5. Vector random variables

- If we have a collection of random variables $Y_1, Y_2, \ldots, Y_n$ we can gather them together into a vector random variable $\mathbf{Y}$.
- Suppose that, for each $i = 1, \ldots, n$ we have $\mathrm{E}[Y_i] = \mu_i$. Then, we write $\mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu}$ for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$.
- Now, write $\mathrm{Cov}(Y_i, Y_j) = V_{ij}$ for $i \neq j$ and $\mathrm{Var}(Y_i) = \mathrm{Cov}(Y_i, Y_i) = V_{ij}$. We call $\mathbb{V} = [V_{ij}]_{n \times n}$ the **variance-covariance matrix** for $\mathbf{Y}$.
- We can also call $\mathbb{V}$ the **covariance matrix** or, more simply, just the **variance matrix**. We write $\mathbb{V} = \mathrm{Var}(\mathbf{Y})$.

Example. Let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ be a vector consisting of n independent random variables, each with mean zero and variance σ^2. This is a common model for **measurement error** on n measurements. We have

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbb{I}$$

where $\mathbf{0} = (0, \ldots, 0)$ and $\mathbb{I}$ is the $n \times n$ identity matrix. The off-diagonal entries of $\mathrm{Var}(\boldsymbol{\epsilon})$ are zero since $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. For measurement error models, we break our usual rule of using upper case letters for random variables.

Example. A population version of the linear model

- First recall the sample version, which is

$$\text{(LM3)} \qquad \mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e},$$

where $\mathbf{y}$ is the measured response, $\mathbb{X}$ is an $n \times p$ matrix of explanatory variables, $\mathbf{b}$ is chosen by least squares, and $\mathbf{e}$ is the resulting vector of residuals.

- We want to build a random vector $\mathbf{Y}$ that provides a population model for the data $\mathbf{y}$. We write this as

$$\boxed{\text{(LM6)} \qquad \mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

where $\mathbb{X}$ is the same explanatory matrix as in (LM3), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is an unknown coefficient vector (we don't know the true population coefficient!) and $\boldsymbol{\epsilon}$ is measurement error with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$.

- Our model (LM6) asserts that the process which generated the response data $\mathbf{y}$ was like drawing a random vector $\mathbf{Y}$ consructed using a random measurement error model with known matrix $\mathbb{X}$ for some fixed but unknown value of $\boldsymbol{\beta}$.

Motivation for finding the means and variances of linear combinations of random variables

- Recall that the main purpose of having a probability model is so that we can investigate the chance variation due to picking the sample.
- Recall that for $(\mathrm{LM3})$, the least squares estimate is $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$.
- This is a **statistic**, which means a function of the data and not a random variable. We cannot properly talk about the mean and variance of $\mathbf{b}$.
- We can work out the mean and variance of $\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$, as long as we know how to work out the mean and variance of linear combinations.
- As long as $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a **useful probability model** for the relationship between the response variable $\mathbf{y}$ and the explanatory variable $\mathbb{X}$, calculations done with this model may be useful.

A digression on “useful” models

“Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an *ideal* gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question ‘Is the model true?’. If *truth* is to be the *whole truth* the answer must be *No*. The only question of interest is ‘Is the model illuminating and useful.’ ” (Box, 1978)

“**Essentially, all models are wrong, but some are useful.**”
(Box and Draper, 1987)

- Perhaps the most useful statistical model ever is $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Anything so widely used is also widely abused. Our task is to understand $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ so that we can be users and not abusers.

Mean of a linear combination, in matrix form

- The linear property of expectation lets us take expectation through a summation sign, and we get

$$\mathrm{E}\Big[\sum_{j=1}^{n} a_{ij} Y_j\Big] = \sum_{j=1}^{n} a_{ij}\mathrm{E}[Y_j].$$

- In matrix form, with $\mathbb{A} = [a_{ij}]$, this is $\boxed{\mathrm{E}\left[\mathbb{A}\mathbf{Y}\right] = \mathbb{A}\mathrm{E}[\mathbf{Y}].}$

Example. For $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have $\boxed{\mathrm{E}[\mathbf{Y}] = \mathbb{X}\boldsymbol{\beta} + \mathrm{E}[\boldsymbol{\epsilon}] = \mathbb{X}\boldsymbol{\beta}}$

Example. For $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$, we have

$$\boxed{\mathrm{E}[\hat{\boldsymbol{\beta}}] = \mathrm{E}\big[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}\big] = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathrm{E}[\mathbf{Y}] = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\beta}}$$

- Interpretation: If the data $\mathbf{y}$ are well modeled as a draw from the probability model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the least squares estimate $\mathbf{b}$ is well modeled by a random vector centered around $\boldsymbol{\beta}$.

Linearity of expectation

- We have seen several versions of the same property that expectations can be moved through sums and multiplicative constants:

$$\begin{aligned}
\mathrm{E}[aX+b] &= a\mathrm{E}[X]+b, \\
\mathrm{E}\left[\sum_{i=1}^{n} a_i Y_i\right] &= \sum_{i=1}^{n} a_i \mathrm{E}[Y_i], \\
\mathrm{E}\left[\sum_{j=1}^{n} a_{ij} Y_j\right] &= \sum_{j=1}^{n} a_{ij} \mathrm{E}[Y_j]. \\
\mathrm{E}[\mathbb{A}\mathbf{Y}] &= \mathbb{A}\mathrm{E}[\mathbf{Y}]
\end{aligned}$$

- These properties are collectively known as **linearity**.
- Why? Maybe because these properties mean that linear equations for random variables lead to linear equations for their expectations.
- The linearity property means E follows a **distributive rule**. We can distribute E across sums just as we are used to doing in basic arithmetic.

Exercises using linearity

Question 5.1. Use basic properties of expectation, and the definition of covariance, to show that $\boxed{\operatorname{Cov}\big(aX+b, cY+d\big) = ac\operatorname{Cov}(X,Y).}$

Question 5.2. Similarly, show that $\operatorname{Cov}\big(Y, \sum_{j=1}^{n} Z_j\big) = \sum_{j=1}^{n} \operatorname{Cov}(Y, Z_j)$.

Moving sums through covariance

Question 5.3. Using Question 5.2, show that

$$\boxed{\mathrm{Cov}\big(\textstyle\sum_{i=1}^m Y_i, \sum_{j=1}^n Z_j\big) = \sum_{i=1}^m \sum_{j=1}^n \mathrm{Cov}(Y_i, Z_j).}$$

- This formula is sometimes called the **bilinearity** of covariance, since $\mathrm{Cov}(Y, Z)$ is linear in Y and linear in Z.
- This is also our first use of double summation.
- Think of $\sum_{i=1}^m \sum_{j=1}^n$ as summing all the entries in an $m \times n$ table, or equivalently, summing entries in an $m \times n$ matrix of covariances.

Variance of a sum

Question 5.4. Using Question 5.3, show that

$$\boxed{\mathrm{Var}\big(\sum_{i=1}^n Y_i\big) = \sum_{i=1}^n \mathrm{Var}(Y_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).}$$

- Thinking of $\sum_{i=1}^n \sum_{j=1}^n$ as summing all the entries in an $n \times n$ table, $\sum_{i<j}$ means summing over all the entries above the diagonal.
- **Covariance is symmetric**, meaning $\mathrm{Cov}(Y, Z) = \mathrm{Cov}(Z, Y)$, and so the table of covariances is symmetric about its diagonal.
- Thinking of the table of covariances as a matrix, the covariance matrix is called a **symmetric matrix**.

The covariance matrix of a linear combination

- Suppose the length n random vector $\mathbf{Y}$ has variance matrix $\mathbb{V}_Y$.
- Let $\mathbb{A} = [a_{ij}]$ be an $m \times n$ matrix and let $\mathbf{Z} = \mathbb{A}\mathbf{Y}$.
- $\mathbf{Z}$ is a length m random vector. Call its variance matrix $\mathbb{V}_Z$.
- Can we find $\mathbb{V}_Z$ if we know $\mathbb{V}_Y$ and $\mathbb{A}$?
- Doing this will let us find the variances and correlations between any collection of linear combinations of $\mathbf{Y}$, a useful thing for working with the linear model.

- To find the entries in the $m \times m$ covariance matrix $\mathbb{V}_Z$, we need to work out $\mathrm{Cov}(Z_i, Z_j)$ for each entry (i, j) in the matrix.
- Recall that $Z_i = \sum_{k=1}^{n} a_{ik} Y_k$.
- Since Z_i and Z_j are linear combinations of $\mathbf{Y}$, we can use our formulas for bilinearity of covariance (a consequence of linearity of expectation combined with the definition of covariance) to find $\mathrm{Cov}(Z_i, Z_j)$.

Covariance of $Z_i = [\mathbb{A}\mathbf{Y}]_i$ and $Z_j = [\mathbb{A}\mathbf{Y}]_j$

Question 5.5. Show that $\mathrm{Cov}(Z_i, Z_j) = \sum_{k=1}^{n} \sum_{\ell=1}^{n} a_{ij} a_{k\ell} [\mathbb{V}_Y]_{k\ell}$

Question 5.6. Show that $\boxed{\mathbb{V}_Z = \mathbb{A}^{\mathrm{T}} \mathbb{V}_Y \mathbb{A}.}$

Covariance of the least squares coefficients

- The covariance matrix formula we just developed can be written as

$$\boxed{\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}^{\mathrm{T}}\mathrm{Var}(\mathbf{Y})\mathbb{A}.}$$

Question 5.7. Consider the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$. Apply this variance formula to $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$ to get

$$\boxed{\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}}$$

Standard errors for the linear model

- The formula $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \big(\mathbb{X}^{\mathrm{T}}\mathbb{X}\big)^{-1}$ needs extra work to be useful for data analysis.
- In practice, we know the model matrix $\mathbb{X}$ but we don't know the measurement standard deviation σ.
- An estimate of the measurement error is the sample standard deviation of the residuals.
- For $\mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e}$ with $\mathbb{X}$ being $n \times p$, an estimate of σ is

$$\boxed{s = \sqrt{\tfrac{1}{n-p}\textstyle\sum_{i=1}^{n}\big(y_i - \hat{y}_i\big)^2} = \sqrt{\tfrac{1}{n-p}\textstyle\sum_{i=1}^{n}\big(y_i - [\mathbb{X}\mathbf{b}]_i\big)^2}}$$

- We will discuss later why we choose to divide by $n-p$.
- The **standard error** of b_k for $k = 1, \dots, p$ is

$$\boxed{\mathrm{SE}(b_k) = s\sqrt{\Big[\big(\mathbb{X}^{\mathrm{T}}\mathbb{X}\big)^{-1}\Big]_{kk}}}$$

- $SE(b_k)$ is an estimate of $\sqrt{\big[\mathrm{Var}(\hat{\boldsymbol{\beta}})\big]_{kk}}$.
- Let's check we now understand how `lm()` gets standard errors in R.