4 Statistical modeling

So far we have been working with arbitrary data points $(1, x_{i1}, \ldots, x_{ip}, y_i)$, $i = 1, \ldots, n$. In other words, the $n \times (p+1)$ matrix $\boldsymbol{X}$ and the vector $\boldsymbol{y}$ consisted of any fixed numbers, except that we assumed that $\boldsymbol{X}$ has full column rank (i.e., the columns of $\boldsymbol{X}$ are linearly independent). If we return to the motivation for the regression problem, recall that we are ultimately interested in learning something about the relationship between the covariate vector $(1, X_{i1}, \ldots, X_{ip})$ and the response Y in some *population*, rather than in the particular dataset (sample) we happened to observe. In other words, we want the least squares regression line that we fit on the sample to *estimate* a 'theoretical' (or 'true') regression line for some target population. This will be possible if we assume that the observations are a *random sample* from the target population. Hence, we will now assume that the data points $(1, x_{i1}, \ldots, x_{ip}, y_i)$ are *i.i.d.* (independent, identically distributed) realizations of a *random vector*

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

Now, we can always write

$$Y_i = \underbrace{\mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)}_{f(X_{i1}, \ldots, X_{ip})} + \underbrace{\left(Y_i - \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)\right)}_{\epsilon_i},$$

where $f(X_{i1}, \ldots, X_{ip})$, the conditional expectation of Y given $X_1, \ldots, X_p$, is the *systematic* part, and ϵ_i, the deviation of Y from its conditional expectation given the X_{ij}'s, $j = 1, \ldots, p$, is the *error* part. Note that

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

i.e., the random variable ϵ has mean zero conditionally on $X_{i1}, \ldots, X_{ip}$ (by the way, this implies that we also have $\mathbb{E}\epsilon_i = 0$ unconditionally).

In general, the function $f(X_{i1}, \ldots, X_{ip})$ can be any function of $(X_{i1}, \ldots, X_{ip})$. From now on, we make the assumption that this is a *linear* (in fact, affine) function of $(x_{i1}, \ldots, x_{ip})$, i.e.,

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

We can summarize all of the above as follows. The general linear model is given by

$$Y_i = \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_{ij}, \qquad \mathbb{E}[\epsilon_i | \text{all } X_{ij}\text{'s}] = 0, \qquad \mathrm{Cov}(\epsilon_k, \epsilon_l | \text{all } X_{ij}\text{'s}) = \begin{cases} \sigma^2, & k = l \\ 0, & i \neq j \end{cases}$$

Actually, throughout the course we will generally treat the X_{ij}'s as *fixed* (nonrandom). In that case, the above is equivalent to

$$Y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_{ij}, \qquad \mathbb{E}[\epsilon_i] = 0, \qquad \mathrm{Cov}(\epsilon_k, \epsilon_l) = \begin{cases} \sigma^2, & k = l \\ 0, & i \neq j \end{cases} \tag{12}$$

Moments of random vectors, algebra of covariance. We are headed toward providing statistical inference for $\boldsymbol{\beta}$ (and σ^2) under the model (12). As this will involve working with random vectors, we begin with some general definitions.

A random vector is a vector $Z = (Z_1, \ldots, Z_n)^\top$ whose components Z_i are random variables with some joint distribution. A random matrix is a matrix

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

whose components Z_{ij} are random variables with some joint distribution.

Definition 2. The expectation of a random $n \times m$ matrix **Z** is defined as the $n \times m$ matrix $\mathbb{E}\boldsymbol{Z}$ whose (i,j)-th entry is

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

In other words,

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

and, as a special case when $m = 1$, for a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$,

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

Properties. $\boldsymbol{Z}, \boldsymbol{W}$ random matrices. For any fixed matrices $\boldsymbol{A}, \boldsymbol{B}$ of compatible dimensions, we have:

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

   Proof. $[\mathbb{E}(\boldsymbol{Z}+\boldsymbol{W})]_{ij} \overset{(1)}{=} \mathbb{E}\left([\boldsymbol{Z}+\boldsymbol{W}]_{ij}\right) \overset{(2)}{=} \mathbb{E}\left(\boldsymbol{Z}_{ij} + \boldsymbol{W}_{ij}\right) \overset{(3)}{=} \mathbb{E}\boldsymbol{Z}_{ij} + \mathbb{E}\boldsymbol{W}_{ij} \overset{(4)}{=} [\mathbb{E}\boldsymbol{Z}]_{ij} + [\mathbb{E}\boldsymbol{W}]_{ij}$ where (1) is due to the definition of the expectation of a matrix; (2) is due to the rule of addition of two matrices (the (i,j)-th element of the sum is the sum of the (i,j)-th elements); (3) is due to linearity of expectation for (univariate) random variables (this is the main step of the proof); and (4) is again due to the definition of the expectation of a matrix. □

2. $\mathbb{E}[\boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$

   Proof. First,

$$[\mathbb{E}(\boldsymbol{A}\boldsymbol{Z})]_{ij} = \mathbb{E}\left(\sum_r \boldsymbol{A}_{ir}\boldsymbol{Z}_{rj}\right) = \sum_r \boldsymbol{A}_{ir}\mathbb{E}\boldsymbol{Z}_{rj} = [\boldsymbol{A}\mathbb{E}\boldsymbol{Z}]_{ij} \Rightarrow \mathbb{E}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\mathbb{E}\boldsymbol{Z}$$

   A similar argument yields

$$\mathbb{E}(\boldsymbol{Z}\boldsymbol{B}) = (\mathbb{E}\mathbf{Z})\boldsymbol{B}$$

   Finally,

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{A}(\boldsymbol{Z}\boldsymbol{B})] \overset{(1)}{=} \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}\boldsymbol{B}] \overset{(2)}{=} \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$$

   □

3. $\mathbb{E}[AU + C] = A\mathbb{E}[U] + C$ (from $1+2$)

Proof. Exercise. $\square$

Reminder. For two random variables Z, W, recall that the covariance of Z and W is

$$\operatorname{Cov}(Z, W) := \mathbb{E}\left(Z - \mu_Z\right)\left(W - \mu_W\right)$$

where $\mu_Z := \mathbb{E}Z, \mu_W := \mathbb{E}W$.

Using linearity of the expectation, we get the identity

$$\operatorname{Cov}(Z, W) = \mathbb{E}[ZW] - \mu_Z \mu_W.$$

In the special case $W = Z$, by the definition we have

$$\operatorname{Cov}(Z, Z) = V(Z) = \mathbb{E}\left(Z - \mu_Z\right)^2$$

Properties of covariance. For any fixed $a \in \mathbb{R}$, we have:

1. $\operatorname{Cov}(W, Z) = \operatorname{Cov}(Z, W)$
2. $\operatorname{Cov}(aZ + R, W) = a \operatorname{Cov}(Z, W) + \operatorname{Cov}(R, W)$

Definition 3. The covariance matrix of a random vector $\boldsymbol{Z} \in \mathbb{R}^n$ with a random vector $\boldsymbol{W} \in \mathbb{R}^m$ is denoted $\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})$, and defined to be the $n \times m$ matrix whose (i, j)-th entry is

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

In the special case where $\boldsymbol{W} = \boldsymbol{Z}$, we denote $\operatorname{cov}(\boldsymbol{Z}) := \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{Z})$, and by the above definition,

$$[\operatorname{cov}(\boldsymbol{Z})]_{ij} := [\operatorname{cov}(\mathbf{Z}, \mathbf{Z})]_{ij} = \operatorname{Cov}\left(Z_i, Z_j\right)$$

Presented differently, with $\mu_Z := \mathbb{E}Z, \mu_W := \mathbb{E}\boldsymbol{W}$, (3.7) and (3.8) can be expressed in matrix notation as

$$\boxed{\operatorname{cov}(\mathbf{Z}, \boldsymbol{W}) := \mathbb{E}\left[\left(\boldsymbol{Z} - \mu_{\boldsymbol{Z}}\right)\left(\boldsymbol{W} - \mu_{\boldsymbol{W}}\right)^\top\right] \in \mathbb{R}^{n \times m}, \qquad (13)}$$

and

$$\boxed{\operatorname{cov}(\mathbf{Z}) := \operatorname{cov}(\mathbf{Z}, \mathbf{Z}) = \mathbb{E}\left[\left(\mathbf{Z} - \mu_{\mathbf{Z}}\right)\left(\mathbf{Z} - \mu_{\mathbf{Z}}\right)^\top\right] \in \mathbb{R}^{n \times n}. \qquad (14)}$$

Using the identity (3.8) for (univariate) random variables Z, W, and the entry-wise definition of the expectation of a matrix, we also have the multivariate counterparts,

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

and

$$\operatorname{cov}(\mathbf{Z}) = \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right] - \mu_{\mathbf{Z}} \mu_{\mathbf{Z}}^\top$$

Properties of covariance matrix. $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{R}$ random vectors; $\boldsymbol{a}$ fixed vector. Then the following properties hold:

1. $\text{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \text{cov}(\boldsymbol{W}, \boldsymbol{Z})^\top$
2. $\text{cov}(\boldsymbol{Z} + \boldsymbol{R}, \boldsymbol{W}) = \text{cov}(\boldsymbol{Z}, \boldsymbol{W}) + \text{cov}(\boldsymbol{R}, \boldsymbol{W})$
3. $\text{cov}(\boldsymbol{A}\boldsymbol{Z}, \boldsymbol{B}\boldsymbol{W}) = A\,\text{cov}(\boldsymbol{Z}, \boldsymbol{W})\boldsymbol{B}^\top$
4. $\quad \text{cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\,\text{cov}(\boldsymbol{Z})\boldsymbol{A}^\top \quad$ (from 3)
5. $\quad V\left(\boldsymbol{a}^\top \boldsymbol{Z}\right) = \boldsymbol{a}^\top \text{cov}(\boldsymbol{Z})\boldsymbol{a} \quad$ (from 4)
6. $\text{cov}(\boldsymbol{Z})$ is a nonnegative definite matrix (from 5)

Now return to the linear model. By the definition of the covariance matrix and expectation, (12) is equivalent to

$$\boxed{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{0}, \quad \text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n,} \tag{15}$$

where $\boldsymbol{X}$ is a *fixed* (nonrandom) $n \times p + 1$ matrix, and $\boldsymbol{\beta},\ \sigma^2$ unknown.

The vector $\boldsymbol{\epsilon}$ is called the *errors*. We will sometimes write $\boldsymbol{\epsilon} \sim \left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right)$ as shorthand for $\mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{0}, \quad \text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. With this notation, (15) can be written even more compactly as: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right)$.