1 Multiple Linear Regression

1.1 Introduction

First it is important to differentiate between inference and prediction: in Inference we are interested in the relation between the response Y and the covariates $X_1, .., X_n$. More specifically

- Which predictors are associated with the response ?
- What is the relationship between the response and each predictor ?
- Can the relationship between response and covariates be adequately summarized using a linear equation, or is the relationship more complicated ?

Hence for intepretation purpose, there is a trade-off between the facility of interpretation and the complexity of the model e.g paramtrizing the relation as linear (yielding linear regression) give a clear interpration of the parameter i.e the relation between Y and $X_1, .., X_n$ whereas parametrizing f in $Y = f(X)$ as a neural network gives a hard time to intrepret the parameter "think black box model". The latter is in contrast useful for prediction. There is also a trade-off in quality of prediction with respect to the model complexity and the number of data we see. As the goal is to have a low error on prediction we need to avoid overfitting on the data we see. In the extreme case of fitting the function f perfectly to the data, we have fitted noise that is not taken into account in the relation between Y and $X_1, .., X_n$.

1.2 Statistical Learning Framework

Definition 1. *A statistical learning problem is a tuple $(\mathscr{H}, \mathscr{Z}, \mathscr{D}, l)$ where*

- *$\mathscr{H}$ is the class of functions $h : \mathscr{X} \to \mathscr{Y}$ and is called the hypotheses class (and h is called a prediction rule, hypothethis or classifier)*
- *$\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$ is the domain, where*
  - *$\mathscr{X}$ is the state space of observation*
  - *$\mathscr{Y}$ is the label space of observations*
- *$\mathscr{D}$ is a probability distribution on $\mathscr{Z}$*
- *l is a measurable loss function $l : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$*

The learner has to find a predictor $h \in \mathscr{H}$ wich minimizes the true loss (Risk)

$$R_{\mathscr{D}}(h) := \mathbb{E}_{(x,y)\sim\mathscr{D}} l(h(x), y)) \tag{1}$$

Note that $\mathscr{D} = \mathscr{D}^x \times \underbrace{\mathscr{D}^{y|x}}_{\text{Unknown}}$ is unknown. Hence we cannot compute the true risk. For this the learner has access to data sample $\mathcal{S} := \{(x_1, y_1), ..., (x_n, y_n)\} \in (\mathscr{X} \times \mathscr{Y})^n$ draw *iid* with respect to $\mathscr{D}$. Given that sample the learner has to find an algorithm $\mathscr{A} : \cup_{n\in\mathbb{N}} (\mathscr{X} \times \mathscr{Y})^n \to \mathscr{H}$ that returns an estimator $h_{\mathcal{S}} := \mathscr{A}(\mathcal{S})$. Note that the underlying distribution is still unknown, hence a common method is to use a proxy i.e the empirical distribution (or measure) yield the ERM methods

$$R(h_{\mathcal{S}}) = \mathbb{E}_{\mathcal{S}\sim\otimes_{i\in[n]} \underbrace{\mathscr{D}}_{\text{unknown}}} \Big[l(h_{\mathcal{S}}, \mathcal{S}) \Big] \tag{2}$$

$$\mathbb{P}_{\mathcal{S}} \to \hat{\mathbb{P}}_{\mathcal{S}} \implies \tag{3}$$

$$\hat{R}(h_{\mathcal{S}}) : = \int_{(\mathscr{X}\times\mathscr{Y})^{\otimes n}} l(h_{\mathcal{S}}, \mathcal{S}))d\hat{\mathbb{P}}_{\mathcal{S}} = \frac{1}{n} \sum_{i\in[n]} l(h(x_i), y_i) \tag{4}$$

$$\text{ERM:} \quad h_{\mathcal{S}}^{\text{ERM}} = \arg\min_{h_{\mathcal{S}}\in\mathscr{H}} \hat{R}(h_{\mathcal{S}}) \tag{5}$$

Note: This is equivalent to the decision theoritic framework, and notation migth change in this document in order to mathc the literatur.
Notice the following two main interesting loss

- 0-1 Loss: assuming $\mathscr{Y} = \{1, ..., k\}$ this corresponds to a classification problem (binary if $k = 2$) then an intuitive cost function is the 0-1 loss i.e $l_{0-1}(h, z) := \mathbb{I}_{\{h(x)\neq y\}}$ which yield

$$R_{\mathscr{D}}(h) = P_{(x,y)\sim\mathscr{D}}(h(x) \neq y) \quad \text{and} \quad \hat{R}_{\mathcal{S}}(h) \frac{1}{n} \sum_{i\in[n]} \mathbb{I}_{\{h(x_i)\neq y_i\}} \tag{6}$$

- Regression and square Loss: in case $\mathscr{Y} = \mathbb{R}^q$ the task is called a regression task (see later chapter) where in this case the squared loss is used $l(h, z) := (h(x) - y)^2$ which yield

$$R_{\mathscr{D}}(h) = \mathbb{E}_{(x,y)\sim\mathscr{D}}(h(x) - y)^2 \quad \text{and} \quad \hat{R}_{\mathcal{S}} = \frac{1}{n} \sum_{i\in[n]} (h(x_i) - y_i)^2 \tag{7}$$

Also note that in Garantees for machine learning we define the Exess Risk as follow

$$\mathscr{E}(h_{\mathcal{S}}, h^*) = R_{\mathscr{D}}(h_{\mathcal{S}}) - \inf_{h\in\mathscr{H}} R_{\mathscr{D}}(h) \tag{8}$$

where h^* defines the true underlying rule $h^* : \mathscr{X} \to \mathscr{Y}$. whre we developped methods to control this excess risk.

1.3 Basics of Regression Analysis and prediction task

Given the settup in the last section we aim at predicting $\underline{\mathbf{Y}}$ given $\underline{\mathbf{X}}$ where $\underline{\mathbf{Y}}$ depends non trivially to $\underline{\mathbf{X}}$. As in the last section this boils down to finding a measurable map $g : \mathscr{X} \to \mathscr{Y}$ such that $g(\underline{\mathbf{X}})$ is close to $\underline{\mathbf{Y}}$ i.e $\underline{\mathbf{Z}} = |g(\underline{\mathbf{X}}) - \underline{\mathbf{Y}}|$ is small. But as both quantity are random variable, it is not clear what is the right notion of closeness. A somewhat arbitrary idea of small is to define a random variable $\underline{\mathbf{Z}}$ small if $\mathbb{E}\underline{\mathbf{Z}}^2 = (\underbrace{\mathbb{E}\underline{\mathbf{Z}}}_{\text{mean}})^2 + \underbrace{var(\underline{\mathbf{Z}})}_{\text{fluctuation around the squared mean}}$ is small, meaning that the mean of the random variable is small as well as the fluctuations around the mean.
Fitting this idea with last paragrgaph we notice that it corresponds to the notion of risk i.e for any measurable function $g : \mathscr{X} \to \mathbb{R}$ the l_2 risk is

$$R_{\mathscr{D}^{(x,y)}}(g) = \mathbb{E}(\underline{\mathbf{Y}} - g(\underline{\mathbf{X}}))^2 \tag{9}$$

where we assumed $q = 1$ or equivalently $\mathscr{Y} = \mathbb{R}$. Now a natural question is what is the best measurable function g i.e the one minimizing the L2 risk ? the answer is

$$f_{best}(x) = \mathbb{E}_{\mathscr{D}^{y|x}}\big[\underline{\mathbf{Y}}|\underline{\mathbf{X}} = x\big] \quad \forall x \in \mathscr{X} \tag{10}$$

Proof:
Assume $\underline{\mathbf{Z}}$ with $\mathbb{E}\underline{\mathbf{Z}} < \infty$ and $var(\underline{\mathbf{Z}}) < \infty$ then

$$\arg\min_{a\in\mathbb{R}} \mathbb{E}(\underline{\mathbf{Z}} - a)^2 = \mathbb{E}\underline{\mathbf{Z}}^2 - 2a\mathbb{E}\underline{\mathbf{Z}} + a^2 \tag{11}$$

$$\leftrightarrow \tag{12}$$

$$\frac{\partial}{\partial a}(\mathbb{E}\underline{\mathbf{Z}}^2 - 2a\mathbb{E}\underline{\mathbf{Z}} + a^2)_{|a=a_{min}} = 0 \tag{13}$$

$$\leftrightarrow \tag{14}$$

$$a_{min} = \mathbb{E}\underline{\mathbf{Z}} \tag{15}$$

Now we can rewrite the l2 Risk as follow

$$\arg\min_{g\in\mathscr{G}} R_{\mathscr{D}}(g) = \arg\min_{g\in\mathscr{G}} \mathbb{E}_{\underline{\mathbf{X}},\underline{\mathbf{Y}}}(\underline{\mathbf{Y}} - g(\underline{\mathbf{X}}))^2 \tag{16}$$

$$= \arg\min_{g\in\mathscr{G}} \mathbb{E}_{\underline{\mathbf{X}}}\mathbb{E}_{\underline{\mathbf{Y}}|\underline{\mathbf{X}}}(\underline{\mathbf{Y}} - \underbrace{g(\underline{\mathbf{X}}|\underline{\mathbf{X}})}_{=:a\in\mathbb{R}})^2 \tag{17}$$

$$\Leftrightarrow \tag{18}$$

$$g_{min}(\underline{\mathbf{X}}) = f_{best}(\underline{\mathbf{X}}) = \mathbb{E}_{\mathscr{D}^{y|x}}\big[\underline{\mathbf{Y}}|\underline{\mathbf{X}}\big] =: \textbf{ Regression function} \tag{19}$$

$$\square. \tag{20}$$

Hence our goal is to have an estimate of the regression function. Taking any estimator or algorithm $\hat{f}_{\mathcal{S}} := \mathscr{A} : \cup_{n\in\mathbb{N}}(\mathscr{X}\times\mathscr{Y})^n \to \mathscr{H}$ given $\mathcal{S} := \{(x_1, y_1), ..., (x_n, y_n)\} \in (\mathscr{X}\times\mathscr{Y})^n$ where now we need to choose $\mathscr{H}$ with objective to approximate $\mathbb{E}_{\mathscr{D}^{y|x}}\big[\underline{\mathbf{Y}}|\underline{\mathbf{X}}\big]$ we can decompose the l2 risk as follow

$$R_{\mathscr{D}}(\underline{\hat{\mathbf{f}}}_{\mathcal{S}}) = \mathbb{E}_{\mathscr{D}}\big[\underline{\mathbf{Y}} - f_{best}(\underline{\mathbf{X}}) + f(\underline{\mathbf{X}}) - \underline{\hat{\mathbf{f}}}_{\mathcal{S}}(\underline{\mathbf{X}})\big]^2 \tag{21}$$

$$= \underbrace{\mathbb{E}_{\mathscr{D}}\big[\underline{\mathbf{Y}} - f_{best}(\underline{\mathbf{X}})\big]^2}_{\text{deterministic and irreductible}} + \underbrace{\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}} - f_{best}\|_2^2}_{\text{random}} \tag{22}$$

hence as a measure of quality we are interesting in the quantity

$$\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}} - f_{best}\|_2^2 = \int_{\mathscr{X}} (\underline{\hat{\mathbf{f}}}_{\mathcal{S}} - f_{best})^2 dP_{\underline{\mathbf{X}}} = \|\underline{\hat{\mathbf{f}}}_{\mathcal{S}} - f_{best}\|^2_{L_2(P_{\underline{\mathbf{X}}})} \tag{23}$$

Note that one of the biggest advantage of using the square loss is that our function space is a Hilbert space ! Hence as we will show later in linear regression, we will make extensive use of orthogonality.
Note that we aim that the quantity of interest goes to Zero as $n \to \infty$
Hence defining a sequence of positive numbers $\{\phi_{n\to\infty}\}_n \to 0$ we bound the random quantity of interest as follow

- Bounds in Expectation: $\mathbb{E}_{\mathcal{S}_n}\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}_n} - f_{best}\|_2^2 \leq \phi_n$ **represent the average behaviour of the estimator for several sample of size n** But notice that this bound does say anything about the deviation of the random quantity hence follows the above bounds
- Bounds with high probability: $\mathbb{P}\big[\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}_n} - f_{best}\|_2^2 > \phi_n(\delta)\big] \leq \delta$ in this setting the bounds control the tail of the quantity if interest. Usually favored in learning theory or PAC learning
- Bounds with high probability: the above bounds usually follows from concentration around the mean of the quantity i.e $\mathbb{P}\big[\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}_n} - f_{best}\|_2 - \mathbb{E}\|\underline{\hat{\mathbf{f}}}_{\mathcal{S}_n} - f_{best}\|_2 > t\big]$

Such techniques were studied in the courses

- Algorithmic foundations of Data Science
- Guarantees for Machine Learning

1.3.1 Random Design

The random design corresponds to the **statistical learning** setup as described in the last sections. Assume the model

$$\underline{\mathbf{y}}_i = f(\underline{\mathbf{x}}_i) + \underline{\epsilon}_i \quad | \quad \underline{\epsilon}_i \sim \mathcal{N}(0, \sigma^2) \tag{24}$$

then let $\mathcal{S}_{n+1} := \underbrace{\{(\underline{\mathbf{y}}_i, \underline{\mathbf{x}}_i)\}_{i\in[n]}}_{=:\underline{\mathcal{S}_{\mathbf{n}}}} \cup (\underline{\mathbf{y}}_{n+1}, \underline{\mathbf{x}}_{n+1})$ of iid samples drawn from $\mathscr{D}$. Using $\underline{\mathcal{S}_{\mathbf{n}}}$ the goal is to construct a function $\underline{\hat{\mathbf{f}}}_{\mathbf{n}}$ such that $\underline{\hat{\mathbf{f}}}_{\mathbf{n}}(\underline{\mathbf{x}}_{n+1})$ is a good predictor of $\underline{\mathbf{y}}_{n+1}$. Keep in mind that to construct our estimator only a realisation of the sample up to n is know/used i.e $\underline{\mathcal{S}_{\mathbf{n}}} = \mathcal{S}_n$. A good measure of performance for a given $\underline{\mathcal{S}_{\mathbf{n}}} = \mathcal{S}_n$ is L2 risk as described in the first section:

$$R(\hat{f}_n) = \mathbb{E}[\underline{\mathbf{y}}_{n+1} - \hat{f}_n(\underline{\mathbf{x}}_{n+1})]^2 = \underbrace{\mathbb{E}[\underline{\mathbf{y}}_{n+1} - f(\underline{\mathbf{x}}_{n+1})]^2}_{=\sigma^2} + \|\hat{f}_n(\underline{\mathbf{x}}_{n+1}) - f(\underline{\mathbf{x}}_{n+1})\|^2_{L^2(P_{x_{n+1}})} \tag{25}$$

where $P_{x_{n+1}}$ is the marginal distribution of $\underline{\mathbf{x}}_{n+1}$.
Interpretation: the squared L2 norm $\|\hat{f}(\underline{\mathbf{x}}_{n+1}) - f(\underline{\mathbf{x}}_{n+1})\|^2_{L^2(P_{x_{n+1}})}$ measures how close $\hat{f}_n$ is to f in average over the realizations of $\underline{\mathbf{x}}_{n+1}$ i.e how good is the prediction of $\underline{\mathbf{y}}_{n+1}$ in average over the realisations of $\underline{\mathbf{x}}_{n+1}$

1.3.2 Fix Design

In the fix design we assume that the vectors $x_1, ..., x_n$ are deterministic. Of course one could see $x_1, .., x_n$ as realisations of random variables. But their is a fundamental difference in performance measure, here we don't have such thing as the marginal over $\underline{\mathbf{x}}_{n+1}$. Since the design points are considered deterministic our goal is to estimate f only at these point. This problem is sometime call denoising since we aim at recovering the $f(x_1), ..., f(x_n)$ from noisy observations. We can define the fix model as follow

$$\underline{\mathbf{y}} = \mu^* + \underline{\epsilon} \quad | \quad \mu_i^* = f(x_i) \quad \underline{\epsilon}_i \sim \mathcal{N}(0, \sigma^2) \tag{26}$$

$$\underline{\mathbf{y}} \in \mathbb{R}^d \quad \mu^* \in \mathbb{R}^d \tag{27}$$

we define the quality of measure known as the **Mean Squared Error**

$$\text{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i\in[n]} (\hat{f}_n(x_i) - f(x_i))^2 = \frac{1}{n}\|\hat{\mu} - \mu^*\|_2^2 \tag{28}$$

1.4 Theory of Linear Regression

1.4.1 Multiple Linear Regression in Compact Form

Model given $\underline{\mathbf{X}} = \mathbb{X} = [x_1, ..., x_n]$:

$$\mathscr{X} = \mathbb{R}^r \quad \mathscr{Y} = \mathbb{R}^q \quad \beta \in \mathbb{R}^d \quad X_i \in \big\{A \in \mathbb{R}^{q\times p} | rank(A) = min(q,d)\big\} \tag{29}$$

$$\underline{\epsilon}_i \sim_{iid} \mathscr{N}(0,\Sigma) \quad \Psi : \mathscr{X} = \mathbb{R}^r \to \mathbb{R}^{q\times p}(\text{feature map}) \quad r,q,d \in \mathbb{N} \tag{30}$$

then we assume the linear model

$$\underline{\mathbf{y}}_i = X_i\beta + \underline{\epsilon}_i = \Psi(x_i)\beta + \underline{\epsilon}_i \quad \forall i \in [n] \tag{31}$$

with hypotheses space

$$\mathscr{H} = \big\{x \to \Psi(x)\beta | \psi : \mathbb{R}^r \to \mathbb{R}^{q\times d}, \beta \in \mathbb{R}^d\big\} \tag{32}$$

MLE: Given iid sample $\underline{\mathcal{S}_\mathbf{n}} = \{(\underline{\mathbf{X_i}}, \underline{\mathbf{y_i}})\}_{i\in[n]} \sim P_\beta^{\otimes n} = \mathbb{P}_\beta$. We notice that maximizing wrt to β mle of $\underline{\mathcal{S}_\mathbf{n}}|\underline{\mathbf{X}} = \mathbb{X}$ and $\underline{\mathcal{S}_\mathbf{n}}$ yield the same programm (random and fix design). Also recall that in the statisical learning framework in section 1 we assumed that $\mathscr{D}^{y|x}$ was the unknown part of the distribution. Hence now we have assumed some model class or hypotheses testing class paramtrized by β. In the above model we assume gaussian distribution which yield

$$\underline{\tilde{\mathbf{y}}}_i := \underline{\mathbf{y}}_i | \underline{\mathbf{X}}_i = x_i \sim \mathscr{N}(X_i\beta, \Sigma) \tag{33}$$

hence knowing it is continous wrt to the lebesgue measure we can write the Likelyhood function as follow

$$\mathscr{L}_{\underline{\mathcal{S}_\mathbf{n}}|\mathbb{X}}(\beta) = \prod_{i\in[n]} f_{\mathscr{N}(X_i\beta,\Sigma)}(\tilde{y}_i) \tag{34}$$

$$= \prod_{i\in[n]} (2\pi)^{-q/2} det(\Sigma)^{-1/2} exp\big(-\frac{1}{2}(\tilde{y}_i - X_i\beta)^T\Sigma^{-1}(\tilde{y}_i - X_i\beta)\big) \tag{35}$$

Now taking the maximum yield

$$\beta_{MLE} = \arg\max_{\beta\in\mathbb{R}^d} \log \mathscr{L}_{\underline{\mathcal{S}_\mathbf{n}}|\mathbb{X}}(\beta) \tag{36}$$

$$= \arg\max_{\beta\in\mathbb{R}^d} \text{const} - \frac{1}{2}\sum_{i\in[n]} \|\Sigma^{-1/2}(\tilde{y}_i - X_i\beta)\|^2 \tag{37}$$

$$= \arg\min_{\beta\in\mathbb{R}^d} \frac{1}{2}\sum_{i\in[n]} \|\Sigma^{-1/2}(\tilde{y}_i - X_i\beta)\|^2 \tag{38}$$

we have a sum of convex functions which is again convex over a convex domain, meaning a global minimum exists and is attained at $\nabla(\cdot)|_{\beta=\beta_{MLE}} = 0$ i.e

$$\nabla_{\beta=\beta_{MLE}} \Big\{\frac{1}{2}\sum_{i\in[n]} \|\Sigma^{-1/2}(\tilde{y}_i - X_i\beta)\|^2\Big\} = \sum_{i\in[n]} (-2X_i^T\Sigma^{-1}\tilde{y}_i + 2X_i^T\Sigma^{-1}X_i\beta) = 0 \tag{39}$$

$$\Leftrightarrow \tag{40}$$

$$\beta_{MLE} = \Big(\sum_{i\in[n]} X_i^T\Sigma^{-1}X_i\Big)^\dagger \Big(X_i^T\Sigma^{-1}\tilde{y}_i\Big) \tag{41}$$

Note: Here we assumed X_i to have full rank, hence the pseudo inverse is an inverse as Σ is psd matrix

Note that when $q = 1$ we have the more standard form

$$\underline{\mathbf{y}}_i = \langle X_i, \beta\rangle + \underline{\epsilon} = \langle \psi(x_i), \beta\rangle + \underline{\epsilon} \tag{42}$$

$$\underbrace{\underline{\mathbf{Y}}}_{\in\mathbb{R}^n} = \underbrace{X}_{\in\mathbb{R}^{n\times d}} \beta + \underline{\epsilon} \quad | \quad X = [\psi(x_1), ..., \psi(x_n)]^T \tag{43}$$

$$\tag{44}$$

1.4.2 Multivariate Linear Regression

Model:

1.5 Theory of Hypotheses Testing

The goal of hypotheses testing in a statistical testing problem is to decide wether a hypothese that has been formulated is correct or not. The choice lie between two decisions, accepting or rejecting the hypotheses. A decision procedure aiming at deciding wether or not a hypothese has to be rejected is called a test of the hypotheses.

1.5.1 Modeling the Problem

Let $\underline{\mathbf{x}} : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathscr{X}, \mathscr{B})$ be a random variable, $(\Omega, \mathscr{A}, \mathbb{P})$ a probability space and $(\mathscr{X}, \mathscr{B})$ a measurable space. The random variable $\underline{\mathbf{x}}$ induces a probability measure P_x on $(\mathscr{X}, \mathscr{B})$ given by $P_x(B) = \mathbb{P}(\underline{\mathbf{x}}(\omega) \in B) \quad \forall B \in \mathscr{B} \quad \forall \omega \in \Omega$. Assume the distribution of $\underline{\mathbf{x}}$ to be parametrized by some unknown θ and belonging to some parametrized family of distribution i.e

$$P_x = P_\theta \in \mathscr{P} := \Big\{P_\theta : \theta \in \Theta\Big\}$$

The decision procedure to accept or reject a hyptohese has to be based on realisations of $\underline{\mathbf{x}}$. Defining the hypotheses $H_0 : \theta \in \Theta_0$ and the alternative $H_1 : \theta \in \Theta_1$ with $\Theta = \Theta_0 \cup \Theta_1; \Theta_0 \cap \Theta_1 = \emptyset$. Note that we could also define an hypotheses as follow

$$g : \Theta \to \Gamma \quad ; H_0 = \gamma_0 \in \Gamma; \quad \theta \in \{\vartheta \in \Theta_0 : g(\vartheta) = \gamma_0\} \tag{45}$$

for sake of simplicity we model g as the identity in this document. As the procedure has to be made on realisations of $\underline{\mathbf{x}}$ we can divide the space $\mathscr{X}$ in the following regions

$$\mathscr{X} = \mathcal{S}_0 \cup \mathcal{S}_1 \tag{46}$$

$$\mathcal{S}_0 := \{x \in \mathscr{X} \quad \text{s.t } H_0 \quad \text{is true}\} = \text{acceptance region} \tag{47}$$

$$\mathcal{S}_1 := \{x \in \mathscr{X} \quad \text{s.t } H_0 \quad \text{is rejected}\} = \text{rejection region} \tag{48}$$

Definition 2 (critical function)**.** *A critical function ϕ is any function of the form $\phi(x) \in [0,1] \quad \forall x \in \mathscr{X}$*

Definition 3 (test function)**.** *A test function is a critical function $\phi(x)$ such that $\forall x \in \mathscr{X}$ we accept H_0 w.p $1 - \phi(x)$ and reject H_1 w.p $\phi(x)$*

A test can made two type of errors: accepting H_0 when it is actually wrong or rejecting H_0 when it is actually true. This can be summarised as follow

Definition 4 (Error Type-I). *"Rejecting when True": for $\theta \in \Theta_0$ the function $\theta \to \mathbb{E}_\theta \phi(\underline{\mathbf{x}})$ is called type-I error*

Definition 5 (Power). *The power of a test procedure is defined as : for $\theta \in \Theta_1$ the function $\beta\theta = \theta \to \mathbb{E}_\theta \phi(\underline{\mathbf{x}})$ is called the power of the testing procedure. "Think: how powerful the test is for rejecting Null Hyp0theses when actually wrong"*

Definition 6 (Error Type-II). *"Accepting when False": the function $1 - \beta(\theta)$ is called error of type-II*

Randomised Test: given $\underline{\mathbf{x}} = x$ and $\phi(x) \in [0,1]$ a random test is defined as:

$$\underline{\delta}(x) = \begin{cases} d_0 & \text{w.p} \quad 1-\phi(x) \\ d_1 & \text{w.p} \quad \phi(x) \end{cases} \qquad \equiv \underline{\delta}(x) \sim Ber(1-\phi(x)) \tag{49}$$

$$d_0 = \text{decision to accept } H_0 \tag{50}$$

$$d_1 = \text{decision to reject } H_0 \tag{51}$$

The term random means that the experiment producing the outcomes decision d_0 and d_1 is random, explaining why the decision are accompanied by probabilities.
We notice that the test is completly characterized by the test function ϕ. Let us show the probability of error type - I for a randomized test "assume $\theta \in \Theta_0$ and reject H_0"

$$\begin{aligned} Pr\{\underline{\delta}(\underline{\mathbf{x}}) = d_1\} &= \mathbb{E}_{\underline{\delta},\underline{\mathbf{x}}} \mathbb{I}\{\underline{\delta}(\underline{\mathbf{x}}) = d_1\} && (52)\\ &= \mathbb{E}_{\underline{\mathbf{x}}} \mathbb{E}_{\underline{\delta}|\underline{\mathbf{x}}=x} \mathbb{I}\{\underline{\delta}(x) = d_1\} && (53)\\ &= \mathbb{E}_{\underline{\mathbf{x}}} Pr\{\underline{\delta}(x) = d_1\} && (54)\\ &= \mathbb{E}_{\underline{\mathbf{x}}} \phi(\underline{\mathbf{x}}) && (55)\\ &= \mathbb{E}_\theta \phi(\underline{\mathbf{x}}) && (56) \end{aligned}$$

Non-Randomised Test: given $\underline{\mathbf{x}} = x$ and $\phi(x) \in \{0,1\}$ the same settup as before yield a non randomised procedure:

$$\delta(x) = \begin{cases} d_0 & \text{w.p} \quad 1-\phi(x) \\ d_1 & \text{w.p} \quad \phi(x) \end{cases} \tag{57}$$

The test is no longer "random" as $\phi(x)$ takes 0 or 1 meaning that the underlying experiment outputs decisions to accept or reject the hypothese (d_0, d_1) with probabilities either 1 or 0. The error of type-I is as follow

$$\begin{aligned} Pr\{\delta(\underline{\mathbf{x}}) = d_1\} &= \mathbb{E}_{\underline{\mathbf{x}}} \mathbb{I}\{\delta(\underline{\mathbf{x}}) = d_1\} && (58)\\ &= \mathbb{E}_\theta \phi(\underline{\mathbf{x}}) && (59) \end{aligned}$$

we also notice

$$\begin{aligned} \mathbb{E}_\theta \phi(\underline{\mathbf{x}}) &= \int_{\mathcal{X}} \phi(x) dP_\theta(x) && (60)\\ &= \int_{\mathcal{S}_0} 0 dP_\theta(x) + \int_{\mathcal{S}_1} 1 dP_\theta(x) && (61)\\ &= \mathbb{I}\{x \in \mathcal{S}_1\} && (62) \end{aligned}$$

were the error of type - I corresponds then to the indicator of the rejection region.

Goal: given the above settup we wish to control for error of type-I and type-II i.e find a function ϕ sucht that

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \phi(\underline{\mathbf{x}}) \leq \alpha \quad | \quad \alpha \in (0,1) \tag{63}$$

$$\text{and} \tag{64}$$

$$\beta(\theta) \text{ is maximal } \forall \theta \in \Theta_1 \tag{65}$$

1.6 Statistical Modelling for Linear Regression

1.6.1 Gramm Schmidt and Multiple Regression from single Regression

- ESLII
- stat modelling slides and maelm summary
-

2 Non-parametric Density Estimation

- Good notes
- script compt stat 2023 (very summarized)
- Elements of computaitonal statistics

3 Non-parametric Regression

- Theory of statistics
- ESL
- Introduction to non parametric estimation

4 Classification

- Book: All of statistics (math compact)
- Book: Intro to Statistical Learning (well written applied)
- mathematics tool for ML (theory)
- Wikipedia for proof of bayes classifier very good
- (`https://en.wikipedia.org/wiki/Bayes_classifier`)