Chapter 3

Means and covariances

1 Matrix notation

Suppose W is an $m \times k$ matrix whose elements are random variables. The expected value of W, written $\mathbb{E}W$, is the $m \times k$ matrix whose (i, j)th element equals $\mathbb{E}W_{i,j}$.

You should convince yourself that, if A is an $\ell \times m$ matrix of constants and B is an $\ell \times k$ matrix of constants then

$$\mathbb{E}(AW + B) = A\mathbb{E}(W) + B.$$

Covariances

If W is an $m \times 1$ vector of random variables with $\mathbb{E}W = \mu_w$ and Z is an $\ell \times 1$ vector of random variables with $\mathbb{E}Z = \mu_z$ then $\text{cov}(W, Z)$ is defined to be the $m \times \ell$ matrix with (i, j)th element

$$\text{cov}(W_i, Z_j) = \mathbb{E}\left((W_i - \mathbb{E}W_i)(Z_j - \mathbb{E}Z_j)\right) = \mathbb{E}(W_i Z_j) - (\mathbb{E}W_i)(\mathbb{E}Z_j).$$

In matrix form,

$$\text{cov}(W, Z) = \mathbb{E}\left((W - \mu_w)(Z - \mu_z)^T\right) = \mathbb{E}(WZ^T) - \mu_w \mu_z^T.$$

If A and B are matrices of constants for which AW and BZ are well defined then

$$\text{cov}(AW, BZ) = A\text{cov}(W, Z)B^T.$$

You should convince yourself of this fact.

For the special case where $W = Z$ the matrix $\text{cov}(W, W)$ is usually denoted by $\text{var}(W)$, so that

$$\text{var}(W)_{i,j} = \begin{cases} \text{var}(W_i) & \text{if } i = j \\ \text{cov}(W_i, W_j) & \text{if } i \neq j \end{cases} .$$

The trace trick

The trace of a square matrix D is defined to be the sum of its diagonal elements, $\text{trace}(D) = \sum_i D_{i,i}$. If F in an $m \times k$ matrix and G is a $k \times m$ matrix then

$$\text{trace}(FG) = \text{trace}(GF)$$

because both sides equal $\sum_{i,j} F_{i,j}G_{j,i}$.

<3.1> **Example.** Suppose y is an $n \times 1$ vector of random variables with $\mathbb{E}y = \mu$ and $\text{var}(y) = V$. Then

$$\mathbb{E}\left(\|y - \mu\|^2\right) = \mathbb{E}\left(\sum\nolimits_i (y_i - \mu_i)^2)\right) = \sum\nolimits_i \text{var}(y_i) = \text{trace}(V).$$

More directly, we could use the fact that

$$\|y - \mu\|^2 = \text{trace}\left((y - \mu)^T(y - \mu)\right) = \text{trace}\left((y - \mu)(y - \mu)^T\right)$$

so that

$$\mathbb{E}\,\|y - \mu\|^2 = \mathbb{E}\text{trace}\left((y - \mu)(y - \mu)^T\right) = \text{trace}\,\mathbb{E}\left((y - \mu)(y - \mu)^T\right) = \text{trace}(V).$$

Here I have used the fact that a number t is the same as the trace of the 1×1 matrix whose only element is t, and the fact that the expected value of a sum is the sum of the expected values.

□

2 The model

Up until now, I have treated least squares as just a method to approximate an $n \times 1$ vector y by a linear combination of the columns of some $n \times p$ matrix X. More succinctly, the problem has been to approximate y by a vector in the subspace $\mathcal{X}$ of $\mathbb{R}^n$ that is spanned by the columns of X. The best approximation can be written as $\widehat{y} = Hy$, where H is the "hat matrix", the matrix that projects vectors orthogonally onto $\mathcal{X}$.

Statistician also regard least squares as a method for estimating a "signal", an unknown vector μ that is assumed to belong to the subspace $\mathcal{X}$, when we observe "signal + noise",

$$y = \mu + \xi \qquad \text{with } \mu \in \mathcal{X}.$$

The simplest model assumes that the noise has zero means ($\mathbb{E}\xi_i = 0$ for all i), constant variances ($\text{var}(\xi_i) = \sigma^2$ for some unknown σ^2) and is uncorrelated ($\text{cov}(\xi_i, \xi_j) = 0$ for $i \neq j$).

If the matrix X is also regarded as random then the expected values, variances, and covariances should all be interpreted as conditional on X:

$$\mathbb{E}(\xi_i \mid X) = 0 \quad \text{AND} \quad \text{cov}(\xi, \xi_j \mid X) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

The calculations of expected values and covariances for fitted values and residuals are much cleaner when expressed in matrix form:

$$\mathbb{E}\xi = 0 \quad \text{AND} \quad \text{var}(\xi) = \sigma^2 I_n,$$

so that

$$\mathbb{E}y = \mu \quad \text{AND} \quad \text{var}(y) = \sigma^2 I_n.$$

It is easy to calculated and covariances for the fitted vector $\widehat{y}$ and the residual vector. First note that $\widehat{y} = \mu + H\xi$ because $H\mu = \mu$. Similarly $r = (I_n - H)y = (I_n - H)\xi$. Notice that, assuming the model is correct (better: under the modeling assumptions),

$$\begin{aligned} \mathbb{E}\widehat{y} &= \mu + H\mathbb{E}(\xi) = \mu \\ \mathbb{E}r &= (I_n - H)\mathbb{E}(\xi) = 0 \\ \text{var}(y) &= H\text{var}(\xi)H^T = \sigma^2 H \\ \text{var}(r) &= (I_n - H)\text{var}(\xi)(I_n - H^T) = \sigma^2(I_n - H) \\ \text{cov}(\widehat{y}, r) &= \text{cov}(H\xi, (I_n - H)\xi) = H\text{var}(\xi)(I - H_n)^T = 0. \end{aligned}$$

Here I have several times used the fact that $H = H^T = H^2$.

Remark. You should not confuse the probabilistic fact that (under the model) $\text{cov}(\widehat{y}, r) = 0$ with the geometric fact that $\langle \widehat{y}, r \rangle = 0$. The first is an assertion about the expected value of a particular $n \times n$ matrix; the second is a fact about a 1×1 matrix.

<3.2> **Example.** (coordinate-free version of Gauss-Markov) Suppose we are interested in estimators for some linear function $c^T\mu$ of the theoretical expected values. (That is, c is a vector of constants.) Suppose also that we are only interested in estimators that are linear functions of y and are ***unbiased*** as estimators of $c^T\mu$, that is,

$$\mathbb{E}_\mu(\ell^T y) = c^T\mu \qquad \text{for all } \mu \in \mathcal{X}.$$

The subscript μ on the $\mathbb{E}$ is to remind you that the unbiasedness is a property that should hold for all possible choices of μ in $\mathcal{X}$.

Problem: Which choice of ℓ makes

$$\text{var}(\ell^T y) = \ell^T \text{var}(y)\ell = \sigma^2 \|\ell\|^2$$

the smallest?

Unbiasedness requires that

$$\mathbb{E}(\ell^T y) = \ell^T\mu = c^T\mu \qquad \text{for all } \mu \text{ in } \mathcal{X}.$$

That is, we require that $(\ell - c)^T\mu = 0$ for all μ in $\mathcal{X}$, which means that $(\ell - c)$ must be orthogonal to $\mathcal{X}$. In other words, $\ell = c + L$ where L is some vector in $\mathcal{X}^\perp$. Equivalently

$$\ell = Hc + (L + (I_n - H)c),$$

a sum of a vector, Hc, in $\mathcal{X}$ and a vector orthogonal to $\mathcal{X}$. The last representation gives

$$\|\ell\|^2 = \|Hc\|^2 + \|L + (I_n - H)c\|^2$$

The right-hand side takes its smallest value when the $\mathcal{X}^\perp$ vector L is chosen to make $L + (I_n - H)c = 0$. That is, the minimum is achieved when $\ell = Hc$, so that $\ell^T y = c^T\widehat{y}$.

□

Remark. The fact that $c^T\widehat{y}$ is the linear function of y that has the smallest variance amongst all unbiased linear estimators of $c^T\mu$ is not particularly surprising, in my opinion. It also ignores two legitimate questions: Why should we consider only linear functions of y? And why should we require unbiasedness? Modern statistical theorists are quick to abandon those requirements for more complicated models.

3 Parameters (full rank case)

Suppose X has rank p with singular value decomposition

$$X = \sum_{i \le p} \lambda_i u_i v_i^T = U_1 \Lambda_1 V_1^T$$

for which $\lambda_i > 0$ for all $i \le p$. Here U_1 is an $n \times p$ matrix whose columns provide an orthonormal basis (onb) for $\mathcal{X}$, and $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_p)$, and V is a $p \times p$ matrix whose columns provide an onb for $\mathbb{R}^p$.

Recall that $H = U_1 U_1^T$ is then the hat matrix and

$$\widehat{b} = V \Lambda_1^{-1} U_1^T y$$

is the unique solution to the equation $Xb = \widehat{y}$. Similarly, the expected value $\mu = \mathbb{E}y$ has a unique representation as $X\beta$, where

$$\beta = V \Lambda^{-1} U_1^T \mu.$$

It is then natural to treat $\widehat{b}$ as an estimator for β. It might be comforting to note that

$$\mathbb{E}\widehat{b} = V \Lambda_1^{-1} U_1^T \mathbb{E}y = V \Lambda_1^{-1} U_1^T \mu = \beta$$

and

$$\text{var}(\widehat{b}) = V \Lambda_1^{-1} U_1^T \text{var}(y) (V \Lambda_1^{-1} U_1^T)^T = \sigma^2 V \Lambda_1^{-2} V^T = \sum_{i \le p} \frac{\sigma^2}{\lambda_i^2} v_i v_i^T.$$

You could use this representation on Homework 3 to find the unit vectors q in $\mathbb{R}^p$ for which $\text{var}(q^T \widehat{b})$ is the largest or smallest.

If you are no longer interested in minimum variance unbiased estimators then you could skip the next example.

<3.3> **Example.** (Gauss-Markov with coordinates) Find the ℓ for which $\ell^T y$ has the smallest variance amongst all unbiased estimators of $d^T \beta$, where d is a specified constant vector in $\mathbb{R}^p$.

This problem is really just a disguised form of Example <3.2> with $c = U_1 \Lambda_1^{-1} V^T d$, the choice for which $c^T \mu = d^T \beta$. We already know that the solution is

$$\ell = Hc = U_1 U_1^T U_1 \Lambda_1^{-1} V^T d = U_1 U_1 \Lambda_1^{-1} V^T d$$

so that

$$\ell^T y = d^T V \Lambda_1^{-1} U_1^T y = d^T \widehat{b}.$$

Surprise!

□

4 Traditional treatment of parameters (full rank case)

Just in case you were wondering, here is the way some of the results in the previous section are usually derived when X has full rank.

The value $\widehat{b}$ is the unique b for which $y - Xb$ is orthogonal to the columns of X. That is, it is the solution to the "normal equations",

$$X^T(y - Xb) = 0.$$

The $p \times p$ matrix $X^TX = V\Lambda_1^2V^T$ has rank p (Why?), which means that it has an inverse and

$$\widehat{b} = (X^TX)^{-1}X^Ty.$$

Compare with the svd representation:

$$(X^TX)^{-1}X^T = (V\Lambda_1^2V^T)^{-1}(V\Lambda_1U_1^T) = V\Lambda^{-1}U^T.$$

Thus

$$\mathbb{E}\widehat{b} = (X^TX)^{-1}X^T\mathbb{E}y = (X^TX)^{-1}X^TX\beta = \beta$$

and

$$\text{var}(\widehat{b}) = (X^TX)^{-1}X^T\text{var}(y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} = \sigma^2V\Lambda_1^{-2}V^T.$$

I think the final representation is more informative because it shows where problems can occur when X is ill-conditioned.

The fitted vector equals

$$X\widehat{b} = X(X^TX)^{-1}X^Ty,$$

which agrees with $\widehat{y} = Hy$ because

$$X(X^TX)^{-1}X^T = U_1\Lambda_1V^TV\Lambda_1^{-2}V^T(V\Lambda_1U_1^T) = U_1U_1^T = H.$$

5 Rank $< p$

Things get much more complicated (and interesting) when the matrix X is not of full rank. See the handout overparametrized.pdf for details.