Lecture 24: Revision Notes

Sofia Olhede

EPFL

December 14, 2020

Worked examples tomorrow

- I have been asked to cover:
- For instance over testing hypothesis on the last exercise sheet;
- Non–parametric regression again;
- more concrete examples on GLMs;
- how to handle the separable data case again, along with jitter residuals;
- Non parametric regression;
- Estimate the unknown function $h(x)$ with the modulators and the wavelets.

Modelling

- This is an abbreviation; the full course is in the regular lecture notes.
- Started by making the distinction between the explanatory and predictive framework.
- Discussion about why data is stochastic.
- Start by specifying a distribution $F(y_1, \ldots, y_n; \theta)$; where $y \in \mathcal{Y}^n$ and $\theta \in \Theta$.
- Assume we observe $Y_1, \ldots, Y_n \in \mathcal{Y}^n$.
- When $F(y_1, \ldots, y_n; \theta)$ is known the problem is parametric; if not then it is non–parametric.
- Example: coin flipping with an unknown success probability θ.
- How do we handle the modelling? We need to understand probability.

Probability

- We model outcomes of experiments. A possible outcome ω is an elementary event.
- The set of total outcomes is written as Ω.
- We always assume $\Omega \neq \emptyset$. Note that $\emptyset$ is a set.
- An event is a subset of Ω.
- The union of two events F_1 and F_2 is $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs.
- The intersection of two events F_1 and F_2 written as $F_1 \cap F_2$ occurs if and only if both of F_1 or F_2 occurs.
- We can define unions of unions and intersections iteratively.

Probability

- The complement of an event F written as F^c contains all the elements in Ω that are not in F.
- Two events F_1 and F_2 are disjoint if they have no elements in common.
- A partition $\{F_n\}$ is a collection of events such that $F_i \cap F_j = \emptyset$ and $\cup_n F_n = \Omega$.
- We can combine these binary operations using De Morgan's laws.
- We go from sets to probability measure. To define this we define the three axioms of probability:

(i) $\Pr\{F\} \geq 0$ for all $F \subset \Omega$.

(ii) $\Pr\{\Omega\} = 1$.

(iii) If an event G is a countable union $G = \cup_n F_n$ for disjoint events F_n then

$$\Pr\{G\} = \sum_n \Pr\{F_n\}.$$

Random Variables

- Conditional probability is the next set of results.
- For any pair of events F_1 and F_2 such that $\Pr\{F_2\} > 0$ then we define the conditional probability of F_1 given F_2:

$$\Pr\{F_1 \mid F_2\} = \frac{\Pr\{F_1 \cap F_2\}}{\Pr\{F_2\}}.$$

- A random variable (RV) X is a real function $X : \Omega \mapsto \mathbb{R}$.
- We for $A \subset \mathbb{R}$ write $\{X \in A\}$ for the event

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

- The distribution function (or cumulative distribution function) $F_X(x)$ is defined as

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

Random Variables

- A continuous random variable X has probability density function $f_X(x)$ for $x \in \mathcal{X}$ such that

$$F_X(b) - F_X(a) = \int_a^b f_X(u)\, du.$$

- $f_x(x)$ on its own is not a probability and so not bounded above.
- For a discrete random variable X we may define its probability mass function (PMF) to be

$$f_X(x) = \Pr\{X = x\}, \quad x \in \mathcal{X}.$$

- Once we understand how to model X we have a model of $Y = g(X)$.
- In real life we never just look at single RVs: we need random vectors.

Random Vectors

- Random vectors: A random vector X for a fixed positive integer d is $X = \begin{pmatrix} X_1 & \dots & X_d \end{pmatrix}^T$ is a finite collection of random variables.
- The joint distribution of the random vector $X = \begin{pmatrix} X_1 & \dots & X_d \end{pmatrix}^T$ is

$$F_X(x_1, x_2, \dots, x_d) = \Pr\{X_1 \leq x_1,\, X_2 \leq x_2, \dots, X_d \leq x_d\}.$$

- One can marginalize distributions by integrating out or summing out variables.
- Everything continuous is multivariate calculus, see e.g. Schaum's Outline of Advanced Calculus, Third Edition.

Random Vectors

- The random variables $X_1, \ldots, X_d$ are called independent if and only if for all $x_1, \ldots, x_d$
$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1)F_{X_2}(x_2)\ldots F_{X_d}(x_d).$$
- Equivalently the random variables $X_1, \ldots, X_d$ are independent if and only if for all $x_1, \ldots, x_d$
$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_d}(x_d).$$
- For two random variables, X and Y, we denote their independence as $X \perp\!\!\!\perp Y$.
- The random vector X in $\mathbb{R}^d$ is called conditionally independent of the random vector Y given the random vector Z written as
$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y|Z,$$
if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$
$$F_{X_1,\ldots,X_d|Z,Y}(x_1, \ldots, x_d) = F_{X_1,\ldots,X_d|Z}(x_1, \ldots, x_d). \tag{1}$$

Random Vectors

- (Sums of random variables). Let X and Y be independent continuous random variables with densities $f_X(x)$ and $f_Y(y)$ respectively. The density of $X+Y$ is the convolution of $f_X(x)$ with $f_Y(y)$. Thus

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

- Define $g: \mathbb{R}^2 \to \mathbb{R}^2$ $(x, y) \overset{g}{\mapsto} (x+y, y)$ with inverse transformation $(u, v) \overset{g^{-1}}{\mapsto} (u-v, v)$. The Jacobian of the inverse is $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, with determinant 1.
- It follows that

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

- Marginalize and you are done.

Expectation

- For a continuous random variable this is defined as

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

- For any $g : \mathbb{R}^d \to \mathbb{R}$ we define

$$\mathbb{E}\{g(X_1, \ldots, X_d)\} = \int_{-\infty}^{\infty} g(x_1, \ldots, x_d) f_X(x) dx_1, \ldots dx_d.$$

- The mean vector of random vector $\boldsymbol{X} = \begin{pmatrix} X_1 & \ldots & X_d \end{pmatrix}^T$ is defined as

$$\mathbb{E}(\mathbb{X}) = \begin{pmatrix} \mathbb{E}(X_1) & \ldots & \mathbb{E}(X_d) \end{pmatrix}^T.$$

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

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\left\{(X - \mathbb{E}(X))^2\right\},$$

if $\mathbb{E}(X^2)$ is finite.

Expectation

- Furthermore the covariance of a random variable X_1 with another random variable X_2 expresses the linear dependence between the two:

$$\mathbb{C}\text{ov}(X_1, X_2) = \mathbb{E}\{(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\}.$$

- The correlation between X_1 and X_2 is defined as

$$\text{corr}(X_1, X_2) = \frac{\mathbb{C}\text{ov}(X_1, X_2)}{\sqrt{\mathbb{V}\text{ar}(X_1)\,\mathbb{V}\text{ar}(X_2)}}.$$

- The correlation conveys equivalent dependence information to the covariance. Advantages: (1) invariant to scale changes, (2) can be understood in absolute terms(ranges in $[-1, 1]$). This is a consequence of the correlation inequality, follows from Cauchy-Schwarz inequality.

Conditional Expectation

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

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum_{x \in \mathcal{X}} x \Pr\{X = x|Y = y\} & \text{if} \quad X \text{ and } Y \text{ discrete} \\ \int_{\mathcal{X}} x f_{X|Y=y}(x|y)\, dx & \text{if} \quad X \text{ and } Y \text{ continuous} \end{cases}$$

- The conditional variance of X given Y is defined as

$$\mathbb{V}\text{ar}\{X|Y\} = \mathbb{E}_Y\left\{\left(X - \mathbb{E}_{X|Y}(X|Y)\right)^2 |Y\right\} = \mathbb{E}(X^2|Y) - \mathbb{E}^2(X|Y).$$

The law of total variance states that

$$\mathbb{V}\text{ar}(X) = \mathbb{E}_Y(\mathbb{V}\text{ar}(X|Y)) + \mathbb{V}\text{ar}_Y(\mathbb{E}(X|Y)).$$

Moment Generating Functions

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

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

  and

$$M_X(t) = \mathbb{E}\Big(e^{tX}\Big).$$

- $M_{X+Y}(t) = M_X(t)M_Y(t)$ when X and Y are independent.

Moment Generating Functions XIII

- Lemma: Let $X \sim N(\mu, \sigma^2)$ and assume $a \neq 0$. Then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
- Corollary: let $X_1, \ldots X_n$ be independent random variables and let $X_i \sim N(\mu_i, \sigma_i^2)$. Take S_n as the sum of the X_i. Then

$$S_n \sim N(\sum_i \mu_i, \sum_i \sigma_i^2).$$

Entropy etc

- The entropy is used to measure the disorder of a random variable.
- The entropy of a random variable X is defined as

$$
\begin{aligned}
H(X) &= -\mathbb{E}\{\log f_X(X)\} \\
&= \begin{cases} -\sum_{x\in\mathcal{X}} f_X(x)\log\{f_X(x)\} & \text{if} \quad X \quad \text{discrete} \\ -\int_{x\in\mathcal{X}} f_X(x)\log\{f_X(x)\}dx & \text{if} \quad X \quad \text{continuous} \end{cases}
\end{aligned}
$$

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

$$
\mathrm{KL}(q||p) \equiv \int_{\mathbb{R}} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx. \tag{2}
$$

Exponential Family

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

$$f(y) = \exp\left\{ \sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi_1, \ldots, \phi_k) + S(y) \right\} \tag{3}$$

where

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

Statistics

- We use sampling theory to understand how functions $T = T(Y_1, \ldots, Y_n)$ carry information about the parameter θ.
- We determine the probability distribution of T and determine how that relates to the distribution of the sample.
- Definition (Statistic). A statistic is any function T of the data whose domain is the sample space $\mathcal{Y}^n$ but which does not depend on any unknown parameters.
- Intuitively any function that can be evaluated from the sample is a statistic.
- Any statistic is a random variable with its own distribution.

Statistics

- Definition (Sampling Distribution) Let $(Y_1 \ldots, Y_n)^T \sim F(y_1, \ldots, y_n; \theta)$ and let T be a q-dimensional statistic

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

The sampling distribution of T under $F(y_1, \ldots, Y_n; \theta)$ is the distribution:

joint distribution from t_1 to t_n

$$F_T(t_1, \ldots t_q) = \Pr(T_1(Y_1 \ldots, Y_n) \leq t_1, \ldots, T_q(Y_1 \ldots, Y_n) \leq t_q).$$

- Definition. Ancillary statistics. A statistic T is ancillary for θ if its distribution does not functionally depend on θ.
- Sufficient Statistic: A Statistic $T = T(\mathrm{Y})$ is said to be sufficient for the parameter θ if the conditional probability distribution of the sample given the statistic

$$F_{\mathrm{Y}|T=y}(y_1, \ldots, y_n) = \Pr\{Y_1 \leq y_1, \ldots, Y_n \leq y_n \,|\, T = t\},$$

does not depend on θ.

Exponential Family XIII

- Thus T is sufficient for θ.
- In general the definition of sufficiency is hard to verify.
- **Theorem (Fisher–Neyman factorization theorem)**: suppose that Y has a joint density or frequency function $f(\mathrm{y};\theta)$, where $\theta \in \boldsymbol{\Theta}$. A Statistic $T = T(\mathrm{Y})$ is sufficient for θ if and only if

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

- 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 $\mathrm{Y} = (Y_1, \ldots, Y_n)$ have joint density or frequency function $f(\mathrm{y};\theta)$ and let $T = T(\mathrm{Y})$ be a statistic. Suppose that $f(\mathrm{y};\theta)/f(\mathrm{z};\theta)$ is independent of θ if and only if $T(\mathrm{y}) = T(\mathrm{z})$. Then T is minimally sufficient for θ.

Sampling Distributions

- By studying sampling distributions we aim to determine what different information do different forms of T carry about θ.
- Theorem: (Sampling Distributions of Gaussian Sufficient Statistics).
  Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and define

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

- The pair $(\bar{Y}, S^2)$ are minimally sufficient for (μ, σ^2) and
  - (a) The sample mean has distribution $\bar{Y} \sim N(\mu, \sigma^2/n)$,
  - (b) The random variables $\bar{Y}$ and S^2 are independent,
  - (c) The random variable S^2 satisfies $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.
- Corollary: (Moments of Sufficient Statistics).
  If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then

$$\mathbb{E}(\bar{Y}) = \mu, \quad \mathbb{V}\text{ar}\{\bar{Y}\} = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2, \quad \mathbb{V}\text{ar}\{S^2\} = \frac{2\sigma^4}{n-1}.$$

Sampling Distributions

- Theorem (Sum of Gaussian Squares) Let $(Z_1, \ldots, Z_k)$ be iid $N(0,1)$ random variables. Then

$$Z_1^2 + \cdots + Z_k^2 \sim \chi_k^2.$$

- Theorem: let $Y_1 \sim \chi^2_{d_1}$ and let $Y_2 \sim \chi^2_{d_2}$ be independent. Then

$$\frac{Y_1/d_1}{Y_2/d_2} \sim F_{d_1,d_2}.$$

Modes of Convergence

convergence allows us to study the distribution as the sample size becomes bigger

=Convergence in Law

- Definition: Convergence in Distribution (Weak Convergence). Let $\{F_n\}_{n\geq 1}$ be a sequence of distribution functions and let G be a distribution function on $\mathbb{R}$. We say that F_n converges weakly or in distribution to G and write $F_n \xrightarrow{\mathcal{L}} G$ whenever

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

  for all y constituting continuity points of G.
- Definition (convergence in probability): When a sequence of random variables satisfies $\Pr\{\|Y_n - Y\| > \epsilon\} \to 0$ for all $\epsilon > 0$ and a given (random variable) Y, then we say that Y_n converges in probability to Y, and write $Y_n \xrightarrow{P} Y$.
- $\xrightarrow{\mathcal{L}}$ relates distribution functions. It says that the probabilistic behaviour of a sequence Y_n becomes more and more alike that of the limit Y.

Modes of Convergence

- Theorem (The Continuous Mapping Theorem)
  Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on the range of Y. Then
  $$(a)\ Y_n \xrightarrow{p} Y \Rightarrow g(Y_n) \xrightarrow{p} g(Y),$$
  $$(b)\, Y_n \xrightarrow{\mathcal{L}} Y \Rightarrow g(Y_n) \xrightarrow{\mathcal{L}} g(Y).$$
- Theorem (Slutsky's theorem): Let $X_n \xrightarrow{\mathcal{L}} X$ and let $Y_n \xrightarrow{\mathcal{L}} c$ where $c \in \mathbb{R}$. Then
  $$(a)\ X_n + Y_n \xrightarrow{\mathcal{L}} X + c.$$
  $$(b)\ X_n Y_n \xrightarrow{\mathcal{L}} Xc.$$
- Theorem (Law of Large Numbers): let Y_n be independent random variables with $\mathbb{E}\, Y_k = \mu$ and $\underbrace{\mathbb{E}\,|Y_k| < \infty}_{\text{finite absolute moment}}$ for all k. Then $n^{-1}(Y_1 + \cdots + Y_n) \xrightarrow{p} \mu$.

Modes of Convergence

- Theorem (Central Limit Theorem). Let $\{Y_n\}$ be a sequence of iid random variables with mean μ and variance σ^2 which is assumed finite. Then
$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)) \xrightarrow{\mathcal{L}} N(0,\sigma^2).$$
- Theorem (Delta method): Let $Z_n = a_n(X_n-\theta) \xrightarrow{\mathcal{L}} Z$ where $a_n \in \mathbb{R}^+$ and $\theta \in \mathbb{R}$ for all n and assume $a_n \to \infty$. Let $g()$ be continuously differentiable at θ. Then
$$a_n\{g(X_n)-g(\theta)\} \xrightarrow{\mathcal{L}} g'(\theta)Z.$$
- Vector versions are also provided.

Estimation

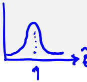
- What is estimation ($\equiv$ "learning" in machine learning)?
- Imagine you assume Y is distributed according to $F(y_1, \ldots, y_n; \theta)$ where $y \in \mathcal{Y}^n$.
- Assume you know the form of $F(y_1, \ldots, y_n; \theta)$ but not the value of θ.
- Guessing θ on having observed $y_1, \ldots, y_n$ is estimation.
- Not that whenever we realise a different set of $Y_1, \ldots, Y_n$ then we realise a different $\widehat{\theta}(Y_1, \ldots, Y_n)$.
- How do we design an estimator $\widehat{\theta}(Y_1, \ldots, Y_n)$?
- A good estimator would produce a value of $\widehat{\theta}(Y_1, \ldots, Y_n)$ near θ.
- We usually address this in terms of the mean and variance of $\widehat{\theta}(Y_1, \ldots, Y_n)$.
- Definition (mean square error): assume that $\widehat{\theta}$ is an estimator of the parameter θ corresponding to the model $F(y; \theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$. The mean square error of $\widehat{\theta}$ is then defined as

$$\mathrm{MSE}\{\widehat{\theta}, \theta\} = \mathbb{E}\left[\|\widehat{\theta} - \theta\|^2\right]. \tag{4}$$

Estimation

UNBIASED ESTIMATOR

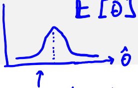
$E[\hat{\theta}]=\theta$



vs

BIASED ESTIMATOR

$E[\hat{\theta}]\neq\theta$



- The mean square error of an estimator, is the bias square plus its variance.
- If the model is not identifiable then you can get the same model with different values of the parameters.
- The Cramér–Rao lower bound provides a bound on the variance of an unbiased estimator.
- Theorem (Rao-Blackwell Theorem): Let $\widehat{\theta}$ be an unbiased estimator of θ that has finite variance. Assume T is sufficient for θ. In this case $\widehat{\theta}^* = \mathbb{E}\{\widehat{\theta}|T\}$ is an unbiased estimator and

$$\mathbb{V}\text{ar}\{\widehat{\theta}^*\} \leq \mathbb{V}\text{ar}\{\widehat{\theta}\}.$$

- Equality is attained if and only if $\Pr\{\widehat{\theta}^* = \widehat{\theta}\} = 1$.

Estimation

- Definition: Let $(Y_1, \ldots, Y_n)$ be a sample of random variables with joint density/frequency $f(y_1, \ldots, y_n; \theta)$ where $\theta \in \mathbb{R}^p$. The likelihood of θ is defined as
$$L(\theta) = f(Y_1, \ldots, Y_n; \theta).$$
- Definition: (Maximum Likelihood Estimation). In the same context, a maximum likelihood estimator (MLE) of $\widehat{\theta}$; is an estimator such that
$$L(\theta) \leq L(\widehat{\theta}), \quad \forall \theta \in \Theta.$$
- Be careful, calculus is not always the answer.
- If the likelihood is twice differentiable in θ, we can verify this by checking

"We should check the Hessian":
$$-\nabla^2_\theta L(\theta)|_{\theta=\widehat{\theta}} \succ 0.$$
- The negative of the Hessian is positive definite.
- When there exists a unique maximum, we speak of the MLE $\widehat{\theta} = \arg_{\theta \in \Theta} \max L(\theta)$.

Estimation

- The next property we shall cover is the equi–variance or invariance of MLEs. If $g(\theta)$ is a bijection, recall that if we are attempting to estimate $\tau = g(\theta)$ then if we form the likelihood

$$L(\theta) = \prod_{j=1}^{n} f(Y_j; \theta),$$

- 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.

Estimation

- Theorem: Let $X_1, \ldots, X_n$ be IID random variables with the same density $f(x;\theta)$. Assume that A1–A6 are satisfied. If the MLE $\widehat{\theta}_n$ exists and is unique, and we have consistency then

centered and re-standardized MLE

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta\right\} \xrightarrow{\mathcal{L}} N\left(0, \mathcal{I}_1(\theta)/\mathscr{I}_1^2(\theta)\right).$$

Furthermore, when we can say that if $\mathcal{I}(\theta) = \mathscr{I}(\theta)$ then

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta\right\} \xrightarrow{\mathcal{L}} N(0, 1/\mathcal{I}_1(\theta)).$$

- For finite samples we often say

$$\widehat{\theta}_n \overset{d}{\approx} N(\theta, 1/\mathcal{I}_n(\theta)).$$

- Despite this, once we allow for bias the MLE is not always the best estimator.
- Need to use decision theory to figure out what to do.

Hypothesis Testing

- Often in science two concurrent theories need to be confronted with the empirical evidence.
- The null hypothesis H_0 which states that $\theta \in \Theta_0$

$$H_0 \; : \; \theta \in \Theta_0.$$

- The alternative hypothesis that postulates $\theta \in \Theta_1$

$$H_1 \; : \; \theta \in \Theta_1.$$

- T is a statistic called a test statistic and;
- C is a subset of the range of T and is called the critical region.
- We can write

$$\delta(\mathrm{Y}) = \mathrm{I}(T(Y_1 \ldots Y_n) \in C).$$

- Take action 0 when H_1 is true–this is a type II error. Take action 1 when H_0 is true–this is a type I error.

Hypothesis Testing

- The Neyman-Pearson Framework
- We declare that we only consider test functions $\delta : \mathcal{X} \mapsto \{0, 1\}$ such that
$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \{\delta : \sup_{\theta \in \Theta_0} \Pr_\theta\{\delta = 1\{\leq \alpha\}.$$
- i.e. rules for which prob of type I error is bounded above by α.
- Jargon: we fix a significance level for our test.
- Within this restricted class of rules, choose δ to minimize prob of type II error:
$$\Pr\{\delta(\boldsymbol{X}) = 0\} = 1 - \Pr\{\delta(\boldsymbol{X}) = 1\}.$$
- Equivalently, maximize the power

$$\beta(\theta, \delta) = \Pr\{\delta(\boldsymbol{X}) = 1\} = \mathbb{E}\,\mathrm{I}\{\delta(\boldsymbol{X}) = 1\} = \mathbb{E}\{\delta(\boldsymbol{X}))\}, \quad \theta \in \Theta_1.$$

Hypothesis Testing

- Neyman-Pearson lemma.
- Likelihood ratio test statistic.
- Score test, Wald test.
- p–values. The p–value is the observed significance level.
- Interval estimation; confidence intervals.
- Multiple testing, Bonferroni, FDR etc.
- Nonparametrics. Kernel Density Estimation.

Bayesian Statistics

- Does not treat the parameter as fixed but unknown, rather models it as random directly.
- Bayes theorem allows us to convert a likelihood to a posterior distribution.
- Minimize the expected posterior loss to arrive at a point estimator.
- Credible intervals permit us to arrive at an interval estimator.

Regression

- Trying to determine the relationship between predictor variables and the response variable.
- Use the linear model to connect the two

$$\mathbb{E}\, Y = X\beta.$$

where $\mathrm{Y} = \begin{pmatrix} Y_1 & \dots & Y_n \end{pmatrix}^T$, is the vector of observations, X is the known $n \times p$ design matrix and $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \dots & \beta_p \end{pmatrix}^T$ is the $p \times 1$ parameter vector.

- We are trying to quantify the systemic variation in Y due to X$\boldsymbol{\beta}$.
- We can also add further assumptions

Second-order assumptions (SOA) var(Y) $= \sigma^2 I_n$ where σ^2 is unknown. Thus var$(Y_i) = \sigma^2$ for all i and the Y_is are uncorrelated.

Normal theory assumptions (NTA) The Y_is are independently and normally distributed with common unknown variance σ^2 so Y $\sim N(\mathrm{X}\beta, \sigma^2 I_n)$.

Regression

- The linear model can be rewritten as

$$Y = X\beta + \epsilon.$$

- Find $\widehat{\beta}$ that minimise the residual sum of squares (RSS), i.e. find

$$\widehat{\beta} = \arg\min_{\beta}(\epsilon^T\epsilon = \sum_{i=1}^{n}\epsilon_i^2).$$

$\epsilon = Y_i - X_i\beta$

- Or

$$\widehat{\beta} = (X^TX)^{-1}X^TY.$$

Why can't we do X^{-1}?

because X is not a square matrix, and may not be invertible. So we assume X^TX has a full rank, so it will have an inverse.

Regression

- $\widehat{\beta}$ is linear in Y, and $\widehat{\beta}$ is unbiased for β.
- Also $var(\widehat{\beta}) = \sigma^2(\mathrm{X}^T\mathrm{X})^{-1}$.
- (Gauss-Markov Theorem) Among all unbiased linear estimates of $\boldsymbol{\beta}$ for a full rank linear model satisfying SOA, any linear combination of the least squares estimator $\widehat{\boldsymbol{\beta}}$ has the smaller or equal variance to that of any other.
- The hat matrix $P = X(X^TX)^{-1}X^T$ is key to understanding the linear model; it is idempotent and its trace is p.

Regression

- You need to understand/be able to reproduce simple linear regression.
- The residual sum of squares (RSS) is defined to be

$$\mathrm{RSS} = Y^T Y - (PY)^T (PY).$$

- With NTA we can make interval estimates, and do testing.
- We can estimate $\widehat{\sigma}^2 = \mathrm{RSS}/(n-p)$.
- Leverage: the ith leverage is p_{ii}. We take notice when $p_{ii} > 2p/n$.
- Weighted least squares.
- Testing for nested models. Likelihood ratio test (see earlier, and tomorrow).
- Outliers and diagnostics.
- Linear algebra. If you need more, look at Schaum's Outline of Linear Algebra, 5th Edition: 568 Solved Problems.

Generalized Linear Regression

- Model selection: forward selection, backwards elimination, cross validation.
- Model selection criteria: AIC, BIC etc.
- Penalization, ridge regression, lasso, shrinkage.
- GLMs. What do we do when the data is not Gaussian?
- Use a link function to connect $\mathbb{E}\, Y_i$ with $x_i^T \beta$.
- Asymptotic normality.
- Deviance.
- Jittered residuals, separation.
- Non–parametric regression for Gaussian data (make link to KDE before). Balancing variance vs bias, and orthogonal function expansion.

Generalized Linear Regression

- GLM setting for non-parametric functions:

$$Y_i = g(x_i) + \epsilon_i \mapsto Y_i \mid x_i \overset{ind}{\sim} \exp\{g(x_i)y_i - \gamma\{g(x_i)\} + S(y_i)\}.$$

- We recognise the latter as a GLM with mean $g(x_i)$.
- Parameterize $g(x)$ using splines, and do penalized max likelihood.
- GLM examples discussed.
- Causal inference and conditional independence representations **not examinable**.

Worked examples tomorrow

- I have been asked to cover:
- For instance over testing hypothesis on the last exercise sheet;
- Non–parametric regression again;
- more concrete examples on GLMs;
- how to handle the separable data case again, along with jitter residuals;
- Non parametric regression;
- Estimate the unknown function $h(x)$ with the modulators and the wavelets.