mathbb{E}[X] = p, \qquad \text{var}[X] = p(1 - p), \qquad M(t) = 1 - p + pe^t.$$

Binomial Distribution

A random variable X is said to follow the Binomial distribution with parameters $p \in (0,1)$ and $n \in \mathbb{N}$, denoted $X \sim \text{Binom}(n,p)$, if

1. $\mathcal{X} = \{0,1,2,...,n\}$,
2. $f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$.

The mean, variance and moment generating function of $X \sim \text{Binom}(n,p)$ are given by

$$\mathbb{E}[X] = np, \qquad \text{var}[X] = np(1-p), \qquad M(t) = (1-p+pe^t)^n.$$

- If $X = \sum_{i=1}^n Y_i$ where $Y_i \overset{iid}{\sim} Bern(p)$, then $X \sim \text{Binom}(n,p)$.

Geometric Distribution

A random variable X is said to follow the Geometric distribution with parameter $p \in (0,1)$, denoted $X \sim \text{Geom}(p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p) = (1-p)^x p$.

The mean, variance and moment generating function of $X \sim \text{Geom}(p)$ are given by

$$\mathbb{E}[X] = \frac{1-p}{p}, \qquad \text{var}[X] = \frac{(1-p)}{p^2}, \qquad M(t) = \frac{p}{1-(1-p)e^t},$$

the latter for $t < -\log(1-p)$.

- Let $\{Y_i\}_{i \geq 1}$ be an infinite collection of random variables, where $Y_i \overset{iid}{\sim} Bern(p)$. Let $T = \min\{k \in \mathbb{N} : Y_k = 1\} - 1$. Then $T \sim Geom(p)$.

Negative Binomial Distribution

A random variable X is said to follow the Negative Binomial distribution with parameters $p \in (0,1)$ and $r > 0$, denoted $X \sim \text{NegBin}(r,p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p,r) = \binom{x+r-1}{x}(1-p)^x p^r$.

The mean, variance and moment generating function of $X \sim \text{NegBin}(r,p)$ are given by

$$\mathbb{E}[X] = r\frac{1-p}{p}, \qquad \text{var}[X] = r\frac{(1-p)}{p^2}, \qquad M(t) = \frac{p^r}{[1-(1-p)e^t]^r},$$

the latter for $t < -\log(1-p)$.

- If $X = \sum_{i=1}^{r} Y_i$ where $Y_i \overset{iid}{\sim} Geom(p)$, then $X \sim \text{NegBin}(r,p)$.

Poisson Distribution

A random variable X is said to follow the Poisson distribution with parameters $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$.

The mean, variance and moment generating function of $X \sim \text{Poisson}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda, \qquad \text{var}[X] = \lambda, \qquad M(t) = \exp\{\lambda(e^t - 1)\}.$$

- Let $\{X_n\}_{n\geq 1}$ be a sequence of $Binom(n, p_n)$ random variables, such that $p_n = \lambda/n$, for some constant $\lambda > 0$. Then $f_{X_n} \overset{n\to\infty}{\longrightarrow} f_Y$, where $Y \sim Poisson(\lambda)$.
- Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent. The conditional distribution of X given $X + Y = k$ is $Binom(k, \lambda/(\lambda+\mu))$ (useful in contingency tables).

Multinomial Distribution

A random vector $\boldsymbol{X}$ in $\mathbb{R}^k$ said to follow the Multinomial distribution with parameters $n \in \mathbb{N}$ and $p = (p_1, \ldots, p_k) \in (0,1)^k$, such that $\sum_{i=1}^k p_i = 1$, denoted $\boldsymbol{X} \sim \text{Multi}(n; p_1, \ldots, p_k)$, if

1. the sample space is $\{0, 1, \ldots, n\}^k$, and
2. $f(x_1, \ldots, x_k; n, \{p_i\}_{i=1}^k) = \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \mathbf{1}\left\{\sum_{i=1}^k x_i = n\right\}$.

The mean, variance covariance and moment generating function are

$$\mathbb{E}[X_i] = np_i, \qquad \text{Var}[X_i] = np_i(1-p_i), \qquad \text{cov}(X_i, X_j) = -np_ip_j$$

$$M(u_1, \ldots, u_k) = \left(\sum_{i=1}^k p_i e^{u_i}\right)^n.$$

Generalises binomial: n independent trials, with k possible outcomes.

Lemma (Poisson and Multinomial)

If $X_i \sim Poiss(\lambda_i)$, $i = 1, ..., k$ are independent, then the conditional distribution of $\boldsymbol{X} = (X_1, ..., X_k)^\top$ given $\sum_{i=1}^k X_i = n$ is Multi$(n; p_1, \ldots, p_k)$, with

$$p_i = \tfrac{\lambda_i}{\lambda_1 + \ldots + \lambda_k}.$$

Uniform Distribution

A random variable X is said to follow the uniform distribution with parameters $-\infty < \theta_1 < \theta_2 < \infty$, denoted $X \sim \text{Unif}(\theta_1, \theta_2)$, if

$$f_X(x; \theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & \text{if } x \in (\theta_1, \theta_2), \\ 0 & \text{otherwise.} \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Unif}(\theta_1, \theta_2)$ are given by

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \qquad \text{var}[X] = (\theta_2 - \theta_1)^2/12, \qquad M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)},\ t \neq 0$$

$$M(0) = 1.$$

Exponential Distribution

A random variable X is said to follow the exponential distribution with parameter $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if

$$f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Exp}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \text{var}[X] = \lambda^{-2}, \qquad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

If X, Y are independent exponential random variables with rates λ_1 and λ_2, then $Z = \min\{X, Y\}$ is also exponential with rate $\lambda_1 + \lambda_2$.

Lack of memory characterisation:

1. Let $X \sim Exp(\lambda)$. Then $\mathbb{P}[X \geq x + t | X \geq t] = \mathbb{P}[X \geq x]$.
2. Conversely: if X is a random variable such that $\mathbb{P}(X > 0) > 0$ and

   $$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \qquad \forall t, s \geq 0,$$

   then there exits a $\lambda > 0$ such that $X \sim Exp(\lambda)$.

Gamma Distribution

A random variable X is said to follow the gamma distribution with parameters $r > 0$ and $\lambda > 0$ (the *shape* and *rate* parameters, respectively), denoted $X \sim \text{Gamma}(r, \lambda)$, if

$$f_X(x; r, \lambda) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Gamma}(r, \lambda)$ are given by

$$\mathbb{E}[X] = r/\lambda, \qquad \text{var}[X] = r/\lambda^2, \qquad M(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad t < \lambda.$$

- If $Y_1, \ldots, Y_r \overset{iid}{\sim} Exp(\lambda)$, then $Y = \sum_{i=1}^r Y_i \sim Gamma(r, \lambda)$ (special case is called Erlang distribution).
- The special case of $Gamma(k/2, 1/2)$ is called the chi-square distribution with k degrees of freedom and denoted by χ^2_k. Will soon see its importance.

A random variable X is said to follow the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ (the *mean* and *variance* parameters, respectively), denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, if

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

The mean, variance and moment generating function of $X \sim \mathrm{N}(\mu, \sigma^2)$ are given by

$$\mathbb{E}[X] = \mu, \qquad \mathrm{var}[X] = \sigma^2, \qquad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

In the special case $Z \sim N(0, 1)$, we use the notation $\varphi(z) = f_Z(z)$ and $\Phi(z) = F_Z(z)$, and call these the *standard normal density* and *standard normal CDF*, respectively.

Lemma

Let $X \sim N(\mu, \sigma^2)$, $a \neq 0$. Then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Condequently, if $X \sim N(\mu, \sigma^2)$, then

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where Φ is the standard normal CDF, $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} \exp\{-z^2/2\} dz$.

Corollary

Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \sim N(\mu_i, \sigma_i^2)$, and let $S_n = \sum_{i=1}^n X_i$. Then,

$$S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Entropy

Entropy

Can one probability model be "more disordered" than another?

The entropy of a random variable X is defined as

$$H(X) = -\mathbb{E}\Big\{\log f_X(X)\Big\} = \begin{cases} -\sum_{x\in\mathcal{X}} f_X(x)\log\{f_X(x)\}, & \text{if } X \text{ is discrete,} \\ -\int_{-\infty}^{+\infty} f_X(x)\log\{f_X(x)\}\,dx, & \text{if } X \text{ is continuous.} \end{cases}$$

- A measure of the intrinsic disorder or unpredictability of a random system.
- Related to but not equivalent to variance.

When X is discrete:

- $H(X) \geq 0$
- $H(g(X)) \leq H(X)$ for any deterministic function g.

KL Divergence (Relative Entropy)

Can we use entropy to compare distributions?

Let $p(x)$ and $q(x)$ be two probability density (frequency) functions on $\mathbb{R}$. We define the Kullback-Leibler divergence or relative entropy of q with respect to p as

$$KL(q||p) := \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

- By Jensen's inequality, for $X \sim p(\cdot)$ we have

$$KL(q||p) = \mathbb{E}\{-\log[q(X)/p(X)]\} \geq -\log\left\{\mathbb{E}\left[\frac{q(X)}{p(X)}\right]\right\} = 0$$

  since q integrates to 1.
- $p = q \iff KL(q||p) = 0$.
- Not a metric (lacks symmetry and violates triangle inequality).

Maximum Entropy Under Constraints

Consider the following variational problem:

Determine the probability distribution f supported on $\mathcal{X}$ with maximum entropy

$$H(f) = -\int_{\mathcal{X}} f(x) \log f(x)\, dx$$

subject to the linear constraints

$$\int_{\mathcal{X}} T_i(x) f(x)\, dx = \alpha_i, \qquad i = 1, ..., k$$

Philosophy: How to choose a probability model for a given situation?

Maximum entropy approach:

- In any given situation, choose the distribution that gives *highest uncertainty* while satisfying situation–specific required constraints.

Proposition.

When a solution to the constrained optimisation problem exists, it is unique and has the form

$$f(\boldsymbol{x}) = Q(\lambda_1, ..., \lambda_k) \exp\left\{\sum_{i=1}^{k} \lambda_i \, T_i(\boldsymbol{x})\right\}$$

Proof.

Let $g(\boldsymbol{x})$ be a density also satisfying the constraints. Then,

$$\begin{aligned} H(g) &= -\int_{\mathcal{X}} g(\boldsymbol{x}) \log g(\boldsymbol{x}) \, d\boldsymbol{x} = -\int_{\mathcal{X}} g(\boldsymbol{x}) \log \left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x})\right] d\boldsymbol{x} \\ &= -\underbrace{KL(g \| f)}_{\geq 0} - \int_{\mathcal{X}} g(\boldsymbol{x}) \log f(\boldsymbol{x}) \, d\boldsymbol{x} \\ &\leq -\log Q \underbrace{\int_{\mathcal{X}} g(\boldsymbol{x}) \, d\boldsymbol{x}}_{=1} - \int_{\mathcal{X}} g(\boldsymbol{x}) \left(\sum_{i=1}^{k} \lambda_i \, T_i(\boldsymbol{x})\right) d\boldsymbol{x} \end{aligned}$$

But g also satisfies the moment constraints, so the last term is

$$\begin{aligned} &= -\log Q - \int_{\mathcal{X}} f(\boldsymbol{x}) \left(\sum_{i=1}^{k} \lambda_i \, T_i(\boldsymbol{x})\right) d\boldsymbol{x} = -\int_{\mathcal{X}} f(\boldsymbol{x}) \log f(\boldsymbol{x}) \, d\boldsymbol{x} \\ &= H(f) \end{aligned}$$

Uniqueness of the solution follows from the fact that strict equality can only follow when $KL(g \| f) = 0$, which happens if and only if $g = f$. □

A probability distribution is said to be a member of a k-parameter exponential family, if its density (or frequency) admits the representation

$$f(y) = \exp\left\{\sum_{i=1}^{k} \phi_i \, T_i(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}$$

where:

1. $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)$ is a k-dimensional parameter in $\Phi \subseteq \mathbb{R}^k$;
2. $T_i : \mathcal{Y} \to \mathbb{R}$, $i = 1, ..., k$, $S : \mathcal{Y} \to \mathbb{R}$, and $\gamma : \mathbb{R}^k \to \mathbb{R}$, are real-valued;
3. The support $\mathcal{Y}$ of f does not depend on $\boldsymbol{\phi}$.

Very rich class of models (sometimes requiring fixing some parameters to satisfy last condition): Binomial, Negative Binomial, Poisson, Gamma, Gaussian, Pareto, Weibull, Laplace, logNormal, inverse Gaussian, inverse Gamma, Normal-Gamma, Beta, Multinomial...

↪ Basis for *Generalised Linear Models (GLM)*.

We will gradually see that such models have magnificent properties.

- $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top$ is called the natural parameter
- But transforming parameter, we can write exponential family in other ways.
- "Natural" is from the mathematics point of view – usual parameter $\boldsymbol{\theta} = \eta^{-1}(\boldsymbol{\phi})$ often different.

Natural vs Usual Parametrization

$$\exp\left\{\sum_{i=1}^{k} \phi_i \, T_i(y) - \gamma(\boldsymbol{\phi}) + S(y)\right\} = \exp\left\{\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta}) \, T_i(y) - d(\boldsymbol{\theta}) + S(y)\right\}.$$

where $\eta : \mathbb{R}^k \to \mathbb{R}^k$ is a C^2 map such that

$$\boldsymbol{\phi} = \eta(\boldsymbol{\theta})$$

and so $\gamma(\boldsymbol{\phi}) = \gamma(\eta(\boldsymbol{\theta})) = d(\boldsymbol{\theta})$, for $d = \gamma \circ \eta$.

- Natural parametrization: great for mathematical manipulation.
- Usual parametrization: more intuitive in context of applications.

Example (Binomial Exponential Family)

Let $Y \sim \text{Binom}(n, p)$. Observe that:

$$\binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{ \log\left(\frac{p}{1-p}\right) y + n \log(1-p) + \log\binom{n}{y} \right\}.$$

Define

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(y) = y,$$

$$S(y) = \log\binom{n}{y}, \quad \gamma(\phi) = n\log(1+e^{\phi}) = -n\log(1-p).$$

Keeping n fixed and allowing only p to vary, the support of f does not depend on ϕ and we get a 1-parameter exponential family. Note that:

$$p = \frac{e^{\phi}}{1+e^{\phi}} \qquad \& \qquad \phi = \underbrace{\log\left(\frac{p}{1-p}\right)}_{=\eta(p)}.$$

so the usual parameter is $p \in (0,1)$, but the natural one is $\phi \in \mathbb{R}$. □

Example (Gaussian Exponential Family)

Let $Y \sim N(\mu, \sigma^2)$. We can write

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} y^2 + \frac{\mu}{\sigma^2} y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2} \right\}. \end{aligned}$$

Define

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$

$$T_1(y) = y, \quad T_2(y) = y^2, \quad S(y) = 0, \quad \gamma(\phi_1, \phi_2) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2}\log\left(-\frac{\pi}{\phi_2}\right),$$

and observe that the support of f is always $\mathbb{R}$. Thus $N(\mu, \sigma^2)$ is a two-parameter exponential family. □

Sampling Theory and Stochastic Convergence

Sampling

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation (sample) of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution. Call this a sample.
4. Use sample $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

Anything we do will be a function $T(Y_1, \ldots, Y_n)$ of the sample

Sampling theory aims to understand:

1. What information do different forms of functions $T : \mathcal{Y}^n \to \mathbb{R}^p$ carry on the parameter θ?
2. What is the probability distribution of $T(Y_1, ..., Y_n)$ and how does it relate to $F(y_1, ..., y_n; \theta)$?

These two questions are closely related.

Definition (Statistic)

A statistic is any function T whose domain is the sample space $\mathcal{Y}^n$ but does not depend on unknown parameters.

↪ Intuitively, any function that can be evaluated on the basis of the sample alone is a statistic.
↪ Any statistic is clearly itself a random variable with its own distribution.

Example

$T(\boldsymbol{Y}) = n^{-1} \sum_{i=1}^{n} Y_i$ is a statistic (since n, the sample size, is known).

Example

$T(\boldsymbol{Y}) = (Y_{(1)}, \ldots, Y_{(n)})$ where $Y_{(1)} \leq Y_{(2)} \leq \ldots, Y_{(n)}$ are the order statistics of $\boldsymbol{Y}$. Since T depends only on the values of $\boldsymbol{Y}$, T is a statistic.

Example

Let $T(\boldsymbol{Y}) = c$, where c is a known constant. Then T is a statistic

Definition (Sampling Distribution)

Let $(Y_1, \ldots, Y_n)^\top \sim F(y_1, ..., y_n; \theta)$ and $T : \mathcal{Y}^n \to \mathbb{R}^q$ be a statistic,

$$T(Y_1, ..., Y_n) = (T_1(Y_1, \ldots, Y_n), \ldots, T_q(Y_1, \ldots, Y_n)).$$

The sampling distribution of T under $F(y_1, ..., y_n; \theta)$ is the distribution

$$F_T(t_1, \ldots, t_p) = \mathbb{P}[T_1(Y_1, \ldots, Y_n) \leq t_1, \ldots, T_q(Y_1, \ldots, Y_n) \leq t_q].$$

Comments:

- We will typically simply write T instead of the cumbersome $T(Y_1, ..., Y_n)$.
- Very often $T : \mathcal{Y}^n \to \mathbb{R}$ (i.e. $q = 1$), in which case the notation simplifies considerably:
$$F_T(t) = \mathbb{P}[T(Y_1, ..., Y_n) \leq t], \quad t \in \mathbb{R}.$$

Key observation:

The sampling distribution of T depends on the unknown θ

The extent and form of this dependence is essential for inference.

- Evident from previous examples: some statistics are more informative and others are less informative regarding the true value of θ
- Any $T(Y_1, ..., Y_n)$ that is not "1-1" carries less information about θ than the original sample $(Y_1, ..., Y_n)$ itself.
- Which are "good" and which are "bad" statistics?

Definition (Ancillary Statistic)

A statistic T is an *ancillary statistic* (for θ) if its distribution does not functionally depend θ

$\hookrightarrow$ So an ancillary statistic has the same distribution $\forall\ \theta \in \Theta$.

Example

Suppose that $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where μ unknown but σ^2 known). Let $T(Y_1, ..., Y_n) = Y_1 - Y_2$; then T has a Normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for the unknown parameter μ. If both μ and σ^2 were unknown, T would not be ancillary for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

- If T is ancillary for θ then T carries no information about θ
- In order to carry any useful information about θ, the sampling distribution F_T must depend explicitly on θ.
- Intuitively, the amount of information T carries on θ increases as the dependence of its sampling distribution F_T on θ increases

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$, $S = \min(Y_1, \ldots, Y_n)$ and $T = \max(Y_1, \ldots, Y_n)$.

- $f_S(y; \theta) = \frac{n}{\theta}\left(1 - \frac{y}{\theta}\right)^{n-1}, \quad 0 \leq y \leq \theta$
- $f_T(y; \theta) = \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1}, \quad 0 \leq y \leq \theta$

$\hookrightarrow$ Neither S nor T are ancillary for θ

$\hookrightarrow$ As $n \uparrow \infty$, f_S becomes concentrated around 0

$\hookrightarrow$ As $n \uparrow \infty$, f_T becomes concentrated around θ while

$\hookrightarrow$ Indicates that T provides more information about θ than does S.

To understand the information carried by statistics on θ, we need to understand how they partition the sample space $\mathcal{Y}^n$.

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} F_\theta$ and $T(\boldsymbol{Y})$ a statistic.

- The *level sets* or *contours* of T are the sets

$$A_t = \{\boldsymbol{y} \in \mathcal{Y}^n : T(\boldsymbol{y}) = t\}.$$

(all potential samples that could have given us the value t for T)

↪ Clearly, T is constant when restricted to a level set.

- Any realization of $\boldsymbol{Y}$ that falls in a given level set is equivalent as far as T is concerned, as T reduces all these values to the same output.
- Any inference drawn through T will be the same within a given level set.
- So let's look at the distribution of $\boldsymbol{Y}$ conditional on a given fibre A_t of T, $F_{\boldsymbol{Y}|T=t}(\boldsymbol{y})$...

- Suppose $F_{\boldsymbol{Y}|T=t}$ changes depending on θ: we are losing information.
- Suppose $F_{\boldsymbol{Y}|T=t}$ is functionally independent of θ

  $\Longrightarrow$ Then $\boldsymbol{Y}$ contains no information about θ on the set A_t
  $\Longrightarrow$ In other words, $\boldsymbol{Y}$ is ancillary for θ on A_t

- If this is true for each $t \in \text{Range}(T)$ then $T(\boldsymbol{Y})$ contains the same information about θ as $\boldsymbol{Y}$ itself does.
  ↪ It does not matter whether we observe $\boldsymbol{Y} = (Y_1, ..., Y_n)$ or just $T(\boldsymbol{Y})$.
  ↪ Knowing the exact value $\boldsymbol{Y}$ in addition to knowing $T(\boldsymbol{Y})$ does not give us any additional information - $\boldsymbol{Y}$ is irrelevant if we already know $T(\boldsymbol{Y})$.

Definition (Sufficient Statistic)

A statistic $T = T(\boldsymbol{Y})$ is said to be *sufficient* for the parameter θ if the conditional probability distribution of the sample given the statistic

$$F_{\boldsymbol{Y}|T(\boldsymbol{Y})=t}(y_1, ..., Y_n) = \mathbb{P}[Y_1 \leq y_1, \ldots, Y_n \leq y_n | T(Y_1, ..., Y_n) = t]$$

does *not* depend on θ.

Example (Coin Tossing)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(\theta)$, and $T(\boldsymbol{Y}) = \sum_{i=1}^n Y_i$. For $\boldsymbol{y} \in \{0,1\}^n$,

$$\begin{aligned}
\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y} | T = t] &= \frac{\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y}]}{\mathbb{P}[T = t]} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \frac{\theta^{\Sigma_{i=1}^n y_i}(1-\theta)^{n-\Sigma_{i=1}^n y_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \binom{n}{t}^{-1} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\}.
\end{aligned}$$

- T is sufficient for $\theta \rightarrow$ Given # of tosses that came heads, knowing *which tosses* came heads is irrelevant in deciding the probability of heads:

$$0\ 0\ 1\ 1\ 1\ 0\ 1 \quad \text{VS} \quad 1\ 0\ 0\ 0\ 1\ 1\ 1 \quad \text{VS} \quad 1\ 0\ 1\ 0\ 1\ 0\ 1$$

- Definition hard to verify (especially for continuous variables)
- Definition does not allow easy identification of sufficient statistics

Theorem (Fisher-Neyman Factorization Theorem)

Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ has a joint density or frequency function $f(\boldsymbol{y}; \theta)$, $\theta \in \Theta$. A statistic $T = T(\boldsymbol{Y})$ is sufficient for θ if and only if

$$f(\boldsymbol{y}; \theta) = g(T(\boldsymbol{y}), \theta) h(\boldsymbol{y}).$$

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$ with pdf $f(y; \theta) = \mathbf{1}\{y \in [0, \theta]\}/\theta$. Then,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\theta^n} \mathbf{1}\{\boldsymbol{y} \in [0, \theta]^n\} = \frac{\mathbf{1}\{\max[y_1, ..., y_n] \leq \theta\} \mathbf{1}\{\min[y_1, ..., y_n] \geq 0\}}{\theta^n}$$

Therefore $T(\boldsymbol{Y}) = Y_{(n)} = \max[Y_1, ..., Y_n]$ is sufficient for θ.

Proof of Neyman-Fisher Theorem - Discrete Real Statistic.

Suppose first that T is sufficient. Then

$$\begin{aligned} f(y;\theta) &= \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}] = \sum_t \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}, T=t] \\ &= \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}, T=T(\boldsymbol{y})] = \mathbb{P}[T=T(\boldsymbol{y})]\mathbb{P}[\boldsymbol{X}=\boldsymbol{y}|\, T=T(\boldsymbol{y})] \end{aligned}$$

Since T is sufficient, $\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\, T=T(\boldsymbol{y})]$ is independent of θ and so $f(y;\theta) = g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})$. Now suppose that $f(y;\theta) = g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})$. Then if $T(\boldsymbol{y}) = t$,

$$\begin{aligned} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|\, T=t] &= \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}, T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}[T=t]}\mathbf{1}\{T(\boldsymbol{y})=t\} \\ &= \frac{g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})\mathbf{1}\{T(\boldsymbol{y})=t\}}{\sum_{z:\,T(z)=t} g(T(\boldsymbol{z});\theta)h(\boldsymbol{z})} = \frac{h(\boldsymbol{y})\mathbf{1}\{T(\boldsymbol{y})=t\}}{\sum_{T(z)=t} h(\boldsymbol{z})}. \end{aligned}$$

which does not depend on θ. □

Example (Sufficient statistics for i.i.d. normal samples)

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Recall that we can write

$$f(y;\mu,\sigma^2) = \frac{e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} = \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}$$

and so

$$f_{Y_1,...,Y_n}(y_1,...,y_n) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n y_i - \frac{n}{2}\log(2\pi\sigma^2) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

Consequently, Fisher-Neyman factorisation implies that the statistic

$$S(\boldsymbol{Y}) = \big(S_1(\boldsymbol{Y}), S_2(\boldsymbol{Y})\big)^\top = \big(\textstyle\sum_{i=1}^n Y_i\,,\, \sum_{i=1}^n Y_i^2\big)^\top = \big(\bar{Y}\,,\, \sum_{i=1}^n Y_i^2\big)^\top$$

is sufficient for the parameter (μ, σ^2) and so is the statistic

$$T(\boldsymbol{Y}) = \big(T_1(\boldsymbol{Y}), T_2(\boldsymbol{Y})\big)^\top = \big(n^{-1}\textstyle\sum_{i=1}^n Y_i\,,\, n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2\big)^\top$$

since T and S are 1-1 functions of each other.

Example (Sufficient statistics for k-parameter exponential families)

More generally, consider a k-parameter exponential family, with density

$$f(y) = \exp\left\{\sum_{j=1}^{k} \phi_j \, T_j(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}, \qquad y \in \mathcal{Y}.$$

Then an i.i.d. sample $(Y_1, ..., Y_n)^\top$ has joint distribution

$$f_{Y_1,...,Y_n}(y_1, \ldots, y_n) = \exp\left\{\sum_{j=1}^{k} \phi_j \tau_j(y_1, \ldots, y_n) - n\gamma(\phi_1, ..., \phi_n) + \sum_{i=1}^{n} S(y_i)\right\}$$

where

$$\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i).$$

So the statistic

$$\tau(Y_1, ..., Y_n) = \big(\tau_1(Y_1, ..., Y_n), ..., \tau_k(Y_1, ..., Y_n)\big)^\top$$

is sufficient for $(\phi_1, ..., \phi_k)$ by Fisher-Neyman factorisation.

Minimally Sufficient Statistics

- Saw that sufficient statistic compresses data without information loss on parameter of interest.
- How much info can we throw away? Is there a "necessary" statistic?

Definition (Minimally Sufficient Statistic)

A statistic $T = T(\boldsymbol{Y})$ is said to be *minimally sufficient* for the parameter θ if it is sufficient for θ and for any other sufficient statistic $S = S(\boldsymbol{Y})$ there exists a function $g(\cdot)$ with

$$T(\boldsymbol{Y}) = g(S(\boldsymbol{Y})).$$

Lemma

If T and S are minimally sufficient statistics for a parameter θ, then there exists injective functions g and h such that $S = g(T)$ and $T = h(S)$.

Theorem

Let $\boldsymbol{Y} = (X_1, ..., X_n)$ have joint density or frequency function $f(\boldsymbol{y}; \theta)$ and $T = T(\boldsymbol{Y})$ be a statistic. Suppose that $f(\boldsymbol{y}; \theta)/f(\boldsymbol{z}; \theta)$ is independent of θ if and only if $T(\boldsymbol{y}) = T(\boldsymbol{z})$. Then T is minimally sufficient for θ.

Proof.(*)

Assume for simplicity that $f(\boldsymbol{y}; \theta) > 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and $\theta \in \Theta$. Let $\mathcal{T} = \{T(\boldsymbol{u}) : u \in \mathbb{R}^n\}$ be the image of $\mathbb{R}^n$ under T and let A_t be the level sets of T. For each t, choose a representative element $\boldsymbol{w}_t \in A_t$. Notice that for any $\boldsymbol{y}$, $\boldsymbol{w}_{T(\boldsymbol{y})}$ is in the same level set as $\boldsymbol{y}$, so that

$$f(\boldsymbol{y}; \theta)/f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta)$$

does not depend on θ by assumption. Let $g(t, \theta) := f(\boldsymbol{w}_t; \theta)$ and notice

$$f(\boldsymbol{y}; \theta) = \frac{f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta) f(\boldsymbol{y}; \theta)}{f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta)} = g(T(\boldsymbol{y}), \theta) h(\boldsymbol{y})$$

and sufficiency follows from th Fisher-Neyman factorization theorem.

[minimality part] Suppose that T' is another sufficient statistic. By the factorization thm: $\exists g', h' : \quad f(\boldsymbol{y}; \theta) = g'(T'(\boldsymbol{y}); \theta) h'(\boldsymbol{y})$. Let $\boldsymbol{y}, \boldsymbol{z}$ be such that $T'(\boldsymbol{y}) = T'(\boldsymbol{z})$. Then

$$\frac{f(\boldsymbol{y}; \theta)}{f(\boldsymbol{z}; \theta)} = \frac{g'(T'(\boldsymbol{y}); \theta) h'(\boldsymbol{y})}{g'(T'(\boldsymbol{z}); \theta) h'(\boldsymbol{z})} = \frac{h'(\boldsymbol{y})}{h'(\boldsymbol{z})}.$$

Since ratio does not depend on θ, we have by assumption $T(\boldsymbol{y}) = T(\boldsymbol{z})$. Hence T is a function of T'; so is minimal by arbitrary choice of T'. □

Example (Bernoulli Trials)

Let $Y_1, ..., Y_n \overset{iid}{\sim}$ Bernoulli(θ). Let $\boldsymbol{z}, \boldsymbol{y} \in \{0, 1\}^n$ be two possible outcomes. Then

$$\frac{f(\boldsymbol{z}; \theta)}{f(\boldsymbol{y}; \theta)} = \frac{\theta^{\Sigma z_i}(1-\theta)^{n - \Sigma z_i}}{\theta^{\Sigma y_i}(1-\theta)^{n - \Sigma y_i}}$$

which is constant if and only if $T(\boldsymbol{z}) = \sum z_i = \sum y_i = T(\boldsymbol{y})$, so that T is minimally sufficient.

Example (Minimal sufficiency for k-parameter exponential families)

An i.i.d. sample $(Y_1, ..., Y_n)^\top$ from an exponential family has joint distribution

$$f_{Y_1,...,Y_n}(y_1, \ldots, y_n) = \exp\left\{\sum_{j=1}^{k} \phi_j \tau_j(y_1, \ldots, y_n) - n\gamma(\phi_1, ..., \phi_n) + \sum_{i=1}^{n} S(y_i)\right\}$$

where $\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i)$, as before. If the $\{T_j\}_{j=1}^{k}$ are non-trivial, the ratio $f(\boldsymbol{y})/f(\boldsymbol{z})$ will be constant with respect to $(\phi_1, \ldots, \phi_k)$ if and only if as $(\phi_1, \ldots, \phi_k)$ varies, the quantity below remains constant.

$$\sum_{j=1}^{k} \phi_j \left(\tau_j(y_1, \ldots, y_n) - \tau_j(z_1, \ldots, z_n)\right)$$

So if $(\phi_1, ..., \phi_k)$ range over an open parameter space of dimension k, this must imply that

$$\tau_j(y_1, \ldots, y_n) = \tau_j(z_1, \ldots, z_n).$$

Conversely, when the latter is true, the density ratio is clearly independent of the parameters, and so the statistic $\tau(\boldsymbol{y}) = (\tau_1(\boldsymbol{y}), ..., \tau_k(\boldsymbol{y}))$ is minimally sufficient for $(\phi_1, ..., \phi_k)$.

Back to Sampling Distributions

Anything we do will be a function $T(Y_1, \ldots, Y_n)$ of the sample

Sampling theory aims to understand:

1. ✓ What information do different forms of functions $T : \mathcal{Y}^n \to \mathbb{R}^p$ carry on the parameter θ? ✓
2. ? What is the probability distribution of $T(Y_1, ..., Y_n)$ and how does it relate to $F(y_1, ..., y_n; \theta)$?

We will now:

(2a) Review important special cases where sampling distributions can be exactly determined, focussing on the iid sampling case.
↪ focussing on sufficient statistics of Gaussian and exponential families.

(2b) Study ways of getting approximations to the sampling behaviour when the precise form is not explicitly available or is tedious (stochastic convergence).

Theorem (Sampling Distribution of Gaussian Sufficient Statistics)

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, and define*

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \& \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

The pair $(\bar{Y}, S^2)$ *is minimally sufficient for* (μ, σ^2) *and:*

1. *The sample mean is distributed as* $\bar{Y} \sim N(\mu, \sigma^2/n)$.
2. *The random variables* $\bar{Y}$ *and* S^2 *are independent.*
3. *The random variable* S^2 *satisfies* $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Corollary (Moments of Sufficient Statistics)

If $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$*, then*

$$\mathbb{E}[\bar{Y}] = \mu, \quad \text{var}(\bar{Y}) = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[S^2\right] = \sigma^2, \quad \text{var}(S^2) = \frac{2\sigma^4}{n-1}.$$

(which is why we typically prefer factor of $(n-1)^{-1}$ instead of n^{-1} in S^2)

Theorem (Sums of Gaussian Squares)

Let $\{Z_1, ..., Z_k\}$ *be i.i.d.* $N(0,1)$ *random variables. Then,*

$$Z_1^2 + \ldots + Z_k^2 \sim \chi^2_k.$$

Recall that a random variable X is said to follow the chi-square distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \chi^2_k$, if it holds that $X \sim Gamma(k/2, 1/2)$. In other words,

$$f_X(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \chi^2_k$ are given by

$$\mathbb{E}[X] = k, \qquad \text{var}[X] = 2k, \qquad M(t) = (1-2t)^{-k/2}, \quad t < \frac{1}{2}.$$

Theorem (Student's Statistic and its Sampling Distribution)

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Then, the empirically standardised mean satisfies*

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim \mathsf{t}_{n-1}.$$

Recall that a random variable X is said to follow Student's t distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \mathsf{t}_k$, if,

$$f_X(x; k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

Assuming $k > 2$, the mean and variance of $X \sim \mathsf{t}_k$ are given by

$$\mathbb{E}[X] = 0, \qquad \text{var}[X] = \frac{k}{k-2}.$$

The mean is undefined for $k = 1$ and the variance is undefined for $k \leq 2$. The moment generating function is undefined for any $k \in \mathbb{N}$.

Theorem (Ratios of Gaussian Sums of Squares and F-Statistic)

Let $Y_1 \sim \chi^2_{d_1}$ *and* $Y_2 \sim \chi^2_{d_2}$ *be independent random variables. Then,*

$$\frac{Y_1/d_1}{Y_2/d_2} \sim \mathsf{F}_{d_1, d_2}.$$

A random variable X is said to follow the Snedecor-Fisher F distribution with parameters $d_1 \in \mathbb{N}$ and $d_2 \in \mathbb{N}$, denoted $X \sim \mathsf{F}_{d_1, d_2}$, if

$$f_X(x; d_1, d_2) = \begin{cases} \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance of $X \sim \mathsf{F}_{d_1, d_2}$ are given by

$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}, \text{ provided } d_2 > 2, \ \text{var}[X] = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 4)(d_2 - 2)^2} \text{ provided } d_2 > 4.$$

The moment generating function does not exist.

Some remarks:

- Notice that in all cases we considered sampling distributions when the sample is iid Normal.
- It's possible to consider samples that are jointly Normally distributed but not iid. This we postpone till later, when we bring in covariates and study regression.
- It's not particularly productive to insist on the formulae for the densities of χ^2, t and F distributions.
  - ↪ Much better to think of them as being implicitly defined via their relation to iid Gaussian sampling (sums of squares and their normalised ratios, etc).
  - ↪ This is what is crucial to remember, along with the relation of their parameters (degrees of freedom) to the setting at hand.

Theorem (Sampling from an Exponential Family)

Let $Y_1, ..., Y_n \overset{iid}{\sim} f$, *where*

$$f(y) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}, \quad \boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top \in \Phi \subseteq \mathbb{R}^k$$

be a density of a k*-parameter exponential family form.*

If Φ *is open, then:*

1. *The minimally sufficient statistic for* $\boldsymbol{\phi}$ *is* $\boldsymbol{\tau} = (\tau_1, ..., \tau_k)$ *where*
$$\tau_j(y_1, ..., y_n) = \textstyle\sum_{i=1}^{n} T_j(y_i).$$
2. *The function* γ *is infinitely differentiable in all* k *of its variables, and*
$$\mathbb{E}[\boldsymbol{\tau}] = n\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \quad \textit{and} \quad \operatorname{cov}[\tau] = n\nabla^2_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}),$$
that is,
$$\mathbb{E}[\tau_j] = n\tfrac{\partial}{\partial\phi_j}\gamma(\phi_1, ..., \phi_k) \quad \textit{and} \quad \operatorname{cov}\{\tau_m, \tau_j\} = n\tfrac{\partial^2}{\partial\phi_m\partial\phi_j}\gamma(\phi_1, ..., \phi_k).$$

(recall that minimal sufficiency was already shown in an example)

Proof. (*)

Focus on case $k = 1$, so that $\tau_j = \tau = \sum_{i=1}^n T(Y_i)$. Let $\phi_0 \in \Phi$. Since Φ is open, there exists s sufficiently small so that $\phi_0 + s \in \Phi$. Now note that the MGF $M_{T(Y_1)}(u) = \mathbb{E}\exp[uT(Y_1)]$ of $T(Y_1)$ evaluated at s is

$$\int_{\mathbb{R}} e^{sT(y)} e^{\phi_0 T(y) - \gamma(\phi_0) + S(y)}\, dy = e^{\gamma(\phi_0+s)-\gamma(\phi_0)} \underbrace{\int_{\mathbb{R}} e^{(\phi_0+s)T(y)-\gamma(\phi_0+s)+S(y)}\, dy}_{=1}$$

1. Therefore $M_{T(Y_1)}(s) < \infty$ for s sufficiently small, and thus:
   - all moments of $T(Y_1)$ exist,
   - and $M_{T(Y_1)}(s)$ is infinitely differentiable on an open neighbourhood of 0.
2. Furthermore, $\gamma(s + \phi_0)$ is infinitely differentiable for s small enough, i.e. γ is infinitely differentiable in an open neighbourhood of ϕ_0. But ϕ_0 is arbitrary so γ is infinitely differentiable everywhere on Φ.

Now we we may differentiate w.r.t. s,and, setting $s = 0$, we get

$$\mathbb{E}[T(Y_1)] = \gamma'(\phi) \text{ and } \text{var}[T(Y_1)] = \gamma''(\phi).$$

The conclusion follows by the fact that $\tau = \sum_{i=1}^n T(Y_i)$. □

Approximate Sampling Behaviour and Stochastic Convergence

Unfortunately, the sampling distribution of a statistic $T(Y_1, \ldots, Y_n)$ isn't always obtainable in a closed/convenient form

↪ Even when T is the sufficient statistic in an exponential family, we may not have a nice workable form for the sampling distribution.

↪ In this case we know that the sampling distribution is again a k-parameter exponential family, but its form may be tedious to work with.

General strategy:

Approximate the sampling distribution $F_{T(Y_1,\ldots,Y_n)}$ by a simpler distribution G

Of course we must make sense of what it means that "the distribution $F_{T(Y_1,\ldots,Y_n)}$ is approximated by the distribution G".

1. We will view $F_{\tau(Y_1,\ldots,Y_n)}$ as a sequence of functions indexed by sample size n.
2. Thus, "approximation by G" will be understood as a form of convergence of F_n to G as $n \to \infty$.
3. En route, we will also discover a stronger form of convergence.

Definition (Convergence in Distribution (Weak Convergence))

Let $\{F_n\}_{n\geq 1}$ be a sequence of distribution functions and G be a distribution function on $\mathbb{R}$. We say that F_n converges in distribution (or weakly) to G, and write $F_n \xrightarrow{d} G$, whenever

$$F_n(y) \overset{n\to\infty}{\longrightarrow} G(y),$$

for all y constituting continuity points of G (i.e. all y such that $\lim_{\varepsilon\to 0} G(y+\varepsilon) = G(y)$).

Example

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}[0,1]$, $M_n = \max\{Y_1, ..., Y_n\}$, and $Q_n = n(1-M_n)$.

$$\mathbb{P}[Q_n \leq y] = \mathbb{P}[M_n \geq 1 - y/n] = 1 - \left(1 - \frac{y}{n}\right)^n \overset{n\to\infty}{\longrightarrow} 1 - e^{-y}$$

for all $y \geq 0$. Hence $Q_n \xrightarrow{d} Q$, with $Q \sim exp(1)$.

Comments:

1. Convergence in distribution $\equiv$ pointwise convergence of distribution function, with the exception that it is not necessary to have convergence at discontinuity points of the limit.

2. When $F_n(y) = \mathbb{P}[Y_n \leq y]$ for a sequence of random variables $\{Y_n\}_{n\geq 1}$ et $G(y) = \mathbb{P}[Z \leq y]$ for another random variable Z, we will abuse notation and write

$$Y_n \xrightarrow{d} Z.$$

3. Our aim of approximating the sampling distribution now translates into finding a random variable Z whose distribution is explicitly known and such that

$$T(Y_1, ..., Y_n) \xrightarrow{d} Z$$

A stronger notion of convergence is convergence in probability:

Definition

When a sequence of random variables $\{Y_n\}$ satisfies $\mathbb{P}[|Y_n - Y| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0$ for all $\epsilon > 0$ and a giveFn random variable Y, we say that Y_n converges in probability to Y, and write $Y_n \xrightarrow{p} Y$.

Example

Let $U_1, \ldots, U_n \overset{iid}{\sim} \mathcal{U}[0,1]$ and $M_n = \max\{U_1, ..., U_n\}$. Fix $\epsilon \in (0,1)$. Then

$$\mathbb{P}[|M_n - 1| > \epsilon] = \mathbb{P}[M_n > 1+\epsilon] + \mathbb{P}[M_n < 1-\epsilon] = 0 + (1-\epsilon)^n \overset{n\to\infty}{\to} 0\,.$$

Also $\mathbb{P}[|M_n - 1| > \epsilon] = 0$ if $\epsilon \geq$. Hence $M_n \xrightarrow{p} 1$ as $n \to \infty$.

Comments:

- Convergence in probability implies convergence in distribution.
- Convergence in distribution does not imply convergence in probability
  - ↪ Consider $Z \sim \mathcal{N}(0,1)$, $-Z + \frac{1}{n} \xrightarrow{d} Z$ but $-Z + \frac{1}{n} \xrightarrow{p} -Z$.
- "$\xrightarrow{d}$" relates *distribution functions*. It says the probabilistic behaviour of a sequence Y_n becomes more and more alike to that of the limit Y.
- "$\xrightarrow{p}$" relates *random variables*. It says that the actual realisations of Y_n can be progressively approximated with high probability by those of Y.
- Both notions of convergence are *metrizable*
  - ↪ i.e. there exist metrics on the space of all random variables that are compatible with the notion of convergence.
  - ↪ Hence can use things such as the triangle inequality etc.

Theorem

(a) $Y_n \xrightarrow{p} Y \implies Y_n \xrightarrow{d} Y$

(b) $Y_n \xrightarrow{d} c \implies Y_n \xrightarrow{p} c,\ c \in \mathbb{R}.$

Proof

(a)Let y be a continuity point of F_Y and $\epsilon > 0$. Then,

$$\begin{aligned} \mathbb{P}[Y_n \leq y] &= \mathbb{P}[Y_n \leq y, |Y_n - Y| \leq \epsilon] + \mathbb{P}[Y_n \leq y, |Y_n - Y| > \epsilon] \\ &\leq \mathbb{P}[Y \leq y + \epsilon] + \mathbb{P}[|Y_n - Y| > \epsilon] \end{aligned}$$

since $\{Y \leq y + \epsilon\}$ contains $\{Y_n \leq y, |Y_n - Y| \leq \epsilon\}$. Similarly,

$$\begin{aligned} \mathbb{P}[Y \leq y - \epsilon] &= \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| \leq \epsilon] + \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| > \epsilon] \\ &\leq \mathbb{P}[Y_n \leq y] + \mathbb{P}[|Y_n - Y| > \epsilon] \end{aligned}$$

which yields

$$\mathbb{P}[Y \leq y - \epsilon] - \mathbb{P}[|Y_n - Y| > \epsilon] \leq \mathbb{P}[Y_n \leq y].$$

Combining the two inequalities and "sandwitching" yields (a).

(b) Let F be the distribution function of a constant r.v. c,

$$F(y) = \mathbb{P}[c \leq y] = \begin{cases} 1 & \text{if } y \geq c, \\ 0 & \text{if } y < c. \end{cases}$$

$$\begin{aligned} \mathbb{P}[|Y_n - c| > \epsilon] &= \mathbb{P}[\{Y_n - c > \epsilon\} \cup \{c - Y_n > \epsilon\}] \\ &= \mathbb{P}[Y_n > c + \epsilon] + \mathbb{P}[Y_n < c - \epsilon] \\ &\leq 1 - \mathbb{P}[Y_n \leq c + \epsilon] + \mathbb{P}[Y_n \leq c - \epsilon] \\ &\xrightarrow{n \to \infty} 1 - F(\underbrace{c + \epsilon}_{\geq c}) + F(\underbrace{c - \epsilon}_{< c}) = 0 \end{aligned}$$

Since $Y_n \xrightarrow{d} c$. □

Now we explore the stability of stochastic convergence notions under transformation.

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous on the range of Y. Then,

(a) $Y_n \xrightarrow{p} Y \implies g(Y_n) \xrightarrow{p} g(Y)$

(b) $Y_n \xrightarrow{d} Y \implies g(Y_n) \xrightarrow{d} g(Y)$

Theorem (Slutsky's Theorem)

Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$. Then

(a) $X_n + Y_n \xrightarrow{d} X + c$

(b) $X_n Y_n \xrightarrow{d} cX$

Proof of Slutsky's Theorem.

(a) We may assume $c = 0$. Let x be a continuity point of F_X. We have

$$\begin{aligned}\mathbb{P}[X_n + Y_n \le x] &= \mathbb{P}[X_n + Y_n \le x, |Y_n| \le \epsilon] + \mathbb{P}[X_n + Y_n \le x, |Y_n| > \epsilon] \\ &\le \mathbb{P}[X_n \le x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]\end{aligned}$$

Similarly, $\mathbb{P}[X_n \le x - \epsilon] \le \mathbb{P}[X_n + Y_n \le x] + \mathbb{P}[|Y_n| > \epsilon]$, therefore,

$$\mathbb{P}[X_n \le x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \le \mathbb{P}[X_n + Y_n \le x] \le \mathbb{P}[X_n \le x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]$$

Since ϵ is arbitrary, this proves (a) by taking $n \to \infty$.

(b) It suffices to assume $c = 0$ (since $(Y_n + c)X_n = X_n Y_n + X_n c$, so if we can show $X_n Y_n \xrightarrow{d} 0$, then (a) gives conclusion). Let $\epsilon, M > 0$:

$$\begin{aligned}\mathbb{P}[|X_n Y_n| > \epsilon] &\le \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| \le 1/M] + \mathbb{P}[|Y_n| \ge 1/M] \\ &\le \mathbb{P}[|X_n| > \epsilon M] + \mathbb{P}[|Y_n| \ge 1/M] \\ &\xrightarrow{n\to\infty} \mathbb{P}[|X| > \epsilon M] + 0\end{aligned}$$

The first term can be made arbitrarily small by letting $M \to \infty$.

□

Theorem (General Version of Slutsky's Theorem)

Let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *be continuous and suppose that* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{d} c \in \mathbb{R}$. *Then,* $g(X_n, Y_n) \to g(X, c)$ *as* $n \to \infty$.

↪Notice that the general version of Slutsky's theorem does not follow immediately from the continuous mapping theorem.

- The continuous mapping theorem would be applicable if (X_n, Y_n) weakly converged jointly (i.e. their joint distribution) to (X, c).
- But here we assume only marginal convergence (i.e. $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ separately, but their joint behaviour is unspecified).
- The key of the proof is that in the special case where $Y_n \xrightarrow{d} c$ where c is a constant, then marginal convergence $\iff$ joint convergence.
- However if $X_n \xrightarrow{d} X$ where X is non-degenerate, and $Y_n \xrightarrow{d} Y$ where Y is non-degenerate, then the theorem fails.
- Notice that even the special cases (addition and multiplication) of Slutsky's theorem fail of both X and Y are non-degenerate.

We will later consider joint stochastic convergence.

Continuous mappings and Slutsky's lemma allow us to get new approximations from old ones.

↪ But how do we get limit theorems in the first place?
Typically these stem from a clever use of the following two fundamental theorems:

Theorem (Law of Large Numbers)

Let $\{Y_n\}$ *be independent random variables with* $\mathbb{E}[Y_k] = \mu$ *and* $\mathbb{E}|Y_k| < \infty$, *for all* $k \geq 1$. *Then,* $n^{-1}(Y_1 + \ldots + Y_n) \xrightarrow{p} \mu$.

Theorem (Central Limit Theorem)

Let $\{Y_n\}$ *be an i.i.d sequence with mean* μ *and variance* $\sigma^2 < \infty$. *Then,*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

Said differently, for large n, $\bar{Y} \approx N(\mu, \sigma^2/n)$ or $Y_1 + \ldots + Y_n \approx N(n\mu, n\sigma^2)$.

The following theorem combines Slutksy's lemma and the continuous mapping theorem in order to allow us to transform central limit theorems:

Theorem (The Delta Method)

Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ *where* $a_n, \theta \in \mathbb{R}$ *for all* n *and* $a_n \uparrow \infty$. *Let* $g(\cdot)$ *be continuously differentiable at* θ. *Then,* $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

Proof

Taylor expanding around θ gives:

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta), \quad \theta_n^* \text{ between } X_n, \theta.$$

Thus $|\theta_n^* - \theta| < |X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ [by Slutsky]
Therefore, $\theta_n^* \xrightarrow{p} \theta$. By the continuous mapping theorem $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$.

$$\begin{aligned} \text{Thus} \qquad a_n(g(X_n) - g(\theta)) &= a_n(g(\theta) + g'(\theta_n^*)(X_n - \theta) - g(\theta)) \\ &= g'(\theta_n^*) a_n(X - \theta) \xrightarrow{d} g'(\theta)Z. \end{aligned}$$

The delta method also applies even when $g'(\theta)$ is not continuous (proof harder).

We can apply this machinery to get the following result for the sampling distribution of a sufficient statistic in a 1-parameter exponential family:

Corollary

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, *where*

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X}$$

with $\phi \in \Phi \subseteq \mathbb{R}$ *and*

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T(X_i) = n^{-1} \tau(X_1, \ldots, X_n).$$

If Φ *is open, then* γ *is infinitely differentiable, and so*

$$\sqrt{n}(\overline{T}_n - \gamma'(\phi)) \xrightarrow{d} N(0, \gamma''(\phi)).$$

The following more general CLT is often useful:

Theorem (Weighted Sum CLT)

Let $\{W_n\}$ be an i.i.d sequence of real random variables, with common mean 0 and variance 1. Let $\{\gamma_n\}$ be a sequence of real constants. Then,

$$\sup_{1\leq j\leq n} \frac{\gamma_j^2}{\sum_{i=1}^n \gamma_i^2} \overset{n\to\infty}{\longrightarrow} 0 \implies \frac{1}{\sqrt{\sum_{i=1}^n \gamma_i^2}} \sum_{i=1}^n \gamma_i W_i \overset{d}{\to} N(0,1).$$

- Supremum condition amounts to saying that, in the limit, any single component contributes a negligible proportion of the total variance.
- Coefficient sequence $\{\gamma_n\}$ might very well diverge, without contradicting the negligibility condition.

To have joint convergence, we need to consider random vectors:

Definition

Let $\{\boldsymbol{Y}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and $\boldsymbol{Y}$ a random vector of $\mathbb{R}^d$ with $\boldsymbol{Y}_n = (Y_n^{(1)}, ..., Y_n^{(d)})^\mathsf{T}$ and $\boldsymbol{Y} = (Y^{(1)}, ..., Y^{(d)})^\mathsf{T}$. Define the distribution functions $F_{\boldsymbol{Y}_n}(\boldsymbol{y}) = \mathbb{P}[Y_n^{(1)} \leq y^{(1)}, ..., Y_n^{(d)} \leq y^{(d)}]$ and $F_{\boldsymbol{Y}}(\boldsymbol{y}) = \mathbb{P}[Y^{(1)} \leq y^{(1)}, ..., Y^{(d)} \leq y^{(d)}]$, for $\boldsymbol{y} = (y^{(1)}, ..., y^{(d)})^\mathsf{T} \in \mathbb{R}^d$. We say that $\boldsymbol{Y}_n$ converges in distribution to $\boldsymbol{Y}$ as $n \to \infty$ (and write $\boldsymbol{Y}_n \overset{d}{\to} \boldsymbol{Y}$) if for every continuity point of $F_{\boldsymbol{Y}}$ we have

$$F_{\boldsymbol{Y}_n}(\boldsymbol{y}) \overset{n\to\infty}{\longrightarrow} F_{\boldsymbol{Y}}(\boldsymbol{y}).$$

There is a link between univariate and multivariate weak convergence:

Theorem (Cramér-Wold Device)

Let $\{\boldsymbol{Y}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and $\boldsymbol{Y}$ a random vector of $\mathbb{R}^d$. Then,

$$\boldsymbol{Y}_n \overset{d}{\to} \boldsymbol{Y} \iff \boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_n \overset{d}{\to} \boldsymbol{u}^\mathsf{T}\boldsymbol{Y},\ \forall \boldsymbol{u} \in \mathbb{R}^d.$$

- Continuous mapping theorem and Slustky's lemma generalise to vector case.
- In either case, continuity is understood in the multidimensional sense:
  1. Continuous mapping: If $g : \mathbb{R}^p \to \mathbb{R}^d$ is continuous on the range of $\boldsymbol{U}$, and if $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$ in $\mathbb{R}^p$, then $g(\boldsymbol{U}_n) \xrightarrow{d} g(\boldsymbol{U})$ in $\mathbb{R}^d$.
  2. Slutsky: If $g : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ is continuous, and if $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$ in $\mathbb{R}^p$ and $\boldsymbol{W}_n \xrightarrow{d} \boldsymbol{u}$ in $\mathbb{R}^q$, for some deterministic $\boldsymbol{u}$, then $g(\boldsymbol{U}_n, \boldsymbol{W}_n) \xrightarrow{d} g(\boldsymbol{U}, \boldsymbol{u})$.

Convergence in probability easily generalises to the vector case:

Definition

When a sequence of random vectors $\{\boldsymbol{Y}_n\}$ in $\mathbb{R}^d$ satisfies $\mathbb{P}[\|\boldsymbol{Y}_n - \boldsymbol{Y}\| > \epsilon] \xrightarrow{n\to\infty} 0$ for all $\epsilon > 0$ and a given random d-vector Y, we say that $\boldsymbol{Y}_n$ converges in probability to $\boldsymbol{Y}$, and write $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{Y}$.

Theorem (Multivariate Law of Large Numbers)

Let $\{\boldsymbol{Y}_n\}$ be iid random vectors with $\mathbb{E}[\boldsymbol{Y}_k] = \boldsymbol{\mu}$ and $\mathbb{E}\|Y_k\| < \infty$, for all $k \geq 1$. Then,

$$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{Y}_k \xrightarrow{p} \boldsymbol{\mu}$$

Theorem (Multivariate CLT)

Let $\{\boldsymbol{X}_n\}$ be an iid sequence of random vectors in $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Omega}$ and define $\bar{\boldsymbol{X}}_n := \sum_{m=1}^{n} \boldsymbol{X}_m/n$. Then,
$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}_d(0, \boldsymbol{\Omega})$ where $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{Y}$ means $F_{\boldsymbol{Y}_n}(u) \to F_{\boldsymbol{Y}}(u)$ for any continuity point $u \in \mathbb{R}^d$ of $F_{\boldsymbol{Y}}$.

OK, but how fast?

Theorem (Berry-Essen)

In the same setting as the previous theorem, take $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Omega} = I$, then

$$\sup_{\boldsymbol{u}\in\mathbb{R}^d} \left| F_{\sqrt{n}\bar{Y}}(\boldsymbol{u}) - F_{\boldsymbol{Z}}(\boldsymbol{u}) \right| \leq Cn^{-1/2} d^{1/4} \mathbb{E}\|\boldsymbol{Y}_i\|^3.$$

We also have a vector version of the Delta Method:

Theorem (Delta Method – vector case)

Let $\boldsymbol{Z}_n := a_n(\boldsymbol{X}_n - \boldsymbol{u}) \xrightarrow{d} \boldsymbol{Z}$ *in* $\mathbb{R}^d$ *where* $a_n \in \mathbb{R}$, $\boldsymbol{u} \in \mathbb{R}^d$ *and* $a_n \uparrow \infty$. *Let* $g : \mathbb{R}^d \to \mathbb{R}^p$ *be continuously differentiable at* $\boldsymbol{u}$. *Then,*

$$a_n(g(X_n) - g(\boldsymbol{u})) \xrightarrow{d} J_g(\boldsymbol{u})\boldsymbol{Z},$$

where $J_g(\boldsymbol{y})$ *is the* $p \times d$ *Jacobian matrix of* g,

$$J_g(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial x_d} g_1(\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_p(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial x_d} g_p(\boldsymbol{y}) \end{bmatrix}.$$

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The first sort of assertion we wish to make is:

1. **Point Estimation**. Given realisation $(Y_1, \ldots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

How? With a point estimator!

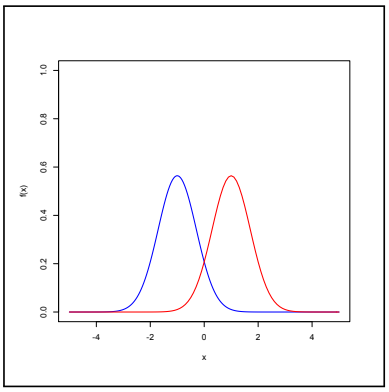
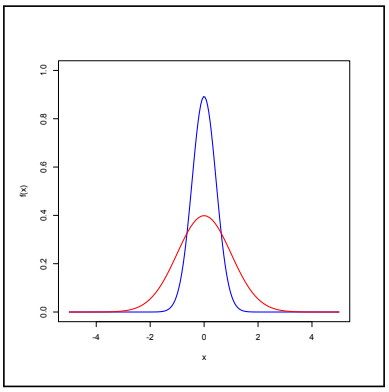
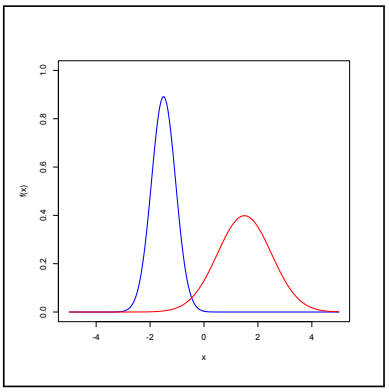
Definition (Point Estimator)

A statistic with codomain Θ is called a *point estimator*, i.e. a point estimator is a statistic $T : \mathcal{Y}^n \to \Theta$.

Since the objective of an estimator is to estimate the θ that generated the data, we typically denote it by $\hat{\theta}(Y_1, ..., Y_n)$, or just $\hat{\theta}$. Note that θ is a deterministic parameter, whereas $\hat{\theta}$ is a random variable.

But which estimator?

- Any statistic taking values in Θ could be used!
- Simpler yet, if we are given some $\hat{\theta}$, how do we judge its quality?
- Since estimators are *random variables*, every different realisation of the sample $(Y_1, \ldots, Y_n)$ will produce a different realised value for $\hat{\theta}$.
- A good estimator should be such that it typically manifests realisations that fall near the true θ.
- More precisely, the sampling distribution of an estimator should be concentrated around the true parameter value θ.

There is a multitude of criteria one can use, but a typical choice is to focus on two basic notions of location and spread for $\hat{\theta}$: its mean and variance

Why?

1. Ease of interpretation. The expectation $\mathbb{E}[\hat{\theta}]$ informs us whether the sampling distribution is located near the truth, whereas the variance $\mathrm{var}[\hat{\theta}]$ quantifies the degree of concentration around the expectation.
2. Central limit theory. Using our theory of stochastic convergence, we can often approximate the sampling distribution of $\hat{\theta}$ by a normal distribution. The latter is fully described by its mean and variance.
3. Concentration inequalities. We can often bound quantities such as $\mathbb{P}\{||\hat{\theta} - \theta|| > \epsilon\}$ by means of moments.

A measure of precision that captures both mean and variance simultaneously is the mean squared error.

Definition (Mean Squared Error)

Let $\hat{\theta}$ be an estimator of a parameter θ corresponding to a model $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. The mean squared error of $\hat{\theta}$ is defined as

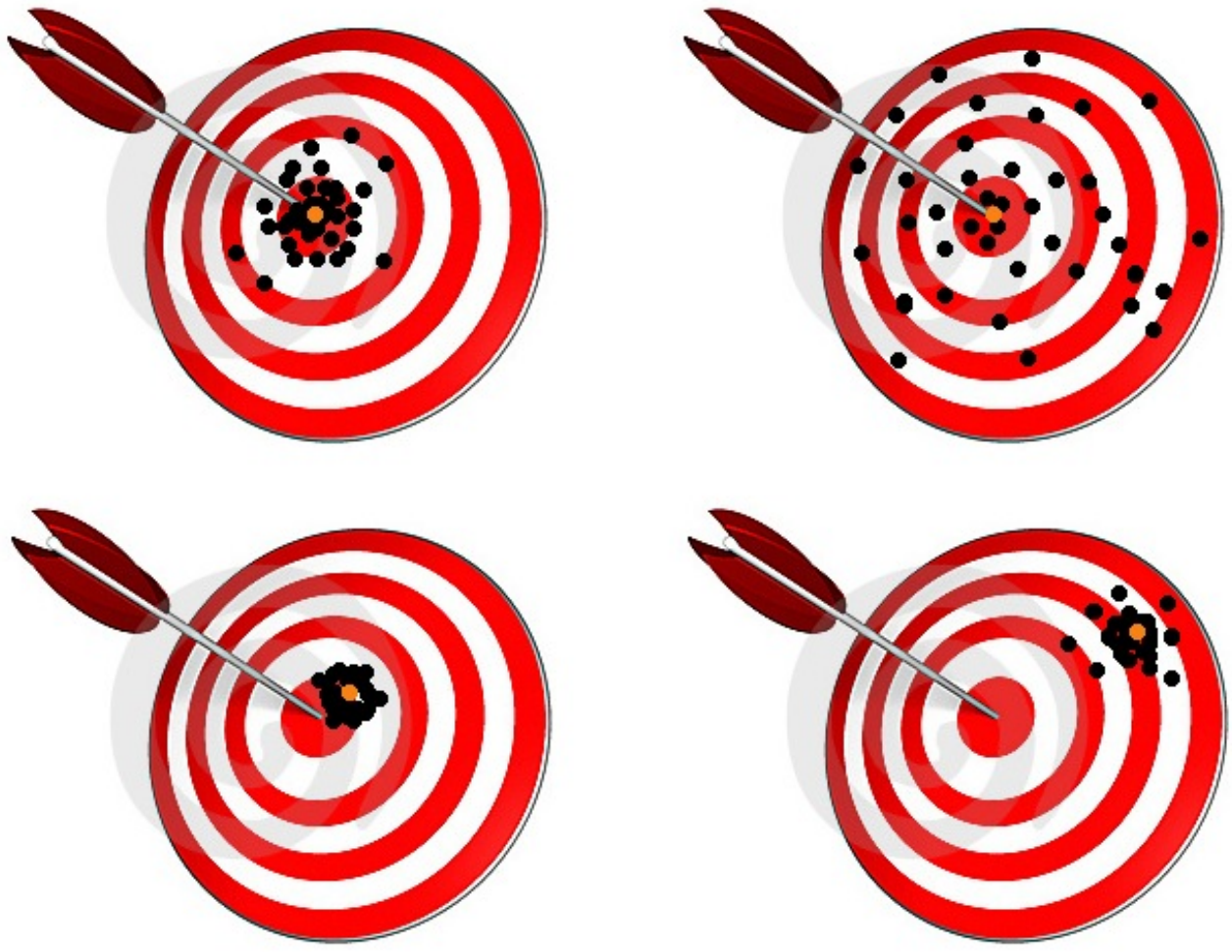
$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right].$$

And here's the relation to means and variances:

Lemma (Bias-Variance Decomposition)

The MSE admits the decomposition

$$\mathrm{MSE}(\hat{\theta}, \theta) = \underbrace{\left\|\mathbb{E}[\hat{\theta}] - \theta\right\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left\|\hat{\theta} - \mathbb{E}(\hat{\theta})\right\|^2\right]}_{\text{variance}}.$$



Proof.

We expand the MSE after adding and subtracting $\mathbb{E}[\hat{\theta}]$:

$$\begin{aligned}
\mathbb{E}[||\hat{\theta} - \theta||^2] &= \mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta||^2] \\
&= \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^\top (\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)\Big] \\
&= ||\mathbb{E}[\hat{\theta}] - \theta||^2 + \mathbb{E}\big[||\hat{\theta} - \mathbb{E}(\hat{\theta})||^2\big] + 2\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^\top (\mathbb{E}[\hat{\theta}] - \theta)\Big] \\
&= ||\mathbb{E}[\hat{\theta}] - \theta||^2 + \mathbb{E}[||\hat{\theta} - \mathbb{E}(\hat{\theta})||^2] + 2\underbrace{(\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}])^\top}_{=0} (\mathbb{E}[\hat{\theta}] - \theta)
\end{aligned}$$

by linearity of the expectation and since $(\mathbb{E}[\hat{\theta}] - \theta)$ is deterministic.

□

As foretold, the concentration of an estimator $\hat{\theta}$ around the true parameter θ can always be bounded by the MSE:

Lemma

Let $\hat{\theta}$ be an estimator of $\theta \in \mathbb{R}^p$. For any $\epsilon > 0$,

$$\mathbb{P}[||\hat{\theta} - \theta|| > \epsilon] \leq \frac{\mathrm{MSE}(\hat{\theta}, \theta)}{\epsilon^2}$$

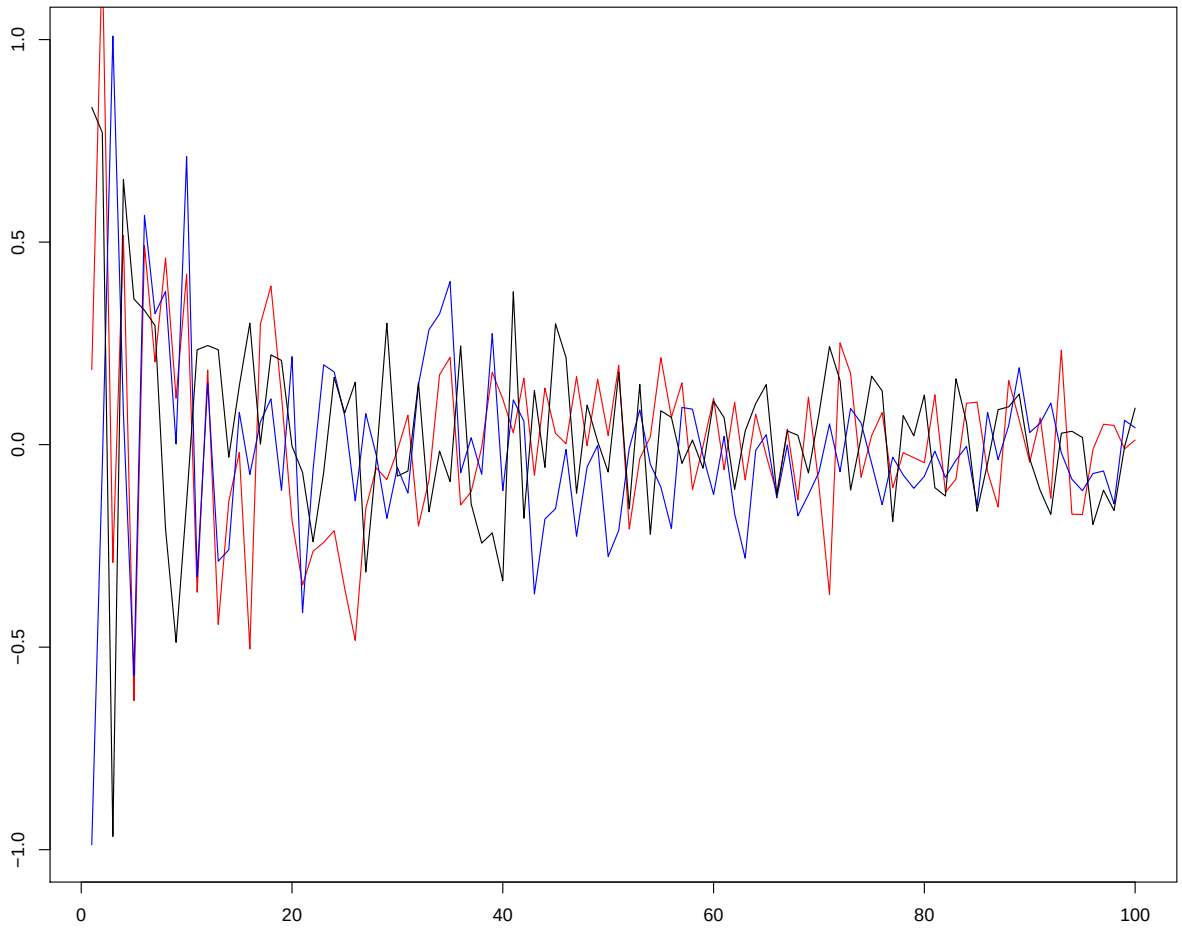
- Note that $\mathrm{MSE}(\hat{\theta}_n, \theta) \overset{n \to \infty}{\longrightarrow} 0 \implies \hat{\theta}_n \xrightarrow{p} \theta$.
- When an estimator has this property, we call it consistent.

Definition (Consistency)

An estimator $\hat{\theta}_n$ of θ, constructed on the basis of a sample of size n, is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

Note that a vanishing MSE implies consistency, but the converse generally fails.

Consistency of sample mean of sample mean of $Y_1, ..., Y_n \sim \mathcal{N}(0, 1)$, towards the true parameter value 0 (by law of large numbers).



Identifiability

Is it always possible to get consistent estimators?

Depends on whether the estimation problem is well-posed

Definition (Identifiability)

A probability model $\{F_\theta\}_{\theta \in \Theta}$ is called identifiable if for any pair $\theta_1, \theta_2 \in \Theta$ we have the implication

$$\theta_1 \neq \theta_2 \implies F_{\theta_1} \neq F_{\theta_2}.$$

- Lack of identifiability means that the same model can be produced by more than one parameter.
- In this case we could never distinguish amongst the parameters that give the same model.
- Example: if we have $N(\mu_1 + \mu_2, \sigma^2)$, we can never identify each μ_i, but only their sum.
- Henceforth we will tacitly assume identifiability (and make special mention if it is at stake).

- We can use the MSE to compare estimators or to gauge their performance.
- But is there a *best possible MSE* for a given problem?
- This is a very difficult problem, equivalent to finding a uniformly optimal estimator: a statistic T_* such that

$$\mathrm{MSE}(T_*, \theta) \leq \mathrm{MSE}(T, \theta)$$

  for all $\theta \in \Theta$ and all other estimators T.
- Here's a simpler question to ask instead:

Among unbiased estimators (bias zero), can we make the MSE arbitrarily small?

- If so, how? (what is the crucial ingredient at play?)

- Rephrasing, we are asking whether there is *fundamental lower bound* for the variance of an unbiased estimator of θ.
- We will concetrate on 1-dimensional parameters for simplicity.

The Question

For $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$ with joint density/frequncy $f(\boldsymbol{y}; \theta)$ depending on an unknown $\theta \in \mathbb{R}$, does there exist some function $\Lambda(\theta) > 0$ such that

$$\mathrm{var}[\hat{\theta}] \geq \Lambda(\theta), \quad \forall \theta$$

for any estimator $\hat{\theta}$ such that $\mathbb{E}[\hat{\theta}] = \theta$?

- At a next step we can ask if this bound is achievable.

Let's assume we can interchange differentiation and integration in the form

$$\frac{d}{d\theta}\int S(\boldsymbol{y})f(\boldsymbol{y};\theta)\,d\boldsymbol{y} \overset{!}{=} \int S(\boldsymbol{y})\tfrac{f(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)}\tfrac{d}{d\theta}f(\boldsymbol{y};\theta)\,d\boldsymbol{y} = \int S(\boldsymbol{y})f(\boldsymbol{y};\theta)\tfrac{d}{d\theta}\log f(\boldsymbol{y};\theta)\,d\boldsymbol{y}$$

whenever an integral such as the one on the left hand side presents itself.

1. Setting $U = \frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)$ and $S(\boldsymbol{y}) = 1$, this gives that

$$\mathbb{E}[U] = \int \frac{\partial}{\partial\theta} f(\boldsymbol{y}) \log f(\boldsymbol{y};\theta)\,d\boldsymbol{y} = \frac{d}{d\theta}\int f(\boldsymbol{y};\theta)\,d\boldsymbol{y} = 0$$

2. Therefore $\operatorname{var}[U] = \mathbb{E}[U^2] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right)^2\right]$

3. For $\hat{\theta}$ unbiased, our "interchange ansatz" with $S(\boldsymbol{y}) = \hat{\theta}(\boldsymbol{y})$ gives

$$\operatorname{cov}\left(\hat{\theta}, U\right) = \mathbb{E}[\hat{\theta}\,U] - \mathbb{E}[\hat{\theta}]\underbrace{\mathbb{E}[U]}_{=0} = \int \hat{\theta}(\boldsymbol{y})f(\boldsymbol{y};\theta)\frac{d}{d\theta}\log f(\boldsymbol{y};\theta)\,d\boldsymbol{y} = \frac{d}{d\theta}\underbrace{\mathbb{E}[\hat{\theta}]}_{=\theta} = 1$$

Now the Cauchy-Schwartz inequality gives

$$\operatorname{var}(\hat{\theta}) \geq \frac{\operatorname{cov}^2\left(\hat{\theta}, U\right)}{\operatorname{var}(U)} \implies \operatorname{var}(\hat{\theta}) \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right)^2\right]}$$

In summary, we have established:

Cramér-Rao Lower Bound

Given sufficient regularity, any unbiased estimator $\hat{\theta}(\boldsymbol{Y})$ of finite variance satisfies:

$$\operatorname{var}[\hat{\theta}(\boldsymbol{Y})] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right)^2\right]} = \frac{1}{\mathcal{I}_n(\theta)}$$

The quantity $\mathcal{I}_n(\theta)$ is fundamental, and called Fisher information.

- If $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ has iid entries, we have $f(\boldsymbol{y};\theta) = \prod_{i=1}^n f(y_i;\theta)$ and so

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta).$$

- By further interchanges of integration/differentiation it typically holds that

$$\mathcal{I}_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{Y};\theta)\right]$$

- The deeper meaning of all this will become clearer when we study likelihood.

Is the Cramér-Rao bound tight (achievable) though?

$$\text{if} \quad \operatorname{var}[\hat{\theta}] = \frac{1}{\mathcal{I}_n(\theta)}$$

$$\text{then} \quad \operatorname{var}[\hat{\theta}] = \frac{\operatorname{cov}^2\left[\hat{\theta}, \frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right]}{\operatorname{var}\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)\right]}$$

which occurs if and only if $\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta)$ is a linear function of $\hat{\theta}$ (correlation 1):

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{Y};\theta) = A(\theta)\hat{\theta}(\boldsymbol{Y}) + B(\theta)$$

Solving this differential equation yields, for all $\boldsymbol{y}$,

$$\log f(\boldsymbol{y};\theta) = A^*(\hat{\theta}) + B^*(\theta) + S(\boldsymbol{y})$$

so that $\operatorname{var}_\theta(\hat{\theta})$ attains the lower bound if and only if the density (frequency) of $\boldsymbol{Y}$ is a one-parameter exponential family with sufficient statistic $\hat{\theta}$.

So what ingredients go into pushing towards this lower bound?

The Rao-Blackwell Theorem tells us that sufficiency is key:

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator of θ with finite variance, and let T be sufficient for θ. Then $\hat{\theta}^* := \mathbb{E}[\hat{\theta}|\, T]$ is also an unbiased estimator of θ and

$$\operatorname{var}(\hat{\theta}^*) \leq \operatorname{var}(\hat{\theta}).$$

Equality is attained if and only if $\mathbb{P}_\theta[\hat{\theta}^* = \hat{\theta}] = 1$.

Comments:

- Throwing away irrelevant aspects of the data improves estimation quality.
- These irrelevant aspects contribute to the variation of the estimator (as they have sampling variation of their own), but without furnishing any useful information on the parameter
- $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|\, T]$ is called a "Rao-Blackwellised" version of $\hat{\theta}$.

Proof.

Since T is sufficient for θ, $\mathbb{E}[\hat{\theta}|\, T = t] = h(t)$ is independent of θ, so that $\hat{\theta}^*$ is well-defined as a statistic (depends only on $\boldsymbol{Y}$ and not θ). Then,

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta}|\, T]] = \mathbb{E}[\hat{\theta}] = \theta.$$

Furthermore, from the law of total variance, we have

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}[\mathbb{E}(\hat{\theta}|\, T)] + \mathbb{E}[\mathrm{var}(\hat{\theta}|\, T)] \geq \mathrm{var}[\mathbb{E}(\hat{\theta}|\, T)] = \mathrm{var}(\hat{\theta}^*)$$

In addition, note that

$$\mathrm{var}(\hat{\theta}|\, T) := \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}|\, T])^2|\, T] = \mathbb{E}[(\hat{\theta} - \hat{\theta}^*)^2|\, T]$$

so that $\mathbb{E}[\mathrm{var}(\hat{\theta}|\, T)] = \mathbb{E}(\hat{\theta} - \hat{\theta}^*)^2 > 0$ unless if $\mathbb{P}(\hat{\theta}^* = \hat{\theta}) = 1$. □

Suppose that $\hat{\theta}$ is an unbiased estimator of $g(\theta)$ and T, S are θ-sufficient.

- What is the relationship between $\mathrm{var}(\underbrace{\mathbb{E}[\hat{\theta}|\, T]}_{\hat{\theta}^*_T}) \overset{?}{\gtreqless} \mathrm{var}(\underbrace{\mathbb{E}[\hat{\theta}|\, S]}_{\hat{\theta}^*_S})$
- Intuition suggests that whichever of T, S carries the least irrelevant information (in addition to the relevant information) should "win"
  ↪ More formally, if $T = h(S)$ then we should expect that $\hat{\theta}^*_T$ dominate $\hat{\theta}^*_S$.

Proposition

For $\hat{\theta}$ an unbiased estimator of θ and T, S two θ-sufficient statistics, define

$$\hat{\theta}^*_T := \mathbb{E}[\hat{\theta}|\, T] \quad \& \quad \hat{\theta}^*_S := \mathbb{E}[\hat{\theta}|\, S].$$

Then, the following implication holds

$$T = h(S) \implies \mathrm{var}(\hat{\theta}^*_T) \leq \mathrm{var}(\hat{\theta}^*_S)$$

- Essentially this means that the best possible "Rao-Blackwellisation" is achieved by conditioning on a minimally sufficient statistic.

Proof.

Recall the *tower property* of conditional expectation: if $Y = f(X)$, then

$$\mathbb{E}[Z|\,Y] = \mathbb{E}\{\mathbb{E}(Z|\,X)|\,Y\}.$$

Since $T = f(S)$ we have

$$\begin{aligned} \hat{\theta}^*_T &= \mathbb{E}[\hat{\theta}|\,T] \\ &= \mathbb{E}[\mathbb{E}(\hat{\theta}|\,S)|\,T] \\ &= \mathbb{E}[\hat{\theta}^*_S|\,T] \end{aligned}$$

The conclusion now follows from the Rao-Blackwell theorem. □

- So now we have a means to judge the quality of an estimator
- In certain cases, we even know what's the best performance we can hope for.
- And (minimal) sufficiency can help us approach it.
- But how can we actually come up with an estimator in the first place?
- We need general methods that can be applied in any model context to yield an estimator.
- Preferably methods that yield good estimators relative to our performance measures/bounds.

↪ The main focus will be on a key method called maximum likelihood.

Motivation: recall our understanding of statistics as "inverse probability".

↪ For the moment, consider the discrete case for simplicity.

Probability Perspective

Given a parameter $\theta \in \Theta$, then for any $(y_1, ..., y_n)^\top \in \mathcal{Y}^n$, we can evaluate

$$(y_1, ..., y_n) \mapsto \mathbb{P}_\theta[Y_1 = y_1, ..., Y_n = y_n]$$

that is, how the probability varies as a function of the sample (=the result).

Statistics Perspective

Given a sample $(y_1, ..., y_n)^\top \in \mathcal{Y}^n$, then for any $\theta \in \Theta$ we can calculate

$$\theta \mapsto \mathbb{P}_\theta[Y_1 = y_1, ..., Y_n = y_n]$$

that is, how the probability varies as a function of θ (=the model).

Intuition: we imagine that, having our sample, the values of θ that are most plausible are those that render the observed sample most probable...

This motivates the following definition...

Definition (Likelihood)

Let $(Y_1, \ldots, Y_n)$ be a sample of random variables with joint density/frequency $f(y_1, ..., y_n; \theta)$, where $\theta \in \mathbb{R}^p$. The likelihood of θ is defined as

$$L(\theta) = f(Y_1, ..., Y_n; \theta).$$

If $(Y_1, ..., Y_n)^\top$ has i.i.d. entries, each with density/frequency $f(y_i; \theta)$ then,

$$L(\theta) = \prod_{i=1}^{n} f(Y_i; \theta)$$

... and the following estimation method

Definition (Maximum Likelihood Estimator)

In the same context, a maximum likelihood estimator (MLE) of $\hat{\theta}$ is an estimator such that

$$L(\theta) \leq L(\hat{\theta}), \qquad \forall \theta \in \Theta.$$

Many comments are in order:

- When there exists a unique maximum, we speak of the MLE $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$
- The likelihood is a random function.
- It is the joint density/frequency of the sample, but viewed as a function of θ.
- It is NOT "the probability of θ"
- $L(\theta)$ is the answer to the question *how does the joint density/probability of the sample vary as we vary θ?*
  - In the discrete case it is exactly "the probability of observing our sample" as a function of θ.
  - In the continuous case, since $F(\boldsymbol{y} + \boldsymbol{\varepsilon}/2\,; \theta) - F(\boldsymbol{y} - \boldsymbol{\varepsilon}/2; \theta) \approx ||\varepsilon|| f(\boldsymbol{y}; \theta)$ as $||\varepsilon|| \downarrow 0$, we can view $||\epsilon|| \times L(\theta)$ as being the "probability of observing something in the neighbourhood of our sample", as a function of θ.

- If a sufficient statistic T exists for θ then Fisher-Neyman factorisation implies

$$L(\theta) = g(T(\boldsymbol{Y}); \theta) h(\boldsymbol{Y}) \propto g(T(\boldsymbol{Y}); \theta)$$

  i.e. any MLE depends on data only through a sufficient statistic.
- Since the sufficient statistic was arbitrary, if a minimally sufficient statistic exists, the MLE will have used an estimator that has achieved the maximal sufficient reduction of the data.
- MLE's are also *equivariant*. If $g : \Theta \to \Theta'$ is a bijection, and if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ (you can take the hat out: $g(\hat{\theta}) = \widehat{g(\theta)}$)
- When the likelihood is differentiable in θ, its maximum $L(\theta)$ must solve the equation

$$\nabla_\theta L(\theta) = 0,$$

- But before declaring a solution as an MLE, we must verify it to be a maximum (and not a minimum!).

- If the likelihood is twice differentiable in θ, we can verify this by checking

$$-\nabla_\theta^2 L(\theta)\big|_{\theta=\hat{\theta}} \succ 0,$$

  i.e that minus the Hessian is positive definite. In one dimension, his reduces to the standard second derivative criterion.

- To solve $\nabla_\theta L(\theta) = 0$ when the Y_i are independent, we must painfully calculate the derivative of an n-fold product.
- To avoid this, we focus instead on the loglikelihood $\ell(\theta) := \log L(\theta)$ instead. Maximisation of ℓ is equivalent to maximisation of L by monotonicity.
- When the Y_i are independent, ℓ has the advantage of being a sum rather than a product

$$\ell(\theta) = \log\left(\prod_{i=1}^n f_{Y_i}(Y_i;\theta)\right) = \sum_{i=1}^n \log f_{Y_i}(Y_i;\theta).$$

- Of course, under twice differentiability, verification of a maximum can be checked again by whether or not

$$\nabla_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} = 0 \quad \& \quad -\nabla_\theta^2 \ell(\theta)\big|_{\theta=\hat{\theta}} \succ 0.$$

Example (MLE for Bernoulli trials)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p)$. The likelihood is

$$L(p) = \prod_{i=1}^n f(Y_i;p) = \prod_{i=1}^n p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^n Y_i}(1-p)^{n-\sum_{i=1}^n Y_i}$$

giving loglikelihood

$$\ell(p) = \log p \sum_{i=1}^n Y_i + \log(1-p)\left(n - \sum_{i=1}^n Y_i\right).$$

This is twice differentiable in p and we calculate

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^n Y_i - (1-p)^{-1}\left(n - \sum_{i=1}^n Y_i\right).$$

Example (MLE for Bernoulli trials, continued)

Solving

$$p^{-1}\sum_{i=1}^{n} Y_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} Y_i\right) = 0,$$

we get the unique root $\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$. Calling this $\hat{p}$, we now verify that

$$\frac{d^2}{dp^2}\ell(p) = -p^2\sum_{i=1}^{n} Y_i - (1-p)^{-2}\left(n - \sum_{i=1}^{n} Y_i\right),$$

which is a negative expression, since $0 \leq \sum_{i=1}^{n} Y_i \leq n$ andt $p \in (0,1)$. Thus

$$\hat{p} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is the unique MLE of p. □

Example (MLE for exponential distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Exp(\lambda)$. The likelihood is

$$L(\lambda) = \prod_{i=1}^{n} f(Y_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda Y_i} = \lambda^n \exp\left\{-\lambda\sum_{i=1}^{n} Y_i\right\}.$$

and the log likelihood is

$$\ell(\lambda) = n\log\lambda - \lambda\sum_{i=1}^{n} Y_i.$$

This is twice differentiable in λ and we calculate

$$\frac{d}{d\lambda}\ell(\lambda) = n\lambda^{-1} - \sum_{i=1}^{n} Y_i.$$

Example (MLE for exponential distribution, continued)

Setting $\ell'(\lambda) = 0$ we get a unique root

$$\left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right)^{-1} = 1/\bar{Y}.$$

Call this $\hat{\lambda}$, and note that

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{n}{\lambda^2}$$

is always negative, since $\lambda > 0$. Thus

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right)^{-1} = 1/\bar{Y}$$

is the unique MLE of λ. □

Example (MLE for Gaussian distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{2\sigma^2} \right\}.$$

giving loglikelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2.$$

All partial second derivatives exist and are

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu) \\ \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \mu)^2. \end{aligned}$$

Example (MLE for Gaussian distribution, continued)

Solving $\nabla_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)=0$ for (μ,σ^2) gives a system of equations in two unknowns, with unique root

$$\left(\bar{Y}, n^{-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right).$$

Call this $(\hat{\mu},\hat{\sigma}^2)$, and let's verify it's a maximum. Note that

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)=-\frac{n}{\sigma^2},\quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(Y_i-\mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)=\frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)=-\frac{\sum_{i=1}^{n}(Y_i-\mu)}{\sigma^4}=\frac{n\mu-n\bar{Y}}{\sigma^4}.$$

Calculating these derivatives at $(\hat{\mu},\hat{\sigma}^2)$, we get

$$\left.\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{\hat{\sigma}^2},\quad \left.\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{2\hat{\sigma}^4}$$

Example (MLE for Gaussian distribution, continued)

$$\left.\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=\left.\frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}=\frac{n\hat{\mu}-n\hat{\mu}}{\hat{\sigma}^4}=0.$$

Thus the matrix

$$\left[-\nabla^2_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)}\right]$$

is diagonal. If both of its diagonal elements are positive, then it will be positive definite. This is indeed the case since $\hat{\sigma}^2>0$ nd so the unique MLE of (μ,σ^2) is given by

$$(\hat{\mu},\hat{\sigma}^2)=\left(\bar{Y}\ ,\ \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right).$$

□

Note that from our Gaussian sampling results we get that σ^2 is biased.

Example (MLE for Poisson Distribution)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Then

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \log(Y_i!)$$

Setting $\nabla_\theta \log L(\theta) = -n + \lambda^{-1} \sum Y_i = 0$ we obtain $\hat{\lambda} = \bar{Y}$ since $\nabla^2_\theta \log L(\theta) = -\lambda^{-2} \sum Y_i < 0$.

Example (MLE for Uniform Distribution – a non-differentiable case)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. The likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq Y_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq Y_{(n)}\}.$$

Hence if $\theta < Y_{(n)}$ the likelihood is zero. In the domain $[Y_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\hat{\theta} = Y_{(n)}$.

Example (Equivariance of the MLE)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, and suppose we're interested in estimating $\mathbb{P}[Y_1 \leq y]$, for a given $y \in \mathbb{R}$. Note that

$$\mathbb{P}[Y_1 \leq y] = \mathbb{P}[Y_1 - \mu \leq y - \mu] = \Phi(y - \mu),$$

where Φ is the standard normal CDF. The mapping $\mu \mapsto \Phi(y - \mu)$ is bijective, since Φ is strictly monotone. So by equivariance, the MLE of $\mathbb{P}[Y_1 \leq y]$ is $\Phi(y - \hat{\mu})$, where $\hat{\mu}$ is the MLE of μ (which by our previous example is $\hat{\mu} = \bar{Y}$).

Example (Equivariance and usual vs natural parameterisation)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, with

$$f(y) = \exp\{\phi\, T(y) - \gamma(\phi) + S(y)\}, \qquad y \in \mathcal{Y}$$

where $\phi \in \Phi \subseteq \mathbb{R}$ is the natural parameter. Suppose we can write $\phi = \eta(\theta)$, where $\theta \in \Theta$ is the usual parameter and $\eta : \Theta \to \Phi$ is a differentiable bijection (so that $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, for $d = \gamma \circ \eta$). In this notation, the density/frequency takes the form

$$\exp\{\phi\, T(y) - \gamma(\phi) + S(y)\} = \exp\{\eta(\theta)\, T(y) - d(\theta) + S(y)\}.$$

Equivariance now implies that if $\hat{\theta}$ is the MLE of θ, then $\eta(\hat{\theta})$ is the MLE of $\phi = \eta(\theta)$. The converse is also true: if $\hat{\phi}$ is the MLE of ϕ, then $\eta^{-1}(\hat{\phi})$ is the MLE of $\theta = \eta^{-1}(\phi)$. □

Examples show that likelihood generally gives sensible estimators – still:

- Beyond intuition, is there a canonical mathematical reason for it?
- What rigorous guarantees can we offer?
  - ↪ Can we get consistency?
  - ↪ Can we approach reasonable MSE performance?

To answer these questions, we go back to entropy and Kullback-Leibler divergence.

Consider the random function

$$\Psi_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}[\log f(\,Y_i;\boldsymbol{u}) - \log f(\,Y_i;\boldsymbol{\theta})]$$

which is maximized at $\hat{\boldsymbol{\theta}}_n$. By the law of large numbers, for each $\boldsymbol{u} \in \Theta$,

$$\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) = \mathbb{E}\left[\log\left(\frac{f(\,Y_i;\boldsymbol{u})}{f(\,Y_i;\boldsymbol{\theta})}\right)\right] = -KL\big(f(\,Y_i;\boldsymbol{u})\big\|f(\,Y_i;\boldsymbol{\theta})\big)$$

- The latter is minimised at $\boldsymbol{\theta}$ and so $\Psi(u)$ is maximized at $\boldsymbol{\theta}$.
- Moreover, unless $f(x;\boldsymbol{u}) = f(x;\boldsymbol{\theta})$ for all $x \in \operatorname{supp} f$, we have $\Psi(\boldsymbol{u}) < 0$
- It follows that Ψ is uniquely maximised at $\boldsymbol{\theta}$

MLE can be regarded as a minimiser of an approximate (empirically constructed) KL-divergence from the truth!

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u})\ \forall\ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.
- If $\theta \in \mathbb{R}$, can prove consistency if f is regular enough & MLE exists uniquely.
- If $\theta \in \mathbb{R}^p$, we need more information on the form of the likelihood function
  - ↪ For instance concavity and existence will usually give us consistency. We will show consistency in exponential families using this approach.
  - ↪ More general situations require stronger forms of convergence of $\Psi_n(u) \to \Psi(u)$ plus additional regularity conditions.

When we can deduce consistency, though, we get some very nice properties for the (asymptotic) sampling distribution of the MLE...

Example (Consistency of MLE in $\theta \in \mathbb{R}$)

Let $Y_1, ..., Y_n \overset{iid}{\sim} f(y; \theta_0)$ where f is C^1 with respect to θ. Assume that $\forall\ n$, there exists a unique MLE $\hat{\theta}_n$. We will show that $\hat{\theta}_n \xrightarrow{p} \theta$.

Define

$$\Xi_n(u) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial}{\partial u} \log \left(\frac{f(Y_i; u)}{f(Y_i; \theta)} \right) \right] \quad \text{and} \quad \Xi(u) = \mathbb{E} \left[\frac{\partial}{\partial u} \log \left(\frac{f(Y_i; u)}{f(Y_i; \theta)} \right) \right],$$

so that

- $\Xi_n(\hat{\theta}_n) = 0$ uniquely, by uniqueness of the MLE.
- $\Xi(\theta) = 0$ uniquely, assuming regularity allowing interchange of $\mathbb{E}$ and $\frac{\partial}{\partial u}$.

Since f is C^1, we have the inequality

$$\mathbb{P}[\Xi_n(\theta - \varepsilon) < 0 \ \& \ \Xi_n(\theta + \varepsilon) > 0] \leq \mathbb{P}[\theta - \varepsilon < \hat{\theta}_n < \theta + \varepsilon]$$

because the event on the left hand side implies that on the right hand side. Finally, the law of large numbers implies that $\Xi_n(u) \xrightarrow{p} \Xi(u)$ for any u, so that the left hand side converges to 1, yielding consistency.

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families)

Consider $Y_1, ..., Y_n \overset{iid}{\sim} f(y; \boldsymbol{\phi})$ from a k-parameter exponential family

$$f(y) = \exp \left\{ \sum_{j=1}^{k} \phi_j \, T_j(y) - \gamma(\phi_1, ..., \phi_k) + S(y) \right\}, \boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top \in \Phi \text{ open.}$$

The likelihood and loglikelihood (up to constants w.r.t. $\boldsymbol{\phi}$) are given by

$$L(\boldsymbol{\phi}) = \exp \left\{ \boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi}) \right\} \quad \& \quad \ell(\boldsymbol{\phi}) = \boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})$$

where

$$\boldsymbol{\tau} = (\tau_1, ..., \tau_k)^\top, \qquad \tau_j(y_1, \ldots, y_n) = \textstyle\sum_{i=1}^{n} T_j(y_i).$$

If it exists, the MLE $\hat{\boldsymbol{\phi}}_n$ must thus satisfy

$$\nabla_{\boldsymbol{\phi}} \ell(\hat{\boldsymbol{\phi}}_n) = 0 \implies \nabla_{\boldsymbol{\phi}} \gamma(\hat{\boldsymbol{\phi}}_n) = n^{-1} \boldsymbol{\tau}.$$

Furthermore, existence of the MLE guarantees uniqueness by strict concavity:

$$-\nabla^2_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}) = n \nabla^2_{\boldsymbol{\phi}} \gamma(\boldsymbol{\phi}) = \text{cov}\{\boldsymbol{\tau}\} \succ 0,$$

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families, ctd)

Now notice that by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} T_j(Y_i) \xrightarrow{p} \mathbb{E}[T_j] = \frac{\partial}{\partial \phi_j}\gamma(\boldsymbol{\phi}), \qquad j = 1, ..., k.$$

It follows that

$$\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n) = n^{-1}\boldsymbol{\tau} \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}).$$

Now if $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ were continuously invertible, with inverse map h, then the continuous mapping theorem would give us:

$$\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n) \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \implies h\big(\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n)\big) \xrightarrow{p} h\big(\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi})\big) \implies \hat{\boldsymbol{\phi}}_n \xrightarrow{p} \boldsymbol{\phi}.$$

In fact, the inverse function theorem tells us that the infinitely differentiable function $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ must admit a continuously differentiable inverse map h locally.

In summary: provided it exists, the MLE of the natural parameter in a k-parameter natural exponential family with open parameter space Φ is consistent.

Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.

For simplicity, assume $X_1, ..., X_n$ are iid with density/frequency) $f(x;\theta)$, $\theta \in \mathbb{R}$.

Introduce the notation:

- $\ell(x_i;\theta) = \log f(x_i;\theta)$
- $\ell'(x_i;\theta)$, $\ell''(x_i;\theta)$ and $\ell'''(x_i;\theta)$ are partial derivatives w.r.t θ.

Regularity Conditions ($*$)

(A1) Θ is an open subset of $\mathbb{R}$.

(A2) The support of f, supp$f$, is independent of θ.

(A3) f is thrice continuously differentiable w.r.t. θ for all $x \in$ supp$f$.

(A4) $\mathbb{E}_\theta[\ell'(X_i;\theta)] = 0\ \forall\theta$ and $\text{var}_\theta[\ell'(X_i;\theta)] = \mathcal{I}_1(\theta) \in (0,\infty)\ \forall\theta$.

(A5) $-\mathbb{E}_\theta[\ell''(X_i;\theta)] = \mathcal{J}_1(\theta) \in (0,\infty)\ \forall\theta$.

(A6) $\exists\ M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}_{\theta_0}[M(X_i)] < \infty$ and

$$|\theta - \theta_0| < \delta \implies |\ell'''(x;\theta)| \leq M(x)$$

Let's demistify these conditions...

- If Θ is open, then for θ_0 the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ_0 (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp}\, f} f(x;\theta)\,dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)\,dx = \int \ell'(x;\theta) f(x;\theta)\,dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

  so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.
- Similarly, (A5) requires that ℓ'' have a first moment for all θ.
- Conditions (A2) and (A6) are smoothness conditions that will allow us to "linearize" the problem, while the other conditions will allow us to "control" the random linearization.
- Furthermore, if we can differentiate twice under the integral sign

$$0 = \int \frac{d}{d\theta}[\ell'(x;\theta) f(x;\theta)]\,dx = \int \ell''(x;\theta) f(x;\theta)\,dx + \int (\ell'(x;\theta))^2 f(x;\theta)\,dx$$

  so that $\mathcal{I}(\theta) = \mathcal{J}(\theta)$.

Asymptotic Normality of the MLE

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(x;\theta)$ and satisfying the stated regularity conditions. If the MLE $\hat{\theta}_n$ exists uniquely and is consistent, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathcal{I}_1(\theta)}{\mathcal{J}_1^2(\theta)}\right).$$

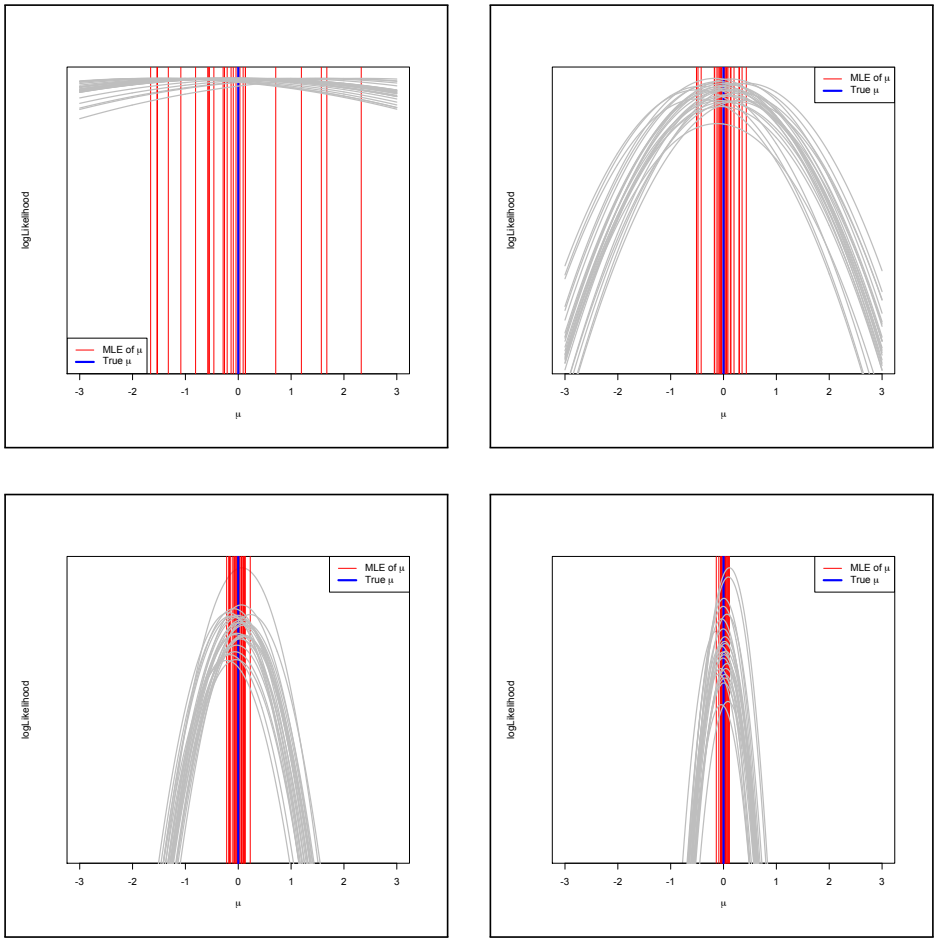
When $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$, we have of course $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$.

- Note that this can be interpreted as

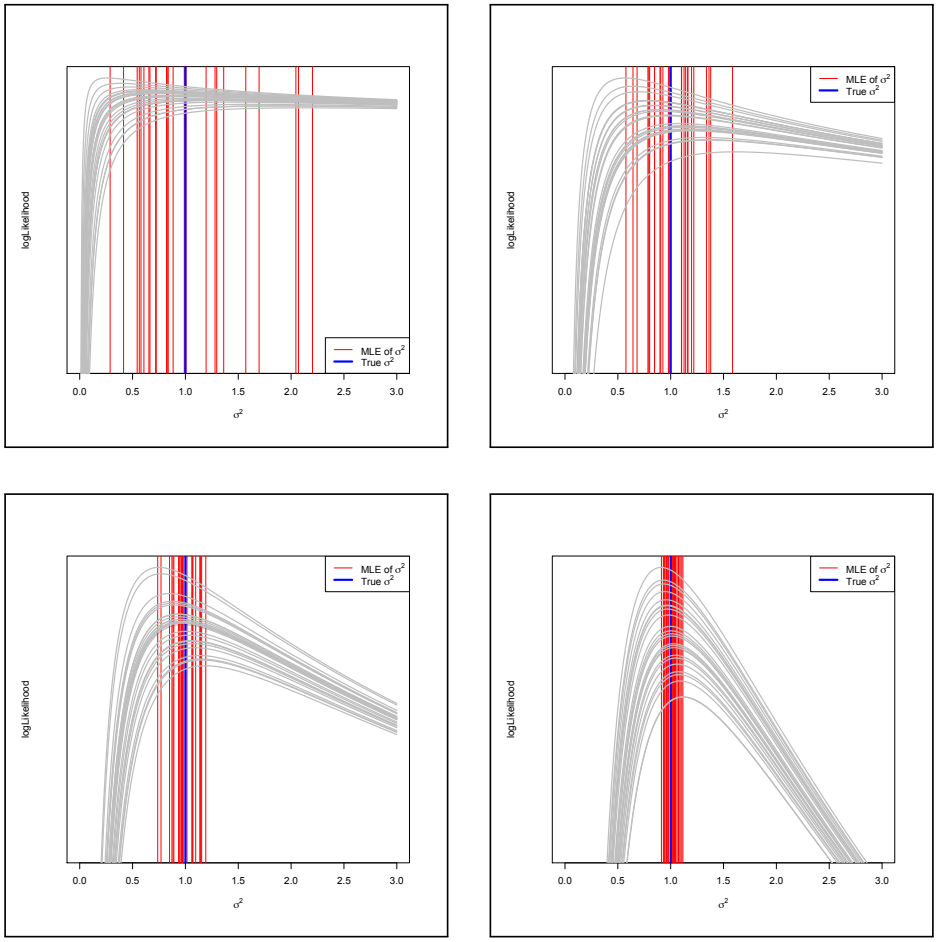
$$\hat{\theta}_n \overset{d}{\approx} N\left(\theta, \frac{1}{n\mathcal{I}_1(\theta)}\right) \equiv N\left(\theta, \frac{1}{\mathcal{I}_n(\theta)}\right).$$

- In order words: the MLE is approximately normally distributed, approximately unbiased, and approximately achieving the Cramér-Rao lower bound!

Why $\mathcal{I}_n(\theta)$? (... curvature)



Why $\mathcal{I}_n(\theta)$? (... curvature)



Proof.

Under conditions (A1)-(A3), if $\hat{\theta}_n$ maximizes the likelihood, we have

$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0.$$

Expanding this equation in a Taylor series, we get

$$\begin{aligned} 0 = \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) &= \sum_{i=1}^{n} \ell'(X_i; \theta) + \\ &\quad + (\hat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta) \\ &\quad + \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*) \end{aligned}$$

with θ_n^* lying between θ and $\hat{\theta}_n$.

Dividing accross by $\sqrt{n}$ yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i; \theta) + \sqrt{n}(\hat{\theta}_n - \theta) \frac{1}{n} \sum_{i=1}^{n} \ell''(X_i; \theta) \\ &\quad + \frac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta)^2 \frac{1}{n} \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*) \end{aligned}$$

which suggests that $\sqrt{n}(\hat{\theta}_n - \theta)$ equals

$$\frac{-n^{-1/2} \sum_{i=1}^{n} \ell'(X_i; \theta)}{n^{-1} \sum_{i=1}^{n} \ell''(X_i; \theta) + (\hat{\theta}_n - \theta)(2n)^{-1} \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)}.$$

Now, from the central limit theorem and condition (A4), it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i; \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\theta)).$$

Next, the weak law of large numbers along with condition (A5) implies

$$\tfrac{1}{n}\sum_{i=1}^{n}\ell''(X_i;\theta) \xrightarrow{p} -\mathcal{J}(\theta).$$

By Slutsky's lemma, the theorem will follow if we show that $R_n \xrightarrow{p} 0$. This is established in the next lemma, which we appeal to, completing the proof. □

Lemma

In the same context as in the previous theorem,

$$R_n = (\hat{\theta}_n - \theta)\tfrac{1}{2n}\sum_{i=1}^{n}\ell'''(X_i;\theta_n^*) \xrightarrow{p} 0$$

for any random variable θ_n^* on the segment joining $\hat{\theta}_n$ and θ.

Proof. $(*)$

We have that for any $\epsilon > 0$

$$\mathbb{P}[|R_n| > \epsilon] = \underbrace{\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| > \delta]}_{\leq \mathbb{P}[|\hat{\theta}_n - \theta| > \delta] \xrightarrow{p} 0} + \mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta]$$

If $|\hat{\theta}_n - \theta| < \delta$, (A6) implies $|R_n| \leq \frac{\delta}{2n}\sum_{i=1}^{n} M(X_i) = \bar{M}_n$.
so we may write

$$\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta] \leq \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta\bar{M}_n]$$

and for $\xi > 0$, the last term can be bounded by

$$\begin{aligned}\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta\bar{M}_n, \bar{M}_n \leq M + \xi] + \\ + \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta\bar{M}_n, \bar{M}_n > M + \xi]\end{aligned}$$

which in turn is bounded by

$$\begin{aligned}\leq \quad & \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[\bar{M}_n > M + \xi] \\ \leq \quad & \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[|\bar{M}_n - M| > \xi]\end{aligned}$$

But the law of large numbers implies that

$$\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} M(X_i) \xrightarrow{p} \mathbb{E}[M(X_1)] < \infty,$$

It follows that

$$\mathbb{P}[|\bar{M}_n - M| > \xi] \to 0.$$

Since we can always choose δ to be as small as we wish, we can make the term

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)]$$

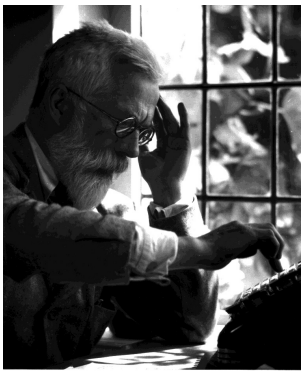
equal to zero. In summary, we have established that $R_n \xrightarrow{p} 0$

□

Optimal Estimation and the Role of Bias in Finite Samples

Does this mean that likelihood estimators are essentially optimal?

- The result holds asymptotically in n, so care must be taken in interpreting it.
- For finite sample size n, the theorem says very little.
- Though bias must vanish asymptotically for consistency to go through...
- ... a little bit of bias can help reduce variance in finite samples.
- The delicate finite-sample tradeoff of bias and variance is decisive.
- Manifested both in parametric and (quite lucidly) nonparametric estimation.



Here's a spectacularly simple (and surprising) counterexample by Charles Stein.

Stein's setup

1. Let $Y_1, ..., Y_n$ be independent random variables.
2. Assume that $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
   - Notice that each Y_i has a different mean but same variance.
3. Suppose that σ^2 is known, say $\sigma^2 = 1$ (wlog)
4. Unknown parameter to estimate: $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\top \in \mathbb{R}^n$
5. Consider mean squared error to judge quality.

↪ Looks like the usual setup, but notice the subtlety: the dimension of the parameter $\dim(\mu)=n$ grows along with the dimension of the sample size.

Is this artificial? No: many modern problems have # parameters comparable to # observations.

↪ Will later see other examples with parameter dimension fixed relative to sample size (ridge regression).

By independence, the loglikelihood in Stein's setup is

$$\ell(\boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu_i)^2$$

and by differentiation and convexity, we have

$$\hat{\boldsymbol{\mu}} = (Y_1, ..., Y_n)^\top$$

is the unique MLE of μ.

- Intuition: we essentially have n Gaussian mean separate problems, each of sample size 1.
- Hence separately estimate each of these means by corresponding sample mean (which is Y_i since there is only 1 observation in each sample)

The MSE of this estimator can be easily calculated to be equal to n:

$$\mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{n}(Y_i - \mu_i)^2 = n.$$

Stein realised that one can always improve this MSE by cleverly introducing bias...

Theorem (James-Stein)

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$ be such that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, I_{n\times n})$, $\boldsymbol{\mu} \in \mathbb{R}^n$ (Stein's setup). Let $\tilde{\boldsymbol{\mu}}_a$ be an estimator defined as

$$\tilde{\boldsymbol{\mu}}_a = \left(1 - \frac{a}{\|\boldsymbol{Y}\|^2}\right) \boldsymbol{Y} = \left(1 - \frac{a}{\|\hat{\boldsymbol{\mu}}\|^2}\right) \hat{\boldsymbol{\mu}},$$

i.e. a *shrunken* version of the MLE $\hat{\boldsymbol{\mu}}$. Then, if $n \geq 3$,

1. for all $a \in (0, 2n-4)$,
$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) \leq \mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$$
2. for $a = n-2$,
$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_{n-2}, 0) < \mathrm{MSE}(\hat{\boldsymbol{\mu}}, 0)$$
3. For all $\boldsymbol{\mu} \in \mathbb{R}^n$ and all $a \in (0, 2n-4)$,
$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_{n-2}, \boldsymbol{\mu}) \leq \mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}).$$

Comments:

- The result is surprising, not just because the MLE is outperformed.
- The JS estimator takes the MLE and shrinks it towards zero.
- The amount of shrinkage depends on $\|\boldsymbol{Y}\|$
- That is, we take into account the estimate of μ_i in order to estimate μ_j ($i \neq j$), even though these are completely unrelated (no "smoothness" assumptions on $\boldsymbol{\mu}$).
- The performance of the MLE as compared to the JS estimator becomes worse and worse as n grows.
- The proof is surprisingly elementary (once one knows what to look for!)

We'll need a simple lemma first.

Lemma $(*)$.

Let $Y \sim \mathcal{N}(\theta, \sigma^2)$ and $h : \mathbb{R} \to \mathbb{R}$ be differentiable. If

1. $\mathbb{E}|h(Y)| < \infty$,
2. $\lim_{y \to \pm\infty} \left\{ h(y) \exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2 \right] \right\} = 0$,

then

$$\mathbb{E}[h(Y)(Y-\theta)] = \sigma^2 \mathbb{E}\left[h'(Y)\right].$$

Proof $(*)$.

By definition, $\mathbb{E}[h(Y)(Y-\theta)] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h(y)(y-\theta) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy$.
Integration by parts transforms the right hand side into

$$\underbrace{-\frac{\sigma^2}{\sigma\sqrt{2\pi}} \left(h(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} \right) \Big|_{-\infty}^{+\infty}}_{=0} + \underbrace{\frac{\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h'(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy}_{=\sigma^2 \mathbb{E}[h'(Y)]}$$

□

Proof of the James-Stein Theorem.

$$\begin{aligned}
\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) &= \mathbb{E} \left\| \left(1 - \frac{a}{\|\boldsymbol{Y}\|^2} \right) \boldsymbol{Y} - \boldsymbol{\mu} \right\|^2 = \mathbb{E} \left\| \boldsymbol{Y} - \boldsymbol{\mu} - \frac{a\,\boldsymbol{Y}}{\|\boldsymbol{Y}\|^2} \right\|^2 \\
&= \mathbb{E} \|\boldsymbol{Y} - \boldsymbol{\mu}\|^2 - 2\mathbb{E} \left(\frac{a\,\boldsymbol{Y}^\top (\boldsymbol{Y} - \boldsymbol{\mu})}{\|\boldsymbol{Y}\|^2} \right) + \mathbb{E} \left[\frac{a^2 \|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4} \right] \\
&= n - 2a \sum_{i=1}^{n} \mathbb{E} \left[\frac{Y_i (Y_i - \mu_i)}{\sum_{j=1}^{n} Y_j^2} \right] + a^2 \mathbb{E} \left[\frac{1}{\|\boldsymbol{Y}\|^2} \right]
\end{aligned}$$

Now define n differentiable functions $h_i : \mathbb{R} \to \mathbb{R}$ by

$$h_i(y) = \frac{y}{y^2 + \sum_{j \neq i}^{n} y_j^2}$$

and observe that, for all $i \in \{1, ..., n\}$,

$$\lim_{y_i \to \pm\infty} \left\{ h_i(y_i) \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu_i)^2 \right] \right\} = 0$$

We now use the tower property and apply our lemma to obtain

$$\mathbb{E}\left[\frac{Y_i(Y_i-\mu_i)}{\sum_{j=1}^n Y_j^2}\right] = \mathbb{E}\left[h_i(Y_i)(Y_i-\mu_i)\right] = \mathbb{E}\Big\{\mathbb{E}\left[h_i(Y_i)(Y_i-\mu_i)|\{Y_j\}_{j\neq i}\right]\Big\}$$
$$= \mathbb{E}\Big\{\mathbb{E}\left[h_i'(Y_i)|\{Y_j\}_{j\neq i}\right]\Big\} = \mathbb{E}\left[h_i'(Y_i)\right] = \mathbb{E}\left[\frac{\sum_{j=1}^n Y_j^2 - 2Y_i^2}{\left(\sum_{j=1}^n Y_j^2\right)^2}\right]$$

It follows that the MSE can be written as

$$\begin{aligned}\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) &= n - 2a\mathbb{E}\left[\frac{n\|\boldsymbol{Y}\|^2 - 2\|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4}\right] + a^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right] \\ &= n + [a^2 - 2a(n-2)]\underbrace{\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right]}_{>0}.\end{aligned}$$

Now, the polynomial $p(a) = a^2 - 2a(n-2)$ is strictly negative in the range $(0, 2n-4)$. Therefore, we have proven part (1). Furthermore, on the same range, $p(a)$ has a unique minimum at $a = n-2$, which proves part (3). For part (2), note that if $\boldsymbol{\mu} = 0$, $\|\boldsymbol{Y}\|^2 \sim \chi^2_n$, so $\mathbb{E}[1/\|\boldsymbol{Y}\|^2] = 1/(n-2)$ (recall that $n \geq 3$). Consequently, $\mathrm{MSE}(\delta_{n-2}, 0) = 2$.

□

Much of our discussion can be extended to cases where the MSE is replaced by some other convex measure of performance.

One can formulate a general framework as follows:

- Replace $\|\hat{\theta} - \theta\|$ by different deviation measure $\mathcal{L}(\hat{\theta}, \theta)$ called a loss function.
- The expected loss is then called the risk,

$$R(\hat{\theta}, \theta) = \mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)].$$

- The choice of loss function can be crucial and must be made judiciously.

Example (Exponential Distribution)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \text{Exponential}(\lambda)$, $n \geq 2$. The MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{Y}}$$

with $\bar{Y}$ the empirical mean. We can easily calculate

$$\mathbb{E}[\hat{\lambda}] = \frac{n\lambda}{n-1}.$$

It follows that $\tilde{\lambda} = (n-1)\hat{\lambda}/n$ is an unbiased estimator of λ. Observe now that

$$\text{MSE}(\tilde{\lambda}) < \text{MSE}(\hat{\lambda})$$

since $\tilde{\lambda}$ is unbiased and $\text{var}(\tilde{\lambda}) < \text{var}(\hat{\lambda})$. Hence the $\hat{\lambda}$ is strictly dominated by $\tilde{\lambda}$.

Observe that the parameter space here is $(0, \infty)$:

- In such cases, quadratic loss penalises over-estimation more heavily than under-estimation
- The maximum possible under-estimation is bounded!
- What happens if we change the loss function to account for that?

Example (Exponential Distribution, continued)

Consider a different loss function

$$\mathcal{L}(a, b) = a/b - 1 - \log(a/b)$$

where, for each fixed a, $lim_{b\to 0}\mathcal{L}(a, b) = lim_{b\to\infty}\mathcal{L}(a, b) = \infty$.
Now, for $n > 1$,

$$\begin{aligned} R(\lambda, \tilde{\lambda}) &= \mathbb{E}_\lambda\left[\frac{n\lambda\bar{Y}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{Y}}{n-1}\right)\right] \\ &= \underbrace{\mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right]}_{R(\lambda,\hat{\lambda})} + \underbrace{\frac{\mathbb{E}_\lambda(\lambda\bar{Y})}{n-1} - \log\left(\frac{n}{n-1}\right)}_{g(n)} \end{aligned}$$

where we wrote $\bar{Y} = \frac{n-1}{n}\bar{Y} + \frac{1}{n}\bar{Y}$. Note that $\mathbb{E}_\lambda[\bar{Y}] = \lambda^{-1}$, so

$$g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).$$

We claim that $g(n) > 0$ for $n \geq 2$.

Example (Exponential Distribution, continued)

Using $\log x = \int_1^x t^{-1}\,dt$, this follows if

$$\begin{aligned} \frac{1}{x} &> \log(x+1) - \log x, \qquad x > 1 \\ \iff \frac{1}{x} &> \int_x^{x+1} t^{-1}\,dt, \qquad x > 1 \end{aligned}$$

which holds by a rectangle area bound on the integral, as follows:

$$\frac{1}{x} = [(x+1) - x]\frac{1}{x} = \int_x^{x+1} \frac{1}{x}\,dt > \int_x^{x+1} \frac{1}{t}\,dt, \quad \text{when } x > 1$$

Consequently, $R(\tilde{\lambda}, \lambda) > R(\hat{\lambda}, \lambda)$ and $\hat{\lambda}$ dominates $\hat{\lambda}$.

We can push generality even further, and obtain an all encompassing framework.

Called decision theory, it views inference as a game between nature and the statistician.

Recall our general framework for statistical inference:

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The decision theory framework formalises step (4) to include estimation, testing, and confidence intervals.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.
- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.
- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves available to the statistician*.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents *how much the statistician has to pay* nature when losing.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{Y}^n \to \mathcal{A}$. These represent the *possible strategies* available to the statistician.

Choice of $\mathcal{A}$ determines what inference we are making. Choice of $\mathcal{D}$ determines what class of procedures we are willing to entertain. Choice of $\mathcal{L}$ determines how we measure our errors.

The statistician would like to pick strategy δ so as to limit his losses. But the losses are random, which is why risk comes into play.

Given a decision rule $\delta : \mathcal{Y}^n \to \mathcal{A}$, the risk is $R(\delta, \theta) = \mathbb{E}\left[\mathcal{L}(\delta(\boldsymbol{Y}), \theta)\right]$.

The key principle of decision theory is that

decision rules should be compared by comparing their risk functions

- Risk varies depending on true state of nature, though.
- So comparisons can be made in different ways:
  1. Uniform (hard). Seek dominance everywhere in Θ.
  2. Minimax (relaxed). Compare worst-case risks over Θ.
  3. Bayes (relaxed). Compare average risk over Θ

Will not go into details, but will give two definitions for educational purposes.

Rather than look at risk at every θ minimax risk concentrates on maximum risk

Definition (Minimax Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If

$$\sup_{\theta\in\Theta} R(\theta,\delta) \leq \sup_{\theta\in\Theta} R(\theta,\delta'), \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a minimax decision rule.

Rather than look at risk at every θ Bayes risk concentrates on average risk

Definition (Bayes Risk)

Let $\pi(\theta)$ be a probability density (frequency) on Θ and let δ be a decision rule for the experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. The π-Bayes risk of δ is defined as

$$r(\pi,\delta) = \int_\Theta R(\theta,\delta)\pi(\theta)\,d\theta = \int_\Theta \int_{\mathcal{X}} \mathcal{L}(\theta,\delta(\boldsymbol{y}))f_\theta(\boldsymbol{y})\,d\boldsymbol{y}\pi(\theta)\,d\theta$$

If $\delta \in \mathcal{D}$ is such that $r(\pi,\delta) \leq r(\pi,\delta')$ for all $\delta' \in \mathcal{D}$, then δ is called a *Bayes decision rule* with respect to π.

The prior $\pi(\theta)$ places different emphasis for different values of θ based on our prior interest/knowedge.

Comments:

- Minimax rules are useful to establish the fundamental inferential complexity or a statistical experiment.
- But using them for more practical purposes requires caution.
- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Makes sense if you are in a zero-sum adversarial game: if your opponent chooses θ to maximize $\mathcal{L}$ then one should look for minimax rules.
- If there is no reason to believe that "nature" is trying to "do her worst", then the minimax approach is overly conservative: it places emphasis on the "bad θ".
- Bayes rules are quite attractive as they can nearly never be uniformly dominated.
- Intuitively, if you can show your rule to be Bayes for a nice prior, you know you're doing reasonably well.

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The first sort of assertion we wish to make is:

1. **Hypothesis Testing**. Given two disjoint regions Θ_0 and Θ_1, which is more plausible to contain the true θ that generated our observation $(Y_1, \ldots, Y_n)^\top$?

The context:

1. We know that the true parameter lies in one of two subsets: Θ_0 or Θ_1, with $\Theta_0 \cap \Theta_1 = \emptyset$.
2. We need to use the sample $(Y_1, .., Y_n)^\top$ at hand to decide between the two possibilities.
3. This situation presents itself often in science, where two concurrent theories need to be confronted with the empirical evidence.
    1. The null hypothesis H_0 which states that $\theta \in \Theta_0$,

$$H_0 : \theta \in \Theta_0,$$

    and

    2. The alternative hypothesis that postulates $\theta \in \Theta_1$,

$$H_1 : \theta \in \Theta_1.$$

Example (Searching for the Higgs)

- One of the biggest questions of the last quarter century in physics: whether the infamous *Higgs boson* existed or not.
- Using the standard model of particle physics, we can calculate how many diphotons would be produced on average in the absence of Higgs' boson. Call this number $b > 0$.
- Similarly, we can calculate the additional mean number of diphotons produce if the Higgs boson existed. Call this number $s > 0$.
- Diphoton events are well-accounted to be Poissonian with mean (say) μ.

Our null hypothesis (no Higgs) is then

$$H_0 : \mu = b,$$

and the competing alternative is

$$H_1 : \mu = b + s.$$

□

Our decision must be based on the sample, so we need to define:

Definition (Test Function)

A test function is a map $\delta : \mathcal{Y}^n \to \{0, 1\}$.

Obtaining 0 or 1 must be decided on whether or not the sample satisfies a certain condition:

$$\delta(Y_1, \ldots, Y_n) = \begin{cases} 1, & \text{if } T(Y_1, \ldots, Y_n) \in C, \\ 0, & \text{if } T(Y_1, \ldots, Y_n) \notin C, \end{cases}$$

where

- T is a statistic called a *test statistic* and
- C is a subset of the range of T, called *critical region*.

In compact form

$$\delta(Y_1, \ldots, Y_n) = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C\}.$$

- To choose good test functions we need to quantify the performance of a test function.

Remark that, obviously, δ is just a Bernoulli random variable:

$$\delta = \begin{cases} 1, & \text{with probability } \mathbb{P}[T(Y_1, \ldots, Y_n) \in C], \\ 0, & \text{with probability } \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C]. \end{cases}$$

- So a good test function must have a sampling distribution concentrated around the right decision.
- The difference from point estimation is that our action space is discrete.
- Can we get an analogue of mean squared error?

Possible errors[5] to be made?

Action / Truth	H_0	H_1
0		Type II Error
1	Type I Error	

By an abuse of terminology, we could define:

$$\mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2], \qquad i \in \{0, 1\}.$$

Since δ is Bernoulli, and i takes values in $\{0, 1\}$, we have

$$\begin{aligned} \mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2] = \mathbb{E}_\theta[|\delta - i|] &= \begin{cases} \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_1. \end{cases} \\ &= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_1. \end{cases} \\ &= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_0, \\ \mathbb{P}_\theta[\delta = 0], & \text{if } \theta \in \Theta_1. \end{cases} \end{aligned}$$

[5] Potential asymmetry in practice: false positive VS false negative. Will return to this.

In decision theory terms, the action space is $\mathcal{A} = \{0, 1\}$ and the loss function is the so-called "0–1" loss,

$$\mathcal{L}(a, \theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 \quad \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 \quad \text{(Type II Error)} \\ 0 & \text{otherwise} \quad \text{(No Error)} \end{cases}$$

i.e. we lose 1 unit whenever committing a type I or type II error.

The risk function then becomes

$$R(\delta, \theta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 \quad \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 \quad \text{(prob of type II error)} \end{cases}$$

In short,

$$R(\delta, \theta) = \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\}$$

Can we hope to simultaneously control both type I and II error probabilities? ↪
Unfortunately the answer is no.

Here's why let $\delta(Y_1, \ldots, Y_n) = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C\}$ and suppose we wish to reduce the type I error probability

$$\mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0,$$

for al $\theta \in \Theta_0$.

To do this, we must replace C by a subset $C_* \subset C$, obtaining

$$\delta_* = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C_*\}.$$

Observe that, $\forall\, \theta \in \Theta_0$,

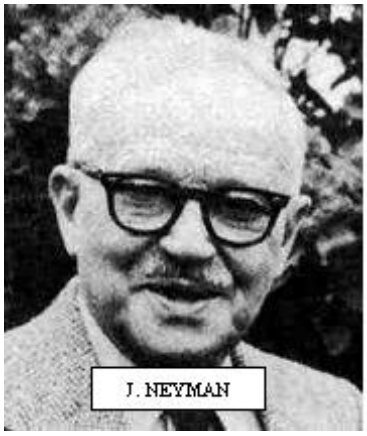
$$\mathbb{P}_\theta[\delta_* = 1] = \mathbb{P}[T(Y_1, \ldots, Y_n) \in C_*] \leq \mathbb{P}[T(Y_1, \ldots, Y_n) \in C] = \mathbb{P}_\theta[\delta = 1]$$

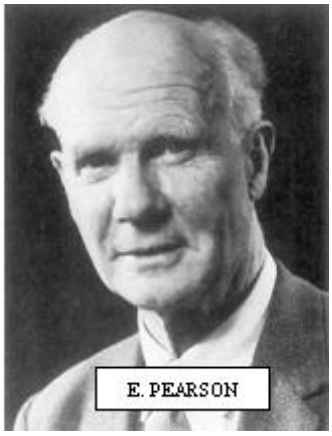
On the other hand $C_* \subset C \implies C_*^c \supset C^c$ and so $\forall\, \theta \in \Theta_1$

$$\mathbb{P}_\theta[\delta_* = 0] = \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C_*] \geq \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C] = \mathbb{P}_\theta[\delta = 0].$$

By reducing the type I error probability we increased the type II error probability

We need to make some concessions...





The fundamental paradigm of *Neyman and Pearson* informally dictates:

1. In applications, one type of error (false positive or negative) is typically more severe.
2. Say this is the type I error, and exploit the asymmetry: fix a tolerance ceiling for the probability of this error.
3. Given this ceiling, consider only test functions that respect it, and focus on minimising type II error (i.e. maximising power).

In mathematical terms:

The Neyman-Pearson Framework

1. We fix an $\alpha \in (0,1)$, usually small (called the significance level)
2. We declare that we only consider test functions $\delta : \mathcal{X} \to \{0,1\}$ such that
$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \left\{\delta : \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha\right\}$$
i.e. rules for which prob of type I error is bounded above by α

↪ *Jargon: we fix a significance level for our test*

3. Within this restricted class of rules, choose δ to minimize prob of type II error:
$$\mathbb{P}_\theta[\delta(\boldsymbol{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1]$$
4. Equivalently, maximize the *power*
$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1] = \mathbb{E}_\theta[\mathbf{1}\{\delta(\boldsymbol{X}) = 1\}] = \mathbb{E}_\theta[\delta(\boldsymbol{X})], \quad \theta \in \Theta_1$$

(since $\delta = 1 \iff \mathbf{1}\{\delta = 1\} = 1$ and $\delta = 0 \iff \mathbf{1}\{\delta = 1\} = 0$)

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Example: Hillary VS Trump 2016. Pollsters gather iid sample $\boldsymbol{Y}$ from Florida with $Y_i = \mathbf{1}\{\text{vote Trump}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: \text{ Trump wins Florida} \\ H_1: \text{ Hillary wins Florida} \end{cases} \quad \text{OR} \quad \begin{cases} H_0: \text{ Hillary wins Florida} \\ H_1: \text{ Trump wins Florida} \end{cases}$$

- Which pair to choose to make a prediction? (confidence intervals?)
- If Trump is conducting poll to decide whether he'll spend more money to campaign in Florida, then his possible losses due to errors are:
  (a) Spend more \$'s to campaign in Florida even though he would win anyway: lose \$'s
  (b) Lose Florida to Hillary because he thought he would win without any extra effort.
- (b) is much worse than (a) (especially since Trump had lots of \$'s)
- Hence Trump would pick $H_0 = \{\text{Hillary wins Florida}\}$ as his null

Finding Good Test Functions

Consider simplest situation:

$$\Theta_0 = \{\theta_0\} \qquad \& \qquad \Theta_1 = \{\theta_1\}$$

The Neyman-Pearson Lemma - Continuous Case

Let $\boldsymbol{Y}$ have joint density/frequency $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1.$$

If $\Lambda(\boldsymbol{Y}) = f_1(\boldsymbol{Y})/f_0(\boldsymbol{Y})$ is a continuous random variable, then there exists a $k > 0$ such that

$$\mathbb{P}_0[\Lambda(\boldsymbol{Y}) \geq k] = \alpha$$

and the test whose test function is given by

$$\delta(\boldsymbol{Y}) = \mathbf{1}\{\Lambda(\boldsymbol{Y}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1. Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$. By assumption, G_0 is a differentiable distribution function, and so is onto $[0,1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1-\alpha\}$ is non-empty for any $\alpha \in (0,1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$ we will have $\mathbb{P}_0[\Lambda \geq k] = \alpha$ and k is simply the $1-\alpha$ quantile of the distribution G_0. Consequently,

$$\mathbb{P}_0[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \geq k])$$

and therefore $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (i.e. δ indeed respects the level α).
To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\boldsymbol{y}) \in \{0,1\}$, then

$$\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \underbrace{\mathbb{E}_0[\delta(\boldsymbol{Y})]}_{=\alpha \text{ (by first part of proof)}} \implies \underbrace{\mathbb{E}_1[\psi(\boldsymbol{Y})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\boldsymbol{Y})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).

WLOG assume that f_0 and f_1 are density functions. Note that

$$f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) \geq 0 \text{ if } \delta(\boldsymbol{y}) = 1 \quad \& \quad f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}) < 0 \text{ if } \delta(\boldsymbol{y}) = 0.$$

Therefore, since ψ can only take the values 0 or 1,

$$\begin{aligned} \psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) &\leq \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y})) \\ \int_{\mathbb{R}^n} \psi(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}))\, d\boldsymbol{y} &\leq \int_{\mathbb{R}^n} \delta(\boldsymbol{y})(f_1(\boldsymbol{y}) - k \cdot f_0(\boldsymbol{y}))\, d\boldsymbol{y} \end{aligned}$$

Rearranging the terms yields

$$\begin{aligned} \int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_1(\boldsymbol{y})\, d\boldsymbol{y} &\leq k \int_{\mathbb{R}^n} (\psi(\boldsymbol{y}) - \delta(\boldsymbol{y})) f_0(\boldsymbol{y})\, d\boldsymbol{y} \\ \implies \mathbb{E}_1[\psi(\boldsymbol{Y})] - \mathbb{E}_1[\delta(\boldsymbol{Y})] &\leq k\,(\mathbb{E}_0[\psi(\boldsymbol{y})] - \mathbb{E}_0[\delta(\boldsymbol{Y})]) \end{aligned}$$

But $k > 0$ by assumption, so when $\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \mathbb{E}_0[\delta(\boldsymbol{Y})]$ the RHS is negative, i.e. δ is an MP test of H_0 vs H_1 at level α. □

- Basically we reject if the likelihood of θ_0 is k times higher than the likelihood of θ_1. This is called a likelihood ratio test, and Λ is the likelihood ratio statistic: *how much more plausible is the alternative than the null?*
- When Λ is a continuous RV, the choice of k is essentially unique. That is, if k' is such that $\delta' = \mathbf{1}\{\Lambda \geq k'\} \in \mathcal{D}(\{\theta_0\}, \alpha)$, then $\delta = \delta'$ almost surely.
- The resulting most powerful test is not necessarily unique.
- Unless Λ is continuous, the most powerful test is not necessarily guaranteed to exist.
- The problem if Λ is a RV with a discontinuous dist is that there may exist no k for which the equation $\mathbb{P}_0[\Lambda \geq k] = \alpha$ has a solution.
- In any case, typically the distribution of the test statistic converges to a continuous limit with large n, so these problems become inessential.

Example (Poisson Distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\mu)$ and for $\mu_1 > \mu_0$ consider the hypotheses:

$$H_0 : \mu = \mu_0 \qquad vs \qquad H_1 : \mu = \mu_1.$$

These correspond to the Higgs example by setting $\mu_0 = b$ and $\mu_1 = b + s$.
Applying the Neyman-Pearson lemma gives a test statistic

$$\delta(Y_1, \ldots, Y_n) = \mathbf{1}\left\{\sum_{i=1}^n Y_i > q_{1-\alpha}\right\},$$

provided α is such that $G_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(Y_1, \ldots, Y_n) \leq q_{1-\alpha}] \overset{!}{=} 1 - \alpha$.
Since the Y_i are independent, one can easily show that

$$\tau(Y_1, \ldots, Y_n) \overset{H_0}{\sim} Poisson(n\mu_0).$$

This being a discrete distribution, the only α for which we get an MP test are

$$e^{-n\mu_0}, e^{-n\mu_0}(1 + n\mu_0), e^{-n\mu_0}\left(1 + n\mu_0 + \tfrac{(n\mu_0)^2}{2}\right), \ldots \text{and so on}$$

Nevertheless notice that as $n \to \infty$, these values become dense near the origin.

When $\{\Theta_0, \Theta_1\}$ are not singletons, choosing a most powerful test is a much stronger requirement:

1. It should respect the level for all $\theta \in \Theta_0$, i.e.

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \mathcal{Y}^n \to \{0,1\} : \mathbb{E}_\theta[\delta] \leq \alpha, \ \forall\, \theta \in \Theta_0\}$$

2. It should be most powerful for all $\theta \in \Theta_1$ (i.e. for all possible simple alternatives),

$$\mathbb{E}_\theta[\delta] \geq \mathbb{E}_\theta[\delta'] \qquad \forall \theta \in \Theta_1 \quad \& \quad \delta' \in \mathcal{D}(\Theta_0, \alpha)$$

Unfortunately UMP tests rarely exist. Why?
↪ Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

- A UMP test must be MP test for any $\theta \neq \theta_1$.
- But the form of the MP test typically differs for $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$!
  ↪ e.g. recall exponential mean example

Example (No UMP test exists)

Let $Y_1, ..., Y_n \sim$Bernoulli(θ) and suppose we want to test:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

at some level α. To this aim, consider first

$$H_0' : \theta = \theta_0 \qquad vs \qquad H_1' : \theta = \theta_1$$

Neyman-Pearson lemma gives test statistics

$$T = \frac{f(\boldsymbol{Y}; \theta_1)}{f(\boldsymbol{Y}; \theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{i=1}^n Y_i}$$

- If $\theta_1 > \theta_0$ then T increasing in $\sum_{i=1}^n Y_i$
  ↪ MP test would reject for large values of $\sum_{i=1}^n Y_i$
- If $\theta_1 < \theta_0$ then T decreasing in $\sum_{i=1}^n Y_i$
  ↪ MP test would reject for small values of $\sum_{i=1}^n Y_i$

So what can we do for more general $\{\Theta_0, \Theta_1\}$?

- One sided hypotheses: when Θ_0 is a an interval of the form $(-\infty, \theta_0]$ or $[\theta_0, +\infty)$ and $\Theta_1 = \Theta_0^c$, there are often uniformly most powerful tests depending on the underlying model.
  - For example, in one-parameter exponential families, one simply uses the Neyman-Pearson lemma, taking the null to be $\theta = \theta_0$ and the alternative $\theta = \theta_1$ for any $\theta_1 \in \Theta_1$ (the form of the test depends only on the direction of the null and the boundary of the null).
  - This generalises to families admitting a so-called "monotone likelihood ratio"
  - In the absence of the "monotone likelihood ratio" property, one can seek locally most powerfull tests, near the hypothesis boundary. It can be shown that the score function (derivative of the logliklihood) at the boundary θ_0 can serve as a test statistic to this aim.
- General hypothesis pairs: we need to abandon optimality, and search for sensible tests. But the likelihood ratio idea can serve us well in this pursuit.

Likelihood Ratio Tests

Consider now the multiparameter case $\boldsymbol{\theta} \in \mathbb{R}^p$ with general Θ_0, Θ_1

- As noted optimality breaks down.
- But we can still seek general-purpose approaches.

The idea: Combine Neyman-Pearson paradigm with Max Likelihood

Definition (Likelihood Ratio)

The *likelihood ratio statistic* corresponding to the pair of hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$ is defined to be

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta} f(\boldsymbol{Y}; \theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} f(\boldsymbol{Y}; \theta)} = \frac{\sup_{\boldsymbol{\theta} \in \Theta} L(\theta)}{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\theta)}$$

- Intuition: choose the "most favourable" $\theta \in \Theta_0$ (in favour of H_0) and compare it against the "most favourable" $\theta \in \Theta_1$ (in favour of H_1) in a simple vs simple setting (applying NP-lemma)
- Typically Θ_0 is a lower dimensional subspace of Θ_1, so taking sup over Θ (rather than Θ_1) incurs no loss. In this case $\Theta_0 \cap \Theta_1 \neq \emptyset$, but $Leb(\Theta_0 \cap \Theta_1) = 0$, which suffices.

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0$$

$$\Lambda(\boldsymbol{Y}) = \frac{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)}{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}^+} f(\boldsymbol{Y};\mu,\sigma^2)} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}} = \left(\frac{\sum_{i=1}^n (Y_i-\mu_0)^2}{\sum_{i=1}^n (Y_i-\bar{Y})^2}\right)^{\frac{n}{2}}$$

So reject when $\Lambda \geq k$, where k is s.t. $\mathbb{P}_0[\Lambda \geq k] = \alpha$. Distribution of Λ? By monotonicity look only at

$$\begin{aligned}\frac{\sum_{i=1}^n (Y_i-\mu_0)^2}{\sum_{i=1}^n (Y_i-\bar{Y})^2} &= 1 + \frac{n(\bar{Y}-\mu_0)^2}{\sum_{i=1}^n (Y_i-\bar{Y})^2} = 1 + \frac{1}{n-1}\left(\frac{n(\bar{Y}-\mu_0)^2}{S^2}\right) \\ &= 1 + \frac{T^2}{n-1}\end{aligned}$$

With $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i-\bar{Y})^2$ and $T = \sqrt{n}(\bar{Y}-\mu_0)/S \overset{H_0}{\sim} t_{n-1}$.

So $T^2 \overset{H_0}{\sim} F_{1,n-1}$ and k may be chosen appropriately.

Example

Let $Y_1, ..., Y_m \overset{iid}{\sim} \text{Exp}(\lambda)$ and $Z_1, ..., Z_n \overset{iid}{\sim} \text{Exp}(\theta)$. Assume $\boldsymbol{Y}$ indep $\boldsymbol{Z}$.

$$\text{Consider:} \quad H_0 : \theta = \lambda \qquad \text{vs} \qquad H_1 : \theta \neq \lambda$$

$$\underset{\sup_{(\lambda,\theta)\in\mathbb{R}^2_+} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Unrestricted MLEs:}} \quad \hat{\lambda} = 1/\bar{Y} \qquad \& \qquad \hat{\theta} = 1/\bar{Z}$$

$$\underset{\sup_{(\lambda,\theta)\in\{(y,z)\in\mathbb{R}^2_+ : y=z\}} f(\boldsymbol{Y},\boldsymbol{Z};\lambda,\theta)}{\text{Restricted MLEs:}} \quad \hat{\lambda}_0 = \hat{\theta}_0 = \left[\frac{m\bar{Y}+n\bar{Z}}{m+n}\right]^{-1}$$

$$\implies \Lambda = \left(\frac{m}{m+n} + \frac{n}{n+m}\frac{\bar{Z}}{\bar{Y}}\right)^m \left(\frac{n}{n+m} + \frac{m}{m+n}\frac{\bar{Y}}{\bar{Z}}\right)^n$$

Depends on $T = \bar{Y}/\bar{Z}$ and can make Λ large/small by varying T.

$\hookrightarrow$ But $T \overset{H_0}{\sim} F_{2m,2n}$ so given α we may find the critical value k.

More often than not, dist(Λ) intractable
$\hookrightarrow$(and no simple dependence on T with tractable distribution either)

Consider asymptotic approximations?

Setup

- Θ open subset of $\mathbb{R}^p$
- either $\Theta_0 = \{\boldsymbol{\theta}_0\}$ or Θ_0 open subset of $\mathbb{R}^s$, where $s < p$
- Concentrate on $\boldsymbol{Y} = (Y_1, ..., Y_n)$ has iid components.
- Initially restrict attention to $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. LR becomes:

$$\Lambda_n(\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{f(Y_i; \hat{\boldsymbol{\theta}}_n)}{f(Y_i; \boldsymbol{\theta}_0)}$$

  where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$.
- Impose regularity conditions from MLE asymptotics

Theorem (Wilks' Theorem, case $p = 1$)

Let $Y_1, ..., Y_n$ be iid random variables with density (frequency) depending on $\theta \in \mathbb{R}$ and satisfying conditions (A1)-(A6), with $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$. If the MLE sequence $\hat{\theta}_n$ is consistent for θ, then the likelihood ratio statistic Λ_n for $H_0 : \theta = \theta_0$ satisfies

$$2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_1$$

when H_0 is true.

- Obviously, knowing approximate distribution of $2 \log \Lambda_n$ is as good as knowing approximate distribution of Λ_n for the purposes of testing (by monotonicity and rejection method).
- Theorem extends immediately and trivially to the case of general p and for a hypothesis pair $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.
  (i.e. when null hypothesis is simple)

Proof (∗).

Under the conditions of the theorem and when H_0 is true,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta_0))$$

Now take logarithms and expand in a Taylor series around $\hat{\theta}_n$,

$$\begin{aligned}\log \Lambda_n &= \sum_{i=1}^{n}[\ell(Y_i;\hat{\theta}_n) - \ell(Y_i;\theta_0)] = \sum_{i=1}^{n}[\ell(Y_i;\hat{\theta}_n) - \ell(Y_i;\hat{\theta}_n)] + \\ &\quad +(\theta_0 - \hat{\theta}_n)\sum_{i=1}^{n}\ell'(Y_i;\hat{\theta}_n) - \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\sum_{i=1}^{n}\ell''(Y_i;\theta_n^*) \\ &= -\frac{1}{2}n(\hat{\theta}_n - \theta_0)^2\frac{1}{n}\sum_{i=1}^{n}\ell''(Y_i;\theta_n^*)\end{aligned}$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0.

If H_0 is true, and since $\hat{\theta}_n$ is a consistent sequence, θ_n^* is sandwiched so

$$\theta_n^* \xrightarrow{p} \theta_0.$$

Hence under assumptions (A1)-(A6), and when H_0 is true, a first order Taylor expansion about θ_0, the continuous mapping theorem and the LLN give

$$\frac{1}{n}\sum_{i=1}^{n}\ell''(Y_i;\theta_n^*) \xrightarrow{p} -\mathbb{E}_{\theta_0}[\ell''(Y_i;\theta_0)] = \mathcal{I}_1(\theta_0)$$

On the other hand, by the continuous mapping theorem,

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{V}{\mathcal{I}_1(\theta_0)}$$

Applying Slutsky's theorem now yields the result. □ □

Theorem (Wilk's theorem, general p, general $s \leq p$)

Let $Y_1, ..., Y_n$ be iid random variables with density (frequency) depending on $\boldsymbol{\theta} \in \mathbb{R}^p$ and satisfying conditions (B1)-(B6), with $\mathcal{I}_1(\boldsymbol{\theta}) = \mathcal{J}_1(\boldsymbol{\theta})$. If the MLE sequence $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$, then the likelihood ratio statistic Λ_n for $H_0 : \{\theta_j = \theta_{j,0}\}_{j=1}^s$ satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_s$ when H_0 is true.

Comments:

- Note that it may potentially be that $s < p$, and this is accommodated by the theore,
- Hypotheses of the form $H_0 : \{g_j(\boldsymbol{\theta}) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:
  - Define $(\phi_1, ..., \phi_p) = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), ..., g_p(\boldsymbol{\theta}))$
  - $g_{s+1}, ..., g_p$ defined so that $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta})$ is 1-1
  - Apply theorem with parameter $\boldsymbol{\phi}$

Many other tests possible. For example:

- Wald's test
  - ↪ For a simple null, may compare the unrestricted MLE with the MLE under the null. Large deviations indicate evidence against null hypothesis. Distributions are approximated for large n via the asymptotic normality of MLEs.
- Score Test
  - ↪ For a simple null, if the null hypothesis is false, then the loglikelihood gradient at the null should not be close to zero, at least when n reasonably large: so measure its deviations form zero. Use asymptotics for distributions (under conditions we end up with a χ^2)
- ...

So far focussed on Neyman-Pearson Framework:

1. Fix a significance level α for the test
2. Consider rules δ respecting this significance level
   ↪ We choose one of those rules, δ^*, based on power considerations
3. We reject at level α if $\delta^*(\boldsymbol{y}) = 1$.

Useful for attempting to determine optimal test statistics
What if we already have a given form of test statistic in mind? (e.g. LRT)
↪ A different perspective on testing (used more in practice) says:

Rather then consider a family of test functions respecting level α...
... consider family of test functions indexed by α

1. Fix a family $\{\delta_\alpha\}_{\alpha\in(0,1)}$ of decision rules, with δ_α having level α
   ↪ for a given $\boldsymbol{y}$ some of these rules reject the null, while others do not
2. Which is the smallest α for which H_0 is rejected given $\boldsymbol{y}$?

Definition (p–Value)

Let $\{\delta_\alpha\}_{\alpha\in(0,1)}$ be a family of test functions satisfying

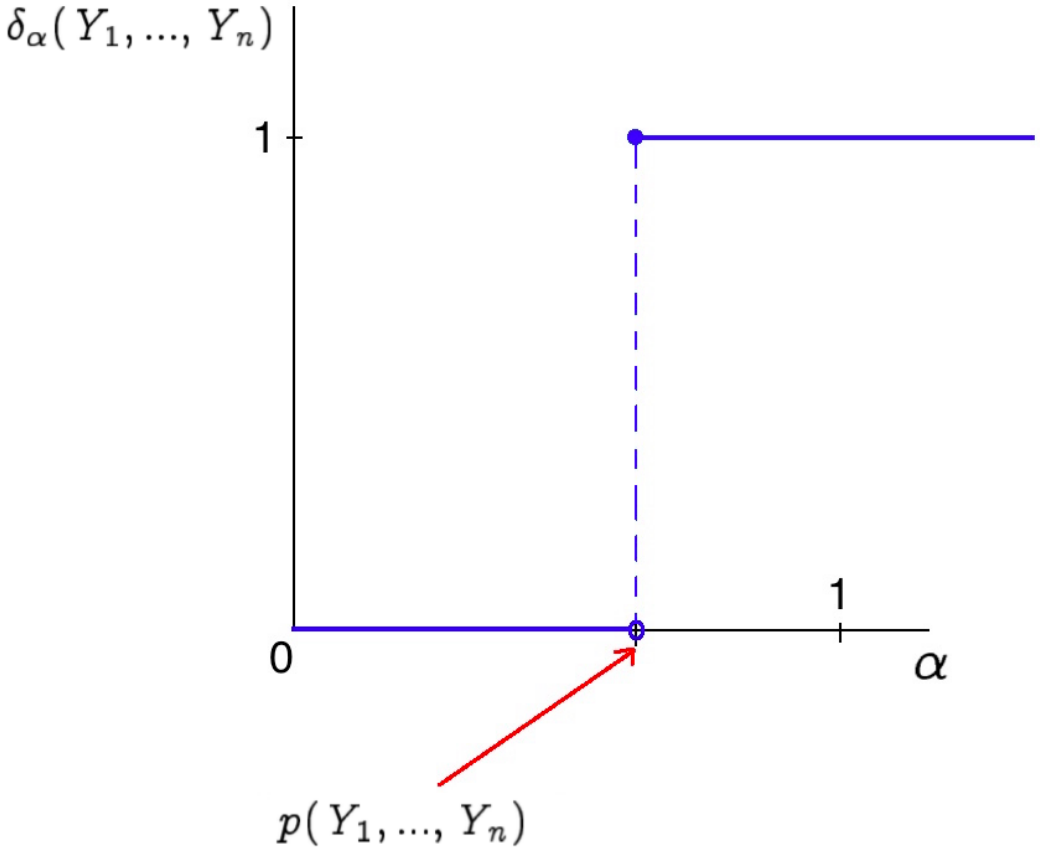
$$\alpha_1 < \alpha_2 \implies \{\boldsymbol{y} \in \mathcal{Y}^n : \delta_{\alpha_1}(\boldsymbol{y}) = 1\} \subseteq \{\boldsymbol{y} \in \mathcal{Y}^n : \delta_{\alpha_2}(\boldsymbol{y}) = 1\}.$$

The p–value (or observed significance level) of the family $\{\delta_\alpha\}$ is

$$p(\boldsymbol{y}) = \inf\{\alpha : \delta_\alpha(\boldsymbol{y}) = 1\}$$

↪ The p–value is the smallest value of α for which the null would be rejected at level α, given $\boldsymbol{Y} = \boldsymbol{y}$.
Most usual setup:

- Have a single test statistic T
- Construct family $\delta_\alpha(\boldsymbol{y}) = \mathbf{1}\{T(\boldsymbol{y}) > k_\alpha\}$
- If $\mathbb{P}_{H_0}[T \leq t] = G(t)$ then $p(\boldsymbol{y}) = \mathbb{P}_{H_0}[T(\boldsymbol{Y}) \geq T(\boldsymbol{y})] = 1 - G(T(\boldsymbol{y}))$



Notice: contrary to NP-framework did not make explicit decision!

- We simply reported a p–value
- The p–value is used as a measure of evidence against H_0
  - ↪ Small p–value provides evidence against H_0
  - ↪ Large p–value provides no evidence against H_0
- How small does "small" mean?
  - ↪ Depends on the specific problem...

Intuition:

- Recall that extreme values of test statistics are those that are "inconsistent" with null (NP-framework)
- p–value is probability of observing a value of the test statistic as extreme as or more extreme than the one we observed, under the null
- If this probability is small, then we have witnessed something quite unusual under the null hypothesis
- Gives evidence against the null hypothesis

Example (Normal Mean)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider:

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0$$

Likelihood ratio test: reject when T^2 large, $T = \sqrt{n}\,\bar{Y}/S \overset{H_0}{\sim} t_{n-1}$.

Since $T^2 \overset{H_0}{\sim} F_{1,n-1}$, p–value is

$$p(\boldsymbol{y}) = \mathbb{P}_{H_0}[\,T^2(\boldsymbol{Y}) \geq T^2(\boldsymbol{y})] = 1 - G_{F_{1,n-1}}(\,T^2(\boldsymbol{y}))$$

Consider two samples (datasets),

$$\boldsymbol{y} = (0.66, 0.28, -0.99, 0.007, -0.29, -1.88, -1.24, 0.94, 0.53, -1.2)$$

$$\boldsymbol{y}' = (1.4, 0.48, 2.86, 1.02, -1.38, 1.42, 2.11, 2.77, 1.02, 1.87)$$

Obtain $p(\boldsymbol{y}) = 0.32$ while $p(\boldsymbol{y}') = 0.006$.

- Reporting a p–value does not necessarily mean making a decision
- A small p–value can simply reflect our "confidence" in rejecting a null

Recall example: Statisticians working for Hillary gather iid sample $\boldsymbol{Y}$ from Florida with $Y_i = \mathbf{1}\{\text{vote Hillary}\}$. Hillary team want to test

$$\begin{cases} H_0 : \text{ Trump wins Florida} \\ H_1 : \text{ Hillary wins Florida} \end{cases}$$

- Will statisticians decide for Hillary?
- Perhaps better to report p–value to him and let him decide...

What if statisticians working for newspaper, not Hillary?

- Something easier to interpret than test/p–value?

A Glance Back at Point Estimation

- Let $Y_1, ..., Y_n$ be iid random variables with density (frequency) $f(\cdot; \theta)$.
- Problem with point estimation: $\mathbb{P}_\theta[\hat{\theta} = \theta]$ typically small (if not zero)
  - ↪ always attach an estimator of variability, e.g. standard error
  - ↪ interpretation?

- Hypothesis tests may provide way to interpret estimator's variability within the setup of a particular problem
  - ↪ e.g. if observe $\hat{P}[\text{Hillary wins}] = 0.52$ can actually see what p–value we get when testing $H_0 : P[\text{Hillary wins}] \geq 1/2$.
- Something more directly interpretable?

Back to our example: What do pollsters do in newspapers?
↪ They announce their point estimate (e.g. 0.52)
↪ They give upper and lower *confidence limits*
What are these and how are they interpreted?

Interval Estimation

Simple underlying idea:

- Instead of estimating θ by a single value
- Present a whole range of values for θ that are consistent with the data
  - ↪ In the sense that they could have produced the data

Definition (Confidence Interval)

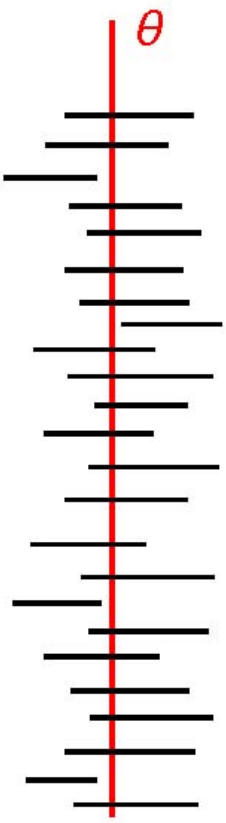
Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be random variables with joint distribution depending on $\theta \in \mathbb{R}$ and let $L(\boldsymbol{Y})$ and $U(\boldsymbol{Y})$ be two statistics with $L(\boldsymbol{Y}) < U(\boldsymbol{Y})$ a.s. Then, the random interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ is called a $100(1-\alpha)\%$ confidence interval for θ if

$$\mathbb{P}_\theta[L(\boldsymbol{Y}) \leq \theta \leq U(\boldsymbol{Y})] \geq 1 - \alpha$$

for all $\theta \in \Theta$, with equality for at least one value of θ.

- $1 - \alpha$ is called the coverage probability or confidence level
- Beware of interpretation!

- Probability statement is NOT made about θ, which is constant.
- Statement is about interval: probability that the interval contains the true value is at least $1-\alpha$.
- Given any realization $\boldsymbol{Y} = \boldsymbol{y}$, the interval $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ will either contain or not contain θ.
- Interpretation: if we construct intervals with this method, then we expect that $100(1-\alpha)\%$ of the time our intervals will engulf the true value.



Example (The example that says all)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Then $\sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0, 1)$, so that

$$\mathbb{P}_\mu[-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96] = 0.95$$

and since

$$-1.96 \leq \sqrt{n}(\bar{Y} - \mu) \leq 1.96 \iff \bar{Y} - 1.96/\sqrt{n} \leq \mu \leq \bar{Y} + 1.96/\sqrt{n}$$

we obviously have

$$\mathbb{P}_\mu \left[\bar{Y} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96}{\sqrt{n}} \right] = 0.95$$

So that the random interval $[L(\boldsymbol{Y}), U(\boldsymbol{Y})] = \left[\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}} \right]$ is a 95% confidence interval for μ.
Central Limit Theorem: same argument can yield approximate 95% CI when $Y_1, ..., Y_n$ are iid, $\mathbb{E}\, Y_i = \mu$ and $\text{var}(Y_i) = 1$, regardless of their distribution.

Example (continued)

Notice that the interval is centred at $\bar{Y}$, the MLE of μ. It's often thus written:

$$\boxed{\bar{Y} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

Observations:

- The length of the interval is $2z_{1-\alpha/2}\sigma/\sqrt{n}$, which depends on σ^2, n and α.
- The parameter σ^2 is beyond our control.
- We can nevertheless control n and $1-\alpha$. Increasing n, the length of the interval decays as $1/\sqrt{n}$.
- Reducing α (i.e. increasing $1-\alpha$) increases the length of the interval (the dependence is quite non-linear, and 5% is chosen as a "sweet spot").

Pivotal Quantities

What can we learn from previous example?

Definition (Pivot)

A random function $g(\boldsymbol{Y}, \theta)$ is said to be a *pivotal quantity* (or simply a *pivot*) if it is a function both of $\boldsymbol{Y}$ and θ whose distribution does not depend on θ.

$\hookrightarrow \sqrt{n}(\bar{Y} - \mu) \sim \mathcal{N}(0,1)$ is a pivot in previous example

Why is a pivot useful?

- $\forall\ \alpha \in (0,1)$ we can find constants $a < b$ independent of θ, such that

$$\mathbb{P}_\theta[a \leq g(\boldsymbol{Y}, \theta) \leq b] = 1 - \alpha \qquad \forall\ \theta \in \Theta$$

- If $g(\boldsymbol{Y}, \theta)$ can be manipulated then the above yields a CI

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. Recall that MLE $\hat{\theta}$ is $\hat{\theta} = Y_{(n)}$, with distribution

$$\mathbb{P}_\theta \left[Y_{(n)} \leq x \right] = F_{Y_{(n)}}(x) = \left(\frac{x}{\theta} \right)^n \implies \mathbb{P}_\theta \left[\frac{Y_{(n)}}{\theta} \leq y \right] = y^n$$

$\rightarrow$ Hence $Y_{(n)}/\theta$ is a pivot for θ. Can now choose $a < b$ such that

$$\mathbb{P}_\theta \left[a \leq \frac{Y_{(n)}}{\theta} \leq b \right] = 1 - \alpha$$

$\rightarrow$ But there are ∞-many such choices!
$\hookrightarrow$ Idea: choose pair (a, b) that minimizes interval's length!

Solution can be seen to be $a = \alpha^{1/n}$ and $b = 1$, yielding

$$\left[Y_{(n)}, \frac{Y_{(n)}}{\alpha^{1/n}} \right]$$

Pivotal method extends to construction of CI for θ_k, when

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_k, ..., \theta_p) \in \mathbb{R}^p$$

and the remaining coordinates are also unknown. $\rightarrow$ Pivotal quantity should now be function $g(\boldsymbol{Y}; \theta_k)$ which

1. Depends on $\boldsymbol{Y}$, θ_k, but no other parameters
2. Has a distribution independent of any of the parameters

$\hookrightarrow$ e.g.: CI for normal mean, when variance unknown

$\rightarrow$ Main difficulties with pivotal method:

- Hard to find exact pivots in general problems
- Exact distributions may be intractable

Resort to asymptotic approximations...
$\hookrightarrow$ Most classic example when have $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$.

What about higher dimensional parameters?

Definition (Confidence Region)

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be random variables with joint distribution depending on $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. A random subset $R(\boldsymbol{Y})$ of Θ depending on $\boldsymbol{Y}$ is called a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ if

$$\mathbb{P}_\theta[R(\boldsymbol{Y}) \ni \boldsymbol{\theta}] \geq 1 - \alpha$$

for all $\boldsymbol{\theta} \in \Theta$, with equality for at least one value of $\boldsymbol{\theta}$.

- No restriction requiring $R(\boldsymbol{Y})$ to be convex or even connected
  - ↪ So when $p = 1$ get more general notion than CI
- Nevertheless, many notions extend immediately to CR case
  - ↪ e.g. notion of a pivotal quantity

Let $g : \mathcal{Y}^n \times \Theta \to \mathbb{R}$ be a function such that dist$[g(\boldsymbol{Y}, \boldsymbol{\theta})]$ independent of $\boldsymbol{\theta}$
↪ Since image space is the real line, can find $a < b$ s.t.

$$\mathbb{P}_{\boldsymbol{\theta}}[a \leq g(\boldsymbol{Y}, \boldsymbol{\theta}) \leq b] = 1 - \alpha$$

$$\implies \mathbb{P}_{\boldsymbol{\theta}}[R(\boldsymbol{Y}) \ni \boldsymbol{\theta}] = 1 - \alpha$$

where $R(\boldsymbol{y}) = \{\boldsymbol{\theta} \in \Theta : g(\boldsymbol{y}, \boldsymbol{\theta}) \in [a, b])\}$

Notice that region can be "wild" since it is a random level set of g

Example

Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$. Two unbiased estimators of $\boldsymbol{\mu}$ and Σ are

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i \\ \hat{\Sigma} &= \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}})^T \end{aligned}$$

Example (cont'd)

Consider the random variable

$$g(\{\boldsymbol{Y}\}_{i=1}^n, \boldsymbol{\mu}) := \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \sim F\text{-dist with k and n-k d.f.}$$

A pivot!
$\hookrightarrow$ If f_q is q-quantile of this distribution, then get $100q\%$ CR as

$$R(\{\boldsymbol{Y}\}_{i=1}^n) = \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \frac{n(n-k)}{k(n-1)}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \leq f_q\right\}$$

- An ellipsoid in $\mathbb{R}^n$
- Ellipsoid centred at $\hat{\boldsymbol{\mu}}$
- Principle axis lengths given by eigenvalues of $\hat{\Sigma}^{-1}$
- Orientation given by eigenvectors of $\hat{\Sigma}^{-1}$

Duality

Visualisation of high-dimensional CR's can be hard

- When these are ellipsoids spectral decomposition helps
- But more generally?

Things especially easy when dealing with rectangles - but they rarely occur!
$\hookrightarrow$ What if we construct a CR as Cartesian product of CI's?

Let $[L_i(\boldsymbol{Y}), U_i(\boldsymbol{Y})]$ be $100q_i\%$ CI's for θ_i, $i = 1, .., p$, and define

$$R(\boldsymbol{Y}) = [L_1(\boldsymbol{Y}), U_1(\boldsymbol{Y})] \times \ldots \times [L_p(\boldsymbol{Y}), U_p(\boldsymbol{Y})]$$

Bonferroni's inequality implies that

$$\mathbb{P}_{\boldsymbol{\theta}}[R(\boldsymbol{Y}) \ni \boldsymbol{\theta}] \geq 1 - \sum_{i=1}^p \mathbb{P}[\theta_i \notin [L_i(\boldsymbol{Y}), U_i(\boldsymbol{Y})]] = 1 - \sum_{i=1}^p (1 - q_i)$$

$\rightarrow$ So pick q_i such that $\sum_{i=1}^p (1 - q_i) = \alpha$ (can be conservative...)

Discussion on CR's → no guidance to choosing "good" regions

But: ∃ close relationship between CR's and HT's!
↪ exploit to transform good testing properties into good CR properties

Suppose $R(\boldsymbol{Y})$ is an exact $100q\%$=$100(1-\alpha)\%$ CR for $\boldsymbol{\theta}$. Consider

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad vs \qquad H_1 : \theta \neq \theta_0$$

Define test function:

$$\delta(\boldsymbol{Y}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}_0 \notin R(\boldsymbol{Y}), \\ 0 & \text{if } \boldsymbol{\theta}_0 \in R(\boldsymbol{Y}). \end{cases}$$

Then, $$\mathbb{E}_{\boldsymbol{\theta}_0}[\delta(\boldsymbol{Y})] = 1 - \mathbb{P}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_0 \in R(\boldsymbol{Y})] \leq \alpha$$

Can use a CR to construct test with significance level α!

Going the other way around, can *invert* tests to get CR's:

Suppose we have tests at level α for any choice of simple null, $\boldsymbol{\theta}_0 \in \Theta$.
↪ Say that $\delta(\boldsymbol{Y}; \boldsymbol{\theta}_0)$ is appropriate test function for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$

Define $$R^*(\boldsymbol{Y}) = \{\boldsymbol{\theta}_0 : \delta(\boldsymbol{Y}; \boldsymbol{\theta}_0) = 0\}$$

Coverage probability of $R^*(\boldsymbol{Y})$ is

$$\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \in R^*(\boldsymbol{Y})] = \mathbb{P}_{\boldsymbol{\theta}}[\delta(\boldsymbol{Y}; \boldsymbol{\theta}) = 0] \geq 1 - \alpha$$

Obtain a $100(1-\alpha)\%$ confidence region by choosing all the $\boldsymbol{\theta}$ for which the null would not be rejected given our data $\boldsymbol{Y}$.
↪ If test inverted is powerful, then get "small" region for given $1-\alpha$.

Modern example: looking for signals in noise

- Interested in detecting presence of a signal $\mu(x_t)$, $t = 1, \ldots, T$ over a discretised domain, $\{x_1, \ldots, x_t\}$, on the basis of noisy measurements
- This is to be detected against some known background, say 0.
- May or may not be specifically interested in detecting the presence of the signal in some particular location x_t, but in detecting whether the a signal is present anywhere in the domain.

Formally:

Does there exist a $t \in \{1, \ldots, T\}$ such that $\mu(x_t) \neq 0$?

or

for which t's is $\mu(x_t) \neq 0$?

More generally:

- Observe
$$Y_t = \mu(x_t) + \varepsilon_t, \qquad t = 1, \ldots, T.$$
- Wish to test, at some significance level α:
$$\begin{cases} H_0 : \mu(x_t) = 0 & \text{for all } t \in \{1, \ldots, T\}, \\ H_A : \mu(x_t) \neq 0 & \text{for some } t \in \{1, \ldots, T\}. \end{cases}$$
- May also be interested in which specific locations signal deviates from zero
- More generally: May have T hypotheses to test simultaneously at level α (they may be related or totally unrelated)
- Suppose we have a test statistic for each individual hypothesis $H_{0,t}$ yielding a p-value p_t.

Bonferroni Method.

If we test each hypothesis individually, we will not maintain the level!

Can we maintain the level α?

Idea: use the same trick as for confidence regions!

Bonferroni

1. Test individual hypotheses separately at level $\alpha_t = \alpha / T$
2. Reject H_0 if at least one of the $\{H_{0,t}\}_{t=1}^{T}$ is rejected

Global level is bounded as follows:

$$\mathbb{P}[\text{not } H_0 | H_0] = \mathbb{P}\left[\bigcup_{t=1}^{T} \{\text{not } H_{0,t}\} \,\middle|\, H_0\right] \leq \sum_{t=1}^{T} \mathbb{P}[\text{not } H_{0,t} | H_0] = T\frac{\alpha}{T} = \alpha$$

Holm-Bonferroni Method.

- Advantage: Works for any (discrete domain) setup!
- Disadvantage: Too conservative when T large

Holm's modification increases average # of hypotheses rejected at level α (but does not increase power for overall rejection of $H_0 = \cap_{t \in T} H_{0,t}$)

Holm's Procedure

1. We reject $H_{0,t}$ for small values of a corresponding p-value, p_t
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. Starting from $t = 1$ and going up, reject all $H_{0,(t)}$ such that $p_{(t)}$ significant at level $\alpha/(T - t + 1)$. Stop rejecting at first insignificant $p_{(t)}$.

Genuine improvement over Bonferroni if want to detect as many signals as possible, not just existence of some signal

Both Holm and Bonferroni reject the global H_0 if and only if $\inf_t p_t$ significant at level α/T.

Taking Advantage of Structure: Independence.

In the (special) case where individual test statistics are independent, one may use Sime's (in)equality,

$$\mathbb{P}\left[p_{(j)} \geq \frac{j\alpha}{T}, \text{ for all } j = 1, ..., T \,\middle|\, H_0 \right] \geq 1 - \alpha$$

(strict equality requires continuous test statistics, otherwise $\leq \alpha$)

Yields Sime's procedure (assuming independence)

1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. If, for some $j = 1, \ldots, T$ the p-value $p_{(j)}$ is significant at level $\frac{j\alpha}{T}$, then reject the global H_0.

Provides a test for the global hypothesis H_0, but does not "localise" the signal at a particular x_t

One can, however, devise a sequential procedure to "localise" Sime's procedure, at the expense of lower power for the global hypothesis H_0:

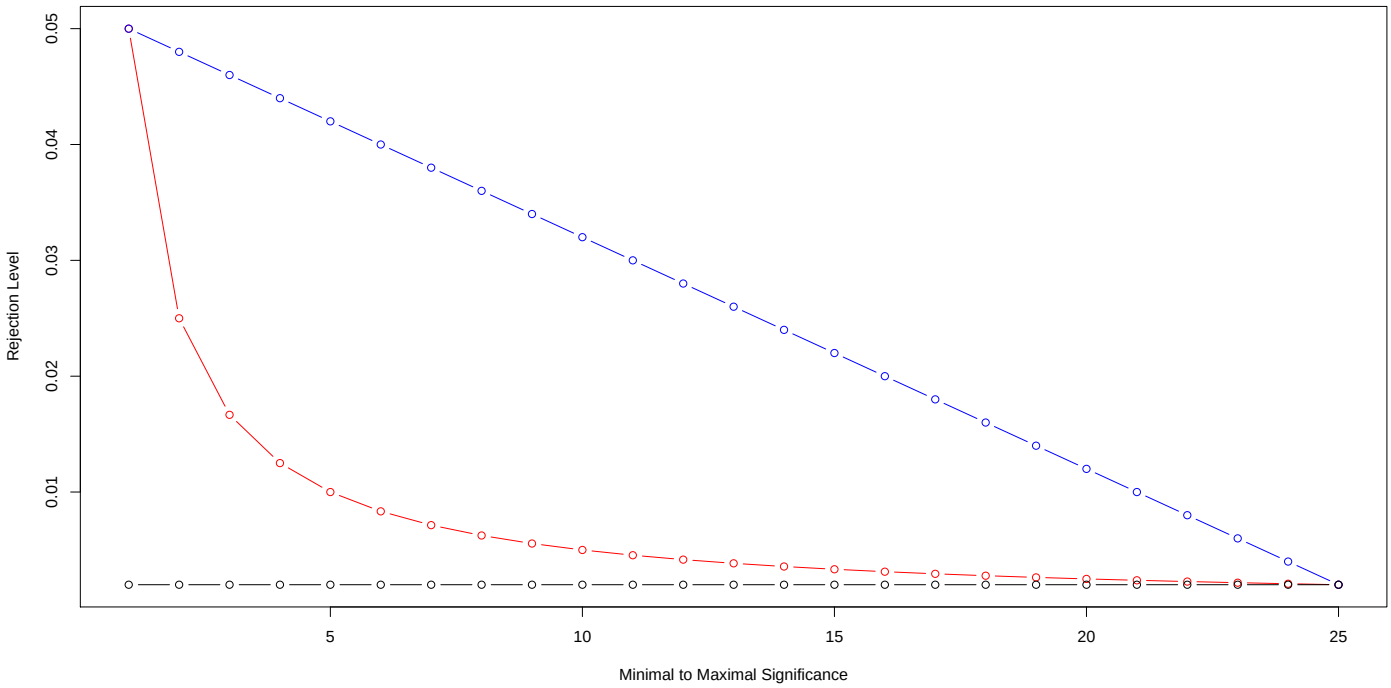
Hochberg's procedure (assuming independence)

1. Suppose we reject $H_{0,j}$ for small values of p_j
2. Order p-values from most to least significant: $p_{(1)} \leq \ldots \leq p_{(T)}$
3. Starting from $j = T, T-1, ...$ and down, accept all $H_{0,(j)}$ such that $p_{(j)}$ insignificant at level $\alpha/(T - j + 1)$.
4. Stop accepting for the first j such that $p_{(j)}$ is significant at level α/j, and reject all the remaining ordered hypotheses past that j going down.

Genuine improvement over Holm-Bonferroni both overall (H_0) and in terms of signal localisation:

1. Rejects "more" individual hypotheses than Holm-Bonferroni
2. Power for overall H_0 "weaker" than Sime's (for $T > 2$), much "stronger" than Holm (for $T > 1$).

Bonferroni, Hochberg, Simes



Nonparametric Estimation

A different idea:

can we estimate the distribution F itself from data $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ without assuming any particular functional form?

- Termed nonparametric estimation as there is no specific parameter θ.
- Otherwise said, $\{F(x) : x \in \mathbb{R}\}$ is itself an infinite-dimensional parameter.

- OK, but how?

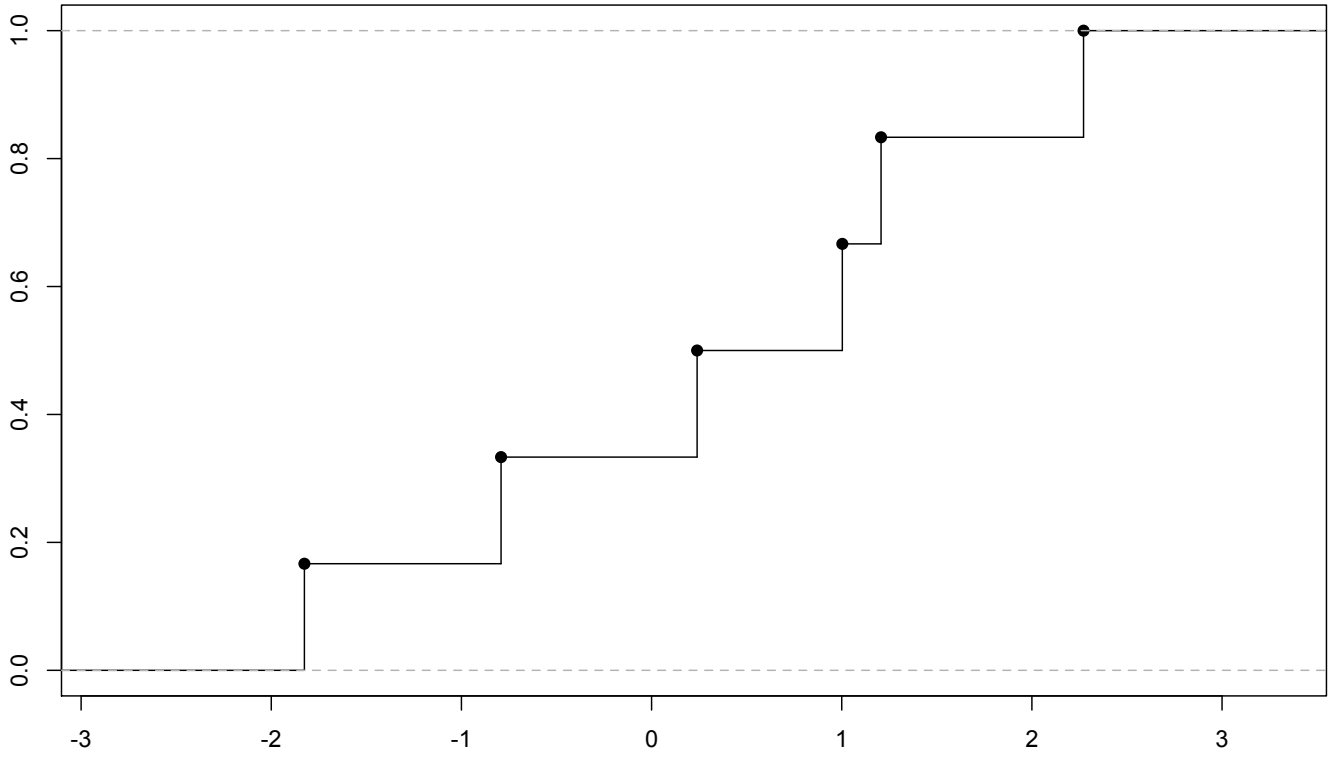
Definition (Empirical Distribution Function)

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$$

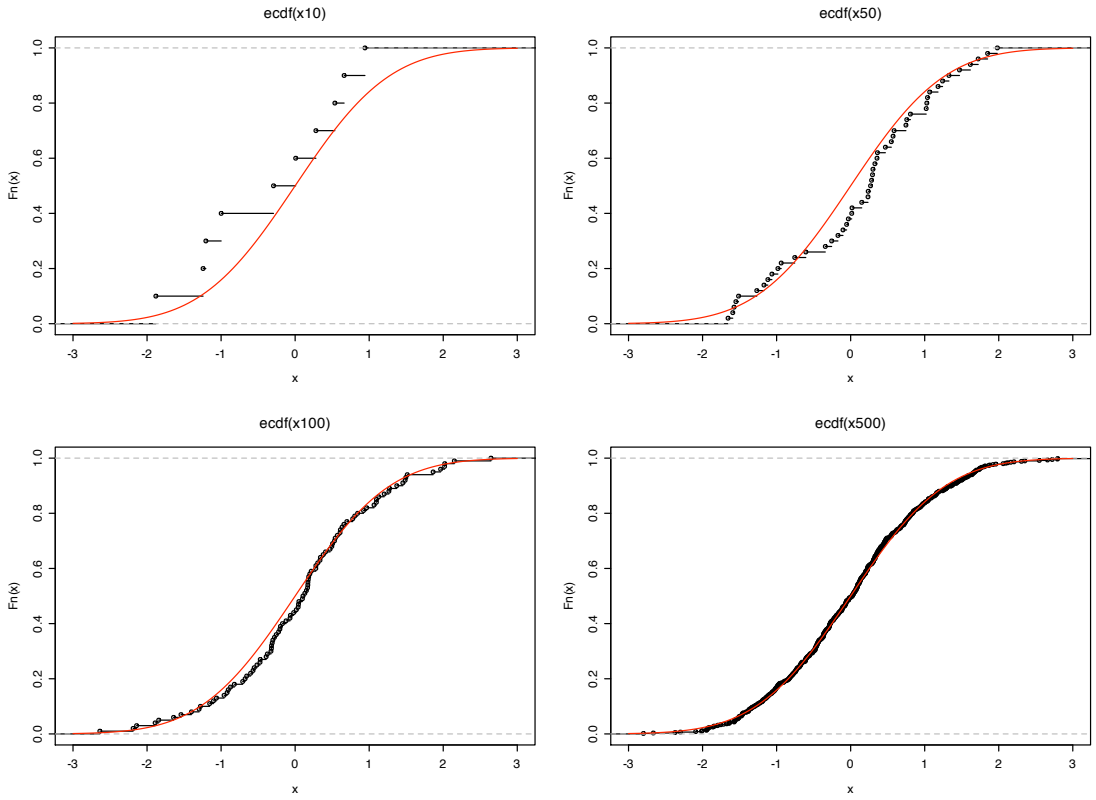
- CDF of the mass function placing mass $1/n$ on location of each Y_i.
- Notice that $W_i(y) := \mathbf{1}\{Y_i \leq y\} \overset{iid}{\sim} \text{Bernoulli}(F(y))$.
- Thus law of large numbers $\implies \hat{F}_n(y) \xrightarrow{a.s.} F(y)$ pointwise $\forall y \in \mathbb{R}$
- Notice how we got consistency without any assumption on form of F!

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$, $n = 6$



- Jump locations at $Y_1, \ldots, Y_n$.
- Jump sizes of $1/n$ ($1/6$ in this case)

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$ for $n = 10,\ 50,\ 100,\ 500$.



Looks like we're doing better than pointwise a.s. convergence...

Theorem (Glivenko-Cantelli)

Let $Y_1, .., Y_n$ be ii random variables with distribution function F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^n \mathbf{1}\{Y_i \leq y\}$ converges uniformly to F with probability 1, i.e.

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Proof ($*$).

Assume first that $F(y) = y\mathbf{1}\{y \in [0,1]\}$. Fix a regular finite partition $0 = x_1 \leq x_2 \leq \ldots \leq x_m = 1$ of [0,1] (so $x_{k+1} - x_k = (m-1)^{-1}$). By monotonicity of $F, \hat{F}_n$

$$\sup_x |\hat{F}_n(x) - F(x)| < \max_k |\hat{F}_n(x_k) - F(x_{k+1})| + \max_k |\hat{F}_n(x_k) - F(x_{k-1})|$$

Adding and subtracting $F(x_k)$ within each term we can bound above by

$$2 \max_k |\hat{F}_n(x_k) - F(x_k)| + \underbrace{\max_k |F(x_k) - F(x_{k+1})| + \max_k |F(x_k) - F(x_{k-1})|}_{= \max_k |x_k - x_{k+1}| + \max_k |x_k - x_{k-1}| = \frac{2}{m-1}}$$

by an application of the triangle inequality to each term.

OK, so let's investigate our bound

$$2 \max_k |\hat{F}_n(x_k) - F(x_k)| + \underbrace{\max_k |F(x_k) - F(x_{k+1})| + \max_k |F(x_k) - F(x_{k-1})|}_{= \max_k |x_k - x_{k+1}| + \max_k |x_k - x_{k-1}| = \frac{2}{m-1}}$$

Letting $n \uparrow \infty$, the SLLN implies that the first term vanishes almost surely. Since m is arbitrary we have proven that, given any $\epsilon > 0$,

$$\lim_{n \to \infty} \left[\sup_x |\hat{F}_n(x) - F(x)| \right] < \epsilon \quad a.s.$$

which gives the result when the cdf F is uniform.

For a general cdf F, we let $U_1, U_2, \ldots \overset{iid}{\sim} \mathcal{U}[0,1]$ and define

$$W_i := F^{-1}(U_i) = \inf\{x : F(x) \geq U_i\}.$$

Observe that

$$W_i \leq x \iff U_i \leq F(x)$$

so that $W_i \overset{d}{=} Y_i$. We may thus assume that $W_i = Y_i$ a.s.

Letting $\hat{G}_n$ be the edf of $(U_1, ..., U_n)$ we note that

$$\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{W_i \leq y\} = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq F(y)\} = \hat{G}_n(F(y)), \quad \text{a.s.}$$

in other words $\hat{F}_n = \hat{G}_n \circ F$, a.s.

Now let $A = F(\mathbb{R}) \subseteq [0, 1]$ so that from the first part of the proof

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in A} |\hat{G}_n(t) - t| \leq \sup_{t \in [0,1]} |\hat{G}_n(t) - t| \overset{a.s.}{\to} 0$$

since obviously $A \subseteq [0, 1]$. □

Some conclusions:

- Empirical distribution converges to true one at same right anywhere on $\mathbb{R}$
- Again, nothing specific was assumed
- Reason is that distribution functions are quite special (recall properties)
- Suggests we should be able to define "confidence bands" depending on n.

Theorem (Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)

Let $Y_1, .., Y_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$ satisfies

$$\mathbb{P}\left\{\sup_{y \in \mathbb{R}} \left|\hat{F}_n(y) - F(y)\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}$$

for all $\epsilon > 0$.

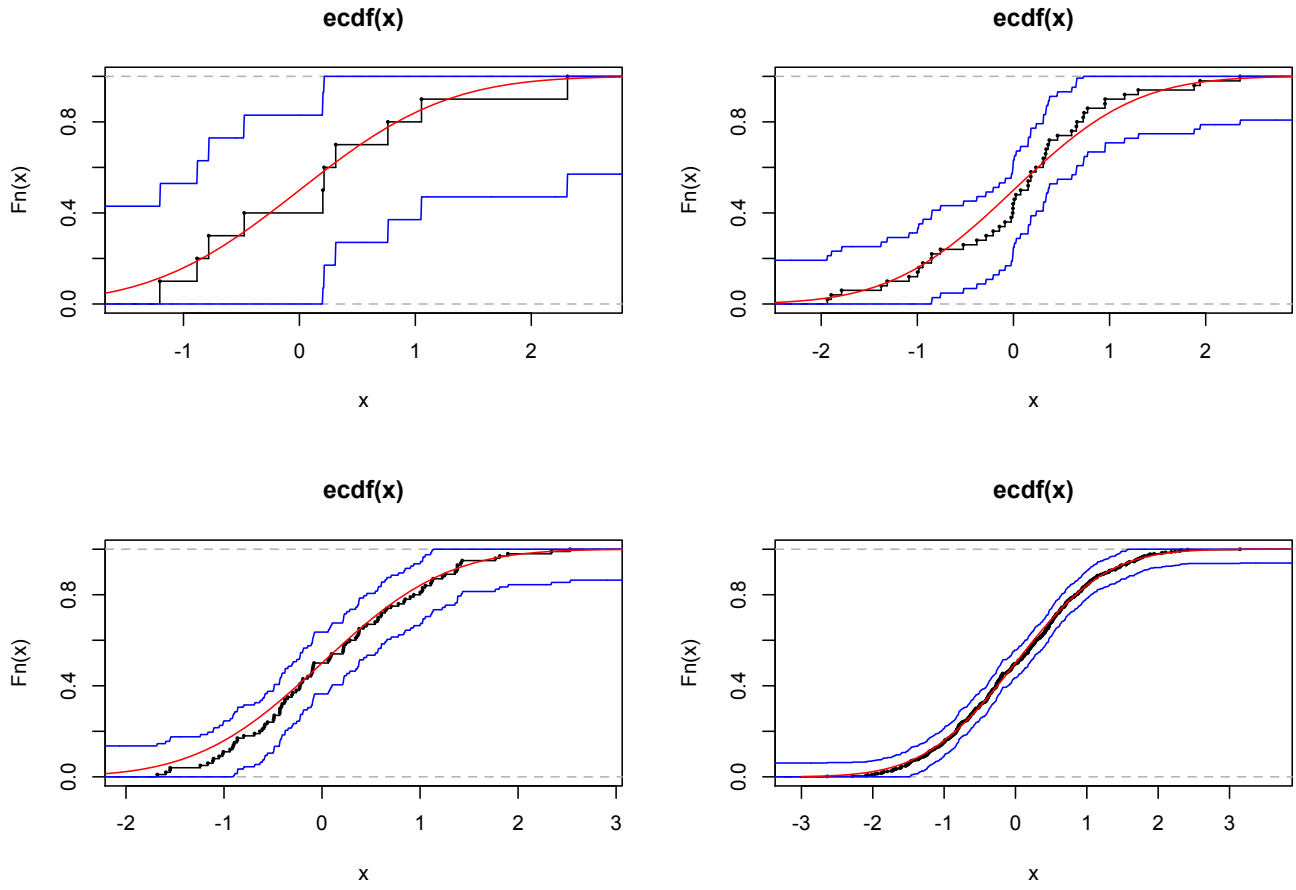
Let's now construct a "confidence interval" for F (known as confidence band)

- For confidence level $\alpha \in (0, 1)$, set $\epsilon_n = \sqrt{\frac{1}{2n} \log(2/\alpha)}$
- Define $L(y) = \max\{\hat{F}_n(y) - \epsilon_n, 0\}$ and $U(y) = \min\{\hat{F}_n(y) + \epsilon_n, 1\}$
- Apply DKW inequality and conclude

$$\mathbb{P}\{L(y) \leq F(y) \leq U(y) \ \forall y \in \mathbb{R}\} \geq 1 - \alpha.$$

- Can also use for hypothesis testing, using duality.

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$ for $n = 10, 50, 100, 500$.
In blue: 95% DKW confidence bands



Estimating Parameters from EDFs

If we'd still like to estimate a parameter, can use the plug-in principle

Let $\nu = \nu(F)$ be a parameter of interest.
We can use $\nu(\hat{F}_n)$ as an estimator of $\nu(F)$, i.e. *plug* $\hat{F}_n$ in $\nu(\cdot)$.

- "Flipped" point of view: viewing parameter ν as a function of F.
- Only sort of parameter we can consider, since no parametric model assumed!

Example (Mean, variance, median)

- For mean $\mu(F) = \int_{-\infty}^{+\infty} y dF(y)$ get
$$\hat{\theta} := \theta(\hat{F}_n) = \int_{-\infty}^{+\infty} y d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$$
- For variance $\sigma^2(F) = \int_{-\infty}^{+\infty} (y - \mu(F))^2 \, dF(y)$ get
$$\sigma^2(\hat{F}_n) = \int_{-\infty}^{+\infty} \left(y - \int_{-\infty}^{+\infty} u d\hat{F}_n(u)\right)^2 d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$$
- For median $m(F) = F^{-1}(1/2)$ and n odd get
$$\hat{m} = m(\hat{F}_n) = Y_{\left(\frac{n+1}{2}\right)}$$

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)\,dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - ↪ Parameters usually interpretable, CDFs are harder to appreciate visually.
- Densities are more easily interpreted – also defined as functional of CDF!
- The density f (when it exists) at $x_0 \in \mathbb{R}$ is $\nu(F) := \frac{d}{dx} F(x)\big|_{x=x_0}$
- Caution: mapping $F \mapsto \nu(F)$ not a "well behaved" mapping in general...

Density Estimation

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)\,dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt} F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a "well-behaved" mapping:
  - If $x \notin \{Y_1, \ldots, Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, \ldots, Y_n\}$ estimator is undefined!
- Problem is that estimator requires differentiation of a function $\hat{F}_n$ with jumps
- We will need a "smoother" estimate of F to plug in instead of $\hat{F}_n$, e.g.

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} \Phi\left(\frac{x-y}{h}\right) d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x-Y_i}{h}\right)$$

  for Φ a standard normal CDF and $h > 0$ a smoothing parameter.
- Transforms flat steps with hard corners to inclined steps with smooth corners (buffs the edges)

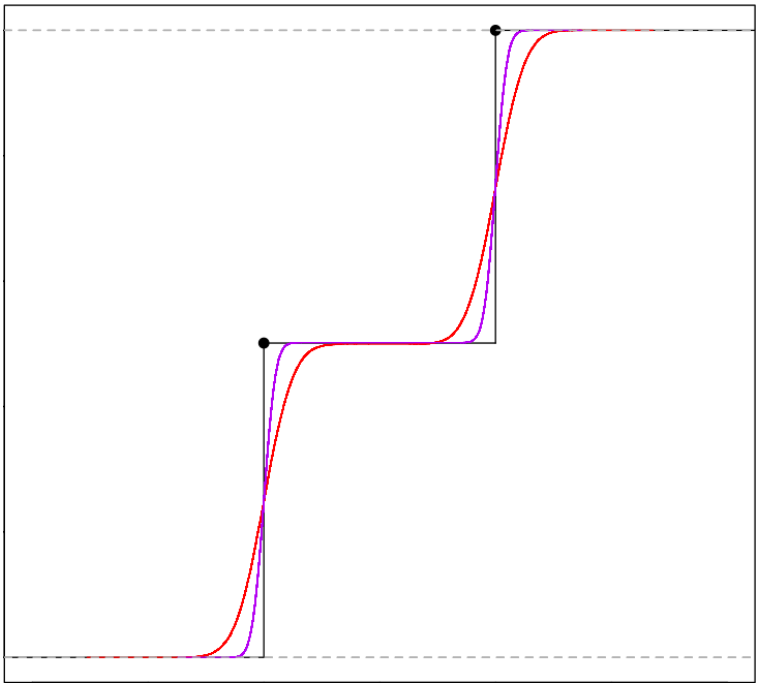


Figure: Empirical distribution function (black) for a size $n = 2$ sample, and "smoothed" approximations by convolution with $\Phi\left(\frac{u}{h}\right)$ for $h = 0.3$ (red) and $h = 0.2$ (purple).

At the level of density, this yields the "smoothed plug-in estimator"

$$\hat{f}(x) = \frac{d}{dx}\tilde{F}_n(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x - Y_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\varphi\left(\frac{x - Y_i}{h}\right)$$

for $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ the standard normal density.

- Nothing special about choice of φ – can choose any smooth unimodal probability density K that is symmetric about zero and has variance 1.
  ↪ Call such a K a kernel.
- Much more important is the choice of $h > 0$ called a *bandwidth* or *smoothing parameter*.

Definition (Kernel Density Estimator)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, where f is a probability density function. A *Kernel Density Estimator* (KDE) $\hat{f}$ of f is a random density function defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - Y_i}{h}\right)$$

for $K : \mathbb{R} \to \mathbb{R}$ a *kernel* and $h > 0$ a *bandwidth* or *smoothing parameter*.

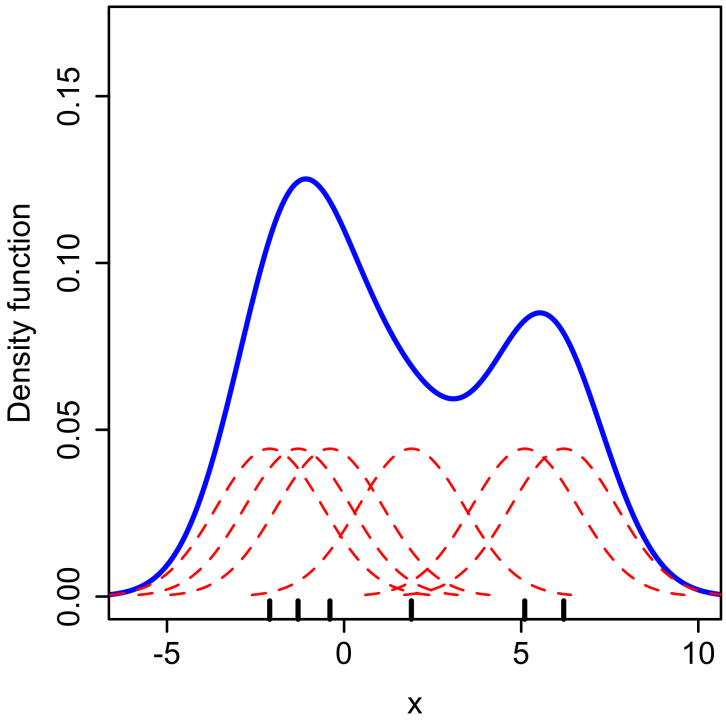


Figure: Schematic Illustration of a kernel density estimator

Only problem: how should we choose arbitrary tuning parameter $h > 0$?

$\hookrightarrow$ Can have decisive effect on quality of estimator.

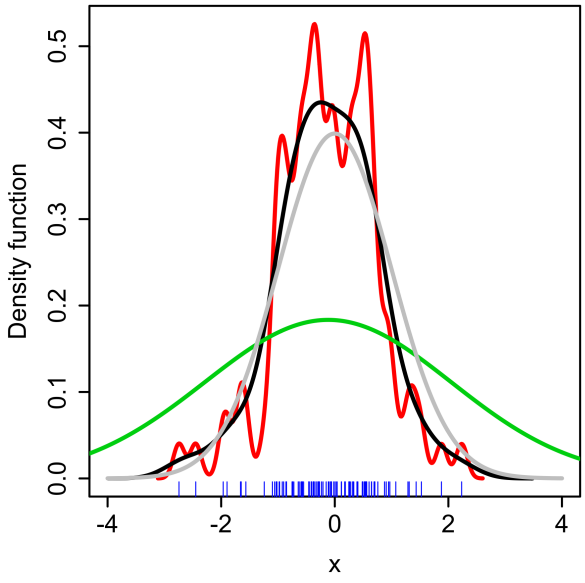


Figure: Effect of bandwidth choice on KDE of standard normal density, $n = 100$. True density in gray.KDE with: $h = 0.05$ in red, $h = 0.337$ in black, $h = 2$ in green.

To select h, need to understand its effect on KDE.

In short, it regulates the bias-variance tradeoff:

- Large h: gives "flattened" estimator (higher bias) but quite stable to small perturbations of the sample values (low variance).
- Small h: gives "wiggly" estimator (lower bias) but overly sensitive to small perturbations of the sample values (high variance).

What bias and variance? Those corresponding to integrated mean squared error:

$$\mathrm{IMSE}(\hat{f}, f) = \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}(x) - f(x)\Big)^2 dx.$$

$$\mathrm{IMSE}(\hat{f}, f) = \underbrace{\int_{\mathbb{R}} \Big(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\Big)^2 dx}_{\text{integrated squared bias}} + \underbrace{\int_{\mathbb{R}} \mathbb{E}\Big\{\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\Big\}^2 dx}_{\text{integrated variance}}$$

To get a useful expression for this we resort to asymptotics.

Theorem (Asymptotic Risk of KDE)

Let $f \in C^3$ be a probability density and $K \in C^2$ a kernel function satisfying

$$\int_{\mathbb{R}} \Big(f''(x)\Big)^2 dx < \infty \quad \int_{\mathbb{R}} \Big|f'''(x)\Big| dx < \infty \quad \& \quad \int_{\mathbb{R}} \Big(K''(x)\Big)^2 dx < \infty.$$

If $\hat{f}_n$ is the KDE of f with iid sample size n, kernel K and bandwidth h,

$$\mathrm{IMSE}(\hat{f}, f) = \tfrac{h^4}{4} \int_{\mathbb{R}} \Big(f''(x)\Big)^2 dx + \tfrac{1}{nh} \int_{\mathbb{R}} K^2(x)\, dx + o\left(h^4 + \tfrac{1}{nh}\right).$$

as $h \to 0$.

Conclusions:

- For consistency, need $h \to 0$ but $nh \to \infty$ as $n \to \infty$.
- Optimal choice of h will unfortunatelydepend on (unknown) f''
- For the record, optimal h is given (after some calculations) by

$$h^* = \left\{ \tfrac{1}{n} \int_{\mathbb{R}} K^2(x)\, dx \Big/ \int_{\mathbb{R}} \Big(f''(x)\Big)^2 dx \right\}^{1/5}$$

- Plugging in the optimal bandwidth yields the a risk of asymptotic order $n^{-4/5}$
- Compare this to parametric model optimal rate of n^{-1}
- Asymptotic bias proportional to curvature of f.

Proof $(*)$.

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)\,dt = \int_{\mathbb{R}} K(y) f(x-hy)\,dy$$

by change of variables $y = (x-t)/h$. Now Taylor expanding f yields

$$f(x-hy) = f(x) - hyf'(x) + \frac{1}{2}h^2y^2f''(x) + o(h^2) \quad \text{as } h \to 0.$$

Plugging into the equation for the expectation, we get that $\mathbb{E}[\hat{f}_n(x)]$ equals

$$f(x)\underbrace{\int_{\mathbb{R}} K(y)\,dy}_{=1} - hf'(x)\underbrace{\int_{\mathbb{R}} yK(y)\,dy}_{=0} + \frac{1}{2}h^2f''(x)\underbrace{\int_{\mathbb{R}} y^2K(y)\,dy}_{=1} + o(h^2)$$

as $h \to 0$ by the kernel properties of K. In summary the pointwise bias is

$$\mathbb{E}\left[\hat{f}_n(x)\right] - f(x) = \frac{1}{2}h^2f''(x) + o(h^2), \quad \text{as } h \to 0.$$

The pointwise variance $\text{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\text{var}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

and by similar manipulations as earlier, and the expression for $\mathbb{E}\left[\hat{f}_n(x)\right]$, we get

$$\text{var}[\hat{f}_n(x)] = \underbrace{\frac{1}{nh}\int_{\mathbb{R}} K^2(y)f(x-hy)\,dy}_{A} - \underbrace{\frac{1}{nh^2}\mathbb{E}^2[\hat{f}_n(x)]}_{B}$$

Now observe that as as $h \to 0$, we have

$$B = \frac{1}{nh^2}(f(x) + \frac{1}{2}h^2f''(x) + o(h^2))^2 = \frac{1}{nh^2}[f(x) + o(h)]^2 = o\left(\frac{1}{n}\right).$$

On the other hand, Taylor expanding $f(x-hy) = f(x) + o(1)$ as $h \to 0$, we have

$$A = \tfrac{1}{nh}\textstyle\int_{\mathbb{R}} K^2(y)[f(x) + o(1)]\,dy = \tfrac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)\,dy + o\left(\tfrac{1}{nh}\right)$$

since $\frac{1}{nh}o(1) = o\left(\frac{1}{nh}\right)$

Putting A and B together gives

$$\operatorname{var}[\hat{f}_n(x)] = \tfrac{1}{nh}\int_{\mathbb{R}} K^2(y)\,dy + o\left(\tfrac{f(x)}{nh}\right) - o\left(\tfrac{1}{n}\right) = \tfrac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)\,dy + o\left(\tfrac{1}{nh}\right)$$

Summing pointwise squared-bias and variance, the pointwise MSE is given by

$$MSE(\hat{f}_n(x), f(x)) = \frac{1}{4}h^4(f''(x))^2 + \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)\,dy + o\left(h^4 + \frac{1}{nh}\right)$$

Finally, integrating over $\mathbb{R}$ and re-arranging yields the sought form

$$\operatorname{IMSE}(\hat{f}, f) = \tfrac{1}{nh}\int_{\mathbb{R}} K^2(x)\,dx + \tfrac{h^4}{4}\int_{\mathbb{R}} \left(f''(x)\right)^2 dx + o\left(h^4 + \tfrac{1}{nh}\right).$$

□

From $n^{-4/5}$ to $n^{-2m/(2m+1)}$

Can we do better than $n^{-4/5}$ by more smoothness assumptions?

Theorem (Minimax Optimal Rates for KDE)

Let $\mathcal{F}(m, r)$ be the subset of m-differentiable densities with mth derivative in an L^2 ball of radius r,

$$\int_{\mathbb{R}} \left(f^{(m)}(x)\right)^2 dx \leq r^2.$$

Then, given any KDE $\hat{f}_n$,

$$\sup_{f \in \mathcal{F}(m,r)} \mathbb{E}\left\{\int_{\mathbb{R}} \left(\hat{f}_n(x) - f(x)\right)^2 dx\right\} \geq Cn^{-\frac{2m}{2m+1}},$$

where the constant $C > 0$ depends only on m and c.

- The smoother the density the better the worst case rate.
- Can never beat n^{-1}, though.
- The price to pay for flexibility!

So how do we choose h in practice? Here's a couple of approches:

- Pilot estimator: use a parametric family (e.g. normal, or mixture) to obtain a preliminary estimator $\check{f}$, and plug this into the optimal bandwidth expression to select a bandwidth.
- Least squares cross-validation: try to construct an unbiased estimator of the IMSE after all, it is an expectation. Then choose h to minimise the estimated IMSE. Also known as unbiased risk estimation.

Let's consider the second approach in more detail. Notice that we can write

$$\begin{aligned} IMSE(\hat{f}_h, f) &= \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}_h(x) - f(x)\Big)^2 dx = \mathbb{E}\left[\int_{\mathbb{R}} \Big(\hat{f}_h(x) - f(x)\Big)^2 dx\right] \\ &= \underbrace{\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)\,dx\right] - 2\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h(x) f(x)\,dx\right]}_{H(\hat{f}_h)} + \mathbb{E}\left[\int_{\mathbb{R}} f^2(x)\,dx\right]. \end{aligned}$$

where the last term does not vary with h.

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)\,dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)\,dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\Big[\hat{f}_{h,-i}(Y_i)\Big] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_i - Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - Y_2}{h}\right)\right] =$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{u - v}{h}\right) f(u) f(v)\,du\,dv = \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - v}{h}\right)\right] f(v)\,dv =$$

$$= \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)\right] f(v)\,dv = \mathbb{E}\left[\int_{\mathbb{R}} \underbrace{\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)}_{=\hat{f}_h(v)} f(v)\,dv\right]$$

Thus $\{\hat{f}_{h,-i}(Y_i)\}_{i=1}^n$ are n variables with mean $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h(x) f(x)\,dx\right]$!

Motivates definition of leave-one-out cross validation estimator

$$LSCV(h) = \int_{\mathbb{R}} \hat{f}_h^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(Y_i)$$

which by construction satisfies

$$\mathbb{E}[LSCV(h)] = H(\hat{f}_h).$$

Strategy: choose h by minimising $LSCV(h)$. Does it work?

Theorem (Stone's Theorem)

In the same context, and under the same assumptions, let h_{CV} denote the bandwidth selected by cross-validation. Then,

$$\frac{\displaystyle\int_{\mathbb{R}} \left(\hat{f}_{h_{CV}}(x) - f(x)\right)^2 dx}{\displaystyle\inf_{h>0} \int_{\mathbb{R}} \left(\hat{f}_h(x) - f(x)\right)^2 dx} \xrightarrow{a.s.} 1,$$

provided that the true density f is bounded.

Higher dimensions?

Conceptually, can generalise KDE very easily to higher dimensions.

- Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n \overset{iid}{\sim} f(\boldsymbol{y})$ be a sample in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to [0, +\infty)$
- Let $\boldsymbol{H} \succeq 0$ be a $d \times d$ symmetric positive-definite bandwidth matrix.
- Let K be a probability density on $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{I}_{d\times d}$.
  - ↪ E.g. $K(x_1, ..., x_n) = \prod_{j=1}^{d} \varphi(x_j)$ for φ the $N(0,1)$ density.

We can define a $d-$dimensional KDE as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{Y}_i)\right), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Once again choice of kernel is secondary but choice of $\boldsymbol{H}$ is paramount.

- Considerably harder: need to choose $d(d+1)/2 \sim d^2$ bandwidth parameters.
- Intuitively: $\boldsymbol{H} = \boldsymbol{U}\text{diag}\{h_1, ..., h_d\}\boldsymbol{U}^\top$ for $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{d\times d}$ and $h_j > 0$.
  - ↪ Choose d smoothing directions, and a bandwidth for each such direction.
- LSCV-type solutions exist for d moderate (computationally intensive).
- Visualisation challenging for $d > 3$.

But what about the quality of estimation?

- Consider simplest special case where $\boldsymbol{H} = h\boldsymbol{I}_{d\times d}$ for $h > 0$.
- Take $K(x_1, ..., x_d) = \prod_{j=1}^{d} \varphi(x_j)$.
- Yields $\hat{f}(x_1, ..., x_d) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{n} \varphi\left(\frac{x_i - Y_{ij}}{h}\right)$

Mimicking our calculations in the 1D case, we can arrive at an approximate risk

$$\mathrm{IMSE}(\hat{f}, f) \approx \frac{h^4}{4} \sum_{j=1}^{d} \sum_{k=1}^{d} \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_j} f(\boldsymbol{x}) \frac{\partial^2}{\partial x_k} f(\boldsymbol{x}) \boldsymbol{x} + \frac{1}{nh^d} \int_{\mathbb{R}^d} K^2(\boldsymbol{x}) d\boldsymbol{x}$$

- Optimal bandwidth now satisfies $h \propto n^{-1/(4d)}$
- Yields rate of convergence of $n^{-\frac{4}{4+d}}$ (*very* bad news)

Table: Equivalent sample sizes n for comparable risk values in different dimensions d

d	1	2	3	4	5	6	7	8	9	10
n	4	19	67	223	768	2'790	10'700	43'700	187'000	842'000

Takehome messages:

- Tradeoff between flexibility and efficiency
- Parametric model enforces rigid form of parsimony (precise formula, few parameters).
- Nonparametric model enforces soft parsimony (smoothness, via bandwidth parameter)
- If model can be confidently assumed, parametric inference is preferable.
- Otherwise, nonparametric methods more flexible and requiring few assumptions.
- Particularly in higher dimensions parametric models more interpretable and efficient.
- But nonparametric curse of dimensionality can be (partially) mitigated by clever approximations (separable/additive models, ridge models, neural networks – more later).
- A very important class of models are semiparametric models. These have some parametric and some nonparametric components.
  - In important cases, can attain parametric efficiency for parametric component.

Regression

Absence or Presence of Covariates

In the beginning we distinguished between:

1. **Marginal Inference.** Here $(Y_1, ..., Y_n)^\top$ has i.i.d. entries each from the same distribution $F(y;\theta)$ with the same parameter θ.
   - In other words, all observations were obtained under identical experimental conditions, and thus depend in the same way on the same unknown θ.
2. **Regression.** Here $(Y_1, ..., Y_n)^\top$ has independent entries, each with distribution $F(y;\theta_i)$ of the same family but with different parameters.
   - Each observation was generated under slightly different experimental conditions. They depend in a similar way on different θ_i.
   - These θ_i correspond to different experimental conditions, say x_i.
   - Each x_i is called a covariate/feature, and is an input that the experimenter can vary. They are known. The index i reminds us that it corresponds to the ith observation Y_i.
   - Usually θ_i is postulated to have a special relationship to x_i, for example $\theta_i = \exp\{\alpha + \beta x_i\}$, for (α, β) uknown parameters.
   - The point here is to understand the effect of varying the covariate/feature on the distribution of the observable.

Statistical model for:

$$\boxed{Y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}}$$

Aim: understand the effect of x on the distribution of random variable Y

General formulation[6]:

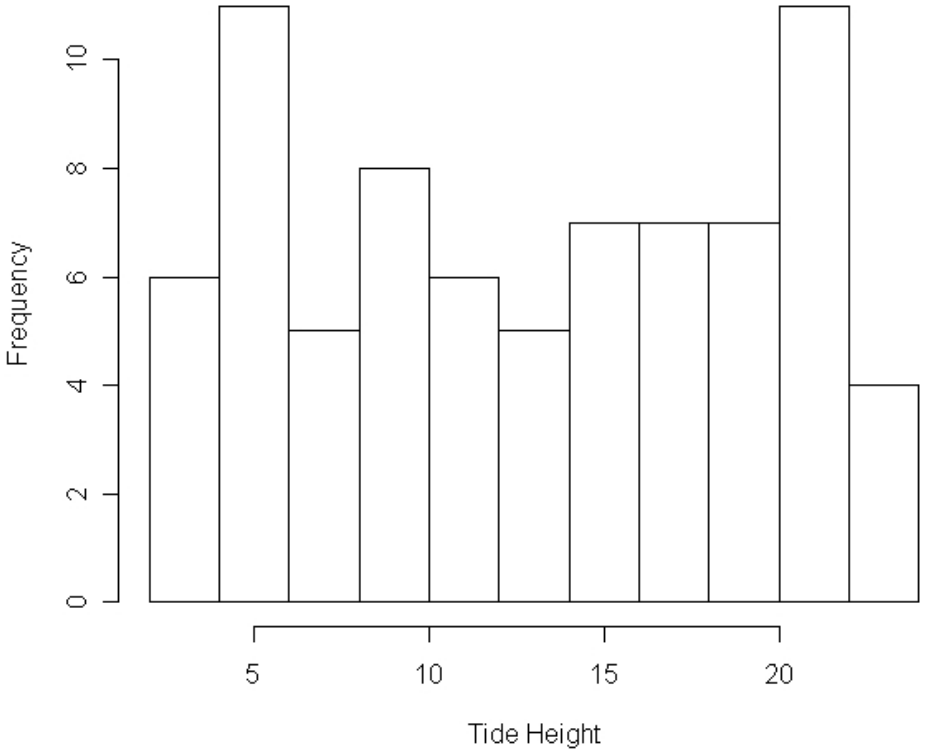
$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

Statistical Problem: Estimate (learn) $g(\cdot)$ from data $\{(x_i, Y_i)\}_{i=1}^n$. Use for:

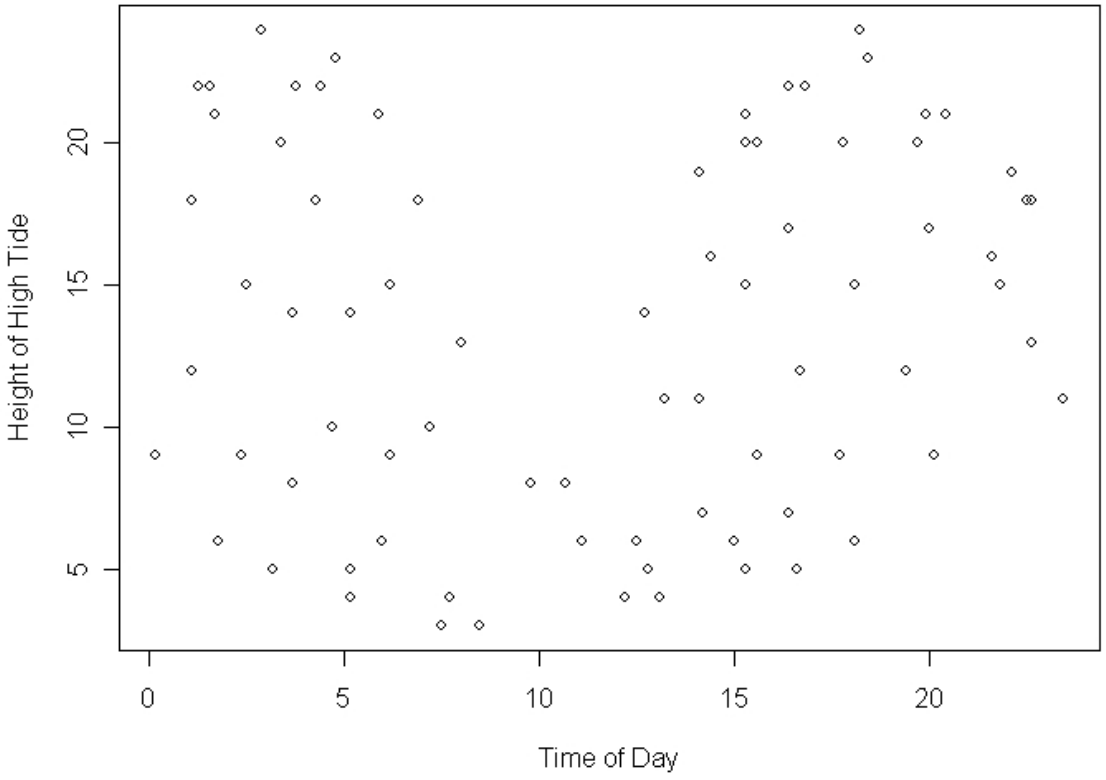
- Inference
- Prediction
- Data compression (parsimonious representations)

[6] Sometimes we write $Y_i | x_i \overset{\text{independent}}{\sim} \text{Distribution}\{g(x_i) = \theta_i\}$ to highlight that the distribution of Y depends on x, but without meaning that (X, Y) are jointly random; such an assumption is unnecessary (e.g., in a designed experiment we choose values for x).

Example: Honolulu tide



Example: Honolulu tide



A bewildering variety of models can be captured by the general specificaiton

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

x_i can be:

- continuous, discrete, categorical, vector ...
- arrive randomly, or be chosen by experimenter, or both
- however x arises, we treat it as constant in the analysis

Distribution can be:

- Gaussian, Laplace, Bernoulli, Poisson, gamma, general exponential family, ...

Function $g(\cdot)$ can be:

- $g(x) = \beta_0 + \beta_1 x$, $g(x) = \sum_{k=-K}^{K} \beta_k e^{-ikx}$, cubic spline, neural net...

Table: A coarse classification of regression models we will consider

Distribution / Function g	$g(\boldsymbol{x}_i^\top) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	g nonparametric
Gaussian	Linear Regression	Smoothing
Exponential Family	GLM	GAM

Fundamental Case: Normal Linear Regression

- $Y, x \in \mathbb{R}$, $g(x) = \beta_0 + \beta_1 x$, Distribution $=$ Gaussian

$$Y \mid x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

$$\Updownarrow$$

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

The second verson is useful for mathematical work, but is puzzling statistically, since we don't observe ϵ.

- Also, covariate could be vector ($Y, \beta_0 \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^p$):

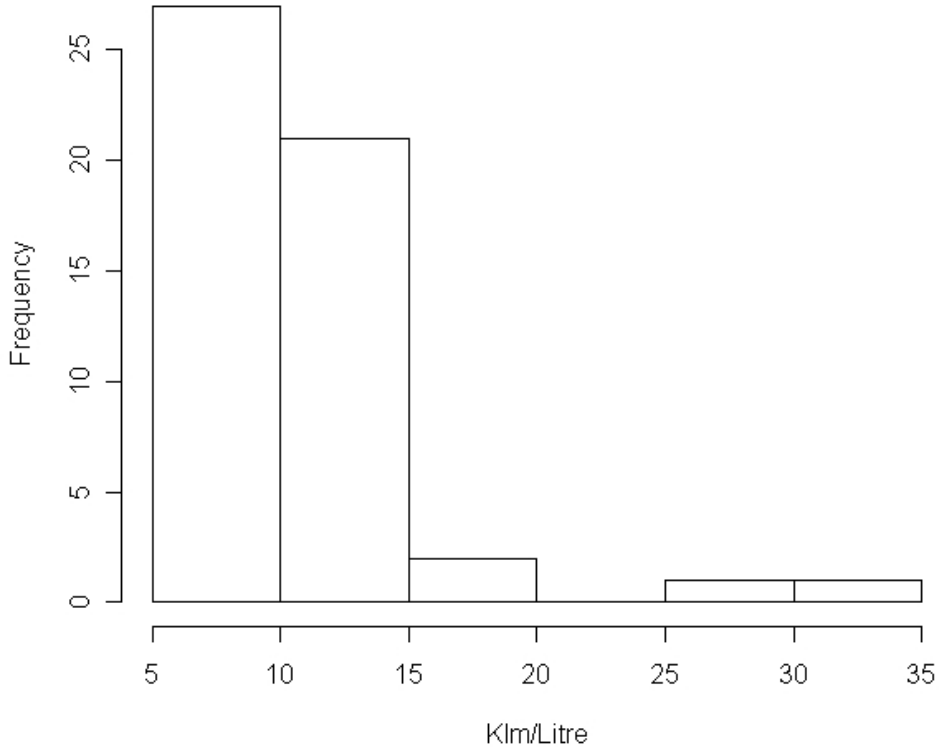
$$Y \mid x \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$$

$$\Updownarrow$$

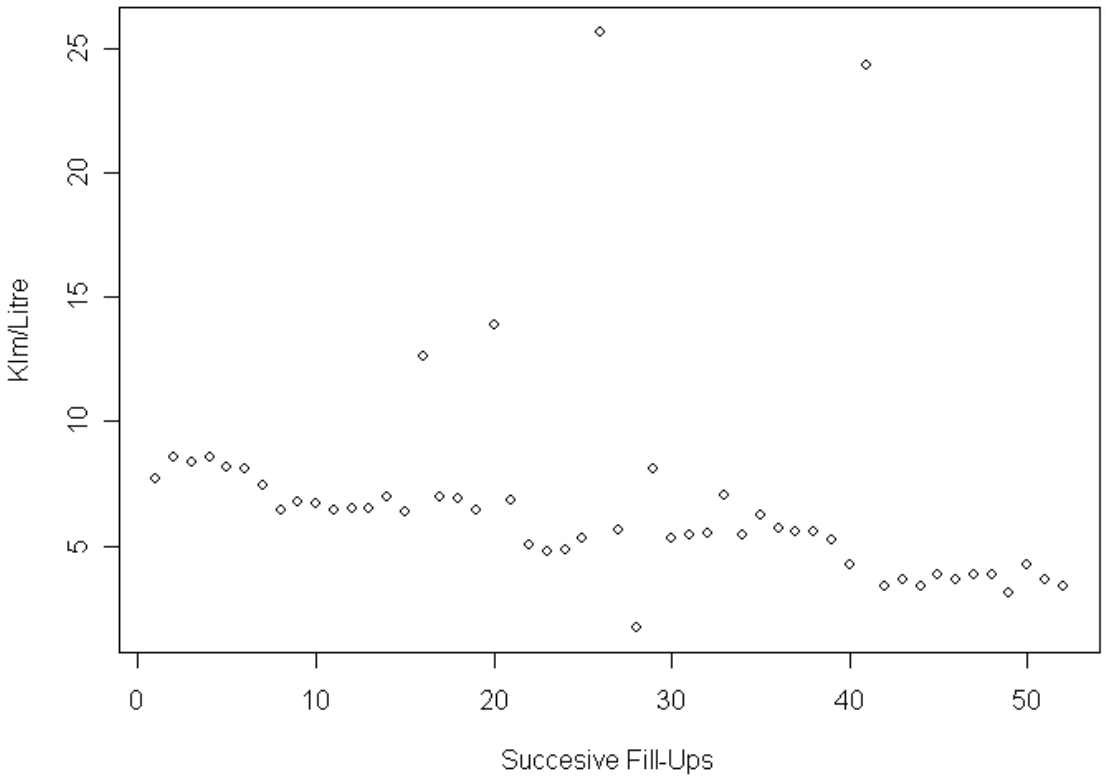
$$Y = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Example: Professor's Van

Example: Professor's Van



Example: Professor's Van



The tools of the trade . . .

Start from Gaussian linear regression then gradually generalise . . .

Obviously: important features of Gaussian linear model are

- Gaussian distribution
- Linearity

These two combine well and give geometric insights to solve the estimation problem. Thus we need to revise some probabilistic linear algebra. . .

- Subpsaces and projection matrices
- Multivariate Gaussian Distribution
- Optimal dimension reduction
- Random quadratic forms

Linear Algebra Intermezzo

Linear Subspaces, Orthogonal Projections, Gaussian Vectors

Subspaces and Spectra

If $\boldsymbol{Q}$ is an $n \times p$ real matrix, we define the column space (or *range*) of $\boldsymbol{Q}$ to be the set spanned by its columns:

$$\mathcal{M}(\boldsymbol{Q}) = \{\boldsymbol{y} \in \mathbb{R}^n : \ \exists \boldsymbol{\beta} \in \mathbb{R}^p, \ \boldsymbol{y} = \boldsymbol{Q}\boldsymbol{\beta}\}.$$

- Recall that $\mathcal{M}(\boldsymbol{Q})$ is a subspace of $\mathbb{R}^n$.
- The columns of $\boldsymbol{Q}$ provide a coordinate system for the subspace $\mathcal{M}(\boldsymbol{Q})$
- If $\boldsymbol{Q}$ is of full column rank (p), then the coordinates $\boldsymbol{\beta}$ corresponding to a $\boldsymbol{y} \in \mathcal{M}(\boldsymbol{Q})$ are unique.
- Allows interpretation of system of linear equations

$$\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{y}.$$

[existence of solution $\leftrightarrow$ is $\boldsymbol{y}$ an element of $\mathcal{M}(\boldsymbol{Q})$?]
[uniqueness of solution $\leftrightarrow$ is there a unique coordinate vector $\boldsymbol{\beta}$?]

Two further important subspaces associated with a real $n \times p$ matrix $\boldsymbol{Q}$:

- the null space (or *kernel*), $\ker(\boldsymbol{Q})$, of $\boldsymbol{Q}$ is the subspace defined as

$$\ker(\boldsymbol{Q}) = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{Q}\boldsymbol{x} = 0\};$$

- the orthogonal complement of $\mathcal{M}(\boldsymbol{Q})$, $\mathcal{M}^{\perp}(\boldsymbol{Q})$, is the subspace defined as

$$\begin{aligned} \mathcal{M}^{\perp}(\boldsymbol{Q}) &= \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y}^\top \boldsymbol{Q}\boldsymbol{x} = 0,\ \forall \boldsymbol{x} \in \mathbb{R}^p\} \\ &= \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y}^\top \boldsymbol{v} = 0,\ \forall \boldsymbol{v} \in \mathcal{M}(\boldsymbol{Q})\}. \end{aligned}$$

The orthogonal complement may be defined for arbitrary subspaces by using the second equality.

Theorem (Spectral Theorem)

A $p \times p$ matrix $\boldsymbol{Q}$ is symmetric if and only if there exists a $p \times p$ orthogonal matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top.$$

In particular:

1. *the columns of $\boldsymbol{U} = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_p)$ are eigenvectors of $\boldsymbol{Q}$, i.e. there exist λ_j such that*
$$\boldsymbol{Q}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \qquad j = 1, \ldots, p;$$
2. *the entries of $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ are the corresponding eigenvalues of $\boldsymbol{Q}$, which are real; and*
3. *the rank of $\boldsymbol{Q}$ is the number of non-zero eigenvalues.*

Note: if the eigenvalues are distinct, the eigenvectors are unique (up to changes in signs).

Theorem (Singular Value Decomposition)

Any $n \times p$ real matrix can be factorised as

$$\underset{n\times p}{\boldsymbol{Q}} = \underset{n\times n}{\boldsymbol{U}}\ \underset{n\times p}{\boldsymbol{\Sigma}}\ \underset{p\times p}{\boldsymbol{V}^\top},$$

where $\boldsymbol{U}$ and $\boldsymbol{V}^\top$ are orthogonal with columns called left singular vectors and right singular vectors, respectively, and Σ is diagonal with real entries called singular values.

1. The left singular vectors are eigenvectors of $\boldsymbol{Q}\boldsymbol{Q}^\top$.
2. The right singular vectors are eigenvectors of $\boldsymbol{Q}^\top\boldsymbol{Q}$.
3. The squares of the singular values are eigenvalues of both $\boldsymbol{Q}\boldsymbol{Q}^\top$ and $\boldsymbol{Q}^\top\boldsymbol{Q}$.
4. The left singular vectors corresponding to non-zero singular values form an orthonormal basis for $\mathcal{M}(\boldsymbol{Q})$.
5. The left singular vectors corresponding to zero singular values form an orthonormal basis for $\mathcal{M}^\perp(\boldsymbol{Q})$.

A matrix $\boldsymbol{Q}$ is called idempotent if $\boldsymbol{Q}^2 = \boldsymbol{Q}$.

An orthogonal projection (henceforth projection) onto a subspace $\mathcal{V}$ is a symmetric idempotent matrix $\boldsymbol{H}$ such that $\mathcal{M}(\boldsymbol{H}) = \mathcal{V}$.

Proposition

The only possible eigenvalues of a projection matrix are 0 and 1.

Proposition

Let $\mathcal{V}$ be a subspace and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then $\boldsymbol{I} - \boldsymbol{H}$ is the projection matrix onto $\mathcal{V}^\perp$.

Proof ($*$).

$(\boldsymbol{I} - \boldsymbol{H})^\top = \boldsymbol{I} - \boldsymbol{H}^\top = \boldsymbol{I} - \boldsymbol{H}$ since $\boldsymbol{H}$ is symmetric and,
$(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I}^2 - 2\boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - \boldsymbol{H}$. Thus $\boldsymbol{I} - \boldsymbol{H}$ is a projection matrix.

It remains to identify the column space of $\boldsymbol{I} - \boldsymbol{H}$. Let $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $\boldsymbol{H}$. Then $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{I} - \boldsymbol{\Lambda})\boldsymbol{U}^\top$.
Hence the column space of $\boldsymbol{I} - \boldsymbol{H}$ is spanned by the eigenvectors of $\boldsymbol{H}$ corresponding to zero eigenvalues of $\boldsymbol{H}$, which coincides with $\mathcal{M}^\perp(\boldsymbol{H}) = \mathcal{V}^\perp$. □

Proposition

Let $\mathcal{V}$ be a subspace and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}$ for all $\boldsymbol{y} \in \mathcal{V}$.

Proposition

If $\boldsymbol{P}$ and $\boldsymbol{Q}$ are projection matrices onto a subspace $\mathcal{V}$, then $\boldsymbol{P} = \boldsymbol{Q}$.

Proposition

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are linearly independent and are such that span$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \mathcal{V}$, then the projection onto $\mathcal{V}$ can be represented as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

where $\boldsymbol{X}$ is a matrix with columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$.

Proposition

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then

$$\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{v}\|, \qquad \forall \boldsymbol{v} \in \mathcal{V}.$$

Proof $(*)$.

Let $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $\boldsymbol{H}$, $\boldsymbol{U} = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_n)$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Letting $p = \dim(\mathcal{V})$,

1. $\lambda_1 = \cdots = \lambda_p = 1$ and $\lambda_{p+1} = \cdots = \lambda_n = 0$,
2. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is an orthonormal basis of $\mathbb{R}^n$,
3. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ is an an orthonormal basis of $\mathcal{V}$.

$$
\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{x}\|^2 &= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - (\boldsymbol{H}\boldsymbol{x})^\top \boldsymbol{u}_i)^2 \qquad \text{[orthonormal basis]} \\
&= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{u}_i)^2 \qquad [H \text{ is symmetric}] \\
&= \sum_{i=1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i - \lambda_i \boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad [u\text{'s are eigenvectors of } H] \\
&= 0 + \sum_{i=p+1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad \text{[eigenvalues 0 or 1]} \\
&\leq \sum_{i=1}^{p} (\boldsymbol{x}^\top \boldsymbol{u}_i - \boldsymbol{v}^\top \boldsymbol{u}_i)^2 + \sum_{i=p+1}^{n} (\boldsymbol{x}^\top \boldsymbol{u}_i)^2 \qquad \forall \boldsymbol{v} \in \mathcal{V} \\
&= \|\boldsymbol{x} - \boldsymbol{v}\|^2.
\end{aligned}
$$

□

Proposition

Let $\mathcal{V}_1 \subseteq \mathcal{V} \subseteq \mathbb{R}^n$ be two nested linear subspaces. If $\boldsymbol{H}_1$ is the projection onto $\mathcal{V}_1$ and $\boldsymbol{H}$ is the projection onto $\mathcal{V}$, then

$$\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1 = \boldsymbol{H}_1\boldsymbol{H}.$$

Proof ($*$).

First we show that $\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1$, and then that $\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}_1$. For all $\boldsymbol{y} \in \mathbb{R}^n$ we have $\boldsymbol{H}_1\boldsymbol{y} \in \mathcal{V}_1$. But then $\boldsymbol{H}_1\boldsymbol{y} \in \mathcal{V}$, since $\mathcal{V}_1 \subseteq \mathcal{V}$.
Therefore $\boldsymbol{H}\boldsymbol{H}_1\boldsymbol{y} = \boldsymbol{H}_1\boldsymbol{y}$. We have shown that $(\boldsymbol{H}\boldsymbol{H}_1 - \boldsymbol{H}_1)\boldsymbol{y} = 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$, so that $\boldsymbol{H}\boldsymbol{H}_1 - \boldsymbol{H}_1 = 0$, as its kernel is all $\mathbb{R}^n$. Hence $\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1$.

To prove that $\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}_1$, note that symmetry of projection matrices and the first part of the proof give

$$\boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}_1^\top \boldsymbol{H}^\top = (\boldsymbol{H}\boldsymbol{H}_1)^\top = (\boldsymbol{H}_1)^\top = \boldsymbol{H}_1 = \boldsymbol{H}\boldsymbol{H}_1.$$

□

Definition (Non-Negative Matrix – Quadratic Form Definition)

A $p \times p$ real symmetric matrix $\boldsymbol{\Omega}$ is called non-negative definite (written $\boldsymbol{\Omega} \succeq 0$) if and only if $\boldsymbol{x}^\top \boldsymbol{\Omega} \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^p$. If $\boldsymbol{x}^\top \boldsymbol{\Omega} \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^p \setminus \{0\}$, then we call $\boldsymbol{\Omega}$ positive definite (written $\boldsymbol{\Omega} \succ 0$).

Definition (Non-Negative Matrix – Spectral Definition)

A $p \times p$ real symmetric matrix $\boldsymbol{\Omega}$ is called non-negative definite (written $\boldsymbol{\Omega} \succeq 0$) if and only the eigenvalues of $\boldsymbol{\Omega}$ are non-negative. If the eigenvalues of $\boldsymbol{\Omega}$ are strictly positive, then $\boldsymbol{\Omega}$ is called positive definite (written $\boldsymbol{\Omega} \succ 0$).

Lemma (Little exercise)

The two definitions are equivalent.

Proposition (Non-Negative and Covariance Matrices)

Let $\boldsymbol{\Omega}$ be a real symmetric matrix. Then $\boldsymbol{\Omega}$ is non-negative definite if and only if $\boldsymbol{\Omega}$ is the covariance matrix of some random vector $\boldsymbol{Y}$.

- Let $\boldsymbol{Y}$ be a random vector in $\mathbb{R}^d$ with covariance matrix $\boldsymbol{\Omega}$.
- Find direction $\boldsymbol{v}_1 \in \mathbb{S}^{d-1}$ such that the projection of $\boldsymbol{Y}$ onto $\boldsymbol{v}_1$ has maximal variance.
- For $j = 2, 3, \ldots, d$, find direction $\boldsymbol{v}_j \perp \{\boldsymbol{v}_1, ..., \boldsymbol{v}_{j-1}\}$ such that projection of $\boldsymbol{Y}$ onto $\boldsymbol{v}_j$ has maximal variance.

Solution: maximise $\text{var}(\boldsymbol{v}_1^\top \boldsymbol{Y}) = \boldsymbol{v}_1^\top \boldsymbol{\Omega} \boldsymbol{v}_1$ over $\|\boldsymbol{v}_1\| = 1$

$$\boldsymbol{v}_1^\top \boldsymbol{\Omega} \boldsymbol{v}_1 = \boldsymbol{v}_1^\top \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{v}_1 = \|\boldsymbol{\Lambda}^{1/2} \boldsymbol{U}^\top \boldsymbol{v}_1\|^2 = \sum_{i=1}^{d} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{v}_1)^2 \qquad \text{[change of basis]}$$

Now $\sum_{i=1}^d (\boldsymbol{u}_i^\top \boldsymbol{v}_1)^2 = \|\boldsymbol{v}_1\|^2 = 1$ so we have a convex combination of $\{\lambda_j\}_{j=1}^d$,

$$\sum_{i=1}^{d} p_i \lambda_i, \qquad \sum_i p_i = 1, \quad p_i \geq 0, \quad i = 1, \ldots, d.$$

But $\lambda_1 \geq \lambda_i \geq 0$ so clearly this sum is maximised when $p_1 = 1$ and $p_j = 0$ $\forall j \neq 1$, i.e. $\boldsymbol{v}_1 = \pm \boldsymbol{u}_1$.

Iteratively, $\boldsymbol{v}_j = \pm \boldsymbol{u}_j$, i.e. principal components are eigenvectors of $\boldsymbol{\Omega}$.

Theorem (Optimal (Linear) Dimension Reduction Theorem)

Let $\boldsymbol{Y}$ be a mean-zero random variable in $\mathbb{R}^d$ with $d \times d$ covariance $\boldsymbol{\Omega}$. Let $\boldsymbol{H}$ be the projection matrix onto the span of the first k eigenvectors of $\boldsymbol{\Omega}$. Then

$$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}\|^2 \leq \mathbb{E}\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{Y}\|^2$$

for any $d \times d$ projection matrix $\boldsymbol{Q}$ or rank at most k.

Intuitively: if you want to approximate a mean-zero random variable taking values $\mathbb{R}^d$ by a random variable that ranges over a subspace of dimension at most $k \leq d$, the optimal choice is the projection of the random variable onto the space spanned by its first k principal components (eigenvectors of the covariance). "Optimal" is with respect to the mean squared error.

For the proof, use lemma below (follows immediately from spectral decomposition)

Lemma

$\boldsymbol{Q}$ is a rank k projection matrix if and only if there exist orthonormal vectors $\{\boldsymbol{v}_j\}_{j=1}^k$ such that $\boldsymbol{Q} = \sum_{j=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$.

Proof of Optimal Linear Dimension Reduction ($*$).

Write $\boldsymbol{Q} = \sum_{j=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$ for some orthonormal $\{\boldsymbol{v}_j\}_{j=1}^k$. Then
$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{Y}\|^2 =$

$$\begin{aligned}
&= \mathbb{E}\left[\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{Q})^\top (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{Y}\right] = \mathbb{E}\left[\text{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{Y}\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{Q})^\top\}\right] \\
&= \text{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right](\boldsymbol{I} - \boldsymbol{Q})^\top\} = \text{tr}\{(\boldsymbol{I} - \boldsymbol{Q})^\top (\boldsymbol{I} - \boldsymbol{Q})\Omega\} \\
&= \text{tr}\{(\boldsymbol{I} - \boldsymbol{Q})\Omega\} = \text{tr}\{\boldsymbol{\Omega}\} - \text{tr}\{\boldsymbol{Q}\boldsymbol{\Omega}\} = \sum_{i=1}^d \lambda_i - \text{tr}\left\{\sum_{j=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{\Omega}\right\} \\
&= \sum_{i=1}^d \lambda_i - \sum_{j=1}^k \text{tr}\left\{\boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{\Omega}\right\} = \sum_{i=1}^d \lambda_i - \sum_{j=1}^k \boldsymbol{v}_i \boldsymbol{\Omega} \boldsymbol{v}_i^\top \\
&= \sum_{i=1}^d \lambda_i - \sum_{j=1}^k \text{var}[\boldsymbol{v}_i^\top \boldsymbol{Y}]
\end{aligned}$$

If we can minimise this expression over all $\{\boldsymbol{v}_j\}_{j=1}^k$ with $\boldsymbol{v}_i^\top \boldsymbol{v}_j = \boldsymbol{1}\{i = j\}$, then we're done. By PCA, this is done by choosing the top k eigenvectors of $\boldsymbol{\Omega}$. □

Corollary (Deterministic Version)

Let $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\} \subset \mathbb{R}^d$ be such that $\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_p = 0$, and let $\boldsymbol{X}$ be the matrix with columns $\{\boldsymbol{x}_i\}_{i=1}^p$. The best approximating k-hyperplane to the points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\}$ is given by the span of the first k eigenvectors of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$, i.e. if $\boldsymbol{H}$ is the projection onto this span, it holds that

$$\sum_{i=1}^{p} \|\boldsymbol{x}_i - \boldsymbol{H}\boldsymbol{x}_i\|^2 \leq \sum_{i=1}^{p} \|\boldsymbol{x}_i - \boldsymbol{Q}\boldsymbol{x}_i\|^2$$

for any $d \times d$ projection operator $\boldsymbol{Q}$ or rank at most k.

Proof.

Define the discrete random vector $\boldsymbol{Y}$ by $\mathbb{P}[\boldsymbol{Y} = \boldsymbol{x}_i] = 1/p$, and use optimal linear dimension reduction as stated earlier. □

Definition (Multivariate Gaussian Distribution)

A random vector $\boldsymbol{Y}$ in $\mathbb{R}^d$ has the multivariate normal distribution if and only if $\boldsymbol{\beta}^\top \boldsymbol{Y}$ has the univariate normal distribution, $\forall \boldsymbol{\beta} \in \mathbb{R}^d$.

How can we use this definition to determine basic properties?

Recall that the *moment generating function* (MGF) of a random vector $\boldsymbol{W}$ in $\mathbb{R}^d$ is defined as

$$M_{\boldsymbol{W}}(\boldsymbol{\theta}) = \mathbb{E}[e^{\boldsymbol{\theta}^\top \boldsymbol{W}}], \qquad \boldsymbol{\theta} \in \mathbb{R}^d,$$

provided the expectation exists. When the MGF exists *it characterises the distribution of the random vector. Furthermore, two random vectors are independent if and only if their joint MGF is the product of their marginal MGF's.*

Most important facts about Gaussian vectors:

1. Moment generating function of $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$:

$$\boxed{M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right).}$$

2. $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ and given $\boldsymbol{B}_{n\times p}$ and $\boldsymbol{\theta}_{n\times 1}$, then
$$\boldsymbol{\theta} + \boldsymbol{B}\,\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top).$$
3. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ density, assuming $\boldsymbol{\Omega}$ nonsingular:

$$\boxed{f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}.}$$

4. Constant density isosurfaces are ellipsoidal
5. Marginals of Gaussian are Gaussian (converse NOT true).
6. $\boldsymbol{\Omega}$ diagonal $\Leftrightarrow$ independent coordinates Y_j.
7. If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$,
$$\boldsymbol{A}\,\boldsymbol{Y} \text{ independent of } \boldsymbol{B}\,\boldsymbol{Y} \iff \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{B}^\top = 0.$$

Proposition (Property 1: Moment Generating Function)

The moment generating function of $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ *is*

$$M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right)$$

Proof (∗).

Let $\boldsymbol{u} \in \mathbb{R}^d$ be arbitrary. Then $\boldsymbol{u}^\top \boldsymbol{Y}$ is Gaussian with mean $\boldsymbol{u}^\top \boldsymbol{\mu}$ and variance $\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}$. Hence it has moment generating function:

$$M_{\boldsymbol{u}^\top \boldsymbol{Y}}(t) = \mathbb{E}\left(e^{t\boldsymbol{u}^\top \boldsymbol{Y}}\right) = \exp\left\{t(\boldsymbol{u}^\top \boldsymbol{\mu}) + \frac{t^2}{2}(\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u})\right\}.$$

Now take $t = 1$ and observe that

$$M_{\boldsymbol{u}^\top \boldsymbol{Y}}(1) = \mathbb{E}\left(e^{\boldsymbol{u}^\top \boldsymbol{Y}}\right) = M_{\boldsymbol{Y}}(u).$$

Combining the two, we conclude that

$$M_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp\left(\boldsymbol{u}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Omega} \boldsymbol{u}\right), \quad \boldsymbol{u} \in \mathbb{R}^d.$$

□

Proposition (Property 2: Affine Transformation)

For $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ *and given* $\boldsymbol{B}_{n\times p}$ *and* $\boldsymbol{\theta}_{n\times 1}$*, we have*

$$\boldsymbol{\theta} + \boldsymbol{B}\,\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)$$

Proof ($*$).

$$\begin{aligned}
M_{\boldsymbol{\theta}+\boldsymbol{B}\,\boldsymbol{Y}}(u) &= \mathbb{E}\left[\exp\{\boldsymbol{u}^\top(\boldsymbol{\theta}+\boldsymbol{B}\,\boldsymbol{Y})\}\right] = \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\}\mathbb{E}\left[\exp\{(\boldsymbol{B}^\top\boldsymbol{u})^\top\boldsymbol{Y}\}\right] \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\} M_{\boldsymbol{Y}}(\boldsymbol{B}^\top\boldsymbol{u}) \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta}\right\}\exp\left\{(\boldsymbol{B}^\top\boldsymbol{u})^\top\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\} \\
&= \exp\left\{\boldsymbol{u}^\top\boldsymbol{\theta} + \boldsymbol{u}^\top(\boldsymbol{B}\boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\} \\
&= \exp\left\{\boldsymbol{u}^\top(\boldsymbol{\theta}+\boldsymbol{B}\boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\boldsymbol{u}\right\}
\end{aligned}$$

And this last expression is the MGF of a $\mathcal{N}(\boldsymbol{\theta}+\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)$ distribution. □

Proposition (Property 3: Density Function)

Let $\boldsymbol{\Omega}_{p\times p}$ *be nonsingular. The density of* $\mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ *is*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \tfrac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}}\exp\left\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}$$

Proof ($*$).

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ be a vector of iid $\mathcal{N}(0,1)$ random variables. Then, because of independence,

(a) the density of $\boldsymbol{Z}$ is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^{p} f_{Z_i}(z_i) = \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\boldsymbol{z}^\top\boldsymbol{z}\right).$$

(b) The MGF of $\boldsymbol{Z}$ is

$$M_{\boldsymbol{Z}}(\boldsymbol{u}) = \mathbb{E}\left\{\exp\left(\sum_{i=1}^{p} u_i Z_i\right)\right\} = \prod_{i=1}^{p}\mathbb{E}\{\exp(u_i Z_i)\} = \exp(\boldsymbol{u}^\top\boldsymbol{u}/2),$$

which is the MGF of a p-variate $\mathcal{N}(0, \boldsymbol{I})$ distribution.

$\overset{(a)+(b)}{\Longrightarrow}$ the $\mathcal{N}(0, \boldsymbol{I})$ density is $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right)$.

By the spectral theorem, $\boldsymbol{\Omega}$ admits a square root, $\boldsymbol{\Omega}^{1/2}$. Furthermore, since $\boldsymbol{\Omega}$ is non-singular, so is $\boldsymbol{\Omega}^{1/2}$.

Now observe that from our Property 2, we have $\boldsymbol{Y} \overset{d}{=} \boldsymbol{\Omega}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

By the change of variables formula,

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}+\boldsymbol{\mu}}(\boldsymbol{y}) \\ &= |\boldsymbol{\Omega}^{-1/2}| f_{\boldsymbol{Z}}\{\Omega^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})\} \\ &= \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}. \end{aligned}$$

[Recall that to obtain the density of $\boldsymbol{W} = g(\boldsymbol{X})$ at $\boldsymbol{w}$, we need to evaluate $f_{\boldsymbol{X}}$ at $g^{-1}(\boldsymbol{w})$ but also multiply by the Jacobian determinant of g^{-1} at $\boldsymbol{w}$.]

□

Proposition (Property 4: Isosurfaces)

The isosurfaces of a $\mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ are $(p-1)$-dimensional ellipsoids centred at $\boldsymbol{\mu}$, with principal axes given by the eigenvectors of $\boldsymbol{\Omega}$ and with anisotropies given by the ratios of the square roots of the corresponding eigenvalues of $\boldsymbol{\Omega}$.

Proof (∗).

Exercise: Use Property 3, and the spectral theorem. □

Proposition (Property 5: Coordinate Distributions)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$. Then $Y_j \sim \mathcal{N}(\mu_j, \Omega_{jj})$.

Proof (∗).

Observe that $Y_j = (0\ ,\ 0\ , \ldots,\ \underbrace{1}_{j^{th}\text{ position}}\ , \ldots,\ 0\ ,\ 0)\,\boldsymbol{Y}$ and use Property 2. □

Proposition (Property 6: Diagonal $\boldsymbol{\Omega} \iff$ Independence)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$. Then the Y_i are mutually independent if and only if $\boldsymbol{\Omega}$ is diagonal.

Proof ($*$).

Suppose that the Y_j are independent. Property 5 yields $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for some $\sigma_j > 0$. Thus the density of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{j=1}^{p} f_{Y_j}(y_j) = \prod_{i=1}^{p} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(y_j-\mu_j)^2}{\sigma_j^2}\right\}$$

$$= \frac{1}{(2\pi)^{p/2}\,|\mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \mathrm{diag}(\sigma_1^{-2},\ldots,\sigma_p^{-2})(\boldsymbol{y}-\boldsymbol{\mu})\right\}.$$

Hence $\boldsymbol{Y} \sim \mathcal{N}\{\boldsymbol{\mu}, \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)\}$, i.e. the covariance $\boldsymbol{\Omega}$ is diagonal.

Conversely, assume $\boldsymbol{\Omega}$ is diagonal, say $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)$. Then we can reverse the steps of the first part to see that the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y})$ can be written as a product of the marginal densities $f_{Y_j}(y_j)$, thus proving independence.

□

Proposition (Property 7: $\boldsymbol{AY}, \boldsymbol{BY}$ indep $\iff \boldsymbol{A\Omega B}^\top = 0$)

If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$, and $\boldsymbol{A}_{m\times p}$, $\boldsymbol{B}_{d\times p}$ be real matrices. Then,

$$\boldsymbol{AY} \text{ independent of } \boldsymbol{BY} \iff \boldsymbol{A\Omega B}^\top = 0.$$

Proof ($*$). [wlog assuming $\boldsymbol{\mu} = 0$ (simplifies the algebra)]

First assume $\boldsymbol{A\Omega B}^\top = 0$. Let $\boldsymbol{W}_{(m+d)\times 1} = \binom{\boldsymbol{AY}}{\boldsymbol{BY}}$ and $\boldsymbol{\theta}_{(m+d)\times 1} = \binom{\boldsymbol{u}_{m\times 1}}{\boldsymbol{v}_{d\times 1}}$.

$$\begin{aligned} M_{\boldsymbol{W}}(\boldsymbol{\theta}) &= \mathbb{E}[\exp\{\boldsymbol{W}^\top\boldsymbol{\theta}\}] = \mathbb{E}\left[\exp\left\{\boldsymbol{Y}^\top\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{Y}^\top\boldsymbol{B}^\top\boldsymbol{v}\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{\boldsymbol{Y}^\top(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})\right\}\right] = M_{\boldsymbol{Y}}(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v}) \\ &= \exp\left\{\frac{1}{2}(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})^\top\boldsymbol{\Omega}(\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{B}^\top\boldsymbol{v})\right\} \end{aligned}$$

$$= \exp\left\{\tfrac{1}{2}\left(\boldsymbol{u}^\top\boldsymbol{A\Omega A}^\top\boldsymbol{u} + \boldsymbol{v}^\top\boldsymbol{B\Omega B}^\top\boldsymbol{v} + \boldsymbol{u}^\top\underbrace{\boldsymbol{A\Omega B}^\top}_{=0}\boldsymbol{v} + \boldsymbol{v}^\top\underbrace{\boldsymbol{B\Omega A}^\top}_{=0}\boldsymbol{u}\right)\right\}$$

$= M_{\boldsymbol{AY}}(\boldsymbol{u})M_{\boldsymbol{BY}}(\boldsymbol{v})$ (joint MGF = product of marginal MGFs, thus independence)

For the converse, assume that $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent. Then, $\forall \boldsymbol{u}, \boldsymbol{v}$,

$$M_{\boldsymbol{W}}(\boldsymbol{\theta}) = M_{\boldsymbol{AY}}(\boldsymbol{u}) M_{\boldsymbol{BY}}(\boldsymbol{v}), \quad \forall \boldsymbol{u}, \boldsymbol{v},$$

$$\implies \exp\left\{\frac{1}{2}\left(\boldsymbol{u}^\top \boldsymbol{A\Omega A}^\top \boldsymbol{u} + \boldsymbol{v}^\top \boldsymbol{B\Omega B}^\top \boldsymbol{v} + \boldsymbol{u}^\top \boldsymbol{A\Omega B}^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{B\Omega A}^\top \boldsymbol{u}\right)\right\}$$

$$= \exp\left\{\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{A\Omega A}^\top \boldsymbol{u}\right\} \exp\left\{\frac{1}{2}\boldsymbol{v}^\top \boldsymbol{B\Omega B}^\top \boldsymbol{v}\right\}$$

$$\implies \exp\left\{\frac{1}{2} \times 2\boldsymbol{v}^\top \boldsymbol{A\Omega B}^\top \boldsymbol{u}\right\} = 1$$

$$\implies \boldsymbol{v}^\top \boldsymbol{A\Omega B}^\top \boldsymbol{u} = 0, \qquad \forall\ \boldsymbol{u}, \boldsymbol{v},$$

$$\implies \boldsymbol{A\Omega B}^\top = 0.$$

□

Reminder:

Definition (χ^2 distribution)

Let $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_{p\times p})$. Then $\|\boldsymbol{Z}\|^2 = \sum_{j=1}^{p} Z_j^2$ is said to have the chi-square (χ^2) distribution with p degrees of freedom; we write $\|\boldsymbol{Z}\|^2 \sim \chi^2_p$.

[Thus, χ^2_p is the distribution of the sum of squares of p real independent standard Gaussian random variates.]

Definition (F distribution)

Let $V \sim \chi^2_p$ and $W \sim \chi^2_q$ be independent random variables. Then $(V/p)/(W/q)$ is said to have the F distribution with p and q degrees of freedom; we write $(V/p)/(W/q) \sim F_{p,q}$.

Proposition (Gaussian Quadratic Forms)

1. *If* $\boldsymbol{Z} \sim \mathcal{N}(0_{p\times 1}, \boldsymbol{I}_{p\times p})$ *and* $\boldsymbol{H}$ *is a projection of rank* $r \leq p$,

$$\boldsymbol{Z}^\top \boldsymbol{H}\, \boldsymbol{Z} \sim \chi^2_r.$$

2. $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Omega}_{p\times p})$ *with* $\boldsymbol{\Omega}$ *nonsingular* $\Longrightarrow$

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Gaussian Linear Regression: Likelihood and Geometry

General formulation:

$$Y_i|x_i \overset{ind}{\sim} \text{Distribution}\{g(x_i)\}, \quad i = 1, \ldots, n.$$

Simple Normal Linear Regression:

$$\begin{cases} \text{Distribution} = \mathcal{N}\{g(x), \sigma^2\} \\ g(x) = \beta_0 + \beta_1 x \end{cases}$$

Resulting Model:

$$Y_i \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\Updownarrow$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$$

Jargon: Y is *response variable* and x is *explanatory variable* (or *covariate*)
Linearity: Linearity is in the *parameters*, not the *explanatory variable*.
Example: Flexibility in what we define as explanatory:

$$Y_j = \beta_0 + \beta_1 \underbrace{\sin(x_j)}_{x_j^*} + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \text{Normal}(0, \sigma^2).$$

Example: Sometimes a transformation may be required:

$$Y_j = \beta_0 e^{\beta_1 x_j} \eta_j, \quad \eta_j \overset{iid}{\sim} \text{Lognormal}$$

$$\log(\cdot) \downarrow \qquad \uparrow \exp(\cdot)$$

$$\log Y_j = \log \beta_0 + \beta_1 x_j + \log \eta_j, \quad \log \eta_j \overset{iid}{\sim} \text{Normal}$$

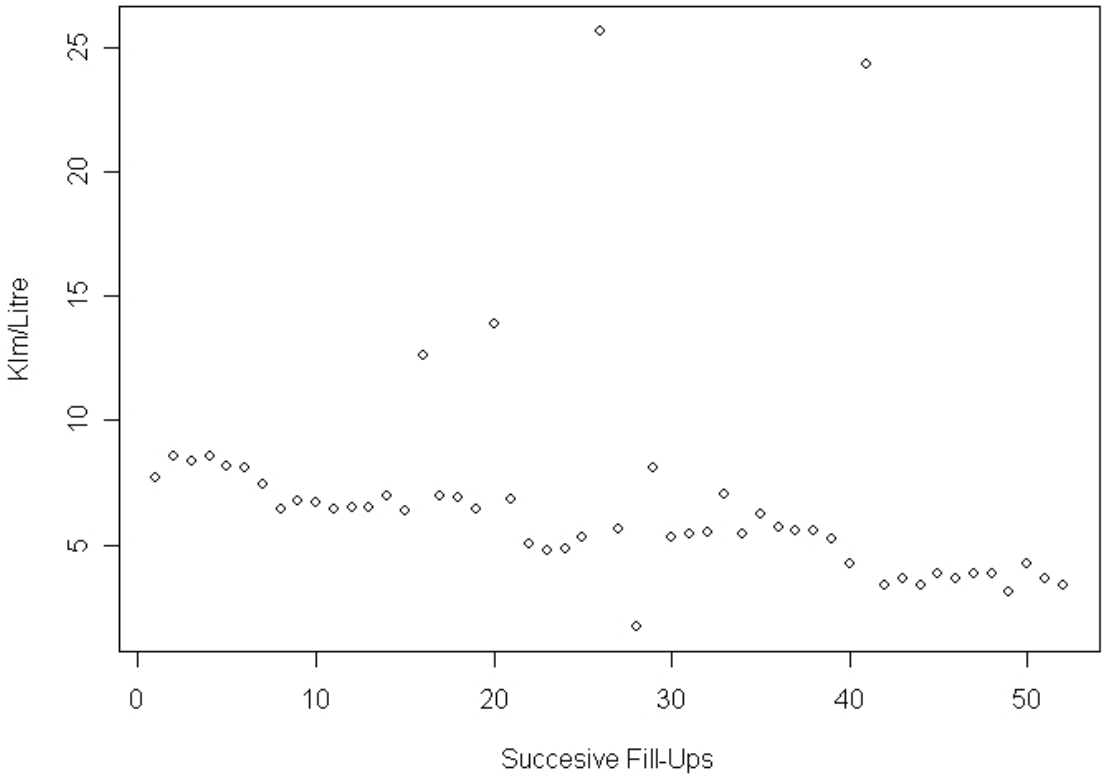
Data Structure:
For $i = 1, \ldots, n$, pairs

$$(x_i, y_i) \longrightarrow \begin{cases} x_i \text{ fixed values of } x \\ Y_i \text{ random output } Y_i \text{ when input is } x_i \end{cases}$$

Example: Professor's Van

Fillup	Km/L
1	7.72
2	8.54
3	8.35
4	8.55
5	8.16
6	8.12
7	7.46
8	6.43
9	6.74
10	6.72

Example: Professor's Van



Instead of $x_i \in \mathbb{R}$ could have $\boldsymbol{x}_i^\top \in \mathbb{R}^q$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2).$$

Letting $p = q + 1$, this can be summarised via matrix notation:

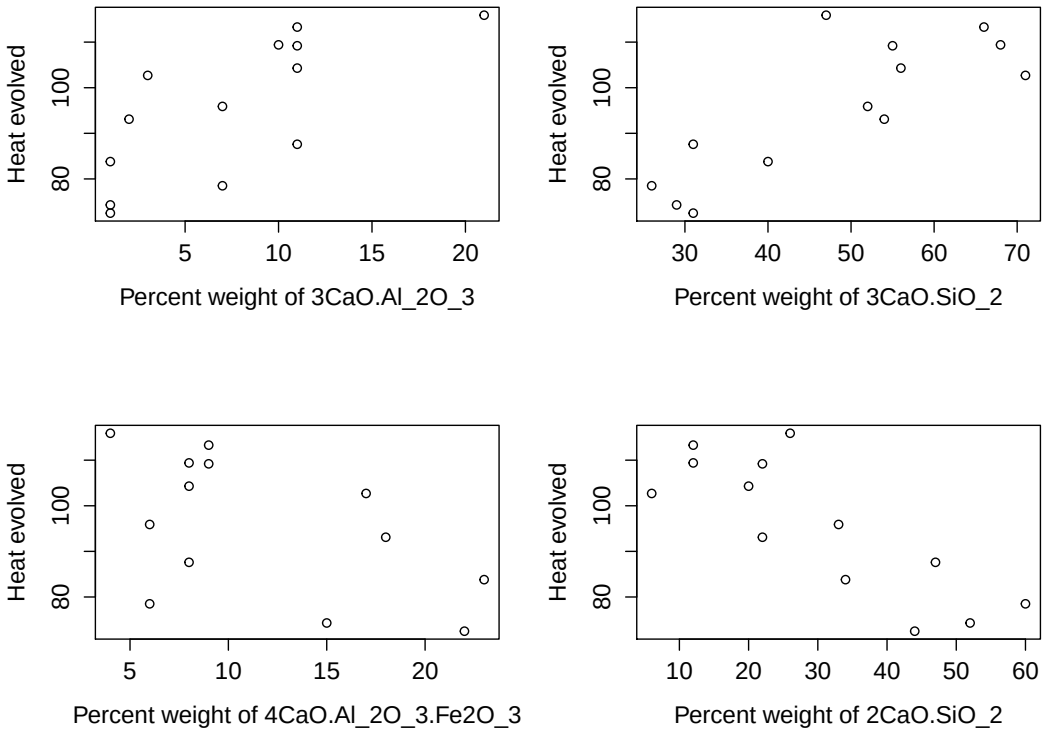
$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1q} \\ 1 & x_{21} & & x_{2q} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nq} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

$$\implies \underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p}{\boldsymbol{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

$\boldsymbol{X}$ is called the design matrix.

Example: Cement Heat Evolution

Case	$3CaO.Al_2O_3$	$3CaO.SiO_2$	$4Cao.Al_2O_3.Fe_2O_3$	$2CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40



Likelihood for Normal Linear Regression

Model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\Updownarrow$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

Observe: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ for given fixed design matrix $\boldsymbol{X}$, i.e.:

$$(Y_1, x_{11}, \ldots, x_{1q}), \ldots, (Y_i, x_{i1}, \ldots, x_{iq}), \ldots, (Y_n, x_{n1}, \ldots, x_{nq})$$

Likelihood and Loglikelihood

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\left\{n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

Whatever the value of σ, the log-likelihood is maximised when $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is minimised. Hence, the MLE of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(-\boldsymbol{X}\boldsymbol{\beta})$$

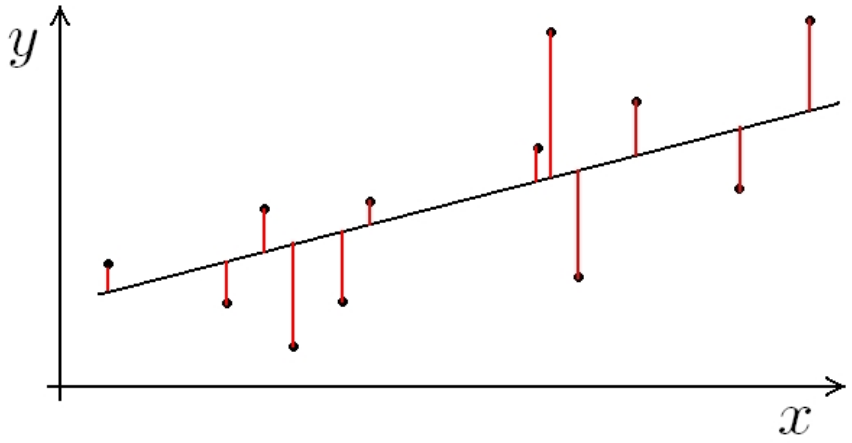
Obtain minimum by solving:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\
0 &= \frac{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})} \quad \text{(chain rule)} \\
0 &= \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{(normal equations)} \\
\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}^\top\boldsymbol{Y} \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \quad \text{(if } \boldsymbol{X} \text{ has rank } p\text{)}
\end{aligned}$$

The MLE $\hat{\boldsymbol{\beta}}$ is called the least squares estimator because it is a result of minimising

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \underbrace{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_q x_{iq})^2}_{\text{sum of squares}}.$$

Thus we are trying to find the $\boldsymbol{\beta}$ that gives the hyperplane with minimum sum of squared vertical distances from our observations.



Residuals: $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, so that $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$, with

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_q x_{iq}$$

"Regression Line" is such that $\sum e_i^2$ is minimised over all $\boldsymbol{\beta}$.

Fitted Values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, so that $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$, with

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$$

Since the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ for all values of σ^2, we have

$$\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2} \left\{ \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \sigma^2) \right\} \\
&= \arg\max_{\sigma^2} \ell(\hat{\boldsymbol{\beta}}, \sigma^2) \\
&= \arg\max_{\sigma^2} \; -\frac{1}{2} \left\{ n \log \sigma^2 + \frac{1}{\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right\}.
\end{aligned}$$

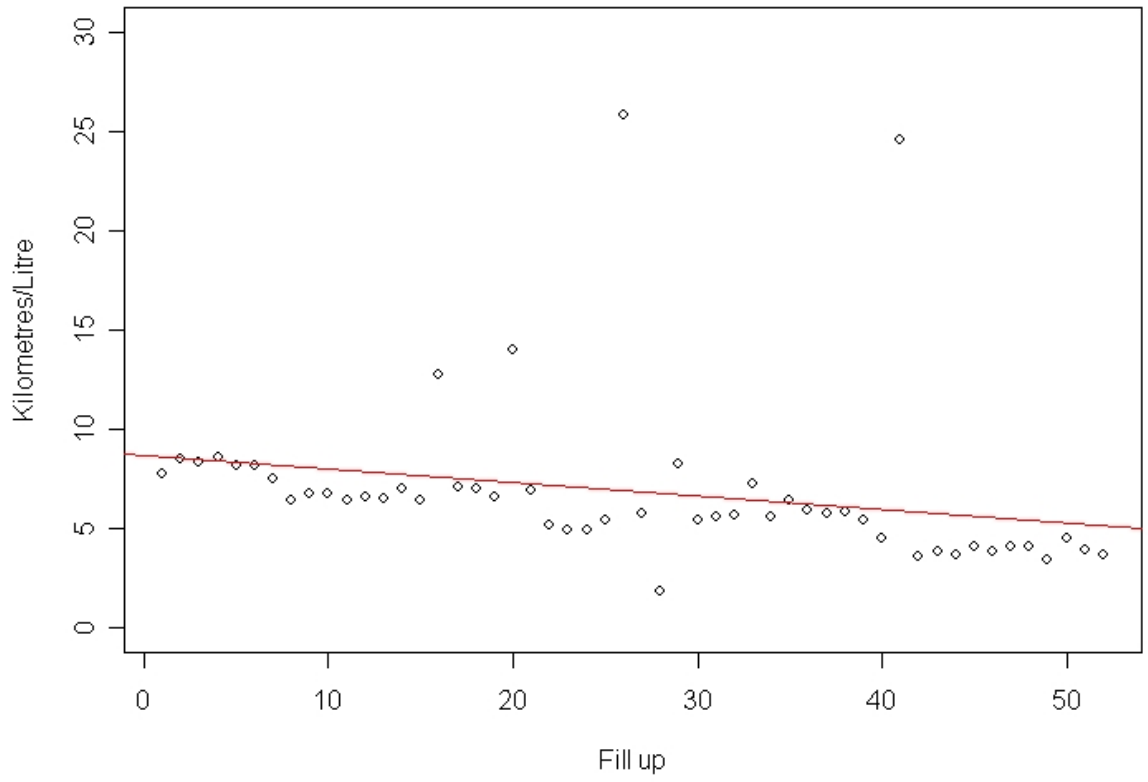
Differentiating and setting equal to zero yields

$$\hat{\sigma}^2 = \frac{1}{n} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

We will soon see that a better (unbiased) estimator is

$$S^2 = \frac{1}{n-p} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Example: Professor's Van



$$\hat{\beta}_0 = 8.6 \quad \hat{\beta}_1 = -0.068 \quad S^2 = 17.4$$

The Geometry of Least Squares

There are two dual geometrical viewpoints that one may adopt:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & & x_{2q} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{(n-1)1} & x_{(n-1)2} & \dots & x_{(n-1)q} \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Row geometry: focus on the n OBSERVATIONS
- Column geometry: focus on the p covariates

Both are useful, usually for different things:

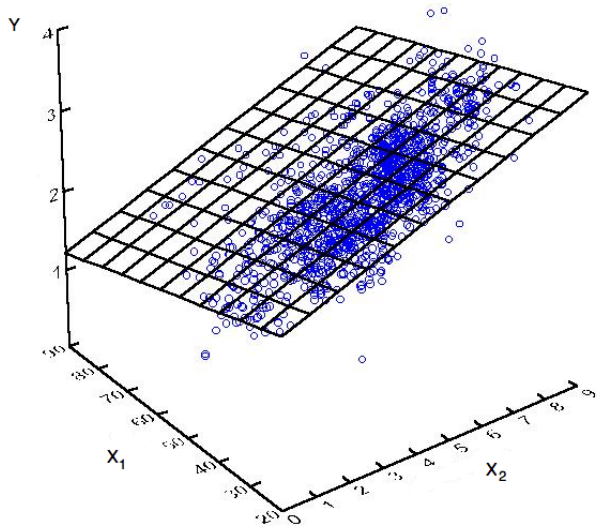
- Row geometry useful for exploratory analysis.
- Column geometry useful for theoretical analysis.

Both geometries give useful, but different, intuitive interpretations of the least squares estimators.

Row Geometry (Observations)

Corresponds to the "scatterplot geometry" – (data space)

- n points in $\mathbb{R}^p$
- each corresponds to an observation
- least squares parameters give parametric equation for a hyperplane
- hyperplane has property that it minimizes the sum of squared vertical distances of observations from the plane itself over all possible hyperplanes



- Fitted values are vertical projections (NOT orthogonal projections!) of observations onto plane, residuals are signed vertical distances of observations from plane.

Column Geometry (Variables)

Adopt the dual perspective:

- Consider the entire vector $\boldsymbol{Y}$ as a single point living in $\mathbb{R}^n$
- Then consider each variable (column of $\boldsymbol{X}$) as a point also in $\mathbb{R}^n$

What is the interpretation of the p-dimensional vector $\hat{\boldsymbol{\beta}}$, and the n-dimensional vectors $\hat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ in this dual space?

Turns out there is another important plane here: the plane spanned by the variable vectors (the column vectors of $\boldsymbol{X}$).

Recall that this is the *column space* of $\boldsymbol{X}$, denoted by $\mathcal{M}(\boldsymbol{X})$.

$$\text{Recall:} \quad \underbrace{\mathcal{M}(\boldsymbol{X})}_{\text{Column Space}} := \{\boldsymbol{X}\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^p\}$$

Q: What does $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ mean?

A: $\boldsymbol{Y}$ is [some element of $\mathcal{M}(\boldsymbol{X})$] + [Gaussian disturbance].

Any realisation of $\boldsymbol{Y}$ will lie outside $\mathcal{M}(\boldsymbol{X})$ (almost surely). MLE estimates $\boldsymbol{\beta}$ by minimising

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

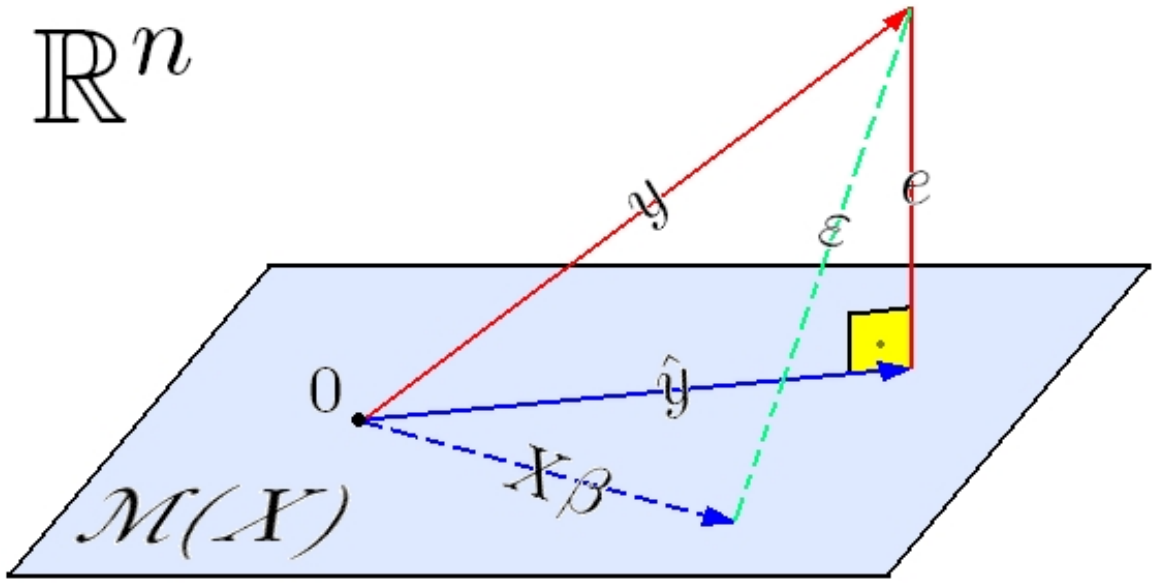
Thus we search for a $\boldsymbol{\beta}$ giving the element of $\mathcal{M}(\boldsymbol{X})$ with the minimum distance from $\boldsymbol{Y}$.
Hence $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{Y}$ onto $\mathcal{M}(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} := \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{\boldsymbol{H}} \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

$\boldsymbol{H}$ is the hat matrix (because it puts a hat on $\boldsymbol{Y}$!)

Leads to geometric derivation of the MLE of $\boldsymbol{\beta}$:

- Choose $\hat{\boldsymbol{\beta}}$ to minimise $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, so
$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$
- $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\gamma} \in \mathcal{M}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{\gamma}\|^2$
- But the unique $\boldsymbol{\gamma}$ that yields $\min_{\boldsymbol{\gamma} \in \mathcal{M}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{\gamma}\|^2$ is $\boldsymbol{\gamma} = \boldsymbol{P}\boldsymbol{Y}$.
- Here $\boldsymbol{P}$ is the projection onto the column space of $\boldsymbol{X}$, $\mathcal{M}(\boldsymbol{X})$.
- Since $\boldsymbol{X}$ is of full rank, $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.
- So $\boldsymbol{\gamma} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$
- $\hat{\boldsymbol{\beta}}$ will now be the unique (since $\boldsymbol{X}$ non-singular) vector of coordinates of $\boldsymbol{\gamma}$ with respect to the basis of columns of $\boldsymbol{X}$.
- So
$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\gamma} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y},$$
which implies that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$



Important facts that will repeatedly be made use of:

1. $e = (I - H)Y = (I - H)\varepsilon$.
2. $\hat{Y}$ and e are orthogonal, i.e. $\hat{Y}^\top e = 0$
3. Pythagoras: $Y^\top Y = \hat{Y}^\top \hat{Y} + e^\top e = Y^\top H Y + \varepsilon^\top (I - H)\varepsilon$

Derivation:

1. $e = Y - X\hat{\beta} = Y - HY = (I - H)Y = (I - H)(X\beta + \varepsilon) = (I - H)X\beta + (I - H)\varepsilon = (I - H)\varepsilon$
2. $e = Y - \hat{Y} = (I - H)Y \implies \hat{Y}^\top e = Y^\top H^\top (I - H)Y = 0$
3. $Y^\top Y = (HY + (I - H)Y)^\top (HY + (I - H)Y) = \hat{Y}^\top \hat{Y} + e^\top e + \underbrace{2YH(I - H)Y}_{=0}$.

Could also assume slightly different model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \frac{\varepsilon_i}{\sqrt{w_i}}, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2), \quad w_i > 0$$

$$\Updownarrow$$

$$Y_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}, \frac{\sigma^2}{w_i}\right).$$

With the w_j *known* weights (example: each Y_j is an average of w_j measurements).

Arises often in practice (e.g., in sample surveys), but also arises in theory (will see in GLM).

Transformation:

$$\boldsymbol{Y}^* = \boldsymbol{W}^{1/2}\boldsymbol{Y}, \quad \boldsymbol{X}^* = \boldsymbol{W}^{1/2}\boldsymbol{X}$$

with

$$\boldsymbol{W}_{n\times n} = \text{diag}(w_1, \ldots, w_n)$$

Leads to usual scenario. In this notation we obtain:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [(\boldsymbol{X}^*)^\top \boldsymbol{X}^*]^{-1}(\boldsymbol{X}^*)^\top \boldsymbol{Y}^* \\ &= (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{Y} \end{aligned}$$

Similarly:

$$S^2 = \frac{1}{n-p}\boldsymbol{Y}^\top \left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\right]\boldsymbol{Y}$$

Distribution Theory of Least Squares

Sampling Distribution of Least Squares Estimators

Gaussian Linear Model:

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$$

We have derived the estimators:

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$
- $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\| \hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2$
- $S^2 = \frac{1}{n-p}\| \hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2$

We need to study the sampling distribution of these estimators for the purpose of:

- Understanding their precision
- Building confidence intervals
- Testing hypotheses
- Comparing them to other candidate estimators
- ...

Theorem (Sampling Distribution of LSE under Gaussian Model)

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$ *and assume that* $\boldsymbol{X}$ *has full rank* $p < n$*. Then,*

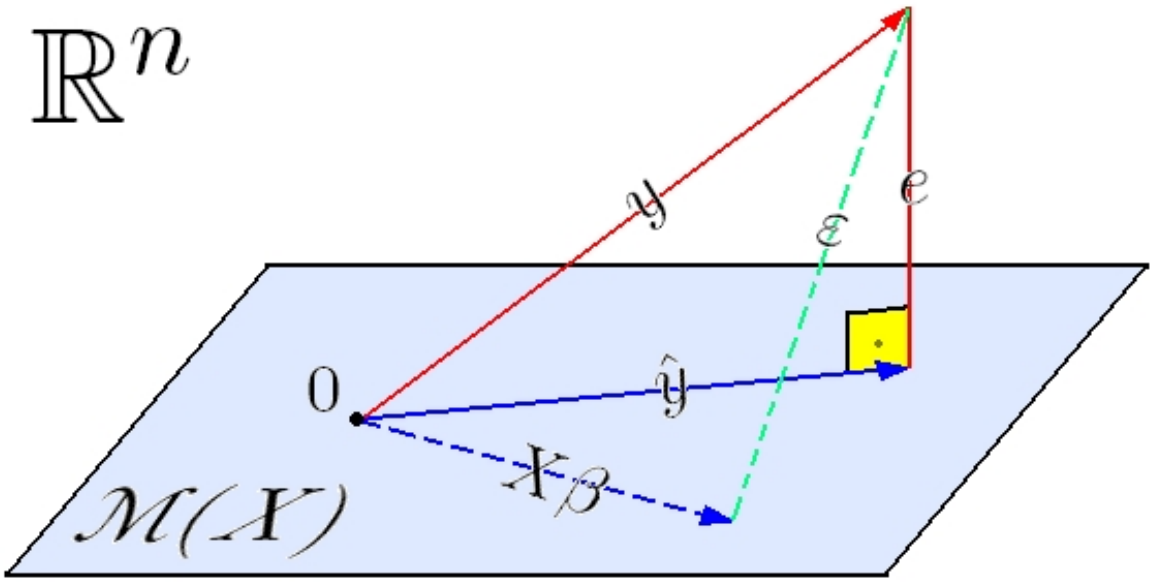
1. $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\}$;
2. *the random variables* $\hat{\boldsymbol{\beta}}$ *and* S^2 *are independent; and*
3. $\frac{n-p}{\sigma^2} S^2 \sim \chi^2_{n-p}$*, where* χ^2_ν *denotes the chi-square distribution with* ν *degrees of freedom.*

Corollary

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n})$*. The statistic* $\boldsymbol{H}\boldsymbol{Y}$ *is sufficient for the parameter* $\boldsymbol{\beta}$*. If* $\boldsymbol{X}$ *has full rank* $p < n$*, then* $\hat{\boldsymbol{\beta}}$ *is also sufficient for* $\boldsymbol{\beta}$*.*

Corollary

S^2 *is unbiased whereas* $\hat{\sigma}^2$ *is biased (so we prefer* S^2*).*



Proof.

1. Recall our results for linear transformations of Gaussian variables:
$$\left.\begin{array}{l}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} \\ \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})\end{array}\right\} \implies \hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\}$$

2. If $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, then $S^2 = \boldsymbol{e}^\top \boldsymbol{e}/(n-p)$ will be independent of $\hat{\boldsymbol{\beta}}$ (why?). Now notice that:

- $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$
- $\hat{\boldsymbol{Y}} = \boldsymbol{H}\,\boldsymbol{Y}$
- $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$

Therefore, from the properties of the Gaussian distribution $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{Y}}$ since $(\boldsymbol{I} - \boldsymbol{H})(\sigma^2 \boldsymbol{I})\boldsymbol{H} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H} = 0$, by idempotency of $\boldsymbol{H}$.

3. For the last part recall that

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon \implies (n-p)S^2 = (n-p)\tfrac{e^\top e}{n-p} = \varepsilon^\top(\boldsymbol{I} - \boldsymbol{H})\varepsilon$$

by idempotency of $\boldsymbol{H}$. But recall that $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$ so $\sigma^{-1}\varepsilon \sim \mathcal{N}_n(0, \boldsymbol{I})$. Therefore, by the properties of normal quadratic forms,

$$\tfrac{(n-p)}{\sigma^2} S^2 = (\sigma^{-1}\varepsilon)^\top(\boldsymbol{I} - \boldsymbol{H})(\sigma^{-1}\varepsilon) \sim \chi^2_{n-p}.$$

Proof of the first Corollary.

Write $\boldsymbol{Y} = \boldsymbol{H}\,\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$.

If we can show that the conditional distribution of the $2n$-dimensional vector $\boldsymbol{W} = (\hat{\boldsymbol{Y}}, \boldsymbol{e})^\top$ given $\hat{\boldsymbol{Y}}$ does not depend on $\boldsymbol{\beta}$, then we will also know that the conditional distribution of $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$ given $\hat{\boldsymbol{Y}}$ does not depend on $\boldsymbol{\beta}$ either, proving the proposition.

But we have proven that $\hat{\boldsymbol{Y}}$ is independent of $\boldsymbol{e}$. Therefore, conditional on $\hat{\boldsymbol{Y}}$, $\boldsymbol{e}$ always has the same distribution $\mathcal{N}(0, (\boldsymbol{I} - \boldsymbol{H})\sigma^2)$. It follows that, conditional on $\hat{\boldsymbol{Y}}$, the vector $\boldsymbol{W}$ has a distribution whose first n coordinates equal $\hat{\boldsymbol{Y}}$ almost surely, and whose last n coordinates are $\mathcal{N}(0, (\boldsymbol{I} - \boldsymbol{H})\sigma^2)$. Neither of those two depend on $\boldsymbol{\beta}$, and the proof is complete.

When $\boldsymbol{X}$ has full rank, $\hat{\boldsymbol{\beta}}$ is a 1-1 function of $\boldsymbol{H}\boldsymbol{y}$, and is also sufficient for $\boldsymbol{\beta}$.

□

Proof of the second Corollary.

Recall that if $Q \sim \chi^2_d$, then $\mathbb{E}[Q] = d$. □

How to construct $1-\alpha$ CI for a linear combination of the parameters, $\boldsymbol{c}^\top\boldsymbol{\beta}$?

- Have $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} \sim \mathcal{N}_1(\boldsymbol{c}^\top\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}) = \mathcal{N}_1(\boldsymbol{c}^\top\boldsymbol{\beta}, \sigma^2\delta)$
- Therefore $Q = (\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta})/(\sigma\sqrt{\delta}) \sim \mathcal{N}_1(0,1)$
- Hence $Q^2 \sim \chi^2_1$
- and Q^2 is independent of S^2 (since $\hat{\beta}$ is independent of S^2)
- while $\frac{n-p}{\sigma^2}S^2 \sim \chi^2_{n-p}$.

In conclusion:

$$\frac{\frac{Q^2}{1}}{\frac{\frac{(n-p)}{\sigma^2}S^2}{n-p}} \sim F_{1,n-p} \Rightarrow \frac{\frac{(\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}-\boldsymbol{c}^\top\boldsymbol{\beta})^2}{\sigma^2\delta}}{\frac{S^2}{\sigma^2}} = \left(\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta}}{\sqrt{S^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}}}\right)^2 \sim F_{1,n-p}$$

- But for real W, $W^2 \sim F_{1,n-p} \iff W \sim t_{n-p}$, so base CI on:

$$\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top\boldsymbol{\beta}}{\sqrt{S^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

- We obtain $(1-\alpha)\times 100\%$ CI:

$$\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2)\sqrt{S^2\boldsymbol{c}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}}.$$

- What about a $(1-\alpha)$ CI for β_r? (rth coordinate)
- Let $\boldsymbol{c}_r = (0, 0, \ldots, 0, \underset{r^{th}\text{ position}}{1}, 0, \ldots, 0)$
- Then $\beta_r = \boldsymbol{c}_r^\top\boldsymbol{\beta}$
- Therefore, base CI on

$$\frac{\boldsymbol{c}_r^\top\hat{\boldsymbol{\beta}} - \boldsymbol{c}_r^\top\boldsymbol{\beta}}{\sqrt{S^2\boldsymbol{c}_r^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{c}_r}} = \frac{\hat{\beta}_r - \beta_r}{\sqrt{S^2 v_{r,r}}} \sim t_{n-p},$$

  where $v_{r,s}$ is the r,s element of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.
- Obtain $(1-\alpha)\times 100\%$ CI:

$$\hat{\beta}_r \pm t_{n-p}(\alpha/2)\sqrt{S^2 v_{rr}}.$$

A prediction interval aims to give confidence bounds on a potential response.

- Suppose we want to predict the value of Y_+ for an $\boldsymbol{x}_+ \in \mathbb{R}^p$
- Our model predicts Y_+ by $\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}}$.
- But $Y_+ = \boldsymbol{x}_+^\top \boldsymbol{\beta} + \varepsilon_+$ so a prediction interval is different from an interval for a linear combination $\boldsymbol{c}^\top \boldsymbol{\beta}$ (extra uncertainty due to ε_+):
  - $\mathbb{E}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} + \varepsilon_+] = \boldsymbol{x}_+^\top \boldsymbol{\beta}$
  - $\text{var}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} + \varepsilon_+] = \text{var}[\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}}] + \text{var}[\varepsilon_+] = \sigma^2[\boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+ + 1]$
- Base prediction interval on:

$$\frac{\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} - Y_+}{\sqrt{S^2\{1 + \boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+\}}} \sim t_{n-p}.$$

- Obtain $(1-\alpha)$ prediction interval:

$$\boldsymbol{x}_+^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2) \sqrt{S^2\{1 + \boldsymbol{x}_+^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_+\}}.$$

The Coefficient of Determination, R^2

R^2 is a *measure of fit* of the model to the data.

- We are trying to best approximate $\boldsymbol{Y}$ through an element of the column-space of $\boldsymbol{X}$.
- How successful are we? Squared error is $\boldsymbol{e}^\top \boldsymbol{e}$.
- How large is this, relative to data variation? Look at

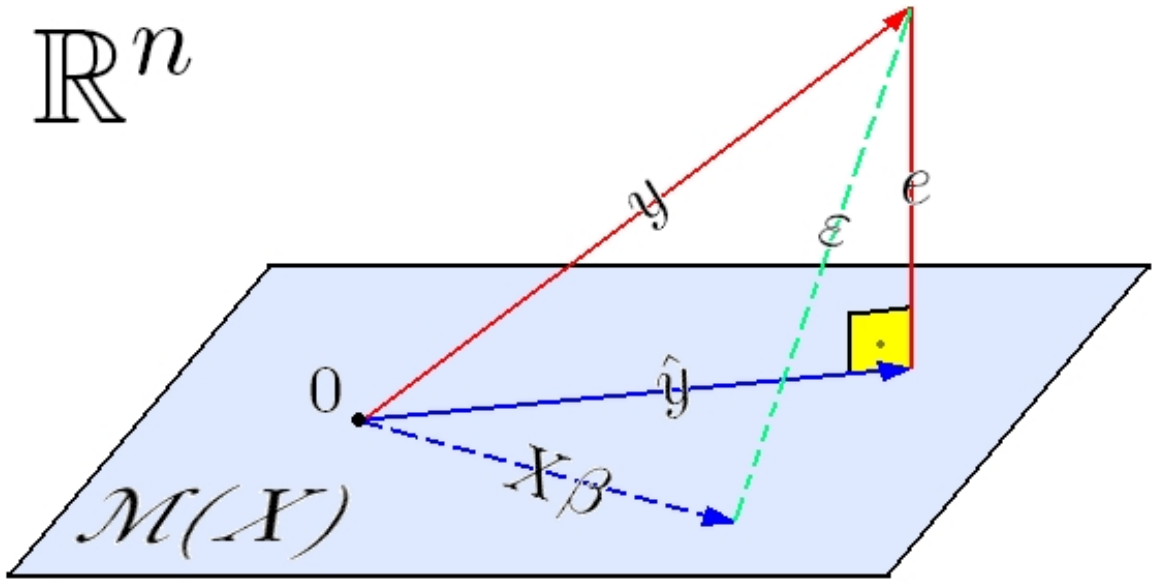
$$\frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{Y}\|^2} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = \frac{\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = 1 - \frac{\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}}}{\boldsymbol{Y}^\top \boldsymbol{Y}}$$

- Define

$$R_0^2 = \frac{\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}}}{\boldsymbol{Y}^\top \boldsymbol{Y}} = \frac{\|\hat{\boldsymbol{Y}}\|^2}{\|\boldsymbol{Y}\|^2}$$

- Note that $0 \leq R_0^2 \leq 1$

Interpretation: what proportion of the squared norm of $\boldsymbol{Y}$ does our fitted value $\hat{\boldsymbol{Y}}$ explain?



"**Centred (in fact, usual)** R^2". Compares empirical variance of $\hat{\boldsymbol{Y}}$ to empirical variance of $\boldsymbol{Y}$, instead of the empirical norms. In other words:

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n}\sum_{i=1}^n(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n \hat{Y}_i^2 - n\,\bar{Y}^2}{\sum_{i=1}^n Y_i^2 - n\,\bar{Y}^2}.$$

(note that $\frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \frac{1}{n}\sum_{i=1}^n(Y_i - e_i) = \bar{Y}$ because $\boldsymbol{e} \perp \mathbf{1}$ (here $\mathbf{1}$ is the vector of 1's = first column of design matrix $\boldsymbol{X}$) so $\sum_i e_i = 0$.

Note that

$$R^2 = \frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\bar{Y}\mathbf{1}\|^2}{\|\boldsymbol{Y}\|^2 - \|\bar{Y}\mathbf{1}\|^2}.$$

- R_0^2 mathematically more natural (does not treat first column of $\boldsymbol{X}$ as special).
- R^2 statistically more relevant (expresses variance—the first column of $\boldsymbol{X}$ usually *is* special, in statistical terms!).
- R_0^2 and R^2 may differ a lot when $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ large.

Geometrical interpretation of R^2: project $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ on orthogonal complement of $\mathbf{1}$, then compare the norms (of the projections):

- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\boldsymbol{Y} = \mathbf{1}n^{-1}\sum_{i=1}^n Y_i = \mathbf{1}\bar{Y}$.
- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\hat{\boldsymbol{Y}} = \mathbf{1}n^{-1}\sum_{i=1}^n \hat{Y}_i = \mathbf{1}\bar{Y}$.

So

$$R^2 = \frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\bar{Y}\mathbf{1}\|^2}{\|\boldsymbol{Y}\|^2 - \|\bar{Y}\mathbf{1}\|^2} = \frac{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})\,\hat{\boldsymbol{Y}}\|^2}{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})\,\boldsymbol{Y}\|^2}$$

Intuition: Should not take into account the part of $\|\boldsymbol{Y}\|$ that is explained by a constant, we only want to see the effect of the explanatory variables.

Note: Statistical packages (e.g., R) provide R^2 (and/or R^2_a, see below), not R^2_0.

Exercise: Show that $R^2 = [\text{corr}(\{\hat{Y}_i\}_{i=1}^n, \{Y_i\}_{i=1}^n)]^2$.
Exercise: Show that $R^2 \leq R^2_0$.

The adjusted R^2 takes into account the number of variables employed. It is defined as:

$$R^2_a = 1 + (1 - R^2)\frac{n-1}{n-p}.$$

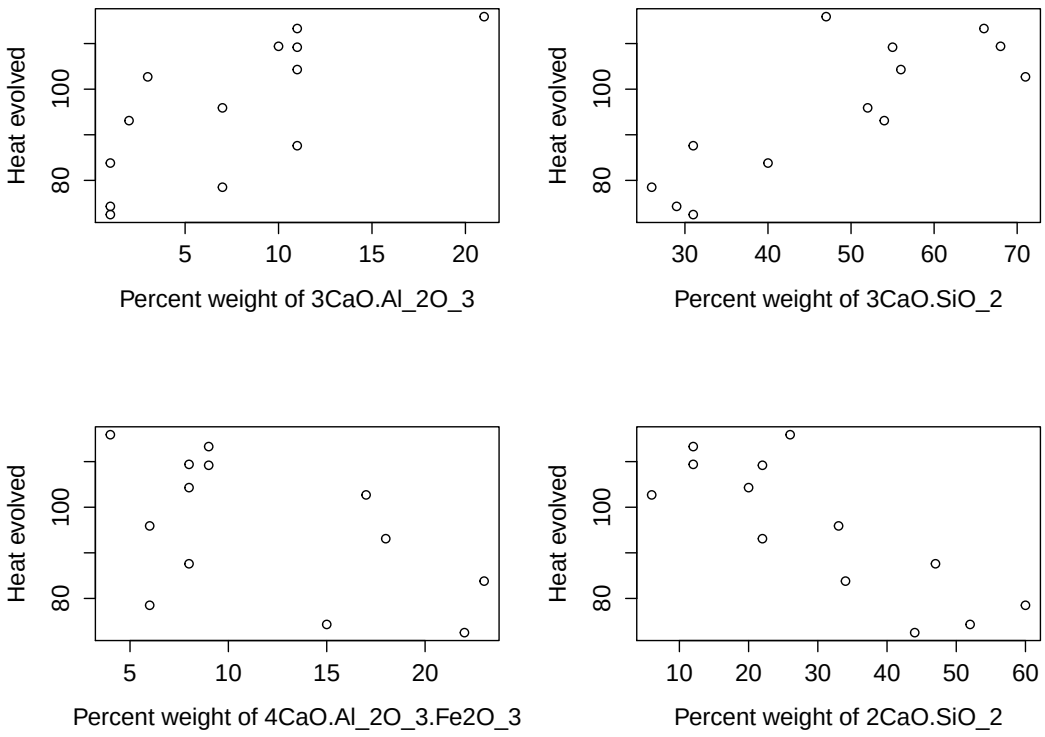
Corrects for the fact that we can always increase R^2 by adding variables. One can also correct the un-centred R^2_0 and take into account instead

$$R^2_{a0} = 1 + (1 - R^2_0)\frac{n}{n-p}.$$

Example: Cement Heat Evolution

Case	$3CaO.Al_2O_3$	$3CaO.SiO_2$	$4Cao.Al_2O_3.Fe_2O_3$	$2CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40

Example: Cement Heat Evolution



```
> cement.lm<-lm(y~1+x1+x2+x3+x4,data=cement)
> summary(cement.lm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.4054	70.0710	0.89	0.3991
x1	1.5511	0.7448	2.08	0.0708
x2	0.5102	0.7238	0.70	0.5009
x3	0.1019	0.7547	0.14	0.8959
x4	−0.1441	0.7091	−0.20	0.8441

```
Residual standard error:  2.446 on 8 degrees of freedom
R-Squared:  0.9824
```



```
> x.plus
[1] 25 25 25 25
predict(cement.lm,x.plus,interval="confidence",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	97.5	128.2

```
predict(cement.lm,x.plus,interval="prediction",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	96.5	129.2

Gauss-Markov & Optimal Estimation

Gaussian Linear Model: Efficiency of LSE (Optimality)

Q: Geometry suggests that the LSE $\hat{\boldsymbol{\beta}}$ is a sensible estimator. But is it the best we can come up with?

A: Yes, in that $\hat{\boldsymbol{\beta}}$ is the *unique minimum variance unbiased estimator* of $\boldsymbol{\beta}$.

($\hat{\boldsymbol{\beta}}$ is sufficient, in fact minimally sufficient, in exponential family)

Thus, in the Gaussian Linear model, the LSE are the best we can do as far as unbiased estimators go.

(actually can show S^2 is optimal unbiased estimator of σ^2, by similar arguments)

The crucial assumption so far was:

- **Normality**: $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$

What if we drop this strong assumption and assume something weaker? (e.g. only moment assumptions?)

- **Uncorrelatedness**: $\mathbb{E}[\varepsilon] = 0$ & $\operatorname{cov}[\varepsilon] = \sigma^2 \boldsymbol{I}$

(notice we do not assume any particular distribution.)

How well do our LSE estimators perform in this case?

(note that in this setup the observations may not be independent — uncorrelatedness implies independence only in the Gaussian case.)

For a start, we retain unbiasedness:

Lemma (Unbiasedness under Moment Assumptions)

If we only assume

$$\mathbb{E}[\varepsilon] = 0 \quad \& \quad \operatorname{var}[\varepsilon] = \sigma^2 \boldsymbol{I}$$

instead of

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}),$$

then the following remain true:

1. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$;
2. $\operatorname{cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$;
3. $\mathbb{E}[S^2] = \sigma^2$.

But what about optimality properties?

Theorem (Gauss-Markov)

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$, *with* $p < n$, $\boldsymbol{X}$ *having rank* p, *and*

- $\mathbb{E}[\varepsilon] = 0$,
- $\mathrm{cov}[\varepsilon] = \sigma^2 I$.

Then, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$ *is the best linear unbiased estimator of* $\boldsymbol{\beta}$, *that is, for any linear unbiased estimator* $\tilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$, *it holds that*

$$\mathrm{cov}(\tilde{\boldsymbol{\beta}}) - \mathrm{cov}(\hat{\boldsymbol{\beta}}) \succeq 0.$$

Proof.

Let $\tilde{\boldsymbol{\beta}}$ be linear and unbiased, in other words: $\begin{cases} \tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}, & \text{for some } \boldsymbol{A}_{p\times n}, \\ \mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}, & \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p. \end{cases}$

These two properties combine to yield,

$$\boldsymbol{\beta} = \mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\varepsilon] = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p$$

$$\implies (\boldsymbol{A}\boldsymbol{X} - \boldsymbol{I})\boldsymbol{\beta} = 0,\ \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

We conclude that the null space of $(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{I})$ is the entire $\mathbb{R}^p$, and so $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$.

$$\begin{aligned} \mathrm{cov}[\tilde{\boldsymbol{\beta}}] - \mathrm{cov}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^\top - \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= \sigma^2\{\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{A}^\top\} \\ &= \sigma^2\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{A}^\top \\ &= \sigma^2\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top\boldsymbol{A}^\top \\ &\succeq 0. \end{aligned}$$

□

If we only assume $\mathbb{E}[\varepsilon] = 0$ and $\text{cov}[\varepsilon] = \sigma^2 \boldsymbol{I}$

↪then Gauss-Markov says $\hat{\boldsymbol{\beta}}$ optimal linear unbiased estimator, regardless of distibution of ε.

Question: *What can we say about the sampling distribution of $\hat{\boldsymbol{\beta}}$ when ε is not necessarily Gaussian?*

Note that we can always write

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

- Since there is a huge variety of candidate distributions for ε that would be compatible with the property $\text{cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$, we cannot say very much about the exact distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \varepsilon$.
- Can we at least hope to say something about this distribution asymptotically, as the sample becomes large?

For this, we need an appropriate asymptotic framework for covariates:

- We let $n \to \infty$ (# rows of $\boldsymbol{X}$ tend to infinity)
- # columns of $\boldsymbol{X}$, i.e., p, (held fixed).

Theorem (Large Sample Distribution of $\hat{\boldsymbol{\beta}}$)

Let $\{\boldsymbol{X}_n\}_{n \geq 1}$ be a sequence of $n \times p$ design matrices, and $\{\varepsilon_n\}_{n \geq 1}$ a sequence of n-vectors, and define $\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \varepsilon_n$. If

1. *$\boldsymbol{X}_n$ is of full rank p for all $n \geq 1$*
2. $\max_{1 \leq i \leq n} [\boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_i] \overset{n \to \infty}{\longrightarrow} 0$,
   (where $\boldsymbol{x}_i^\top$ is the ith row of $\boldsymbol{X}_n$)
3. *$\mathbb{E}[\varepsilon_n] = 0$ and $\text{cov}[\varepsilon_n] = \sigma^2 \boldsymbol{I}_{n \times n}$ for all $n \geq 1$,*

then the least squares estimator $\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{Y}_n$ satisfies

$$(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{1/2} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\longrightarrow} \mathcal{N}_p(0, \sigma^2 \boldsymbol{I}_{p \times p}).$$

Conclusion can be interpreted as:

$$\text{for } n \text{ "large enough", } \hat{\boldsymbol{\beta}} \overset{d}{\approx} \mathcal{N}\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\}$$

- i.e. distribution of $\hat{\boldsymbol{\beta}}$ gradually becomes the same as what it would be if $\boldsymbol{\varepsilon}$ were Gaussian
- ... provided design matrix $\boldsymbol{X}$ satisfies extra condition (2).
- Can be shown equivalent to: *diagonal elements of* $\boldsymbol{H}_n = \boldsymbol{X}_n(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top$ *, say* $h_{jj}(n)$ *converge to zero uniformly in* j *as* $n \to \infty$
- Note that trace$(\boldsymbol{H}) = p$, so that the average $\sum h_{jj}(n)/n \to 0$ — the question is do all the $h_{jj}(n) \to 0$ uniformly?

Has a very clear interpretation in terms of the form of the design that we will see when we discuss the notions of leverage and influence.

To understand Condition (2), consider simple linear model

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Here, $p = 2$. Can show that

$$h_{jj}(n) = \frac{1}{n} + \frac{(t_j - \bar{t})^2}{\sum_{k=1}^n (t_k - \bar{t})^2}$$

- Suppose $t_i = i$, for $i = 1, \ldots, n$ (regular grid). Then

$$h_{jj}(n) = \frac{1}{n} + \frac{\{j - (n+1)/2\}^2}{(n^2 - n)/12}$$

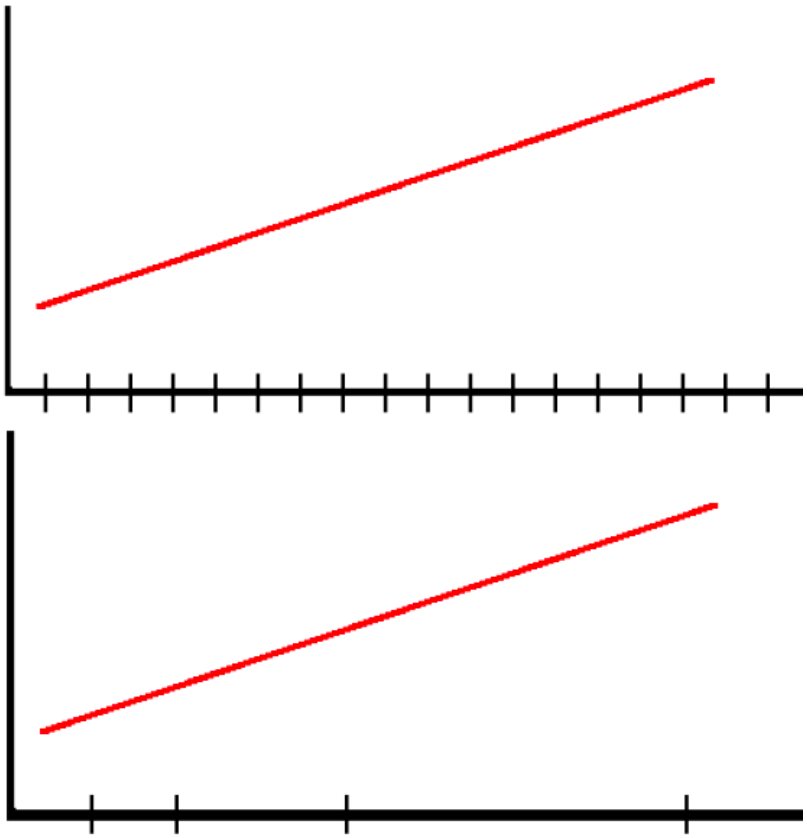
$$\text{so} \qquad \max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) = \frac{1}{n} + \frac{6(n-1)}{n(n+1)} \overset{n \to \infty}{\longrightarrow} 0.$$

- Now consider $t_i = 2^i$ (grid points spread apart as n grows).
  The centre of mass and sum of squares of the grid points is now

$$\bar{t} = \frac{2(2^n - 1)}{n}, \quad \sum_{i=1}^{n} (t_i - \bar{t})^2 = \frac{4^{n+1} - 4}{3} - \frac{4^{n+1} + 4 - 2^{n+3}}{n}$$

  and so

$$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) \overset{n \to \infty}{\longrightarrow} \frac{3}{4}.$$



Proof.

Recall that $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{\varepsilon}_n$. We will show that for any unit vector $\boldsymbol{u}$,

$$\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{X}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} N(0, \sigma^2),$$

and then the theorem will be proven by the Cramér-Wold device[a]. Now notice that

$$\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{X}_n^\top \boldsymbol{\varepsilon}_n = \boldsymbol{\gamma}_n^\top \boldsymbol{\varepsilon}_n$$

where:

1. $\boldsymbol{\gamma}_n = (\gamma_{n,1}, \ldots, \gamma_{n,n})^\top = \left(\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{x}_1, \ldots, \boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{x}_n \right)^\top$
2. $\gamma_{n,i}^2 \leq \|\boldsymbol{u}\|^2 \left\| (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{x}_i \right\|^2 = \boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_i$ (Cauchy-Schwarz)
3. $\boldsymbol{\gamma}_n^\top \boldsymbol{\gamma}_n = \boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} (\boldsymbol{X}_n^\top \boldsymbol{X}_n) (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2} \boldsymbol{u} = 1.$

Consequently, the result follows from the weighted sum CLT upon noticing:

$$\max_{1 \leq i \leq n} \gamma_{n,i}^2 \Big/ \sum_{k=1}^n \gamma_{n,k}^2 \leq \max_{1 \leq i \leq n} \boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_i \to 0$$

[a] Cramér-Wold: $\boldsymbol{\xi}_n \xrightarrow{d} \boldsymbol{\xi}$ in $\mathbb{R}^d$ if and only if $\boldsymbol{u}^\top \boldsymbol{\xi} \xrightarrow{d} \boldsymbol{u}^\top \boldsymbol{\xi}$ in $\mathbb{R}$ for all unit vectors $\boldsymbol{u}$.

Diagnostics

Four basic assumptions inherent in the Gaussian linear regression model:

- Linearity: $\mathbb{E}[Y]$ is linear in X.
- Homoskedasticity: $\text{var}[\varepsilon_j] = \sigma^2$ for all $j = 1, \ldots, n$.
- Gaussian Distribution: errors are Normally distributed.
- Uncorrelated Errors: ε_i uncorrelated with ε_j for $i \neq j$.

When one of these assumptions fails clearly, then Gaussian linear regression is inappropriate as a model for the data.

Isolated problems, such as outliers and influential observations also deserve investigation. They *may or may not* decisively affect model validity.

Scientific reasoning: impossible to *validate* model assumptions.

Cannot *prove* that the assumptions hold. Can only provide evidence in favour (or against!) them.

Strategy:

- Find implications of each assumption that we can check graphically (mostly concerning residuals).
- Construct appropriate plots and assess them (requires experience).

"Magical Thinking": Beware of overinterpreting plots!

Residuals $\boldsymbol{e}$: Basic tool for checking assumptions.

$$\text{Recall: } \boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon$$

Intuition: the residuals represent the aspects of $\boldsymbol{Y}$ that cannot be explained by the columns of $\boldsymbol{X}$.

Since $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$, if the model is correct we should have $\boldsymbol{e} \sim \mathcal{N}_n\{0, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\}$. (if model correct, residuals are ancillary)

$$\text{So if assumptions hold} \rightarrow \begin{cases} e_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\} \\ \operatorname{cov}(e_i, e_j) = -\sigma^2 h_{ij} \end{cases}$$

Note the residuals are correlated, and that they have unequal variances. Define the *standardised residuals*:

$$r_i := \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n.$$

These are still correlated but have variance ≈ 1.
(can decorrelate by $\boldsymbol{U}^\top \boldsymbol{e}$, where $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$) – why?

A first impression can be drawn by looking at plots of the response against each of the explanatory variables.
Other plots to look at?
Notice that under the assumption of linearity we have

$$\boldsymbol{X}^\top \boldsymbol{e} = 0.$$

Hence, no correlation should appear between explanatory variables and residuals.

- Plot standardised residuals $\boldsymbol{r}$ against each covariate (columns of $\boldsymbol{X}$).
  - ↪ No systematic patterns should appear in these plots. A systematic pattern would suggest incorrect dependence of the response on the particular explanatory (e.g. need to add a transformation of that explanatory as an additional variable).
- Plot standardised residuals $\boldsymbol{r}$ against covariates left out of the model.
  - ↪ No systematic patterns should appear in these plots. A systematic pattern suggests that we have left out an explanatory variable that should have been included.

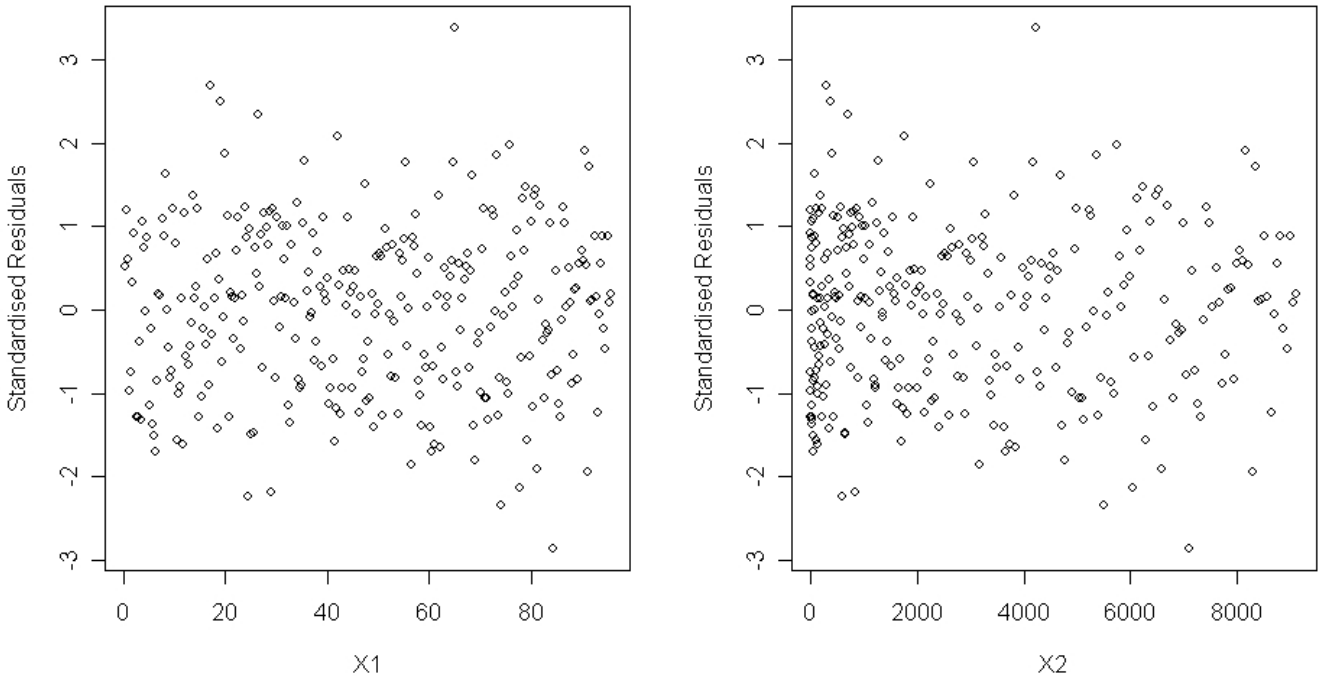


Figure: Linearity OK

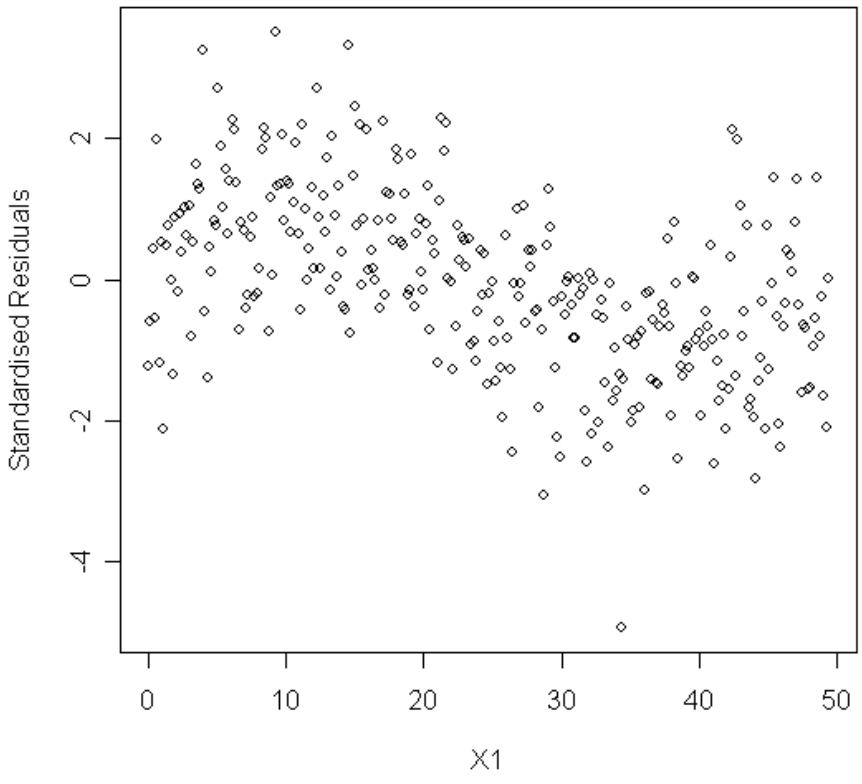


Figure: Linearity NOT OK – need to add $\sin(x_1)$ in model

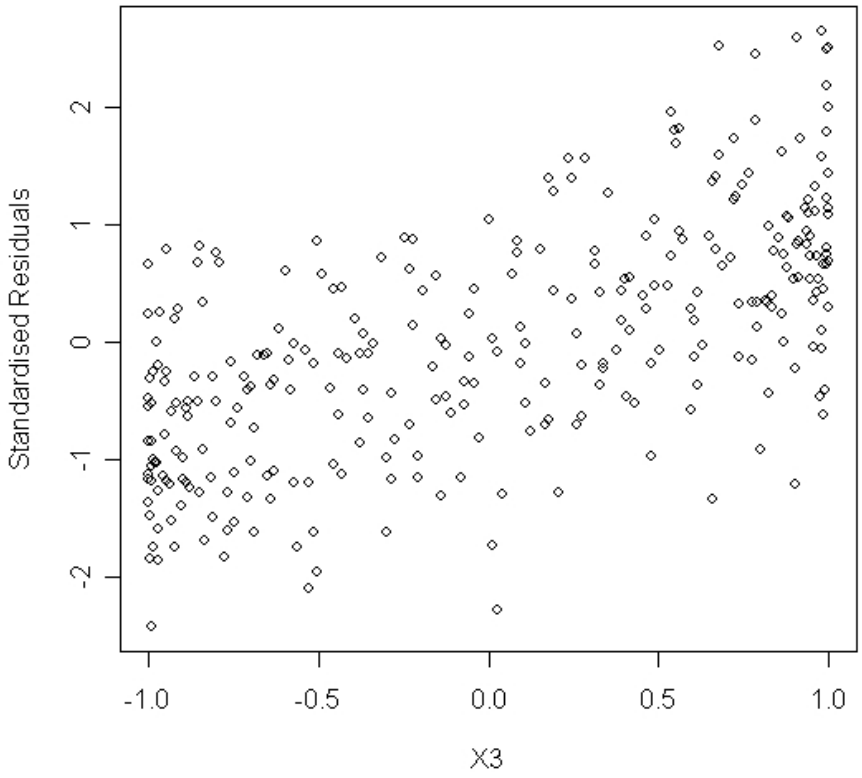


Figure: Important Covariate Left out

Checking for Homoskedasticity

$$\text{Homoskedastic} = \underbrace{\acute{o}\mu o}_{same} + \underbrace{\sigma\kappa\varepsilon\delta\alpha\sigma\mu\grave{o}\varsigma}_{spread}$$

According to our model assumptions, the variance of the errors ε_j should be the same across indices:

$$\text{var}(\varepsilon_j) = \sigma^2$$

- Plot $\boldsymbol{r}$ against the fitted values $\hat{\boldsymbol{Y}}$. (why not against $\boldsymbol{Y}$?)
  - ↪ A random scatter should appear, with approximately constant spread of the values of r for the different values of $\hat{\boldsymbol{Y}}$. "Trumpet" or "bulging" effects indicate failure of the homoskedasticity assumption.
  - ↪ Since $\hat{\boldsymbol{Y}}^\top \boldsymbol{e} = 0$, this plot can also be used to check linearity, as before.

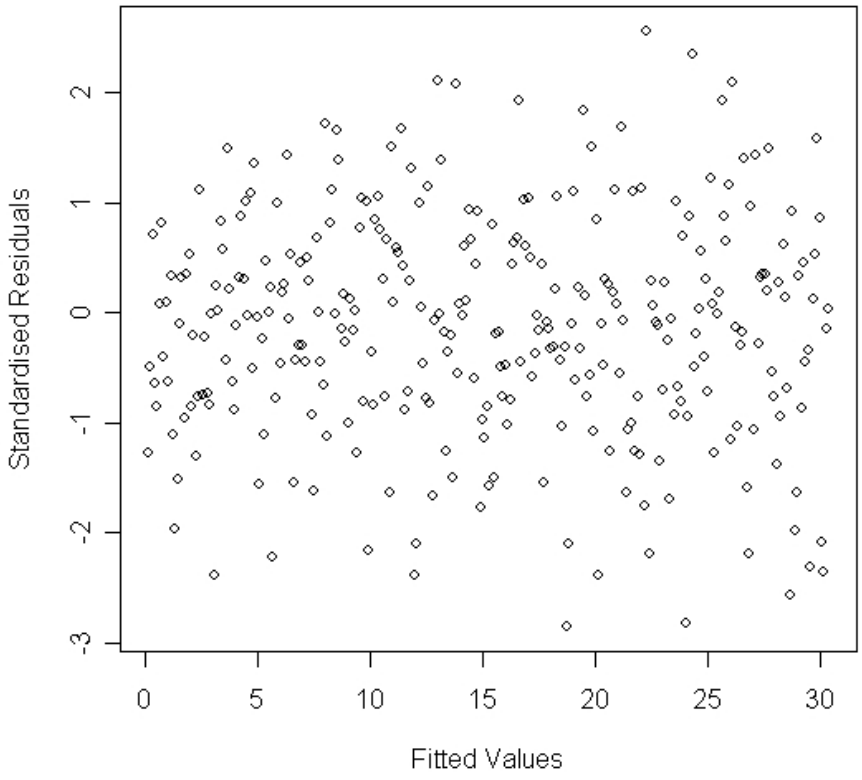


Figure: Homoskedasticity OK

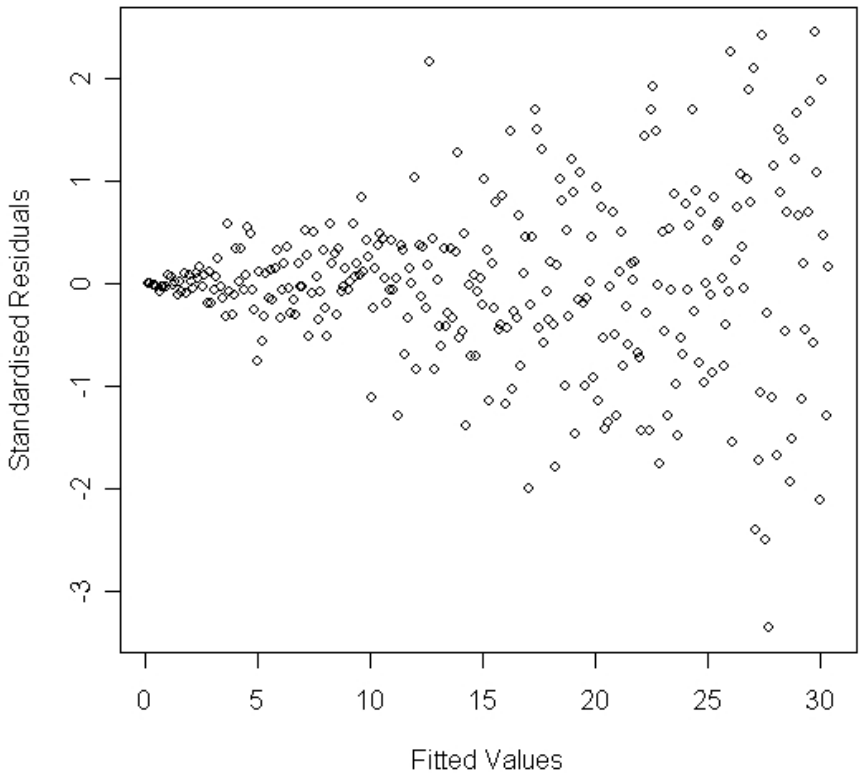


Figure: Heteroskedasticity (i.e. lack of Homoskedasticity)

Idea: compare the distribution of standardised residuals against a Normal distribution.

How?

Compare empirical vs theoretical quantiles ...

Reminer: The α-*quantile* ($\alpha \in [0,1]$) of a distribution F is the value $F^-(\alpha)$ defined as

$$F^-(\alpha) := \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}.$$

Given a sample $W_1, \ldots, W_n$, the *empirical* α *quantile* is the value defined as

$$\hat{F}^-(\alpha) := \inf\{t \in \mathbb{R} : \hat{F}(t) \geq \alpha\} = \inf\left\{t \in \mathbb{R} : \frac{\#\{W_i \leq t\}}{n} \geq \alpha\right\}.$$

where $\hat{F}$ is the empirical distribution function (as defined before).

A quantile plot for a given sample plots certain empirical quantiles against the corresponding theoretical quantiles (i.e. those under the assumed distribution).

If the sample at hand originates from F, then we expect that the points of the plot fall close to the 45° line.

- Plot the empirical $\{k/n\}_{k=1}^n$ quantiles of standardised residuals

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$$

against theoretical quantiles $\Phi^{-1}\{1/(n+1)\}, \ldots, \Phi^{-1}\{n/(n+1)\}$ of a $\mathcal{N}(0,1)$ distribution.

  - ↪ Think why we pick $\Phi^{-1}\left(\frac{k}{n+1}\right)$ instead of $\Phi^{-1}\left(\frac{k}{n}\right)$.
  - ↪ If the points of the quantile plot deviate significantly from the 45° line, there is evidence against the normality assumption. Outliers, skewness and heavy tails easily revealed.
  - ↪ If we plot the empirical quantiles of the unstandardised residuals against those of a $N(0,1)$, then we compare against a line with slope equal to stdev(e) and intercept zero.

Beware of overinterpretation when n is small!

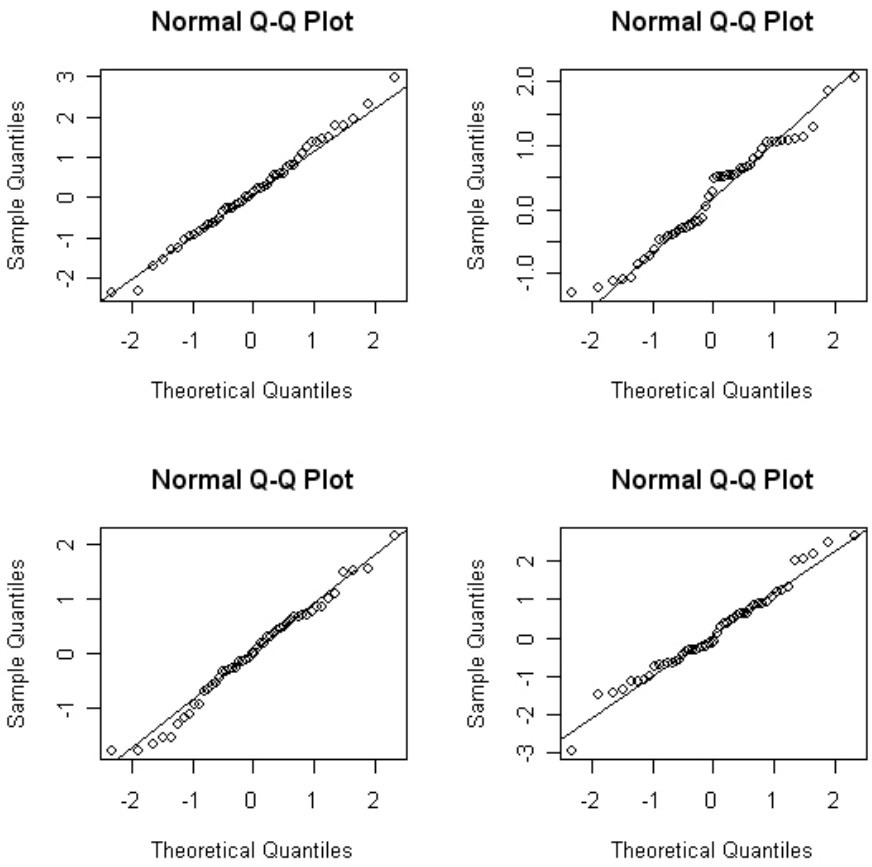


Figure: QQ Plot for $n = 50$

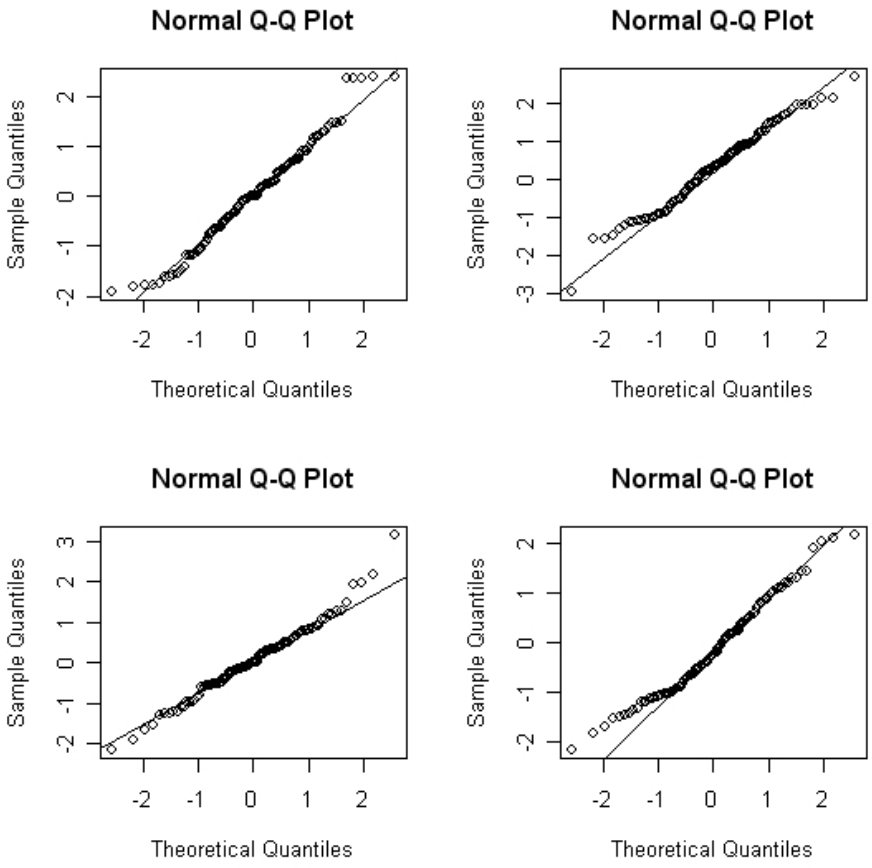


Figure: QQ Plot for $n = 100$

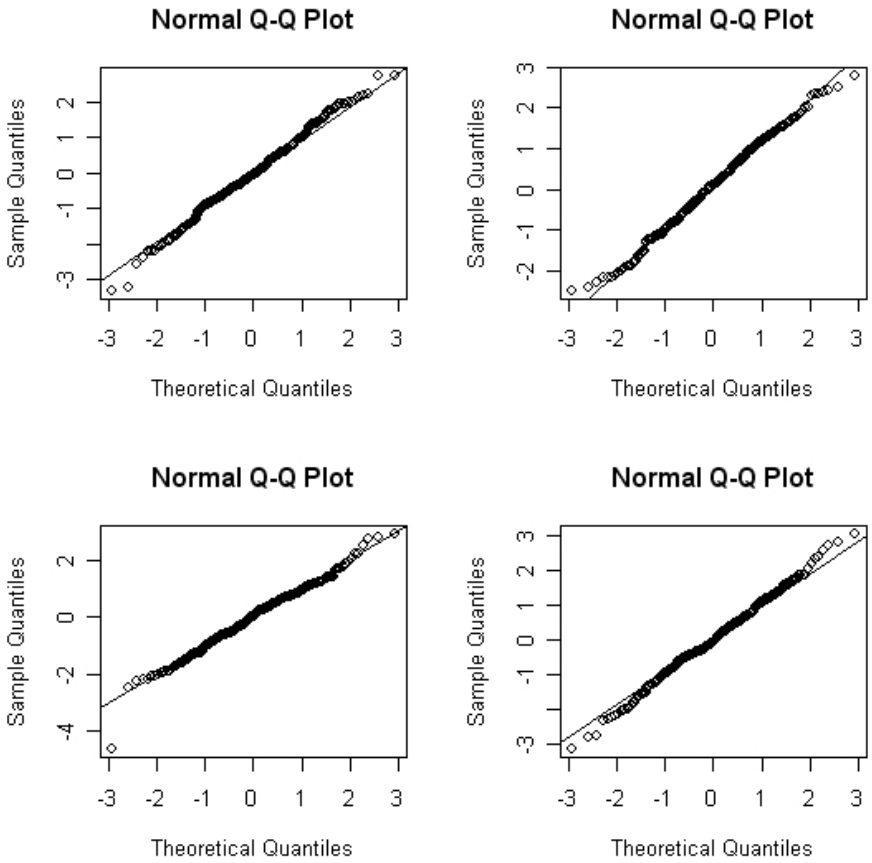


Figure: QQ Plot for $n = 300$

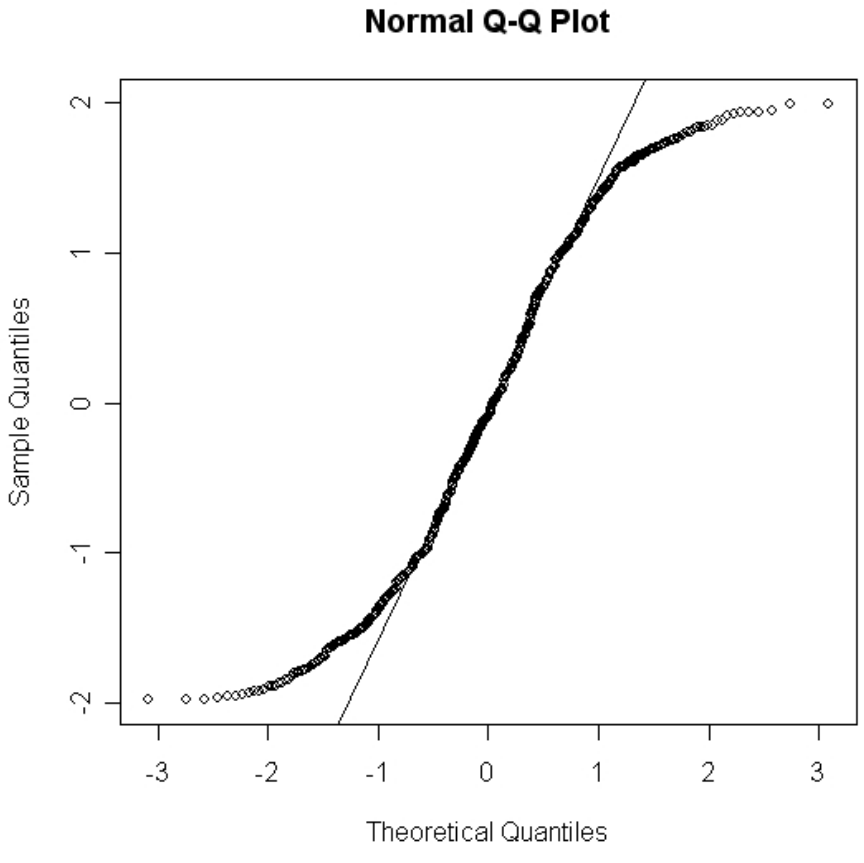


Figure: Normality not OK

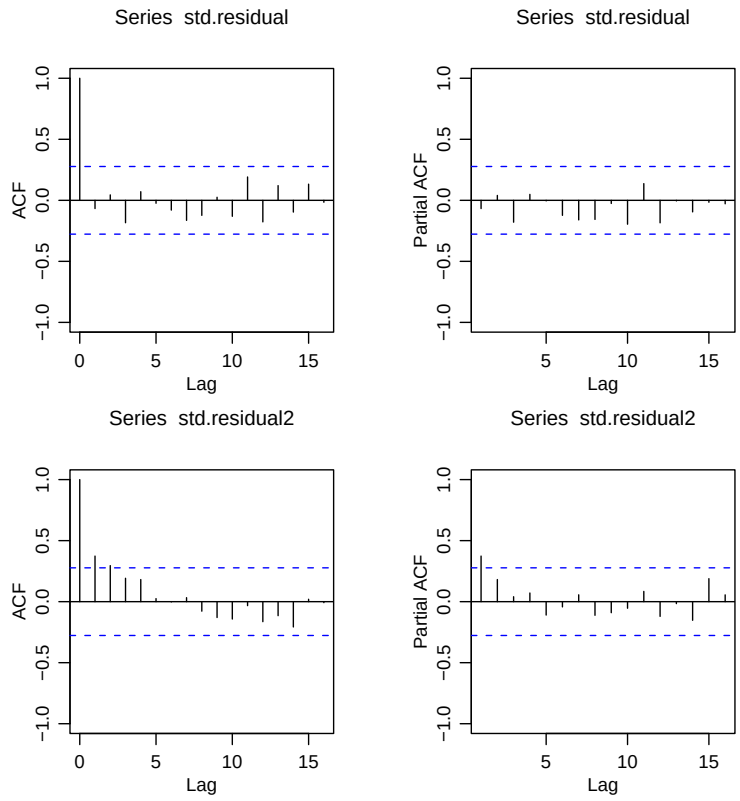
- It is assumed that $\text{cov}[\varepsilon] = \sigma^2 \boldsymbol{I}$.
- Under assumption of normality this is equivalent to independence

Difficult to check this assumption in practice.

- One thing to check for is clustering, which may suggest dependence.
  ↪ e.g. identifying groups of related individuals with correlated responses
- When observations are time-ordered can look at correlation $\text{corr}[r_t, r_{t+k}]$ or partial correlation $\text{corr}[r_t, r_{t+k} | r_{t+1}, \ldots, r_{t+k-1}]$. When such correlations exist, we enter the domain of *time series*.

Existence of correlation:

- seriously affects estimator reliability
- inflates standard errors



An influential observation can usually be categorised as an:

- outlier (relatively easier to spot by eye)

  OR

- leverage point (not as easy to spot by eye)

Influential observations

- May or may not decisively affect model validity.
- Require scrutiny on an individual basis and consultation with the data expert.

David Brillinger (Berkeley): *You will not find your Nobel prize in the fit, you will find it in the outliers!*

Influential observations may reveal unanticipated aspects of the scientific problem that are worth studying, and so must not simply be scorned as "non-conformists"!

An *outlier* is an observation that stands out in some way from the rest of the observations, causing *Surprise*! Exact mathematical definition exists (Tukey) but we will not pursue it.

- In regression, outliers are points falling far from the cloud surrounding the regression line (or surface).
- They have the effect of "pulling" the regression line (surface) toward them.

Outliers can be checked for visually through:

- The regression scatterplot.
  - ↪ Points that can be seen to fall relatively far from the point cloud surrounding the regression line (surface)
- Residual Plots.
  - ↪ Points that fall beyond $(-2, 2)$ in the $(\hat{\boldsymbol{Y}}, \boldsymbol{r})$ plot.

Outliers may result from a data registration error, or a single extreme event. They can, however, result because of a deeper inadequacy of our model (especially if there are many!).

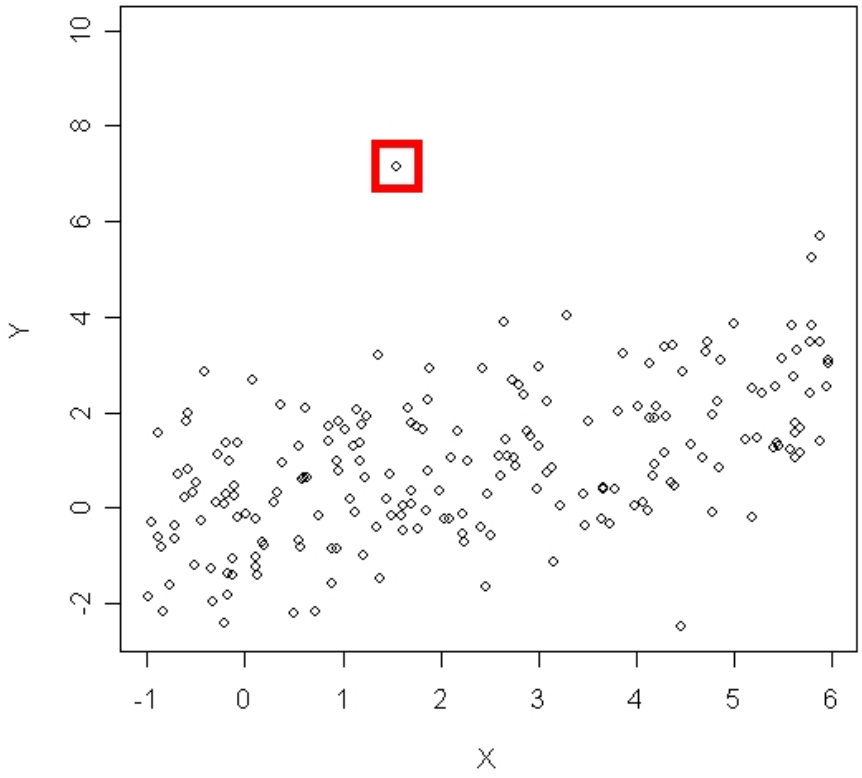


Figure: An Outlier

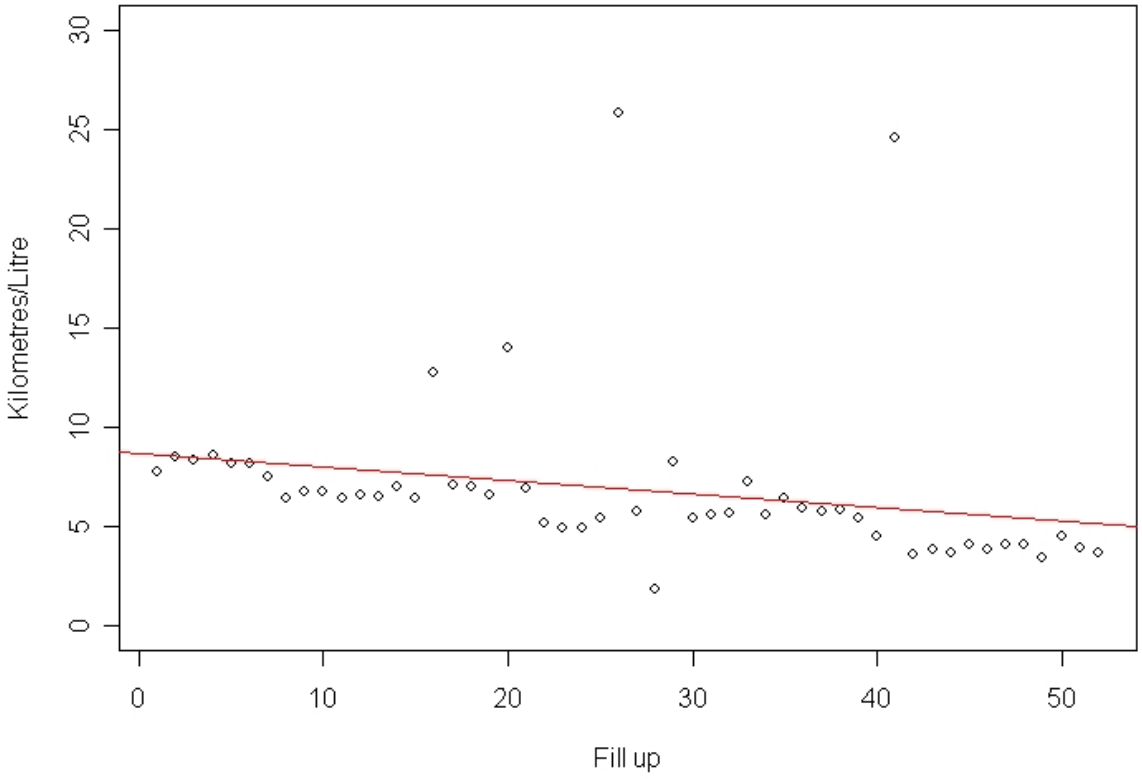


Figure: Professor's Van: Outliers

- Outliers may be influential: they "stand out" in the "y-dimension".
- However an observation may also be influential because of unusual values in the "x-dimension".
- Such influential observations cannot be so easily detected through plots.

Call $(\boldsymbol{x}_j^\top, Y_j)$ the *j-th case* and notice that

$$\mathrm{var}(Y_j - \hat{Y}_j) = \mathrm{var}(e_j) = \sigma^2(1 - h_{jj}).$$

If $h_{jj} \approx 1$, then the model is constrained so $\hat{Y}_j = \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} \simeq Y_j$! (i.e., need a separate parameter entirely devoted to fitting this observation!)

- h_{jj} is called the leverage of the j-th case.
- since $\mathrm{trace}(\boldsymbol{H}) = \sum_{j=1}^n h_{jj} = p$, cannot have low leverage for all cases
- a good (=balanced) design corresponds to $h_{jj} \simeq p/n$ for all j

  (i.e. assumption $\max_{j \leq n} h_{jj} \overset{n\to\infty}{\to} 0$ satisfied in asymptotic thm).

Leverage point: (rule of thumb) if $h_{jj} > 2p/n$ observation needs further scrutiny—e.g., fitting again without j-th case and studying effect.
Outlier+Leverage Point = TROUBLE

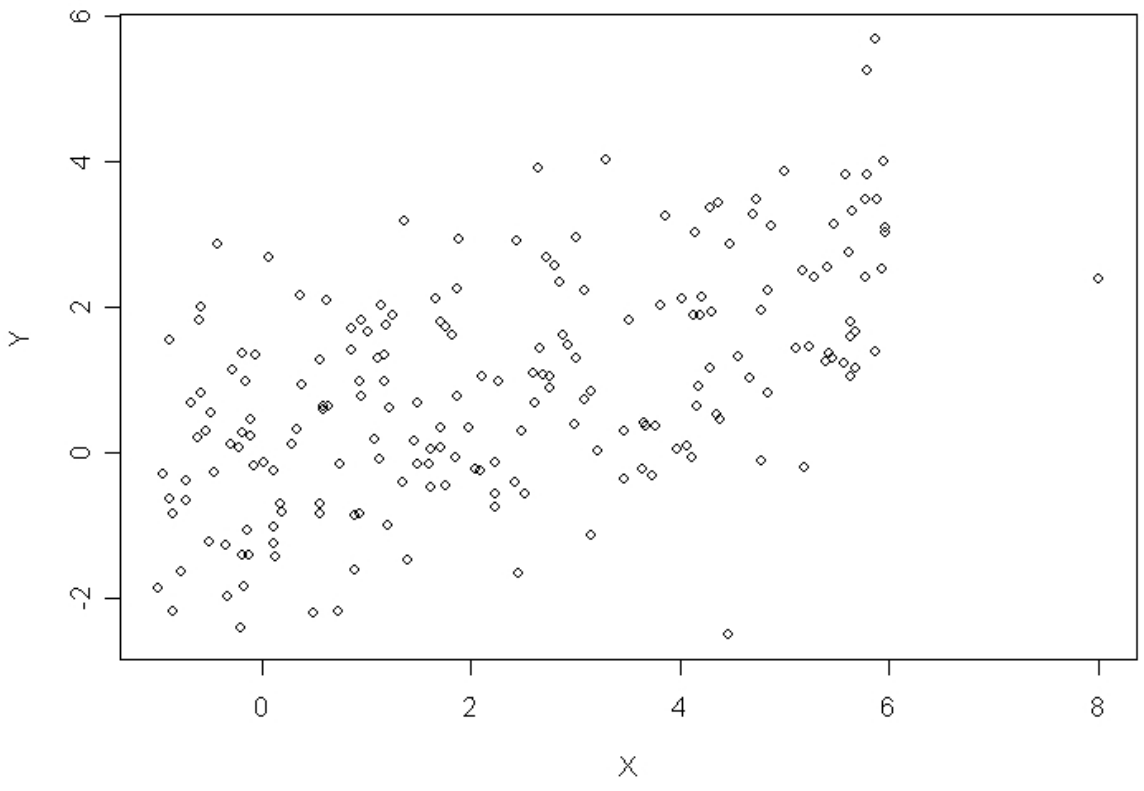


Figure: A (very) Noticeable Leverage Point

- How to find cases having strong effect on fitted model?
- Idea: see effect when case j, i.e., $(\boldsymbol{x}_j^\top, Y_j)$, is dropped.
- Let $\hat{\boldsymbol{\beta}}_{-j}$ be the LSE when model is fitted to data without case j, and let $\hat{\boldsymbol{Y}}_{-j} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-j}$ be the corresponding fitted value.
- Define *Cook's distance*

$$C_j = \frac{1}{ps^2}(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{-j})^\top(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{-j}),$$

  which measures scaled distance between $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_{-j}$.
- Can show that

$$C_j = \frac{r_j^2 h_{jj}}{p(1-h_{jj})},$$

  so large C_j implies large r_j and/or large h_{jj}.
- Cases with $C_j > 8/(n-2p)$ worth a closer look (rule of thumb)
- Plot C_j against index $j = 1, \ldots, n$ and compare with $8/(n-2p)$ level.

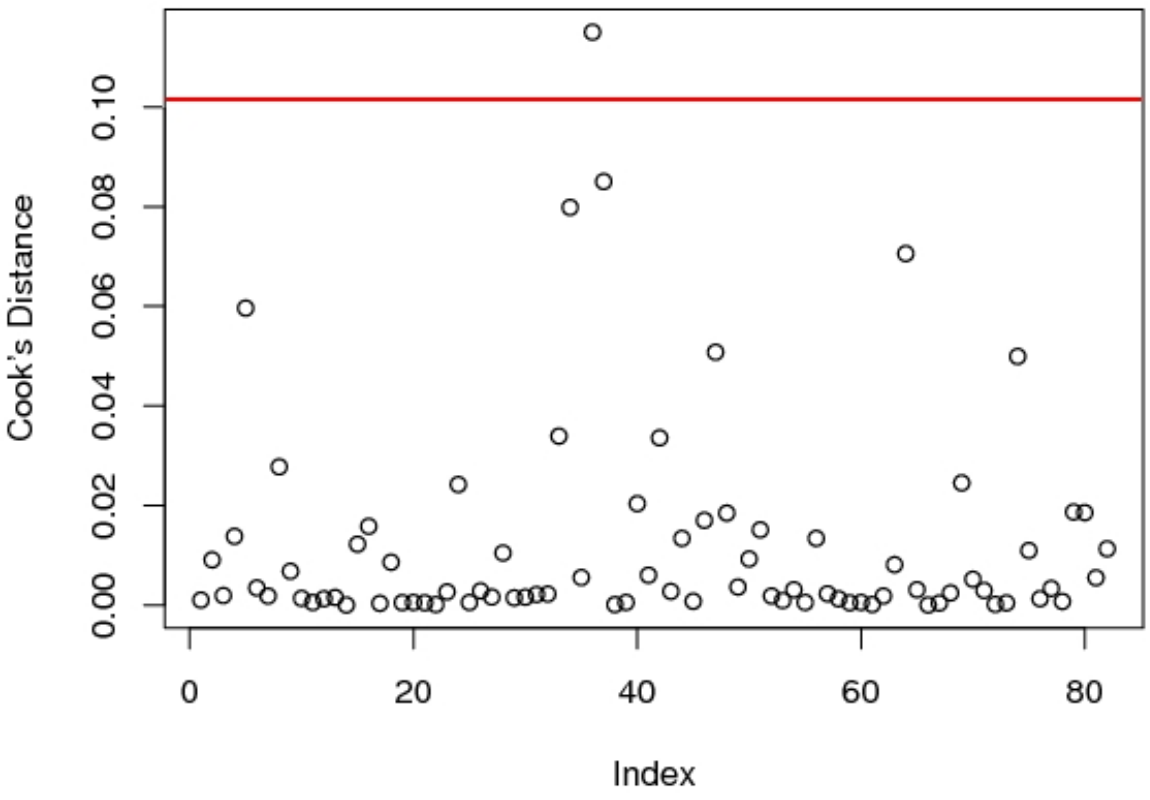


Figure: A Cook Distance Plot

Diagnostic plots usually constructed:

- $\boldsymbol{Y}$ against columns of $\boldsymbol{X}$
  - ↪ check for linearity and outliers
- standardized residual $\boldsymbol{r}$ against columns of $\boldsymbol{X}$
  - ↪ check for linearity
- $\boldsymbol{r}$ against covariates not included
  - ↪ check for variables left out
- $\boldsymbol{r}$ against fitted value $\hat{\boldsymbol{Y}}$
  - ↪ check for homoskedasticity
- Normal quantile plot
  - ↪ check for normality
- Cook's distance plot
  - ↪ check for influential observations

Nested Model Selection & ANOVA

Consider the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

This will always have higher R^2 than the sub-model:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

- Why? (think of geometry...)
- The question is: is the first model *significantly* better than the second one?
  - ↪ i.e. does the first model explain the variation adequately enough, or should we incorporate extra explanatory variables? Need a quantitative answer.

Model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$. Estimator:

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

Interpretation: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\,\boldsymbol{Y}$ is the projection of $\boldsymbol{Y}$ into the column space of $\boldsymbol{X}$, $\mathcal{M}(\boldsymbol{X})$. This subspace has dimension p, when $\boldsymbol{X}$ is of full column rank p. Now for $q < p$ write $\boldsymbol{X}$ in block notation as

$$\boldsymbol{X} = \underset{n\times q}{\boldsymbol{X}_1} \quad \underset{n\times(p-q)}{\boldsymbol{X}_2}.$$

Interpretation: $\boldsymbol{X}_1$ is built by the first q columns of $\boldsymbol{X}$ and $\boldsymbol{X}_2$ by the rest. Similarly write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1\ \boldsymbol{\beta}_2)^\top$ so that:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon = (\boldsymbol{X}_1\ \boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \varepsilon = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon.$$

Our question can now be stated as:

- Is $\boldsymbol{\beta}_2 = 0$?

Let $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$, and $\hat{\boldsymbol{Y}}_1 = \boldsymbol{H}_1 \boldsymbol{Y}$, $\boldsymbol{e}_1 = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_1$.
Pythagoras tells us that:

$$\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_1\|^2}_{RSS(\hat{\boldsymbol{\beta}}_1)=\|\boldsymbol{e}_1\|^2} = \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}_{RSS(\hat{\boldsymbol{\beta}})=\|\boldsymbol{e}\|^2} + \underbrace{\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_1\|^2}_{RSS(\hat{\boldsymbol{\beta}}_1)-RSS(\hat{\boldsymbol{\beta}})=\|\boldsymbol{e}-\boldsymbol{e}_1\|^2}$$

Notice that $RSS(\hat{\boldsymbol{\beta}}_1) \geq RSS(\hat{\boldsymbol{\beta}})$ always (think why!)

So the idea is simple: to see if it is worthwhile to include $\boldsymbol{\beta}_2$ we will compare how much larger $RSS(\hat{\boldsymbol{\beta}}_1)$ is compared to $RSS(\hat{\boldsymbol{\beta}})$.

- Equivalently, we can look at a ratio like $\{RSS(\hat{\boldsymbol{\beta}}_1) - RSS(\hat{\boldsymbol{\beta}})\}/RSS(\hat{\boldsymbol{\beta}})$
- This is in fact the likelihood ratio test statistic for our hypothesis.
- To construct a test based on this quantity, we need to figure its sampling distributions . . .

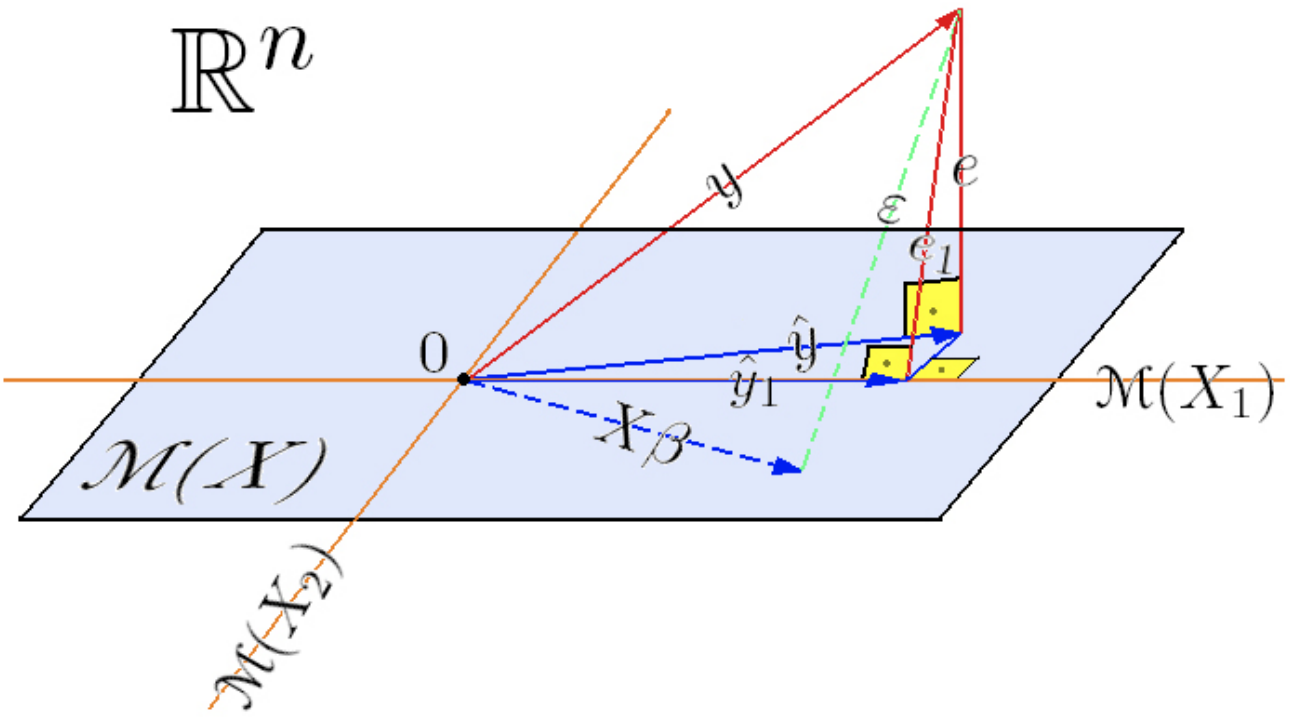


Figure: Geometry Revisited

Theorem (Sampling Distributions for Sums of Squares)

We have the following properties:

(A) $\boldsymbol{e} - \boldsymbol{e}_1 \perp \boldsymbol{e}$;

(B) $\|\boldsymbol{e}\|^2 = RSS(\hat{\boldsymbol{\beta}})$ *and* $\|\boldsymbol{e}_1 - \boldsymbol{e}\|^2 = RSS(\hat{\boldsymbol{\beta}}_1) - RSS(\hat{\boldsymbol{\beta}})$ *are independent;*

(C) $\|\boldsymbol{e}\|^2 \sim \sigma^2 \chi^2_{n-p}$;

(D) *under the hypothesis* $H_0 : \boldsymbol{\beta}_2 = 0$, $\|\boldsymbol{e}_1 - \boldsymbol{e}\|^2 \sim \sigma^2 \chi^2_{p-q}$.

Proof.

(A) holds since $\boldsymbol{e} - \boldsymbol{e}_1 = \boldsymbol{Y} - \hat{\boldsymbol{Y}} - \boldsymbol{Y} + \hat{\boldsymbol{Y}}_1 = -\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}_1 \in \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ but $\boldsymbol{e} \in [\mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)]^\perp$.

To show (B), we notice that

$$\boldsymbol{e}_1 = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H}_1\boldsymbol{H})\boldsymbol{Y}$$

because $\mathcal{M}(\boldsymbol{X}_1) \subset \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$.

Therefore,

$$\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{H}_1\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} - \boldsymbol{Y} + \boldsymbol{H}_1\boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{Y}.$$

But recall that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{1}\beta, \sigma^2\boldsymbol{I})$. Therefore, to prove independence of $\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, we need to show that

$$(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}[\sigma^2\boldsymbol{I}](\boldsymbol{I} - \boldsymbol{H})^\top = 0.$$

This is clearly the case since $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = 0$, proving (B).

(C) follows immediately, since we have already proven last time that $\forall \boldsymbol{\beta}$ (even when $\boldsymbol{\beta}_2 = 0$)

$$RSS(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{n-p}$$

To prove (D), we note that

$$\boldsymbol{e} - \boldsymbol{e}_1 = (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\,\boldsymbol{Y} \sim \mathcal{N}\{(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\,\boldsymbol{X}\beta, \sigma^2 \underbrace{(\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{H}\boldsymbol{H}^\top(\boldsymbol{H}_1 - \boldsymbol{I})^\top}_{=\boldsymbol{H}-\boldsymbol{H}_1}\}.$$

But $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}$. So, in block notation,

$$\boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}((\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{H}_1 - \boldsymbol{I})\boldsymbol{X}_2\boldsymbol{\beta}_2, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)).$$

Now $(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 = 0$ always, since $\boldsymbol{I} - \boldsymbol{H}_1$ projects onto $\mathcal{M}^\perp(\boldsymbol{X}_1)$. Therefore,

$$\boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}(0, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)), \qquad \text{when } \boldsymbol{\beta}_2 = 0.$$

Now observe that $(\boldsymbol{H} - \boldsymbol{H}_1)^\top = (\boldsymbol{H} - \boldsymbol{H}_1)$ and $(\boldsymbol{H} - \boldsymbol{H}_1)^2 = (\boldsymbol{H} - \boldsymbol{H}_1)$ (because $\mathcal{M}(\boldsymbol{X}_1) \subset \mathcal{M}(\boldsymbol{X}_1, \boldsymbol{X}_2)$). Thus,

$$\begin{aligned} \boldsymbol{e} - \boldsymbol{e}_1 \sim \mathcal{N}(0, \sigma^2(\boldsymbol{H} - \boldsymbol{H}_1)^2) \implies \boldsymbol{e} - \boldsymbol{e}_1 &\overset{d}{=} (\boldsymbol{H} - \boldsymbol{H}_1)\varepsilon \\ \implies RSS(\hat{\boldsymbol{\beta}}_1) - RSS(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{e} - \boldsymbol{e}_1\|^2 &\overset{d}{=} \varepsilon^\top(\boldsymbol{H} - \boldsymbol{H}_1)\varepsilon \sim \sigma^2\chi^2_{p-q}. \end{aligned}$$

since $(\boldsymbol{H} - \boldsymbol{H}_1)$ is symmetric idempotent with trace $p - q$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$.

Corollary

We conclude that, under the hypothesis $\boldsymbol{\beta}_2 = 0$,

$$\frac{\left(\dfrac{RSS(\hat{\boldsymbol{\beta}}_1) - RSS(\hat{\boldsymbol{\beta}})}{p - q}\right)}{\left(\dfrac{RSS(\hat{\boldsymbol{\beta}})}{n - p}\right)} \sim F_{p-q,n-p}$$

Distributional results suggest the following test:

- Have $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \sigma^2\boldsymbol{I})$
- $H_0 : \boldsymbol{\beta}_2 = 0$
- Data: $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2)$.
- Test statistic: $T = \dfrac{\left(\dfrac{RSS(\hat{\boldsymbol{\beta}}_1) - RSS(\hat{\boldsymbol{\beta}})}{p-q}\right)}{\left(\dfrac{RSS(\hat{\boldsymbol{\beta}})}{n-p}\right)}$

Then, under H_0, it holds that $T \sim F_{p-q,n-p}$. Suppose we observe $T = \tau$. Then,

$$p = \mathbb{P}_{H_0}[T(Y) \geq \tau] = \mathbb{P}[F_{p-q,n-p} \geq \tau]$$

Reject the null hypothesis if $p < \alpha$, some small α, usually 0.05.

Example (Nested Models in Cement Data)

▶We fitted the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

▶ But would the following simpler model be in fact adequate?

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

▶ Intuitively: is the extra explanatory power of the "larger" model significant enough in order to justify its use instead of a simpler model? (i.e., is the residual vector for the "larger" model significantly smaller than that of the simpler model?)
▶ In this case, $n = 13$, $p = 5$, $q = 2$ and

$$RSS(\hat{\beta}) = 47.86, \qquad RSS(\hat{\beta}_1) = 1265.7$$

yielding

$$\tau = \frac{(1265.7 - 47.86)/(5-2)}{(47.86)/(13-5)} = 67.86$$

▶$p = \mathbb{P}[F_{3,8} \geq 67.86] = 4.95 \times 10^{-6}$, so we reject the hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

- Let $\mathbf{1}$, $\underset{n\times q_1}{\boldsymbol{X}_1}$, …, $\underset{n\times q_r}{\boldsymbol{X}_r}$ be groups of columns of $\boldsymbol{X}$ (the "terms"), such that

$$\boldsymbol{X} = (\underset{n\times 1}{\mathbf{1}} \; \underset{n\times q_1}{\boldsymbol{X}_1} \; \underset{n\times q_2}{\boldsymbol{X}_2} \; \dots \; \underset{n\times q_r}{\boldsymbol{X}_r}), \quad \beta = (\underset{1\times 1}{\beta_0} \; \underset{1\times q_1}{\boldsymbol{\beta}_1} \; \underset{1\times q_2}{\boldsymbol{\beta}_2} \; \dots \; \underset{1\times q_r}{\boldsymbol{\beta}_r})^\top$$

We have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$$

- Would like to do the same "F-test investigation", but this time do it term-by-term. That is, we want to look at the following sequence of nested models:
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{\varepsilon}$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$
  - $\vdots$
  - $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \cdots + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$

Proceed similarly as before. Define:

- $\boldsymbol{X}_0 := \mathbf{1}$ and $\mathcal{X}_k = (\boldsymbol{X}_0 \; \boldsymbol{X}_1 \; \boldsymbol{X}_2 \; \dots \; \boldsymbol{X}_k), \quad k \in \{0, \dots, r\}$
- $\mathcal{H}_k := \mathcal{X}_k(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}\mathcal{X}_k^\top, \quad k \in \{0, \dots, r\}$
- $\hat{\boldsymbol{Y}}_k := \mathcal{H}_k \boldsymbol{Y}, \quad k \in \{0, \dots, r\}$
- $\boldsymbol{e}_k = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_k, \quad k \in \{0, \dots, r\}$
- Note that $\hat{\boldsymbol{Y}}_0 = \bar{Y}\mathbf{1}$.

- As before, Pythagoras implies

$$\begin{aligned}
\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_0\|^2}_{\|e_0\|^2} &= \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_r\|^2}_{\|e_r\|^2} + \underbrace{\underbrace{\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{r-1}\|^2}_{\|e_r - e_{r-1}\|^2} + \cdots + \underbrace{\|\hat{\boldsymbol{Y}}_1 - \hat{\boldsymbol{Y}}_0\|^2}_{\|e_1 - e_0\|^2}}_{\|e_r - e_0\|^2} \\
&= \underbrace{\|\boldsymbol{e}_r\|^2}_{RSS_r} + \sum_{k=0}^{r-1} \underbrace{\|\boldsymbol{e}_{k+1} - \boldsymbol{e}_k\|^2}_{RSS_k - RSS_{k+1}}
\end{aligned}$$

with RSS_k the residual sum of squares for $\hat{\boldsymbol{Y}}_k$, with ν_k degrees of freedom.

Some observations:

- $RSS_k - RSS_{k+1}$ is the reduction in residual sum of squares caused by adding $\boldsymbol{X}_{k+1}$, when the model already contains $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_k$.
- RSS_r and $\{RSS_k - RSS_{k+1}\}_{k=0}^{r-1}$ are all mutually independent.
- Obviously, $\nu_0 \geq \nu_1 \geq \nu_2 \geq \cdots \geq \nu_r$
- $\nu_{k+1} = \nu_k$ if $\boldsymbol{X}_{k+1} \in \mathcal{M}(\mathcal{X}_k)$.

▶Given this information, we want to see how adding each term in the model sequentially, affects the explanatory capacity of the model.

↪ In other words, we want to investigate the reduction in the residual sum of squares achieved by adding each term to the model. Is this significant?

ANOVA Table

Terms	df	Residual RSS	Terms added	df	Reduction in RSS	F-test
$\mathbf{1}$	$n-1$	RSS_0				
$\mathbf{1}, \boldsymbol{X}_1$	ν_1	RSS_1	$\boldsymbol{X}_1$	$n-1-\nu_1$	$RSS_0 - RSS_1$	
$\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2$	ν_2	RSS_2	$\boldsymbol{X}_2$	$\nu_1 - \nu_2$	$RSS_1 - RSS_2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_r$	ν_r	RSS_r	$\boldsymbol{X}_r$	$\nu_{r-1} - \nu_r$	$RSS_{r-1} - RSS_r$	

The F-statistic for testing the significance of the reduction in RSS when $\boldsymbol{X}_k$ is added to the model containing terms $1, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ is

$$F_k = \frac{(RSS_{k-1} - RSS_k)/(\nu_{k-1} - \nu_k)}{RSS_r/\nu_r},$$

and $F_k \sim F_{\nu_{k-1}-\nu_k, \nu_r}$ under the null hypothesis $H_0 : \beta_k = 0$.

Large values of F_k relative to the null distribution are evidence against H_0.

Example

Nested Sequence in Cement Data

- Reductions in overall sum of squares when sequentially entering terms x_1, x_2, x_3 and x_4.
- Does adding extra variables improve model significantly?

	Df	Red Sum Sq	F value (τ)	p-value
x_1	1	1450.08	242.37	2.88×10^{-7}
x_2	1	1207.78	201.87	5.86×10^{-7}
x_3	1	9.79	1.64	0.2366
x_4	1	0.25	0.04	0.8441
Residual SSq	8	47.86		

▶ In this case, each term is a single column (variable).

Warning!

- Significance of entering a term depends on how the sequence is defined: when entering terms in different order get different results! (why?)
- When a term is entered "early" and is significant, this does not tell us much (why?)
- When a term is entered "late" is significant, then this is quite informative (why?)

▶ Why is this true? Are there special cases when the order of entering terms doesn't matter?

▶Consider terms $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2$ from $\boldsymbol{X}$, so

$$\underset{n\times 1}{\boldsymbol{X}} = (\underset{n\times 1}{\boldsymbol{X}_0}\ \underset{n\times q_1}{\boldsymbol{X}_1}\ \underset{n\times q_2}{\boldsymbol{X}_2}), \quad \boldsymbol{\beta} = (\underset{1\times 1}{\beta_0}\ \underset{1\times q_1}{\boldsymbol{\beta}_1}\ \underset{1\times q_2}{\boldsymbol{\beta}_2})^\top$$

▶Assume orthogonality of terms, i.e. $\boldsymbol{X}_i^\top \boldsymbol{X}_j = 0, \quad i \neq j$
Notice that in this case

$$\hat{\beta} = \begin{pmatrix} \boldsymbol{X}_0^\top \boldsymbol{X}_0 & 0 & 0 \\ 0 & \boldsymbol{X}_1^\top \boldsymbol{X}_1 & 0 \\ 0 & 0 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix}^\top \boldsymbol{Y}$$

$$\implies \hat{\beta}_0 = \bar{Y},\ \hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{Y},\ \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{Y}$$

It follows that the reductions of sums of squares are unique, in the sense that they do not depend upon the order of entry of the terms in the model. (show this!)
Intuition: $\boldsymbol{X}_i$ contains completely linearly independent information from $\boldsymbol{X}_j$ for $\boldsymbol{Y}$, $i \neq j$

Model Selection / Collinearity / Regularisation

- **Theory:** We are given a relationship

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

and asked to provide estimators, tests, confidence intervals, optimality properties ...

... and we can do it with complete success!

- **Practice**: We are given data $(\boldsymbol{Y}, \boldsymbol{X})$ and suspect a linear relationship between $\boldsymbol{Y}$ and some of the columns of $\boldsymbol{X}$. We don't know a priori which exactly!
  - ↪ Need to select a "most appropriate" subset of the columns of $\boldsymbol{X}$
  - ↪ General principle: parsimony (Latin *parsimōnia*: sparingness; simplicity and least number of requisites and assumptions; economy or frugality of components and associations).

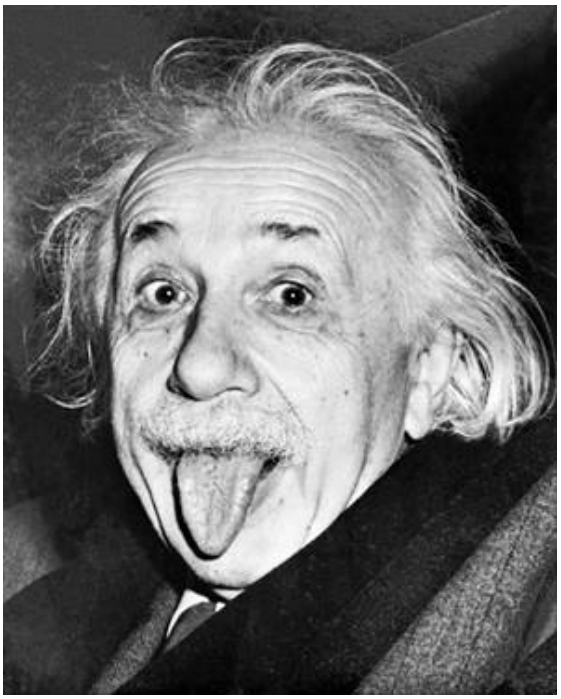


Figure: Albert Einstein (1879–1955): *'Everything should be made as simple as possible, but no simpler.'*

Figure: William of Ockham (?1285–1347): ' *It is vain to do with more what can be done with fewer*' (Occam's razor: Given several explanations of the same phenomenon, we should prefer the simplest.

Beyond Exploratory Data Analysis

Graphical exploration $\rightsquigarrow$ provides initial picture:

- plots of $\boldsymbol{Y}$ against candidate variables;
- plots of transformations of $\boldsymbol{Y}$ against candidate variables;
- plots of transformations of certain variables against $\boldsymbol{Y}$;
- plots of pairs of candidate variables.

This will often provide a starting point, but:

- Automatic Model Selection: Need objective model comparison criteria, as a screening device.
  - $\hookrightarrow$ We saw how to do an F-test, but what if models to be compared are not nested?
- Automatic Model Building: Situations when p large, so there are *lots* of possible models.
  - $\hookrightarrow$ Automatic methods for building a model? We saw that ANOVA depends on the order of entry of variables in the model ...

Consider design matrix $\boldsymbol{X}$ with p variables.

- 2^p possible models!
- Denote set of all models generated by $\boldsymbol{X}$ by $2^{\boldsymbol{X}}$ (model powerset)
- If wish to consider k different transformations of each variable, then p becomes $(1+k)p$
- Fast algorithms (branch and bound, `leaps` in `R`) exist to fit them, but they don't work for *large* p, and anyway ...
- ... need criterion for comparison.

So given a collection of models, we need an automatic (objective) way to pick out a "best" one (unfortunately cannot look carefully at all of them, but nothing can replace careful scrutiny of the final model by an experienced researcher).

Many possible choices, none universally accepted. Some (classical) possibilities:

- Prediction error based criteria (CV)
- Information criteria (AIC, BIC, ...)
- Mallow's C_p statistic

Before looking at these, let's introduce terminology: Suppose that the truth is

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ but with $\boldsymbol{\beta}_r = 0$ for some subset $\boldsymbol{\beta}_r$ of $\boldsymbol{\beta}$.

- The true model contains only the columns for which $\boldsymbol{\beta}_r \neq 0$
  - ↪ Equivalently, the true model uses $\boldsymbol{X}_\heartsuit$ as the design matrix, the latter being the matrix of columns of $\boldsymbol{X}$ corresponding to non-zero coefficients.
- A correct model is the true model plus extra columns.
  - ↪ Equivalently, a correct model has a design matrix $\boldsymbol{X}_\diamondsuit$, such that $\mathcal{M}(\boldsymbol{X}_\heartsuit) \subset \mathcal{M}(\boldsymbol{X}_\diamondsuit)$.
- A wrong model is a model that does not contain all the columns of the true model.
  - ↪ Equivalently, a wrong model has a design matrix $\boldsymbol{X}_\diamondsuit$, such that $\mathcal{M}(\boldsymbol{X}_\heartsuit) \cap \mathcal{M}(\boldsymbol{X}_\diamondsuit) \neq \mathcal{M}(\boldsymbol{X}_\heartsuit)$.

▶ We may wish to choose a model by minimising the error we make on average, when predicting a future observation given our model.
Our "experiment" is:

- Design matrix $\boldsymbol{X}$
- response $\boldsymbol{Y}$ at $\boldsymbol{X}$

Every model $f \in 2^{\boldsymbol{X}}$, will yield fitted values $\hat{\boldsymbol{Y}}(f) = \boldsymbol{H}_f \boldsymbol{Y}$. And suppose we now obtain new independent responses $\boldsymbol{Y}_+$ for the same "experimental setup" $\boldsymbol{X}$.
Then, one approach is to select the model

$$f^* = \arg\min_{f \in 2^{\boldsymbol{X}}} \underbrace{\frac{1}{n}\mathbb{E}\left\{\| \boldsymbol{Y}_+ - \hat{\boldsymbol{Y}}(f)\|^2\right\}}_{\Delta(f)},$$

where expectation is taken over both $\boldsymbol{Y}$ and $\boldsymbol{Y}_+$.

Let $\boldsymbol{X}$ be a design matrix, and let $\boldsymbol{X}_\diamondsuit$ $(n \times p)$ and $\boldsymbol{X}_\heartsuit$ $(n \times q)$ be matrices built using columns of $\boldsymbol{X}$. Suppose that the true relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is

$$\boldsymbol{Y} = \underbrace{\boldsymbol{X}_\heartsuit \beta}_{\mu} + \varepsilon$$

but we use the matrix $\boldsymbol{X}_\diamondsuit$ instead of $\boldsymbol{X}_\heartsuit$ (i.e., we fit a different model).
Therefore our fitted values are

$$\hat{\boldsymbol{Y}} = (\boldsymbol{X}_\diamondsuit^\top \boldsymbol{X}_\diamondsuit)^{-1} \boldsymbol{X}_\diamondsuit^\top \boldsymbol{Y} = \boldsymbol{H}_\diamondsuit \boldsymbol{Y}.$$

Now suppose that we obtain new observations $\boldsymbol{Y}_+$ corresponding to the same design $\boldsymbol{X}$

$$\boldsymbol{Y}_+ = \boldsymbol{X}_\heartsuit \boldsymbol{\beta} + \varepsilon_+ = \boldsymbol{\mu} + \varepsilon_+.$$

Then, observe that

$$\begin{aligned} \boldsymbol{Y}_+ - \hat{\boldsymbol{Y}} &= \boldsymbol{\mu} + \varepsilon_+ - \boldsymbol{H}_\diamondsuit(\boldsymbol{\mu} + \varepsilon) \\ &= (\boldsymbol{I} - \boldsymbol{H}_\diamondsuit)\boldsymbol{\mu} + \varepsilon_+ - \boldsymbol{H}_\diamondsuit \varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned}\|\boldsymbol{y}_+ - \hat{\boldsymbol{Y}}\|^2 &= (\boldsymbol{Y}_+ - \hat{\boldsymbol{Y}})^\top(\boldsymbol{Y}_+ - \hat{\boldsymbol{Y}}) \\ &= \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_\diamond)\boldsymbol{\mu} + \boldsymbol{\varepsilon}^\top \boldsymbol{H}_\diamond \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}_+^\top \boldsymbol{\varepsilon}_+ + [\text{cross terms}].\end{aligned}$$

Since $\mathbb{E}[\text{cross terms}] = 0$ (why?), we observe that

$$\Delta = \begin{cases} n^{-1}\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_\diamond)\boldsymbol{\mu} + (1 + p/n)\sigma^2, & \text{if model wrong,} \\ (1 + p/n)\sigma^2, & \text{if model correct,} \\ (1 + q/n)\sigma^2, & \text{if model true.} \end{cases}$$

- Selecting a correct model instead of the true model brings in additional variance, because $q < p$.
- Selecting a wrong model instead of the true model results in bias, since $(\boldsymbol{I} - \boldsymbol{H}_\diamond)\boldsymbol{\mu} \neq 0$ when $\boldsymbol{\mu}$ is not in the column space of $\boldsymbol{X}_\diamond$.
- Must find a balance between small variance (few columns in the model) and small bias (all columns in the model).

Cross Validation, Again.

▶ Impossible to calculate Δ (depends on unknown $\boldsymbol{\mu}$ and σ^2), so we must find a proxy (estimator) $\widehat{\Delta}$.
Suppose that n is large so that we can split the data in two pieces:

- $\boldsymbol{X}^*$, $\boldsymbol{Y}^*$ used to estimate the model
- $\boldsymbol{X}'$, $\boldsymbol{Y}'$ used to estimate the prediction error for the model

The estimator of the prediction error will be

$$\widehat{\Delta} = (n')^{-1}\| \boldsymbol{Y}' - \boldsymbol{X}'\hat{\boldsymbol{\beta}}^*\|^2.$$

In practice n might be small and we often cannot afford to split the data (variance of $\hat{\Delta}$ is too large).
Instead we use the leave-one-out cross validation:

$$n\widehat{\Delta}_{CV} = CV = \sum_{j=1}^{n}(Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_{-j})^2,$$

where $\hat{\boldsymbol{\beta}}_{-j}$ is the estimate produced when dropping the jth case (line).

Does this mean we need to fit n different models?

No! Rank 1 perturbation theory proves that

$$CV = \sum_{j=1}^{n} \frac{(Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})^2}{(1-h_{jj})^2},$$

so the full regression may be used! Alternatively one may use a more stable version:

$$GCV = \sum_{j=1}^{n} \frac{(Y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})^2}{(1-\text{trace}(\boldsymbol{H})/n)^2},$$

where "G" stands for "generalised", and we guard against any $h_{jj} \approx 1$.
It can be shown that:

$$\mathbb{E}[GCV] = \frac{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\mu}}{(1-p/n)^2} + \frac{n\sigma^2}{1-p/n} \approx n\Delta.$$

▷ Suggests strategy: pick variables to minimise (G)CV.

Akaike's Information Criterion

Criteria can be obtained based on the notion of *relative entropy (KL divergence)*.

- Same basic idea as for prediction error: aim to choose candidate model $f(\boldsymbol{y})$ to minimise *information distance*:

  $$\int \log\left\{\frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}\right\} g(\boldsymbol{y})\, dy \geq 0,$$

  where $g(\boldsymbol{y})$ represents true model—equivalent to maximising expected log likelihood

  $$\int \log f(\boldsymbol{y})\, g(\boldsymbol{y})\, dy.$$

- Can show that (apart from constants) information distance is estimated by

  $$\text{AIC} = -2\hat{\ell} + 2p \quad (\equiv n \log \hat{\sigma}^2 + 2p \text{ in linear model})$$

  where $\hat{\ell}$ is maximised log likelihood for given model, and p is number of parameters.

There are many flavours of such critera:

- Improved (corrected) version of AIC for regression problems:

$$\mathrm{AIC}_c \equiv \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}.$$

- Also can use *Bayes' information criterion*

$$\mathrm{BIC} = -2\hat{\ell} + p \log n.$$

- Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.
- Comments:
  - AIC tends to choose models that are too complicated, buts AIC_c cures this somewhat;
  - BIC is *model selection consistent*—if the true model is among those fitted, BIC chooses it with probability $\to 1$ as $n \to \infty$ (for fixed p).

Example (Simulation Experiment)

For each $n \in \{10, 20, 40\}$ we construct 20 $n \times 7$ design matrices. We multiply each of these design matrices from the right with $\beta = (1, 2, 3, 0, 0, 0, 0)^\top$ and we add a $n \times 1$ Gaussian error. We do this independently 50 times, obtaining 1000 regressions with $p = 3$. Selected models with 1 or 2 covariates have a bias term, and those with 4 or more covariates have excess variance.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	AIC_c	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	AIC_c		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	AIC_c			786	105	52	41	16

▶ We saw so far:
Automatic Model Selection: build a set of models and select the "best" one.

▶ Now look at different philosophy:
Automatic Model Building: construct a single model in a way that would hopefully provide a good one.

There golden oldies for doing this are:

- Forward Selection
- Backward Elimination
- Stepwise Selection

Caution: Although widely used, these have little theoretical basis. Element of arbitrariness . . .

- *Forward selection*: starting from the model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to step 1.

- *Backward elimination*: starting from the model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to step 1.

- *Stepwise*: starting from an arbitary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model, and choose the most significant option;
  2. if model unchanged, stop; otherwise go to step 1.

Some thoughts:

- Each procedure may produce a different model.
- Systematic search minimising Prediction Error, AIC or similar over all possible models is preferable— BUT not always feasible (e.g., when p large).
- Stepwise methods can fit 'highly significant' models to purely random data! Main problem is lack of objective function.
- Can be improved by comparing Prediction Error/AIC for different models at each step — uses objective function, but no systematic search.

Example (Nuclear Power Station Data)

Data on light water reactors (LWR) constructed in the USA. The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
⋮											
32	270.71	67.83	7	80	886	1	0	0	1	11	1

Example (Nuclear Power Station Data, continued)

	Full model		Backward		Forward	
	Est	t	Est	t	Est	t
Int.	−14.24	−3.37	−13.26	−4.22	−7.62	−2.66
`date`	0.2	3.21	0.21	4.91	0.13	3.38
`logT1`	0.092	0.38				
`logT2`	0.29	1.05				
`logcap`	0.694	5.10	0.72	6.09	0.67	4.75
`PR`	−0.092	−1.20				
`NE`	0.25	3.35	0.24	3.36		
`CT`	0.12	1.82	0.14			
`BW`	0.033	0.33				
`log(N)`	−0.08	−1.74	−0.08	−2.11		
`PT`	−0.22	−1.83	−0.22	−1.99	−0.49	−4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Dangers of "Big" Models

Recall: $\hat{\boldsymbol{Y}}$ is projection of $\boldsymbol{Y}$ onto $\mathcal{M}(\boldsymbol{X})$

↪ Adding more variables (columns) into $\boldsymbol{X}$ "enlarges" $\mathcal{M}(\boldsymbol{X})$
. . . if the rank increases by the # of new variables

Consider two extremes

- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}^{\perp}(\boldsymbol{X})$
  ↪ Gives us completely "new" information.
- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}(\boldsymbol{X})$
  ↪ Gives no "new" information — cannot even do least squares (why not?)

What if we are between the two extremes? What if

$$\boldsymbol{X}_{p+1} \notin \mathcal{M}(\boldsymbol{X}) \quad \text{but} \quad \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}_{p+1} = \boldsymbol{H}\boldsymbol{X}_{p+1} \simeq \boldsymbol{X}_{p+1}?$$

We can certainly fit the regression, but what will happen?

Using block matrix properties, have

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left[(\boldsymbol{X} \ \boldsymbol{X}_{p+1})^\top (\boldsymbol{X} \ \boldsymbol{X}_{p+1}) \right]^{-1}$$

with

$$\left[(\boldsymbol{X} \ \boldsymbol{X}_{p+1})^\top (\boldsymbol{X} \ \boldsymbol{X}_{p+1}) \right]^{-1} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X^\top \boldsymbol{X}_{p+1} \\
&\quad \times (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{B} &= -(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}_{p+1} (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}, \\
\boldsymbol{C} &= -(\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{D} &= (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}.
\end{aligned}$$

Multicollinearity

Multicollinearity: when p covariates concentrate around a subspace of dimension $q < p$

[simplest case: pairs of variables that are correlated]

But: might exist even if pairs of variables appear uncorrelated!

Can be caused by:

- Poor design [can try designing again],
- Inherent relationships [other remedies needed].

So what are the results?

- Huge variances of the estimators!
  - ↪ Can even flip signs for different data, to give the impression of inverse effects.
- Individual coefficients insignificant:
  - ↪ t-test p-values inflated.
- But global F-test might give significant result!

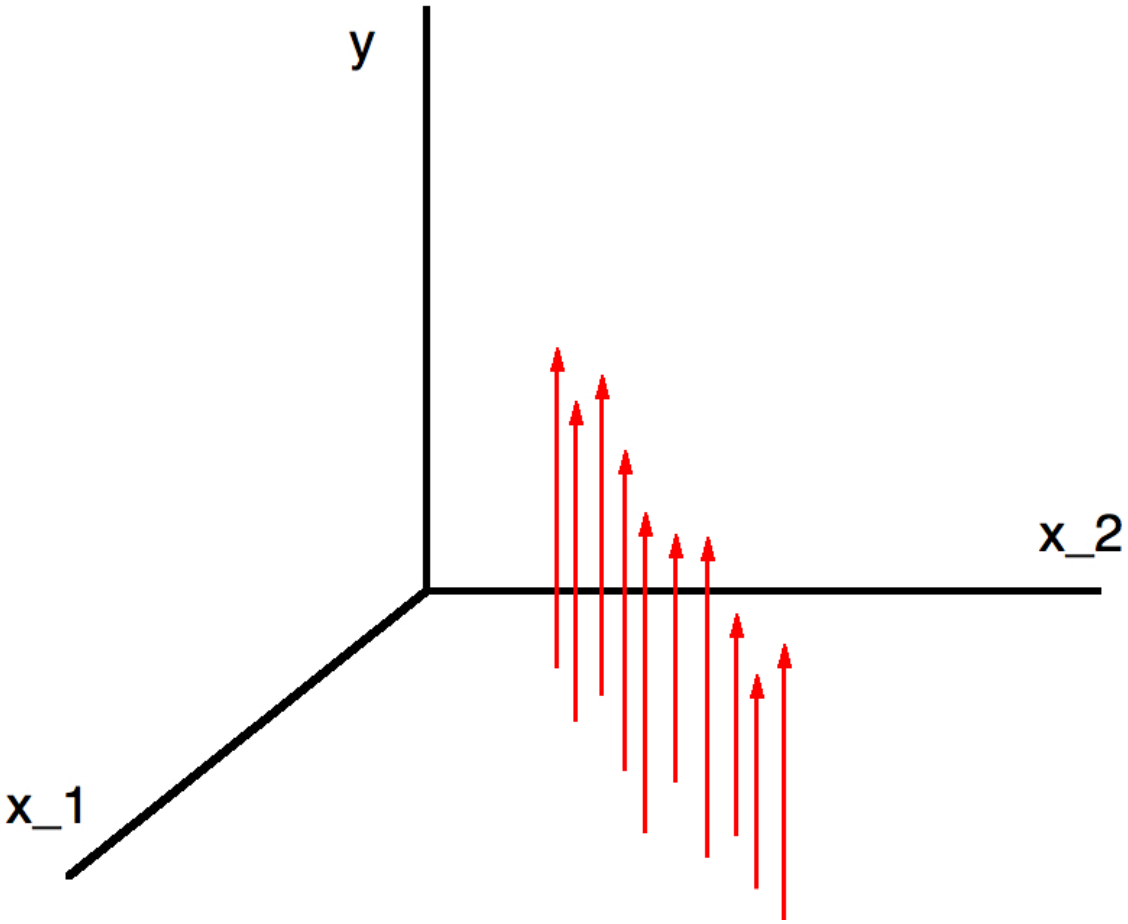


Figure: The Picket-Fence (Hocking & Pendleton)

Diagnosing Multicollinearity: Variance Inflation Factors

Simple first steps:

- Look at scatterplots,
- Look at correlation matrix of covariates,

Might not reveal more complex linear constraints, though.

- Look at the *variance inflation factors*:

$$VIF_j = \frac{\text{var}(\hat{\beta}_j)\|\boldsymbol{X}_j\|^2}{\sigma^2} = \|\boldsymbol{X}_j\|^2 \left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right]_{jj}.$$

Can show that

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination for the regression

$$\text{of } \boldsymbol{X}_j \text{ on } \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p\},$$

measuring linear dependence of $\boldsymbol{X}_j$ on the other columns of $\boldsymbol{X}$.

Let $\boldsymbol{X}_{-j}$ be the design matrix without the j-th variable. Then

$$R_j^2 = \frac{\|\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top \boldsymbol{X}_j\|^2}{\|\boldsymbol{X}_j\|^2} \in [0,1]$$

is close to 1 if $\underbrace{\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}}_{\boldsymbol{H}_{-j}}\boldsymbol{X}_j \simeq \boldsymbol{X}_j$.

Large values of VIF_j indicate that $\boldsymbol{X}_j$ is linearly dependent on the other columns of the design matrix.

Interpretation: how much the variance is inflated when including variable j as compared to the variance we would obtain if $\boldsymbol{X}_j$ were orthogonal to the other variables—how much worse are we doing as compared to the ideal case.

Rule of thumb: $VIF_j > 5$ or $VIF_j > 10$ considered to be "large".

Diagnosing Multicollinearity: Condition Indices and Numbers

Consider the spectral decomposition of $\boldsymbol{X}^\top \boldsymbol{X}$, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $\boldsymbol{U}^\top \boldsymbol{U} = I$. Then

$$\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \#\{j : \lambda_j \neq 0\}, \qquad \det(\boldsymbol{X}^\top \boldsymbol{X}) = \prod_{j=1}^{p} \lambda_j.$$

Hence "small" λ_j's mean "almost" reduced rank, revealing the effect of collinearity. Measure using condition index:

$$CI_j(\boldsymbol{X}^\top \boldsymbol{X}) := \sqrt{\lambda_{\text{max}}/\lambda_j}$$

Global "instability" measured by the condition number,

$$CN(\boldsymbol{X}^\top \boldsymbol{X}) = \sqrt{\lambda_{\text{max}}/\lambda_{\text{min}}}$$

Rule of thumb: $CN > 30$ indicates moderate to significant collinearity, $CN > 100$ indicates severe collinearity (choices vary).

Remedies?
If design faulty, may redesign.

↪ Otherwise? Inherent relationships between covariates.

- Variable deletion - attempt to remove problematic variables
  → E.g., by backward elimination.

- Choose an orthogonal basis for $\mathcal{M}(\boldsymbol{X})$ and use its elements as covariates
  → Use columns of $\boldsymbol{U}$ from spectrum, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$
  → OK for prediction
  → Problem: lose interpretability

Other approaches?

Example (Body Fat Data)

Body fat is measure of health → not easy to measure!
Collect 252 measurements on body fat and some explanatory variables.

Can we use measuring tape and scales only to find body fat?

Explanatory variables:

- age
- weight
- height
- biceps
- neck
- chest
- abdomen
- forearm
- hip
- thigh
- knee a
- ankle
- wrist

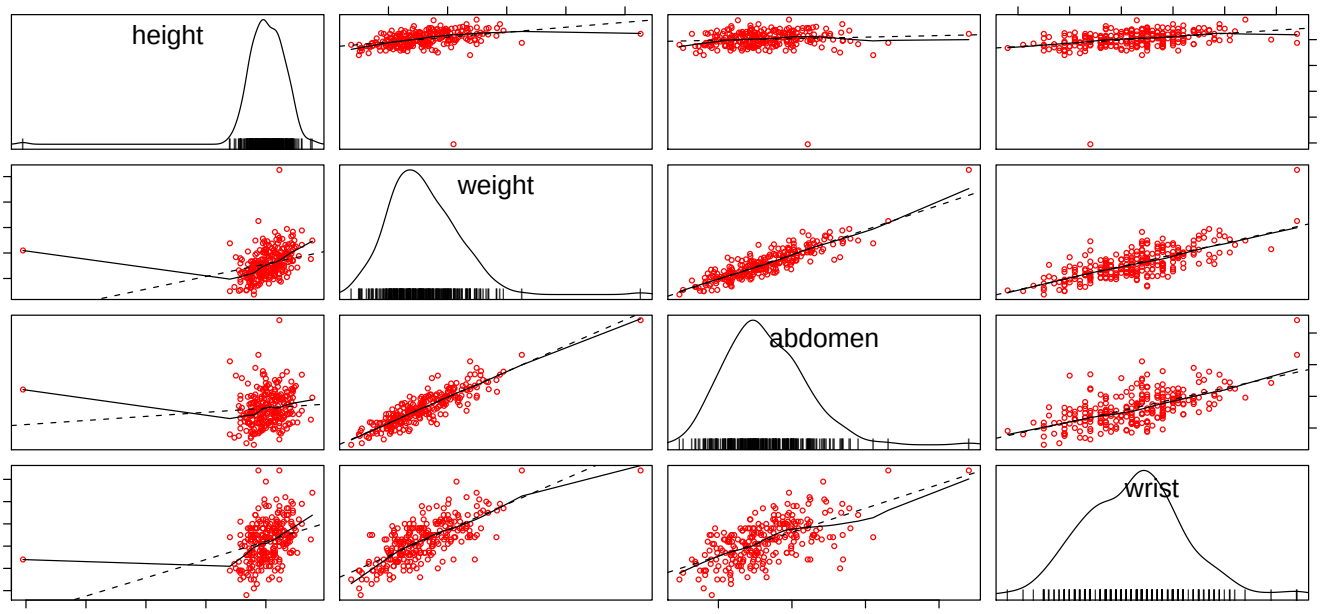


Figure: Some Scatterplots [library(car);scatterplot.matrix(. . .)]. Looks like we're in trouble. Let's go ahead and fit anyway . . .

Model Fit Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
age	0.0621	0.0323	1.92	0.0562
weight	−0.0884	0.0535	−1.65	0.0998
height	−0.0696	0.0960	−0.72	0.4693
neck	−0.4706	0.2325	−2.02	0.0440
chest	−0.0239	0.0991	−0.24	0.8100
abdomen	0.9548	0.0864	11.04	0.0000
hip	−0.2075	0.1459	−1.42	0.1562
thigh	0.2361	0.1444	1.64	0.1033
knee	0.0153	0.2420	0.06	0.9497
ankle	0.1740	0.2215	0.79	0.4329
biceps	0.1816	0.1711	1.06	0.2897
forearm	0.4520	0.1991	2.27	0.0241
wrist	−1.6206	0.5349	−3.03	0.0027

$R^2 = 0.749$, F-test: $p < 2.2 \times 10^{-16}$.

Split Data in Two and Fit Separately (Picket Fence)

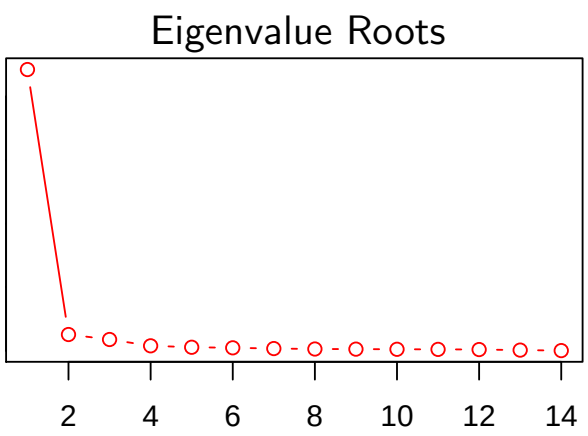
	Estimate	Pr(>|t|)
(Intercept)	−32.6564	0.1393
age	0.1048	0.0153
weight	−0.1285	0.0502
height	−0.0666	0.5207
neck	−0.5086	0.0721
chest	0.0168	0.9002
abdomen	0.9750	0.0000
hip	−0.2891	0.1265
thigh	0.3850	0.0565
knee	0.2218	0.5111
ankle	0.4377	0.0694
biceps	−0.1297	0.5485
forearm	0.8871	0.0174
wrist	−1.7378	0.0309

Estimate	Pr(>|t|)
−1.2221	0.9730
0.0256	0.6252
−0.0237	0.8223
−0.1005	0.7284
−0.4619	0.2635
−0.0910	0.5877
0.8924	0.0000
−0.0265	0.9130
0.0334	0.8793
−0.1310	0.7366
−0.5037	0.3516
0.4458	0.1179
0.2247	0.3750
−1.5902	0.0560

Diagnostic Check

	VIF
age	2.25
weight	33.51
height	1.67
neck	4.32
chest	9.46
abdomen	11.77
hip	14.80
thigh	7.78
knee	4.61
ankle	1.91
biceps	3.62
forearm	2.19
wrist	3.38

	CI
1	1.00
2	17.47
3	25.30
4	58.61
5	83.59
6	100.63
7	137.90
8	175.29
9	192.62
10	213.01
11	228.16
12	268.21
13	555.67



Condition Number $\simeq$ 556 !

```
 Multiple R-Squared:  0.7466,
F-statistic p-value:  < 2.2e-16
```

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	−22.6564	11.7139	−1.93	0.0543	
age	0.0658	0.0308	2.14	0.0336	2.05
weight	−0.0899	0.0399	−2.25	0.0252	18.82
neck	−0.4666	0.2246	−2.08	0.0388	4.08
abdomen	0.9448	0.0719	13.13	0.0000	8.23
hip	−0.1954	0.1385	−1.41	0.1594	13.47
thigh	0.3024	0.1290	2.34	0.0199	6.28
forearm	0.5157	0.1863	2.77	0.0061	1.94
wrist	−1.5367	0.5094	−3.02	0.0028	3.09

Define $\boldsymbol{Z} = \boldsymbol{X}\,\boldsymbol{U}$ as design matrix. R^2=0.749, F-test p-value<2.2×10^{-16}

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
Z[, 1]	−0.1353	0.0619	−2.19	0.0297
Z[, 2]	−0.0168	0.0916	−0.18	0.8546
Z[, 3]	0.2372	0.1070	2.22	0.0276
Z[, 4]	−0.7188	0.0571	−12.58	0.0000
Z[, 5]	0.0248	0.0827	0.30	0.7649
Z[, 6]	0.4546	0.1001	4.54	0.0000
Z[, 7]	0.5903	0.1366	4.32	0.0000
Z[, 8]	−0.1207	0.1742	−0.69	0.4890
Z[, 9]	−0.0836	0.1914	−0.44	0.6627
Z[, 10]	0.5043	0.2082	2.42	0.0162
Z[, 11]	−0.5735	0.2254	−2.54	0.0116
Z[, 12]	0.3007	0.2628	1.14	0.2536
Z[, 13]	1.5168	0.5447	2.78	0.0058

- Eigenvector approach rotates space so as to "free" the dependence of one coefficient β_j on others $\{\beta_i\}_{i \neq j}$

  ↪ Imposes constraint on $\boldsymbol{X}$ (orthogonal columns)

Problem: lose interpretability! (prediction OK)

  - Example: most significant "rotated" term in fat data: `Z[,4]=-0.01*age -0.058*weight -0.011*height +0.46*neck -0.144*chest -0.441*abdomen +0.586*hip +0.22*thigh -0.197*knee -0.044*ankle -0.07*biceps -0.33*forearm -0.249*wrist`

- Other approach to reduce this strong dependence?

  ↪ Impose constraint on $\boldsymbol{\beta}$! How? (introduces bias)

Ridge Regression: From Rotation to Shrinkage

Multicollinearity problem is that $\det\left[\boldsymbol{X}^\top \boldsymbol{X}\right] \approx 0$

[i.e. $\boldsymbol{X}^\top \boldsymbol{X}$ *almost not invertible*]

A Solution: add a "small amount" of a full rank matrix to $\boldsymbol{X}^\top \boldsymbol{X}$.

For reasons to become clear soon, we *standardise* the design matrix:

- Write $\boldsymbol{X} = (\mathbf{1} \ \boldsymbol{W})$, $\beta = (\beta_0 \ \boldsymbol{\gamma})^\top$
- Recentre/rescale the covariates (columns) defining: $\boldsymbol{Z}_j = \frac{\sqrt{n}}{\mathrm{sd}(\boldsymbol{W}_j)}(\boldsymbol{W}_j - \mathbf{1}\overline{\boldsymbol{W}}_j)$
  - ↪ Coefficients now have common scale
  - ↪ Interpretation of β_j slightly different: not "mean impact on response per unit change of explanatory variable", but now "mean impact on response per unit deviation of explanatory variable from its mean, measured in units of standard deviation"
- The $\boldsymbol{Z}_j$ are all orthogonal to $\mathbf{1}$ and are of unit norm.

- Since $\boldsymbol{Z}_j \perp \mathbf{1}$ for all, j, we can estimate β_0 and $\boldsymbol{\gamma}$ by two separate regressions (orthogonality).
- Least squares estimators become

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}.$$

- Ridge regression replaces $\boldsymbol{Z}^\top \boldsymbol{Z}$ by $\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_{(p-1)\times(p-1)}$ (i.e. adds a "ridge")

$$\boxed{\hat{\beta}_0 = \bar{Y}, \quad \hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}}.$$

Adding $\lambda \boldsymbol{I}_{(p-1)\times(p-1)}$ to $\boldsymbol{Z}^\top \boldsymbol{Z}$ makes inversion more stable
$\hookrightarrow$ λ called *ridge parameter*.

$\rightarrow$ Ridge term $\lambda \boldsymbol{I}$ seems slightly ad-hoc. Motivation?

$\hookrightarrow$ Can see that $(\hat{\beta}_0 \quad \hat{\boldsymbol{\gamma}}) = (\bar{Y} \quad (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y})$ minimizes

$$\|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_2^2$$

or equivalently

$$\|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} \gamma_j^2 = \|\boldsymbol{\gamma}\|_2^2 \leq r(\lambda)$$

instead of least squares estimator which minimizes

$$\|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2.$$

Idea: in the presence of collinearity, coefficients are ill-defined: a wildly positive coefficient can be cancelled out by a largely negative coefficient (many coefficient combinations can produce the same effect). By imposing a *size* constraint, we limit the possible coefficient combinations!

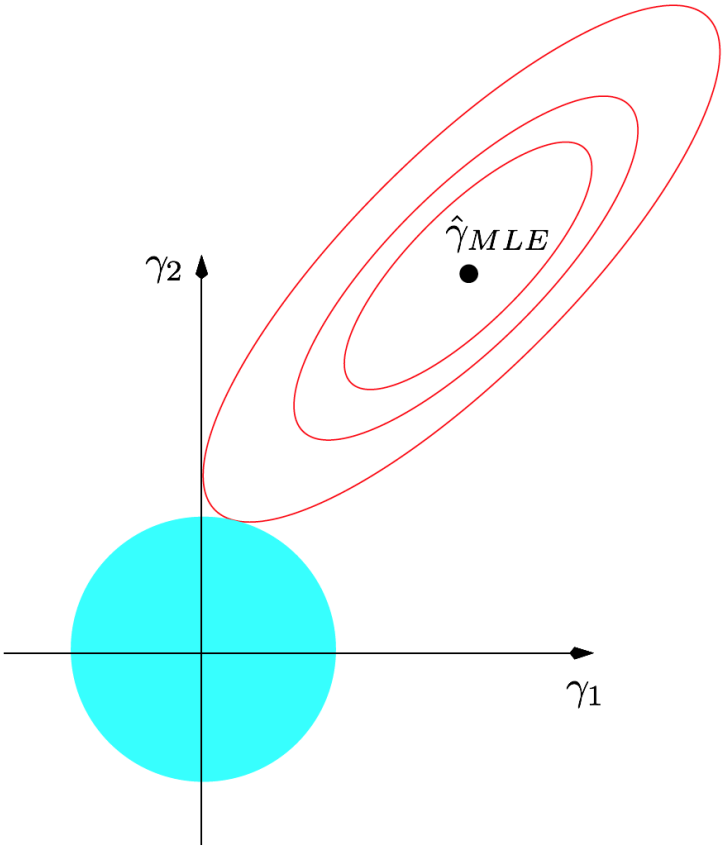


Figure: L^2 Shrinkage [Ridge Regression]

Theorem

Let $\boldsymbol{Z}_{n\times q}$ be a matrix of rank $r \leq q$ with centred column vectors of unit norm. Given $\lambda > 0$, the unique minimiser of

$$Q(\alpha, \boldsymbol{\xi}) = \| \boldsymbol{Y} - \alpha \mathbf{1} - \boldsymbol{Z}\boldsymbol{\xi}\|_2^2 + \lambda \|\boldsymbol{\xi}\|_2^2$$

is

$$(\hat{\beta}_0 \, , \, \hat{\gamma}) = (\, \bar{Y} \, , \, (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}).$$

Proof.

Write

$$\boldsymbol{Y} = \underbrace{(\boldsymbol{Y} - \bar{Y}\mathbf{1})}_{=\boldsymbol{Y}^* \in \mathcal{M}^\perp(\mathbf{1})} + \underbrace{\bar{Y}\mathbf{1}}_{\in \mathcal{M}(\mathbf{1})}$$

Note also that by assumption $\mathbf{1} \in \mathcal{M}^\perp(\boldsymbol{Z})$. Therefore by Pythagoras' theorem

$$\| \boldsymbol{Y} - \hat{\beta}_0 \mathbf{1} - \boldsymbol{Z}\hat{\gamma}\|_2^2 = \| \underbrace{(\bar{Y} - \hat{\beta}_0)\mathbf{1}}_{\in \mathcal{M}(\mathbf{1})} + \underbrace{(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\gamma})}_{\in \mathcal{M}(\boldsymbol{Z})} \|_2^2 = \|(\bar{Y} - \hat{\beta}_0)\mathbf{1}\|_2^2 + \|(\boldsymbol{Y}^* - \boldsymbol{Z}\hat{\gamma})\|_2^2.$$

Therefore, $\min_{\alpha,\boldsymbol{\xi}} Q(\alpha,\boldsymbol{\xi}) = \min_{\alpha} \|(\bar{Y}-\alpha)\mathbf{1}\|_2^2 + \min_{\boldsymbol{\xi}} \left\{ \|(\boldsymbol{Y}^* - \boldsymbol{Z}\boldsymbol{\xi})\|_2^2 + \lambda\|\boldsymbol{\xi}\|_2^2 \right\}$

Clearly, $\arg\min_\alpha \|(\bar{Y}-\alpha)\mathbf{1}\|_2^2 = \bar{Y}$ while the second component can be written

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^q} \left\| \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda} I_{q\times q} \end{pmatrix} \boldsymbol{\xi} - \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} \right\|_2^2$$

using block notation. This is the usual least squares problem with solution

$$\left[(\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda}\boldsymbol{I}_{q\times q} \end{pmatrix} \right]^{-1} (\boldsymbol{Z}^\top, \sqrt{\lambda}\boldsymbol{I}_{q\times q}) \begin{pmatrix} \boldsymbol{Y}^* \\ \mathbf{0}_{q\times 1} \end{pmatrix} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}^*$$

Note that $\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I}$ is indeed invertible. Writing $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we have

$$\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\boldsymbol{I} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{U}(\lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_{q\times q})\boldsymbol{U}^\top$$

and $\boldsymbol{\Lambda} = \text{diag}\{\underbrace{\lambda_1,\ldots,\lambda_r}_{>0}, \underbrace{\lambda_{r+1},\ldots,\lambda_q}_{=0}\}$ ($\boldsymbol{Z}^\top\boldsymbol{Z} \succeq 0$ & $\texttt{rank}(\boldsymbol{Z}^\top\boldsymbol{Z}) = \texttt{rank}(\boldsymbol{Z})$).

To complete the proof, observe that $\boldsymbol{Z}^\top\boldsymbol{Y}^* = \boldsymbol{Z}^\top\boldsymbol{Y} - \bar{Y}\boldsymbol{Z}^\top\mathbf{1} = \boldsymbol{Z}^\top\boldsymbol{Y}$. □

The Effect of Shrinkage: Bias and Variance

Note that if the SVD of $\boldsymbol{Z}$ is $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, last steps of previous proof may be used to show that

$$\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \frac{\lambda_j}{\lambda_j^2 + \lambda} (\boldsymbol{V}_j^\top\boldsymbol{Y})\boldsymbol{U}_j,$$

where the $\boldsymbol{V}_j$s and $\boldsymbol{U}_j$s are the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$, respectively.

Compare this to the ordinary least squares solution, when $\lambda = 0$:

$$\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \frac{1}{\lambda_j} (\boldsymbol{V}_j^\top\boldsymbol{Y})\boldsymbol{U}_j,$$

which is not even defined if $\boldsymbol{Z}$ is of reduced rank.

Role of λ is to reduce the size of $1/\lambda_j$ when λ_j becomes very small.

Proposition

Let $\hat{\gamma}$ be the ridge regression estimator of γ. Then

$$\text{bias}(\hat{\gamma}, \gamma) = -\lambda \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_{q \times q} \right)^{-1} \gamma$$

and

$$\text{cov}(\hat{\gamma}) = \sigma^2 (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}.$$

Proof.

Since $\mathbb{E}(\hat{\gamma}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} \gamma$, the bias is

$$\begin{aligned}
\text{bias}(\hat{\gamma}, \gamma) &= \mathbb{E}(\hat{\gamma}) - \gamma = \{ (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Z} - \boldsymbol{I} \} \gamma \\
&= \left\{ \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} - \boldsymbol{I} \right) - \boldsymbol{I} \right\} \gamma \\
&= \left\{ \boldsymbol{I} - \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} - \boldsymbol{I} \right\} \gamma = - \left(\frac{1}{\lambda} \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I} \right)^{-1} \gamma.
\end{aligned}$$

The covariance term is straightforward. □

Corollary (Domination over Least Squares)

Assume that $\texttt{rank}(\boldsymbol{Z}_{n \times q}) = q$ and let

$$\tilde{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} \qquad \& \qquad \hat{\gamma}_\lambda = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}$$

be the least squares estimator and ridge estimator, respectively. Then,

$$\mathbb{E}\left\{ (\tilde{\gamma} - \gamma)(\tilde{\gamma} - \gamma)^\top \right\} - \mathbb{E}\left\{ (\hat{\gamma}_\lambda - \gamma)(\hat{\gamma}_\lambda - \gamma)^\top \right\} \succeq 0$$

for all $\lambda \in (0, 2\sigma^2 / \|\gamma\|^2)$.

Ridge estimator uniformly better than least squares! How can this be?
(What happened to Gauss-Markov?)

- Gauss-Markov only covers unbiased estimators – but ridge estimator biased.
- A bit of bias can improve the MSE by reducing variance!
- Also, there is a catch! The "right" range for λ depends on unknowns.
- Choosing a good λ is all about balancing bias and variance.

Proof.

From our bias/variance calculations on the ridge estimator, we have

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} =$$

$$= \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1} + \lambda\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}_{q\times q}\right)^{-1}\gamma$$

$$= \lambda(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\left(\sigma^2(2\boldsymbol{I}+\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}) - \lambda\gamma\gamma^\top\right)(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}.$$

This last term can be made positive definite if

$$2\sigma^2\boldsymbol{I} + \sigma^2\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - \lambda\gamma\gamma^\top \succeq 0.$$

Noting that we can always write

$$\boldsymbol{I} = \frac{\gamma\gamma^\top}{\|\gamma\|^2} + \sum_{j=1}^{q-1}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^\top$$

for $\{\gamma/\|\gamma\|, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{q-1}\}$ an orthonormal basis of $\mathbb{R}^q$ we see that $\lambda \in (0, 2\sigma^2/\|\gamma\|^2)$ suffices for positive definiteness to hold true.

□

Bias–Variance Tradeoff, Again.

Role of λ: Regulates Bias–Variance tradeoff

- $\lambda\uparrow$ decreases variance (collinearity) but increases bias
- $\lambda\downarrow$ decreases bias but variance inflated if collinearity exists

Recall:

$$\mathbb{E}\|\hat{\gamma}-\gamma\|^2 = \underbrace{\|\mathbb{E}\hat{\gamma}-\gamma\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\|\hat{\gamma}-\mathbb{E}\hat{\gamma}\|^2}_{\text{variance}=trace[cov(\hat{\gamma})]} + \underbrace{2(\mathbb{E}\hat{\gamma}-\gamma)^\top\mathbb{E}[\hat{\gamma}-\mathbb{E}\hat{\gamma}]}_{=0}$$

Writing $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ $\text{trace}\{\text{cov}(\hat{\gamma})\} = \sum_{j=1}^{q}\frac{\lambda_i}{\lambda_i^2+\lambda}\sigma^2$

So choose λ so as to optimally to balance bias/variance

Use cross validation!

Motivated from Ridge Regression formulation can consider:

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} |\gamma_j| = \|\boldsymbol{\gamma}\|_1 \leq r(\lambda)$$

$$\Longleftrightarrow$$

$$\text{min!} \qquad \|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1.$$

Shrinks coefficient *size* by different version of *magnitude*.

- Resulting estimator non–linear in $\boldsymbol{Y}$
- No explicit form available (unless $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}$), needs quadratic programming algorithm

Why choose a different type of norm?
L^1 penalty (almost) produces a "continuous" model selection!

When the explanatory variables are orthogonal (i.e. $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}$), then the LASSO[7] exactly performs model selection via thresholding:

Theorem (Orthogonal Design $\leftrightarrow$ Model Selection)

Consider the linear model

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{1\times 1}{\beta_0}\,\underset{n\times 1}{\mathbf{1}} + \underset{n\times(p-1)}{\boldsymbol{Z}}\,\underset{(p-1)\times 1}{\boldsymbol{\gamma}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{Z}^\top \mathbf{1} = 0$ *and* $\boldsymbol{Z}^\top \boldsymbol{Z} = I$. *Let* $\hat{\boldsymbol{\gamma}}$ *be the least squares estimator of* $\boldsymbol{\gamma}$,

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} = \boldsymbol{Z}^\top \boldsymbol{Y}.$$

Then, the unique solution to the LASSO problem

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \left\{ \|\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right\}$$

is given by $(\hat{\beta}_0, \check{\boldsymbol{\gamma}}) = (\beta_0, \check{\gamma}_1, \ldots, \check{\gamma}_{p-1})$, *defined as*

$$\hat{\beta}_0 = \bar{Y} \qquad \& \qquad \check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i) \left(|\hat{\gamma}_i| - \frac{\lambda}{2} \right)_+, \quad i = 1, \ldots, p-1.$$

[7] Lease Absolute Shrinkage and Selection Operator.

Proof (∗).

Note that since $Z^\top \mathbf{1} = 0$ and since β_0 does not appear in the L^1 penalty, we have

$$\hat{\beta}_0 = (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1} Y = \bar{Y}.$$

Thus, the LASSO problem reduces to

$$\min_{\beta_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{p-1}} \left\{ \| \boldsymbol{Y} - \beta_0 \mathbf{1} - Z\gamma \|_2^2 + \lambda \|\gamma\|_1 \right\} = \min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \| \boldsymbol{u} - \boldsymbol{Z}\gamma \|_2^2 + \lambda \|\gamma\|_1 \right\}.$$

where $\boldsymbol{u} = \boldsymbol{Y} - \bar{Y}\mathbf{1}$ for tidiness. Expanding $\|\boldsymbol{u} - \boldsymbol{Z}\gamma\|_2^2$ gives

$$\boldsymbol{u}^\top \boldsymbol{u} - 2\boldsymbol{u}^\top \boldsymbol{Z}\gamma + \gamma^\top \underbrace{(\boldsymbol{Z}^\top \boldsymbol{Z})}_{=I} \gamma = \boldsymbol{u}^\top \boldsymbol{u} - 2\underbrace{\boldsymbol{Y}^\top \boldsymbol{Z}}_{=\hat{\gamma}^\top} \gamma + 2\bar{Y} \underbrace{\mathbf{1}^\top \boldsymbol{Z}}_{=0} \gamma + \gamma^\top \gamma$$

Since $\boldsymbol{u}^\top \boldsymbol{u}$ does not depend on γ, we see that the LASSO objective function is

$$-2\hat{\gamma}^\top \gamma + \|\gamma\|_2^2 + \lambda \|\gamma\|_1.$$

Clearly, this has the same minimizer if multiplied across by $1/2$, i.e.

$$-\hat{\gamma}^\top \gamma + \tfrac{1}{2}\|\gamma\|_2^2 + \tfrac{1}{2}\lambda \|\gamma\|_1 = \textstyle\sum_{i=1}^{p-1} \left(-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i| \right).$$

Notice that we now have a sum of $p-1$ objective functions, each depending only on one γ_i. We can thus optimise each separately. That is, for any given $i \leq p-1$, we must minimise

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|.$$

We distinguish 3 cases:

1. Case $\hat{\gamma}_i = 0$. In this case, the objective function becomes $\frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ and it is clear that it is minimised when $\gamma_i = 0$. So in this case $\check{\gamma}_i = 0$.
2. Case $\hat{\gamma}_i > 0$. In this case, the objective function $-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in [0, \infty)$ because the term $-\hat{\gamma}_i \gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \geq 0$, the objective function becomes

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}\gamma_i = \left(\frac{\lambda}{2} - \hat{\gamma}_i \right) \gamma_i + \frac{1}{2}\gamma_i^2.$$

   If $\frac{\lambda}{2} - \hat{\gamma}_i \geq 0$, then the minimum over $\gamma_i \in [0, \infty)$ is clearly at $\gamma_i = 0$. Otherwise, when $\frac{\lambda}{2} - \hat{\gamma}_i < 0$, we differentiate and find the minimum at $\gamma_i = \hat{\gamma}_i - \lambda/2 > 0$. In summary, $\check{\gamma}_i = (\hat{\gamma}_i - \lambda/2)_+ = \operatorname{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

3. Case $\hat{\gamma}_i < 0$. In this case, the objective function $-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in (-\infty, 0]$ because the term $-\hat{\gamma}_i\gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \leq 0$, the objective function becomes

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}(-\gamma_i) = \left(\frac{\lambda}{2} + \hat{\gamma}_i\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2 = \left(\frac{\lambda}{2} - |\hat{\gamma}_i|\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2.$$

If $\frac{\lambda}{2} - |\hat{\gamma}_i| \geq 0$, then the minimum over $\gamma_i \in (-\infty, 0]$ is clearly at $\gamma_i = 0$, since $-\gamma_i$ ranges over $[0, \infty)$. Otherwise, when $\frac{\lambda}{2} - |\hat{\gamma}_i| < 0$, we differentiate and find the minimum at $\gamma_i = -|\hat{\gamma}_i| + \lambda/2 < 0$, which we may re-write as:

$$-|\hat{\gamma}_i| + \lambda/2 = -\left(|\hat{\gamma}_i| - \lambda/2\right) = \operatorname{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \lambda/2\right).$$

In summary, $\check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

The proof is now complete, as we can see that all three cases yield

$$\check{\gamma}_i = \operatorname{sgn}(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \frac{\lambda}{2}\right)_+, \quad i = 1, ..., p-1.$$

□

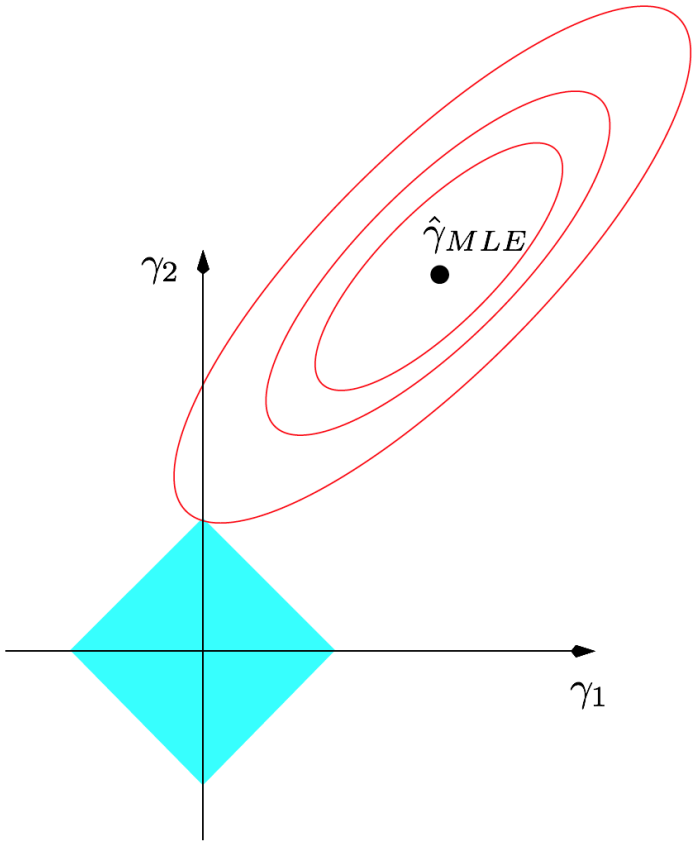


Figure: L^1 Shrinkage (the LASSO)

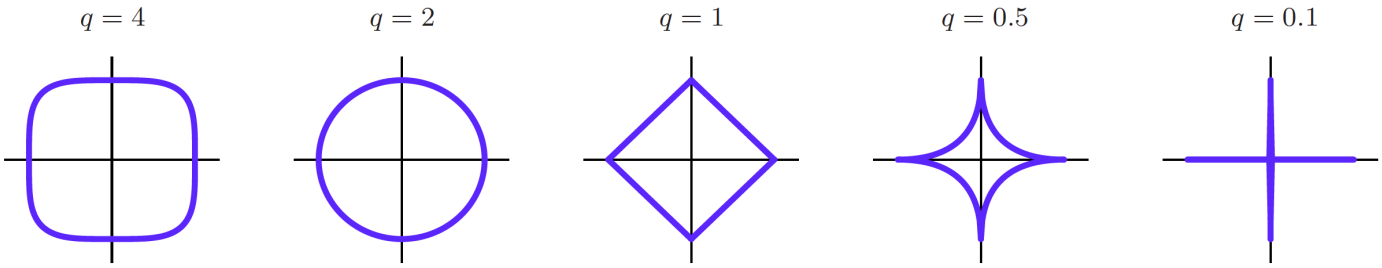
Intuition: L_1 norm induces "sharp" balls!

- Balls more concentrated around the axes
- Induces model selection by regulating the LASSO (through λ)

Extreme case: L^0 "Norm", gives best subset selection!

$$\|\gamma\|_0 = \sum_{j=1}^{p-1} |\gamma_j|^0 = \sum_{j=1}^{p-1} \mathbf{1}_{\{\gamma_j \neq 0\}} = \#\{j : \gamma_j \neq 0\}$$

Generally: $\|\gamma\|_q^q = \sum_{j=1}^{p-1} |\gamma_j|^q$, sharp balls for $0 < q \leq 1$



But L^1 gives sharpest convex ball among these.

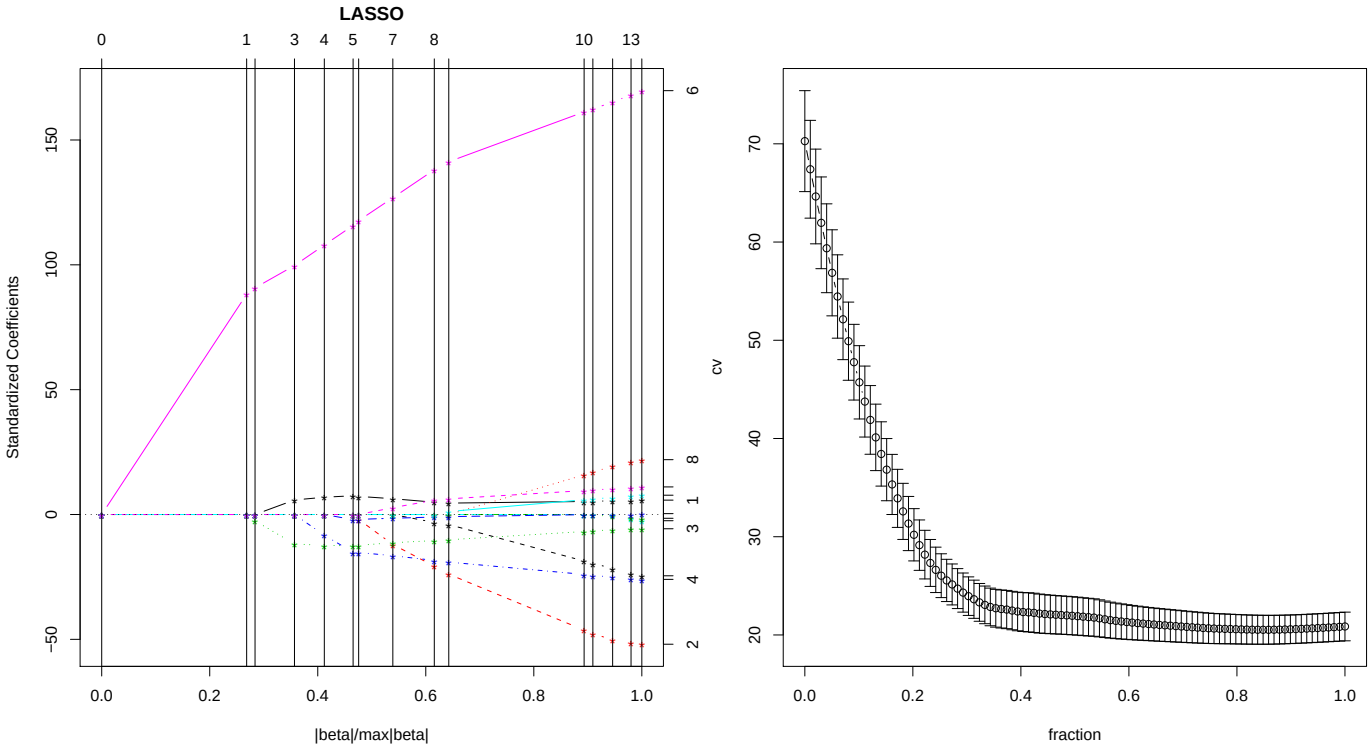


Figure: LASSO and CV for different values of $r(\lambda)/\|\hat{\gamma}\|_1$ for the bodyfat data (LARS algorithm)

Generalised Linear Models

Generalised Linear Models (GLM)

Back to the big picture:

$$\boxed{Y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}}$$

General formulation:

$$Y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

Distribution / Function g	$g(x_i^\top) = x_i^\top \beta$	g nonparametric
Gaussian	Linear Regression✓	Smoothing
Exponential Family	GLM ←	GAM

Generalised Linear Models: regression with exponential family responses!

GLM for $Y_1, ..., Y_n$

Response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ has independent entries with joint law

$$f(\boldsymbol{y}; \boldsymbol{\phi}) = \prod_{i=1}^{n} \exp\{\phi_i y_i - \gamma(\phi_i) + S(y_i)\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{y} - \sum_{i=1}^{n} \gamma(\phi_i) + \sum_{i=1}^{n} S(y_i)\right\},$$

where $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^n$ is the natural parameter with Φ open. The parameter varies as a function of the covariates via

$$\boldsymbol{\phi} = \boldsymbol{X}_n \boldsymbol{\beta},$$

for $\boldsymbol{X}_n$ the $n \times p$ covariate matrix of rank p and $\boldsymbol{\beta}$ a p-dimensional parameter.

In our general notation

$$Y_i \overset{\text{independent}}{\sim} f(y_i; \phi_i) = \exp\{\underbrace{(\boldsymbol{x}_i^\top \boldsymbol{\beta})}_{=\phi_i} y_i - \gamma(\underbrace{\boldsymbol{x}_i^\top \boldsymbol{\beta}}_{=\phi_i}) + S(y_i)\}, \qquad i = 1, ..., n.$$

where the row vector $\boldsymbol{x}_i^\top$ is the ith row of $\boldsymbol{X}_n$.

Comments:

- Notice that the the sufficient statistic for each marginal distribution $f(y; \phi_i)$ was taken to be the identity $T(Y) = Y$.
- This does not incur any loss of generality for two reasons:
  1. In the three main GLM of interest (Gaussian, Bernoulli, and Poisson) the natural statistic is for $f(y; \phi_i)$ is indeed the identity.
  2. More generally, since we only observe a single observation from each $f(y; \phi_i)$, if the natural statistic were $T(Y_i) \neq Y_i$, we could re-define the response to just be $T_i = T(Y_i)$. The sampling distribution of T_i can be shown to also be a one-parameter exponential family with the same natural parameter.
- Recall from our sampling theory results:
  - $\mu_i = \mathbb{E}[Y_i] = \dfrac{\partial}{\partial \phi_i} \gamma(\phi_i)$
  - $\text{var}[Y_i] = \dfrac{\partial^2}{\partial \phi_i^2} \gamma(\phi_i)$
  - So if γ is invertible (which it is for the three main examples), the variance can be written as a function of the mean:

$$\text{var}[Y_i] = \gamma''([\gamma']^{-1}(\mu_i)) = V(\mu_i).$$

Interpretation of $\phi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$?

- In key cases ϕ_i is directly interpretable. If not, can switch perspective using the mean μ_i as defining parameter, connected to the linear predictor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ via

$$[\gamma']^{-1}(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \phi_i$$

- The function $[\gamma']^{-1}(\cdot)$ is called the natural link function.
- Instead of $[\gamma']^{-1}$ can use other link functions $g(\cdot)$ and postulate

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

  This will also yield a GLM, but now the natural parameter will not be equal to the linear predictor but to some function $u(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ of it.
- In summary, the nomenclature is:
  - $g(\cdot)$ is the link function
  - $h = g^{-1}$ is the inverse link function
  - $g(\cdot) = [\gamma']^{-1}(\cdot)$ is the natural link function
- Will focus on natural links but methods/results generalise quite easily.

Likelihood and IRLS

With natural link, the loglikelihood (up to constants w.r.t. $\boldsymbol{\beta}$) is

$$\ell_n(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^n \gamma(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

for $\boldsymbol{x}_i^\top$ the ith row of $\boldsymbol{X}_n$. The corresponding $p \times 1$ derivative (*score function*) is

$$\nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{Y} - \sum_{i=1}^n \boldsymbol{x}_i \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i \left(Y_i - \mu_i \right) = \boldsymbol{X}_n^\top \left(\boldsymbol{Y} - \boldsymbol{\mu} \right)$$

with $p \times p$ covariance equaling the information matrix and given by

$$\begin{aligned} \operatorname{cov} \left\{ \nabla_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) \right\} &= \sum_{i=1}^n \boldsymbol{x}_i^\top \operatorname{cov} \left\{ Y_i - \mu_i \right\} \boldsymbol{x}_i = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}) \boldsymbol{X}_n \\ &= -\nabla^2_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \mathcal{I}_n(\boldsymbol{\beta}), \end{aligned}$$

where $\operatorname{cov}\{\boldsymbol{Y}\} = \boldsymbol{V}(\boldsymbol{\beta}) \succ 0$ is diagonal, with ith diagonal element

$$\operatorname{var}\{Y_i\} = \gamma''(\phi_i) = \gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Thus, if the MLE $\hat{\boldsymbol{\beta}}$ exists it is also unique, and must satisfy:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \left(Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \right) = 0$$

By a first order Taylor expansion of γ', we have

$$\gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \approx \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})$$

for some guesstimate $\tilde{\boldsymbol{\beta}}$ near $\hat{\boldsymbol{\beta}}$. Plugging into the score equation yields

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top \left(Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) + \boldsymbol{x}_i^\top (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \right) \approx 0$$

$$\implies \sum_{i=1}^{n} \gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \boldsymbol{x}_i^\top (Z_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{Z}} - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}) \approx 0$$

where we defined $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_n)^\top$ to be the adjusted response

$$\tilde{Z}_i = \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} + \frac{1}{\gamma''(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})}(Y_i - \gamma'(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})) \quad \text{so} \quad \tilde{\boldsymbol{Z}} = \boldsymbol{X}_n \tilde{\boldsymbol{\beta}} + \boldsymbol{V}^{-1}(\tilde{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})).$$

The last expression for the score expression now yields:

$$\hat{\boldsymbol{\beta}} \approx (\boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}}) \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{V}(\tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{Z}}$$

Just a weighted least squares estimate! Where's the catch?

- Weight matrix $\boldsymbol{V}(\tilde{\boldsymbol{\beta}})$ requires specification of an initial guesstimate $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}$ sufficiently close to $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$.
- Luckily we can give such a guesstimate by recalling that $\mu_i = \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ so that estimating μ_i by Y_i yields the guesstimate $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} \equiv (\gamma')^{-1}(Y_i)$.
- Suggests the following Iteratively Reweighted Least Squares (IRLS)

IRLS

1. Initialize with $\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)} \leftarrow \gamma'(Y_i)$ and $Z_i^{(0)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)} + \frac{(Y_i - \gamma'(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)}))}{\gamma''(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(0)})}$
2. Update with $\boldsymbol{\beta}^{(j+1)} \leftarrow (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}^{(j)}) \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta}^{(j)}) \boldsymbol{Z}^{(j)}$

- Equivalent to Newton-Raphson iteration.
- Not always guaranteed to converge.

Heuristics:

- Suppose we had started iteration at true $\boldsymbol{\beta}$ and stopped at first iterate:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{Z}$$

  where

$$Z_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \frac{1}{\gamma''(\boldsymbol{x}_i^\top\boldsymbol{\beta})}(Y_i - \gamma'(\boldsymbol{x}_i^\top\boldsymbol{\beta})) \quad \text{so} \quad \boldsymbol{Z} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{V}^{-1}(\boldsymbol{\beta})(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

- This would give us,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

- So we would expect $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\text{cov}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}_n^\top \boldsymbol{V}^{-1}(\boldsymbol{\beta})\boldsymbol{X}_n)^{-1} = \mathcal{I}_n^{-1}(\boldsymbol{\beta})$.
- And we would conjecture a Gaussian limiting law (paralleling the IID setting)

Under conditions, this is indeed what we obtain.

Result stated in form valid for more general (sufficiently regular) link functions.

Theorem (Asymptotic Normality of MLE in GLM)

In the same context and notation as before, assume that:

(C1) $\boldsymbol{\beta} \in B$ *for* B *an open convex subset of* $\mathbb{R}^p$.

(C2) *The* $p \times p$ *matrix* $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ *is of full rank for all* n.

(C3) *The information diverges, i.e.* $\lambda_{min}\Big(\mathcal{I}_n(\boldsymbol{\beta})\Big) \to \infty$ *as* $n \to \infty$ *for* $\lambda_{\min}(\cdot)$ *the smallest eigenvalue.*

(C4) *Given any parameter* $\boldsymbol{\beta} \in \mathbb{R}^p$ *it holds that*

$$\sup_{\boldsymbol{\alpha} \in N_\delta(\boldsymbol{\beta})} \left\| \mathcal{I}_n^{-1/2}(\boldsymbol{\beta})\mathcal{I}_n^{1/2}(\boldsymbol{\alpha}) - \boldsymbol{I}_{p\times p} \right\| \to 0$$

$\forall\, \delta > 0$, *where* $N_\delta(\boldsymbol{\beta}) = \{\boldsymbol{\alpha} \in \mathbb{R}^p : (\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathcal{I}_n(\boldsymbol{\beta})(\boldsymbol{\alpha} - \boldsymbol{\beta}) \leq \delta\}$.

Then, as $n \to \infty$, *provided it exists, the MLE* $\hat{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$ *is unique & satisfies*

$$\mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{I}_{p\times p}).$$

- Recall that under canonical link $\mathcal{I}_n(\boldsymbol{\beta}) = \boldsymbol{X}_n^\top \boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}_n$.

Comments on conditions:

- (C1) implies that $\boldsymbol{X}_n\boldsymbol{\beta}$ ranges over an open set, and so γ is infinitely differentiable, and our exponential family possesses all moments.
- (C2) is as with our linear model, and essentially means that our covariates should not be perfectly correlated.
- (C3) is similar to the "balanced design" assumption we had for the asymptotics of a non-Gaussian linear model.
- (C1-C3) are up to us: they depend on the design matrix $\boldsymbol{X}_n$, which in principle is for us to choose.
- (C4) asks that the (root) information matrix converge uniformly on compact ellipsoids centred at the true parameter.

Comments on conclusion:

- Can also be read as saying that for n sufficiently large,

$$\hat{\boldsymbol{\beta}}_n \overset{d}{\approx} N(\boldsymbol{\beta}, \mathcal{I}_n^{-1}(\boldsymbol{\beta}))$$

- Conclusion also immediately implies that

$$(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \mathcal{I}_n(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \chi^2_p.$$

- Allows adapting testing/CI developed before using LR and Wald statistics.

Goodness of Fit and Nested Models: Deviance

In Gaussian linear regression we used sums of squares to measure fit and compare nested models. What about in GLM?

Idea: compare best possible to observed maximised loglikelihood

- Let $\ell_n(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ be the maximised loglikelihood.
- Define the saturated model to be that which has

$$\#\text{parameters} = \#\text{observations}$$

  i.e. where we replace $\boldsymbol{X}_n\boldsymbol{\beta}$ by some unconstrained $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top \in \mathbb{R}^n$. (and thus we also replace $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ by η_i).
- Let $\hat{\boldsymbol{\eta}}$ be the maximiser of $\ell_n(\boldsymbol{\eta}) = \boldsymbol{\eta}^\top \boldsymbol{Y} + \sum_{i=1}^n \gamma(\eta_i)$ w.r.t. $\boldsymbol{\eta}$.
- Define the saturated loglikelihood as $\ell_n(\hat{\boldsymbol{\eta}})$.

Definition ((Scaled) Deviance)

$$D = 2\big(\ell_n(\hat{\boldsymbol{\eta}}) - \ell_n(\hat{\boldsymbol{\beta}})\big) = 2\Big((\hat{\boldsymbol{\eta}} - \boldsymbol{X}_n\hat{\boldsymbol{\beta}})^\top \boldsymbol{Y} + \sum_{i=1}^{n}(\gamma(\hat{\eta}_i) - \gamma(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))\Big)$$

Comments:

- Always $D \geq 0$ (why?)
- Small D implies a good model fit ($\boldsymbol{X}_n\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\eta}}$).
- Large D implies poor fit.
- In Gaussian case: deviance $\equiv$ residual sum of squares (exercise).
- Can now use deviance differences to mimic sum-of-square ratios and construct a GLM variant of the F-test.

Deviance Comparisons for Nested Models

- Consider the problem of comparing two nested models:
  - Model A: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ vary freely — MLE $\hat{\boldsymbol{\beta}}^A$
  - Model B: for $q < p$, $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence only q free parameters, with MLE $\hat{\boldsymbol{\beta}}^B$
- Model B is *nested within* model A: B can be obtained by restrictions on A
  - More generally, could have Model B with $\boldsymbol{\beta}$ constrained to vary in a subspace $\mathcal{V}$ of dimension $q < p$, which we can write as $\boldsymbol{\beta} = \boldsymbol{Q}_{p\times q}\underbrace{\boldsymbol{\zeta}}_{q\times 1}$ for $\mathcal{M}(\boldsymbol{Q}) = \mathcal{V}$.
- Likelihood ratio test statistic for comparing the models is

$$2\big(\ell_n(\hat{\boldsymbol{\beta}}^A) - \ell_n(\hat{\boldsymbol{\beta}}^B)\big) = D_B - D_A,$$

and when model B is correct $D_B - D_A \xrightarrow{d} \chi^2_{p-q}$.

Main idea: use deviance instead of sums-of-squares and use final iterate of IWLS to get hat matrix

- Leverage h_{jj} defined as jth diagonal element of

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n(\boldsymbol{X}_n^\top \boldsymbol{V}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}}),$$

- Cook statistic now becomes the change in deviance

$$2p^{-1}\left\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{-j})\right\},$$

  where $\hat{\boldsymbol{\beta}}_{-j}$ is MLE when jth case $(\boldsymbol{x}_j^\top, Y_j)$ is dropped.

- Cook statistic can be approximated by

$$C_j = \frac{h_{jj}}{p(1-h_{jj})} r_{Pj}^2,$$

  where r_{Pj} is standardised Pearson residual (to be defined in next slide).

- Deviance residual:

$$d_j = \text{sgn}(\hat{\eta}_j - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j)\Big[2\{\underbrace{\eta_j\, Y_j + \gamma(\eta_j)}_{\ell_j(\hat{\boldsymbol{\eta}})} - \underbrace{[(\boldsymbol{x}_j^\top \boldsymbol{\beta})\, Y_j + \gamma(\boldsymbol{x}_j^\top \boldsymbol{\beta})]}_{\ell_j(\hat{\boldsymbol{\beta}})}\}\Big]^{1/2},$$

  for which we note that

$$\sum_{j=1}^{n} d_j^2 = D$$

  gives the deviance (in analogy with RSS in Gaussian linear regression).

- Pearson residual:

$$p_j = \frac{Y_i - \gamma'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{\sqrt{\gamma''(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad \text{so } \boldsymbol{r}_P = \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})).$$

- Standardised versions:

$$r_{Dj} = \frac{d_j}{(1-h_{jj})^{1/2}} \qquad \& \qquad r_{Pj} = \frac{p_j}{(1-h_{jj})^{1/2}}.$$

We will now consider two fundamental specific GLM families:

1. Logistic Regression for Binary Data (Bernoulli GLM with natural link)
2. Loglinear Regression for Count Data (Poisson GLM with natural link)

These will give us concrete situations to keep in mind, demonstrating concepts already presented in generality.

We will conclude with remarks on the notion of a scale parameter.

$$\text{Very often have response} \to Y = \begin{cases} 1, & \text{“success”} \\ 0, & \text{“failure”} \end{cases}$$

So Y has a very simple Bernoulli structure:

$$\mathbb{P}[Y = 1] = \pi = 1 - \mathbb{P}[Y = 0], \quad \mathbb{E}\{Y\} = \pi$$

- Regression: need to connect response Y with an explanatory x.
- Use GLM. Can postulate that $g(\pi) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for some link g. Why?
- Intuition: Depending on circumstances, can imagine Y arising as

$$Y = \mathbf{1}\{Z > 0\} \implies \mathbb{P}[Y = 1] = \pi = 1 - F_Z(0)$$

  i.e. describing the level of a “hidden” variable Z:
- Now suppose $Z = \boldsymbol{x}^\top \boldsymbol{\alpha} + \sigma\varepsilon$. Then

$$\pi_i = 1 - F_Z(0) = 1 - F_\varepsilon\left(-\boldsymbol{x}^\top(\sigma^{-1}\boldsymbol{\alpha})\right) = 1 - F_\varepsilon(-\boldsymbol{x}^\top\boldsymbol{\beta})$$

  ($(\boldsymbol{\alpha}, \sigma)$ unidentifiable, but $\boldsymbol{\beta} = \sigma^{-1}\boldsymbol{\alpha}$ is ok)

So $g(x) = -F^{-1}(1-x)$ can serve as a link function

$$1 - \pi = F(-\boldsymbol{x}^\top \boldsymbol{\beta}) \implies -F^{-1}(1-\pi) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

Choice of Link $\iff$ Choice of Error Distribution F_ε

Distribution $F_\varepsilon(u)$		Link function $g(\pi)$	
Logistic	$e^u/(1+e^u)$	Logit	$\log\{\pi/(1-\pi)\}$
Normal	$\Phi(u)$	Probit	$\Phi^{-1}(\pi)$
Log Weibull	$1-\exp(-\exp(u))$	Log-log	$-\log\{-\log(\pi)\}$
Gumbel	$\exp\{-\exp(-u)\}$	Complementary log-log	$\log\{-\log(1-\pi)\}$

- Logit and probit symmetric, hard to distinguish in practice
- Log-log and complementary log-log are asymmetric
- Logit (canonical link) is usual choice, with nice interpretation

Assuming independence:

$$\mathbb{P}[Y_i = y] \overset{ind}{\sim} \pi_i^y (1-\pi_i)^{1-y}, \quad y \in \{0,1\}, \qquad \text{with } g(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^p$$

▶ Suppose $\{1,...,N\} = M_1 \cup M_2 \cup ... \cup M_n, \quad M_k \cap M_q = \emptyset,\ k \neq q$

with $\boldsymbol{x}_i = \boldsymbol{c}_k$ for $i \in M_k$. Then we have a Binomial GLM:

$$\underbrace{R_j}_{\in[0,1]} | x_j \overset{ind}{\sim} \exp\left[m_j \left\{ \frac{r}{m_j} \log\left(\frac{\pi_j}{1-\pi_j} \right) + \log(1-\pi_j) \right\} + \log\binom{m_j}{r} \right]$$

$$\to g(\pi_j) = \boldsymbol{x}_j^\top \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad j = 1,...,n$$

with $m_j = |M_j|$, $j = 1,...,n \quad (\sum_j m_j = N)$.

M's are called covariate classes - se why important later.

Example (Challenger Catastrophe)

Challenger Space-shuttle exploded at launch on 28/1/1986, killing all seven astronauts on board. US Presidential Commission (incuding Richard Feynman) concluded that the cause was the leakage of gas due to behaviour of O-rings under low temp (temp at launch was an unusual $31^\circ F$).

Table: O-Ring Data

Mission #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Temp - °F	53	57	58	63	66	67	67	67	68	69	70
# Damaged	5	1	1	1	0	0	0	0	0	0	1
# Intact	1	5	5	5	6	6	6	6	6	6	5

Mission #	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Temp - °F	70	70	70	72	73	75	76	76	76	78	79	81
# Damaged	0	1	0	0	0	0	1	0	0	0	0	0
# Intact	6	5	6	6	6	6	5	6	6	6	6	6

Binary regression with natural (logit link): $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$

Interpretation?

- Unit change in x_{jk} yields additive change of logodds by β_k.
- Equivalently, unit change in x_{jk} results in multiplicative change of odds by e^{β_k}.
- In terms of parameter:

$$\pi_j = \frac{\exp\{\beta_0 + \beta_1 x_{j1} + \ldots + \beta_q x_{jq}\}}{1 + \exp\{\beta_0 + \beta_1 x_{j1} + \ldots + \beta_q x_{jq}\}}, \qquad p = q + 1$$

So

$$\frac{\partial}{\partial x_{jk}} \pi_j = \beta_k \pi_j (1 - \pi_j) \quad \text{(logistic equation!)}$$

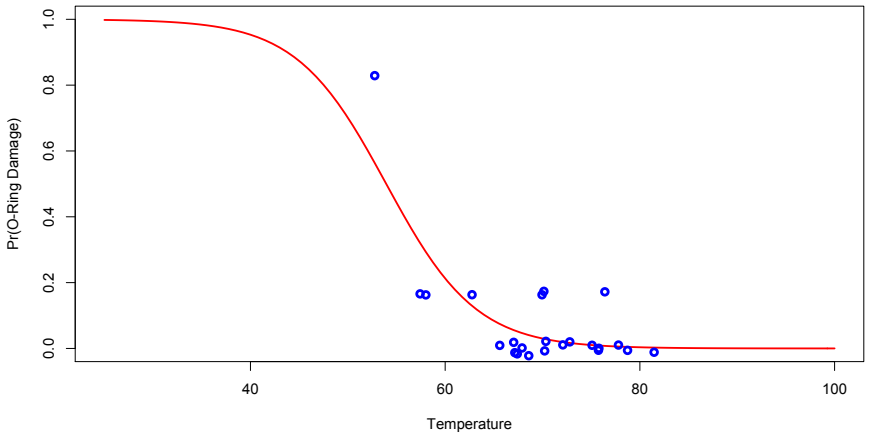
- Thus effects larger when π near $1/2$ than near endpoints of $[0, 1]$.

Example (Challenger Catastrophe, continued)

Logistic regression fit for probability of damage with temp as covariate:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 11.663 - 0.2162 \times t_i$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	−0.2162	0.0532	−4.07	0.0000



The Effect of Sparseness

Sparseness: covariate classes $\{M_j\}_{j=1}^n$ are "small":

- i.e. n is of the order of N
  - ↪ extreme: continuous covariate, $m_k = 1\ \forall k$, so $n = N$.

Sparseness affects interpretability of deviance:

- In extreme case deviance is only a function of $\hat{\pi}$ (exercise)
  - ↪ No contrast with data! (no information about fit in absolute sense). Similar problems with Pearson statistic. Problems with residuals also.
- $D \sim \chi^2_{N-p}$ breaks down even in non-extreme case, as this requires $m_i \to \infty$ as $N \to \infty$, so small m's can hurt us.
- Deviance reduction is reasonable for comparing nested models, though.
- Interpretability and accuracy of estimators remains the same!
- Rule of thumb: sparseness when $m_k \leq 5$ for several classes.
- A solution: grouping data! (i.e. merge into covariate classes)

Example (Birth Weight Data)

Births at US hospital: study the effect of various factors on the birth weight.

Response: Binary$= \begin{cases} 1, & \text{if birth weight is less than 2.5 kg;} \\ 0, & \text{otherwise.} \end{cases}$

Have data on 189 births. Variables are:

- `low` (response): birth weight less than 2.5 kg (binary)
- `age`: age of mother in years.
- `lwt`: weight of mother (lbs) at menstrual period
- `race`: white/black/other
- `smoke`: smoking status during pregnancy
- `ptd`: number of previous premature labours (none/at least one)
- `ht`: history of hypertension (0/1)
- `ui`: has uterine irritability
- `ftv`: number of physician visits the last trimester (0/1/at least two)

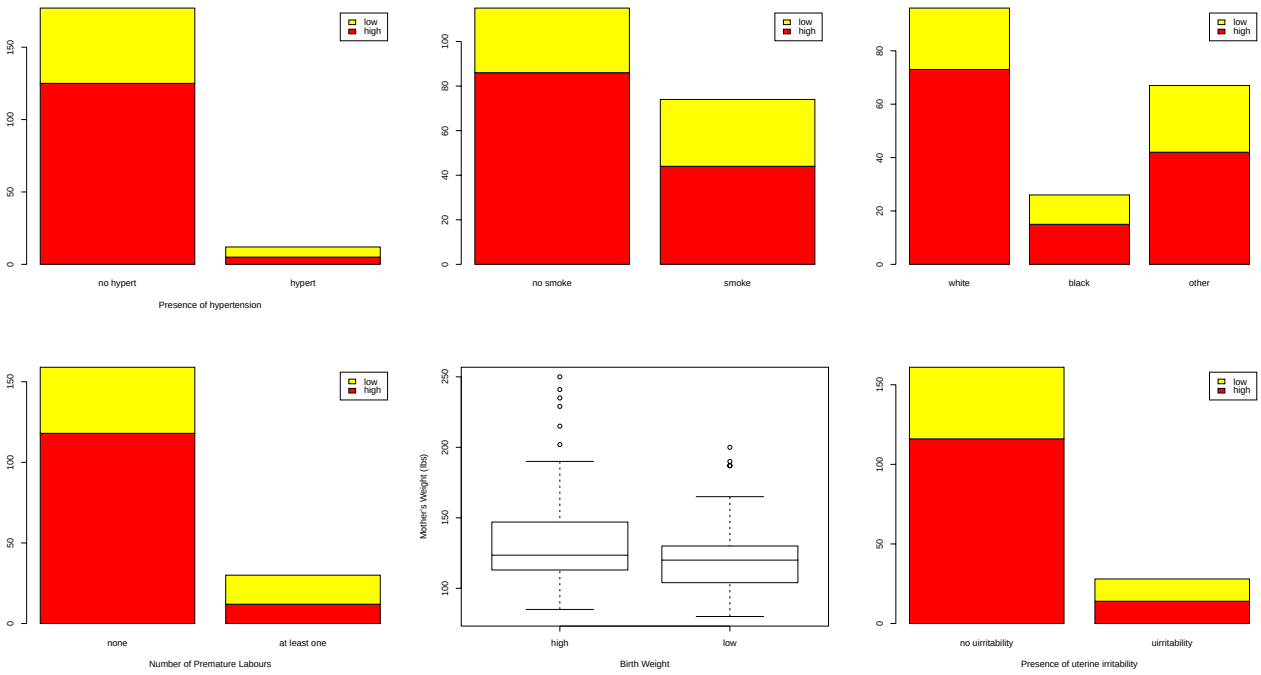


Figure: Exploratory analysis

Example (Birth Weight Data, continued)

Choose natural link function:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

Possible choice of covariates (based on exploratory analysis):

$$\texttt{low} \sim 1 + \texttt{lwt} + \texttt{race} + \texttt{ptd} + \texttt{ht} + \texttt{smoke} + \texttt{ui}$$

One may use AIC model selection on all of the possible variables$\longrightarrow$`stepAIC()`.
Thus we write:

```
model<-glm(low~1+lwt+race+ptd+ht+smoke+ui,family=binomial)
```


Example (Birth Weight Data, continued)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.0466	0.9705	−0.05	0.9617
lwt	−0.0172	0.0071	−2.43	0.0152
raceblack	1.2776	0.5320	2.40	0.0163
raceother	0.8555	0.4401	1.94	0.0519
ptdat least one	1.1586	0.4499	2.58	0.0100
hthypert	1.9197	0.7107	2.70	0.0069
smokesmoke	1.0498	0.4393	2.39	0.0169
uiuirritability	1.1917	0.6076	1.96	0.0498

- Null deviance: 234.67 on 188 degrees of freedom
- Residual deviance: 196.68 on 180 degrees of freedom
- AIC: 214.68

Brillinger & Preisler (1983), Brillinger (1996) suggest the use of "jittered quantile residuals" for binary responses (a.k.a *quantile residuals*).

Idea:

- If $Y \sim F(y;\theta)$ continuous, then $U := F(Y;\theta)$ has a uniform distribution on $(0,1)$.
- If $\hat{\theta} \simeq \theta \implies \hat{U} := F(Y|\hat{\theta}) \overset{d}{\approx} \text{Unif}(0,1)$.
- Hence, obtain quantile residuals $R_i = \Phi^{-1}(F(Y_i;\hat{\theta})) \overset{d}{\approx} N(0,1)$.

In discrete case must be a little more careful:

- let $\hat{U}_i^{(1)} \sim \text{Unif}[0,\hat{\pi}_i]$ and $\hat{U}_i^{(2)} \sim \text{Unif}[\hat{\pi}_i,1]$.
- Jittering randomisation: $\hat{U}_i = \hat{U}_i^{(1)} Y_i + \hat{U}_i^{(2)}(1-Y_i)$
- And so $\Phi^{-1}(\hat{U}_i) \overset{d}{\approx} N(0,1)$ distribution, if fit is good and model correct.

May now construct some plots, e.g. normal probability plots, plots against covariates, plots against "linked fits".

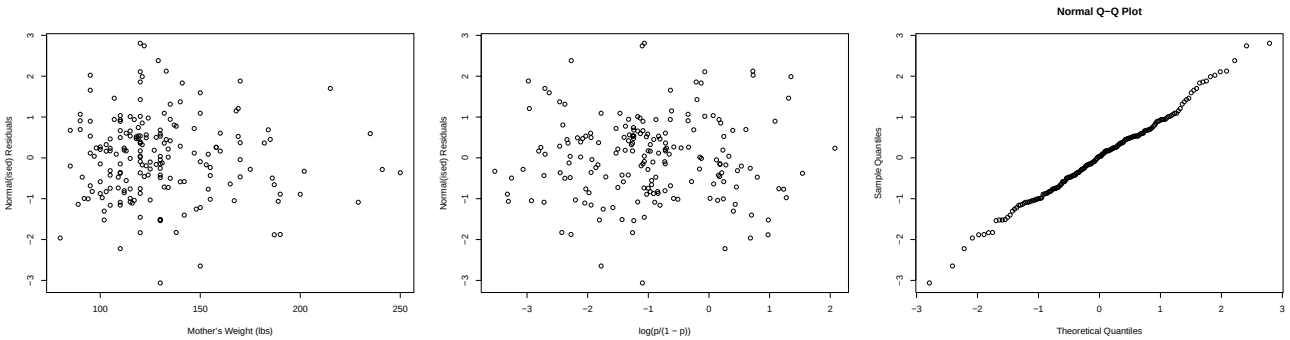


Figure: Jittered Quantile Residuals for Birth Weight Data

Example (The Donner Party)

Group of settlers ("pioneers") who set out for California in 1846. Failed to reach the Sierra Nevada before autumn, and had to face the winter head-on. Of 87 expedition members only 37 survived.

Model impact of age and sex on the survival probability of the 45 adults:

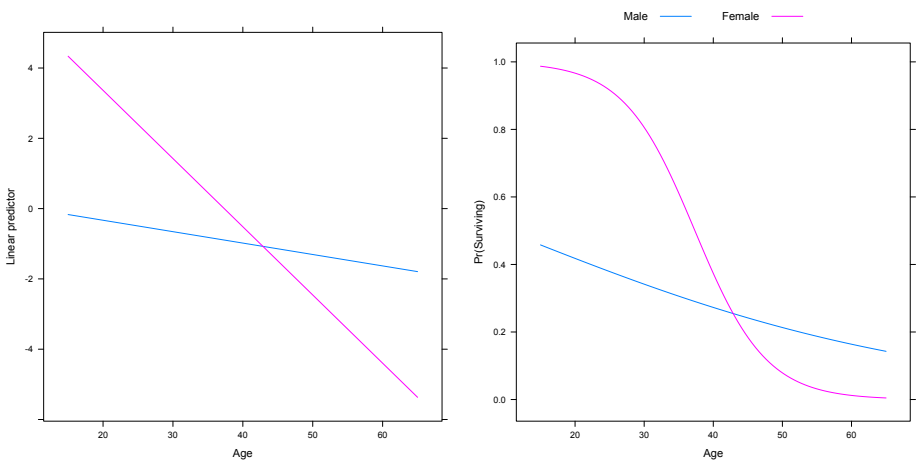
Table: Donner Party Survival Data

Status	**Age**	**Sex**
Died	23	Male
Survived	40	Female
Survived	40	Male
Died	30	Male
⋮	⋮	⋮
Survived	25	Female

Example (The Donner Party, continued)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.2464	3.2052	2.26	0.0238
Age	-0.1941	0.0874	-2.22	0.0264
SexMale	-6.9280	3.3989	-2.04	0.0415
Age:SexMale	0.1616	0.0943	1.71	0.0865

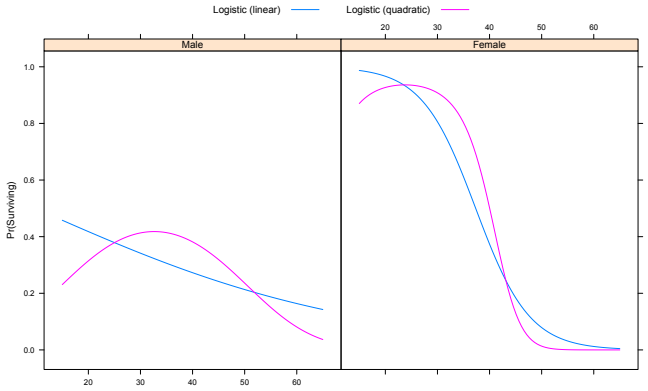
- Null deviance: 61.827 on 44 degrees of freedom
- Residual deviance: 47.346 on 41 degrees of freedom



Example (The Donner Party, continued)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0532	9.6843	-0.32	0.7526
Age	0.4829	0.6581	0.73	0.4631
SexMale	-0.2653	10.4552	-0.03	0.9798
Age^2	-0.0102	0.0103	-0.99	0.3222
Age:SexMale	-0.2999	0.6960	-0.43	0.6666
SexMale:Age^2	0.0074	0.0107	0.69	0.4913

- Null deviance: 61.827 on 44 degrees of freedom
- Residual deviance: 45.361 on 39 degrees of freedom



The Problem of Separation

Suppose we have data:

x_i	-0.14	2.13	1.11	-0.53	-6.25	-3.29	-0.04	1.07	0.55
Y_i	0	1	1	0	0	0	0	1	1

Logistic regression loglikelihood can be written as:

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - Y_i) \log(1 - \pi_i) \\ &= \sum_{j \in \mathcal{P}} \log\left(\frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}}\right) - \sum_{j \in \mathcal{P}^c} \log\left(1 + e^{\beta_0 + x_i\beta_1}\right).\end{aligned}$$

where $\mathcal{P} = \{i : x_i > 0\}$. For given β_0, what happens as $\beta_1 \to \infty$?

- Loglikelihood converges to zero! (likelihood converges to 1).
- So MLE does not exist!. Why? The problem is perfect separation.
- $\exists$ hyperplane perfectly separating covariates corresponding to 0's and 1's.
- More likely to occur when p is large relative to n.

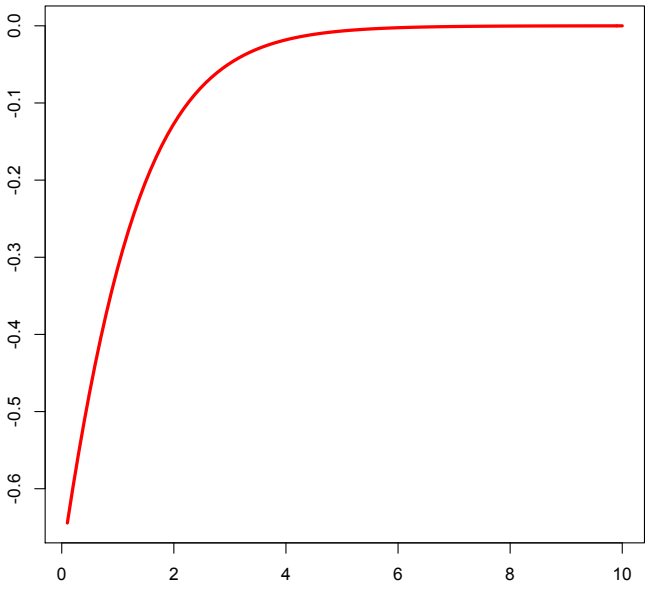


Figure: $\beta_1 \mapsto \ell(\beta_1)$

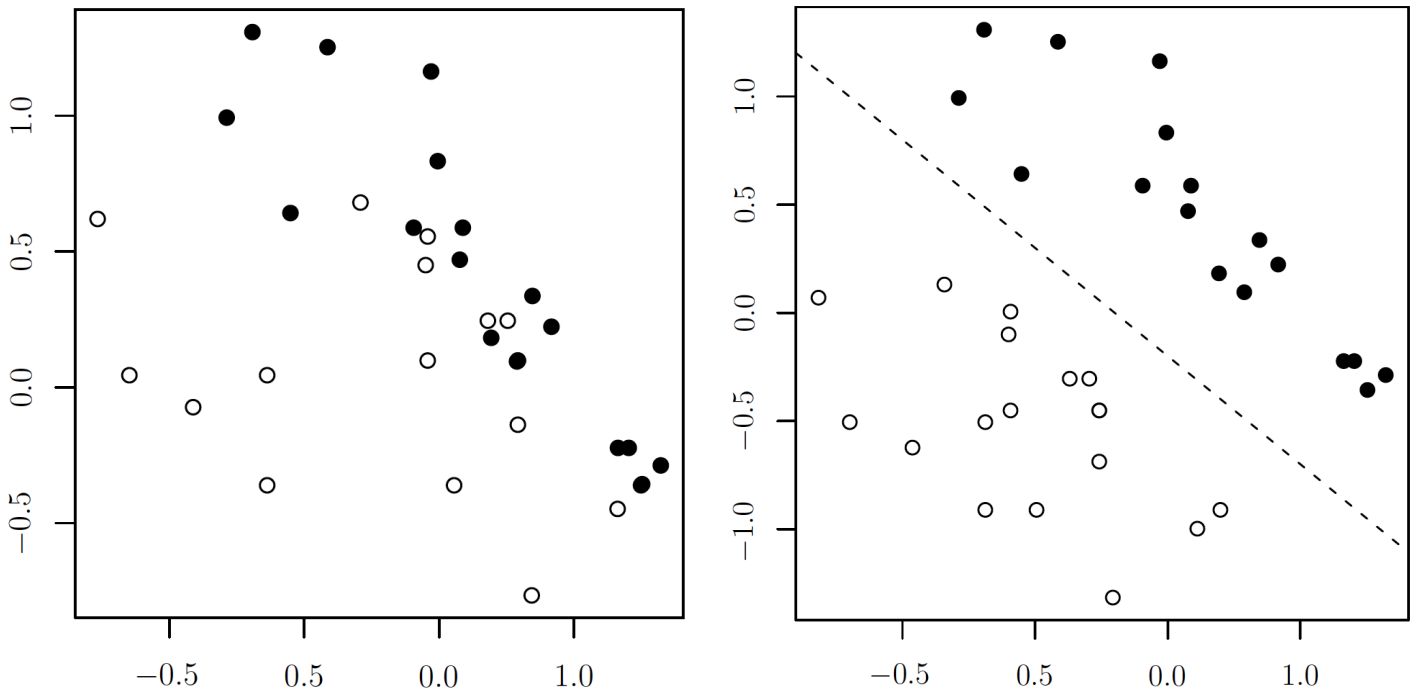


Figure: Overlap (left) vs Complete Separation (right)

We have complete separation when there exists $\gamma \in \mathbb{R}^p$ such that for all i

$$\{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma > 0\} \quad \& \quad \{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma < 0\}$$

Theorem

In the complete separation regime, the logistic regression MLE does not exist, and

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\beta) = 1.$$

Proof.

Let γ be such that $\{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma > 0\} \quad \& \quad \{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma < 0\}$.
Then we may write the loglikelihood of $t\gamma$ (for some $t > 0$) as

$$\ell(t\gamma) = \sum_{j \in \mathcal{P}} \log \left(\frac{e^{t(\boldsymbol{x}_i^\top \gamma)}}{1 + e^{t(\boldsymbol{x}_i^\top \gamma)}} \right) - \sum_{j \in \mathcal{P}^c} \log \left(1 + e^{t(\boldsymbol{x}_i^\top \gamma)} \right).$$

where $\mathcal{P} = \{i : \boldsymbol{x}_i^\top \gamma > 0\} = \{i : Y_i = 1\}$. The proof is complete upon noting:

1. For $t > 0$, $\boldsymbol{x}_i^\top \gamma > 0 \iff t(\boldsymbol{x}_i^\top \gamma) > 0$ and $\boldsymbol{x}_i^\top \gamma < 0 \iff t(\boldsymbol{x}_i^\top \gamma) < 0$.
2. As $t \to \infty$, $\ell(t\gamma) \to 0$.
3. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\ell(\boldsymbol{\beta}) < 0$ (replace $t(\boldsymbol{x}_i^\top \gamma)$ by $\boldsymbol{\beta}$ above and verify).

□

Ramifications:

- IWLS will fail to converge, with weights converging to zero.
- Standard errors will blow up.
- In a sense a design issue.
  - In Gaussian linear regression, $\boldsymbol{X}$ full rank $\implies$ MLE exists.
  - Binary regression is more subtle, and rank conditions alone do not suffice and instabilities can manifest in ways more subtle than multicollinearity.

Diagnostics and Remedies?

- Often get warning that iterations stopped after maxing out.
- But best keep track both of the likelihood value and the parameter values as the iteration evolves.
- Can remedy by imposing a penalty. Motivates regularised logistic regression:

$$\sum_{i=1}^{n} Y_i(\gamma_0 + \boldsymbol{x}_i^\top \gamma) - \log \left(1 + e^{\gamma_0 + \boldsymbol{x}_i^\top \gamma} \right) + \lambda \|\gamma\|_q^2$$

  for $q = 2$ (ridge) or $q = 1$ (lasso). Assuming $\boldsymbol{X}$ has been standardised, and

$$\boldsymbol{\beta}^\top = (\gamma_0, \gamma^\top).$$

Can we at least hope that MLE exists in the overlapping regime?

Theorem (Existence and Uniqueness)

In logistic regression with an intercept term and full rank design, the maximum likelihood estimator uniquely exists if and only if the covariates overlap.

The theorem actually applies more generally to other link functions than logit.

If the model postulates that $\pi_i = g(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, and the design includes an intercept, then overlap is a necessary and sufficient provided that:

1. $-\log(g^{-1}(t))$ and $-\log(1 - g^{-1}(t))$ are convex.
2. $g^{-1}(t)$ is strictly increasing at every t.
3. $0 < g^{-1}(t) < 1$ for all t.

2×2 Tables

Special case of Bernoulli/Binomial GLM: 2×2 Contingency Tables

- How does a single binary covariate affect a binary response?
- Say $x \in \{0, 1\}$ (control/case), $y \in \{0, 1\}$ (failure/success)
- Simple model: individuals are independent, with m_0 and m_1 persons in categories of $x \in \{0, 1\}$ and with success probabilities

$$\pi_0 = \frac{e^{\lambda}}{1 + e^{\lambda}}, \quad \pi_1 = \frac{e^{\lambda+\psi}}{1 + e^{\lambda+\psi}}$$

Yields independent binomial variables

$$W_1 \sim \text{Binomial}(m_1, \pi_1), \quad W_0 \sim \text{Binomial}(m_0, \pi_0)$$

and likelihood

$$L(\psi, \lambda) \propto \frac{e^{(r_0+r_1)\lambda + r_1\psi}}{(1 + e^{\lambda+\psi})^{m_1}(1 + e^{\lambda})^{m_0}}.$$

Table: Notation for 2×2 table.

	Success	Failure	Total
Case	R_1	$m_1 - R_1$	m_1
Control	R_0	$m_0 - R_0$	m_0
Total	$R_1 + R_0$	$m_1 + m_0 - R_1 - R_0$	$m_1 + m_0$

Example (Challenger Catastrophe, continued)

Table: Damage vs Temperature.

	Intact	Damaged	Total
High Temp ($\geq 18^\circ$C)	111	3	114
Low Temp ($< 18^\circ$C)	16	8	24
Total	127	11	138

- Key question: Does treatment affect success probability?
- In mathematics: is it true that $\pi_1 = \pi_0$? If not, by how much do they differ?
- Could consider absolute difference of risks, or probability ratio

$$\pi_1 - \pi_0, \quad \pi_1/\pi_0$$

- More common to consider difference of log odds

$$\psi = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right).$$

- This is natural parameter of exponential family.
- One must be quite careful with 2×2 tables – as the next example will illustrate.

Example (Women and Smoking)

- Data gathered by surveying people on electoral register in 1972–74.habits, ...
- Follow-up study 20 years later: see how many people had died.
- 162 women had smoked before 1972 but had stopped by 1972, and smoking habits were unknown for 18 women; these 180 women were excluded.

Table: Twenty-year survival and smoking status for 1314 women . The smoker and non-smoker columns contain number dead/total (% dead).

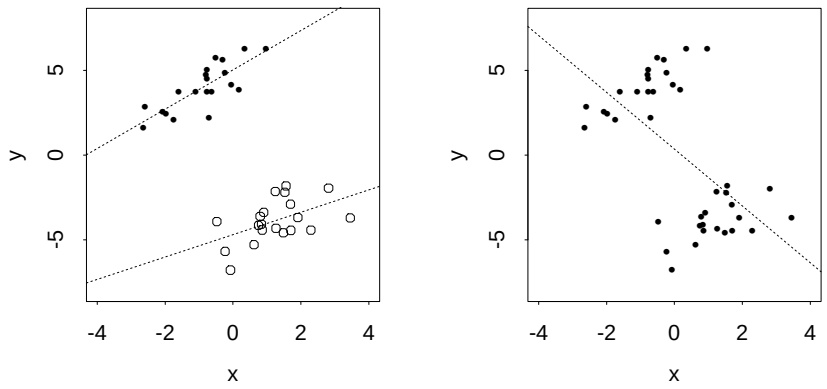
Age (years)	Smokers	Non-smokers
Overall	139/582 (24%)	230/732 (31%)
18–24	2/55 (4%)	1/62 (2%)
25–34	3/124 (2%)	5/157 (3%)
35–44	14/109 (13%)	7/121 (6%)
45–54	27/130 (21%)	12/78 (15%)
55–64	51/115 (44%)	40/121 (33%)
65–74	29/36 (81%)	101/129 (78%)
75+	13/13 (100%)	64/64 (100%)

Example (Women and Smoking, continued)

```
> data(smoking)
> summary(glm(cbind(dead,alive)~smoker,data=smoking,binomial))

Call:
glm(formula = cbind(dead, alive) ~ smoker, family = binomial,
    data = smoking)

..

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78052    0.07962  -9.803  < 2e-16 ***
smoker      -0.37858    0.12566  -3.013  0.00259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.5  on 13  degrees of freedom
Residual deviance: 632.3  on 12  degrees of freedom
```

Figure: Women and Smoking - Simpson's Paradox

- Problem: Smoking seems to reduce the death rate! Why?
- Explanation: Inappropriate marginalisation!
- Interested in how x affects Y, but a third binary variable z is lurking



- Marginalisation over z gives misleading inference about how Y depends on x: $\mathbb{E}[Y|X=x,Z=z]$ increases with x for each z (left panel), but $\mathbb{E}[Y|X=x]$ decreases with x (right panel)

Example (Women and Smoking, continued)

```
> summary(glm(cbind(dead,alive)~age+smoker-1,data=smoking,binomial))

Call:
glm(formula = cbind(dead, alive) ~ age + smoker - 1, family = binomial,
    data = smoking)

..

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
age18-24   -3.8601     0.5939  -6.500 8.05e-11 ***
age25-34   -3.7401     0.3715 -10.067  < 2e-16 ***
age35-44   -2.5190     0.2499 -10.079  < 2e-16 ***
age45-54   -1.7468     0.2157  -8.097 5.62e-16 ***
age55-64   -0.6793     0.1621  -4.190 2.78e-05 ***
age65-74    1.2279     0.1934   6.349 2.17e-10 ***
age75+     23.9472 11293.1430   0.002   0.9983
smoker      0.4274     0.1770   2.414   0.0158 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 902.7701  on 14  degrees of freedom
Residual deviance:   2.3809  on  6  degrees of freedom
```

Figure: Women and Smoking - Age Included

Assume response variables of interest Y_i takes values $y \in \{0, 1, 2, \ldots\}$ – perhaps with upper bound m

↪ depending on sampling scheme/experiment

Three standard models:

- unconstrained responses $Y_i \overset{indep}{\sim} \text{Poisson}(\mu_i)$
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j=1}^{d} Y_j = m$ having multinomial distribution, with probabilities $(\pi_1, \ldots, \pi_d)$ and denominator m.
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j \in I_k} Y_j = m_k$ (for disjoint index parition sets $\{I_k : k = 1, ..., K\}$) having product multinomial.

These models are *very* closely related.

Lemma (Poisson and Multinomial)

Let $Y_1, \ldots Y_d$ be independently distributed as Poisson, with means $\mu_1, ..., \mu_d$, respectively. Then the conditional distribution of

$$(Y_1, ..., Y_d) \text{ given } \sum_{k=1}^{d} Y_i = m$$

is multinomial with denominator m and probabilities $\pi_i = \mu_i / \sum_j \mu_j$.

Proof.

Using Bayes' formula, $\mathbb{P}\left[\bigcap_{i=1}^{d}\{Y_i = y_i\} \,\middle|\, \sum_j Y_j = m\right]$ equals

$$\begin{aligned}
&= \frac{\mathbb{P}[\sum_j Y_j = m | \bigcap_{i=1}^{d}\{Y_i = y_i\}]\mathbb{P}[\bigcap_{i=1}^{d}\{Y_i = y_i\}]}{\mathbb{P}[\sum_j Y_j = m]} \\
&= \mathbf{1}[\Sigma_{j=1}^{d} y_j = m] \frac{\prod_{j=1}^{d} e^{-\mu_j} \frac{\mu_j^{y_j}}{y_j!}}{e^{-\sum \mu_j} \frac{(\sum \mu_j)^m}{m!}} \\
&= \mathbf{1}[m = \Sigma_{j=1}^{d} y_j] \frac{\prod_{j=1}^{d} e^{-\mu_j} \frac{\mu_j^{y_j}}{y_j!}}{e^{-\sum \mu_j} \frac{(\sum \mu_j)^{\sum y_j}}{m!}} \\
&= \mathbf{1}[\Sigma_{j=1}^{d} y_j = m] \frac{m!}{y_1! \ldots y_d!} \prod_{i=1}^{d} \left(\frac{\mu_i}{\sum_{j=1}^{d} \mu_j}\right)^{y_i}
\end{aligned}$$

□

Example (Example: Smoking data (Doll and Hill))

Table: Lung cancer deaths in British male physicians. The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Example (Jacamar Data)

Table: Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	Aphrissa boisduvalli N/S/E	*Phoebis argante* N/S/E	*Dryas iulia* N/S/E	*Pierella luna* N/S/E	*Consul fabius* N/S/E	*Siproeta stelenes*† N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Assume that $Y_i \overset{indep}{\sim} \text{Pois}(\mu_i)$

$$\mathbb{P}[Y_i = y] = e^{-\mu}\frac{\mu^y}{y!}, \quad y \in \mathbb{Z}_+, \mu > 0$$

- Exponential family
- Natural parameter $\phi = \log \mu$
- Can fit GLM via some link function $g(\mu)$

$$\underbrace{Y_i}_{\in \mathbb{Z}_+} | x_i \overset{ind}{\sim} \text{Poisson}(\mu_i) \quad \text{such that} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \ \boldsymbol{\beta} \in \mathbb{R}^p, \quad i = 1, ..., n.$$

Log-linear model

$\updownarrow$

Poisson GLM with canonical logarithmic link:

$$\boldsymbol{x}_i^\top \boldsymbol{\beta} = \log \mu_i$$

- Occasionally Y_i counts the events of a Poisson process up to time T_i, so

$$\mathbb{E}[Y_i] = \mu_i = \lambda_i T_i$$

with λ_i the intensity of the process. In this case one sets

$$g(\mu_i) = \log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log T_i$$

- $\log T_i$ is the so-called offset term and is treat as a known constant.

Looks fairly straightforward. What's the big deal?

Earlier lemma suggests intimate relationship with categorical data.

↪ Consider again the binary case, $d = 2$.

$$Y_2 | \{ Y_1 + Y_2 = m \} \quad \sim \quad \text{Binomial}\left(m, \pi = \frac{\mu_2}{\mu_1 + \mu_2} \right)$$

- Hence if $\mu_1 = \exp(\gamma + \boldsymbol{x}_1^\top \boldsymbol{\beta})$, $\mu_2 = \exp(\gamma + \boldsymbol{x}_2^\top \boldsymbol{\beta})$,

$$\pi = \frac{\mu_2}{\mu_1 + \mu_2} = \frac{\mu_2 \mu_1^{-1}}{1 + \mu_2 \mu_1^{-1}} = \frac{\exp\{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^\top \boldsymbol{\beta}\}}{1 + \exp\{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^\top \boldsymbol{\beta}\}}.$$

- So we can estimate $\boldsymbol{\beta}$ using either a loglinear model or logistic model
  ↪ but can't estimate γ from logistic model (lose absolute information)

- This is particularly convenient for fitting more general contingency tables.

- Contingency table entries: count data cross-classified by different categories
  ↪ Example: jacamar data cross-classify butterflies by

  6 species × 8 colours × 3 fates

  yielding 144 categories total, each with count $\in \{0, 1, \ldots, 14\}$. Sampling scheme may fix certain totals — in the jacamar data the total for each species and colour is fixed, so reponses are trinomial: (not eaten, sampled, eaten)

Poisson vs multinomial vs product multinomial likelihoods (r=row, c=column):

- Poisson $\left(\prod_{r,c} \left\{ e^{-\mu_{rc}} \frac{\mu_{rc}^{Y_{rc}}}{Y_{rc}!} \right\} \right)$
  ↪ Just collect data, then arrange into table. Yields poisson distribution for each cell.
- Multinomial $\left(\frac{m!}{\prod_{r,c} Y_{rc}!} \prod_{r,c} \pi_{rc}^{Y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1 \right)$
  ↪ Keep collecting until $m = \sum_{rc} Y_{rc}$ is reached. Yields multinomial distribution for table entries.
- Product multinomial $\left(\prod_r \left\{ \frac{m_r!}{\prod_c Y_{rc}!} \prod_c \pi_{rc}^{Y_{rc}} \right\}, \ \sum_c \pi_{rc} = 1, \forall r \right)$
  ↪ Fix row totals alone in advance (e.g. fix # of butterflies in each colour/species category). In effect this treats row categories as independent subpopulations, i.e. independent multinomials for table entries of each row.

All three models can be easily fitted using Poisson GLM (with appropriate offsets).

- For multinomial settings, arrange as two-way layout with row totals fixed (single row in multinomial layout, several rows in product multinomial).
  - ↪ In Jacamar data, create new variable `species*colour` with 48 categories – yields 48 rows r, and leaves 3 columns c corresponding to `fate`.
- Model (r, c)-the cell as independent Poisson with mean

$$\mu_{rc} = \exp(\gamma_r + \boldsymbol{x}_{rc}^\top \boldsymbol{\beta})$$

  - γ_r acounts for the overall mean row count.
  - $\boldsymbol{x}_{rc}$ is such that $\sum_c \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} = \beta_{rc}$ which accounts for deviations of the cth column from the overall row count.
  - Interest focuses on $\boldsymbol{\beta}$, not γ_r, so will not worry about identifiability constraints.
- Conditioning on row totals being m_r get (product) multinomial model with probabilities $\{\pi_{rc} : \pi_{rc} \geq 0, \sum_c \pi_{rc} = 1\}$,

$$\pi_{rc} = \frac{\mu_{rc}}{\sum_d \mu_{rd}} = \frac{\exp(\gamma_r + \boldsymbol{x}_{rc}^\top \boldsymbol{\beta})}{\sum_d \exp(\gamma_r + \boldsymbol{x}_{rd}^\top \boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_{rc}^\top \boldsymbol{\beta})}{\sum_d \exp(\boldsymbol{x}_{rd}^\top \boldsymbol{\beta})},$$

  and the $\{\gamma_r\}$ parameters become irrelevant.

Thus multinomial loglikelihood is (up to constants)

$$\begin{aligned} \ell_{\text{Mult}}\left(\boldsymbol{\beta}; \boldsymbol{y} \middle| \sum_c Y_{rc} = m_r\right) &\equiv \sum_{r,c} Y_{rc} \log \pi_{rc} \\ &= \sum_r \left\{ \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - m_r \log\left(\sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}}\right)\right\}. \end{aligned}$$

The unconstrained Poisson model, would give loglikelihood (up to constants)

$$\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{r,c} Y_{rc} \log \mu_{rc} - \mu_{rc} = \sum_r \left(M_r \gamma_r + \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - e^{\gamma_r} \sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} \right)$$

where $M_r = \sum_c Y_{rc}$ is not given (i.e. is Poisson random variable). Writing

$$\tau_r = \sum_c \mu_{rc} = e^{\gamma_r} \sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} = \mathbb{E}[M_r]$$

for the row total means and using[8] $\gamma_r = \log \tau_r - \log\left\{\sum_c \exp(\boldsymbol{x}_{rc}^\top \boldsymbol{\beta})\right\}$ yields

$$\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \left(\sum_r M_r \log \tau_r - \tau_r \right) + \sum_r \left\{ \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - M_r \log\left(\sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}}\right)\right\}.$$

[8] Under identifiability constraints $\boldsymbol{\gamma} \leftrightarrow \boldsymbol{\tau}$ is 1-1 function.

So using Bayes' theorem, we obtain:

$$\boxed{\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \ell_{\text{Poiss}}(\boldsymbol{\tau}; \boldsymbol{m}) + \ell_{\text{Mult}}(\boldsymbol{\beta}; \boldsymbol{Y} | \boldsymbol{M} = \boldsymbol{m})}$$

- Hence inferences on $\boldsymbol{\beta}$ using the multinomial model are equivalent to those based on the Poisson model, provided the row parameters γ_r are included.
- A more detailed calculation shows that the MLE $\hat{\boldsymbol{\beta}}$ and its sampling distribution are identical under the two models.

Example (Smoking Data (Doll and Hill), continued.)

Table: Lung cancer deaths in British male physicians. The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Example (Smoking Data (Doll and Hill), continued.)

- Suppose number of deaths y has Poisson distribution, mean $T\lambda(d,t)$, where T is man-years at risk, d is number of cigarettes smoked daily and t is time smoking (years).
- Log-linear model
  - $\lambda_{rc} = \exp(\gamma_r + \beta_c)$
  - deviance 51.47 on 48 df, one parameter for each row and column
- Model taking into account nature of rows/columns:

$$\lambda(d,t) = \{e^{\gamma_0} + \exp(\gamma_1 + \beta_2 \log d)\} \exp(\beta_3 \log t)$$

- deviance is 59.58 on 59 df with just 4 parameters overall

Table: Parameter estimates (standard errors) for lung cancer data.

	γ_0	γ_1	β_2	β_3
Smokers only	0.96 (25.4)	2.15 (1.45)	1.20 (0.40)	4.50 (0.34)
All data	2.94 (0.58)	1.82 (0.66)	1.29 (0.20)	4.46 (0.33)

Perfect Separation – Again

Perfect Separation can also affect multinomial regression:

- If any one class is separated from all the rest by a covariate hyperplane, then by reduction to the binary case we can see that the MLE for that class will fail to exist.
- Detection more subtle: now there are $\binom{p}{2}$ cases to determine.
- Simple heuristic: insist that the iterative method used to maximize the likelihood terminate only after both the value of the likelihood function and the parameter vector stop changing.
- Different inference approaches such as extended logistic regression (Clarkson & Jennrich, 1991), bias-reduced ML (Firth, 1993), and exact logistic regression (Mehta & Patel, 1995) can be used that are stable to separation (as long as we know we have a problem!).
- Can also fit penalised loglinear regression with ridge/lasso penalty.

Nonparametric Regression

A More Flexible Regression Model

So far we have discussed the following setup:

$$Y_i \mid \boldsymbol{x}_i^\top \overset{ind}{\sim} \text{Dist}[\phi_i] \to \begin{cases} \phi_i = g(\boldsymbol{x}_i^\top) = \boldsymbol{x}_i^\top \boldsymbol{\beta}), \\ \boldsymbol{\beta} \in \mathbb{R}^p, \end{cases}$$

with $\boldsymbol{\beta}$ to be estimated from data, e.g.

- Dist $= \mathcal{N}(\mu_i, \sigma^2)$ and $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (Gaussian linear regression)
- Dist $\in \texttt{ExpFamily}(\phi_i)$ and $\phi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta})$ (GLM)

Would now like to extend model to a more flexible dependence:

$$Y_i \mid \boldsymbol{x}_i^\top \overset{ind}{\sim} \text{Dist}[\phi_i] \to \begin{cases} \phi_i = g(\boldsymbol{x}_i^\top), \\ g \in \mathcal{F} \subset L^2(\mathbb{R}^p) \text{ (say)}, \end{cases}$$

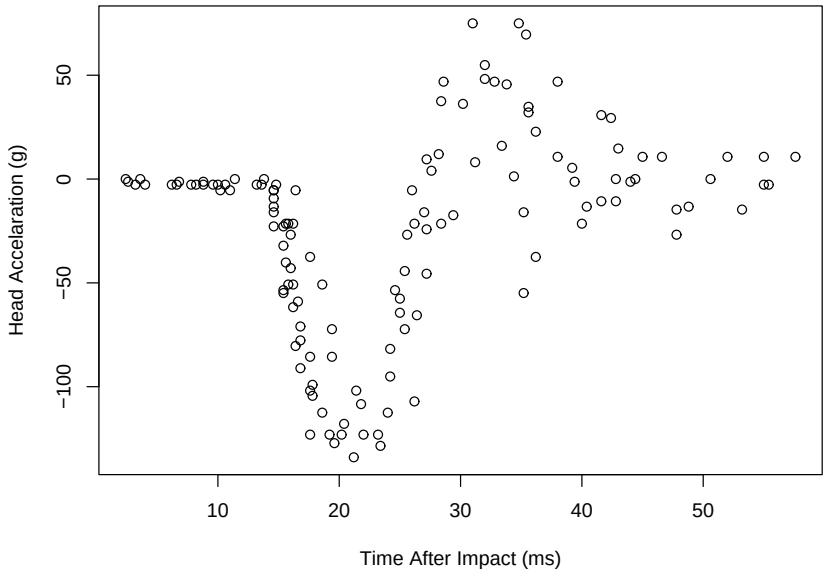
with $g : \mathbb{R}^p \to \mathbb{R}$ unknown, to be estimated given data $\{(Y_i, \boldsymbol{x}_i^\top)\}_{i=1}^n$.

- A *nonparametric* problem (parameter ∞-dimensional)!
- How to estimate g in this context?
- $\mathcal{F}$ is usually assumed to be a class of smooth functions (e.g., C^k).

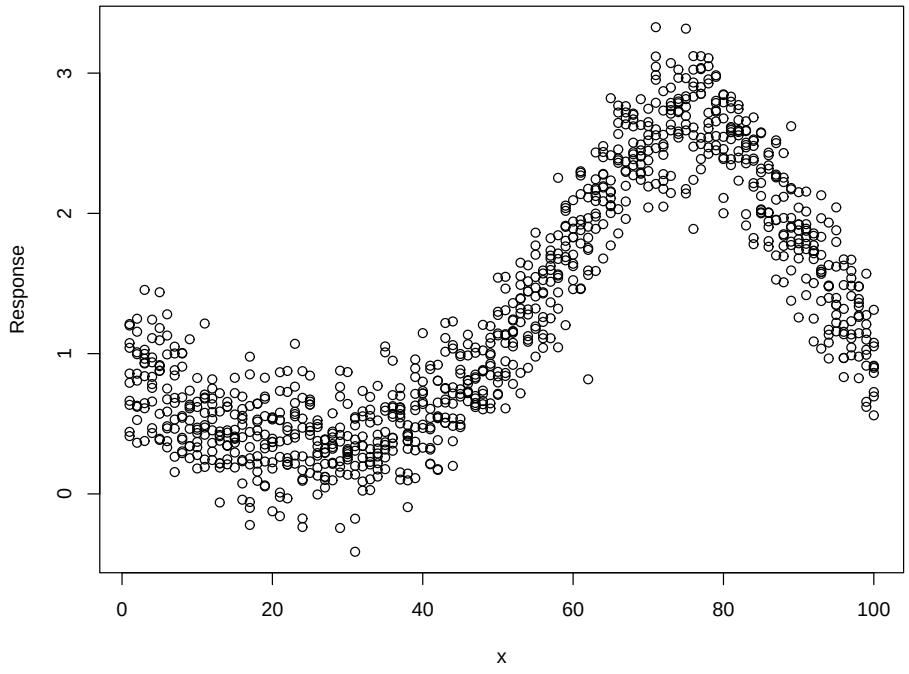
Start from simplest problem:

$$\left.\begin{array}{l}\text{Dist} \equiv \mathcal{N}(\mu_i, \sigma^2) \\ x_i \in \mathbb{R}\end{array}\right\} \Longrightarrow \quad Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Figure: Motorcycle Accident Data

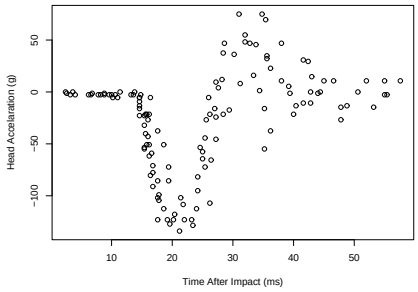


- Ideally: large sample plus multiple x_i with same value (many large covariate classes):



- Then average Y_i's at each covariate class and interpolate . . .
- But this is never the case in practice... . . .

- Usually unique x_i distinct:



- Here is where the smoothness assumption comes in
- Since have distinct value for each x_i, need to borrow information from nearby ...
- ... use continuity!!! (or even better, *smoothness*)

▶ Recall: A function $g : \mathbb{R} \to \mathbb{R}$ is *continuous* if:

$$\forall\ \epsilon > 0\ \exists\ \delta > 0 :\ |x - x_0| < \delta \implies |g(x) - g(x_0)| < \epsilon.$$

▶ So maybe average Y_i's corresponding to x_i's in a δ-neighbourhood of x.
▶ Motivates the use of a kernel smoother ...

Naive idea: $\hat{g}(x_0)$ should be the average of Y_i-values with x_i's "close" to x_0.

$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^n \mathbf{1}\{|x_i - x_0| \leq \lambda\}} \sum_{i=1}^n Y_i \mathbf{1}\{|x_i - x_0| \leq \lambda\}.$$

A weighted average! Choose other weights? Kernel estimator:

$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{\lambda}\right)} \sum_{i=1}^n Y_i K\left(\frac{x_i - x_0}{\lambda}\right) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i Y_i.$$

- K is a kernel function.
  ↪ E.g. standard Gaussian pdf, $K(x) = \varphi(x)$.
- λ is the bandwidth parameter
  ↪ small λ gives local behaviour, large λ gives global behaviour
- Choice of K not so important, choice of λ very important.
- The resulting fitted values are linear in the responses, i.e., $\hat{\boldsymbol{Y}} = \boldsymbol{S}_\lambda \boldsymbol{Y}$, where the smoothing matrix $\boldsymbol{S}_\lambda$ depends on $x_1, \ldots, x_n$, K, and λ. Analogous to a projection matrix in linear regression, but $\boldsymbol{S}_\lambda$ is not a projection.

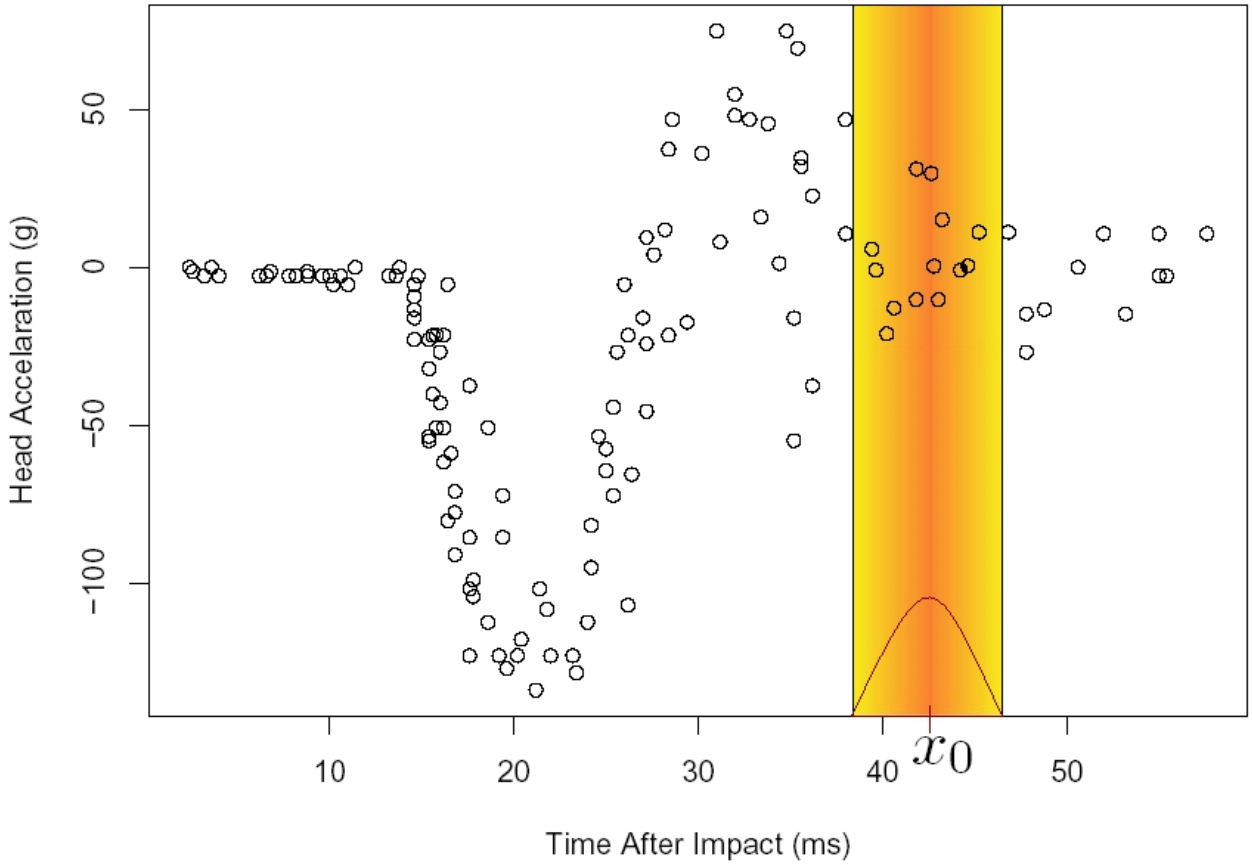


Figure: Visualising a kernel smoother at work


```
> plot(time,accel,xlab="Time After Impact (ms)",ylab="Head Accelaration (g)")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=0.7))
> lines(ksmooth(time,accel,kernel="normal",bandwidth=5),col="red")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=10),col="blue")
```

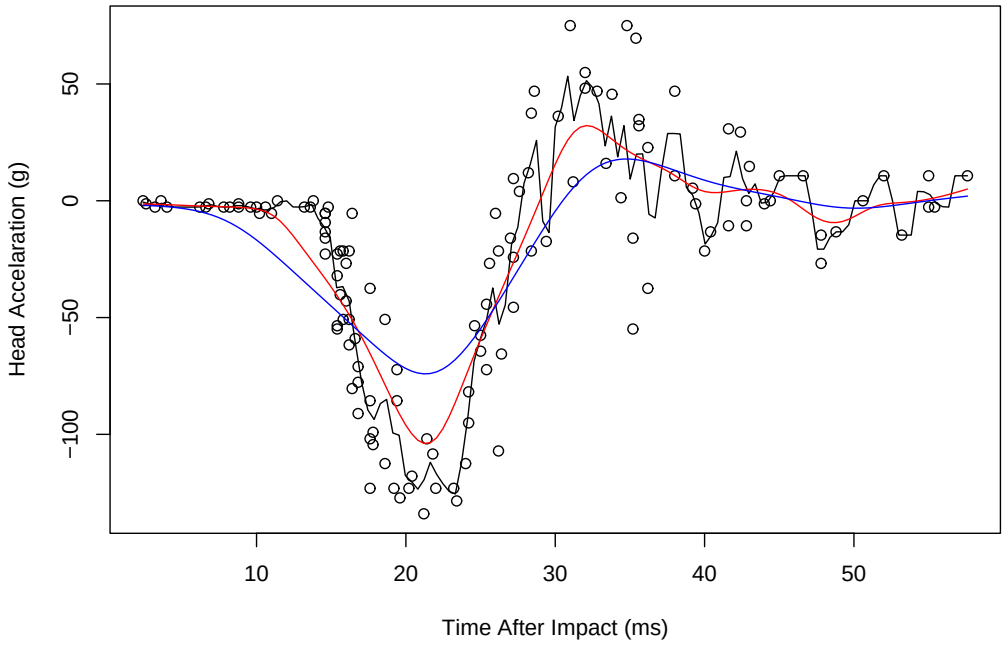


Figure: Motorcycle data kernel smooth for varying bandwidths

Whatever happened to likelihood, though? Find $h \in C^2$ that minimises

$$\underbrace{\sum_{i=1}^{n} \{Y_i - h(x_i)\}^2}_{\text{Fit Penalty}} + \underbrace{\lambda \int_I \{h''(t)\}^2 \, dt}_{\text{Roughness Penalty}}$$

- This is a Gaussian likelihood with a roughness penalty
  ↪ If use only likelihood, any interpolating function is an MLE!
- λ to balance fidelity to the data and smoothness of the estimated h.

Remarkably, problem has unique explicit solution!
↪ Natural Cubic Spline with knots at $\{x_i\}_{i=1}^n$:

- piecewise polynomials of degree 3,
- with pieces defined at the knots,
- with two continuous derivatives at the knots,
- and linear outside the data boundary.

Can represent splines via natural spline basis functions B_j, as

$$s(x) = \sum_{j=1}^{n} \gamma_j B_j(x).$$

Defining matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ as

$$B_{ij} = B_j(x_i), \quad \Omega_{ij} = \int B_i''(x) B_j''(x) \, dx,$$

our penalised likelihood becomes

$$\text{min!} \left\{ (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\gamma})^\top (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\gamma} \right\}.$$

Differentiating and equating with zero yields

$$(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega}) \hat{\boldsymbol{\gamma}} = \boldsymbol{B}^\top \boldsymbol{Y} \implies \hat{\boldsymbol{\gamma}} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top \boldsymbol{Y}.$$

- The *smoothing matrix* is $\boldsymbol{S}_\lambda = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top$.
- The cubic spline fit is approximately a kernel smoother (keyword: equivalent kernel).

```
lines(smooth.spline(time,accel),col="red")
```

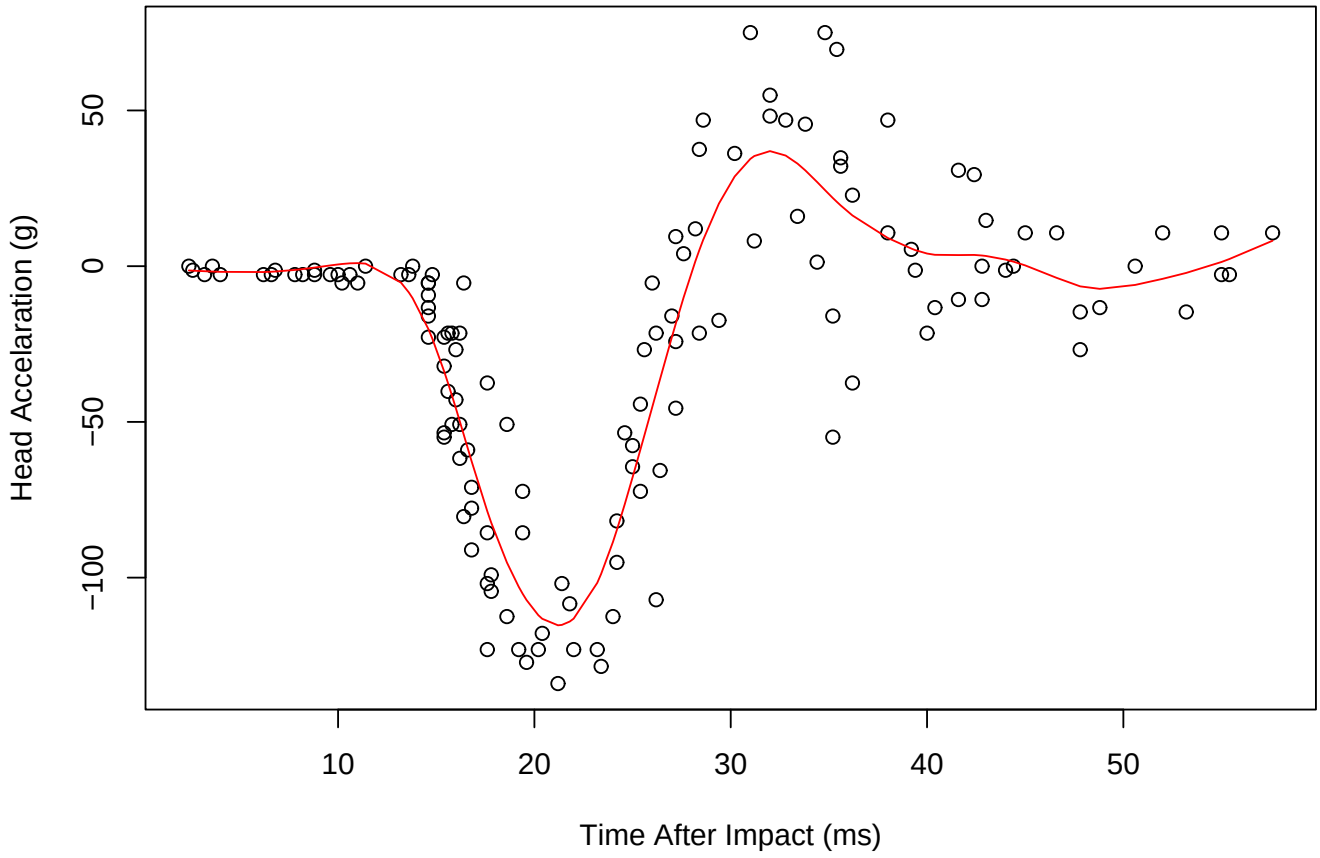


Figure: Motorcycle Example Cubic Spline Fit

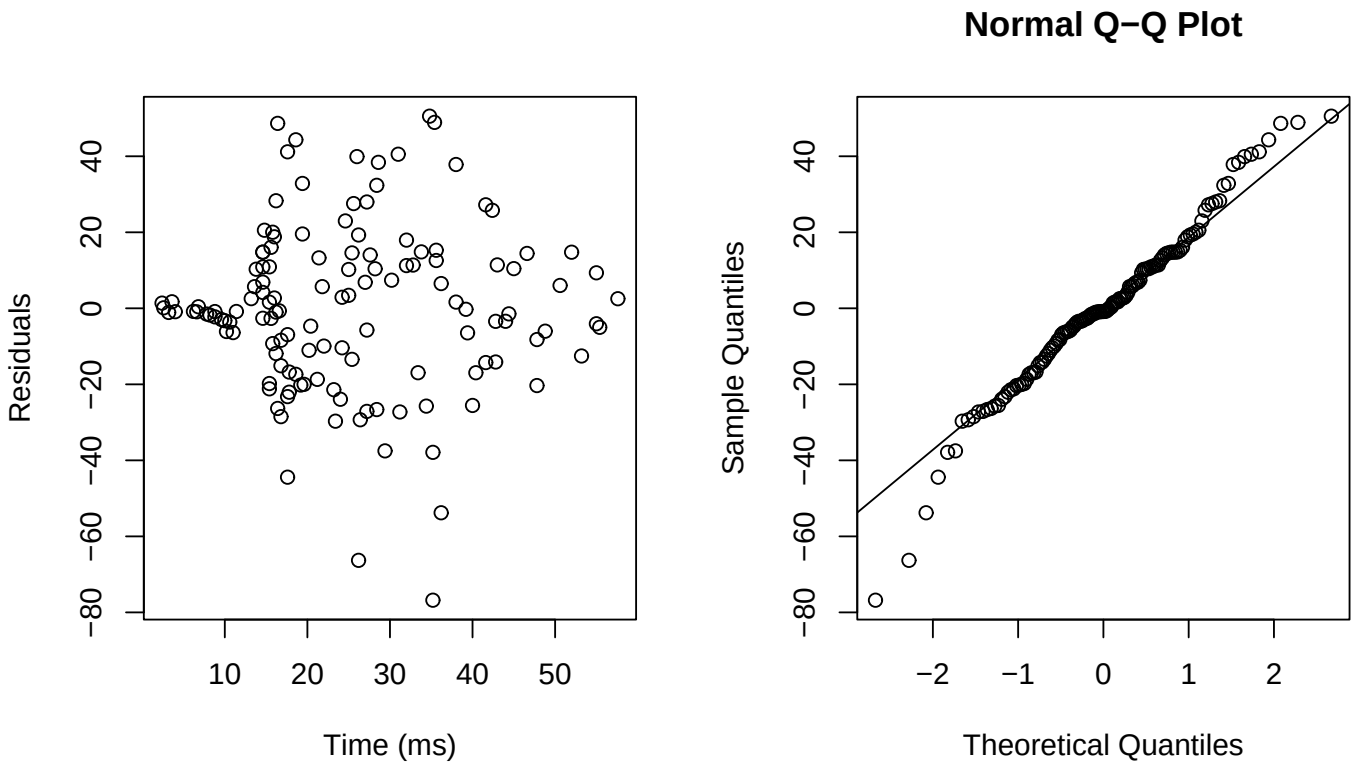


Figure: Motorcycle Example Cubic Spline Residuals

Equivalent degrees of freedom

- Least squares estimation: $\boldsymbol{Y} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we have $\hat{\boldsymbol{Y}} = \boldsymbol{H}\,\boldsymbol{Y}$, with $\mathrm{trace}(\boldsymbol{H}) = p$, in terms of the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.
- In spline smoothing

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B}^\top}_{\boldsymbol{S}_\lambda}\ \boldsymbol{Y}.$$

  suggesting definition of equivalent degrees of freedom of smoother as

$$\mathrm{edf} = \mathrm{trace}(\boldsymbol{S}_\lambda)$$

- $\mathrm{trace}(\boldsymbol{S}_\lambda)$ is monotone decreasing in λ, with $\mathrm{trace}(\boldsymbol{S}_\lambda) \to 2$ as $\lambda \to \infty$ (will always have two nonzero eigenvalues) and $\mathrm{trace}(\boldsymbol{S}_\lambda) \to n$ as $\lambda \to 0$.
- Note 1–1 map $\lambda \leftrightarrow \mathrm{trace}(\boldsymbol{S}_\lambda) =$ df, so usually determine roughness using edf (interpretation easier).
- Each eigenvalue of $\boldsymbol{S}_\lambda$ lies in $(0, 1)$, so this is a smoothing matrix, not a projection matrix.

Bias/Variance Tradeoff and Cross Validation

Focus on the fit for the given grid $x_1, \ldots, x_n$:

$$\hat{\mathbf{g}} = (\hat{g}(x_1), \ldots, \hat{g}(x_n)), \quad \mathbf{g} = (g(x_1), \ldots, g(x_n))$$

Consider the mean squared error:

$$\mathbb{E}(\|\mathbf{g} - \hat{\mathbf{g}}\|^2) = \underbrace{\mathbb{E}\{\|\mathbb{E}(\hat{\mathbf{g}}) - \hat{\mathbf{g}}\|^2\}}_{\text{variance}} + \underbrace{\|\mathbf{g} - \mathbb{E}(\hat{\mathbf{g}})\|^2}_{\text{bias}^2}.$$

In the case of a linear smoother, for which $\hat{\mathbf{g}} = \boldsymbol{S}_\lambda\,\boldsymbol{Y}$, we easily calculate

$$\mathbb{E}(\|\mathbf{g} - \hat{\mathbf{g}}\|^2) = \frac{\mathrm{trace}(\boldsymbol{S}_\lambda\boldsymbol{S}_\lambda^\top)}{n}\sigma^2 + \frac{(\mathbf{g} - \boldsymbol{S}_\lambda\mathbf{g})^\top(\mathbf{g} - \boldsymbol{S}_\lambda\mathbf{g})}{n},$$

so

- $\lambda \uparrow \implies$ variance $\downarrow$ but bias $\uparrow$,
- $\lambda \downarrow \implies$ bias $\downarrow$ but variance $\uparrow$.
- Would like to choose λ to find optimal bias-variance tradeoff:
  - ↪ Unfortunately, optimal λ will depend on unknown g!

- Fitted values are $\hat{\boldsymbol{Y}} = \boldsymbol{S}_\lambda \boldsymbol{Y}$.
- Fitted value $\hat{Y}_j^-$ obtained when (Y_j, x_j) is dropped from fit is

$$S_{jj}(\lambda)(Y_j - \hat{Y}_j^-) = \hat{Y}_j - \hat{Y}_j^-.$$

- Cross-validation sum of squares is

$$\text{CV}(\lambda) = \sum_{j=1}^{n} (Y_j - \hat{Y}_j^-)^2 = \sum_{j=1}^{n} \left\{ \frac{Y_j - \hat{Y}_j}{1 - S_{jj}(\lambda)} \right\}^2,$$

  and generalised cross-validation sum of squares is

$$\text{GCV}(\lambda) = \sum_{j=1}^{n} \left\{ \frac{Y_j - \hat{Y}_j}{1 - \text{trace}(S_\lambda)/n} \right\}^2,$$

  where $S_{jj}(\lambda)$ is (j, j) element of $\boldsymbol{S}_\lambda$.

Orthogonal Series: "Parametrising" The Problem

If $\mathcal{F} \ni g(\cdot)$ is a separable Hilbert space,we can write:

$$g(x) = \sum_{k \in \mathbb{Z}} \beta_k \psi_k(x) \quad \text{(in an appropriate sense)},$$

with $\{\psi\}_{k=1}^{\infty}$ known (orthogonal) basis functions for $\mathcal{F}$, e.g.,

- $\mathcal{F} = L^2(-\pi, \pi)$,
- $\{\psi_k\} = \{e^{-ikx}\}_{k \in \mathbb{Z}}$, $\psi_i \perp \psi_j$, $i \neq j$.
- Gives Fourier series expansion, $\beta_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) e^{-ikx} \, dx$.

If we truncate series, then we reduce to linear regression:

$$Y_i = \sum_{|k| < \tau} \beta_k \psi_k(x_i) + \varepsilon_i, \quad \tau < \infty$$

Notice: truncation has implications, e.g., in Fourier case:

- Truncating implies assume $g \in \text{span}\{\psi_{-\tau}, ..., \psi_\tau\} \subset L^2$.
- Interpret this as a smoothness assumption on g.
- How to choose τ optimally?

Convolution: Series Truncation $\overset{?}{\simeq}$ Smoothing

Classical exercise in Fourier analysis shows that

$$\sum_{k=-\tau}^{\tau} \beta_k \, e^{-ikx} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(y) D_\tau(x-y) \, dy$$

with the Dirichlet kernel of order τ, $D_\tau(u) = \sin\{(\tau + 1/2)\, u\}/\sin(u/2)$.

Recall kernel smoother:

$$\hat{g}(x_0) = \sum_{i=1}^{n} \frac{Y_i K_\lambda(x_i - x_0)}{\sum_{i=1}^{n} K_\lambda(x_i - x_0)} = \frac{1}{c} \int_I y(x) K_\lambda(x - x_0) \, dx,$$

with

$$y(x) = \sum_{i=1}^{n} Y_i \delta(x - x_i).$$

- So if K is the Dirichlet kernel, we can do series approximation via kernel smoothing.
- Works for other series expansions with other kernels (e.g., Fourier with convergence factors)

The Dirichlet kernel

So far: how to estimate $g : \mathbb{R} \to \mathbb{R}$ (assumed smooth) in

$$Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \text{ given data } \quad \{(Y_i, x_i)\}_{i=1}^n.$$

Can extend to GLM setting as:

$$Y_i | x_i \overset{indep}{\sim} \exp\{g(x_i)y - \gamma(g(x_i)) + S(y)\}$$

- Parametrise candidate g via spline

$$s(x) = \sum_{j=1}^{n} \gamma_j B_j(x).$$

- Define matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ as before,

$$B_{ij} = B_j(x_i), \quad \Omega_{ij} = \int B_i''(x) B_j''(x)\, dx$$

- And consider penalised likelihood, similarly as with penalised GLM

$$\ell_n(\gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma = \gamma^\top \boldsymbol{B}^\top \boldsymbol{Y} - \sum_{i=1}^{n} \gamma(\boldsymbol{b}_i^\top \gamma) + \lambda \gamma^\top \boldsymbol{\Omega} \gamma.$$

How can we generalise to multivariate covariates?

- "Immediate" Generalisation: $g : \mathbb{R}^p \to \mathbb{R}$ (smooth)

$$Y_j = g(x_{j1}, \ldots, x_{jp}) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Estimation by (e.g.) multivariate kernel method.

- Two basic drawbacks of this approach ...

↪ Shape of kernel? (definition of *local*)

↪ *Curse of dimensionality*

What is "local" in $\mathbb{R}^p$, though?

$\rightarrow$ Need some definition of "local" in the space of covariates

$\hookrightarrow$ Use some metric on $\mathbb{R}^p \ni (x_1, \ldots, x_p)^\top$!

But which one?

- Choice of metric $\iff$ choice of geometry
  - $\hookrightarrow$ e.g., curvature reflects intertwining of dimensions
- Geometry $\implies$ reflects structure in the covariates
  - potentially different units of measurement (variable stretching of space)
  - g may be of higher variation in some dimensions (need finer neighbourhoods there)
  - statistical dependencies present in the covariates ("local" should reflect these)

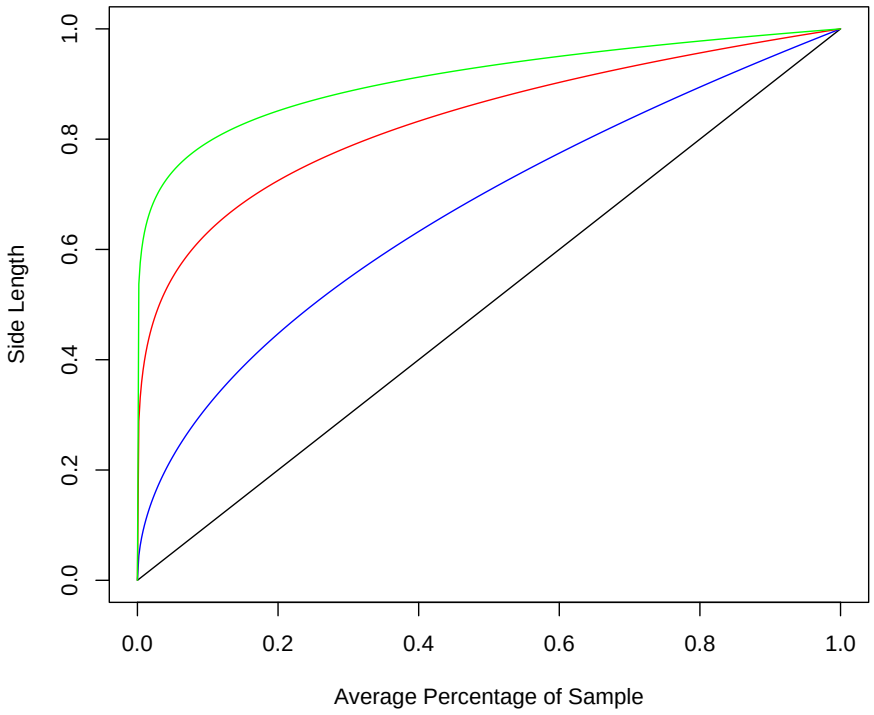


Figure: Curse of Dimensionality ($\mathrm{Unif}[0,1]^p$): $p = 1$, $p = 2$, $p = 5$, $p = 10$

Curse of Dimensionality

> *"neighbourhoods with a fixed number of points become less local as the dimensions increase"*
>
> *Bellman (1961)*

- Notion of local in terms of % of data: fails in high dimensions
  $\hookrightarrow$ There is too much space!
- Hence to allow for reasonably small bandwidths
  $\hookrightarrow$ Density of sampling must increase.
- Need to have ever larger samples as dimension grows.

Tackling the Dimensionality Issue

Attempt to find a link/compromise between:

- our mastery of 1D case (at least we can do that well . . .),
- and higher dimensional covariates (and associated difficulties).

Perhaps something that can be fitted/interpreted variable-by-variable?

- Compromise: Additive Model

$$Y_i = \alpha_i + \sum_{k=1}^{p} f_k(x_{ik}) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

with f_k's univariate smooth functions, $\sum_i f_k(x_{ik}) = 0$.

- Can extend to Generalised Additive Model:

$$Y_i | \boldsymbol{x}_i^\top \overset{indep}{\sim} \exp\left\{ \alpha_i y + y \sum_{k=1}^{p} f_k(x_{ik}) - \gamma\left(\alpha_i + \sum_{k=1}^{p} f_k(x_{ik}) \right) + S(y) \right\}$$

- How to fit additive model? Consider Gaussian case only for simplicity.
  - ↪ Know how to fit each f_k separately quite well
  - ↪ Take advantage of this ...

- Consider ith response:

$$\mathbb{E}\left[Y_i - \alpha - \sum_{m \neq k} f_m(x_{im})\right] = f_k(x_{ik})$$

- Suggests the *Backfitting Algorithm*:

(1) `Initialise:` $\alpha = \bar{Y}$, $f_k = f_k^0$, $k = 1, \ldots, p$.

(2) `Cycle:` Get f_k by 1D smoothing of partial residual scatterplot

$$\left\{\left(Y_i - \alpha - \sum_{m \neq k} f_m(x_{im}), x_{ik}\right)\right\}_{i=1}^{n} = \{e_{ik}, x_{ik}\}_{i=1}^{n}.$$

(3) `Stop:` when individual functions don't change

- Any smoother can be used, usually splines.

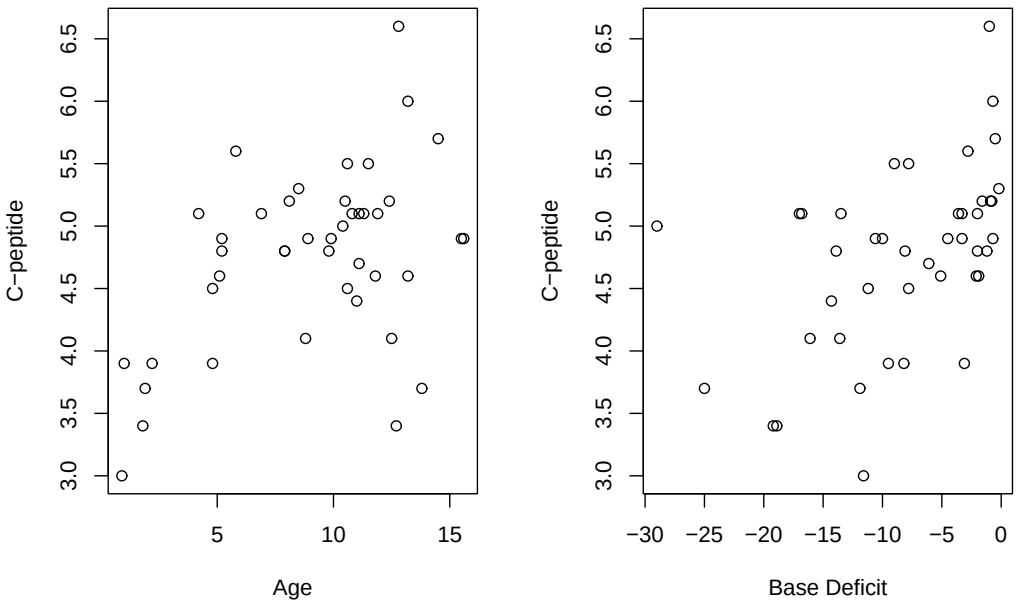


Figure: Example: Diabetes Data

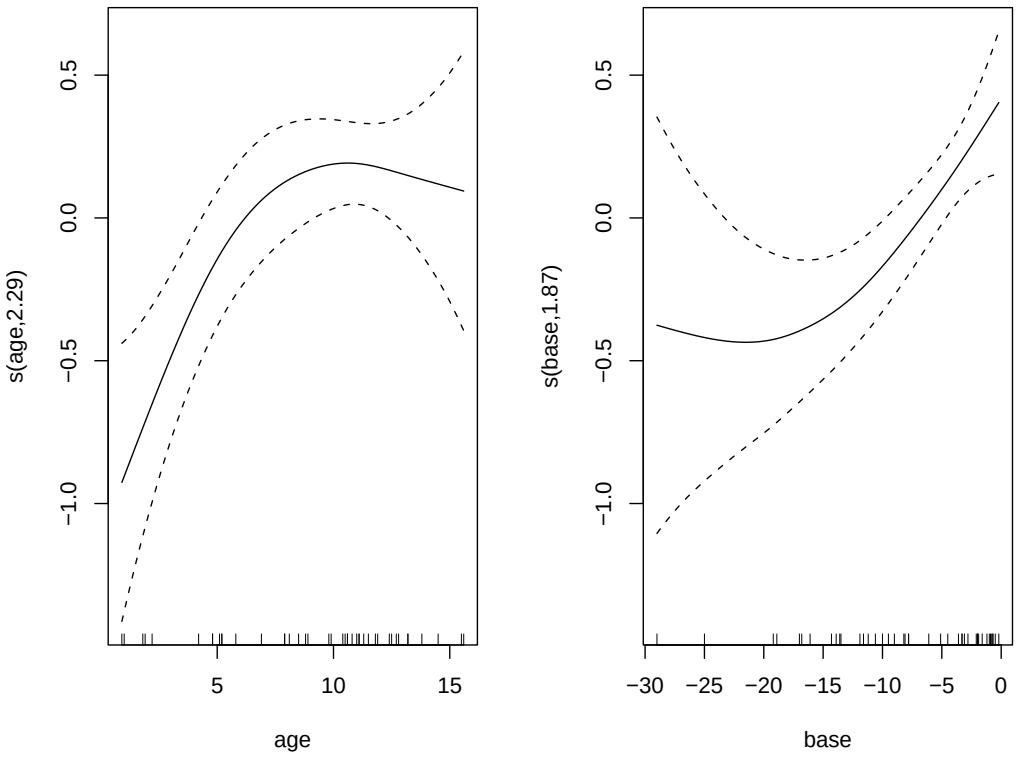


Figure: Example: Diabetes Data Additive Fit

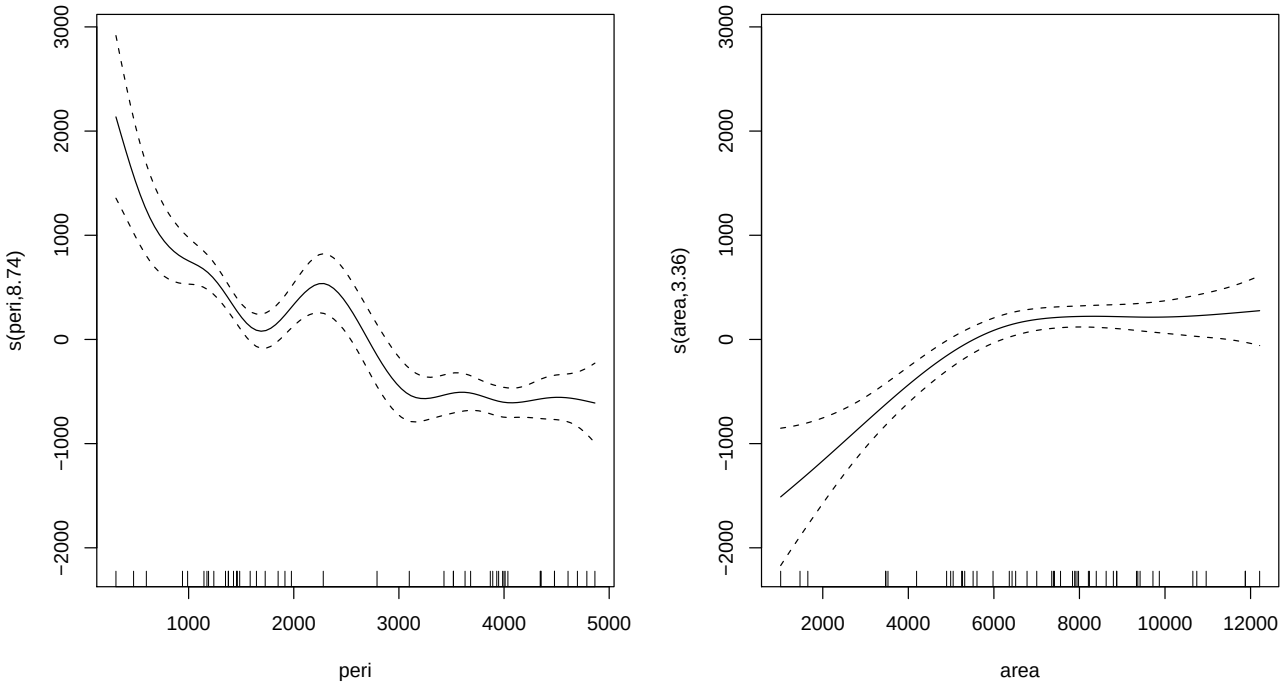


Figure: Example: Rock Permeability Data (measurements on 48 rock samples from a petroleum reservoir)

```
Family: gaussian
Link function: identity

Formula:
perm ~ 1 + s(peri) + s(area)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.45      27.18   15.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
          edf Est.rank      F  p-value
s(peri) 8.739        9 18.286 9.49e-11 ***
s(area) 3.357        7  6.364 7.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.815   Deviance explained = 86.3%
```


Projection Pursuit Regression

A different approach is inspired by tomography. Model Gaussian response as:

$$Y_i = \underbrace{\sum_{k=1}^{K} h_k(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}_{=g(\boldsymbol{x}_i^\top)} + \varepsilon_i, \quad \|\boldsymbol{\beta}_k\| = 1, \ \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

- Also additively decomposes g into smooth functions $h_k : \mathbb{R} \to \mathbb{R}$.
- But each function now depends on a global linear feature $\boldsymbol{x}_i^\top \boldsymbol{\beta}_k$
  $\hookrightarrow$ a linear combination of the covariates
  $\hookrightarrow$ $\|\boldsymbol{\beta}_k\| = 1$ for identifiability.
- Projections directions to be chosen for best fit (nonlinear problem)
- Each h_k is a ridge function of $\boldsymbol{x}_i^\top$: varies only in the direction defined by $\boldsymbol{\beta}_k$

Pros and Cons:

(+) By classical Fourier series, can show that any $C^1([0,1]^p) \to \mathbb{R}$ function is uniformly approximated arbitrarily well as $K \to \infty$. Useful for prediction.

(−) Interpretability? What do terms mean within problem?

How is the model fitted to data?

Assume only one term, $K = 1$ and consider penalized likelihood:

$$\min_{h_1 \in C^2[0,1], \|\boldsymbol{\beta}\|=1} \left\{ \sum_{i=1}^{n} \{ Y_i - h_1(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \}^2 + \int_0^1 \{h_1''(t)\}^2 dt \right\}.$$

Two steps:

- *Smooth*: Given a direction $\boldsymbol{\beta}$, fitting $h_1(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is done via 1D smoothing.
- *Pursue*: Given h_1, have a non-linear regression problem w.r.t. $\boldsymbol{\beta}$.

Hence, iterate between the two steps

↪ Complication is that h_1 not explicitly known, so need numerical derivatives.

↪ Computationally intensive (impractical in the '80's but doable today).

↪ Can separate second step by looking for non-Gaussian projection directions.

Further terms added in forward stepwise manner.

Embracing Nonlinearity (at the expense of interpretability)

If $\boldsymbol{\beta}_k$ needs to be estimated non-linearly anyway...

$$g(\boldsymbol{x}_i^\top) \approx \sum_{k=1}^{K} h_k(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)$$

... do we really need to estimate the h_k or can we fix them?

Theorem (Nonlinear Sigmoidal Approximation)

Let $\Psi : \mathbb{R} \to [0, 1]$ be a strictly increasing distribution function and $g : [0, 1]^p \to \mathbb{R}$ be an arbitrary continuous function. Then, for any $\epsilon > 0$, there exists $K < \infty$ and vectors $\boldsymbol{\alpha}, \boldsymbol{t} \in \mathbb{R}^K$ and $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\} \subset \mathbb{R}^p$ such that

$$\sup_{\boldsymbol{x} \in [0,1]^d} \left| g(\boldsymbol{x}) - \sum_{k=1}^{K} \alpha_k \Psi(t_k + \boldsymbol{x}^\top \boldsymbol{\beta}_k) \right| < \epsilon.$$

- Can take h_k to be translations of the same known function Ψ!
- The tradeoff is that K may need to be quite large (interpretability?)
- Called a (single layer) neural network by analogy to synaptic function.
- A parametric model with many parameters – fit by nonlinear least squares (gradient descent)

What about including transformations of the original covariates?

1. Can of course include J transformations $w_j : \mathbb{R}^p \to \mathbb{R}$

$$(u_1, ..., u_p) \mapsto w_j(u_1, ..., u_n), \qquad j = 1, ..., J,$$

   of the original variables as additional covariates by suitably enlarging the design matrix $\boldsymbol{X}$.
2. We simply adjoin to $\boldsymbol{X}$ another J columns of dimension $n \times 1$ each:

$$\begin{pmatrix} w_j(\boldsymbol{x}_1^\top) \\ \vdots \\ w_j(\boldsymbol{x}_n^\top) \end{pmatrix} \qquad j = 1, ..., J.$$

3. Which functions w_j should we pick though?

Since we've gone nonlinear anyway,

why not attempt to learn which transformations to include from the data?

How?

- Instead of including our original covariates (p columns of $\boldsymbol{X}$)...
- ... use q derived covariates (q can be larger than p)

$$\begin{pmatrix} w_1(\boldsymbol{x}_1^\top) \\ \vdots \\ w_1(\boldsymbol{x}_n^\top) \end{pmatrix}, \begin{pmatrix} w_2(\boldsymbol{x}_1^\top) \\ \vdots \\ w_2(\boldsymbol{x}_n^\top) \end{pmatrix}, \quad \cdots \quad , \begin{pmatrix} w_q(\boldsymbol{x}_1^\top) \\ \vdots \\ w_q(\boldsymbol{x}_n^\top) \end{pmatrix}$$

- ... where the q transformations $\{w_j\}_{j=1}^q$ are to be estimated from the data.

Recycling our nonlinear approximation theorem, write

$$w_j(\boldsymbol{x}^\top) \approx \sum_{m=1}^{M_j} \delta_{m,j} \Psi(s_{m,j} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{m,j})$$

using the same Ψ, and needing to estimate $(\boldsymbol{\delta}_j, \boldsymbol{s}_j, \boldsymbol{\gamma}_{1,j}, ..., \boldsymbol{\gamma}_{M_j,j})$, for $j = 1, ..., q$.

Assuming that we've constructed our new variables, we have a new design matrix

$$\begin{pmatrix} w_1(\boldsymbol{x}_1^\top) & \dots & w_q(\boldsymbol{x}_1^\top) \\ \vdots & & \vdots \\ w_1(\boldsymbol{x}_n^\top) & \dots & w_q(\boldsymbol{x}_n^\top) \end{pmatrix}.$$

Summarising, we have defined a hierarchical nonlinear regression model:

$$Y_i = \sum_{k=1}^{K} \alpha_k \Psi\Big(t_k + (w_i(\boldsymbol{x}_1^\top), ..., w_i(\boldsymbol{x}_n^\top))\boldsymbol{\beta}_k \Big) + \varepsilon_i =$$

$$= \sum_{k=1}^{K} \alpha_k \Psi \left(t_k + \left(\sum_{m=1}^{M_1} \delta_{m,1} \Psi(s_{m,1} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{m,1}), ..., \sum_{l=1}^{M_q} \delta_{l,q} \Psi(s_{l,q} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{l,q}) \right) \boldsymbol{\beta}_k \right) + \varepsilon_i$$

... known these days as a two-layer neural network.

- Can add more layers ("deep neural network").
- Highly non-linear and non-convex – cascade of simple nonlinearities applied to linear transformations.
- More easily perceived visually through a graphical representation

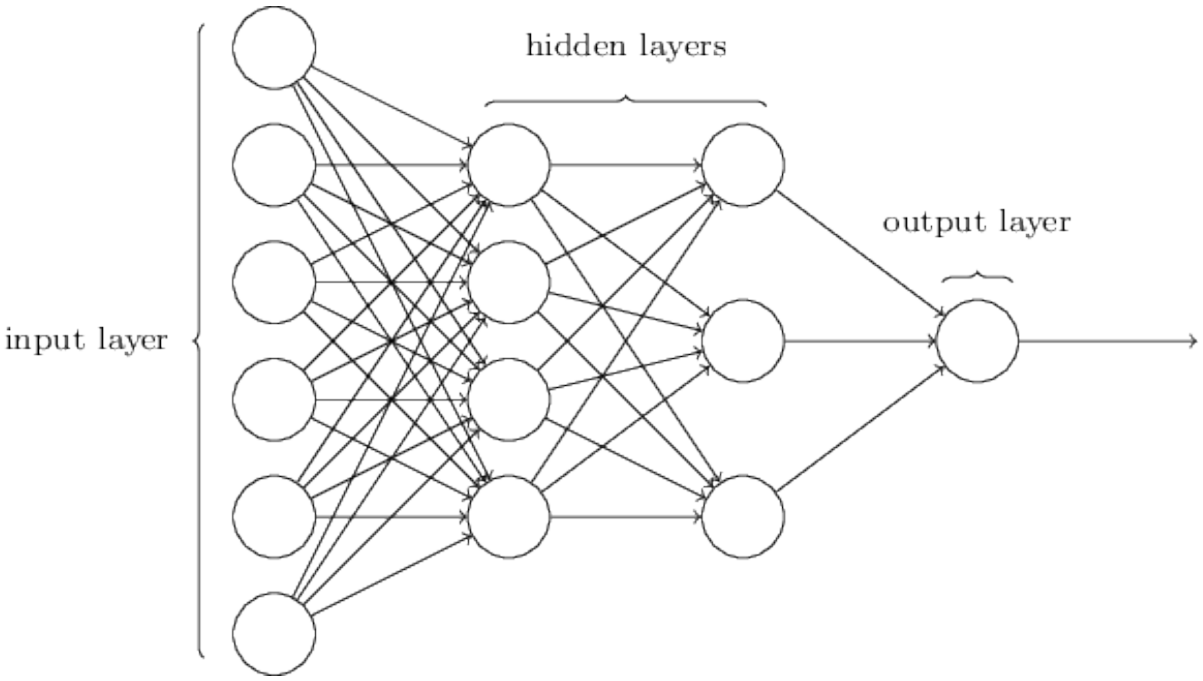


Figure: A multilayer neural network. The "input layer" corresponds to the original covariates. Each "hidden layer" corresponds to successively derived covariates via sigmoid superposition. The "output layer" is the final sigmoid superposition yielding the response.

Seems like we're asking a lot from the data...

...and indeed we are

- NN fitted by optimising least square fit of Y_i to model w.r.t. parameters.
- Typically carried out with some flavour of gradient descent.
- Representation can be highly non-unique and many local optima can exist.
- Large number of parameters requires very large number of observations.
- Sampling distribution essentially totally unknown, practically no theory available.
- Seldom useful for interpretation – typically used for prediction/classification.
- Often requires context-dependent tuning and optimisation architecture.