Regression and Analysis of Variance

Fall 2025, Math 533 Course Notes

McGill University

By Jiajun Zhang

September 2, 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 Chen, 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) $\Omega \in \mathbb{F}$; (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:

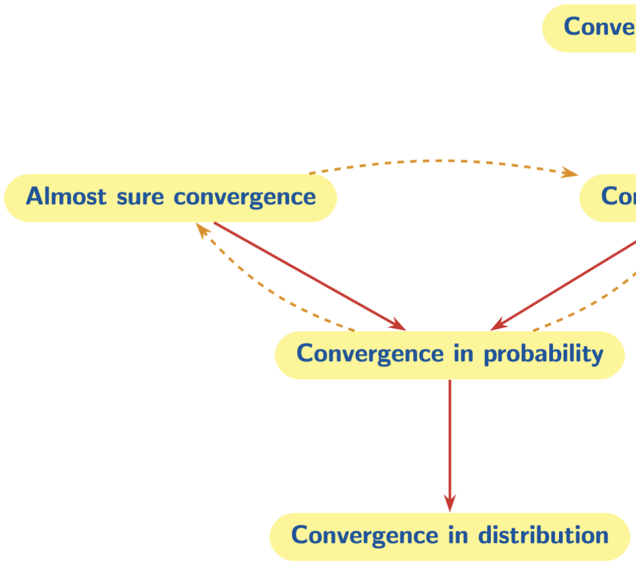
Convergence in L^q, $q > p$
Almost sure convergence
Convergence in L^p
Convergence in probability
Convergence in distribution

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

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:

Corollary

Corollary 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

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

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 $X_n \overset{m}{\to} X$, 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 $f(X_n) \overset{m}{\to} f(X)$.*

Proof. I am not proving this. ■

Theorem

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

Proof. I am not proving this. ■

Linear Regression & Regression Analysis

2.1 An Introduction

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

Definition

Definition 1. *The coordinates of $\mathbf{x}$, 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 $\mathbf{x}$, 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 $\mathbf{x}$ and the response y. We recall that $\mathbb{P}_{\mathbf{x},y}$ is the joint distribution of $\mathbf{x}, y$, which can be written as

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

through a conditional probability argument, in above $\mathbb{P}_{\mathbf{x}}$ is the marginal distribution of all covariates and $\mathbb{P}_{y|\mathbf{x}}$ is the conditional distribution of y given $\mathbf{x}$. It is not easy to find the conditional distribution, but we can work with conditional expectation $\mathbb{E}[y|\mathbf{x}]$ 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 $m : \mathbf{u} \to \mathbb{E}[\mathbf{y}|\mathbf{x} = \mathbf{u}] \in \mathcal{M}$, then*

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

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

Proof. We have

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

where

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

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

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

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