Regression and Analysis of Variance

Fall 2025, Math 533 Course Notes

McGill University

By Jiajun Zhang

September 12, 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

Review of Asymptotic Statistics

1.1 Random Variables and Convergence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space, where Ω is some arbitrary non-empty set (we usually denote as the sample space). $\mathcal{F}$ is another set which contains a collection of subsets of Ω that satisfies: (i) $\Omega \in \mathcal{F}$; (ii) Closed under set compliments; (iii) Closed under countable unions. $\mathcal{F}$ 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) $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that: (i) $\mathbb{P}(\Omega) = 1$; (ii) If $\{X_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ 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 $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $\forall B \subseteq \mathfrak{B}(\mathbb{R})$, its pre-image $X^{-1}(B) \subseteq \mathcal{F}$. We will work with a sequence of random variables $\{X_i\}_{i=1}^{\infty}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $X_n \xrightarrow{L} X$, 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 $X_n \xrightarrow{P} X$, meaning X_n converges to X in probability, if $\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \to 0$. We write $X_n \xrightarrow{L^p} X$, 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 $X_n \xrightarrow{a.s} X$.

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:

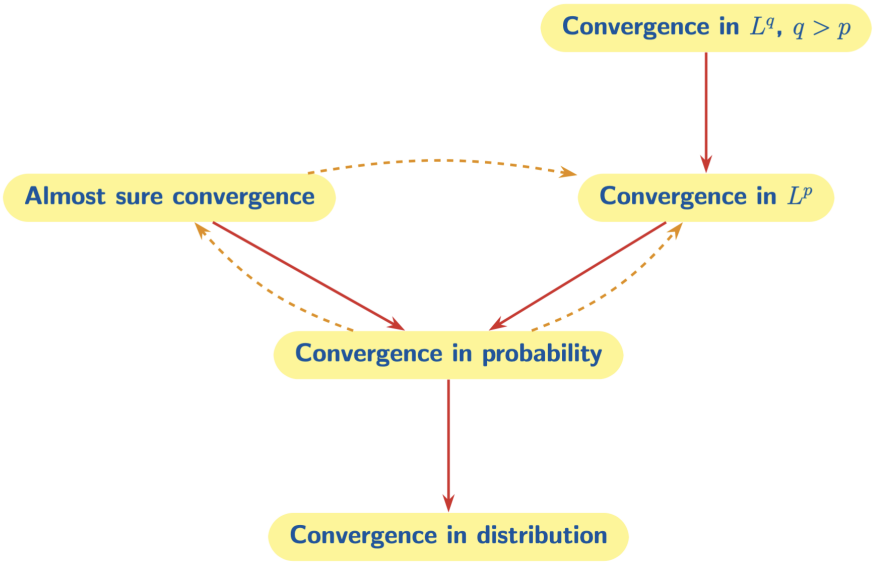

Figure 1: The diagram shows the relations between different modes of convergence. The arrows in red means direct implication without any further condition, while arrows in orange will hold if extra conditions are given. The general structure is that, convergence in probability implies convergence almost 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 $(\Omega, \mathcal{F}, \mathbb{P})$.

Theorem

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

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 $X_{n_k} \xrightarrow{a.s} X$. ■

Theorem

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

Proof. The proof is straightforward, we have

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

■

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

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

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

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

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 $\mathcal{L}^p$), 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) $X_n + Y_n \xrightarrow{L} X + c$; $(ii) X_n Y_n \xrightarrow{L} cX$; (iii) $X_n / Y_n \xrightarrow{L} X/c$.*

Proof. I am not proving this. ■

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 $(\Omega, \mathcal{F}, \mathbb{P})$, 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

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

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

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

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

Proof. We have

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

where

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

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

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

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

We can also prove the theorem using projection theorem:

Theorem

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

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

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

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

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:

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

Definition

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

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

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 $\mathcal{P}$ such that each $p(x) \in \mathcal{P}$ 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 $\mathcal{C} \subseteq \mathcal{M}$ and indeed try to find a minimizing f in the class $\mathcal{C}$. This class can depend on n. As we can see, if $\mathcal{C}_n$ gets closer and closer to $\mathcal{M}$, we will get more estimation errors because of the problem of overfitting I discussed; also if $\mathcal{C}_n$ 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 $\mathcal{C}_n$ be a class of functions depending on $\mathcal{D}_n$, if $\hat{m}_n := \hat{m}(\cdot, \mathcal{D}_n)$ satisfies*

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

then

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

Proof. By a similar argument , we can show that

$$\begin{aligned} \mathbb{E}\Big[\{\hat{m}_n(\boldsymbol{x}) - m(\boldsymbol{x})\}^2 \Big| \mathcal{D}_n\Big] &= \mathbb{E}\Big[(y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \mathbb{E}\Big[(y - m(\boldsymbol{x}))^2\Big] && (2.11) \\ &= \mathbb{E}\Big[(y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \inf_{f \in \mathcal{C}_n} \mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big] \\ &\quad + \inf_{f \in \mathcal{C}_n} \Big\{\mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big] - \mathbb{E}\Big[(y - m(\boldsymbol{x}))^2\Big]\Big\} && (2.12) \\ &= \mathbb{E}\Big[(y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \inf_{f \in \mathcal{C}_n} \mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big] \\ &\quad + \inf_{f \in \mathcal{C}_n} \mathbb{E}\Big[(f(\boldsymbol{x}) - m(\boldsymbol{x}))^2\Big] && (2.13) \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}\Big[(y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \inf_{f \in \mathcal{C}_n} \mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big] \\ &= \sup_{f \in \mathcal{C}_n} \Big\{\mathbb{E}\Big[(y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big]\Big\} && (2.14) \\ &= \sup_{f \in \mathcal{C}_n} \Bigg(\mathbb{E}\Big[y - \hat{m}_n(\boldsymbol{x}))^2 \Big| \mathcal{D}_n\Big] - \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}_n(\boldsymbol{x}_i))^2 + \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}_n(\boldsymbol{x}_i))^2 \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 - \mathbb{E}\Big[(y - f(\boldsymbol{x}))^2\Big]\Bigg) && (2.15) \end{aligned}$$

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

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

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

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

■

We may wonder which class $\mathcal{C}_n$ to choose from? If $\mathcal{C}_n$ is too large, overfitting may occur and the estimation error will likely increase; however if $\mathcal{C}_n$ 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:

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

Definition

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

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

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 $f(\boldsymbol{x}) = \beta_0 + \beta_1 x + \beta_x x_2 + \beta_3 x_2^2$, 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

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

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

Definition

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

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

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. Then we use

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

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 F is strictly convex and finite:

$$F(\boldsymbol{u}) = \mathbb{E}\left[(y - \boldsymbol{x}^\top \boldsymbol{u})^2\right] \tag{2.23}$$

proof still under construction

Next we have

$$F(\boldsymbol{u}) = \underbrace{\mathbb{E}[y^2]}_{F_1(\boldsymbol{u})} \underbrace{-2\boldsymbol{u}\mathbb{E}[yx^T]}_{F_2(\boldsymbol{u})} + \underbrace{\boldsymbol{u}^T\boldsymbol{u}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]}_{F_3(\boldsymbol{u})} \tag{2.24}$$