Week 3 Lecture Notes

Statistical Modeling: From Data to Random Vectors

1 Motivation: From Observed Data to Statistical Modeling

Before we dive into the mathematics, let us pause and remember **why** we study regression. Up to now, our focus has been on fitting a line (or, more generally, a hyperplane) to a given set of data points, treating the entries of our data matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ as just fixed numbers. The only constraint was that the columns of $\boldsymbol{X}$ must be linearly independent (meaning $\boldsymbol{X}$ has full column rank).

However, in real-world applications, we are rarely interested in just *explaining* the values of the data at hand. Rather, our goal is to *learn about the broader relationship* between the covariates (the X_{ij}) and the response (Y) as it exists in some population of interest. In other words, we want the line we fit to the sample to tell us as much as possible about the "true" relationship in the population.

Statistical Modeling: Introducing Randomness

How can we make such a leap from a single dataset to statements about a population? The answer lies in embracing statistical modeling. Specifically, we now make a fundamental assumption:

The data points $(1, x_{i1}, \ldots, x_{ip}, y_i)$ are not just arbitrary numbers, but are produced as ***independent and identically distributed*** *(i.i.d.) samples from some underlying random process.*

Formally, our observed data are n i.i.d. realizations of a random vector:

$$(1, X_1, \ldots, X_p, Y) \sim P$$

where P denotes the (unknown) joint distribution governing the population.

Regression as Decomposition: Systematic and Random Parts

Given this statistical mindset, it is always true (for any joint distribution P) that we can "split" a response Y_i as follows:

$$Y_i = \underbrace{\mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)}_{\text{systematic part } f(X_{i1},\ldots,X_{ip})} + \underbrace{\left(Y_i - \mathbb{E}[Y_i \mid X_{i1}, \ldots, X_{ip}]\right)}_{\text{random noise } \epsilon_i} \tag{1}$$

Here:

- $f(X_{i1}, \ldots, X_{ip}) = \mathbb{E}[Y_i \mid X_{i1}, \ldots, X_{ip}]$ captures the (possibly complicated) *systematic* relationship between the predictors and response in the population,
- ϵ_i captures the *unexplained variation*: how much the actual observation Y_i deviates from the best prediction based on $X_{i1}, \ldots, X_{ip}$.

A key property: This error ϵ_i always satisfies (by the definition of conditional expectation)

$$\begin{aligned}\mathbb{E}\big[\epsilon_i \bigm| X_{i1}, \ldots, X_{ip}\big] &= \mathbb{E}\left[Y_i - \mathbb{E}(Y_i \mid X_{i1}, \ldots, X_{ip}) \mid X_{i1}, \ldots, X_{ip}\right] \\ &= \mathbb{E}(Y_i \mid X_{i1}, \ldots, X_{ip}) - \mathbb{E}(Y_i \mid X_{i1}, \ldots, X_{ip}) = 0.\end{aligned}$$

In words: The error ϵ_i has mean zero, once you know the predictor values (and indeed also unconditionally).

1.1 Linear Model Assumption

While $f(X_{i1}, \ldots, X_{ip})$ could be any function, *we now make the crucial linearity assumption*:

The model:

$$f(1, X_{i1}, \ldots, X_{ip}) = \sum_{j=0}^{p} \beta_j X_{ij}$$

That is, we assume the true conditional mean of Y is an **affine** (linear plus intercept) function of the predictors.

Putting this all together, our **general linear model** is

$$Y_i = \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_i, \qquad \mathbb{E}\big[\epsilon_i \mid X_{i1}, \ldots, X_{ip}\big] = 0, \tag{2}$$

$$\operatorname{Cov}(\epsilon_k, \epsilon_\ell \mid \text{all } X_{ij}) = \begin{cases} \sigma^2, & k = \ell; \\ 0, & k \neq \ell. \end{cases} \tag{3}$$

Interpretation:

- The **errors** ϵ_i are random, have mean 0, variance σ^2, and are uncorrelated *conditional* on all the predictors.
- The parameters β_j and σ^2 are (unknown) quantities describing the population relationship.

A Technical Note: Treating Predictors as Fixed

Throughout most of the course, we will usually treat the data matrix $\boldsymbol{X}$ as *fixed*, not random. (This is called the "fixed design" or "conditional on X" perspective.)

Thus, for the remainder, the model is written as

$$Y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \operatorname{Cov}(\epsilon_k, \epsilon_\ell) = \begin{cases} \sigma^2, & k = \ell \\ 0, & k \neq \ell \end{cases} \tag{4}$$

Where are we going? Our ultimate aim is to learn about (i.e., statistically infer) the population parameters $\boldsymbol{\beta}$ *and* σ^2 in this model, given our observed data.

2 Random Vectors and Random Matrices: Moments and Covariance Algebra

2.1 Random Vectors and Matrices: Definitions

Let's now lay out some formal probability language that will be essential for rigorous inference.

Definition 2.1 (Random Vector and Matrix)**.**

- *A* random vector *is an $n \times 1$ column vector*

$$\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$$

where each Z_i is a random variable and together they have some joint distribution.

- *A* random matrix *is an $n \times m$ matrix*

$$\boldsymbol{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1m} \\ Z_{21} & Z_{22} & \cdots & Z_{2m} \\ \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nm} \end{bmatrix}$$

where every entry Z_{ij} is a random variable, and together they have a joint distribution.

Definition 2.2 (Expectation of a Random Matrix)**.** *Let $\boldsymbol{Z}$ be a random $n \times m$ matrix. Its expectation (mean) $\mathbb{E}\boldsymbol{Z}$ is the $n \times m$ matrix whose (i,j)th entry is $\mathbb{E}Z_{ij}$:*

$$[\mathbb{E}\boldsymbol{Z}]_{ij} = \mathbb{E}[Z_{ij}].$$

In formulas:

$$\mathbb{E}\boldsymbol{Z} = \begin{bmatrix} \mathbb{E}Z_{11} & \mathbb{E}Z_{12} & \cdots & \mathbb{E}Z_{1m} \\ \mathbb{E}Z_{21} & \mathbb{E}Z_{22} & \cdots & \mathbb{E}Z_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbb{E}Z_{n1} & \mathbb{E}Z_{n2} & \cdots & \mathbb{E}Z_{nm} \end{bmatrix}$$

As a special case, when $m = 1$, for a random vector $\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$,

$$\mathbb{E}\boldsymbol{Z} = \begin{bmatrix} \mathbb{E}Z_1 \\ \mathbb{E}Z_2 \\ \vdots \\ \mathbb{E}Z_n \end{bmatrix}$$

2.2 Properties of Expectation (Random Matrices and Vectors)

Let $\boldsymbol{Z}$ and $\boldsymbol{W}$ be random matrices (of compatible sizes), and let $\boldsymbol{A}, \boldsymbol{B}$ be fixed (deterministic) matrices of appropriate dimensions.

1. **Linearity**: $\mathbb{E}[\boldsymbol{Z} + \boldsymbol{W}] = \mathbb{E}[\boldsymbol{Z}] + \mathbb{E}[\boldsymbol{W}]$.

   Hint of Proof. For a fixed entry:

$$[\mathbb{E}(\boldsymbol{Z} + \boldsymbol{W})]_{ij} = \mathbb{E}(Z_{ij} + W_{ij}) = \mathbb{E}Z_{ij} + \mathbb{E}W_{ij} = [\mathbb{E}\boldsymbol{Z}]_{ij} + [\mathbb{E}\boldsymbol{W}]_{ij}.$$

□

2. **Compatibility with Linear Maps:**

   - $\mathbb{E}[\boldsymbol{AZ}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]$,
   - $\mathbb{E}[\boldsymbol{ZB}] = \mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$,
   - $\mathbb{E}[\boldsymbol{AZB}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$.

   Hint of Proof. This follows since expectation is computed entrywise, and all the constants (i.e., elements of $\boldsymbol{A}$ and $\boldsymbol{B}$) can be factored out. □

3. **Affine Linearity:** For any vector U (random), matrix A (fixed), constant C,

$$\mathbb{E}[AU + C] = A\,\mathbb{E}[U] + C.$$

   (Exercise: try proving this from the above.)

2.3 Recap: Classical Variance and Covariance (Univariate Case)

Recall for Z, W (random variables):

- **Covariance:**

$$\mathrm{Cov}(Z, W) := \mathbb{E}\left[(Z - \mu_Z)(W - \mu_W)\right], \qquad \mu_Z := \mathbb{E}Z,\ \mu_W := \mathbb{E}W$$

- **Alternative formula:**

$$\mathrm{Cov}(Z, W) = \mathbb{E}[ZW] - \mathbb{E}Z \cdot \mathbb{E}W$$

- **Variance:** $\mathrm{Cov}(Z, Z) = V(Z) = \mathbb{E}\left[(Z - \mu_Z)^2\right]$

Properties:

- $\mathrm{Cov}(Z, W) = \mathrm{Cov}(W, Z)$
- For any real a, $\mathrm{Cov}(aZ + R, W) = a\,\mathrm{Cov}(Z, W) + \mathrm{Cov}(R, W)$

2.4 Multivariate Generalization: Covariance Matrix

Definition 2.3 (Covariance Matrix of Random Vectors)**.** *Let $\boldsymbol{Z} \in \mathbb{R}^n$ and $\boldsymbol{W} \in \mathbb{R}^m$ be random vectors. The* covariance matrix $\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})$ *is the $n \times m$ matrix whose (i, j) entry is*

$$[\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})]_{ij} := \operatorname{Cov}(Z_i, W_j)$$

When $\boldsymbol{W} = \boldsymbol{Z}$, write

$$\operatorname{cov}(\boldsymbol{Z}) := \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{Z})$$

One can equivalently express (using matrix notation):

$$\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{E}\left[(\boldsymbol{Z} - \mu_{\boldsymbol{Z}})(\boldsymbol{W} - \mu_{\boldsymbol{W}})^\top\right] \in \mathbb{R}^{n \times m} \tag{5}$$

$$\operatorname{cov}(\boldsymbol{Z}) = \mathbb{E}\left[(\boldsymbol{Z} - \mu_{\boldsymbol{Z}})(\boldsymbol{Z} - \mu_{\boldsymbol{Z}})^\top\right] \in \mathbb{R}^{n \times n} \tag{6}$$

Alternatively, by distributing expectation,

$$\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{W}^\top] - \mu_{\boldsymbol{Z}}\mu_{\boldsymbol{W}}^\top \tag{7}$$

$$\operatorname{cov}(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top] - \mu_{\boldsymbol{Z}}\mu_{\boldsymbol{Z}}^\top \tag{8}$$

Intuitive Note: The covariance matrix encodes not just the variance of each component (on the diagonal), but also how different components "move together" (the off-diagonal entries), generalizing the familiar scalar variance to vectors.

Properties of the Covariance Matrix

Let $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{R}$ be random column vectors, and fix any matrix $\boldsymbol{A}$ (of suitable dimensions), and any deterministic vector $\boldsymbol{a}$. Then:

1. $\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \operatorname{cov}(\boldsymbol{W}, \boldsymbol{Z})^\top$
2. $\operatorname{cov}(\boldsymbol{Z} + \boldsymbol{R}, \boldsymbol{W}) = \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) + \operatorname{cov}(\boldsymbol{R}, \boldsymbol{W})$
3. $\operatorname{cov}(\boldsymbol{A}\boldsymbol{Z}, \boldsymbol{B}\boldsymbol{W}) = \boldsymbol{A} \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) \boldsymbol{B}^\top$
4. $\operatorname{cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A} \operatorname{cov}(\boldsymbol{Z}) \boldsymbol{A}^\top$
5. $V(\boldsymbol{a}^\top \boldsymbol{Z}) = \boldsymbol{a}^\top \operatorname{cov}(\boldsymbol{Z})\boldsymbol{a}$
6. $\operatorname{cov}(\boldsymbol{Z})$ is always a non-negative definite matrix[1].

Remark 2.4. *These properties are powerful: for example, property 4 tells us that applying a linear transformation to a random vector simply transforms its covariance in a predictable matrix way.*

[1] That is, for any vector $\boldsymbol{a}$, $\boldsymbol{a}^\top \operatorname{cov}(\boldsymbol{Z})\boldsymbol{a} \geq 0$.

2.5 Application: Covariance in the Linear Model

Returning to the general linear model, in vector and matrix notation, we may write:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with

$$\mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{0}, \qquad \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$$

where:

- $\boldsymbol{X}$ is a fixed (non-random) $n \times (p+1)$ matrix (including intercept),
- $\boldsymbol{\beta}$ and σ^2 are unknown population parameters,
- $\boldsymbol{\epsilon}$ is a random n-vector of errors, independent and identically distributed, with mean 0 and variance σ^2.

Compact Notation: Sometimes we write $\boldsymbol{\epsilon} \sim (\boldsymbol{0},\, \sigma^2\boldsymbol{I}_n)$ as shorthand for "mean zero and covariance $\sigma^2\boldsymbol{I}_n$."

Thus, the full model may be compactly summarized as:

$$\boxed{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (\boldsymbol{0},\, \sigma^2\boldsymbol{I}_n)}$$

Remark 2.5. *The framework above sets the stage for nearly all modern statistical inference in linear regression — both for estimation and for assessing uncertainty in our estimated parameters.*