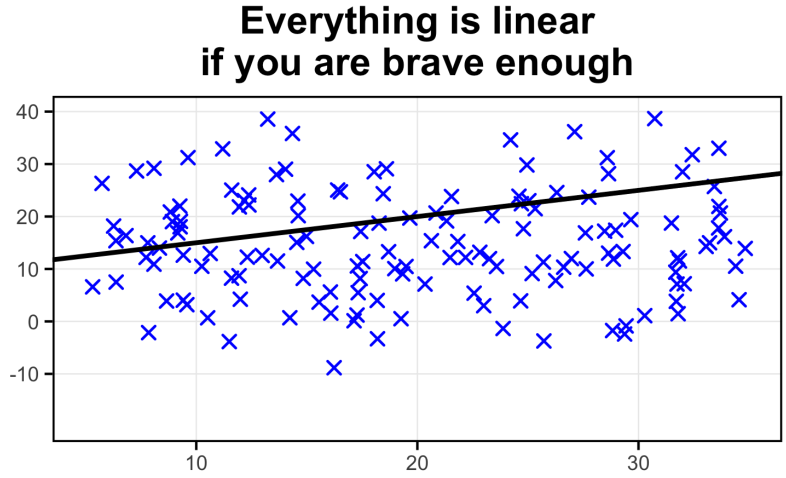
Regression and Analysis of Variance

Fall 2025, Math 533 Course Notes

McGill University



By Jiajun Zhang

December 1, 2025

Acknowledgments

I would like to extend my deepest thanks and appreciation to the following people, without whose support this note would not have been possible:

[Mehdi Dagdoug], *Assistant Professor, McGill University*
I would like to express my deepest gratitude to Professor Dagdoug,
the instructor for this course, whose guidance and expertise
were invaluable throughout the development of my notes.

Despite all efforts, there may still be some typos, unclear explanations, etc. If you find potential mistakes, or any suggestions regarding concepts or formats, etc., feel free to reach out to the author at *zhangjohnson729@gmail.com*.

Contents

1 Review of Asymptotic Statistics **3**

1.1 Random Variables and Convergence 3

1.2 Law of Large Numbers and Central Limit Theorem 5

1.3 Convergence as Functions of Random Variables 7

1.4 Some Random Facts 7

2 Linear Regression & Regression Analysis **9**

2.1 An Introduction 9

2.2 Least Square Loss 11

2.3 Linear Projection Coefficient 13

2.4 Design Matrix and Linear Models 19

2.5 Closing Remark 21

3 The Ordinary Least Squares Estimator **21**

3.1 Derivations of the OLS Estimator 21

3.2 Hat Matrix and Annihilator Matrix 27

3.3 Properties of OLS and Gauss-Markov Theorem 31

4 Inference in the Gaussian Regression Model **37**

4.1 Basic Theory of Testing 37

4.2 t-Statistic and F-Statistic for testing 39

4.3 Confidence Interval for Predictions 44

5 Asymptotic Properties and Inference **47**

6 Analysis of Variance **51**

7 Appendix: Statistical Inference **53**

7.1 Basic Results of Random Samples 53

7.2 Large Sample Theory 55

7.3 Common Types of Confidence Intervals in $\mathbb{R}$ 62

7.4 Univariate Hypothesis Testing 70

Review of Asymptotic Statistics

1.1 Random Variables and Convergence

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be the probability space, where Ω is some arbitrary non-empty set (we usually denote as the sample space). $\mathscr{F}$ is another set which contains a collection of subsets of Ω that satisfies: (i) $\Omega \in \mathscr{F}$; (ii) Closed under set compliments; (iii) Closed under countable unions. $\mathscr{F}$ is also called a σ-algebra of Ω. We denote $\mathfrak{B}(\mathbb{R})$ as the σ-algebra generated by all the open sets of $\mathbb{R}$, which is called Borel σ-algebra. $\mathbb{P}$ is the probability measure (a set function) $\mathbb{P} : \mathscr{F} \to [0,1]$ such that: (i) $\mathbb{P}(\Omega) = 1$; (ii) If $\{X_i\}_{i=1}^{\infty} \subseteq \mathscr{F}$ and $X_i \cap X_j = \varnothing$ whenever $i \neq j$ then $\mathbb{P}\Big(\bigcup_{i=1}^{\infty} X_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(X_i)$.

A random variable X is also a function $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $\forall B \subseteq \mathfrak{B}(\mathbb{R})$, its pre-image $X^{-1}(B) \subseteq \mathscr{F}$. We will work with a sequence of random variables $\{X_i\}_{i=1}^{\infty}$ defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. There are four types of convergences we are interested in, namely **weak convergence**, **convergence in probability**, **convergence in L^p**, **convergence almost surely**. We write $X_n \overset{L}{\to} X$, meaning X_n converges to X weakly, (or in law, in distribution) if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ for all x such that $x \mapsto \mathbb{P}(X \leq x)$ is continuous, or by saying $F_n(x) \to F(x)$ where F represents the cumulative distribution function. We write $X_n \overset{P}{\to} X$, meaning X_n converges to X in probability, if $\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \to 0$. We write $X_n \overset{L^p}{\to} X$, meaning X_n converges to X in L^p $(p \geq 1)$, if $\mathbb{E}|X_n - X|^p \to 0$. Lastly if X_n converges to X almost surely, we have $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$ and denote as $X_n \overset{a.s}{\to} X$.

Theorem

Theorem 1. *(Markov's Inequality) Let $X \geq 0$ a.s, then for all $t \geq 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \tag{1.1}$$

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X \cdot \mathbb{1}\{X \geq t\}] + \mathbb{E}[X \cdot \mathbb{1}\{X < t\}] \\ &\geq \mathbb{E}[X \cdot \mathbb{1}\{X \geq t\}] \\ &= t\mathbb{P}(X \geq t).\end{aligned}$$

■

We may use Markov's inequality to deduce Chebyshev's inequality: If $\mathbb{E}[X^2] < \infty$ then $\forall t > 0$,

$$\mathbb{P}\Big(|X - \mathbb{E}X| > t\Big) \leq \frac{\mathbb{V}(X)}{t^2} \tag{1.2}$$

where we have

$$\mathbb{P}\Big(|X - \mathbb{E}X| > t\Big) = \mathbb{P}\Big(|X - \mathbb{E}X|^2 > t^2\Big) \leq \frac{\mathbb{V}(X)}{t^2} \tag{1.3}$$

where by definition $\mathbb{E}[|X-\mathbb{E}X|^2] := \mathbb{V}(X)$.

In general, the different modes of convergence can be related by the following diagram:

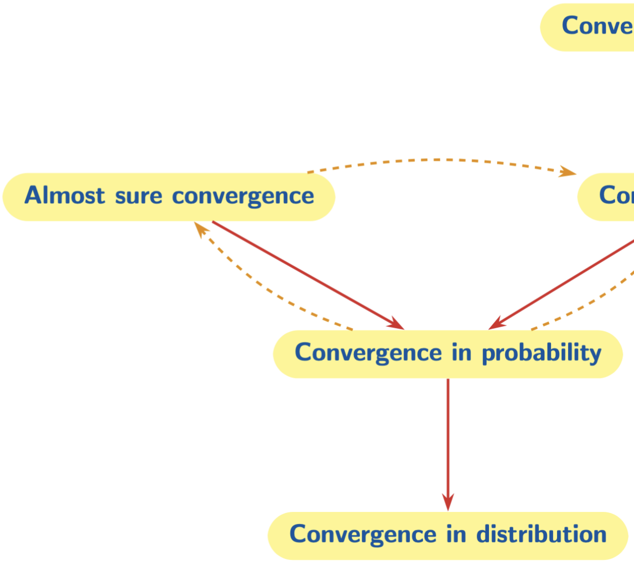

Convergence in L^q, $q > p$
Almost sure convergence
Convergence in L^p
Convergence in probability
Convergence in distribution

Figure 1: The diagram shows the relations between different modes of convergence. The arrows in red means direct implication without any further condition, while arrows in orange will hold if extra conditions are given. The general structure is that, convergence in probability implies convergence almost surely along a subsequence; Convergence in probability with uniform integrability would imply convergence in L^p; Convergence almost surely when dominated convergence theorem applies will imply convergence in L^p.

The next few theorems will show the proofs for some arrows. Denote $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$.

Theorem

Theorem 2. *If $X_n \xrightarrow{P} X$, then there exists a subsequence n_k of $\mathbb{N}$ such that $X_{n_k} \overset{a.s}{\rightarrow} X$.*

Proof. Assume $X_n \xrightarrow{P} X$, then $\forall \varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$, meaning that $\forall k \geq 1$, $\exists n_k$ such that $\mathbb{P}(\{X_{n_k} - X| > 1/k\}) \leq 1/k^2$, denote $A_k := \{X_{n_k} - X| > 1/k\}$, then by *Borel-Cantelli Lemma*, we have

$$\mathbb{P}\left(\bigcap_{l=1}^{\infty}\bigcup_{k=l}^{\infty} A_k\right) = \lim_{l\to\infty} \mathbb{P}\left(\bigcup_{k=l}^{\infty} A_k\right) \leq \lim_{l\to\infty} \sum_{k=l}^{\infty} \mathbb{P}(A_k) = 0, \tag{1.4}$$

meaning that for almost everywhere, $\exists l$, such that $\forall k \geq l : |X_{n_k} - X| \leq 1/k$, which means that for almost everywhere, $\lim_{k\to\infty} |X_{n_k} - X| = 0$, thus $X_{n_k} \overset{a.s}{\rightarrow} X$. ■

Theorem

Theorem 3. *If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{P} X$.*

Proof. The proof is straightforward, we have

$$\mathbb{P}\{|X_n - X| > \varepsilon\} = \mathbb{P}\{|X_n - X|^p > \varepsilon^p\} \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} \to 0. \tag{1.5}$$

■

1.2 Law of Large Numbers and Central Limit Theorem

Assume we have a sequence of random variables $\{X_i\}_{i=1}^{\infty} \overset{i.i.d}{\sim} \mathbb{P}_x$, then in this section we will introduce some important theorems in probability.

Theorem

Theorem 4. *(Weak Law of Large Numbers) Assume $\mathbb{E}|X| < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X]. \tag{1.6}$$

Proof. Our task is much easier if we assume $\mathbb{E}|X|^2 < \infty$. Then Chebyshev's inequality states that

$$\begin{aligned} \mathbb{P}\left\{\left|\overline{X}_n - \mathbb{E}|X|\right| > \varepsilon\right\} &\leq \frac{\mathbb{V}(\overline{X}_n)}{\varepsilon^2} \\ &= \frac{\mathbb{V}(X)}{n\varepsilon^2} \to 0. \end{aligned}$$

■

In fact a stronger statement can be shown, known as the Strong Law of Large Numbers (SLLN), where

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \mathbb{E}[X]. \tag{1.7}$$

So far don't think about the proof of (1.6). If you really want some torture, check out Probability Theory by Daniel Stroock, Section 1.4.. Next we introduce a technical lemma to prove central limit theorem:

Lemma

Lemma 1. *(Levy Continuity Theorem) The characteristic function of X is defined as*

$$\mathbb{1}_X(t) := \mathbb{E}[\exp(itX)]. \tag{1.8}$$

Then $X_n \xrightarrow{L} X$ iff $f_{X_n}(t) \overset{pointwise}{\to} f_X(t)$ for all $t \in \mathbb{R}$.

Now we state the central limit theorem:

Theorem

Theorem 5. *Let* $\{X_i\}_{i=1}^{\infty} \overset{i.i.d}{\sim} f$ *and assume* $\mathbb{E}[X^2] < \infty$*, then*

$$\sqrt{n}\Big(\overline{X}_n - \mathbb{E}[X]\Big) \xrightarrow{L} N(0, \mathbb{V}(X)). \tag{1.9}$$

Proof. WLOG assume $\mathbb{E}X = 0, \mathbb{V}(X) = 1$ then

$$\begin{aligned}
\mathbb{1}_{\sqrt{n}\overline{X}_n}(t) &= \mathbb{E}\Big[\exp\Big(\sqrt{n}it\overline{X}_n\Big)\Big] \\
&= \mathbb{E}\left[\exp\left(\frac{it(X_1 + \cdots + X_n)}{\sqrt{n}}\right)\right] \\
&= \left(\mathbb{E}\left[\exp\left(\frac{itX}{\sqrt{n}}\right)\right]\right)^n \\
&= \left[\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right)\right]^n.
\end{aligned}$$

Then a Taylor expansion around 0 will yield

$$\begin{aligned}
\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right) &= \mathbb{1}_X(0) + \mathbb{1}'_X(0)\cdot\frac{t}{\sqrt{n}} + \mathbb{1}''_X \cdot \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \\
&= 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)
\end{aligned}$$

This is because

$$\begin{aligned}
\mathbb{1}'_X(t)\Bigg|_{t=0} &= \frac{d}{dt}\Bigg|_{t=0} \int_B f(x)\cdot\exp(itX)dx \qquad (1.10)\\
&= \int_B \frac{d}{dt}\Bigg|_{t=0} f(x)\cdot\exp(itX)dx \qquad (1.11)\\
&= \int_B ixf(x)\cdot\exp(itX)\Bigg|_{t=0} dx \qquad (1.12)\\
&:= i\mathbb{E}[X] \qquad (1.13)\\
&= 0. \qquad (1.14)
\end{aligned}$$

A similar statement can be drawn: $\mathbb{1}''_X(0) = i^2\mathbb{E}[X^2] = -1$. Use the fact that $\left(1+\frac{x}{n}\right)^n \sim e^x$ for large n, we have

$$\begin{aligned}
\mathbb{1}_{\sqrt{n}\overline{X}_n}(t) &= \left[\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \\
&= \exp\left(-\frac{t^2}{2}\right). \qquad (1.15)
\end{aligned}$$

By the uniqueness of the characteristic function, (1.15) is the characteristic function of $N(0,1)$.

■

1.3 Convergence as Functions of Random Variables

The main theorems we will introduce are continuous mapping theorem and Slutsky's theorem:

Theorem

Theorem 6. *(Continuous Mapping Theorem) Assume $X_n \xrightarrow{m} X$, where m represents any mode of convergence (i.e in distribution, in probability, almost surely, in $\mathscr{L}^p$), and let f be continuous at x, then $f(X_n) \xrightarrow{m} f(X)$.*

Proof. I am not proving this. ■

Theorem

Theorem 7. *(Slutsy's Theorem) Assume $X_n \xrightarrow{L} X$, $Y_n \xrightarrow{P} c$ for some constant c, then: (i) $X_n + Y_n \xrightarrow{L} X + c; (ii) X_nY_n \xrightarrow{L} cX$; (iii) $X_n/Y_n \xrightarrow{L} X/c$.*

Proof. I am not proving this. ■

1.4 Some Random Facts

Proposition

Proposition 1. *(Law of Total Expectation) Let X, Y be random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and g be a measurable function, then $\mathbb{E}[g(X)] = \mathbb{E}[\mathbb{E}[g(X)|Y]]$.*

Proof. Wlog, let X, Y be continuous with density function $f(x), g(y)$, then

$$\mathbb{E}[g(X)] = \int \mathbb{E}[g(x)|Y = y]dy = \int\int g(x) f_{X|Y}(x|y) dydx := \mathbb{E}[\mathbb{E}[g(X)|Y]]. \tag{1.16}$$

■

Proposition

Proposition 2. *(Variance Decomposition) Let X be square integrable, then*

$$\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y]). \tag{1.17}$$

Proof. By direct computation, one would have

$$\begin{aligned}
\mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y]) &= \mathbb{E}\Big[\mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2\Big] + \mathbb{E}\Big[\mathbb{E}[X|Y]\Big]^2 - \Big(\mathbb{E}[\mathbb{E}[X|Y]]\Big)^2 \qquad (1.18)\\
&= \mathbb{E}[X^2] - \mathbb{E}(\mathbb{E}[X|Y])^2 + \mathbb{E}(\mathbb{E}[X|Y])^2 - (\mathbb{E}[X])^2 \qquad (1.19)\\
&:= \mathbb{V}(X). \qquad (1.20)
\end{aligned}$$

where in (1.19) we used the law of total expectation. ■

Linear Regression & Regression Analysis

2.1 An Introduction

Our basic set up: Let $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ be a random vector defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, let $\mathbb{P}_{\boldsymbol{x},y}$ denote the joint distribution of $\boldsymbol{x}, y$. Without specification, all random variables are square-integrable, that is, $\mathbb{E}|X|^2 < \infty$.

Definition

Definition 1. *The coordinates of $\boldsymbol{x}$, denoted $\{x_j\}_{j\in[p]}$ is called the* ***covariates****, or independent variables; y is called the dependent variable, or the* ***response****, or the variable of interest; p is the dimension of the covariates.*

In this course, we will consider all response y are continuous. Recall that a numeric variable is said to be discrete if its support is at most countable; otherwise it is said to be continuous. The continuous response we can think of are income, weight. For covariates $\boldsymbol{x}$, it normally has the following types:

(i) Quantitative continuous covariates, like income, weight, etc.

(ii) Transformations of quantitative inputs, like different functions of a original covariate. Say $x_2 = x_1^2, x_3 = \log(x_1)$, etc.

(iii) Functions of original covariates, say $x_3 = x_1 + x_2$.

(iv) Categorical covariates, usually coded as dummy variables. Like for gender, we use $X = 1$ for male and $X = 0$ for female, for example.

The goal of regression is to analysis and find the relation between the covariates $\boldsymbol{x}$ and the response y. We recall that $\mathbb{P}_{\boldsymbol{x},y}$ is the joint distribution of $\boldsymbol{x}, y$, which can be written as

$$\mathbb{P}_{\boldsymbol{x},y} = \mathbb{P}_{\boldsymbol{x}} \cdot \mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}} \tag{2.1}$$

through a conditional probability argument, in above $\mathbb{P}_{\boldsymbol{x}}$ is the marginal distribution of all covariates and $\mathbb{P}_{y|\boldsymbol{x}}$ is the conditional distribution of y given $\boldsymbol{x}$. It is not easy to find the conditional distribution, but we can work with conditional expectation $\mathbb{E}[y|\boldsymbol{x}]$ instead.

Theorem

Theorem 8. *Let $\mathscr{M}$ denote the set of measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$ and denote by m the function $m : \boldsymbol{u} \to \mathbb{E}[\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{u}] \in \mathscr{M}$, then*

$$m = \arg\min_{f\in\mathscr{M}} \mathbb{E}\left[\{y - f(\boldsymbol{x})\}^2\right]. \tag{2.2}$$

This is known as the best prediction property under the L^2 risk.

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\{y-f(\boldsymbol{x})\}^2\right] &= \mathbb{E}\left[\{y-m(\boldsymbol{x})+m(\boldsymbol{x})-f(\boldsymbol{x})\}^2\right]\\ &= \mathbb{E}\left[\{y-m(\boldsymbol{x})\}^2\right]+\mathbb{E}\left[\{m(\boldsymbol{x})-f(\boldsymbol{x})\}^2\right]+2\mathbb{E}\left[\{y-m(\boldsymbol{x})\}\cdot\{m(\boldsymbol{x})-f(\boldsymbol{x})\}\right],\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[\{y-m(\boldsymbol{x})\}\cdot\{m(\boldsymbol{x})-f(\boldsymbol{x})\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\{y-m(\boldsymbol{x})\}\cdot\{m(\boldsymbol{x})-f(\boldsymbol{x})\}\middle|\boldsymbol{x}\right]\right]\\ &= \mathbb{E}\left[\{m(\boldsymbol{x})-f(\boldsymbol{x})\}\cdot\mathbb{E}\left[\{y-m(\boldsymbol{x})\}\middle|\boldsymbol{x}\right]\right]\\ &= \mathbb{E}\left[\{m(\boldsymbol{x})-f(\boldsymbol{x})\}\cdot\left(\mathbb{E}\left[y\middle|\boldsymbol{x}\right]-m(\boldsymbol{x})\right)\right]\end{aligned}$$

and by definition, $m(\boldsymbol{x})=\mathbb{E}[y|\boldsymbol{x}]$ so the above term will become zero. Hence now we have

$$\mathbb{E}\left[\{y-f(\boldsymbol{x})\}^2\right]=\mathbb{E}\left[\{y-m(\boldsymbol{x})\}^2\right]+\mathbb{E}\left[\{m(\boldsymbol{x})-f(\boldsymbol{x})\}^2\right] \tag{2.3}$$

as a function of f, so it is easy to see it will attain the minimum when $f(\boldsymbol{x})=m(\boldsymbol{x})=\mathbb{E}[y|\boldsymbol{x}]$. $\blacksquare$

We can also prove the theorem using projection theorem:

Theorem

Theorem 9. *Let $\mathscr{H}$ be a Hilbert space, $\mathscr{W}\subseteq\mathscr{H}$ closed, and $\forall x\in\mathscr{H}$ there is a unique $P_{\mathscr{W}}(x)\in\mathscr{W}$ such that*

$$||x-P_{\mathscr{W}}(x)||=\inf_{y\in\mathscr{W}}||x-y||. \tag{2.4}$$

Then consider $\mathscr{M}$ denote the set of all measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$, and $L^2(\mathscr{M}):=\{z=f(x),f:\mathbb{R}^p\to\mathbb{R}\}\subseteq L^2$, we know L^2 is a Hilbert space with inner product defined by $\{Z_1,Z_2\}=\mathbb{E}[Z_1Z_2]$ and then using projection theorem $\mathbb{E}[(y-P_{\mathscr{M}}(y))f(x)]=0$ for all f, and we choose such g and m then $\mathbb{E}[(m(x)-g(x))f(x)]=0$, which is $\mathbb{E}[(m(x)-g(x))^2]=0$ which implies $m(x)=g(x)$ a.e.

We may also write $y=m(\boldsymbol{x})+(y-m(\boldsymbol{x}))=m(\boldsymbol{x})+\varepsilon$ as a derivation, and ε is defined as the noise term, $m(\boldsymbol{x})$ is the regression function, if $\varepsilon=0$ almost surely, the model is said to be noiseless.

In practice, we do not have access to the true conditional distribution $\mathbb{P}_{y|\boldsymbol{x}}$, what we have is an observable sample $D_n=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)\}$ of n i.i.d random variables with joint distribution $\mathbb{P}_{\boldsymbol{x},y}$. We say D_n is the observed sample, and n is the sample size. (Unless stated, we will assume $p<n$). The goal of regression is to estimate and understand the relationship between $\boldsymbol{x}$ and y, using only the observed sample.

The covariates can choose to be either fixed or random. For fixed design the covariates are constant and they do not change; while in a random design the covariates are random. Both methods are valid in different contexts: Fixed design regression is especially appropriate when the data has been generated rather than observed; while random design regression allows for a more general treatment and is especially suitable for non-experimental sciences such as econometrics. A example is that, suppose we want to study the relationship between the person's age (covariate) and salary (the response y), then a fixed covariate design would be choose people from different ages by design; however a random design would be choose people at random and then we have accesss to their age. This selection process is treated as random.

2.2 Least Square Loss

Recall that if we want to use $f(\boldsymbol{x})$ to model the response y, we have the mean squared error (MSE) defined by

$$\mathbb{E}[(y-f(\boldsymbol{x}))^2] = \mathbb{E}[(y-m(\boldsymbol{x}))^2] + \mathbb{E}[(m(x)-f(x))^2] \tag{2.5}$$

where $m(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{X}=\boldsymbol{x}]$, then based on our observable data $\mathscr{D}_n$, how can we find such a function f? Note that the first term on the right hand side above does not depend on f and it suffices to find

$$f^* = \arg\min_{f\in\mathscr{M}} \mathbb{E}[(y-f(\boldsymbol{x}))^2] \tag{2.6}$$

Also note that in reality we do not have access to the moment either, since we don't know the true distribution. But instead we can replace by its empirical estimate:

$$\widehat{f} := \arg\min_{f\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - f(x_i)\Big)^2 \tag{2.7}$$

Definition

Definition 2. *Let $f \in \mathscr{M}$ be a given function, the least square loss of f over $\mathscr{D}_n$ is*

$$\mathscr{L}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2. \tag{2.8}$$

Now there is another issue: If $\mathbb{P}_{\boldsymbol{x}}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, then $\mathbb{P}(\boldsymbol{X}_i \neq \boldsymbol{X}_j) = 1$ almost surely for $i \neq j$. Then, each $\boldsymbol{x}_i$ must be distinct, and then we can in fact find a class of polynomials $\mathscr{P}$ such that each $p(x) \in \mathscr{P}$ will passes through all $\boldsymbol{x}_i$ exactly, then we can minimize the term to 0 exactly. Then we might have some "bumpy" polynomials that jumps back and forth and hence it is not a good represent of our data. Remember our goal is to find such a f such that, if we have a new input $\boldsymbol{x}_{n+1}$, what would be a good fit for its y_{n+1}, based on the model we designed. So by using those "bumpy" polynomials is absolutely not a good choice. This phenomenon is known as **overfitting**. To avoid the problem of interpolating functions, we can then define a subclass $\mathscr{C} \subseteq \mathscr{M}$ and indeed try to find a minimizing f in the class $\mathscr{C}$. This class can depend on n. As we can see, if $\mathscr{C}_n$ gets closer and closer to $\mathscr{M}$, we will get more estimation errors because of the problem of overfitting I discussed; also if $\mathscr{C}_n$ is too small, then although we might be able to find f under this class easily, but it might be a bad approximation and not we want. The following theorem tells us we can indeed bound the MSE by the “estimation error” and the “approximation error":

Theorem

Theorem 10. *Let $\mathscr{C}_n$ be a class of functions depending on $\mathscr{D}_n$, if $\widehat{m}_n := \widehat{m}(\cdot,\mathscr{D}_n)$ satisfies*

$$\widehat{m}(\cdot,\mathscr{D}_n) := \arg\min_{f\in\mathscr{C}} \frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i,\mathscr{D}_n)\}^2, \tag{2.9}$$

then

$$\begin{aligned}\mathbb{E}\left[\{\widehat{m}_n(\boldsymbol{x}) - m(\boldsymbol{x})\}^2 \middle| \mathscr{D}_n\right] \leq & 2\sup_{f\in\mathscr{C}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\boldsymbol{x}_i)\}^2 - \mathbb{E}\left[\{y - f(\boldsymbol{x})\}^2\right]\right| \\ & + \inf_{f\in\mathscr{C}_n}\mathbb{E}\left[\{f(\boldsymbol{x}) - m(\boldsymbol{x})\}^2\right].\end{aligned} \tag{2.10}$$

Proof. By a similar argument , we can show that

$$\mathbb{E}\left[\{\widehat{m}_n(\boldsymbol{x}) - m(\boldsymbol{x})\}^2 \middle| \mathscr{D}_n\right] = \mathbb{E}\left[(y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \mathbb{E}\left[(y - m(\boldsymbol{x}))^2\right] \tag{2.11}$$

$$\begin{aligned}&= \mathbb{E}\left[(y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \inf_{f\in\mathscr{C}_n}\mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right] \\ &+ \inf_{f\in\mathscr{C}_n}\left\{\mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right] - \mathbb{E}\left[(y - m(\boldsymbol{x}))^2\right]\right\}\end{aligned} \tag{2.12}$$

$$\begin{aligned}&= \mathbb{E}\left[(y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \inf_{f\in\mathscr{C}_n}\mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right] \\ &+ \inf_{f\in\mathscr{C}_n}\mathbb{E}\left[(f(\boldsymbol{x}) - m(\boldsymbol{x}))^2\right]\end{aligned} \tag{2.13}$$

Note we already obtained the last term in (2.10), now we continue to bound the rest:

$$\mathbb{E}\left[(y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \inf_{f\in\mathscr{C}_n}\mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right]$$

$$= \sup_{f\in\mathscr{C}_n}\left\{\mathbb{E}\left[(y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right]\right\} \tag{2.14}$$

$$\begin{aligned}&= \sup_{f\in\mathscr{C}_n}\Bigg(\mathbb{E}\left[y - \widehat{m}_n(\boldsymbol{x}))^2 \middle| \mathscr{D}_n\right] - \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}_n(\boldsymbol{x}_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}_n(\boldsymbol{x}_i))^2 \\ &- \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 - \mathbb{E}\left[(y - f(\boldsymbol{x}))^2\right]\Bigg)\end{aligned} \tag{2.15}$$

Note that by definition $\widehat{m}_n := \arg\min_{f\in\mathscr{C}} f$ so it follows that

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}_n(\boldsymbol{x}_i))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 \leq 0 \tag{2.16}$$

and denote the right hand side of (2.15) by I, we now have

$$\begin{aligned}
I &\leq \sup_{f\in\mathscr{C}_n}\left|\mathbb{E}\left[(y-\widehat{m}_n(\boldsymbol{x}))^2\middle|\mathscr{D}_n\right]-\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{m}_n(\boldsymbol{x}_i))^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i))^2-\mathbb{E}\left[(y-f(\boldsymbol{x}))^2\right]\right| \\
&\leq \left|\mathbb{E}\left[(y-\widehat{m}_n(\boldsymbol{x}))^2\middle|\mathscr{D}_n\right]-\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{m}_n(\boldsymbol{x}_i))^2\right|+\sup_{f\in\mathscr{C}_n}\left|\frac{1}{n}\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i))^2-\mathbb{E}\left[(y-f(\boldsymbol{x}))^2\right]\right| \\
&\leq 2\sup_{f\in\mathscr{C}_n}\left|\frac{1}{n}\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i))^2-\mathbb{E}\left[(y-f(\boldsymbol{x}))^2\right]\right|. \qquad (2.17)
\end{aligned}$$

■

2.3 Linear Projection Coefficient

We may wonder which class $\mathscr{C}_n$ to choose from? If $\mathscr{C}_n$ is too large, overfitting may occur and the estimation error will likely increase; however if $\mathscr{C}_n$ is too small the estimation error may decrease but the approximation error is likely to increase. We will now focus on the class $\mathbb{L}$ of linear functions:

$$\mathbb{L}:\left\{f:\mathbb{R}^p\to\mathbb{R},\forall\boldsymbol{x}\in\mathbb{R}^p,\exists\boldsymbol{\beta}\in\mathbb{R}^p,f(\boldsymbol{x})=\boldsymbol{x}^\top\boldsymbol{\beta}\right\} \qquad (2.18)$$

Definition

Definition 3. *The best linear predictor of y given $\boldsymbol{x}$ is, if it exists, the function I^* satisfying*

$$I^*=\arg\min_{f\in\mathbb{L}}\mathbb{E}\left[\{y-f(\boldsymbol{x})\}^2\right]. \qquad (2.19)$$

Since $I^*\in\mathbb{L}$, there exists $\boldsymbol{\beta}\in\mathbb{R}^p$ such that $I^*(\boldsymbol{x})=\boldsymbol{x}^\top\boldsymbol{\beta}$, the vector $\boldsymbol{\beta}$ is called the **linear projection coefficient**. Note that if we have a function $f(x)=\beta_0+\beta_1x+\beta_xx_2+\beta_3x_2^2$, then it is not linear if the covariates are chosen as $\boldsymbol{x}=(x_1,x_2)^\top$. But it would be linear if we choose the covariates to be $\boldsymbol{x}=(x_1,x_2,x_2^2)^\top$. So it is important to specify our covariates in advance, and we do not change them during the study.

Theorem

Theorem 11. *Under regularity conditions:*

(H1) $\mathbb{E}[y^2]<\infty$; (H2) $\mathbb{E}[||\boldsymbol{x}||_2^2]<\infty$; (H3) $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ is positive definite.

Then the best linear predictor of y given $\boldsymbol{x}$, denoted I^* exists and is unique, and is fully determined by the unique linear projection coefficient

$$\boldsymbol{\beta}:=\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right)^{-1}\mathbb{E}[\boldsymbol{x}y]. \qquad (2.20)$$

We first provide some technical lemma in convex optimization. We first say a set $\mathscr{C}\subseteq\mathbb{R}^p$ is **convex**, if $\forall\boldsymbol{x},\boldsymbol{y}\in\mathscr{C},t\in[0,1]$, we have $t\boldsymbol{x}+(1-t)\boldsymbol{y}\in\mathscr{C}$.

Definition

Definition 4. *A function $f : \mathbb{R}^p \to \mathbb{R}^p$ defined on a convex domain is convex, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{C}, t \in [0,1]$,*

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}). \tag{2.21}$$

If strict inequality holds then we say f is strictly convex, f is strictly concave if $-f$ is strictly convex.

The theory we will use is that: if f is convex, then any local minimizer is also a global minimizer, if f is strictly convex, then the minimizer (if exists) is unique. Its gradient can be found by Taylor expansion:

$$f(\boldsymbol{u}+\boldsymbol{h}) = f(\boldsymbol{u}) + \langle \boldsymbol{h}, \nabla f(\boldsymbol{u}) \rangle_2 + o(||\boldsymbol{h}||_2). \tag{2.22}$$

Hence we can easily find the gradient in this way, and solve for $\nabla f(\boldsymbol{u}) = 0$. The sketch of the proof will be: First show $F : \mathbb{R}^p \to \mathbb{R}, \boldsymbol{u} \to \mathbb{E}\left[(y - \boldsymbol{x}^\top \boldsymbol{u})^2\right]$ is strictly convex and finite, then solve $\nabla F(\boldsymbol{u}) = 0$ and in this case $\boldsymbol{u}$ will be the $\arg\min$ by the theory we just developed.

We first show $\boldsymbol{\beta}$ in the above theorem is indeed the minimizer.

Proof. Note that we have

$$F(\boldsymbol{u}) = \underbrace{\mathbb{E}[y^2]}_{F_1(\boldsymbol{u})} \underbrace{-2\boldsymbol{u}\mathbb{E}[y\boldsymbol{x}^\top]}_{F_2(\boldsymbol{u})} + \underbrace{\boldsymbol{u}^\top \boldsymbol{u}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]}_{F_3(\boldsymbol{u})} \tag{2.23}$$

We later will show F is indeed finite and convex, but now it suffices to find the gradient for each term first and compute the minimizer. In F_1, since it is independent of $\boldsymbol{u}$ so the answer is simply zero. In f_2, we have

$$\begin{aligned} F_2(\boldsymbol{u}+\boldsymbol{h}) &= -2\mathbb{E}[y\boldsymbol{x}^\top](\boldsymbol{u}+\boldsymbol{h}) && (2.24)\\ &= -2\mathbb{E}[y\boldsymbol{x}^\top]\boldsymbol{u} - 2\mathbb{E}[y\boldsymbol{x}^\top]\boldsymbol{h} && (2.25)\\ &= F_2(\boldsymbol{u}) + \boldsymbol{h}^\top(-2\mathbb{E}[y\boldsymbol{x}]) && (2.26) \end{aligned}$$

where $\nabla F_2(\boldsymbol{u}) = -2\mathbb{E}[\boldsymbol{x}y]$, and similarly

$$\begin{aligned} F_3(\boldsymbol{u}+\boldsymbol{h}) &= (\boldsymbol{u}+\boldsymbol{h})^\top(\boldsymbol{u}+\boldsymbol{h})\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] && (2.27)\\ &= (\boldsymbol{u}^\top\boldsymbol{u} + 2\boldsymbol{u}^\top\boldsymbol{h} + \boldsymbol{h}^\top\boldsymbol{h})\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] && (2.28)\\ &= \boldsymbol{u}^\top\boldsymbol{u}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] + 2\boldsymbol{u}^\top\boldsymbol{h}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] + o(||\boldsymbol{h}||_2) && (2.29) \end{aligned}$$

and hence $\nabla F_3(\boldsymbol{u}) = 2\boldsymbol{u}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, and finally we solve $\nabla F = 0$, that is,

$$-2\mathbb{E}[\boldsymbol{x}y] + 2\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{u} = 0. \tag{2.30}$$

By assumption $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ is positive definite, so it is invertible and hence we have

$$\boldsymbol{u} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1} \cdot \mathbb{E}[\boldsymbol{x}y]. \tag{2.31}$$

■

We then show F is strictly convex and finite:

Proof. Note that, the function

$$F(\boldsymbol{u}) = \mathbb{E}\left[(y - \boldsymbol{x}^\top \boldsymbol{u})^2\right] \tag{2.32}$$

can be written as

$$F(\boldsymbol{u}) = \underbrace{\mathbb{E}[y^2]}_{F_1(\boldsymbol{u})} \underbrace{-2\boldsymbol{u}\mathbb{E}[y\boldsymbol{x}^\top]}_{F_2(\boldsymbol{u})} + \underbrace{\boldsymbol{u}^\top \boldsymbol{u}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]}_{F_3(\boldsymbol{u})} \tag{2.33}$$

where by assumption the first term is already finite. In the second term, note that by Jensen's inequality we have

$$||\mathbb{E}[\boldsymbol{x}y]||_2 \leq \mathbb{E}[||\boldsymbol{x}y||_2] = \mathbb{E}[|y| \cdot ||\boldsymbol{x}||_2] \tag{2.34}$$

and by Cauchy-Schwarz inequality, where the inner product is defined by $\langle Z_1, Z_2 \rangle = \mathbb{E}[Z_1 Z_2]$, we have

$$\mathbb{E}[|y| \cdot ||\boldsymbol{x}||_2] \leq \mathbb{E}[|y|^2]^{1/2} \cdot \mathbb{E}[||\boldsymbol{x}||_2^2]^{1/2} \tag{2.35}$$

which is finite by assumptions. Finally for the last term we again use Jensen's inequality,

$$\boldsymbol{u}^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{u} \leq ||\boldsymbol{u}||_2^2 \cdot ||\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]||_2 \leq \mathbb{E}[||\boldsymbol{x}\boldsymbol{x}^\top||_2] = \mathbb{E}[||\boldsymbol{x}||_2^2] < \infty \tag{2.36}$$

by assumption. Finally, $\forall \theta \in (0,1)$, we have

$$\begin{aligned}
F(\theta \boldsymbol{u} + (1-\theta)\boldsymbol{v}) &= \mathbb{E}\left[(y - \theta \boldsymbol{x}^\top \boldsymbol{u} - (1-\theta)\boldsymbol{x}^\top \boldsymbol{v})^2\right] && (2.37) \\
&= \mathbb{E}\left[(\theta y + (1-\theta)y - \theta \boldsymbol{x}^\top \boldsymbol{u} - (1-\theta)\boldsymbol{x}^\top \boldsymbol{v})^2\right] && (2.38) \\
&= \mathbb{E}\left[\Big(\theta(y - \boldsymbol{x}^\top \boldsymbol{u}) + (1-\theta)(y - \boldsymbol{x}^\top \boldsymbol{v})\Big)^2\right] && (2.39) \\
&\leq \mathbb{E}\left[\theta(y - \boldsymbol{x}^\top \boldsymbol{u})^2 + (1-\theta)(y - \boldsymbol{x}^\top \boldsymbol{v})^2\right] && (2.40) \\
&= \theta \mathbb{E}\left[(y - \boldsymbol{x}^\top \boldsymbol{u})^2\right] + (1-\theta)\mathbb{E}\left[(-\boldsymbol{x}^\top \boldsymbol{v})^2\right]. && (2.41)
\end{aligned}$$

■

Remarks

1. We assumed that $\boldsymbol{x} \in \mathbb{R}^p$ so in this case $\boldsymbol{x}^\top \boldsymbol{x} \in \mathbb{R}$ is a scalar and $\boldsymbol{x}\boldsymbol{x}^\top$ is a $p \times p$ matrix. The matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ is also known as the random matrix defined by

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] = \begin{pmatrix} \mathbb{E}[X_{11}] & \cdots & \mathbb{E}[X_{1n}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_{n1}] & \cdots & \mathbb{E}[X_{nn}] \end{pmatrix} \tag{2.42}$$

2. *Jensen's inequality:* Let $f(x)$ be convex, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(x)]$, and if f is concave we have $\geq$ sign. When f is strictly convex/concave, we drop the equal sign.

3. *Cauchy-Schwarz inequality:* Let $\langle \cdot, \cdot \rangle$ be an inner product defined on the inner product space V. Then $\forall x, y \in V$, $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$.

Proposition

Proposition 3. *With the same assumption in theorem 11, the linear projection coefficient*

$$\boldsymbol{\beta} = \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\right)^{-1} \cdot \mathbb{E}[\boldsymbol{x}y] \tag{2.43}$$

also solves

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{u}\in\mathbb{R}^p} \mathbb{E}\left[\left(m(\boldsymbol{x}) - \boldsymbol{x}^\top\boldsymbol{u}\right)^2\right] \tag{2.44}$$

where $m(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$.

Proof. The proof is straightforward. We need to know in fact $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$, then

$$\begin{aligned}
\mathbb{E}\left[\left(m(\boldsymbol{x}) - \boldsymbol{x}^\top\boldsymbol{u}\right)^2\right] &= \mathbb{E}\left[m(\boldsymbol{x})^2 - 2m(\boldsymbol{x})\boldsymbol{x}^\top\boldsymbol{u} + (\boldsymbol{x}^\top\boldsymbol{u})^2\right] && (2.45)\\
&= \mathbb{E}\left[\mathbb{E}^2[y|\boldsymbol{x}] - 2\mathbb{E}[y|\boldsymbol{x}]\boldsymbol{x}^\top\boldsymbol{u} + (\boldsymbol{x}^\top\boldsymbol{u})^2\right] && (2.46)\\
&= \mathbb{E}[\mathbb{E}^2[y|\boldsymbol{x}]] - 2\mathbb{E}[\mathbb{E}[y|\boldsymbol{x}]\boldsymbol{x}^\top\boldsymbol{u}] + \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{u})^2] && (2.47)\\
&= \mathbb{E}[y^2] - \mathbb{E}[\mathbb{V}(y|\boldsymbol{x})] - 2\mathbb{E}[y\boldsymbol{x}^\top\boldsymbol{u}] + \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{u})^2] && (2.48)\\
&= \mathbb{E}\left[(y - \boldsymbol{x}^\top\boldsymbol{u})^2\right] - \mathbb{E}[\mathbb{V}(y|\boldsymbol{x})] && (2.49)
\end{aligned}$$

where the last term is independent of $\boldsymbol{u}$. ■

With our linear coefficient $\boldsymbol{x}^\top\boldsymbol{\beta}$, we recall previously we defined

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n \tag{2.50}$$

we now can write

$$y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + e_i, \qquad i = 1, 2, \cdots, n \tag{2.51}$$

where $e_i = y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}$, and it is denoted as the "noise term". *(Note that our notation is kinda unclear,* $\boldsymbol{x}_i$ *is the ith sample, not covariates, since* $i = [n]$ *and* $j = [p]$ *for covariates.)*

Proposition

Proposition 4. *For* $i = 1, 2, \cdots, n$, $\mathbb{E}[\boldsymbol{x}_i e_i] = \mathbf{0} \in \mathbb{R}^p$.

Proof. By direct computation, we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_i e_i] &= \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})] && (2.52)\\
&= \mathbb{E}[\boldsymbol{x}_i\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\beta}] && (2.53)\\
&= \mathbb{E}[\boldsymbol{x}_i y_i] - \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top]\boldsymbol{\beta} && (2.54)\\
&= \mathbb{E}[\boldsymbol{x}_i y_i] - \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top]\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}\mathbb{E}[\boldsymbol{x}_i y_i] && (2.55)\\
&= \mathbb{E}[\boldsymbol{x}_i y_i] - \mathbf{1}_p\mathbb{E}[\boldsymbol{x}_i y_i] && (2.56)\\
&= \mathbf{0}_p. && (2.57)
\end{aligned}$$

■

Definition

Definition 5. *An intercept is any constant covariate $\boldsymbol{x}_j$, i.e, there exists $C \in \mathbb{R}$ such that $\boldsymbol{x}_j = C$ almost surely.*

From the previous proposition, we can see that

$$\mathbb{E}[\boldsymbol{x}_i e_i] = \begin{pmatrix} \mathbb{E}[\boldsymbol{x}_{i1} e_i] \\ \mathbb{E}[\boldsymbol{x}_{i2} e_i] \\ \vdots \\ \mathbb{E}[\boldsymbol{x}_{1p} e_i] \end{pmatrix} = \mathbf{0}_p \tag{2.58}$$

while if $\boldsymbol{x}_{ij} = C$ for some covariates, then we can see $\mathbb{E}[e_i] = 0$ must be true. Usually, with out loss of generality the intercept is usually taken to be equal to 1 in the first position. Now consider another decomposition:

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \underbrace{(m(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{\beta})}_{\eta(x)} + \underbrace{(y - m(\boldsymbol{x}))}_{\varepsilon} \tag{2.59}$$

where $m(\boldsymbol{x}) := \mathbb{E}[y|\boldsymbol{x}]$, which takes the form $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \eta(\boldsymbol{x}) + \varepsilon$, we then study some properties of the form (2.59) with intercept introduced:

Proposition

Proposition 5. *(Weak Orthogonality) Let $x_j \in [p]$ be covariates, then $\mathbb{E}[\eta(\boldsymbol{x})x_j] = 0$ for all $j = [p]$.*

Proof. Previously we have shown that $\mathbb{E}[\boldsymbol{x}(y - \boldsymbol{x}^\top \boldsymbol{\beta})] = \mathbf{0}_p$, so it must also hold for x_j, so $\mathbb{E}[x_j(y - \boldsymbol{x}^\top \boldsymbol{\beta})] = 0$. Note that $y - \boldsymbol{x}^\top \boldsymbol{\beta} = \eta(x) + \varepsilon$, so we have

$$\mathbb{E}[x_j(y - \boldsymbol{x}^\top \boldsymbol{\beta})] = \mathbb{E}[x_j \eta(\boldsymbol{x})] + \mathbb{E}[x_j \varepsilon] = 0, \tag{2.60}$$

where

$$\mathbb{E}[x_j \varepsilon] = \mathbb{E}[\mathbb{E}[x_j \varepsilon | \boldsymbol{x}]] = \mathbb{E}[x_j \mathbb{E}[\varepsilon | \boldsymbol{x}]] = \mathbb{E}[x_j \cdot 0] = 0, \tag{2.61}$$

which implies $\mathbb{E}[\eta(\boldsymbol{x})x_j] = 0$ for all $j = [p]$. ■

Proposition

Proposition 6. *(Strong Orthogonality) For all $f \in L^2(\mathbb{P}_{\boldsymbol{x}})$, $\mathbb{E}[f(\boldsymbol{x})\varepsilon] = 0$.*

Proof. From (2.59), we see that $\varepsilon = y - \boldsymbol{x}^\top \boldsymbol{\beta} - m(\boldsymbol{x}) + \boldsymbol{x}^\top \boldsymbol{\beta} = y - m(\boldsymbol{x})$, so we have

$$\mathbb{E}[f(\boldsymbol{x})\varepsilon] = \mathbb{E}[f(\boldsymbol{x})(y - m(\boldsymbol{x}))], \tag{2.62}$$

also note that

$$\mathbb{E}[f(\boldsymbol{x})(y - m(\boldsymbol{x}))] = \mathbb{E}[\mathbb{E}[f(x)(y - m(\boldsymbol{x}))|\boldsymbol{x}]] = \mathbb{E}[f(x)(\mathbb{E}[y|\boldsymbol{x}] - m(\boldsymbol{x}))] \equiv 0 \tag{2.63}$$

where $m(\boldsymbol{x}) := \mathbb{E}[y|\boldsymbol{x}]$. ■

Proposition

Proposition 7. *Suppose intercept is included among the covariates, then in the decomposition*

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \underbrace{(m(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{\beta})}_{\eta(x)} + \underbrace{(y - m(\boldsymbol{x}))}_{\varepsilon}, \tag{2.64}$$

we have $\mathbb{E}[\eta(\boldsymbol{x})] = 0$ *and* $\mathbb{E}[\varepsilon\eta(\boldsymbol{x})] = 0$.

Proof. As we have shown before, if intercept is included then $\mathbb{E}[e] = 0$ when $y = \boldsymbol{x}^\top \boldsymbol{\beta} + e$, in this case we have $e = m(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{\beta} + y - m(\boldsymbol{x}) = y - \boldsymbol{x}^\top \boldsymbol{\beta}$ so $\mathbb{E}[y - \boldsymbol{x}^\top \boldsymbol{\beta}] = 0$, then by a conditional argument we have

$$\mathbb{E}[y - \boldsymbol{x}^\top \boldsymbol{\beta}] = \mathbb{E}[\mathbb{E}[y|\boldsymbol{x}] - \boldsymbol{x}^\top \boldsymbol{\beta}] := \mathbb{E}[\eta(\boldsymbol{x})] = 0. \quad (m(\boldsymbol{x}) := \mathbb{E}[y|\boldsymbol{x}]) \tag{2.65}$$

As for the second statement, we have

$$\mathbb{E}[\varepsilon\eta(\boldsymbol{x})] = \mathbb{E}[\mathbb{E}[\varepsilon\eta(\boldsymbol{x})|\boldsymbol{x}]] = \mathbb{E}[\eta(\boldsymbol{x}) \cdot \mathbb{E}[\varepsilon|\boldsymbol{x}]] = \mathbb{E}[0 \cdot 0] \equiv 0. \tag{2.66}$$

■

Remarks

We may think of expectation as an inner product. Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we define $L^2 := L^2(\Omega, \mathscr{F}, \mathbb{P})$ as the space of square-integrable random variables with

$$\mathbb{E}[X^2] = \int_\Omega X(\omega)^2 \mathbb{P}(d\omega) < \infty. \tag{2.67}$$

We claim that L^2 defines a Hilbert space (a complete inner product space). Then the inner product is defined by $\langle X, Y \rangle := \mathbb{E}[XY]$ and $\langle X, Y \rangle = 0$ if and only if X, Y are independent (or called orthogonal). The norm equipped on L^2 is simply defined by $||X||_{L^2} := \mathbb{E}[X^2]^{1/2}$.

In a Hilbert space F, let $E \subseteq F$ closed, then the projection theorem says that $\forall x \in F$, there is a unique element $P_E(x) \in E$ such that

$$||x - P_E(x)|| = \inf_{y \in E} ||x - y|| \tag{2.68}$$

if $h \in E$, $\langle x - h, y \rangle = 0$ for every $y \in E$, then $h = P_E(x)$, denote by the orthogonal projection of x onto E. Based on that, we know every $x \in F$ has a unique decomposition $x = P_E(x) + P_{E^\perp}(x)$ where $P_E(x) \in E, P_{E^\perp}(x) \in E^\perp$. The mapping $\mathscr{P} : x \mapsto P_E(x)$ is called the orthogonal projection operator.

Then we can use projection argument to prove the above propositions easily. Note that to show $\mathbb{E}[\boldsymbol{x}\varepsilon] = 0$, recall our definition that $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$, it can also be viewed as a projection:

$$y = \mathscr{P}_{\mathscr{L}}(y) + \mathscr{P}_{\mathscr{L}^\perp}(y) \tag{2.69}$$

where $\mathscr{L}$ denotes the space of linear functions, and we can see that $\boldsymbol{x}^\top\boldsymbol{\beta} \in \mathscr{L}, \boldsymbol{\varepsilon} \in \mathscr{L}^\perp$ and we have

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{\varepsilon}] = \langle \boldsymbol{x}, \boldsymbol{\varepsilon} \rangle = \mathbf{0}_p. \tag{2.70}$$

And similarly for $\mathbb{E}[f(\boldsymbol{x})\boldsymbol{\varepsilon}]$ we have

$$y = \mathscr{P}_{\mathscr{G}}(y) + \mathscr{P}_{\mathscr{G}^\perp}(y) \tag{2.71}$$

where $\mathscr{G}$ denotes the space of functions $f \in L^2(\mathbb{P}_{\boldsymbol{x}})$ and hence $\mathbb{E}[f(\boldsymbol{x})\varepsilon] = 0$.

2.4 Design Matrix and Linear Models

Definition

Definition 6. *Given data $\mathscr{D}_n := \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, we say the model is homoscedastic if*

$$\mathbb{V}(\varepsilon|\boldsymbol{x}) = \sigma^2). \tag{2.72}$$

If it is not a constant, the model is said to be heteroscedastic.

Further, we may present the data $(\boldsymbol{x}_i, y_i)$ in a matrix form:

Definition

Definition 7. *The covariates are stored in the design matrix $\boldsymbol{X} \in \mathscr{M}_{n,p}$ where $\boldsymbol{X}_{ij}$ denotes the value of covariate j for observation $i \in [n]$, we have*

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,2} & \cdots & X_{1,p} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & X_{n,2} & \cdots & X_{n,p} \end{pmatrix} \tag{2.73}$$

where we assume intercept is included, and we put it as the first covariate. The response is represented by a vector $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$.

The design matrix $\boldsymbol{X}$ is said to be **full rank** if $\mathbb{P}[\text{rank}(\boldsymbol{X}) = p] = 1$. The covariates are said to be **multicollinear** if it is not full rank, and in this case we may discard some of the covariates since they are considered as "redundant" information". The good thing about full rank is that, if $\boldsymbol{X}$ is full rank, then $\boldsymbol{X}^\top\boldsymbol{X}$ will be invertible, which can make our life easier later. In the setting of a random design, $\boldsymbol{X}$ is random; while in a fixed design $\boldsymbol{X}$ is deterministic, and the intercept column corresponds to a degenerate distribution at 1.

Definition

Definition 8. *We denote $\mathscr{D}_n = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ to be the i.i.d data set we get, we say the data follow a linear model if $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x}^\top\boldsymbol{\beta}$, i.e, $m(\boldsymbol{x}) \in \mathbb{L}$. The model is homoscedastic if $\mathbb{V}(y|\boldsymbol{x}) = \mathbb{V}(\varepsilon|\boldsymbol{x}) = \sigma^2$ a constant, and the model is heteroscedastic if it is not a constant.*

Proposition

Proposition 8. *If $\mathbb{P}_x$ has a continuous distribution with respect to the Lebesgue measure on $\mathbb{R}^p$, then the design matrix $\mathbf{X}$ is full rank.*

Proof. ■

We now introduce two different regression models, the **jointly Gaussian linear model** and **normal linear model**:

Jointly Gaussian linear model: This model is often viewed as the main motivation for the full-rank, homoscedastic linear model. We assume $(\boldsymbol{x}, y)^\top$ is jointly Gaussian in the linear model

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon \tag{2.74}$$

where we have

$$\mathbf{Cov}(\boldsymbol{x}, \varepsilon) = \mathbb{E}[\boldsymbol{x}\varepsilon] - \mathbb{E}[\boldsymbol{x}] \cdot \mathbb{E}[\varepsilon] = \mathbf{0}_p. \tag{2.75}$$

This is because by definition $\mathbb{E}[\varepsilon] = \mathbb{E}[\mathbb{E}[\varepsilon|\boldsymbol{x}]] = \mathbb{E}[0] = 0$ and furthermore

$$\mathbb{E}[\boldsymbol{x}\varepsilon] = \mathbb{E}[\mathbb{E}[\boldsymbol{x}\varepsilon|\boldsymbol{x}]] = \mathbb{E}[\boldsymbol{x}\mathbb{E}[\varepsilon|\boldsymbol{x}]] = \mathbb{E}[\boldsymbol{x} \cdot 0] = 0. \tag{2.76}$$

In a jointly Gaussian model, once we have the covariance is zero, we may conclude that $\boldsymbol{x}, \varepsilon$ are independent, and since Gussianity is preserved under linear transformations, it follows that $(\boldsymbol{x}, \varepsilon)^\top$ is also a jointly Gaussian, and by independence, $\mathbb{V}(\varepsilon|\boldsymbol{x}) = \mathbb{V}(\varepsilon)$ which means the model is homoscedastic.

Normal Linear Model: This is a relaxation of the jointly Gaussian model, it assumes that the residuals are Gaussian, and the covariates are independent of the residuals. It is less restrictive then the jointly Gaussian model because here the covariates may follow any distribution, and normal linear model is therefore still a particular case of a homoscedastic linear model.

In the case of a linear model, one of the main focuses is to make inference about the unknown vector of coefficients $\boldsymbol{\beta}$.

Definition

Definition 9. *A statistical model is a class of distributions $\{\mathbb{P}_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ that we believe might have generated the data $\mathscr{D}_n$, the set Θ is called the parameter space and its elements are the parameters of the model.*

Parameters can often be partitioned as $\boldsymbol{\theta} = [\boldsymbol{\gamma}^\top, \boldsymbol{\eta}^\top]^\top$ where our primarily interest is called **parameter of interest** $\boldsymbol{\gamma}$ and the remaining component $\boldsymbol{\eta}$ is not of direct interest, but is still required to fully specify the distribution and it called a **nuisance parameter**. There are three models:

(i) If the parameter space is finite dimensional, the model is called **parametric**;

(ii) If the parameter of interest lies in a finite dimensional space, but there exists a nuisance parameter described by an infinite dimensional space, the model is called **semi-parametric**;

(iii) If the parameter of interest cannot be indexed by a finite dimensional space, the model is called **non-parametric**.

Proposition

Proposition 9. *A random design linear model is semi-parametric; A fixed design linear model with Gaussian residuals is a parametric model.*

2.5 Closing Remark

Linear regression analysis may have several objectives, for example:(i) Estimate the regression function $m(\boldsymbol{x})$; (ii) Make predictions on new/unobserved points, especially in random design;(iii) Recover the true signal $m(\boldsymbol{x}_i)$ from noisy observations $y_i = m(\boldsymbol{x}_i) + \varepsilon_i$, and this is sometimes called denoising;(iv) Understand how a covariate $\boldsymbol{X}_j$ affects the response y.

However, there are limitations of traditional linear models: (i) The regression function may not be linear at all; (ii) The data may exhibit dependence; (iii) The data may not be identically distributed; (iv) The conditional variance may not be constant; (v) The data may be high-dimensional where $p >> n$.

That motivates us to continue our study on regression.

The Ordinary Least Squares Estimator

3.1 Derivations of the OLS Estimator

Previously, we have shown that if we assume a linear model, then the regression function can be written as

$$m(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} = \sum_{j=1}^{p} x_j \beta_j \tag{3.1}$$

and the problem of estimating the regression function reduces to the problem of estimating $\boldsymbol{\beta}$, given by $\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y]$. However in reality we do not have access to the true expectation

$\mathbb{E}[\cdot]$, what we normally do is to replace $\mathbb{E}[\cdot]$ by its empirical estimate

$$\widehat{\mathbb{E}} = \frac{1}{n}\sum_{i\in[n]}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2. \tag{3.2}$$

Definition

Definition 10. *The ordinary least squares loss over $\mathscr{D}_n$ is the function $\mathscr{O}_{LS}(\cdot, \mathscr{D}_n) := \mathscr{O}_{LS,n}$ from $\mathbb{R}^p$ to $\mathbb{R}$ defined by*

$$\mathscr{O}_{LS,n}(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i\in[n]}\{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\}^2 := \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2. \tag{3.3}$$

An ordinary least squares estimators of $\boldsymbol{\beta}$ is any vector $\widehat{\boldsymbol{\beta}}_{ols}(\mathscr{D}_n) := \widehat{\boldsymbol{\beta}}_{ols,n}$ satisfying

$$\widehat{\boldsymbol{\beta}}_{ols,n} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathscr{O}_{LS,n}(\boldsymbol{\beta}). \tag{3.4}$$

Note that sometimes the residual sum of squares (RSS) is considered to be the loss of reference, where

$$RSS(\boldsymbol{\beta}) = \sum_{i\in[n]}\{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\}^2 = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 = n\cdot\mathscr{O}_{LS,n}. \tag{3.5}$$

but they are considered equivalent since are up to some constants.

Theorem

Theorem 12. *The ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ satisfies the normal equations*

$$\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}}_n := \boldsymbol{X}^\top\boldsymbol{y}. \tag{3.6}$$

Additionally, if the design matrix is full rank, then the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_{ols,n} := \widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ exists, and it is unique, given by

$$\widehat{\boldsymbol{\beta}}_n := \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \left(\sum_{i\in[n]}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\cdot\sum_{i\in[n]}\boldsymbol{x}_i y_i. \tag{3.7}$$

Proof. The proof is similar to **theorem 11**, where we will use the fact that the function $\mathscr{O}_{LS,n}(\boldsymbol{\beta})$ is strictly convex, and it hence suffices to compute the gradient and equate to zero to solve for $\boldsymbol{\beta}$.

Let

$$F(\boldsymbol{u}) = \frac{1}{n}\sum_{i\in[n]}\{y_i - \boldsymbol{x}_i^\top \boldsymbol{u}\}^2 \tag{3.8}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\boldsymbol{x}_i^\top \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u}\right) \tag{3.9}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i^2}_{F_1(\boldsymbol{u})} \underbrace{-\frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i^\top \boldsymbol{u}}_{F_2(\boldsymbol{u})} + \underbrace{\frac{1}{n}\boldsymbol{u}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u}}_{F_3(\boldsymbol{u})} \tag{3.10}$$

where we compute the gradient separately. Note that $F_1(\boldsymbol{u})$ is independent of $\boldsymbol{u}$ so we simply have $\nabla F_1(\boldsymbol{u}) = 0$. Then since the function F is strictly convex, we just need to find $\boldsymbol{u}^*$ that cancels out the gradient and $\boldsymbol{u}^*$ is the unique global minimum. We have

$$F_2(\boldsymbol{u}+\boldsymbol{h}) = -\frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i^\top(\boldsymbol{u}+\boldsymbol{h}) \tag{3.11}$$

$$= -\frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i^\top \boldsymbol{u} - \boldsymbol{h}^\top \frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i \tag{3.12}$$

$$:= F_2(\boldsymbol{u}) + \boldsymbol{h}^\top \nabla F_2(\boldsymbol{u}), \tag{3.13}$$

and we have

$$\nabla F_2(\boldsymbol{u}) = -\frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i. \tag{3.14}$$

Similarly for $F_3(\boldsymbol{u})$:

$$F_3(\boldsymbol{u}+h) = (\boldsymbol{u}+\boldsymbol{h})^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{u}+\boldsymbol{h}) \tag{3.15}$$

$$= \boldsymbol{u}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u} + \boldsymbol{h}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{u}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{h} + \boldsymbol{h}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u} \tag{3.16}$$

$$= F_3(\boldsymbol{u}) + \mathcal{O}(||\boldsymbol{h}||_2) + \frac{2}{n}\boldsymbol{h}^\top \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u} \tag{3.17}$$

and again we the have

$$\nabla F_3(\boldsymbol{u}) = \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u}. \tag{3.18}$$

We finally set $\nabla F(\boldsymbol{u}) = \boldsymbol{0}_p$ and we have

$$-\frac{2}{n}\sum_{i=1}^{n} y_i\boldsymbol{x}_i + \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u} = \boldsymbol{0}_p \tag{3.19}$$

we solve for $\boldsymbol{u}$ and we get

$$\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{u} = \sum_{i=1}^{n} y_i\boldsymbol{x}_i \tag{3.20}$$

which is equivalent as $\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u}$. Further if $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, then we can solve $\boldsymbol{u}$ as

$$\boldsymbol{u} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{3.21}$$

■

Remark

We have used several properties from linear algebra: Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, then (i) $\boldsymbol{a}^\top \boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{a}$; (ii) $\boldsymbol{a}\boldsymbol{a}^\top \in M_{p\times p}(\mathbb{R})$ is symmetric; (iii) $(ABC)^\top = C^\top B^\top A^\top$ for matrices.

We may implement **theorem 12** to a simple linear model:

Proposition

Proposition 10. *Suppose we have observable data $\mathscr{D} := \{(x_1, y_1), \cdots, (x_n, y_n)\}$ ($(x_i, y_i) \in \mathbb{R}^2$) and suppose we have a regression model $y = \beta_0 + \beta_1 x + \varepsilon$ with $\beta_0, \beta_1 \in \mathbb{R}$ as parameters, then*

$$\widehat{\beta}_0 = \bar{y}_n - \bar{x}_n \widehat{\beta}_1 \qquad \widehat{\beta}_1 = \frac{\sum_{i\in[n]}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i\in[n]}(x_i - \bar{x}_n)^2} \tag{3.22}$$

Proof. Note that we have

$$\mathscr{O}_{LS}(\beta_1, \beta_2) = \sum_{i=1}^n \{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\}^2 \tag{3.23}$$

$$= \arg\min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^n \{y_i - (\beta_0 + \beta_1 x_i)\}^2 \tag{3.24}$$

$$= \arg\min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^n \{y_i^2 + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2 - 2y_i\beta_0 - 2\beta_1 y_i x_i\} \tag{3.25}$$

Further, y_i^2 is independent of β_0, β_1, and we use the fact that

$$\frac{1}{n}\sum_{i=1}^n x_i := \bar{x}_n \qquad \frac{1}{n}\sum_{i=1}^n y_i := \bar{y}_n \tag{3.26}$$

to obtain

$$\mathscr{O}_{LS}(\beta_1, \beta_2) := \arg\min_{\beta_0,\beta_1} \left\{ \beta_0^2 + \frac{1}{n}\sum_{i=1}^n x_i^2 \beta_1^2 + 2\bar{x}_n\beta_0\beta_1 - 2\bar{y}_n\beta_0 - \frac{2}{n}\sum_{i=1}^n x_i y_i \beta_1 \right\}. \tag{3.27}$$

Define a multivariate function $f(\beta_0, \beta_1)$ by

$$f(\beta_0, \beta_1) = \beta_0^2 + \frac{1}{n}\sum_{i=1}^n x_i^2 \beta_1^2 + 2\bar{x}_n\beta_0\beta_1 - 2\bar{y}_n\beta_0 - \frac{2}{n}\sum_{i=1}^n x_i y_i \beta_1 \tag{3.28}$$

and the gradient of f is given by

$$\nabla f = \left(\frac{\partial f}{\partial \beta_0}, \frac{\partial f}{\partial \beta_1}\right) = \left(2\beta_0 + 2\bar{x}_n\beta_1 - 2\bar{y}_n \ , \ \frac{2}{n}\sum_{i=1}^n x_i^2\beta_1 + 2\bar{x}_n\beta_0 - \frac{2}{n}\sum_{i=1}^n x_i y_i\right). \tag{3.29}$$

Set $\nabla f = 0$, we solve the system

$$\begin{cases} 2\beta_0 + 2\bar{x}_n\beta_1 - 2\bar{y}_n & = 0 \\ \frac{2}{n}\sum_{i=1}^n x_i^2\beta_1 + 2\bar{x}_n\beta_0 - \frac{2}{n}\sum_{i=1}^n x_i y_i & = 0 \end{cases} \tag{3.30}$$

where the first equation will directly yield

$$\widehat{\beta}_0 = \bar{y}_n - \bar{x}_n\widehat{\beta}_1. \tag{3.31}$$

Now use (3.31) to solve for $\widehat{\beta}_1$ in the second equation, we have

$$\frac{1}{n}\sum_{i=1}^n x_i^2\widehat{\beta}_1 + \bar{x}_n(\bar{y}_n - \bar{x}_n\widehat{\beta}_1) - \frac{1}{n}\sum_{i=1}^n x_i y_i = 0 \tag{3.32}$$

which gives us

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}. \tag{3.33}$$

Note that (3.33) is just a reformation of the result, since one have

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) &= \sum_{i=1}^n x_i y_i - \bar{y}_n\sum_{i=1}^n x_i - \bar{x}_n\sum_{i=1}^n y_i + n\bar{x}_n\bar{y}_n && (3.34) \\ &= \sum_{i=1}^n x_i y_i - n\bar{y}_n\bar{x}_n - n\bar{y}_n\bar{x}_n + n\bar{x}_n\bar{y}_n && (3.35) \\ &= \sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n && (3.36) \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}_n)^2 &= \sum_{i=1}^n x_i^2 + n\bar{x}_n^2 - 2\bar{x}_n\sum_{i=1}^n x_i && (3.37) \\ &= \sum_{i=1}^n x_i^2 + n\bar{x}_n^2 - 2\bar{x}_n^2 && (3.38) \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}_n^2. && (3.39) \end{aligned}$$

Since such $(\widehat{\beta}_0, \widehat{\beta}_1)$ is the unique point that cancels the gradient, followed by the fact that $\mathscr{O}_{LS}$ is strictly convex, so such $(\widehat{\beta}_0, \widehat{\beta}_1)$ is the unique minimizer, where

$$\widehat{\beta}_0 = \widehat{\beta}_0 = \bar{y}_n - \bar{x}_n\widehat{\beta}_1 \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}. \tag{3.40}$$

■

Conceptually, the proof is not hard but it requires a lot of computations. One can also use the matrix representation form but that would be more complicated since it involves inverting a matrix.

Proposition

Proposition 11. *Let the design matrix $\boldsymbol{X}$ be full rank and assume $\boldsymbol{X}$ is orthogonal, then the coefficients of $\widehat{\boldsymbol{\beta}}_n$ are equal to those obtained by fitting p independent simple linear regressions.*

We need a remark from linear algebra for this proposition:

Remark

Suppose $\boldsymbol{X}$ is an orthonormal matrix, then $\boldsymbol{X}^\top \boldsymbol{X} = \mathbb{I}_p$. If normal condition is dropped and we only have orthogonal matrix, then

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \sum_{i=1}^p X_{i1}^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sum_{i=1}^p X_{ip}^2 \end{pmatrix} \tag{3.41}$$

Corollary

Corollary 1. *(Indicator Regression) Let $\mathscr{D} := \{i \in [n] : (\boldsymbol{x}_i, y_i)\} \subseteq \mathbb{R}^p \times \mathbb{R}$ be the set of observations, let $T \in [n-1]$ be fixed, $\{A_j\}_{j \in [T]}$ be a partition of $\mathbb{R}^p$, define the map $\mathscr{M} : \mathbb{R}^p \to \mathbb{R}^T$ by*

$$\boldsymbol{x}_i \mapsto \boldsymbol{z}_i := \begin{bmatrix} \mathbb{1}_{A_1}(\boldsymbol{x}_i) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_i) \end{bmatrix}^\top \tag{3.42}$$

then the OLS estimator $\widehat{\boldsymbol{\beta}}_z := (\widehat{\beta}_{z1}, \cdots, \widehat{\beta}_{zT})^\top \in \mathbb{R}^T$ of the regression function of y on $\boldsymbol{z}$ is

$$\widehat{\beta}_{zj} := \frac{\sum_{i=1}^n \mathbb{1}_{A_j}(\boldsymbol{x}_i) y_i}{\sum_{i=1}^n \mathbb{1}_{A_j}(\boldsymbol{x}_i)}. \tag{3.43}$$

Proof. We know the normal equation is defined by $\boldsymbol{Z}^\top \boldsymbol{Z} \widehat{\boldsymbol{\beta}}_T = \boldsymbol{Z}^\top \boldsymbol{y}$ where the design matrix $\boldsymbol{Z}$ takes the form

$$\boldsymbol{Z} = \begin{pmatrix} \mathbb{1}_{A_1}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \mathbb{1}_{A_1}(\boldsymbol{x}_n) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_n) \end{pmatrix} \in M_{n \times T}(\mathbb{R}) \tag{3.44}$$

Hence the normal equation can be written as

$$\begin{pmatrix} \mathbb{1}_{A_1}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_1}(\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbb{1}_{A_T}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_n) \end{pmatrix} \cdot \begin{pmatrix} \mathbb{1}_{A_1}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \mathbb{1}_{A_1}(\boldsymbol{x}_n) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_n) \end{pmatrix} \cdot \begin{pmatrix} \widehat{\beta}_{z1} \\ \vdots \\ \widehat{\beta}_{zT} \end{pmatrix} \tag{3.45}$$

$$= \begin{pmatrix} \mathbb{1}_{A_1}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_1}(\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbb{1}_{A_T}(\boldsymbol{x}_1) & \cdots & \mathbb{1}_{A_T}(\boldsymbol{x}_n) \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{3.46}$$

which simplifies to

$$\operatorname{diag}\left(\sum_{i=1}^{n}\mathbb{1}_{A_1}(\boldsymbol{x}_i),\cdots,\sum_{i=1}^{n}\mathbb{1}_{A_T}(\boldsymbol{x}_i)\right)\begin{pmatrix}\widehat{\beta}_{z1}\\ \vdots\\ \widehat{\beta}_{zT}\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{n}\mathbb{1}_{A_1}(\boldsymbol{x}_i)y_i\\ \vdots\\ \sum_{i=1}^{n}\mathbb{1}_{A_T}(\boldsymbol{x}_i)y_i\end{pmatrix} \tag{3.47}$$

Where (if) the diagonal matrix is invertible, then the coefficient $\widehat{\beta}_{zj}$ is given by

$$\widehat{\beta}_{zj} := \frac{\sum\limits_{i=1}^{n}\mathbb{1}_{A_j}(\boldsymbol{x}_i)y_i}{\sum\limits_{i=1}^{n}\mathbb{1}_{A_j}(\boldsymbol{x}_i)}. \tag{3.48}$$

■

3.2 Hat Matrix and Annihilator Matrix

Note that although we have obtained the estimate $\widehat{\boldsymbol{\beta}}$, but our final goal is to say something about $\boldsymbol{\beta}$. However $\boldsymbol{\beta}$ is never observed! So our goal would be to find an estimator $\widehat{\boldsymbol{\beta}}$ with low variance and low bias. The figure below illustrates two different estimates: One with low variance but high bias, one with high variance but low bias:

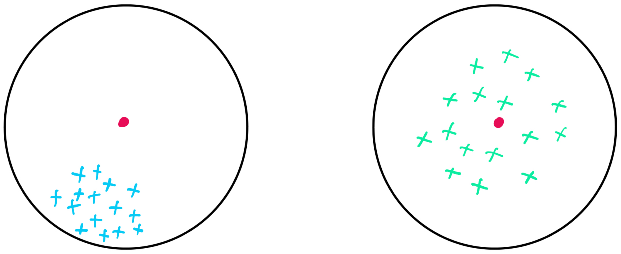


Figure 2: In the figure above, the red dot is our target parameter and crosses marked in blue and green are two estimates. The first one (left) has low variance but high bias; the second one (right) has high variance but low bias.

Once we have obtained the OLS estimator $\widehat{\boldsymbol{\beta}}_n$, the predicted value at a new given point $\boldsymbol{x} \in \mathbb{R}^p$ is now given by

$$\widehat{m}(\boldsymbol{x}) := \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}. \tag{3.49}$$

Definition

Definition 11. *When $\boldsymbol{x} = \boldsymbol{x}_i$ for some $i \in [n]$, we say $\widehat{m}(\boldsymbol{x}_i) = \widehat{y}_i$ is the fitted value of element i, and the OLS sample residual of element $i \in [n]$ is defined as $\widehat{e}_i := y_i - \widehat{y}_i$. Additionally we define the hat matrix to be $\boldsymbol{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and the annihilator matrix to be $\boldsymbol{M} := \mathbb{I} - \boldsymbol{H}$.*

Then, we see that the fitted values $\widehat{\boldsymbol{y}} := (\widehat{y}_1, \cdots, \widehat{y}_n)^\top$ can be written as $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. *That's probably why we call it a "Hat" matrix since it puts a hat on* $\boldsymbol{y}$

Proposition

Proposition 12. *The hat matrix $\boldsymbol{H}$ and the annihilator matrix $\boldsymbol{M}$ are both orthogonal projection matrices (idempotent and symmetric).*

Remark: Idempotent means we have $A^2 = A$, as for symmetric, you know that.

Proof. We prove it for $\boldsymbol{H}$ first. Note that

$$\boldsymbol{H}^\top = \left(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)^\top = (\boldsymbol{X}^\top)^\top\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)^\top\boldsymbol{X}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \tag{3.50}$$

where we used the fact that $(ABC)^\top = C^\top B^\top A^\top$, and $(A^{-1})^\top = (A^\top)^{-1}$ (given A is invertible). Then,

$$\boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}_{\text{note that this is } \mathbb{I}}\boldsymbol{X}^\top := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top. \tag{3.51}$$

Hence $\boldsymbol{H}$ is an orthogonal projection, and it follows that $\boldsymbol{M} = \mathbb{I} - \boldsymbol{H}$ is also an orthogonal projection. ■

Proposition

Proposition 13. *$\boldsymbol{H}$ projects onto $Im(\boldsymbol{X})$ and $\boldsymbol{M}$ projects onto $Ker(\boldsymbol{X}^\top)$;*

Proof. We see that $\forall \boldsymbol{y} \in \mathbb{R}^n$, denote $\widehat{\boldsymbol{\beta}}_n$ to be the OLS estimate, then by definition we have indeed

$$\boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \in Im(\boldsymbol{X}) \tag{3.52}$$

We further use the fact that $Im(\boldsymbol{X}^\perp) = Ker(\boldsymbol{X}^\top)$ to conclude $\boldsymbol{M} := \mathbb{I} - \boldsymbol{H}$ is the projection onto $Ker(\boldsymbol{X}^\top)$. Since if $\boldsymbol{X}$ projects onto S then $\mathbb{I} - \boldsymbol{X}$ projects onto $S^\perp$. ■

Proposition

Proposition 14. *$tr(\boldsymbol{H}) = p, tr(\boldsymbol{M}) = n - p$. (Recall that p is the number of covariates and n is the number of sample)*

Proof. We use the fact that $tr(AB) = tr(BA)$. So we have

$$tr(\boldsymbol{H}) = tr(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = tr(\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) = tr(\mathbb{I}_p) = p. \tag{3.53}$$

Note that $\boldsymbol{X} \in M_{n\times p}(\mathbb{R}), \boldsymbol{H} \in M_{n\times n}(\mathbb{R})$. Also we have $tr(A+B) = tr(A) + tr(B)$, so we have

$$tr(\boldsymbol{M}) = tr(\mathbb{I}_n - \boldsymbol{H}) = tr(\mathbb{I}_n) - tr(\boldsymbol{H}) = n - p. \tag{3.54}$$

■

Proposition

Proposition 15. *Suppose intercept is included in the covariates, then* $(\widehat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\top \widehat{\boldsymbol{e}} = 0$*, where*

$$\widehat{\boldsymbol{y}} = (\widehat{y}_1, \cdots, \widehat{y}_n)^\top, \bar{\boldsymbol{y}} = (\bar{y}_1, \cdots, \bar{y}_n)^\top, \widehat{\boldsymbol{e}} = (\widehat{e}_1, \cdots, \widehat{e}_n)^\top. \tag{3.55}$$

Proof. We already know that $\widehat{\boldsymbol{y}} \in Im(\boldsymbol{H})$, $\widehat{\boldsymbol{e}} \in Im(\boldsymbol{M})$ and it is clear that $Im(\boldsymbol{H})$ and $Im(\boldsymbol{M})$ are orthogonal to each other hence $\widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{e}} = 0$ and we only need to show $\bar{\boldsymbol{y}} \in Im(\boldsymbol{H})$. Since intercept is included, denote the design matrix $\boldsymbol{X}$ by

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{3.56}$$

hence $(1, \cdots, 1)^\top \in Im(\boldsymbol{X})$, thus $(1, \cdots, 1)^\top \in Im(\boldsymbol{H})$ (*recall that* $\boldsymbol{H}$ *projects onto* $\boldsymbol{X}$), and so $\bar{\boldsymbol{y}}_n \in Im(H)$ since it's just a difference by a scalar. ■

Definition

Definition 12. *The leverage of element* $i \in [n]$ *is defined*

$$h_{ii} := \boldsymbol{H}_{(ii)} := \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \tag{3.57}$$

and an extension yields

$$h_{ij} := \boldsymbol{H}_{(ij)} = \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_j \tag{3.58}$$

Leverages appear in many places in linear regression analysis, especially when p is large. We study some properties of leverage:

Proposition

Proposition 16. *The leverages are positive and bounded:* $0 \leq h_{ii} \leq 1$*;*

Proof. First of all since $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite, so $\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i \geq 0$ for all $\boldsymbol{x}_i$ hence $h_{ii} \geq 0$, also since $\boldsymbol{H}$ is an orthogonal projection, we have $\boldsymbol{H}^2 = \boldsymbol{H}$ and moreover, $\boldsymbol{H}^2_{(ii)} = \boldsymbol{H}_{ii}$, meaning we only need to show

$$\sum_{j=1}^{n} h_{ij}^2 = h_{ii} \leq 1. \tag{3.59}$$

which is,

$$h_{ii} - h_{ii}^2 = \sum_{j \in [n] \setminus \{i\}} h_{ij}^2 > 0 \tag{3.60}$$

which means $h_{ii} \in [0, 1]$. *Well I thought that's a Cauchy-Schwarz argument* ■

Proposition

Proposition 17. *If intercept is included, then denote* $\boldsymbol{x} = (1, x_{(1)}, \cdots)^\top$, *then*

$$\sum_{i\in[n]} \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i = 1 \tag{3.61}$$

Proof. Using the fact that $\boldsymbol{x}_i^\top \boldsymbol{e}_i = 1$ where $\boldsymbol{e}_i$ is the i-th unit vector, we have

$$\begin{aligned}
\sum_{i\in[n]} \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i &= \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \sum_{i\in[n]} \boldsymbol{x}_i & (3.62)\\
&= \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \sum_{i\in[n]} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{e}_i & (3.63)\\
&= \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{e}_i & (3.64)\\
&= \boldsymbol{x}^\top \boldsymbol{e}_i & (3.65)\\
&= 1. & (3.66)
\end{aligned}$$

■

REMARK

The proposition above have a particular case: That is, for all $j \in [n]$, we have $\sum_{i\in[n]} h_{ij} = 1$ (when intercept is included).

Recall that we have introduced hat matrix and annihilator matrix, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, $\boldsymbol{M} = \mathbb{I} - \boldsymbol{H}$, we indeed have

$$\begin{aligned}
\boldsymbol{y} &= (\boldsymbol{H} + \boldsymbol{M})\boldsymbol{y} & (3.67)\\
&= \boldsymbol{H}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} & (3.68)\\
&= \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) & (3.69)\\
&= \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{e} & (3.70)
\end{aligned}$$

which we may then rewrite $\boldsymbol{y}$ as $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}$, where we define $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ as *sample residuals*.

Proposition

Proposition 18. *The sample covariates are orthogonal to the sample residuals, that is,*

$$\sum_{i\in[n]} \boldsymbol{x}_i \widehat{e}_i = \boldsymbol{0}_p \tag{3.71}$$

Proof. We have that

$$\sum_{i\in[n]} \boldsymbol{x}_i \widehat{e}_i = \boldsymbol{X}^\top \widehat{\boldsymbol{e}} \tag{3.72}$$

$$= \boldsymbol{X}^\top \boldsymbol{M}\boldsymbol{e} \tag{3.73}$$

$$= \boldsymbol{X}^\top (\mathbb{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{e} \tag{3.74}$$

$$= \boldsymbol{X}^\top \boldsymbol{e} - \boldsymbol{X}^\top \boldsymbol{e} \tag{3.75}$$

$$= \mathbf{0}_p. \tag{3.76}$$

■

Proposition

Proposition 19. *If intercept is included in the model, then the sample residuals are centered, that is,*

$$\sum_{i\in[n]} \widehat{e}_i = 0. \tag{3.77}$$

Proof. Consider the design matrix as

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{np} \end{pmatrix}. \tag{3.78}$$

Then observe that

$$\sum_{i\in[n]} \widehat{e}_i = \langle \mathbb{1}_n, \widehat{\boldsymbol{e}} \rangle. \tag{3.79}$$

where $\mathbb{1}_n$ is the unit vector in $\mathbb{R}^n$ and $\mathbb{1}_n \in Im(\boldsymbol{H})$, also $\widehat{\boldsymbol{e}} \in Im(\boldsymbol{M})$ so the inner product will yield zero. ■

In a linear model, we have in fact

$$RSS(\widehat{\boldsymbol{\beta}}_n) = \widehat{\boldsymbol{e}}^\top \widehat{\boldsymbol{e}} = \boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{e}. \tag{3.80}$$

This is because

$$\widehat{\boldsymbol{e}}^\top \widehat{\boldsymbol{e}} = (\boldsymbol{M}\boldsymbol{e})^\top \cdot (\boldsymbol{M}\boldsymbol{e}) = \boldsymbol{e}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{e} := \boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{e} \tag{3.81}$$

where we used the fact that $\boldsymbol{M}$ is symmetric and idempotent.

3.3 Properties of OLS and Gauss-Markov Theorem

We will investigate several properties of the OLS estimator given by

$$\widehat{\boldsymbol{\beta}}_n := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{3.82}$$

Theorem

Theorem 13. *The OLS estimator is (conditional) unbiased given* $\boldsymbol{X}$, *that is,*

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right] = \boldsymbol{\beta}. \tag{3.83}$$

Also the covariance of the OLS estimator is given by

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \tag{3.84}$$

where $\sigma^2 := \mathbb{V}(\boldsymbol{\varepsilon}) = \mathbb{V}(\boldsymbol{y}|\boldsymbol{X})$.

Proof. Note that we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right] &= \mathbb{E}\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \middle| \boldsymbol{X}\right] && (3.85)\\
&= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbb{E}\left[\boldsymbol{y} \middle| \boldsymbol{X}\right] && (3.86)\\
&= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbb{E}\left[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \middle| \boldsymbol{X}\right] && (3.87)\\
&= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \mathbb{E}\left[\boldsymbol{\beta} \middle| \boldsymbol{X}\right] && (3.88)\\
&= \boldsymbol{\beta}. && (3.89)
\end{aligned}$$

where we used the fact that $\mathbb{E}[\boldsymbol{e}|\boldsymbol{X}] = \boldsymbol{0}$. For the variance term, note that we have

$$\begin{aligned}
\mathbb{V}\left(\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right) &= \mathbb{V}\left((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \middle| \boldsymbol{X}\right) && (3.90)\\
&= \left((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right) \cdot \mathbb{V}(\boldsymbol{y}|\boldsymbol{X}) \cdot \left((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right)^\top && (3.91)\\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \cdot \sigma^2 \cdot \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} && (3.92)\\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma^2. && (3.93)
\end{aligned}$$

■

Remark

(i) For a response given by one data $y_i \in \mathbb{R}$, we have $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i$, when put them into a vector, we then have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{3.94}$$

(ii) The conditional unbiased and variance will also imply the unconditional case:

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_n\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right]\right] = \mathbb{E}[\boldsymbol{\beta}] \tag{3.95}$$

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}_n) = \mathbb{E}[\mathbb{V}(\widehat{\boldsymbol{\beta}}_n|\boldsymbol{X})] + \mathbb{V}(\mathbb{E}[\widehat{\boldsymbol{\beta}}_n|\boldsymbol{X}]) = \sigma^2 \mathbb{E}\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]. \tag{3.96}$$

Definition

Definition 13. *An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is said to be linear (in y_i) if there exists a weight matrix $\boldsymbol{W} \in \mathscr{M}_{p,n}$ such that $\widehat{\boldsymbol{\beta}} = \boldsymbol{W}\boldsymbol{y}$.*

Note that for the OLS estimator $\widehat{\boldsymbol{\beta}}$, it is also linear since we have

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} := \widehat{\boldsymbol{W}}\boldsymbol{y} \tag{3.97}$$

where $\widehat{\boldsymbol{W}} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \in \mathbb{R}^{p\times n}$.

Theorem

Theorem 14 (Gauss-Markov)**.** *Assume a full rank homoscedastic linear model, the OLS estimator $\widehat{\boldsymbol{\beta}}$ is the* ***Best Linear Unbiased Estimator (BLUE)*** *estimator of $\boldsymbol{\beta}$. Which is, for any linear unbiased estimator $\widetilde{\boldsymbol{\beta}}_n$, we have*

$$\mathbb{V}\left(\widetilde{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right) \succeq \mathbb{V}\left(\widehat{\boldsymbol{\beta}}_n \middle| \boldsymbol{X}\right). \tag{3.98}$$

Note that $\succeq$ is the sense in terms of matrix, if $\boldsymbol{A} \succeq \boldsymbol{0}$ then it means $\boldsymbol{A}$ is positive definite.

Proof. Let $\widetilde{\boldsymbol{\beta}}$ be linear and unbiased, then $\exists \widetilde{\boldsymbol{W}} \in \mathbb{R}^{p\times n}$ such that $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{W}}\boldsymbol{y}$ also $\mathbb{E}[\widetilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Then we can relate $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}$ by

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}} + \widehat{\boldsymbol{W}})\boldsymbol{y} := (\boldsymbol{D} + \widehat{\boldsymbol{W}})\boldsymbol{y} = \boldsymbol{D}\boldsymbol{y} + \widehat{\boldsymbol{\beta}} \tag{3.99}$$

where $\boldsymbol{D} := \widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}}$. We further have

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\beta}} \tag{3.100}$$

and taking expectation on both sides will yield

$$\mathbb{E}[\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}] + \mathbb{E}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] \tag{3.101}$$

using the fact that $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}$ are unbiased, with the property of conditional expectation, we have $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$. Then we have

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} && (3.102)\\
&= (\widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}})\boldsymbol{y} + \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} && (3.103)\\
&= \boldsymbol{D}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \widehat{\boldsymbol{W}}\boldsymbol{y} - \boldsymbol{\beta} && (3.104)\\
&= \boldsymbol{D}\boldsymbol{\varepsilon} + \widehat{\boldsymbol{W}}\boldsymbol{\varepsilon} + \widehat{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} && (3.105)\\
&= (\boldsymbol{D} + \widehat{\boldsymbol{W}})\boldsymbol{\varepsilon} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} && (3.106)\\
&= (\boldsymbol{D} + \widehat{\boldsymbol{W}})\boldsymbol{\varepsilon}. && (3.107)
\end{aligned}$$

Now finally

$$\begin{aligned}\mathbb{V}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}|\boldsymbol{X}) &= \mathbb{V}((\boldsymbol{D}+\widehat{\boldsymbol{W}})\boldsymbol{\varepsilon}|\boldsymbol{X}) && (3.108)\\ &= (\boldsymbol{D}+\widehat{\boldsymbol{W}})\cdot\sigma^2\cdot(\boldsymbol{D}+\widehat{\boldsymbol{W}})^\top && (3.109)\\ &= \sigma^2(\boldsymbol{D}\boldsymbol{D}^\top+\boldsymbol{D}\widehat{\boldsymbol{W}}^\top+\widehat{\boldsymbol{W}}\boldsymbol{D}^\top+\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^\top) && (3.110)\\ &= \sigma^2(\boldsymbol{D}\boldsymbol{D}^\top+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) && (3.111)\\ &= \sigma^2\cdot\boldsymbol{D}\boldsymbol{D}^\top+\mathbb{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}). && (3.112)\end{aligned}$$

Since $\boldsymbol{D}\boldsymbol{D}^\top$ is positive definite, it follows that

$$\mathbb{V}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}) \succeq \mathbb{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}). \tag{3.113}$$

■

Corollary

Corollary 2 (Unconditional Gauss-Markov)**.** *Let $\widetilde{\boldsymbol{\beta}}$ be a linear unbiased estimator of $\boldsymbol{\beta}$, then*

$$\mathbb{V}(\widetilde{\boldsymbol{\beta}}) \succeq \mathbb{V}(\widehat{\boldsymbol{\beta}}) \tag{3.114}$$

Proof. Note that we have

$$\mathbb{V}(\boldsymbol{X}) = \mathbb{V}(\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}])+\mathbb{E}[\mathbb{V}(\boldsymbol{X}|\boldsymbol{Y})] \tag{3.115}$$

so we will have

$$\mathbb{V}(\widetilde{\boldsymbol{\beta}}) = \mathbb{V}(\mathbb{E}[\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}])+\mathbb{E}[\mathbb{V}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})] = \mathbb{V}(\boldsymbol{\beta})+\mathbb{V}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}) := \mathbb{V}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}), \tag{3.116}$$

and similarly for $\mathbb{V}(\widehat{\boldsymbol{\beta}})$ we have $\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \mathbb{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$, so the result follows. ■

Recall that $e_i = \varepsilon_i = y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}$ and $\widehat{e}_i = y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}$. We further study some properties of the sample residuals:

Proposition

Proposition 20. *The sample residual is conditionally unbiased, that is, $\mathbb{E}[\widehat{e}_i|\boldsymbol{X}] = \mathbb{E}[\varepsilon_i|\boldsymbol{X}] = 0$.*

Proof. We in fact let $\widehat{\boldsymbol{e}}_i = (\widehat{e}_1,\cdots,\widehat{e}_n)^\top$ then

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{e}}|\boldsymbol{X}] &= \mathbb{E}[\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] && (3.117)\\ &= \mathbb{E}[\boldsymbol{y}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}|\boldsymbol{X}] && (3.118)\\ &= \mathbb{E}[(\mathbb{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}|\boldsymbol{X}] && (3.119)\\ &= \mathbb{E}[\boldsymbol{M}\boldsymbol{y}|\boldsymbol{X}] && (3.120)\\ &= \boldsymbol{0}. && (3.121)\end{aligned}$$

So it follows that $\mathbb{E}[\widehat{e}_i|\boldsymbol{X}] = 0$. ■

Proposition

Proposition 21. *The conditional variance of the sample residuals are given by*

$$\mathbb{V}(\widehat{e}_i|\boldsymbol{X}) = \mathbb{E}[\widehat{e}_i^2|\boldsymbol{X}] = \sigma^2(1-h_{ii}) \tag{3.122}$$

Proof. The first equality is easy, since one have

$$\mathbb{V}(\widehat{e}_i|\boldsymbol{X}) = \mathbb{E}[\widehat{e}_i^2|\boldsymbol{X}] + \Big(\mathbb{E}[\widehat{e}_i|\boldsymbol{X}]\Big)^2 = \mathbb{E}[\widehat{e}_i^2|\boldsymbol{X}]. \tag{3.123}$$

For the other equality, note that

$$\begin{aligned}
\mathbb{V}(\widehat{\boldsymbol{e}}_n|\boldsymbol{X}) &= \mathbb{V}(\boldsymbol{Me}|\boldsymbol{X}) && (3.124)\\
&= \boldsymbol{M}\mathbb{V}(\boldsymbol{e}|\boldsymbol{X})\boldsymbol{M}^\top && (3.125)\\
&= \sigma^2\boldsymbol{M}^2 && (3.126)\\
&= \sigma^2(\mathbb{I}-\boldsymbol{H}). && (3.127)
\end{aligned}$$

Hence

$$\mathbb{V}(\widehat{e}_i|\boldsymbol{X}) = \Big[\sigma^2\Big(\mathbb{I}_n - \boldsymbol{H}\Big)\Big]_{ii} = \sigma^2(1-h_{ii}). \tag{3.128}$$

■

If we wish to estimate $\mathbb{E}[\varepsilon^2]$, then a natural estimator is defined by

$$\widehat{\sigma}^2_{naive} := \frac{1}{n}\sum_{i\in[n]}\widehat{e}_i^2. \tag{3.129}$$

However this estimator is not ideal.

Theorem

Theorem 15. *The estimator $\widehat{\sigma}^2_{naive}$ is negatively biased, that is,*

$$\mathbb{E}\Big[\widehat{\sigma}^2_{naive}\Big] = \frac{n-p}{n}\sigma^2 < \sigma^2. \tag{3.130}$$

and the bias corrected estimator is defined by

$$\widehat{\sigma}^2_{cor} := \frac{1}{n-p}\sum_{i\in[n]}\widehat{e}_i^2. \tag{3.131}$$

Proof. Note that

$$\frac{1}{n}\sum_{i\in[n]}\widehat{e}_i^2 = \frac{1}{n}\widehat{\boldsymbol{e}}^\top\widehat{\boldsymbol{e}} = \frac{1}{n}\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon} = \frac{1}{n}\mathrm{tr}(\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}), \tag{3.132}$$

then we have

$$\begin{aligned}\frac{1}{n}\mathbb{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}] &= \frac{1}{n}\mathbb{E}\Big[\mathrm{tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{M})\Big|\boldsymbol{X}\Big] && (3.133)\\ &= \frac{1}{n}\mathrm{tr}\Big(\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{M}\Big) && (3.134)\\ &= \frac{1}{n}(\sigma^2(n-p)) && (3.135)\end{aligned}$$

as desired. ■

Theorem

Theorem 16. *Consider a Gaussian regression model, then the conditional likelihood of the model is*

$$L_n(\boldsymbol{\beta},\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\sum_{i\in[n]}\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right) \tag{3.136}$$

Inference in the Gaussian Regression Model

4.1 Basic Theory of Testing

In the previous chapter, we provide an estimate of $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$. We would like to know if $\widehat{\boldsymbol{\beta}}$ is "good enough". We partition $\mathbb{R}^p = \mathscr{H}_0 \cup \mathscr{H}_1$ as two hypothesis and design a test statistics $t(\mathscr{D}_n)$ based on the data. The test statistic is designed in a way, such that if the null hypothesis $\mathscr{H}_0$ is true, we can determine the distribution. So whenever it is true we can then see how likely it was to observe $t(\mathscr{D}_n)$.

For example, suppose that $t(\mathscr{D}_n) \sim N(0,1)$ when the null hypothesis ($\mathscr{H}_0$) is true, and we want to see how likely is it that $t = 100$? Under this normal assumption it is extremely unlikely, so we may see $\mathscr{H}_0$ is extremely unlikely to be true.

For a two-sided test with symmetric distribution, for a realization t' of $t(\mathscr{D}_n)$, the p-value can be interpreted as

$$p = \mathbb{P}_{\theta_0}\Big(|t(\mathscr{D}_n)| \geq |t'|\Big), \tag{4.1}$$

that is, the probability under $\mathscr{H}_0$ that the test statistic is larger (as large) than what we observed. Also we define the significance level of a test as the probability of rejecting $\mathscr{H}_0$ when it is true.

In a linear model, if we predict $\widehat{m}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_n$, it is unlikely that the true value y is the same as predicted $\widehat{m}(\boldsymbol{x})$, and we wish to construct a random interval $\widehat{C} := [a,b]$ for some $a < b$ such that

$$\mathbb{P}\Big(y \in \widehat{C}\Big) = 1 - \alpha \tag{4.2}$$

where α is the significance level, or the miscoverage level. Assume we have a centered symmetric distribution, then we define $z_{\alpha/2}$ ()the quantile of order $\alpha/2$) by

$$\mathbb{P}\Big(-z_{\alpha/2} \leq t(\mathscr{D}_n) \leq z_{\alpha/2}\Big) = 1 - \alpha \tag{4.3}$$

that is, $z_{\alpha/2} = F^{-1}(\alpha/2)$ for the cumulative distribution function F. The value $z_{\alpha/2}$ is also called the critical value of the test, and we reject $\mathscr{H}_0$ if the observation $t(\mathscr{D}_n)$ does not belong in the interval $[-z_{\alpha/2}, z_{\alpha/2}]$. *Note that it is only for centered symmetric distribution, for example standard normal. For χ^2 distribution we need to specify two different tails and the rejection region is different. See the appendix for the construction of confidence intervals as well as hypothesis tests.*

Theorem

Theorem 17 (Distribution of OLS and Sample Residuals). *Consider a gaussian linear model, that is, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, then given $\boldsymbol{X}$, the vectors $\widehat{\boldsymbol{e}}$ and $\widehat{\boldsymbol{\beta}}$ are independent and jointly normal:*

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ \widehat{\boldsymbol{e}} \end{pmatrix} \Bigg| \boldsymbol{X} \sim \mathcal{N}\left(\mathbf{0}_{p+n}, \sigma^2 \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \mathbf{0}_{p,n} \\ \mathbf{0}_{n,p} & \boldsymbol{M} \end{pmatrix}\right) \tag{4.4}$$

Proof. Their independence is ensured by the model it self. Note that

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\beta} + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \tag{4.5}$$

hence we have

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\beta}, \quad \widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{M}\boldsymbol{y} = \cancel{\boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta}} + \boldsymbol{M}\boldsymbol{\varepsilon} \tag{4.6}$$

By the unbiasedness of $\widehat{\boldsymbol{\beta}}$ and the zero mean property of $\widehat{\boldsymbol{e}}$, we only need to find the covariance matrix of the random vector $((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top, (\widehat{\boldsymbol{e}})^\top)^\top$, where we have

$$\begin{aligned}
\mathbb{V}\left(\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{e}} \end{pmatrix} \Bigg| \boldsymbol{X}\right) &= \mathbb{V}\left(\begin{pmatrix} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \\ \boldsymbol{M} \end{pmatrix} \boldsymbol{\varepsilon} \Bigg| \boldsymbol{X}\right) && (4.7) \\
&= \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \\ \boldsymbol{M} \end{pmatrix} \sigma^2 \mathbb{I}_n \begin{pmatrix} \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \boldsymbol{M} \end{pmatrix} && (4.8) \\
&= \sigma^2 \mathbb{I}_n \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} & (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M} \\ \boldsymbol{M}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \boldsymbol{M} \end{pmatrix} && (4.9) \\
&= \sigma^2 \cdot \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \mathbf{0}_{p,n} \\ \mathbf{0}_{n,p} & \boldsymbol{M} \end{pmatrix}. && (4.10)
\end{aligned}$$

■

Theorem

Theorem 18 (Distribution of the Scaled Variance Estimator). *Consider a Gaussian linear model, then the rescaled estimator $\widehat{\sigma}^2_{cor}$ defined in* (3.131) *satisfies*

$$(n-p) \cdot \frac{\widehat{\sigma}^2_{cor}}{\sigma^2} \sim \chi^2(n-p) \tag{4.11}$$

which is a χ^2 distribution with $(n-p)$ degrees of freedom.

Remark: If $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}_p, \mathbb{I}_p)$ and $\boldsymbol{A} \in \mathcal{M}_p$ is determinstic and idepotent, then $\boldsymbol{z}^\top \boldsymbol{A} \boldsymbol{z} \sim \chi^2(\operatorname{rank}(\boldsymbol{A}))$.

Proof. By direct compitation we have

$$(n-p) \cdot \frac{\widehat{\sigma}^2_{cor}}{\sigma^2} = (n-p) \cdot \frac{1}{\sigma^2} \frac{1}{n-p} \sum_{i \in [n]} \widehat{e}_i^2 = \frac{1}{\sigma^2} \widehat{\boldsymbol{e}}^\top \widehat{\boldsymbol{e}} = \left(\frac{\boldsymbol{e}}{\sigma}\right)^\top \boldsymbol{M} \frac{\boldsymbol{e}}{\sigma}, \tag{4.12}$$

and we are done since $\operatorname{rank}(\boldsymbol{M}) = n - p$ and $\boldsymbol{e}/\sigma \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. ■

4.2 t-Statistic and F-Statistic for testing

In this section, we would like to test some of the covariates β_j. Suppose we now wish to test a single covariate with hypothesis $\mathscr{H}_0 : \beta_j = b_j$ against $\mathscr{H}_1 : \beta_j \neq b_j$ for some $b_j \in \mathbb{R}$. Under null hypothesis, assume we have a Gaussian linear model, we simply have

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}} = \frac{\widehat{\beta}_j - b_j}{\sigma\sqrt{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}} \sim \mathscr{N}(0,1) \tag{4.13}$$

One drawback is that the true variance σ is not known. So one would replace σ by its estimate $\widehat{\sigma}_{cor}$, and we call the resulting statistic the t-statistic (or t-ratio).

Theorem

Theorem 19. *The statistic t_j is defined by*

$$t_j := \frac{\widehat{\beta}_j - b_j}{\widehat{\sigma}_{cor}\sqrt{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}} \tag{4.14}$$

Under $\mathscr{H}_0$ and assume a Gaussian linear model, the t-statistic has a student distribution $t(n-p)$ with $n-p$ degrees of freedom.

Remark: The student t-distribution can be obtained by a standard normal distribution Z and a $V \sim \chi(\nu)^2$ where

$$T = \frac{Z}{\sqrt{V/\nu}} \tag{4.15}$$

(See the appendix for more on distribution theory).

Proof. By direct computation,

$$\begin{aligned} t_n &= \frac{\widehat{\beta}_j - b_j}{\widehat{\sigma}_{cor}\sqrt{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}} && (4.16)\\ &= \frac{\widehat{\beta}_j - b_j}{\sqrt{\frac{1}{n-p}\sum_{i\in[n]} \widehat{e}_i^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}} && (4.17)\\ &= \frac{\widehat{\beta}_j - b_j}{\sigma\sqrt{\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}_{jj}}}\left(\sqrt{\frac{\boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{e}}{\sigma^2(n-p)}}\right)^{-1} && (4.18)\\ & && (4.19) \end{aligned}$$

where the results follows. ■

Hence to test the hypothesis $\mathscr{H}_0 : \beta_j = b_j$ against $\mathscr{H}_1 : \beta_j \neq b_j$ we proceed as follows:

- Compute the t-statistic for a parameter β_j with hypothesized value b_j;
- Choose the significance level α;
- Use the table of $t(n-p)$, find the critical value $z_{\alpha/2}$;
- Compare the test statistic t wto the critical value, if we have $|t| > z_{\alpha/2}$, we reject $\mathscr{H}_0$, and if $|t| < z_{\alpha/2}$, we accept $\mathscr{H}_0$.
- Note that we can also compute the p value $p := 2\mathbb{P}(t(n-p) > |t|)$ and reject $\mathscr{H}_0$ if $p \leq \alpha$, and accept $\mathscr{H}_0$ if $p > \alpha$

Remark

Note that accepting $\mathscr{H}_0$ truely means **we do not have enough evidence to reject** $\mathscr{H}_0$

Next, we consider a testing where we wish to know if several coefficients are equal to a given value, say 0. Without loss of generality we partition the covariates into S_1, S_2 of respective cardinality p_1, p_2, so that $\beta_j \neq 0$ for all $j \in S_1$ and $\beta_{j'} = 0$ for all $j' \in S_2$, then we may fit the model as

$$M_{1,2} : y = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{4.20}$$

then we wish to test if the model is in fact

$$M_1 : y = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{4.21}$$

that is, all coefficients in S_2 are zero, and we wish to test

$$\mathscr{H}_0 := \{\text{The model } M_1 \text{ is correct}\} \quad \mathscr{H}_1 := \{\text{The model } M_{12} \text{ is correct}\} \tag{4.22}$$

By assuming a Gaussian linear model, we may use likelihood ratio test (see appendix if not familiar) and we would result in the F-statistic.

Theorem

Theorem 20. *The F-statistic for a set of parameters S_1 is defined by*

$$F := \frac{(\widehat{\sigma}_1^2 - \widehat{\sigma}_{1,2}^2)/p_2}{\widehat{\sigma}_{1,2}^2/(n-p)} \tag{4.23}$$

(it is called the F-statistic with respect to S_1) Under the assumption of a Gaussian linear model, F has a Fisher distribution $F(p_2, n-p)$ wit parameters $p_2, n-p$.

Remark: The F distribution can be viewed as the ratio of two χ-squared distribution, where

$$F(m,n) \sim \frac{\chi_m^2/m}{\chi_n^2/n} \tag{4.24}$$

(*See appendix for more*)

Proof. We know that

$$F := \frac{n\sigma^{-2}(\widehat{\sigma}_1^2 - \widehat{\sigma}_{1,2}^2)/p_2}{n\sigma^{-2}\widehat{\sigma}_{1,2}^2/(n-p)} \tag{4.25}$$

and we also know that the denominator is just $\chi^2(n-p)/(n-p)$ because we know

$$\frac{n}{n-p} \cdot \frac{\widehat{\sigma}_{1,2}^2}{\sigma^2} = \frac{\widehat{\sigma}_{cor}^2}{\sigma^2} \sim \frac{\chi^2(n-p)}{n-p}. \tag{4.26}$$

Now for the numerator, under null hypothesis $\mathscr{H}_0$:

$$\begin{aligned}(\widehat{\sigma}_1^2 - \widehat{\sigma}_{1,2}^2) \cdot n &= \sum_{i \in [n]} \left(\widehat{e}_{i1}^2 + \widehat{e}_{1,2,i}^2\right) &(4.27)\\ &= \boldsymbol{\varepsilon}^\top (\boldsymbol{M}_1 - \boldsymbol{M}_{1,2})\boldsymbol{\varepsilon} &(4.28)\end{aligned}$$

Indeed we can show $\boldsymbol{M}_1 - \boldsymbol{M}_{1,2}$ is idempotent and determinstic, then we have

$$\operatorname{rank}(\boldsymbol{M}_1 - \boldsymbol{M}_{1,2}) = (n - p_1) - (n - (p_1 + p_2)) = p_2 \tag{4.29}$$

with the scaling on $\boldsymbol{\varepsilon}$, we finish the proof. ■

We may generalize to arbitrary linear hypothesis of the form

$$\mathscr{H}_0 : \boldsymbol{R\beta} = \boldsymbol{r} \tag{4.30}$$

where $\boldsymbol{R} \in \mathscr{M}_{kp}, \boldsymbol{r} \in \mathbb{R}^k$ with k denoting the number of constraints. The constraints we consider must satisfy (1) No redundacy, i.e $\boldsymbol{R}$ is full row rank; (2) The constraints do not contradict eah other.

Proposition

Proposition 22. *Under $\mathscr{H}_0$, the statistic*

$$F = \frac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^\top \left\{\boldsymbol{R}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{R}^\top\right\}^{-1} \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right)/k}{\widehat{\sigma}_{cor}^2} \tag{4.31}$$

follows a Fisher distribution $F_{k,n-p}$.

Proof. Under $\mathscr{H}_0$, we simply have $\boldsymbol{R\beta} = \boldsymbol{r}$, also we know that

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathscr{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right) \tag{4.32}$$

hence

$$\begin{aligned}\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\Big|\boldsymbol{X} &\sim \mathscr{N}\left(\boldsymbol{R\beta} - \boldsymbol{r}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}\right) &(4.33)\\ &\sim \mathscr{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}\right) &(4.34)\end{aligned}$$

hence we have

$$\left.\frac{\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}}{\sigma}\right|\boldsymbol{X} \sim \mathcal{N}\left(0, \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}\right) \tag{4.35}$$

and since $\dim(\boldsymbol{R}\widehat{\boldsymbol{\beta}}) = k$, so it follows that

$$\left(\frac{\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}}{\sigma}\right)^\top \left\{\boldsymbol{R}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{R}^\top\right\}^{-1}\left(\frac{\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}}{\sigma}\right)/k \sim \chi^2(k)/k \tag{4.36}$$

and we finally use the property that

$$(n-p)\cdot\frac{\widehat{\sigma}^2_{cor}}{\sigma^2} \sim \chi^2(n-p) \tag{4.37}$$

to conclude

$$\begin{aligned} F &= \frac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^\top \left\{\boldsymbol{R}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{R}^\top\right\}^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)/k}{\widehat{\sigma}^2_{cor}} && (4.38)\\ &= \frac{\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^\top \left\{\boldsymbol{R}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{R}^\top\right\}^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}-\boldsymbol{r}\right)/\sigma^2 k}{\widehat{\sigma}^2_{cor}/\sigma^2} && (4.39)\\ &\sim \frac{\chi^2(k)/k}{\chi^2(n-p)/(n-p)}. && (4.40) \end{aligned}$$

■

Remark

If $\boldsymbol{z} \sim \mathcal{N}_p(\boldsymbol{0}, \mathbb{I})$ and $\boldsymbol{A}$ is determinstic and idempotent, then

$$\boldsymbol{z}^\top\boldsymbol{A}\boldsymbol{z} \sim \chi^2(\operatorname{rank}(\boldsymbol{A})) \tag{4.41}$$

If $\boldsymbol{z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z}-\boldsymbol{\mu}) \sim \mathcal{N}_p(\boldsymbol{0}, \mathbb{I}) \quad (\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \sim \chi^2(p) \tag{4.42}$$

Theorem

Theorem 21. *Consider a Gaussian linear model, fix $j \in [p]$, denote $z_{\alpha/2}$ as the $\alpha/2$ quantile of the student distribution with $n-p$ degress of freedom, then the interval*

$$I_\alpha = \left[\widehat{\beta}_j - z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}_{jj}}, \widehat{\beta}_j + z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}_{jj}}\right] \tag{4.43}$$

has exact coverage probability of $1-\alpha$.

The testing procedure consisting at rejecting the null hypothesis $\mathscr{H}_0 : \beta_j = b_j$ when b_j is not in I_α is equivalent to the t-test. Now how can we construct confidence intervals for several coefficients?

Proposition

Proposition 23 (Bonferonnni Bound). *Assume that $\mathbb{P}(\beta_j \in I_j) = 1 - \alpha_j$ for each $j \in [p]$, then*

$$\mathbb{P}(\boldsymbol{\beta} \in I) \geq 1 - \sum_{j \in [p]} \alpha_j \tag{4.44}$$

Proof. We have

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\beta} \in I) &:= \bigcap_{j \in [p]} \mathbb{P}(\beta_j \in I_j) && (4.45)\\
&= 1 - \bigcup_{j \in [p]} \mathbb{P}(\beta_j \notin I_j) && (4.46)\\
&\geq 1 - \sum_{j \in p} \mathbb{P}(\beta_j \notin I_j) && (4.47)\\
&= 1 - \sum_{j \in [p]} \alpha_j. && (4.48)
\end{aligned}$$

■

Theorem

Theorem 22 (Boferonni Confidence Intervals). *Consider a Gaussian linear model, fix $j \in [p]$ and denote $z_{\alpha/2p}$ the $\alpha/2p$ quantile of the student distribution with $n-p$ degress of freedom. Let $I_{\alpha,j}$ be the $1-\alpha/p$ confidence interval for β_j, then*

$$I_\alpha = I_{alpha.p,1} \times \cdots \times I_{\alpha/p,p} \tag{4.49}$$

has coverage probability greater than α, $\mathbb{P}(\boldsymbol{\beta} \in I_\alpha) \geq 1 - \alpha$.

The Bonferonni confidence interval is not an exact coverage, when p is small (e.g $p \leq 5$)as well as α, the Bonferonni bound is relatively sharp, i other scenarios, the confidence interval may ve conservative, meaning $\mathbb{P}(\boldsymbol{\beta} \in I_\alpha) >> 1 - \alpha$ and lead to a very broad, uninformative intervals.

Proposition

Proposition 24. *The statistic*

$$\frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \left(\mathbf{X}^\top \mathbf{X}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) / p}{\widehat{\sigma}^2_{cor}} \tag{4.50}$$

follows a Fisher distribution $F_{p,n-p}$.

Proof. The idea is the same as proposition 22. ■

Theorem

Theorem 23. *The confidence interval of $\boldsymbol{\beta}$ is defined by*

$$I_\alpha := \left\{ \boldsymbol{\beta} \in \mathbb{R}^p ; \frac{\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \left(\mathbf{X}^\top \mathbf{X}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\widehat{\sigma}_{cor}^2} \leq pF_{p,n-p}(1-\alpha) \right\} \tag{4.51}$$

which has covergae probability $1-\alpha$ for $\boldsymbol{\beta}$.

The confidence interval can be viewed as an ellipsoid centered at $\widehat{\boldsymbol{\beta}}$ and with radius $\sqrt{pF_{p,n-p}(1-\alpha)}$.

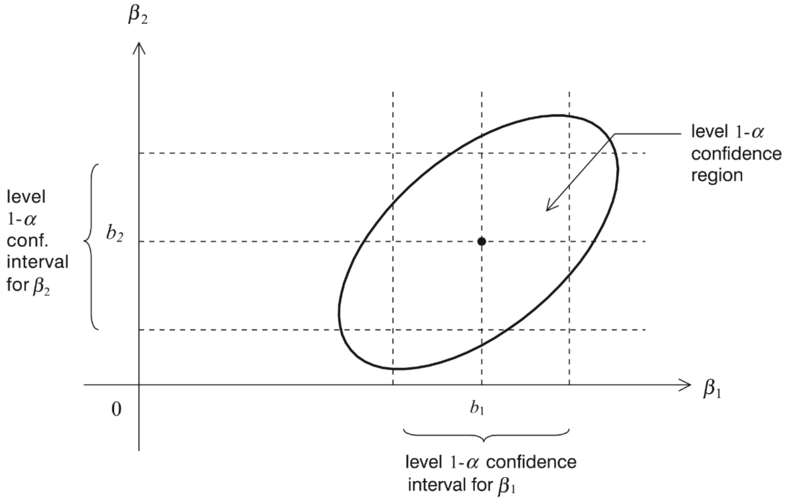


Figure 3: Difference between indivivual and simultaneous confidence ontervals

4.3 Confidence Interval for Predictions

From a sample $\mathscr{D}_n$, we already showed how to find the estimator $\widehat{\boldsymbol{\beta}}$, and now asume we have a new observaition $(\boldsymbol{x}_{new}, y_{new}) \notin \mathscr{D}_n$, where

$$y_{new} = \boldsymbol{x}_{new}^\top \boldsymbol{\beta} + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.52}$$

we want to create confidence intervals for both y_{new} and $m(\boldsymbol{x}_{new}) := \boldsymbol{x}_{new}^\top \boldsymbol{\beta}$.

Proposition

Proposition 25. *Assume a Gaussian linear model, then*

$$\widehat{m}(\boldsymbol{x}_{new}) := \boldsymbol{x}_{new}^\top \widehat{\boldsymbol{\beta}} \Big| \mathbf{X}, \boldsymbol{x}_{new} \sim \mathcal{N}\left(m(\boldsymbol{x}_{new}), \sigma^2 \boldsymbol{x}_{new}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_{new}\right) \tag{4.53}$$

also given $\mathbf{X}, \boldsymbol{x}_{new}$ we have $\widehat{m}(\boldsymbol{x}_{new}) \perp\!\!\!\perp \widehat{\sigma}_{cor} \perp\!\!\!\perp y_{new}$.

Proof. We know that $\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\boldsymbol{e}}|\boldsymbol{X}$ so it also holds for any function of them. We know $\widehat{\sigma}_{cor}$ is a function of $\widehat{\boldsymbol{e}}$, $\widehat{m}(\boldsymbol{x}_{new})$ is a function of $\boldsymbol{\beta}$, so the independence holds. We know a linear combination of normal distribution is still a normal distribution, so

$$\boldsymbol{x}_{new}^{\top}\widehat{\boldsymbol{\beta}} \sim \mathcal{N} \tag{4.54}$$

where

$$\mathbb{E}\left[\boldsymbol{x}_{new}^{\top}\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X},\boldsymbol{x}_{new}\right] = \boldsymbol{x}_{new}^{\top}\boldsymbol{\beta} \tag{4.55}$$

by using the fact that $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right] = \boldsymbol{\beta}, \boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{x}_{new}$, also

$$\mathbb{V}\left(\boldsymbol{x}_{new}^{\top}\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X},\boldsymbol{x}_{new}\right) = \boldsymbol{x}_{new}^{\top}\cdot\sigma^2\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\cdot\boldsymbol{x}_{new} \tag{4.56}$$

hence we have

$$\boldsymbol{x}_{new}^{\top}\widehat{\boldsymbol{\beta}}\Big|\boldsymbol{X},\boldsymbol{x}_{new} \sim \mathcal{N}\left(\boldsymbol{x}_{new}^{\top}\boldsymbol{\beta},\sigma^2\boldsymbol{x}_{new}^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{x}_{new}\right) \tag{4.57}$$

■

Theorem

Theorem 24. *Assume a Gaussian linear model and denote by $z_{\alpha/2}$ the $\alpha/2$ quantile of the student distribution with $n-p$ degress of freedom, then $\widehat{m}(\boldsymbol{x}_{new})$ is the best unbiased estimator of $m(\boldsymbol{x})$, and the interval*

$$I_{\alpha} = \left[\widehat{m}(\boldsymbol{x}_{new}) - z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{h_{xx}} \quad , \quad \widehat{m}(\boldsymbol{x}_{new}) + z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{h_{xx}}\right] \tag{4.58}$$

has exact coverage probability of $1-\alpha$ for $\widehat{m}(\boldsymbol{x}_{new})$.

To show $\widehat{m}(\boldsymbol{x}_{new})$ is the best, we may just apply Gauss-Markov theorem, and the confidence interval follows naturally from inverting a t statistic.

Theorem

Theorem 25 (Confidence interval for y_{new})**.** *Assume a Gaussian linear model, and denote by $\alpha/2$ the $\alpha/2$ quantile of the student distribution with $n-p$ degress of freedom. The interval*

$$I_{\alpha} = \left[\widehat{m}(\boldsymbol{x}_{new}) - z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{1+h_{xx}} \quad , \quad \widehat{m}\boldsymbol{x}_{new}) + z_{\alpha/2}\widehat{\sigma}_{cor}\sqrt{1+h_{xx}}\right] \tag{4.59}$$

has exact coverage probability of $1-\alpha$ for y_{new}.

Proof. We know that $y_{new} = \boldsymbol{x}_{new}^{\top}\boldsymbol{\beta} + \varepsilon_{new}$ where $\varepsilon_{new} \sim \mathcal{N}(0,\sigma^2)$ so

$$y_{new} - \boldsymbol{x}_{new}^{\top}\boldsymbol{\beta}\Big|\boldsymbol{x}_{new} \sim \mathcal{N}\left(0,\sigma^2\right) \tag{4.60}$$

from the previous theorem, we know that

$$\boldsymbol{x}_{new}^{\top}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_{new}^{\top}\boldsymbol{\beta}\Big|\boldsymbol{X},\boldsymbol{x}_{new} \sim \mathcal{N}\left(0,\sigma^2 h_{xx}\right) \tag{4.61}$$

where $h_{xx} := \boldsymbol{x}_{new}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{new}$, hence

$$-(y_{new} - \boldsymbol{x}_{new}^\top)\widehat{\boldsymbol{\beta}}\Big|\boldsymbol{X}, \boldsymbol{x}_{new} \sim \mathscr{N}\left(0, \sigma^2(1+h_{xx})\right) \tag{4.62}$$

and by symmetry

$$\frac{y_{new} - \widehat{m}(\boldsymbol{x}_{new})}{\sigma\sqrt{1+h_{xx}}} \sim \mathscr{N}(0,1) \implies \frac{y_{new} - \widehat{m}(\boldsymbol{x}_{new})}{\widehat{\sigma}_{cor}\sqrt{1+h_{xx}}} \sim t(n-p) \tag{4.63}$$

and hence the confidence interval can be obtained by inverting the t-statistic. ■

Asymptotic Properties and Inference

In the previous chapter, we discussed the inference for Gaussian linear model, under the assumption that $\mathbb{E}[y|x] = \boldsymbol{x}^\top \boldsymbol{\beta}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$. In this chapter, we want to know the same results as in the previous chapter, but we do not assume a Gaussian linear model. The main idea to allow such statements is to use the poewe of asymptotics.

Proposition

Proposition 26. *Assume that (H1) $\mathbb{E}[y^2] < \infty$; (H2) $\mathbb{E}[||x||_2^2] < \infty$; (H3) $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ is positive definite. Then the followings hold:*

$$\frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i\boldsymbol{x}_i^\top \xrightarrow{\mathbb{P}} \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] \tag{5.1}$$

$$\frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i y_i \xrightarrow{\mathbb{P}} \mathbb{E}[\boldsymbol{x}y] \tag{5.2}$$

$$\frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i e_i \xrightarrow{\mathbb{P}} \mathbb{E}[\boldsymbol{x}e] \tag{5.3}$$

The results follows directly from the law of large numbers and independence.

Theorem

Theorem 26 (Convergence of OLS estimator)**.** *Let $\{\widehat{\boldsymbol{\beta}}\}_{n\in\mathbb{N}}$ be a sequence of OLS estimators, then under (H1),(H2),(H3), we have*

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\beta} := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1} \cdot \mathbb{E}[\boldsymbol{x}y]. \tag{5.4}$$

Proof. We know that

$$\widehat{\boldsymbol{\beta}} := (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} := \left(\frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i y\right) \tag{5.5}$$

by using continuous mapping theorem and the previous proposition, we will have the desired result. ■

Note that we don't even require a linear model! If the model is indeed truly linear, then

$$\widehat{m}_n(\boldsymbol{x}) := \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{x}^\top \boldsymbol{\beta} = m(\boldsymbol{x}), \tag{5.6}$$

otherwise, we have the decomposition

$$m(\boldsymbol{x}) = \underbrace{\boldsymbol{x}^\top \boldsymbol{\beta}}_{l(x)} + \underbrace{m(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{\beta}}_{\eta(x)} \tag{5.7}$$

and we have

$$\widehat{m}_n(\boldsymbol{x}) := \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{x}^\top \boldsymbol{\beta} := I(\boldsymbol{x}) \tag{5.8}$$

where $I(\boldsymbol{x})$ is the closest linear function of $\boldsymbol{x}$ to y (or m) and η represents non-linearity.

Theorem

Theorem 27. *Let $\{\widehat{\boldsymbol{\beta}}\}_{n\in\mathbb{N}}$ be a sequenceof OLS estimators, assume an homoskedastic linear model and (H1),(H2),(H3), we have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_p, \sigma^2 \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]^{-1}\right). \tag{5.9}$$

Proof. It can be shown that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \cdot \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i\right). \tag{5.10}$$

Using central limit theorem and the fact that $\mathbb{E}[\boldsymbol{x}_i e_i] = 0$, we have

$$\frac{1}{\sqrt{n}} \cdot \sum_{i=1}^{n} \boldsymbol{x}_i e_i \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}(\boldsymbol{x}_i \varepsilon_i)\right), \tag{5.11}$$

and by some computations we have

$$\mathbb{V}(\boldsymbol{x}_i e_i) = \mathbb{E}[\boldsymbol{x}_i \varepsilon_i^2 \boldsymbol{x}_i^\top] - \left(\mathbb{E}[\boldsymbol{x}_i \varepsilon_i]\right)^2 = \mathbb{E}\left[\boldsymbol{x}_i \mathbb{E}[\varepsilon_i^2 | \boldsymbol{x}_i] \boldsymbol{x}_i^\top\right] = \sigma^2 \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] \tag{5.12}$$

so using Slutsky's theorem we will have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]^{-1} \cdot \mathcal{N}\left(\mathbf{0}_p, \sigma^2 \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\right) := \mathcal{N}\left(\mathbf{0}_p, \sigma^2 \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\right). \tag{5.13}$$

■

Note that if we assume a Gaussian model, then we will have

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \boldsymbol{X} \sim \mathcal{N}\left(\mathbf{0}_p, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right). \tag{5.14}$$

Now we will need to obtain a consistent variance estimator:

Theorem

Theorem 28. *Let $\{\widehat{\sigma}_{naive}\}, \{\widehat{\sigma}_{cor}\}$ be sequences of naive and corrected variance estimators, and assume an homoskedastic linear model and (H1),(H2),(H3), then*

$$\widehat{\sigma}_{naive} \xrightarrow{\mathbb{P}} \sigma^2 \quad \widehat{\sigma}_{cor} \xrightarrow{\mathbb{P}} \sigma^2. \tag{5.15}$$

Note that when p is fixed, deviding by n or $n-p$ does not matter so we only prove for $\widehat{\sigma}_{naive}$.

Proof. By direct computation we have that

$$\widehat{\sigma}^2_{naive} := \frac{1}{n}\sum_{i\in[n]}(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})^2 \tag{5.16}$$

$$= \frac{1}{n}\sum_{i\in[n]}\left(y_i - \boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)^2 \tag{5.17}$$

$$= \frac{1}{n}\sum_{i\in[n]} e_i^2 + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top\frac{2}{n}\sum_{i\in[n]} e_i\boldsymbol{x}_i + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top\left(\frac{1}{n}\sum_{i\in[n]}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}). \tag{5.18}$$

In the first term, since e_is are independent from our assumptions, so by WLLN, it will converge to $\mathbb{E}[e^2] = \sigma^2$ in probability. Also we know $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ in probability so the second term and the third term will go to zero asymptotically. Hence $\widehat{\sigma}^2_{naive} \to \sigma^2$ in probability. ■

Since we know that $\mathbb{V}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, so a natural choice of variance estimator is

$$\widehat{V} := \widehat{\sigma}^2_{cor}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{5.19}$$

Theorem

Theorem 29. *Let $\{\widehat{V}_n\}$ be a sequence of variance estimators of the OLS, assume a homoskedastic model and (H1),(H2),(H3), we have*

$$n\widehat{V}_n \xrightarrow{\mathbb{P}} \sigma^2\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1} \tag{5.20}$$

Proof. We see that

$$n\widehat{V}_n = \left(\frac{1}{n}\sum_{i\in[n]}\widehat{e}_i^2\right)\cdot\left(\frac{1}{n}\sum_{i\in[n]}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}, \tag{5.21}$$

then using the previous theorems, the result is immediate. ■

Now, assume we are interested in some function (not necessarily linear) of $\boldsymbol{\beta}$, say $\theta := f(\boldsymbol{\beta})$ for $f:\mathbb{R}^p \to \mathbb{R}$.

Theorem

Theorem 30. *Let $\widehat{\theta}_n = f(\widehat{\boldsymbol{\beta}}_n)$, and $\{\widehat{\theta}_n\}$ be a sequence of estimators of θ, assume a homoskedastic linear model and (H1), (H2), (H3), and furthermore $f \in \mathscr{C}^1(\boldsymbol{\beta}), \nabla_f(\boldsymbol{\beta}) \neq \mathbf{0}_p$, then*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\nabla_f(\boldsymbol{\beta})^\top\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]^{-1}\nabla_f(\boldsymbol{\beta})\right) \tag{5.22}$$

We will first recall **mean value property** in $\mathbb{R}^p$: Let $f:\mathbb{R}^p\to\mathbb{R}$, $f\in\mathscr{C}^1([\boldsymbol{a},\boldsymbol{b}])$ defined as $[\boldsymbol{a},\boldsymbol{b}] := t\boldsymbol{a}+(1-t)\boldsymbol{b}, t\in(0,1)$, then $\exists c \in (\boldsymbol{a},\boldsymbol{b})$ such that $f(\boldsymbol{a}) - f(\boldsymbol{b}) = \nabla f(\boldsymbol{c})^\top(\boldsymbol{a}-\boldsymbol{b})$.

Proof. Using mean value theorem, it can be shown that

$$\widehat{\theta}_n - \theta = \nabla f(\boldsymbol{c}_n)^\top (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \tag{5.23}$$

for some $\boldsymbol{c}_n \in (\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta})$. So we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \nabla f(\boldsymbol{\beta}) + \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top (\nabla f(\boldsymbol{c}_n) - \nabla f(\boldsymbol{\beta})). \tag{5.24}$$

Using central limit theorem and slutsky's theorem, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \nabla f(\boldsymbol{\beta}) \to \nabla f(\boldsymbol{\beta})^\top \mathcal{N}\left(\boldsymbol{0}_p, \sigma^2 \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\right). \tag{5.25}$$

Also since $\boldsymbol{c}_n \in (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, $\exists t \in [0,1]$ such that

$$\boldsymbol{c}_n = t\widehat{\boldsymbol{\beta}}_n + (1-t)\boldsymbol{\beta} \xrightarrow{\mathbb{P}} t\boldsymbol{\beta} + (1-t)\boldsymbol{\beta} = \boldsymbol{\beta} \tag{5.26}$$

by continuous mapping theorem, we have $\nabla f(\boldsymbol{c}_n) - \nabla f(\boldsymbol{\beta}) = o(1)$, so the result follows. ∎

Theorem

Theorem 31. *Let $\{\widehat{V}_n\}$ be a sequence of variance estimators of θ, assume an homoskedastic linear model and (H1),(H2),(H3), if $f \in \mathscr{C}^1(\boldsymbol{\beta})$ and $\nabla f(\boldsymbol{\beta}) \neq \boldsymbol{0}_p$, then*

$$n\widehat{V}_n \xrightarrow{\mathbb{P}} \sigma^2 \nabla f(\boldsymbol{\beta})^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1} \nabla f(\boldsymbol{\beta}) \tag{5.27}$$

The proof is essentially a consequence of the previous theorem, together with continuous mapping theorem.

Theorem

Theorem 32. *Let $\{t_n\}$ be a sequence of t-statistics defined by*

$$t_\theta = \frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{V}_\theta}} \tag{5.28}$$

then under the assumptions as before,

$$t_{\theta,n} \xrightarrow{d} \mathcal{N}(0,1). \tag{5.29}$$

Analysis of Variance

In this section, we will restrict our attention to the case where the covariates $x_1, \cdots, x_p$ are **categorical**. We will use dummy variables to transform the covariates so the design matrix $\boldsymbol{X}$ is numerical. Assume that $p = 1$ and x is a categorical variables with G categories, the covariate support is finite, with elements to be understood as **labels** :$Supp(x) = \{1, \cdots, G\}$. Then we may partition $Supp(x)$ by creating G new covariates such that

$$z_g = \mathbb{1}_{X=g}, g \in [G] \tag{6.1}$$

and $\boldsymbol{Z}$ as he new design matrix with columns given by the new covariates $z_1, \cdots, z_p$. For example, with only one covariate, suppose the category is "Bad, Medium, Good", and our observation (design matrix) conststing 4 samples is given by

$$\boldsymbol{X} = \begin{bmatrix} \text{Good} \\ \text{Bad} \\ \text{Bad} \\ \text{Medium} \end{bmatrix} \tag{6.2}$$

then under the transformation, the new design matrix $\boldsymbol{Z}$ is given by

$$\boldsymbol{Z} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{6.3}$$

where the first column is the indicator for "Bad", second column for "Medium" and third column for "Good". The ith row is just the ith observation. With intercept intrduced, we have

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \tag{6.4}$$

where the first column shown in blue is the intercept.

Proposition

Proposition 27. *Under the transformation where $\boldsymbol{Z}$ is the new design matrix, with the intercept the matrix $\boldsymbol{Z}$ has rank G.*

Proof. This is simply a linear algebra argument. ■

So we see that if we assume intercept, then the design matrix $\boldsymbol{Z}$ is not full rank, this problem is often called the "dummy variable trap". To adoid this issue we may just delete one column.
Now the regression function $m(x)$ is defined on a finite set $\{1, \cdots, G\}$ and we have

$$m(i) = \mathbb{E}[y|x = i] := \beta_i \quad i \in [G] \tag{6.5}$$

which is the expectation of y for elements in the group i. We define the group sizes by

$$n_j = \sum_{i\in[n]} \mathbb{1}_{x_i=j} \tag{6.6}$$

as the number of element in group j, $j \in [G]$.

Definition

Definition 14. ***One-way Classification*** *refers to the situation of one categorical covariate with a numeric response.*

Following the traditional literature, we sort the observations in the following manner:

- We use the first n_1 rows to place elements in g_1
- We use the rows n_1 up to n_1+n_2 to place n_2 elements in g_2, etc.

We use a double subscript to index the individual observations and merge responses with a common covariate value. We have $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_G)^\top$ where

$$\boldsymbol{y}_j = (y_{j,1}, \cdots, y_{j,n_j})^\top = (y_{n_1+\cdots+n_{j-1}+1}, \cdots, y_{n_1+\cdots+n_j})^\top \quad j \in [G] \tag{6.7}$$

Definition

Definition 15. *When only qualitative covariates are involved in a linear model, we refers to it as an* ***Analysis Of VAriance (ANOVA)*** *mode. In cases where both qualitative and quantitative covariates are involved, we refer to it as a* ***Analysis of COVAriance*** *model.*

An example: A health researcher wishes to compare the effects of fouranti-inflammatory drugs on arthritis patients. She takes a randomsample of patients and divides them randomly into four groups ofequal size, each of which receives one of the drugs. In the courseof the study several patients became seriously ill and had towithdraw, leaving four unequal-sized groups. We thus have fourindependent samples, each receiving a di!erent treatment.

Definition

Definition 16. *In the ANOVA literature, a categorical variable is often called a* ***factor****. The different categories of the covariate are called the* ***levels*** *of the factor.*

In the previous example, the drug is the factor, and the levelsare the four di!erent anti-inflammatory drugs. We can view the data as G samples where in each sample s_g we have access to the data $\{y_{ig}\}_{i\in[n_g]}$. Each sample s_g is then regarded as coming from its own underlying hypothetical population, distributed i.i.d. according to $\mathbb{P}_{y|\boldsymbol{x}=g}$. It is typically assumed that

$$y_{ig} = \beta_g + \varepsilon_{ig}, \quad i \in [n_g], g \in [G], \varepsilon_{ig} \sim \mathcal{N}(0, \sigma^2). \tag{6.8}$$

We may also write

$$y_{ig} = \beta + \alpha_g + \varepsilon_{ig} \tag{6.9}$$

where $\alpha_g = \beta_g - \beta$. This can be seen as a linear model since for $i \in s_g$,

$$y_i = \boldsymbol{z}_i^\top \boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{G} z_{ij}\beta_j + \varepsilon_i = \beta_g + \varepsilon_i. \tag{6.10}$$

Appendix: Statistical Inference

7.1 Basic Results of Random Samples

Below, denote $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ as a random sample. Then we know that

1. $\overline{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$;

2. $\overline{X}_n \perp\!\!\!\perp S_n^2$, and indeed $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$;

The t distribution can be obtained by a standard normal distribution $Z \sim N(0,1)$ and a $V \sim \chi^2_\nu$ distribution, where

$$T = \frac{Z}{\sqrt{V/\nu}} \overset{\text{pdf}}{=} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right). \tag{7.1}$$

Recall from CLT, where in a random sample $X_1, \cdots, X_n$ with $\mathbb{E}X = \mu, Var(X) = \sigma^2$, we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1). \tag{7.2}$$

We are also able to obtain T by a variation of central limit theorem, in a random sample $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, we have

$$T = \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t(n-1) \tag{7.3}$$

The F distribution can be viewed as the ratio of two χ-squared distribution, where

$$F(m,n) \sim \frac{\chi^2_m/m}{\chi^2_n/n} \tag{7.4}$$

where if we consider two mutually independent random samples $X_1, \cdots, X_m \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_n \sim N(\mu_2, \sigma_2^2)$, then

$$F = \frac{S_m^2/\sigma_1^2}{S_n^2/\sigma_n^2} \sim F(m-1, n-1). \tag{7.5}$$

Here are some useful transformations of random variables: Below, the independence of random variables is assumed.

The most basic ones are that the square of a standard normal $\mathcal{N}(0,1)$ distribution is a $\chi^2(1)$ distribution (a χ^2 distribution with one degree of freedom); using moment generating functions or convolutions, if $X_1, \cdots, X_n \overset{i.i.d}{\sim} \chi^2(1)$ then $\sum X_i \sim \chi^2(n)$. The same way works for many other samples like standard normal.

1. If $X \sim Exp(\theta)$ where $f(x) = \theta e^{-\theta x}$, then $2\theta X \sim \chi^2_2$, where the square of a standard normal is χ^2_1, and $\chi^2_a + \chi^2_b = \chi^2_{a+b}$.

2. If $X \sim Exp(\theta)$ where $f(x) = \theta e^{-\theta x}$, then $X_1 + \cdots + X_n \sim Gamma(n, \theta)$.

3. (Cochran's Theorem) Suppose we have $Z_1, \cdots, Z_n \overset{i.i.d}{\sim} N(0,1)$, then

$$\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2 \sim \chi^2_{n-1} \tag{7.6}$$

where $\overline{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \sim N\left(0, \frac{1}{n}\right)$.

4. The sample mean of a χ^2 distribution family is distributed according to a Gamma distribution. That is, if $X_1, \cdots, X_n \overset{i.i.d}{\sim} \chi^2_k$, then

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim Gamma\left(\frac{nk}{2}, \frac{2}{n}\right) \tag{7.7}$$

where $Gamma(\alpha, \beta) \sim \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$. That is, we have the following relations between Gamma distribution and Chi-squared distribution: If $Q \sim \chi^2_\nu$ and $c > 0$, then $cQ \sim Gamma\left(\frac{\nu}{2}, 2c\right)$.

5. If $X \sim \chi^2_a$ and $Y \sim \chi^2_b$, then $\frac{X}{X+Y} \sim Beta\left(\frac{a}{2}, \frac{b}{2}\right)$, a similar result is also used for the ratio of Gamma distributions: If $X \sim Gamma(\alpha, \theta)$ and $Y \sim Gamma(\beta, \theta)$, then $\frac{X}{X+Y} \sim Beta(\alpha, \beta)$. Furthermore, using **Basu's Theorem**, $\frac{X}{X+Y} \perp\!\!\!\perp X + Y$, and $X + Y \sim Gamma(\alpha + \beta, \theta)$.

6. Let $X \sim Uniform(0,1)$, then $-2\log X \sim \chi^2_2$.

7. (Poisson Process with Gamma Distribution) If $X \sim Gamma(\alpha, \beta)$ and $Y \sim Poisson\left(\frac{x}{\beta}\right)$ for any x, then $P(X \leq x) = P(Y \geq \alpha)$.

7.2 Large Sample Theory

This section was cited from my notes in MATH 357, the goal is for the reader to get familiar with the basic concepts in statistical inference. Definitions include: Convergence in Probability, convergence in distribution, CLT, WLLN, Slutsky's Theorem, Continuous Mapping Theorem, (First/Second) order Delta-Method. I am listing some of them:

Theorem

Theorem 33. *(Delta Methods)*
Let $\{X_n\}_1^\infty$ *be a sequence of random variables such that* $\mathbb{E}X_n = \mu$, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} v$ *as* $n \to \infty$ *and* g *be a real valued function such that* $g'(\mu)$ *exists and non-zero. Then*

$$\sqrt{n}(g(X_n 0 - g(\mu)) \xrightarrow{d} g'(\mu) \cdot V. \tag{7.8}$$

If $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, *and* $g'(\mu) = 0, g''(\mu) \neq 0$, *then we have*

$$n(g(X_n) - g(\mu)) \xrightarrow{d} \frac{1}{2}\sigma^2 g''(\mu)\chi_1^2. \tag{7.9}$$

Given a random sample $X_1, \cdots, X_n \sim f(x, \theta)$, a function $T(\boldsymbol{X}) = T(X_1, \cdots, X_n)$ is an estimator of θ, if it does not depend on the unknown parameter θ, such a $T(\boldsymbol{X})$ is also known as a statistic. The estimator is called an **unbiased estimator**, if $\mathbb{E}[T(\boldsymbol{X})] = \theta$.

Definition

Definition 17. *(Consistency of an Estimator)*
An estimator $T(\boldsymbol{X})$ *of* θ *is consistent, if* $T(\boldsymbol{X}) \xrightarrow{P} \theta$.

Exercise 1. *Consider the random sample* $X_1, \cdots, X_n \sim Uniform[0, \theta]$, *verify the consistency of the estimators* $T_1(\boldsymbol{X}) = 2\overline{X}_n, T_2(\boldsymbol{X}) = \frac{n+1}{n} X_{(n)}$. *Which one is better? If the random sample is changed to* $N(\mu, \sigma^2)$, *verify the consistency for* $\overline{X}_n$ *and* S_n^2.

Exercise 2. *Consider our favourite random sample (apart from* $\mathbf{GL}_3(\mathbb{Z}/2\mathbb{Z})$*), is* $X_1, \cdots, X_n \sim f(x, \theta) = e^{-(x-\theta)}, x \geq \theta$, *propose two different consistent estimator of* θ, *one based on* $X_{(1)}$, *one based on* $\overline{X}_n$. *Which one is better?*

In fact we have so many questions for this random sample in exercise 2! Like what is the sufficient statistic? Minimal sufficient statistic? UMVUE? MLE? Method of moment estimator?

In a random sample $X_1, \cdots, X_n \sim f(x, \theta)$ where θ is one-dimensional, we give the following **regularity conditions**:

(i) The family $\left\{ f(x, \theta) : \theta \in \Theta \subseteq \mathbb{R} \right\}$ has a common support that does not depend on θ.

(ii) $\dfrac{d}{d\theta} \log f(x, \theta)$ always exists.

(iii) For any statistic $h(\boldsymbol{X}) \in L^1$,

$$\frac{d}{d\theta}\int_{\mathscr{S}} h(\boldsymbol{x})f(\boldsymbol{x},\theta)dx = \int_{\mathscr{S}} h(\boldsymbol{x})\frac{d}{d\theta}f(\boldsymbol{x},\theta)dx \tag{7.10}$$

Theorem

Theorem 34. *(Crémer-Rao Lower Bound)*

Under regularly conditions, given a random sample $X_1,\cdots,X_n \sim f(x,\theta)$ with joint pdf denotes as $p_\theta(\boldsymbol{x})$, suppose $T(\boldsymbol{X})$ is an unbiased estimator for $\tau(\theta)$, then

$$Var(T(\boldsymbol{X})) \geq \frac{[\tau'(\theta)]^2}{\mathbb{E}\left\{\left(\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right)^2\right\}} \tag{7.11}$$

The quantity $\mathscr{I}_n = \mathbb{E}\left\{\left(\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right)^2\right\}$ is known as the Fisher information, and $\mathscr{I}_n = n\mathscr{I}_1$.

Exercise 3. *Given the random sample $X_1,\cdots,X_n \sim Bernoulli(\theta)$, what is the CRLB for all unbiased estimator of θ? Can you find an estimator such that the variance is attained at CRLB?*

Under regularity conditions, we have:

First Bartlett Identity:

$$\mathbb{E}\left\{\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right\} = 0, \forall\theta\in\Theta; \tag{7.12}$$

Second Bartlett Identity:

$$\mathbb{E}\left\{\left(\frac{d}{d\theta}\log p_\theta(\boldsymbol{x})\right)^2\right\} = -\mathbb{E}\left\{\frac{d^2}{d\theta^2}\log p_\theta(\boldsymbol{x})\right\}. \tag{7.13}$$

Theorem

Theorem 35. *(Conditions to attain CRLB)*

Suppose that a random sample $X_1,\cdots,X_n \sim p_\theta(\boldsymbol{x})$, and $T(\boldsymbol{X})$ be an unbiased estimator for $\tau(\theta)$. Then the variance of $T(\boldsymbol{X})$ attains the CRLB iff

$$a(\theta)\cdot\{T(\boldsymbol{X})-\tau(\theta)\} = \frac{d}{d\theta}\log p_\theta(\boldsymbol{x}) \tag{7.14}$$

for some function $a(\theta)$ for all $\theta\in\Theta$.

Definition

Definition 18. *(Exponential Family)*

A random sample $X_1, \cdots, X_n \sim f(x, \theta)$ is said to be of the exponential family, if the joint pdf $p_\theta(\boldsymbol{x})$ takes the form

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \cdot c(\theta) \cdot \exp\{\omega(\theta) T(\boldsymbol{X})\} \tag{7.15}$$

where $h(\boldsymbol{x})$ is a non-negative function of $\boldsymbol{x}$ and $c(\theta)$ a non-negative function of θ, and $T(\boldsymbol{X})$ is a function of $X_1, \cdots, X_n$, $\omega(\theta)$ could be any function of θ. Most importantly, the support $\boldsymbol{x} \in \mathscr{S}$ does not depend on θ. Most most importantly, in this family $\frac{1}{n}\sum T(\boldsymbol{X}_i)$ is the UMVUE of $\tau(\theta) = -\frac{c'(\theta)}{c(\theta)\omega'(\theta)}$.

Exercise 4. *Try to show that the random sample $X_1, \cdots, X_n \sim Poisson(\lambda)$ is of exponential family, and indeed the sample mean $\overline{X}_n$ is the UMVUE of λ.*

Theorem

Theorem 36. *(Neyman-Fisher Factorization Theorem)*

Let a random sample $X_1, \cdots, X_n \sim p_\theta(\boldsymbol{x})$, a statistics $T(\boldsymbol{X})$ is sufficient iff there exists functions g, h such that

$$p_\theta(\boldsymbol{x}) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x}) \tag{7.16}$$

for all $\theta \in \Theta$ and $\boldsymbol{x} \in \mathscr{X}$. Note that h must be a function that purely depends on $\boldsymbol{x}$, and g depends on $\boldsymbol{x}$ only through $T(\boldsymbol{x})$.

Exercise 5. *Find a sufficient statistic for the given random samples: $X_1, \cdots, X_m \sim Bernoulli(\theta)$, and $Y_1, \cdots, Y_n \sim Uniform[0, \theta]$, and $Z_1, \cdots, Z_l \sim e^{-(x-\theta)}, x \geq \theta$.*

Have you forgot the indicator function?

Exercise 6. *Now consider the normal family of random samples. If $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, then what is a sufficient statistic of (μ, σ^2)? If I have two mutually independent random samples $X_1, \cdots, X_m \sim N(\mu_1, \sigma^2)$ and $Y_1, \cdots, Y_n \sim N(\mu_2, \sigma^2)$, then what is a sufficient statistic of (μ_1, μ_2, σ^2)?*

Theorem

Theorem 37. *(Sufficiency under Transformation)*

If $T(\boldsymbol{X})$ is a sufficient statistics of θ, and we have $T(\boldsymbol{X}) = \varphi(T^*(\boldsymbol{X}))$ where φ is a measurable function and $T^*(\boldsymbol{X})$ is a statistic, then $T^*(\boldsymbol{X})$ is also sufficient. Or, under any one-to-one transformation g, $g(T(\boldsymbol{X})$ is also sufficient of $g(\theta)$.

Exercise 7. *Convince yourself, that for an exponential family, a sufficient statistic of θ can be $T(\boldsymbol{X}) = \sum X_i$. You will see later, that $T(\boldsymbol{X})$ is also complete!*

Theorem

Theorem 38. *(Lehmann-Scheffé for Minimum Sufficient Statistic)*

For a parametric family $p_\theta(\cdot)$, suppose a statistic $T(\boldsymbol{X})$ is such that $\forall x, y \in \mathscr{X}$, $T(\boldsymbol{x}) = T(\boldsymbol{y})$ iff $\dfrac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{y})}$ does not depend on θ, then $T(\boldsymbol{X})$ is a minimum sufficient statistic. In fact any one-to-one function of a minimum sufficient statistics is also minimum sufficient.

Exercise 8. *Again, consider the random samples $X_1, \cdots, X_m \sim Uniform[0, \theta]$ and $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$. What are the minimal sufficient statistic for those two?*

Definition

Definition 19. *(Completeness of Sufficient Statistic)*

Let X be a random variable with a pdf belonging to a parametric family $\mathscr{F} : \{f_\theta : \theta \in \Theta\}$. This family is said to be complete, if for any measurable function g with $\mathbb{E}[g(x)]$ exists, we have

$$\mathbb{E}[g(x)] = 0, \forall \theta \in \Theta \implies P(g(x) = 0) = 1 \text{ almost surely.} \tag{7.17}$$

A statistic $T(\boldsymbol{X})$ is complete, if its family of distributions is complete.

Theorem

Theorem 39. *(Rao-Blackwell)*

Let $U(\boldsymbol{X})$ be an unbiased estimator for $\tau(\theta)$, let $T(\boldsymbol{X})$ be a complete sufficient statistic for the parametric family, then

$$\mathbb{E}\{U(\boldsymbol{X}) | T(\boldsymbol{X}) = t\}, \forall t \in \mathscr{S}_T \tag{7.18}$$

is the UMVUE for $\tau(\theta)$. If the sufficient statistic is not complete, then conditioning on it will only get a better estimator, and it is not necessarily the UMVUE.

Theorem

Theorem 40. *(Lehmann-Scheffé Uniqueness Theorem)*

Let $T(\boldsymbol{X})$ be a complete sufficient statistic, also let $U(\boldsymbol{X}) = h(T(\boldsymbol{X}))$ for a measurable function h, be an unbiased estimator of $\tau(\theta)$ such that $\mathbb{E}\{U^2(\boldsymbol{X})\} < \infty$. Then $U(\boldsymbol{X})$ is the unique UMVUE.

Exercise 9. *Use Lehmann-Scheffé, find the UMVUE for each parameter of the following random samples:*

$$X_1, \cdots, X_n \sim Bernoulli(\theta), Poisson(\lambda), N(\mu, \sigma^2). \tag{7.19}$$

Exercise 10. *In the previous exercise, we had two random samples $X_1, \cdots, X_m \sim Bernoulli(\theta)$ and $Y_1, \cdots, Y_n \sim Poisson(\lambda)$. Now find the UMVUE for $\tau(\theta) = \theta(1-\theta)$, and $\tau(\lambda) = e^{-\lambda}$.*

Theorem

Theorem 41. *(Relation Between UMVUE and Other Estimators)*

An estimator $U(\boldsymbol{X})$ of $\tau(\theta)$ is the UMVUE iff it is un-correlated with all unbiased estimators of zero, that is,

$$Cov(U(\boldsymbol{X}), \delta(\boldsymbol{X})) = 0 \tag{7.20}$$

for all $\delta(\boldsymbol{X})$ such that $\mathbb{E}\{\delta(\boldsymbol{X})\} = 0$, $\delta(\boldsymbol{X})$ is called the unbiased estimator of zero.

Exercise 11. *(Make sure to look at this! Just in case Khalili really puts this question on the final)*

Let $X \sim Uniform\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right), \theta \in \mathbb{R}$. Then show that the UMVUE of $\tau(\theta)$ must be a constant function.

Assume $\delta(X)$ is an unbiased estimator of zero, hence $\mathbb{E}[\delta(X)] = 0$. So we have

$$\int_{\theta-1/2}^{\theta+1/2} \delta(X) = 0 \tag{7.21}$$

and by Fundamental Theorem of Calculus we have we have

$$\delta\left(\theta + \frac{1}{2}\right) - \delta\left(\theta - \frac{1}{2}\right) = \frac{d}{d\theta} \int_{\theta-1/2}^{\theta+1/2} \delta(X) = 0 \tag{7.22}$$

Hence we have $\delta(x) = \delta(x+1)$. Now let $U(X)$ be the UMVUE of $\tau(\theta)$, then it is easy to see that $U(X)\delta(X)$ is also an unbiased estimator of zero, and by the previous theorem, we have

$$Cov(U(X), \delta(X)) = \mathbb{E}\{U(X)\delta(X)\} = 0 \tag{7.23}$$

So now we actually have $U(x)\delta(x) = U(x+1)\delta(x+1)$. Hence $U(x) = U(x+1)$, and by definition,

$$\mathbb{E}[U(X)] = \int_{\theta-1/2}^{\theta+1/2} U(X)dx = \tau(\theta) \tag{7.24}$$

and we use Fundamental Theorem of calculus to get

$$U\left(\theta + \frac{1}{2}\right) - U\left(\theta - \frac{1}{2}\right) = \tau'(\theta) \tag{7.25}$$

and we know that $\tau'(\theta) = 0$, meaning $\tau(\theta) = C$ where it is a constant! So the UMVUE of $\tau(\theta)$ must be a constant.

Theorem

Theorem 42. *(Method of Moment Estimator)*

The set up is that, we match the first k moments of $f(x,\boldsymbol{\theta})$ where $\dim(\boldsymbol{\theta}) = k$ with the first k population mean, if exists. We set them to be equal to set our estimate. i.e we solve the linear system

$$\begin{aligned} \frac{X_1 + \cdots + X_n}{n} &\overset{set}{=} \mathbb{E}X \\ \frac{X_1^2 + \cdots + X_n^2}{n} &\overset{set}{=} \mathbb{E}X^2 \\ &\vdots \\ \frac{X_1^k + \cdots + X_n^k}{n} &\overset{set}{=} \mathbb{E}X^k \end{aligned}$$

Exercise 12. *You better get it, it is so easy. One remark is that is may not always exist, when $\mathbb{E}X = \infty$, like the random sample with pdf $f(x,\theta) = x^{-2}\theta$, where you will see why soon. In a special case where we have a random sample of $Uniform[-\theta,\theta]$, we see that $\mathbb{E}X = 0$ and so we match the second moment.*

Theorem

Theorem 43. *(Bayesian Estimation)*

Let $\pi(\theta)$ reflect the prior distribution of the unknown parameter θ and $p(\boldsymbol{x}|\theta)$ is the joint pdf of the random sample, then an estimate of θ can be done by

$$\pi(\theta|\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}) \cdot \pi(\theta)}{\int_\Theta p_\theta(\boldsymbol{x}) \cdot \pi(\theta) d\theta} \tag{7.26}$$

where $\pi(\theta|\boldsymbol{x})$ is the posterior distribution. Also under the squared error loss function condition, the Bayes estimate of θ is given by $\mathbb{E}[\pi(\theta|\boldsymbol{x})]$. To find $\pi(\theta|x)$, we may just ignore all the constants as well as the denominator, they are all known as the "normalizing constants", then we may refer to the table to find the distribution.

Exercise 13. *Consider the random sample $X_1, \cdots, X_n \sim Bernoulli(\theta)$, and the prior distribution for θ is given by*

$$\pi(\theta) \sim Beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \alpha,\beta > 0 \tag{7.27}$$

Find the Bayes estimator of θ under the squared error loss.

Exercise 14. *Consider the random sample $X_1, \cdots, X_n \sim Uniform(0,\theta)$, and the prior distribution for θ is given by*

$$\pi(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}, \alpha,\beta > 0; \theta > 1 \tag{7.28}$$

Find the Bayes estimator of θ under the squared error loss. You can find the answer to this question in Q9 of the next part.

Theorem

Theorem 44. *(Invariance of MLE)*

If $\widehat{\theta}_n$ is the MLE of θ, then for any function $\tau(\theta)$, $\tau(\widehat{\theta}_n)$ is the MLE for $\tau(\theta)$.

Theorem

Theorem 45. *(Asymptotic Normality of MLE)*

Under all the regularity conditions, if $\widehat{\theta}_n$ is the MLE for θ, then

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathscr{I}_1^{-1}(\theta)). \tag{7.29}$$

Exercise 15. *Write a joke such that it contains at least* 4 *out of the following words:* 1 $\mathbf{GL}_3(\mathbb{F}_2)$*;* 2 *MIT;* 3 *Cows;* 4 *Electric Cars;* 5 *Missile;* 6 *Trivial.*

Finally, here are some distributions, that you might encounter, make sure you know everything about them!!!

- $Bernoulli(\theta)$ **Distribution:**

The pmf is $f(x,\theta) = \theta^x(1-\theta)^{1-x}, 0 < \theta < 1, x\ \{0,1\}$, where θ usually represents the probability of success. The sufficient statistic of θ is $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$, the MLE of θ is $\overline{X}_n$, the Fisher information (for one distribution) is $\mathscr{I}_1(\theta) = \dfrac{1}{\theta(1-\theta)}$, the UMVUE of θ^k is $\binom{n-k}{t-k}/\binom{n}{t}$, where $t = \sum X_i$ is the sufficient statistic!

- $Poisson(\lambda)$ **Distribution:**

The pmf is $f(x,\lambda) = e^{-\lambda}\dfrac{\lambda^x}{x!}, x = 0,1,2,\cdots; \lambda > 0$. The sufficient statistic of θ is $T(\boldsymbol{X}) = \sum X_i$, the MLE of θ is $\overline{X}_n$, the Fisher information (for one distribution) is $\mathscr{I}_1(\lambda) = \dfrac{1}{\lambda}$, the UMVUE of λ^k is $\dfrac{T(T-1)\cdots(T-k+1)}{n^k}$ where $T = \sum X_i$ is the sufficient statistic.

- $Geometric(\theta)$ **Distribution:**

The pmf is $f(x,\theta) = \theta(1-\theta)^{x-1}, x = 1,2,\cdots; 0 < \theta < 1$. The sufficient statistic is $T(\boldsymbol{X}) = \sum X_i$, the MLE is $n/\sum X_i$, the fisher information (for one distribution) is $\mathscr{I}_1(\theta) = \dfrac{1}{\theta^2(1-\theta)}$, and the UMVUE of θ is $\binom{t-2}{n-2}/\binom{t-1}{n-1}$ where $t = \sum X_i$ is the sufficient statistic.

- $Uniform(0,\theta)$ **Distribution:**

The pdf is $f(x,\theta) = \dfrac{1}{\theta} \cdot \chi\{0 < X < \theta\}$, a sufficient statistic is $X_{(n)}$, the MLE is $X_{(n)}$, the UMVUE of θ^k is $\dfrac{n+k}{n} X_{(n)}^k$.

- **Shifted Exponential Distribution:**

The pdf is $f(x,\theta) = e^{-(x-\theta)} \cdot \chi\{x > \theta\}$, in this family we have $\mathbb{E}X = \theta + 1$ and $Var(X) = 1$, and a sufficient statistic is $X_{(1)}$, the cdf of $X_{(1)}$ is $F(x) = 1 - e^{n(\theta-x)}$ and the pdf of $X_{(1)}$ is $f(x) = ne^{n(\theta-x)}$, the UMVUE of θ is $X_{(1)} - \frac{1}{n}$.

- **Exponential Distribution:**

The pdf is $f(x,\theta) = \theta e^{-\theta x}, x > 0$, in this family we have $\mathbb{E}X = \frac{1}{\theta}$, and $Var(X) = \frac{1}{\theta^2}$, a sufficient statistic is $T(\boldsymbol{X}) = \sum X_i$, the MLE is $\frac{n}{\sum X_i}$, the Fisher information (for one distribution) is $\mathscr{I}_1(\theta) = \frac{1}{\theta^2}$, the UMVUE of θ is $\frac{n-1}{\sum X_i}$. In this family, a well known integral is important:

$$\int_0^\infty x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \Gamma(\alpha)\beta^\alpha \tag{7.30}$$

Furthermore, you can see the UMVUE of $\theta^k, k < n$ is

$$\frac{\Gamma(n)}{\Gamma(n-k)} \cdot \frac{1}{n^k} \cdot \frac{n}{\sum X_i}. \tag{7.31}$$

7.3 Common Types of Confidence Intervals in $\mathbb{R}$

This section was cited from my notes in MATH 357, the goal is for the reader to get familiar with the basic concepts in confidence intervals.

Definition

Definition 20. *(Interval Estimator)*

Let $L(\boldsymbol{X}), U(\boldsymbol{X})$ be two statistics such that $L(\boldsymbol{x}) \le U(\boldsymbol{x}), \forall x \in \mathscr{X}$. A random interval $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ is called an interval estimator or confidence interval with confidence level $1-\alpha, \alpha \in (0,1)$ if $P(L(\boldsymbol{X}) \le \theta \le U(\boldsymbol{X})) = 1 - \alpha$.

Note that it is **wrong** to say that $(L(\boldsymbol{x}), U(\boldsymbol{x}))$ (post-experimental data) captures θ with probability $1-\alpha$. The interpretation is that *this interval estimator either includes θ or not, basically it captures θ with probability* 0 *or* 1. *if we were to repeat the experiment and compute similar confidence intervals for θ, we expect that* $100(1-\alpha)\%$ *of those post-experimental intervals to capture θ.*

Definition

Definition 21. *(Pivot Quantity)*

A random function $Q(\boldsymbol{x},\theta)$ is called a pivot quantity (PQ) iff its distribution does not depend on the parameter θ, and Q is a function of $\boldsymbol{X}$ and θ only.

Note that the function Q might include θ, but its overall distribution $Q(\boldsymbol{x},\theta)$ is free of any parameters! For example if $X \sim N(0,\theta^2)$, then $Q(x,\theta) = \frac{1}{\theta}X \sim N(0,1)$, which is free of any parameter and it is a known distribution!

Once we have a PQ, and the confidence level $1-\alpha$ is given, we may find constants c_1, c_2 such that

$$P(c_1 \le Q(\boldsymbol{x},\theta) \le c_2) = 1-\alpha, \tag{7.32}$$

then having c_1, c_2 we can solve in terms of θ to get

$$P(L(\boldsymbol{X}) \le \theta \le U(\boldsymbol{X})) = 1-\alpha. \tag{7.33}$$

Mostly, the PQ is chosen based on a sufficient statistic.
Below are common random samples and their corresponding PQ and confidence intervals:

(i) **Confidence Interval for μ in a normal family**:

• If σ^2 is known, then

$$Q(\boldsymbol{X},\mu) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1) \tag{7.34}$$

and the $100(1-\alpha)\%$ C.I is given by

$$\left(\overline{X}_n - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) \tag{7.35}$$

• If σ^2 is unknown, then

$$Q(\boldsymbol{X},\mu) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t(n-1) \tag{7.36}$$

and the $100(1-\alpha)\%$ C.I is given by

$$\left(\overline{X}_n - t(n-1,\alpha/2)\cdot\frac{S_n}{\sqrt{n}}, \overline{X}_n + t(n-1,\alpha/2)\cdot\frac{S_n}{\sqrt{n}}\right) \tag{7.37}$$

Note: z_p is called the p-th quantile in a standard normal distribution, and $P(Z < z_p) = p$, where $Z \sim N(0,1)$. The same idea applies in $t(n-1,\alpha/2)$, where $n-1$ indicates the degrees of freedom, and $\alpha/2$ is the quantle.

Exercise 16. *A manufacturer developed a new gunpowder and tested it in eight shells. The resulting muzzle velocities, in feet per second, were:*

$$3005, 2925, 2935, 2965, 2995, 3005, 2937, 2905. \tag{7.38}$$

Assume that the velocities are iid sample from a $N(\mu, \sigma^2)$. Compute a 95% confidence interval for μ. Provide interpretation for your interval.

So in this example, we see that σ^2 is unknown. So we replace it by S_n. Note that the S_n of this sample can be calculated by

$$S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{7}\sum_{i=1}^{8}(X_i - \overline{X_n})^2} \approx 39.09 \tag{7.39}$$

From the C.I above, we construct

$$\left(2959 - t(7, 0.025) \cdot \frac{39.09}{\sqrt{8}}, 2959 + t(7, 0.025) \cdot \frac{39.09}{\sqrt{8}}\right). \tag{7.40}$$

From the t-table, $t(7, 0.025) = 2.365$, so for the 95% confidence interval, our final answer is $(2926.38, 2991.62)$. The interpretation is that, *this interval estimator either includes θ or not, basically it captures θ with probability* 0 *or* 1. *if we were to repeat the experiment and compute similar confidence intervals for μ, we expect that* $100(1-\alpha)\%$ *of those post-experimental intervals to capture μ.*

(ii) **Confidence interval for σ^2 in a normal family:**

- If μ is unknown, then

$$Q(\boldsymbol{X}, \sigma^2) = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1} \tag{7.41}$$

and the $100(1-\alpha)\%$ C.I is given by

$$\left(\frac{(n-1)S_n^2}{\chi^2_{(n-1,\alpha/2)}}, \frac{(n-1)S_n^2}{\chi^2_{(n-1,1-\alpha/2)}}\right). \tag{7.42}$$

- if μ is known, then

$$Q(\boldsymbol{X}, \sigma^2) = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n \tag{7.43}$$

and the $100(1-\alpha)\%$ C.I is given by

$$\left(\frac{(n-1)\sum(X_i - \mu)^2}{\chi^2_{(n,\alpha/2)}}, \frac{(n-1)\sum(X_i - \mu)^2}{\chi^2_{(n,1-\alpha/2)}}\right) \tag{7.44}$$

Exercise 17. *Assume that the number of days needed to hatch an egg of a certain type of a rare lizard is distributed Normally. Using incubator,* 13 *eggs from different nests separately hatched. The sample mean is* 18.97 *weeks and the sample standard deviation is* $\sqrt{10.7}$ *weeks. Find a* 90% *confidence interval for the population variance. Provide interpretation for the interval.*

Here, it is the case that both μ, σ^2 are unknown, so we construct the C.I based on

$$\left(\frac{(n-1)S_n^2}{\chi^2_{(n-1,\alpha/2)}}, \frac{(n-1)S_n^2}{\chi^2_{(n-1,1-\alpha/2)}}\right) = \left(\frac{12 \cdot 10.7}{\chi^2_{(12,0.05)}}, \frac{12 \cdot 10.7}{\chi^2_{(12,0.95)}}\right) = (6.107, 24.569). \tag{7.45}$$

(iii) **Confidence Interval Approximation for μ of non-normal large sample:**

We will assume that in a large random sample (*usually $n > 25$ is good enough*) we have $X_1, \cdots, X_n \sim f$, $\mathbb{E}X = \mu$, and $VarX = \sigma^2$, $EX^4 < \infty$, and we consider the asymptotic approximate C.I for μ, as $n \to \infty$.

• If σ is known, then we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1) \tag{7.46}$$

so the $100(1-\alpha)\%$ C.I approximate is

$$\left(\overline{X}_n - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right). \tag{7.47}$$

• If σ is unknown, then we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1) \tag{7.48}$$

so the $100(1-\alpha)\%$ C.I approximate is

$$\left(\overline{X}_n - z_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}\right). \tag{7.49}$$

Exercise 18. *Shopping times of* 64 *randomly selected customers in a supermarket averaged* 33 *minutes with a standard deviation of* 16 *minutes. Construct an approximate* 90% *confidence interval for the mean shopping time per customer. Provide interpretation for the interval.*

Standard deviation is S_n, not the square!!!

Here, if we use large sample theory and the central limit theorem, we will have

$$\frac{\sqrt{64}(33 - \mu)}{16} \xrightarrow{d} N(0,1) \tag{7.50}$$

Here $\alpha = 0.1$, so we look at the normal table and see that $z_{\alpha/2} = z_{0.05} \approx 1.64$, hence we have the C.I

$$\left(33 - 1.64 \times \frac{16}{8}, 33 + 1.64 \times \frac{15}{8}\right) = (29.72, 36.28). \tag{7.51}$$

Remark: *Since we are interested in $z_{0.05}$. if we can not find the exact value, i.e we know $z_{0.0495} = 1.65$ and $z_{0.505=1.64}$, we then may take the average to get a better approximate of $z_{0.05} \approx 1.645$.*

(iv) **Mean Difference in Two Mutually Independent Normal Random Samples:**

Given two mutually independent random samples $X_1, \cdots, X_m \sim N(\mu_1, \sigma^2)$ and $Y_1, \cdots, Y_n \sim N(\mu_2, \sigma^2)$, we are interested in constructing the confidence interval for $\mu_1 - \mu_2$. The PQ is

$$Q(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\mu}) = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2) \tag{7.52}$$

where

$$S^2 = \frac{1}{m+n-2}\left\{\sum_{i=1}^{m}(X_i - \overline{X}_m)^2 + \sum_{j=1}^{n}(Y_i - \overline{Y}_n)^2\right\} \tag{7.53}$$

so the $100(1-\alpha)\%$ C.I is given by

$$\left((\overline{X}_m - \overline{Y}_n) - t(m+n-2, \alpha/2) \cdot S\sqrt{\frac{1}{m} + \frac{1}{n}}, (\overline{X}_m - \overline{Y}_n) + t(n+m-2, \alpha/2) \cdot S\sqrt{\frac{1}{m} + \frac{1}{n}}\right) \tag{7.54}$$

Exercise 19. *In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results in seconds are:*

$$\begin{aligned} &\textit{old}: 42.7, 43.8, 42.5, 43.1, 44.0, 43.6, 43.3, 43.5, 41.7, 44.1; \\ &\textit{new}: 42.1, 41.3, 42.4, 43.2, 41.8, 41.0, 41.8, 42.8, 42.3, 42.7. \end{aligned}$$

Construct a 95% confidence interval for the difference in the respective means. Provide interpretation for the interval. (Assume that the timings for the old and new machines are independent i.i.d samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, and $\sigma_1 = \sigma_2$.

Here, denote X_i to be the old sample, and Y_j to be the new sample. Then from the data provided we can compute that $\overline{X} = 43.23$ and $\overline{Y} = 42.14$, so $\overline{X} - \overline{Y} = 1.09$, and

$$S^2 = \frac{1}{18}\left\{\sum_{i=1}^{10}(\overline{X}_i - 43.23)^2 + \sum_{j=1}^{10}(\overline{Y}_j - 42.14)^2\right\} \tag{7.55}$$

(v) **Mean Difference in Two Mutually Independent Non-normal Random Samples:**

This set up is roughly the same as the previous one, but here we removed the normal restriction, and we use central limit theorem to get the approximate of C.I for large n, m. We know that the PQ is

$$Q(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \xrightarrow{d} N(0,1). \tag{7.56}$$

and the $100(1-\alpha)\%$ C.I is given by

$$\left(\overline{X}_m - \overline{Y}_n - z_{\alpha/2} \cdot \sqrt{\frac{S_m^2}{m} + \frac{S_n^2}{n}}, \overline{X}_m - \overline{Y}_n + z_{\alpha/2} \cdot \sqrt{\frac{S_m^2}{m} + \frac{S_n^2}{n}}\right) \tag{7.57}$$

Note that as $n \to \infty$, $S_m^2 \xrightarrow{P} \sigma_1^2$ and $S_n^2 \xrightarrow{P} \sigma_2^2$, so we can swap to whichever is easier for us to compute.

Exercise 20. *We wish to compare the daily intake of selenium in two regions. In each region,* 30 *adults were tested and the results (in mg/day) were:* $\bar{x}_n = 167.1, s_n = 24.3, \bar{y}_m = 140.9, s_m = 17.6$ *Find a* 95% *approximate confidence interval for the difference in mean daily intake of selenium in the two regions. Provide interpretation for the interval.*

This is the case where we will find the C.I for the difference of mean in two non-normal random samples. So we apply the above formula, we know that $\overline{X}_m - \overline{Y}_n = 26.2$, $\sqrt{\frac{s_m^2}{m} + \frac{s_n^2}{n}} \approx 5.477986$ and here $\alpha = 0.05$, so $z_{\alpha/2} = z_{0.025}$, and hence we have the 95% C.I to be $26.2 \pm z_{0.025} \cdot 5.477986$. From the normal table, $z_{0.025} = 1.96$, so finally we have $(15.46315, 36.93685)$.

(vi) **Population Proportion:**

Now consider we have a random sample $X_1, \cdots, X_n \sim Bernoulli(\theta)$, then we know that by central limit theorem, a PQ is

$$Q(\boldsymbol{X}, \theta) = \frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \xrightarrow{d} N(0,1) \tag{7.58}$$

also for large sample, we have $\overline{X}_n \xrightarrow{P} \theta$, denote $\widehat{\theta}_n = \overline{X}_n$, we also have an alternative PQ

$$Q(\boldsymbol{X}, \theta) = \frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}} \xrightarrow{d} N(0,1). \tag{7.59}$$

The advantage for this PQ is that we can easily separate θ and get the $100(1-\alpha)\%$ C.I:

$$\left(\widehat{\theta}_n - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}, \widehat{\theta}_n + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\right) \tag{7.60}$$

Exercise 21. *A sample of* $n = 1000$ *voters, randomly selected from a city, showed* 560 *in favour of candidate Jones. Find an approximate* 99% *confidence interval for the population proportion in favour of candidate Jones. Provide interpretation for the interval.*

Here we can easily see that in this Bernoulli random sample, we have $\widehat{\theta}_n = \overline{X}_n = 0.56$, hence we construct the C.I by

$$\left(0.56 - z_{0.005} \cdot \sqrt{\frac{0.56 \cdot 0.44}{1000}}, 0.56 + z_{0.005} \cdot \sqrt{\frac{0.56 \cdot 0.44}{1000}}\right) = 0.56 \pm z_{0.005} \cdot 0.0157 \tag{7.61}$$

From the normal table, we see that $z_{0.005} \approx 2.575$.

(vii) **Difference in Two Population Proportion:**

Here, we have two mutually independent random samples $X_1, \cdots, X_m \sim Bernoulli(\theta_1)$ and $Y_1, \cdots, Y_n \sim Bernoulli(\theta_2)$. We are interested in the C.I of $\theta_1 - \theta_2$. Similarly, denote $\widehat{\theta}_1 = \overline{X}_m, \widehat{\theta}_2 = \overline{Y}_n$, and the improved PQ is

$$Q(\boldsymbol{X},\boldsymbol{Y},\theta)=\frac{(\widehat{\theta}_1-\theta_1)-(\widehat{\theta}_2-\theta_2)}{\sqrt{\widehat{\theta}_1(1-\widehat{\theta}_1)/m+\widehat{\theta}_2(1-\widehat{\theta}_2)/n}}\xrightarrow{d}N(0,1). \tag{7.62}$$

and the $100(1-\alpha)\%$ C.I is now

$$\left((\widehat{\theta}_1-\widehat{\theta}_2)-z_{\alpha/2}\cdot\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{m}+\frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n}},(\widehat{\theta}_1-\widehat{\theta}_2)+z_{\alpha/2}\cdot\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{m}+\frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n}}\right) \tag{7.63}$$

Exercise 22. *A medical researcher conjectures that smoking can result in wrinkled skin around the eyes. The researcher recruited* 150 *smokers and* 250 *nonsmokers to take part in an observational study and found that* 95 *of the smokers and* 105 *of the nonsmokers were seen to have prominent wrinkles around the eyes (based on a standardized wrinkle score administered by a person who did not know if the subject smoked or not). Find an approximate* 95% *confidence interval for the difference in the proportions of people who have wrinkled skin around their eyes in the two populations. Provide interpretation for the interval.*

Here, we have two independent random samples, let $X_1,\cdots,X_{150}\sim Bernoulli(\theta_1)$ to be the smokers sample and $Y_1,\cdots,Y_{250}\sim Bernoulli(\theta_2)$ to be the nonsmokers sample, where θ_1,θ_2 denotes the proportion of populations who have wrinkled skin, thus $\widehat{\theta}_1=95/150=0.633$ and $\widehat{\theta}_2=105/250=0.42$. Also in this case we have $\alpha=0.05$ hence $\alpha/2=0.025$ and from the normal table we have $z_{\alpha/2}=z_{0.025}=1.96$, also $\widehat{\theta}_1-\widehat{\theta}_2=0.213$, $\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{m}+\frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n}}\approx 0.0502$, and thus our 95% C.I is $(0.1146,0.3114)$.

(viii) **Approximate Using MLE Theory:**

Recall that, let a random sample $X_1,\cdots,X_n\sim f(x,\theta)$ where θ is a one-dimensional parameter, and $\widehat{\theta}_n$ to be the MLE of θ, then we know that

$$\sqrt{n}(\widehat{\theta}_n-\theta)\xrightarrow{d}N\left(0,\mathscr{I}_1^{-1}(\widehat{\theta}_n)\right) \tag{7.64}$$

One remark is that since we have convergence in probability $\widehat{\theta}_n\xrightarrow{P}\theta$, so we can replaced by $\widehat{\theta}_n$ in Fisher information, which would be easier to separate θ in the calculation.

So we have the $100(1-\alpha)\%$ given by

$$\left(\widehat{\theta}_n-z_{\alpha/2}\cdot\sqrt{\frac{1}{n}[\mathscr{I}_1(\widehat{\theta}_n)]^{-1}},\widehat{\theta}_n+z_{\alpha/2}\cdot\sqrt{\frac{1}{n}[\mathscr{I}(\widehat{\theta}_n)]^{-1}}\right). \tag{7.65}$$

Also we may use the "empirical Fisher" to get our estimate of $\mathscr{I}_1(\widehat{\theta})$:

$$\mathscr{I}_1(\theta)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log f(x_i,\theta)\Big|_{\theta=\widehat{\theta}_{MLE}}\right)^2\overset{\text{under regularity conditions}}{=}-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta)\Big|_{\theta=\widehat{\theta}_{MLE}}\right) \tag{7.66}$$

Recall the delta-method and the invariance of MLE. If $\widehat{\theta}_n$ is the MLE of θ, and then for any function g, $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$, and the first order delta-method shows us that if $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathscr{I}_1^{-1}(\widehat{\theta}_n))$, then

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2 \cdot \mathscr{I}_1^{-1}(\widehat{\theta}_n)). \tag{7.67}$$

and the $100(1-\alpha)\%$ C.I approximate is

$$\left(g(\widehat{\theta}_n) - z_{\alpha/2} \cdot \sqrt{\frac{1}{n} \mathscr{I}_1^{-1}(\widehat{\theta}_n) \cdot |g'(\widehat{\theta}_n)|^2} \ , \ g(\widehat{\theta}_n) + z_{\alpha/2} \cdot \sqrt{\frac{1}{n} \mathscr{I}_1^{-1}(\widehat{\theta}_n) \cdot |g'(\widehat{\theta}_n)|^2} \right) \tag{7.68}$$

Exercise 23. *Suppose a random sample $X_1, \cdots, X_n \sim Poisson(\lambda)$ and λ is the unknown parameter. Using the MLE theory, construct a $100(1-\alpha)\%$ approximate two-sided confidence interval for λ. Then find the $100(1-\alpha)\%$ C.I for λ^2, and $e^{-\lambda}$.*

Here, we can easily see that the MLE of a Poisson random sample is just the sample mean, $\widehat{\theta}_n = \overline{X}_n$. Now we find the Fisher information. The log-pdf of one sample is

$$\log f(X = k, \lambda) = \log e^{-\lambda} \frac{\lambda^k}{k!} = -\lambda + k \log \lambda - \log k!. \tag{7.69}$$

The second partial derivative is

$$\frac{\partial^2 \log f(X = k, \lambda)}{\partial \lambda^2} = -k \frac{1}{\lambda^2} \tag{7.70}$$

and hence we have

$$\mathscr{I}_1(\lambda) = -\mathbb{E}\left\{ -X \frac{1}{\lambda^2} \right\} = \frac{1}{\lambda}, \mathscr{I}_1^{-1}(\lambda) = \lambda \tag{7.71}$$

So the MLE theory says that

$$\sqrt{n}\left(\overline{X}_n - \lambda\right) \xrightarrow{d} N\left(0, \mathscr{I}_1^{-1}(\widehat{\theta}_n)\right) \tag{7.72}$$

and given that $\mathscr{I}_1^{-1}(\widehat{\theta}) = \overline{X}_n$, hence the $100(1-\alpha)\%$ C.I is

$$\left(\overline{X}_n - z_{\alpha/2} \cdot \sqrt{\frac{1}{n} \cdot \overline{X}_n} \ , \ \overline{X}_n + z_{\alpha/2} \cdot \sqrt{\frac{1}{n} \cdot \overline{X}_n} \right). \tag{7.73}$$

For λ^2 and $e^{-\lambda}$, we will then apply delta-method to find the C.I, we define $g(\lambda) = \lambda^2$, and $h(\lambda) = e^{-\lambda}$, hence it is easy to see that $\widehat{\lambda}^2_{MLE} = \overline{X}_n^2$ and $e^{-\lambda}_{MLE} = e^{-\overline{X}_n}$. Then delta-method says that

$$\sqrt{n}(\overline{X}_n^2 - \lambda^2) \xrightarrow{d} N\left(0, 4\lambda^2 \cdot \overline{X}_n\right) \tag{7.74}$$

which is,

$$\frac{\sqrt{n}(\overline{X}_n^2 - \lambda^2)}{2\lambda\sqrt{\overline{X}_n}} \xrightarrow{d} N(0,1) \tag{7.75}$$

use our large sample theory, we have $\overline{X}_n \xrightarrow{P} \lambda$, so we replace the λ in the denominator by $\overline{X}_n$, and we get the C.I as

$$\left(\overline{X}_n^2 - z_{\alpha/2} \cdot 2\overline{X}_n \sqrt{\frac{\overline{X}_n}{n}}, \overline{X}_n^2 + z_{\alpha/2} \cdot 2\overline{X}_n \sqrt{\frac{\overline{X}_n}{n}}\right). \tag{7.76}$$

Similarly, in $e^{-\lambda}$, we have

$$\frac{\sqrt{n}(e^{-\overline{X}_n} - e^{-\lambda})}{\lambda e^{-\lambda} \cdot \sqrt{e^{-\overline{X}_n}}} \xrightarrow{d} N(0,1) \tag{7.77}$$

and the C.I is

$$\left(e^{-\overline{X}_n} - z_{\alpha/2} \cdot 2e^{-\overline{X}_n} \cdot e^{-e^{\overline{X}_n}} \cdot \sqrt{\frac{e^{-\overline{X}_n}}{n}}\ ,\ e^{-\overline{X}_n} + z_{\alpha/2} \cdot 2e^{-\overline{X}_n} \cdot e^{-e^{\overline{X}_n}} \cdot \sqrt{\frac{e^{-\overline{X}_n}}{n}}\right). \tag{7.78}$$

7.4 Univariate Hypothesis Testing

This section was cited from my notes in MATH 357, the goal is for the reader to get familiar with the basic concepts in univariate hypothesis testing.

A statistical hypothesis test is a decision rule that uses the data to infer which two mutually exclusive hypothesis, that reflect two completing hypothetical states of the nature is correct. The decision rule partitions the sample space $\mathscr{X}$ into 2 disjoint regions that respectively reflect support for the null hypothesis $\mathscr{H}_0$ and the alternative hypothesis $\mathscr{H}_1$, where $\mathscr{X} = \mathscr{H}_0 \bigcup \mathscr{H}_1$ and $\mathscr{H}_0 \bigcap \mathscr{H}_1 = \varnothing$. Our goal would be to use the data to decide whether the parameter of interest θ is whether $\theta \in \mathscr{H}_0$ or $\theta \in \mathscr{H}_1$. Unlike point estimation and confidence interval, we do not perform any estimate on θ. In the field of machine learning, it is referred as *classification*.

Definition

Definition 22 (Test Function)**.** *Suppose a test $\mathscr{H}_0$ and $\mathscr{H}_1$ partitioning the sample space $\mathscr{X}$ into two disjoint regions $\mathscr{R}$ and $\mathscr{R}^C$, and we will reject $\mathscr{H}_0$ if $\boldsymbol{x} \in \mathscr{R}$, such $\mathscr{R}$ is called the critical region. We may formulate the test as a function:*

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in \mathscr{R}, \text{and we reject } \mathscr{H}_0 \\ 0 & \text{if } \boldsymbol{x} \in \mathscr{R}^C, \text{and we do not reject } \mathscr{H}_0 \end{cases} \tag{7.79}$$

There are two types of errors: Type one error is when $\mathscr{H}_0$ is rejected when $\mathscr{H}_0$ is indeed true; Type two error is $\mathscr{H}_0$ is not rejected when $\mathscr{H}_1$ is indeed true.

Theorem

Theorem 46 (Neyman-Pearson Lemma)**.** *Let* $\varphi(\boldsymbol{X}) = \begin{cases} 1 & \text{if } p(\boldsymbol{x},\theta_1) > kp(x,\theta_0) \\ \\ 0 & \text{if } p(\boldsymbol{x},\theta_1) < kp(\boldsymbol{x},\theta_0) \end{cases}$*, and* k *is such that* $P(\text{rejecting } \mathscr{H}_0) = \alpha$*, (we reject* $\mathscr{H}_0$ *if* $\varphi(\boldsymbol{X}) = 1$*)then* φ *is the UMP test in the class of all tests* φ^* *with the same level* α*,* α *is called the significance level. Hence the UMP test has a rejection region*

$$\mathscr{R} := \left\{ \boldsymbol{x} \in \mathscr{X}, \frac{p(\boldsymbol{x},\mathscr{H}_1)}{p(\boldsymbol{x},\mathscr{H}_0)} > k \right\}, \ k \text{ is chosen such that } P(\boldsymbol{x} \in \mathscr{R}) \leq \alpha. \tag{7.80}$$

Exercise 24. *Suppose we have a random sample* $X_1,\cdots,X_n \sim N(\mu,1)$*, and* $\mu = \{0,1\}$*. So we would like to test that* $\mathscr{H}_0 : \mu = 0$ *and* $\mathscr{H}_1 : \mu = 1$*.*

To use NP lemma, we first construct the ratio:

$$\frac{p(\boldsymbol{x},\mu=1)}{p(\boldsymbol{x},\mu=0)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum(x_i-1)^2\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum(x_i)^2\right\}} = e^{-\frac{1}{2}\sum(x_i-1)^2+\frac{1}{2}\sum x_i^2} = e^{n\overline{X}_n - n/2}$$

Then the NP lemma says that we will reject $\mathscr{H}_0$ if $\dfrac{p(\boldsymbol{x},\mu=1)}{p(\boldsymbol{x},\mu=0)} > k$ for some k, so we solve for $e^{n\overline{X}_n - n/2} > k$, and we get $\bar{x}_n > k^* = \frac{\ln k}{n} + \frac{1}{2}$, and this is the rejection region. Now given significance level α, the type one error is given by

$$P(\text{Reject } \mathscr{H}_0 \text{ when it is true}) = P(\bar{x}_n > k \big| \mu = 0) = \alpha. \tag{7.81}$$

Then we use the fact that $\overline{X}_n \sim N\left(0, \dfrac{1}{n}\right)$ (the case when $\mathscr{H}_0$ is true), and we have

$$P\left(\frac{\sqrt{n}(\overline{X}_n - 0)}{1} > \sqrt{n}k^*\right) = \alpha \tag{7.82}$$

and then we can refer to the normal table to solve for k^*. The same idea applies for type two error, where in this case we have

$$P(\text{Not rejecting } \mathscr{H}_0 \text{ when it is false}) = P(\bar{x}_n < k^* \big| \mu = 1). \tag{7.83}$$

Theorem

Theorem 47 (Likelihood Ratio (LR) Test)**.** *Consider the random sample $X_1,\cdots,X_n \sim f(x,\theta)$, and we have $\mathscr{H}_0$ and $\mathscr{H}_1$ as two tests, and we define the likelihood ratio (LR) statistic to be*

$$\lambda_n(\boldsymbol{X}) = \frac{L_n(\widehat{\theta}_{MLE,\mathscr{H}_0})}{L_n(\widehat{\theta}_{MLE,\Theta})} \tag{7.84}$$

A test based on LR statistic has the following form

$$\varphi(\boldsymbol{X}) = \begin{cases} 1 & \lambda_n(\boldsymbol{X}) < C \text{ (reject } \mathscr{H}_0\text{)} \\ 0 & \lambda_n(\boldsymbol{X}) > C \end{cases} \tag{7.85}$$

for $C \in [0,1]$ and the rejection region takes the form $\mathscr{R} := \{\boldsymbol{x} \in \mathscr{X} : \lambda_n(\boldsymbol{X}) < C\}$.

Theorem

Theorem 48 (Asymptotic Property of LR test)**.** *At significance level α, the rejection region ($\lambda_n(\boldsymbol{X}) < C$) of the LR-based test under regularity conditions for large n is approximately*

$$\mathscr{R} := \left\{\boldsymbol{x} \in \mathscr{X} : -2\log[\lambda_n(\boldsymbol{X})] \geq \chi^2_{d,\alpha}\right\}, d = \dim\Theta - \dim\Theta_0 \tag{7.86}$$

where

$$-2\log[\lambda_n(\boldsymbol{X})] = 2\left\{\sup_{\theta\in\Theta} \text{ł}_n(\theta) - \sup_{\theta\in\Theta_0} \text{ł}_n(\theta)\right\} \tag{7.87}$$

Also, we have a general formula to solve for hypothesis testing. We list some cases here:

(i) **Testing the Mean in a (Asymptotically) Normal Sample with Known Variance**

Suppose we know the random sample takes the form $N(\mu,\sigma^2)$ where σ^2 is known, and we wish to test:

$$\mathscr{H}_0 : \mu = \mu_0, \mathscr{H}_1 : \mu \neq \mu_0 (\mu > \mu_0)(\mu < \mu_0) \tag{7.88}$$

and we first compute the value under the null hypothesis:

$$z = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\sigma} \text{ In a large sample we may replace } (\sigma \text{ by } s_n) \tag{7.89}$$

and we will reject $\mathscr{H}_0$ at significance level α if:

$$|z| > z_{\alpha/2} (z > z_\alpha)(z < -z_\alpha). \tag{7.90}$$

(ii) **Testing the Mean in a Normal Sample with Unknown Variance**

Suppose we know the random sample takes the form $N(\mu,\sigma^2)$ where σ^2 is unknown, and we wish to test:

$$\mathscr{H}_0 : \mu = \mu_0, \mathscr{H}_1 : \mu \neq \mu_0 (\mu > \mu_0)(\mu < \mu_0) \tag{7.91}$$

and we first compute the value under the null hypothesis:

$$t = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{s_n} \tag{7.92}$$

and we will reject $\mathscr{H}_0$ at significance level α if:

$$|t| > t_{\alpha/2,n-1}(t > t_{\alpha,n-1})(t < -t_{\alpha,n-1}). \tag{7.93}$$

(iii) **Testing the Variance in a Normal Sample with Unknown Mean and Variance**

Suppose we know the random sample takes the form $N(\mu, \sigma^2)$ where μ, σ^2 is unknown, and we wish to test:

$$\mathscr{H}_0 : \sigma^2 = \sigma_0^2, \mathscr{H}_1 : \sigma^2 \neq \sigma_0^2(\sigma^2 > \sigma_0^2)(\sigma^2 < \sigma_0^2) \tag{7.94}$$

and we first compute the value under the null hypothesis:

$$\chi = \frac{(n-1)s^2}{\sigma_0^2} \tag{7.95}$$

and we will reject $\mathscr{H}_0$ at significance level α if:

$$\chi^2 > \chi^2_{\alpha/2,n-1}\text{or} < \chi^2_{1-\alpha/2,n-1}(\chi^2 > \chi^2_{\alpha,n-1})(\chi^2 < \chi^2_{1-\alpha,n-1}) \tag{7.96}$$

(iv) **Testing the Ratio in a Bernoulli Sample**

Suppose we know the random sample takes the form $Bernoulli(\theta)$ where θ is unknown, and we wish to test:

$$\mathscr{H}_0 : \theta = \theta_0, \mathscr{H}_1 : \theta \neq \theta_0(\theta > \theta_0)(\theta < \theta_0) \tag{7.97}$$

and we first compute the value under the null hypothesis:

$$z = \frac{\widehat{\theta} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \tag{7.98}$$

and we will reject $\mathscr{H}_0$ at significance level α if:

$$|z| > z_{\alpha/2}(z > z_\alpha)(z < -z_\alpha). \tag{7.99}$$

(v) **Testing the Difference in Mean of Two Bernoulli Samples**

Suppose we have two mutually independent large random samples (≥ 25) $X_1, \cdots, X_m \sim Bernoulli(\theta_1)$ and $Y_1, \cdots, Y_n \sim Bernoulli(\theta_2)$, and we wish to test

$$\mathscr{H}_0 : \theta_1 - \theta_2 = D_0; \mathscr{H}_1 : \theta_1 - \theta_2 \neq D_0(\theta_1 - \theta_2 > D_0)(\theta_1 - \theta_2 < D_0) \tag{7.100}$$

and we first compute the value under the null hypothesis:

If $D_0 = 0$, then

$$z = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{m} + \frac{\widehat{\theta}(1-\widehat{\theta})}{n}}}, \widehat{\theta} = \frac{x+y}{m+n}, \text{where } x/m = \widehat{\theta}_1, y/n = \widehat{\theta}_2 \tag{7.101}$$

If $D_0 \neq 0$, then

$$z = \frac{\widehat{\theta}_1 - \widehat{\theta}_2 - D_0}{\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{m} + \frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n}}} \tag{7.102}$$

and we will reject $\mathscr{H}_0$ at significance level α if

$$|z| > z_{\alpha/2}(z > z_\alpha)(z < z_\alpha). \tag{7.103}$$

(vi) **Testing the Difference in Mean of Two Normal Samples with Known Variance**

Suppose we have two mutually independent large samples (> 25) $X_1, \cdots, X_m \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_n \sim N(\mu, \sigma_2^2)$ where σ_1, σ_2 are known. We wish to test

$$\mathscr{H}_0 : \mu_1 - \mu_2 = D_0; \mathscr{H}_1 : \mu_1 - \mu_2 \neq D_0(\mu_1 - \mu_2 > D_0)(\mu_1 - \mu_2 < D_0) \tag{7.104}$$

We first compute the value under the null hypothesis:

$$z = \frac{\overline{x}_m - \overline{y}_n - D_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \tag{7.105}$$

and we will reject $\mathscr{H}_0$ at significance level α if

$$|z| > z_{\alpha/2}(z > z_\alpha)(z < -z_\alpha) \tag{7.106}$$

(vii) **Testing the Difference in Mean of Two Normal Samples with Unknown Variance**

Suppose we have two mutually independent large samples (> 25) $X_1, \cdots, X_m \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_n \sim N(\mu, \sigma_2^2)$ where σ_1, σ_2 are unknown, but we have $\sigma_1^2 = \sigma_2^2$. We wish to test

$$\mathscr{H}_0 : \mu_1 - \mu_2 = D_0; \mathscr{H}_1 : \mu_1 - \mu_2 \neq D_0(\mu_1 - \mu_2 > D_0)(\mu_1 - \mu_2 < D_0) \tag{7.107}$$

We first compute the value under the null hypothesis:

$$t = \frac{\overline{x}_m - \overline{y}_n - D_0}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}, s_p^2 = \frac{(m-1)s_m^2 + (n-1)s_n^2}{m+n-2} \tag{7.108}$$

and we will reject $\mathscr{H}_0$ at significance level α if

$$|t| > t_{\alpha/2, m+n-2}(t > t_{\alpha, m+n-2})(t < -t_{\alpha, m+n-2}) \tag{7.109}$$

Now I have listed a bunch of exercises, do them carefully, and identify which of the above 7 cases! The values of z, t, χ can all be found in the standard table.

Exercise 25. *Atlantic bluefin tuna is the largest and most endangered of the tuna species; the concern is that this species has been overfished and that the mean weight has decreased. Suppose a random sample of 12 Atlantic blue fin tuna was obtained from commercial fishing boats and weighted. The sample is normally distributed with $x_n = 535.7$ and $s_n = 37.8$. Is there any evidence that the mean weight is less than 550 pounds? Use the significance level $\alpha = 0.05$.*

A general approach is that, we always make null hypothesis to be simple, i.e it is equal to something. So $\mathscr{H}_0 : \mu = 550$. Also we want to see if its the weight has been reduced, so we make the alternative hypothesis to be $\mathscr{H}_1 : \mu < 550$. Then according to the table, here both μ, σ unknown, we compute the value under the null hypothesis $\mathscr{H}_0$ first:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{12}} = \frac{535.7 - 550}{37.8/\sqrt{12}} = -1.310493468 \tag{7.110}$$

now, in order to reject $\mathscr{H}_0$ under level α, we need $t < -t_{\alpha,n-1}$, which according to the t-table we have $t_{0.05,11} = 1.796$, where we have $-1.31049 > -1.796$, hence we do not reject $\mathscr{H}_0$, so at significance level $\alpha = 0.05$, we do not see any evidence that the mean weight is less than 550 pounds.

Exercise 26. *Despite a sophisticated recycling system, a water park informs the city water department of their need for 1 million liter of water per day. The city water department selected a random sample of $n = 21$ days; the mean and sample standard deviation of the park?s water usage (in thousands of liter) were $x_n = 927.43$, $s_n = 154.45$. Assuming the usage is normally distributed, is there evidence to suggest the mean water usage is different from 1 million liter per day? Use the significance level $\alpha = 0.05$.*

Here, we have a similar case as the previous exercise: We're in a normal random sample with both μ, σ^2 unknown. So let $\mathscr{H}_0 : \mu = 1000$ and $\mathscr{H}_1 := \mu \neq 1000$. So again we compute the value under $\mathscr{H}_0$, we have

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{21}} = \frac{927.43 - 1000}{154.45/\sqrt{21}} = -2.153 \tag{7.111}$$

now in order to reject $\mathscr{H}_0$ under level α, we need $|t| > t_{\alpha/2,n-1}$ and from the t-table we have $t_{\alpha/2,n-1} = t_{0.025,20} = 2.086$, and since we do have $|t| = 2.153 > 2.086$, we will hence reject $\mathscr{H}_0$ and in favor of $\mathscr{H}_1$, that is, there is evidence that the mean water usage is different from 1 million liter per day.

Exercise 27. *A study conducted by the Florida Game and Fish Commission aims at assessing the amounts of the DDT insecticide in the brain tissue of brown pelicans. Approximately Normal and independent samples of $n = 10$ juveniles and $m = 13$ nestlings gave (in parts per million), and*

$$\bar{x}_n = 0.041, s_n = 0.017, \bar{y}_m = 0.026, s_m = 0.006. \tag{7.112}$$

Test whether the mean amounts of DDT in juveniles and nestlings are the same. Use the significance level $\sigma = 0.05$.

In this example, we have two mutually independent random samples, and we design $\mathscr{H}_0 : \mu_1 - \mu_2 = 0$ while $\mathscr{H}_1 : \mu_1 - \mu_2 \neq 0$, and in this two samples we have $\sigma_1^2 = \sigma_2^2$ unknown, so we first compute the value under the null hypothesis, which is

$$t = \frac{\bar{x}_n - \bar{y}_m - 0}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}, s_p^2 = \frac{(n-1)s_n^2 + (m-1)s_m^2}{n+m-2} \tag{7.113}$$

where we compute $s_p = \sqrt{\frac{9 \times 0.017^2 + 12 \times 0.006^2}{10+13-2}} = 0.012$, and we have

$$t = \frac{0.041 - 0.026}{0.012 \times \sqrt{\frac{1}{10} + \frac{1}{13}}} = 2.97178 \tag{7.114}$$

while we will reject $\mathscr{H}_0$ if $|t| > t_{\alpha/2, n+m-2} = t_{0.025,21} = 2.08$, which we see that clearly we should reject $\mathscr{H}_0$, meaning that the amount of DDT in juveniles and nestlings are not the same.

Exercise 28. *A company produces machine engine parts that are supposed to have a diameter variance no larger than* 0.0002. *A random sample of $n = 10$ parts gave a sample variance of* 0.0003. *We wish to test $\mathscr{H}_0 : \sigma^2 = 0.0002$, $\mathscr{H}_1 : \sigma^2 > 0.0002$. at the significance level $\alpha = 0.05$. Assume that the random sample is iid from $N(\mu, \sigma^2)$ with both parameters unknown.*

In this example, we will first compute the value under the null hypothesis $\mathscr{H}_0$:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \times 0.0003}{0.0002} = 13.5 \tag{7.115}$$

In this case we will reject $\mathscr{H}_0$ if we have $\chi^2 > \chi^2_{\alpha, n-1} = \chi^2_{0.05,9} = 16.918$, where we see that we do not satisfy this condition, and hence we will not reject $\mathscr{H}_0$, and we conclude that at significance level $\alpha = 0.05$, we do not have much information that $\sigma^2 > 0.0002$.

Exercise 29. *An experimenter was convinced that the variability in his/her measuring equipment results in a standard deviation of 2; $n = 16$ measurements yielded $s^2 = 6.1$. Do the data disagree with his/her claim? Use the significance level $\alpha = 0.05$. Assume the measurements are normally distributed with both mean and variance unknown.*

In this example, we will test $\mathscr{H}_0 : \sigma^2 = 4$ and $\mathscr{H}_1 : \sigma^2 \neq 4$. *Note that standard deviation is σ!!!.* In this normal sample with both mean and variance unknown, we first compute the value under the null hypothesis $\mathscr{H}_0$:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{15 \times 6.1}{4} = 22.875 \tag{7.116}$$

Under this condition we will reject $\mathscr{H}_0$ if $\chi^2 > \chi^2_{\alpha/2,n-1}$ or $\chi^2 < \chi^2_{1-\alpha/2,n-1}$. Here we have $\alpha = 0.05, n-1 = 15$, and $\chi^2_{0.025,15} = 27.48839$ and $\chi^2_{0.975,15} = 6.26214$, and we see that χ^2 do not satisfy any of these two conditions, so we will not reject $\mathscr{H}_0$, at significance level 0.05.

Exercise 30. *A study published in* 2004 *in Current Allergy and Clinical Immunology concerns the allergy to the powder on latex gloves. Among other things, the exposure to the powder of* $n = 46$ *hospital employees with diagnosed latex allergy was investigated. The number of latex gloves used per week by these sampled workers is summarized as* $\bar{x}_n = 19.3$ *,* $s_n = 11.9$. *Is there evidence to conclude that the mean number of latex gloves used per week by hospital employees with latex allergy is more than* 15*? Use* $\alpha = 0.01$.

Here since we have a large sample so we could use a normal approximation for this sample. We will test $\mathscr{H}_0 : \mu = 15$ and $\mathscr{H}_1 : \mu > 15$. We first compute the value under the null hypothesis:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{19.3 - 15}{11.9/\sqrt{46}} = 2.4507 \tag{7.117}$$

We will reject $\mathscr{H}_0$ if $t > t_{\alpha,n-1} = t_{0.01,45} = 2.326$, where we will reject $\mathscr{H}_0$ under this case, at a significance level $\alpha = 0.01$.

Exercise 31. *A machine in a factory produces* 10% *of defectives among a large lot of items that it produces in a day. A random sample of* 100 *items from the day?s production contains* 15 *defectives, and the supervisor says that the machine must be repaired. Is there evidence that the machine produces more than* 10% *of defectives on average? Use* $\alpha = 0.05$.

Here, we have a Bernoulli random sample, and we test $\mathscr{H}_0 : p = 0.1$ and $\mathscr{H}_1 : p > 0.1$. We first compute the value under the null hypothesis:

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.15 - 0.1}{\sqrt{0.1 \times 0.9/100}} = 1.66667 \tag{7.118}$$

and we will reject $\mathscr{H}_0$ if $z > z_\alpha = z_{0.005} = 1.645$, which means we will reject $\mathscr{H}_0$ under significance level $\alpha = 0.05$.

Exercise 32. *Lipitor is a drug that is used to control cholesterol. In a randomized clinical trial,* 94 *subjects were treated with Lipitor and* 270 *independently selected subjects were given a placebo. Among* 94 *treated with Lipitor,* 7 *developed infections, while among* 270 *given a placebo,* 27 *developed infections. Is there a difference between the infection rates for the two drugs? Use* $\alpha = 0.05$.

Here, we have two Bernoulli random samples, let X_i be the sample treated with Lipitor and Y_j be the sample treated with placebo. We have $X_1, \cdots, X_{94} \sim Bernoulli(\theta_1)$, $\theta_1 = 0.074468$ and $Y_1, \cdots, Y_{270} \sim Bernoulli(\theta_2), \theta_2 = 0.1$, we test $\mathscr{H}_0 : \mu_1 - \mu_2 = 0$, and $\mathscr{H}_1 : \mu_1 - \mu_2 \neq 0$. We first compute the value under the null hypothesis:

$$z = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n_1} + \dfrac{\widehat{p}(1-\widehat{p})}{n_2}}}, \widehat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{7+27}{94+270} = 0.09341 \tag{7.119}$$

and we get

$$z = \frac{0.074468 - 0.1}{\sqrt{0.09341(1-0.09341)/94 + 0.09341(1-0.09341)/270}} = -0.73262 \tag{7.120}$$

We will reject $\mathscr{H}_0$ if $|z| > z_{\alpha/2} = z_{0.025} = 1.96$, and hence we do not reject $\mathscr{H}_0$, under the significance level $\alpha = 0.05$.