4.8: Expected Value and Covariance Matrices

The main purpose of this section is a discussion of expected value and covariance for random matrices and vectors. These topics are somewhat specialized, but are particularly important in multivariate statistical models and for the multivariate normal distribution. This section requires some prerequisite knowledge of linear algebra.

We assume that the various indices m, n, p, k that occur in this section are positive integers. Also we assume that expected values of real-valued random variables that we reference exist as real numbers, although extensions to cases where expected values are ∞ or $-\infty$ are straightforward, as long as we avoid the dreaded indeterminate form $\infty - \infty$.

Basic Theory

Linear Algebra

We will follow our usual convention of denoting random variables by upper case letters and nonrandom variables and constants by lower case letters. In this section, that convention leads to notation that is a bit nonstandard, since the objects that we will be dealing with are vectors and matrices. On the other hand, the notation we will use works well for illustrating the similarities between results for random matrices and the corresponding results in the one-dimensional case. Also, we will try to be careful to explicitly point out the underlying spaces where various objects live.

Let $\mathbb{R}^{m \times n}$ denote the space of all $m \times n$ matrices of real numbers. The (i, j) entry of $\boldsymbol{a} \in \mathbb{R}^{m \times n}$ is denoted a_{ij} for $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$. We will identify $\mathbb{R}^n$ with $\mathbb{R}^{n \times 1}$, so that an ordered n-tuple can also be thought of as an $n \times 1$ column vector. The transpose of a matrix $\boldsymbol{a} \in \mathbb{R}^{m \times n}$ is denoted $\boldsymbol{a}^T$—the $n \times m$ matrix whose (i, j) entry is the (j, i) entry of $\boldsymbol{a}$. Recall the definitions of matrix addition, scalar multiplication, and matrix multiplication. Recall also the standard *inner product* (or *dot product*) of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \tag{4.8.1}$$

The *outer product* of $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\boldsymbol{x}\boldsymbol{y}^T$, the $n \times n$ matrix whose (i, j) entry is $x_i y_j$. Note that the inner product is the *trace* (sum of the diagonal entries) of the outer product. Finally recall the standard *norm* on $\mathbb{R}^n$, given by

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{4.8.2}$$

Recall that inner product is *bilinear*, that is, linear (preserving addition and scalar multiplication) in each argument separately. As a consequence, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle \tag{4.8.3}$$

Expected Value of a Random Matrix

As usual, our starting point is a random experiment modeled by a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. So to review, Ω is the set of outcomes, $\mathscr{F}$ the collection of events, and $\mathbb{P}$ the probability measure on the sample space $(\Omega, \mathscr{F})$. It's natural to define the expected value of a random matrix in a component-wise manner.

Suppose that $\boldsymbol{X}$ is an $m \times n$ matrix of real-valued random variables, whose (i, j) entry is denoted X_{ij}. Equivalently, $\boldsymbol{X}$ is as a random $m \times n$ matrix, that is, a random variable with values in $\mathbb{R}^{m \times n}$. The *expected value* $\mathbb{E}(\boldsymbol{X})$ is defined to be the $m \times n$ matrix whose (i, j) entry is $\mathbb{E}(X_{ij})$, the expected value of X_{ij}.

Many of the basic properties of expected value of random variables have analogous results for expected value of random matrices, with matrix operation replacing the ordinary ones. Our first two properties are the critically important *linearity properties*. The first part is the *additive property*—the expected value of a sum is the sum of the expected values.

$\mathbb{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y})$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random $m \times n$ matrices.

Proof

This is true by definition of the matrix expected value and the ordinary additive property. Note that $\mathbb{E}(X_{ij} + Y_{ij}) = \mathbb{E}(X_{ij}) + \mathbb{E}(Y_{ij})$. The left side is the (i, j) entry of $\mathbb{E}(\boldsymbol{X} + \boldsymbol{Y})$ and the right side is the (i, j) entry of $\mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y})$.

The next part of the linearity properties is the *scaling property*—a nonrandom matrix factor can be pulled out of the expected value.

Suppose that $\boldsymbol{X}$ is a random $n \times p$ matrix.

1. $\mathbb{E}(\boldsymbol{a}\boldsymbol{X}) = \boldsymbol{a}\mathbb{E}(\boldsymbol{X})$ if $\boldsymbol{a} \in \mathbb{R}^{m \times n}$.
2. $\mathbb{E}(\boldsymbol{X}\boldsymbol{a}) = \mathbb{E}(\boldsymbol{X})\boldsymbol{a}$ if $\boldsymbol{a} \in \mathbb{R}^{p \times n}$.

Proof

1. By the ordinary linearity and scaling properties, $\mathbb{E}\left(\sum_{j=1}^{n} a_{ij} X_{jk}\right) = \sum_{j=1}^{n} a_{ij} \mathbb{E}(X_{jk})$. The left side is the (i, k) entry of $\mathbb{E}(\boldsymbol{a}\boldsymbol{X})$ and the right side is the (i, k) entry of $\boldsymbol{a}\mathbb{E}(\boldsymbol{X})$.
2. The proof is similar to (a).

Recall that for independent, real-valued variables, the expected value of the product is the product of the expected values. Here is the analogous result for random matrices.

$\mathbb{E}(\boldsymbol{X}\boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X})\mathbb{E}(\boldsymbol{Y})$ if $\boldsymbol{X}$ is a random $m \times n$ matrix, $\boldsymbol{Y}$ is a random $n \times p$ matrix, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

Proof

By the ordinary linearity properties and by the independence assumption,

$$\mathbb{E}\left(\sum_{j=1}^{n} X_{ij} Y_{jk}\right) = \sum_{j=1}^{n} \mathbb{E}\left(X_{ij} Y_{jk}\right) = \sum_{j=1}^{n} \mathbb{E}\left(X_{ij}\right) \mathbb{E}\left(Y_{jk}\right) \tag{4.8.4}$$

The left side is the (i, k) entry of $\mathbb{E}(\boldsymbol{X}\boldsymbol{Y})$ and the right side is the (i, k) entry of $\mathbb{E}(\boldsymbol{X})\mathbb{E}(\boldsymbol{Y})$.

Actually the previous result holds if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are simply uncorrelated in the sense that X_{ij} and Y_{jk} are uncorrelated for each $i \in \{1, \ldots, m\}$, $j \in \{1, 2, \ldots, n\}$ and $k \in \{1, 2, \ldots p\}$. We will study covariance of random vectors in the next subsection.

Covariance Matrices

Our next goal is to define and study the covariance of two random vectors.

Suppose that $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$ and $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^n$.

1. The *covariance matrix* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the $m \times n$ matrix $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$ whose (i, j) entry is $\operatorname{cov}\left(X_i, Y_j\right)$ the ordinary covariance of X_i and Y_j.
2. Assuming that the coordinates of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have positive variance, the *correlation matrix* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the $m \times n$ matrix $\operatorname{cor}(\boldsymbol{X}, \boldsymbol{Y})$ whose (i, j) entry is $\operatorname{cor}\left(X_i, Y_j\right)$, the ordinary correlation of X_i and Y_j

Many of the standard properties of covariance and correlation for real-valued random variables have extensions to random vectors. For the following three results, $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$ and $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^n$.

$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left(\left[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})\right]\left[\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\right]^T\right)$

Proof

By the definition of the expected value of a random vector and by the defintion of matrix multiplication, the (i, j) entry of $\left[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})\right]\left[\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\right]^T$ is simply $\left[X_i - \mathbb{E}\left(X_i\right)\right]\left[Y_j - \mathbb{E}\left(Y_j\right)\right]$. The expected value of this entry is $\operatorname{cov}\left(X_i, Y_j\right)$, which in turn, is the (i, j) entry of $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$

Thus, the covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the expected value of the outer product of $\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})$ and $\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})$. Our next result is the computational formula for covariance: the expected value of the outer product of $\boldsymbol{X}$ and $\boldsymbol{Y}$ minus the outer product of the expected values.

$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left(\boldsymbol{X}\boldsymbol{Y}^T\right) - \mathbb{E}(\boldsymbol{X})[\mathbb{E}(\boldsymbol{Y})]^T$.

Proof

The (i, j) entry of $\mathbb{E}\left(\boldsymbol{X}\boldsymbol{Y}^T\right) - \mathbb{E}(\boldsymbol{X})[\mathbb{E}(\boldsymbol{Y})]^T$ is $\mathbb{E}\left(X_i, Y_j\right) - \mathbb{E}\left(X_i\right)\mathbb{E}\left(Y_j\right)$, which by the standard computational formula, is $\operatorname{cov}\left(X_i, Y_j\right)$, which in turn is the (i, j) entry of $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$.

The next result is the matrix version of the symmetry property.

$\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) = [\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})]^T$.

Proof

The (i, j) entry of $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$ is $\operatorname{cov}\left(X_i, Y_j\right)$, which is the (j, i) entry of $\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})$.

In the following result, $\boldsymbol{0}$ denotes the $m \times n$ zero matrix.

$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}$ if and only if $\operatorname{cov}\left(X_i, Y_j\right) = 0$ for each i and j, so that each coordinate of $\boldsymbol{X}$ is uncorrelated with each coordinate of $\boldsymbol{Y}$.

Proof

This follows immediately from the definition of $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$.

Naturally, when $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}$, we say that the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *uncorrelated*. In particular, if the random vectors are independent, then they are uncorrelated. The following results establish the *bi-linear* properties of covariance.

The additive properties.

1. $\operatorname{cov}(\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{Z}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Z}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{Z})$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors in $\mathbb{R}^m$ and $\boldsymbol{Z}$ is a random vector in $\mathbb{R}^n$.
2. $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{Z}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Z})$ if $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$, and $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are random vectors in $\mathbb{R}^n$.

Proof

1. From the ordinary additive property of covariance, $\operatorname{cov}\left(X_i + Y_i, Z_j\right) = \operatorname{cov}\left(X_i, Z_j\right) + \operatorname{cov}\left(Y_i, Z_j\right)$. The left side is the (i, j) entry of $\operatorname{cov}(\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{Z})$ and the right side is the (i, j) entry of $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Z}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{Z})$.
2. The proof is similar to (a), using the additivity of covariance in the second argument.

The scaling properties

1. $\operatorname{cov}(\boldsymbol{a}\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{a}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$ if $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$, $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^p$, and $\boldsymbol{a} \in \mathbb{R}^{m \times n}$.
2. $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{a}\boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{a}^T$ if $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$, $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^n$, and $\boldsymbol{a} \in \mathbb{R}^{k \times n}$.

Proof

1. Using the ordinary linearity properties of covariance in the first argument, we have

$$\operatorname{cov}\left(\sum_{j=1}^{n} a_{ij} X_j, Y_k\right) = \sum_{j=1}^{n} a_{ij} \operatorname{cov}(X_j, Y_k) \tag{4.8.5}$$

The left side is the (i, k) entry of $\operatorname{cov}(\boldsymbol{a}\boldsymbol{X}, \boldsymbol{Y})$ and the right side is the (i, k) entry of $\boldsymbol{a}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$.

2. The proof is similar to (a), using the linearity of covariance in the second argument.

Variance-Covariance Matrices

Suppose that $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$. The covariance matrix of $\boldsymbol{X}$ with itself is called the *variance-covariance matrix* of $\boldsymbol{X}$:

$$\operatorname{vc}(\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathbb{E}\left(\left[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})\right]\left[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})\right]^T\right) \tag{4.8.6}$$

Recall that for an ordinary real-valued random variable X, $\operatorname{var}(X) = \operatorname{cov}(X, X)$. Thus the variance-covariance matrix of a random vector in some sense plays the same role that variance does for a random variable.

$\operatorname{vc}(\boldsymbol{X})$ is a symmetric $n \times n$ matrix with $(\operatorname{var}(X_1), \operatorname{var}(X_2), \ldots, \operatorname{var}(X_n))$ on the diagonal.

Proof

Recall that $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$. Also, the (i, i) entry of $\operatorname{vc}(\boldsymbol{X})$ is $\operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i)$.

The following result is the formula for the variance-covariance matrix of a sum, analogous to the formula for the variance of a sum of real-valued variables.

$\operatorname{vc}(\boldsymbol{X} + \boldsymbol{Y}) = \operatorname{vc}(\boldsymbol{X}) + \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) + \operatorname{vc}(\boldsymbol{Y})$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors in $\mathbb{R}^n$.

Proof

This follows from the additive property of covariance:

$$\operatorname{vc}(\boldsymbol{X} + \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{X} + \boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) + \operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{Y}) \tag{4.8.7}$$

Recall that $\operatorname{var}(aX) = a^2 \operatorname{var}(X)$ if X is a real-valued random variable and $a \in \mathbb{R}$. Here is the analogous result for the variance-covariance matrix of a random vector.

$\operatorname{vc}(\boldsymbol{a}\boldsymbol{X}) = \boldsymbol{a}\operatorname{vc}(\boldsymbol{X})\boldsymbol{a}^T$ if $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$ and $\boldsymbol{a} \in \mathbb{R}^{m \times n}$.

Proof

This follows from the scaling property of covariance:

$$\operatorname{vc}(\boldsymbol{a}\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{a}\boldsymbol{X}, \boldsymbol{a}\boldsymbol{X}) = \boldsymbol{a}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{a}^T \tag{4.8.8}$$

Recall that if X is a random variable, then $\operatorname{var}(X) \geq 0$, and $\operatorname{var}(X) = 0$ if and only if X is a constant (with probability 1). Here is the analogous result for a random vector:

Suppose that $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$.

1. $\operatorname{vc}(\boldsymbol{X})$ is either positive semi-definite or positive definite.
2. $\operatorname{vc}(\boldsymbol{X})$ is positive semi-definite but not positive definite if and only if there exists $\boldsymbol{a} \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that, with probability 1, $\boldsymbol{a}^T\boldsymbol{X} = \sum_{i=1}^n a_i X_i = c$

Proof

1. From the previous result, $0 \leq \operatorname{var}\left(\boldsymbol{a}^T\boldsymbol{X}\right) = \operatorname{vc}\left(\boldsymbol{a}^T\boldsymbol{X}\right) = \boldsymbol{a}^T\operatorname{vc}(\boldsymbol{X})\boldsymbol{a}$ for every $\boldsymbol{a} \in \mathbb{R}^n$. Thus, by definition, $\operatorname{vc}(\boldsymbol{X})$ is either positive semi-definite or positive definite.
2. In light of (a), $\operatorname{vc}(\boldsymbol{X})$ is positive semi-definite but not positive definite if and only if there exists $\boldsymbol{a} \in \mathbb{R}^n$ such that $\boldsymbol{a}^T\operatorname{vc}(\boldsymbol{X})\boldsymbol{a} = \operatorname{var}\left(\boldsymbol{a}^T\boldsymbol{X}\right) = 0$. But in turn, this is true if and only if $\boldsymbol{a}^T\boldsymbol{X}$ is constant with probability 1.

Recall that since $\operatorname{vc}(\boldsymbol{X})$ is either positive semi-definite or positive definite, the eigenvalues and the determinant of $\operatorname{vc}(\boldsymbol{X})$ are nonnegative. Moreover, if $\operatorname{vc}(\boldsymbol{X})$ is positive semi-definite but not positive definite, then one of the coordinates of $\boldsymbol{X}$ can be written as a linear transformation of the other coordinates (and hence can usually be eliminated in the underlying model). By contrast, if $\operatorname{vc}(\boldsymbol{X})$ is positive definite, then this cannot happen; $\operatorname{vc}(\boldsymbol{X})$ has positive eigenvalues and determinant and is invertible.

Best Linear Predictor

Suppose that $\boldsymbol{X}$ is a random vector in $\mathbb{R}^m$ and that $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^n$. We are interested in finding the function of $\boldsymbol{X}$ of the form $\boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$, where $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^{n \times m}$, that is closest to $\boldsymbol{Y}$ in the mean square sense. Functions of this form are analogous to linear functions in the single variable case. However, unless $\boldsymbol{a} = \boldsymbol{0}$, such functions are not *linear transformations* in the sense of linear algebra, so the correct term is *affine function* of $\boldsymbol{X}$. This problem is of fundamental importance in statistics when random vector $\boldsymbol{X}$, the *predictor vector* is observable, but not random vector $\boldsymbol{Y}$, the *response vector*. Our discussion here generalizes the one-dimensional case, when X and Y are random variables. That problem was solved in the section on Covariance and Correlation. We will assume that $\operatorname{vc}(\boldsymbol{X})$ is positive definite, so that $\operatorname{vc}(\boldsymbol{X})$ is invertible, and none of the coordinates of $\boldsymbol{X}$ can be written as an affine function of the other coordinates. We write $\operatorname{vc}^{-1}(\boldsymbol{X})$ for the inverse instead of the clunkier $[\operatorname{vc}(\boldsymbol{X})]^{-1}$.

As with the single variable case, the solution turns out to be the affine function that has the same expected value as $\boldsymbol{Y}$, and whose covariance with $\boldsymbol{X}$ is the same as that of $\boldsymbol{Y}$.

Define $L(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\left[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})\right]$. Then $L(\boldsymbol{Y} \mid \boldsymbol{X})$ is the only affine function of $\boldsymbol{X}$ in $\mathbb{R}^n$ satisfying

1. $\mathbb{E}[L(\boldsymbol{Y} \mid \boldsymbol{X})] = \mathbb{E}(\boldsymbol{Y})$
2. $\operatorname{cov}[L(\boldsymbol{Y} \mid \boldsymbol{X}), \boldsymbol{X}] = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})$

Proof

From linearity,

$$\mathbb{E}[L(\boldsymbol{Y} \mid \boldsymbol{X})] = E(\boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})[\mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{X})] = 0 \tag{4.8.9}$$

From linearity and the fact that a constant vector is independent (and hence uncorrelated) with any random vector,

$$\operatorname{cov}[L(\boldsymbol{Y} \mid \boldsymbol{X}), \boldsymbol{X}] = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{vc}(\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) \tag{4.8.10}$$

Conversely, suppose that $\boldsymbol{U} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$ for some $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^{m \times n}$, and that $\mathbb{E}(\boldsymbol{U}) = \mathbb{E}(\boldsymbol{Y})$ and $\operatorname{cov}(\boldsymbol{U}, \boldsymbol{X}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})$. From the second equation, again using linearity and the uncorrelated property of constant vectors, we get $\boldsymbol{b}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})$ and therefore $\boldsymbol{b} = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})$. Then from the first equation, $\boldsymbol{a} + \boldsymbol{b}\mathbb{E}(\boldsymbol{X}) = \boldsymbol{Y}$ so $\boldsymbol{a} = \mathbb{E}(\boldsymbol{Y}) - \boldsymbol{b}\mathbb{E}(\boldsymbol{X})$.

A simple corollary is the $\boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X})$ is uncorrelated with any affine function of $\boldsymbol{X}$:

If $\boldsymbol{U}$ is an affine function of $\boldsymbol{X}$ then

1. $\operatorname{cov}[\boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X}), \boldsymbol{U}] = \boldsymbol{0}$
2. $\mathbb{E}(\langle \boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X}), \boldsymbol{U} \rangle) = 0$

Proof

Suppose that $\boldsymbol{U} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{X}$ where $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^{m \times n}$. For simplicity, let $\boldsymbol{L} = L(\boldsymbol{Y} \mid \boldsymbol{X})$

1. From the previous result, $\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) = \operatorname{cov}(\boldsymbol{L}, \boldsymbol{X})$. Hence using linearity,
$$\operatorname{cov}(\boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{U}) = \operatorname{cov}(\boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{a}) + \operatorname{cov}(\boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{X})\boldsymbol{b}^T = \boldsymbol{0} + [\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X}) - \operatorname{cov}(\boldsymbol{L}, \boldsymbol{X})] = \boldsymbol{0} \tag{4.8.11}$$
2. Recall that $\langle \boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{U} \rangle$ is the trace of $\operatorname{cov}(\boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{U})$ and hence has expected value 0 by part (a).

The variance-covariance matrix of $L(\boldsymbol{Y} \mid \boldsymbol{X})$, and its covariance matrix with $\boldsymbol{Y}$ turn out to be the same, again analogous to the single variable case.

Additional properties of $L(\boldsymbol{Y} \mid \boldsymbol{X})$:

1. $\operatorname{cov}[\boldsymbol{Y}, L(\boldsymbol{Y} \mid \boldsymbol{X})] = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$
2. $\operatorname{vc}[L(\boldsymbol{Y} \mid \boldsymbol{X})] = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$

Proof

Recall that $L(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}) + \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})[\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})]$

1. Using basic properties of covariance,
$$\operatorname{cov}[Y, L(\boldsymbol{Y} \mid \boldsymbol{X})] = \operatorname{cov}[\boldsymbol{Y}, \boldsymbol{X} - \mathbb{E}(\boldsymbol{X})]\left[\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\right]^T = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) \tag{4.8.12}$$
2. Using basic properties of variance-covariance,
$$\operatorname{vc}[L(\boldsymbol{Y} \mid \boldsymbol{X})] = \operatorname{vc}\left[\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\boldsymbol{X}\right] = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{vc}(\boldsymbol{X})\left[\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\right]^T = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) \tag{4.8.13}$$

Next is the fundamental result that $L(\boldsymbol{Y} \mid \boldsymbol{X})$ is the affine function of $\boldsymbol{X}$ that is closest to $\boldsymbol{Y}$ in the mean square sense.

Suppose that $\boldsymbol{U} \in \mathbb{R}^n$ is an affine function of $\boldsymbol{X}$. Then

1. $\mathbb{E}\left(\|\boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X})\|^2\right) \le \mathbb{E}\left(\|\boldsymbol{Y} - \boldsymbol{U}\|^2\right)$
2. Equality holds in (a) if and only if $\boldsymbol{U} = L(\boldsymbol{Y} \mid \boldsymbol{X})$ with probability 1.

Proof

Again, let $\boldsymbol{L} = L(\boldsymbol{Y} \mid \boldsymbol{X})$ for simplicity and let $\boldsymbol{U} \in \mathbb{R}^n$ be an affine function of $\boldsymbol{X}$.

1. Using the linearity of expected value, note that
$$\mathbb{E}\left(\|\boldsymbol{Y} - \boldsymbol{U}\|^2\right) = \mathbb{E}\left[\|(\boldsymbol{Y} - \boldsymbol{L}) + (\boldsymbol{L} - \boldsymbol{U})\|^2\right] = \mathbb{E}\left(\|\boldsymbol{Y} - \boldsymbol{L}\|^2\right) + 2\mathbb{E}(\langle \boldsymbol{Y} - \boldsymbol{L}, \boldsymbol{L} - \boldsymbol{U} \rangle) + \mathbb{E}\left(\|\boldsymbol{L} - \boldsymbol{U}\|^2\right) \tag{4.8.14}$$
But $\boldsymbol{L} - \boldsymbol{U}$ is an affine function of $\boldsymbol{X}$ and hence the middle term is 0 by our previous corollary. Hence
$\mathbb{E}\left(\|\boldsymbol{Y} - \boldsymbol{U}\|^2\right) = \mathbb{E}\left(\|\boldsymbol{L} - \boldsymbol{Y}\|^2\right) + \mathbb{E}\left(\|\boldsymbol{L} - \boldsymbol{U}\|^2\right) \ge \mathbb{E}\left(\|\boldsymbol{L} - \boldsymbol{Y}\|^2\right)$
2. From (a), equality holds in the inequality if and only if $\mathbb{E}\left(\|\boldsymbol{L} - \boldsymbol{U}\|^2\right) = 0$ if and only if $\mathbb{P}(\boldsymbol{L} = \boldsymbol{U}) = 1$.

The variance-covariance matrix of the difference between $\boldsymbol{Y}$ and the best affine approximation is given in the next theorem.

$\operatorname{vc}[\boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X})] = \operatorname{vc}(\boldsymbol{Y}) - \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$

Proof

Again, we abbreviate $L(\boldsymbol{Y} \mid \boldsymbol{X})$ by $\boldsymbol{L}$. Using basic properties of variance-covariance matrices,

$$\operatorname{vc}(\boldsymbol{Y} - \boldsymbol{L}) = \operatorname{vc}(\boldsymbol{Y}) - \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{L}) - \operatorname{cov}(\boldsymbol{L}, \boldsymbol{Y}) + \operatorname{vc}(\boldsymbol{L}) \tag{4.8.15}$$

But $\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{L}) = \operatorname{cov}(\boldsymbol{L}, \boldsymbol{Y}) = \operatorname{vc}(\boldsymbol{L}) = \operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})\operatorname{vc}^{-1}(\boldsymbol{X})\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{X})$. Substituting gives the result.

The actual mean square error when we use $L(\boldsymbol{Y} \mid \boldsymbol{X})$ to approximate $\boldsymbol{Y}$, namely $\mathbb{E}\left(\|\boldsymbol{Y} - L(\boldsymbol{Y} \mid \boldsymbol{X})\|^2\right)$, is the trace (sum of the diagonal entries) of the variance-covariance matrix above. The function of $\boldsymbol{x}$ given by

$$L(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(\boldsymbol{Y}) + \cov(\boldsymbol{Y}, \boldsymbol{X}) \vc^{-1}(\boldsymbol{X}) [\boldsymbol{x} - \mathbb{E}(\boldsymbol{X})] \tag{4.8.16}$$

is known as the *(distribution) linear regression function*. If we observe $\boldsymbol{x}$ then $L(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ is our best affine prediction of $\boldsymbol{Y}$.

Multiple linear regression is more powerful than it may at first appear, because it can be applied to non-linear transformations of the random vectors. That is, if $g: \mathbb{R}^m \to \mathbb{R}^j$ and $h: \mathbb{R}^n \to \mathbb{R}^k$ then $L[h(\boldsymbol{Y}) \mid g(\boldsymbol{X})]$ is the affine function of $g(\boldsymbol{X})$ that is closest to $h(\boldsymbol{Y})$ in the mean square sense. Of course, we must be able to compute the appropriate means, variances, and covariances.

Moreover, *Non-linear regression* with a single, real-valued predictor variable can be thought of as a special case of multiple linear regression. Thus, suppose that X is the predictor variable, Y is the response variable, and that $(g_1, g_2, \ldots, g_n)$ is a sequence of real-valued functions. We can apply the results of this section to find the linear function of $(g_1(X), g_2(X), \ldots, g_n(X))$ that is closest to Y in the mean square sense. We just replace X_i with $g_i(X)$ for each i. Again, we must be able to compute the appropriate means, variances, and covariances to do this.

Examples and Applications

Suppose that (X, Y) has probability density function f defined by $f(x, y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$. Find each of the following:

1. $\mathbb{E}(X, Y)$
2. $\vc(X, Y)$

Answer

1. $\left(\frac{7}{12}, \frac{7}{12}\right)$
2. $\begin{bmatrix} \frac{11}{144} & -\frac{1}{144} \\ -\frac{1}{144} & \frac{11}{144} \end{bmatrix}$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 2(x + y)$ for $0 \le x \le y \le 1$. Find each of the following:

1. $\mathbb{E}(X, Y)$
2. $\vc(X, Y)$

Answer

1. $\left(\frac{5}{12}, \frac{3}{4}\right)$
2. $\begin{bmatrix} \frac{43}{720} & \frac{1}{48} \\ \frac{1}{48} & \frac{3}{80} \end{bmatrix}$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 6x^2 y$ for $0 \le x \le 1$, $0 \le y \le 1$. Find each of the following:

1. $\mathbb{E}(X, Y)$
2. $\vc(X, Y)$

Answer

Note that X and Y are independent.

1. $\left(\frac{3}{4}, \frac{2}{3}\right)$
2. $\begin{bmatrix} \frac{3}{80} & 0 \\ 0 & \frac{1}{18} \end{bmatrix}$

Suppose that (X, Y) has probability density function f defined by $f(x, y) = 15x^2 y$ for $0 \le x \le y \le 1$. Find each of the following:

1. $\mathbb{E}(X, Y)$
2. $\vc(X, Y)$
3. $L(Y \mid X)$
4. $L\left[Y \mid \left(X, X^2\right)\right]$
5. Sketch the regression curves on the same set of axes.

Answer

1. $\left(\frac{5}{8}, \frac{5}{6}\right)$
2. $\begin{bmatrix} \frac{17}{448} & \frac{5}{336} \\ \frac{5}{336} & \frac{5}{252} \end{bmatrix}$
3. $\frac{10}{17} + \frac{20}{51} X$
4. $\frac{49}{76} + \frac{10}{57} X + \frac{7}{38} X^2$

Suppose that (X, Y, Z) is uniformly distributed on the region $\left\{(x, y, z) \in \mathbb{R}^3 : 0 \le x \le y \le z \le 1\right\}$. Find each of the following:

1. $\mathbb{E}(X, Y, Z)$
2. $\vc(X, Y, Z)$

3. $L[Z \mid (X, Y)]$
4. $L[Y \mid (X, Z)]$
5. $L[X \mid (Y, Z)]$
6. $L[(Y, Z) \mid X]$

Answer

1. $\left(\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right)$
2. $\begin{bmatrix} \frac{3}{80} & \frac{1}{40} & \frac{1}{80} \\ \frac{1}{40} & \frac{1}{20} & \frac{1}{40} \\ \frac{1}{80} & \frac{1}{40} & \frac{3}{80} \end{bmatrix}$
3. $\frac{1}{2} + \frac{1}{2}Y$. Note that there is no X term.
4. $\frac{1}{2}X + \frac{1}{2}Z$. Note that this is the midpoint of the interval $[X, Z]$.
5. $\frac{1}{2}Y$. Note that there is no Z term.
6. $\begin{bmatrix} \frac{1}{3} + \frac{2}{3}X \\ \frac{2}{3} + \frac{1}{3}X \end{bmatrix}$

Suppose that X is uniformly distributed on $(0, 1)$, and that given X, random variable Y is uniformly distributed on $(0, X)$. Find each of the following:

1. $\mathbb{E}(X, Y)$
2. $\text{vc}(X, Y)$

Answer

1. $\left(\frac{1}{2}, \frac{1}{4}\right)$
2. $\begin{bmatrix} \frac{1}{12} & \frac{1}{24} \\ \frac{1}{24} & \frac{7}{144} \end{bmatrix}$