3 Random vectors and multivariate normal distribution

As we saw in Chapter 1, a natural way to think about repeated measurement data is as a series of **random vectors**, one vector corresponding to each unit. Because the way in which these vectors of measurements turn out is governed by probability, we need to discuss extensions of usual **univariate** probability distributions for (scalar) random variables to **multivariate** probability distributions governing random vectors.

3.1 Preliminaries

First, it is wise to review the important concepts of random variable and probability distribution and how we use these to model individual observations.

RANDOM VARIABLE: We may think of a **random variable** Y as a characteristic whose values may **vary**. The way it takes on values is described by a **probability distribution**.

CONVENTION, REPEATED: It is customary to use upper case letters, e.g Y, to denote a generic random variable and lower case letters, e.g. y, to denote a particular value that the random variable may take on or that may be observed (data).

EXAMPLE: Suppose we are interested in the characteristic "body weight of rats" in the population of all possible rats of a certain age, gender, and type. We might let

$$Y = \text{ body weight of a (randomly chosen) rat}$$

from this population. Y is a random variable.

We may conceptualize that body weights of rats are **distributed** in this population in the sense that some values are more common (i.e. more rats have them) than others. If we randomly select a rat from the population, then the chance it has a certain body weight will be governed by this distribution of weights in the population. Formally, values that Y may take on are **distributed** in the population according to an associated **probability distribution** that describes how likely the values are in the population.

In a moment, we will consider more carefully **why** rat weights we might see **vary**. First, we recall the following.

(POPULATION) MEAN AND VARIANCE: Recall that the **mean** and **variance** of a probability distribution summarize notions of "center" and "spread" or "variability" of all possible values. Consider a random variable Y with an associated probability distribution.

The **population mean** may be thought of as the average of all possible values that Y could take on, so the average of all possible values across the entire distribution. Note that some values occur more frequently (are more likely) than others, so this average reflects this. We write

$$E(Y). \tag{3.1}$$

to denote this average, the **population mean**. The **expectation operator** E denotes that the "averaging" operation over all possible values of its argument is to be carried out. Formally, the average may be thought of as a "weighted" average, where each possible value is represented in accordance to the **probability** with which it occurs in the population. The symbol "μ" is often used.

The population mean may be thought of as a way of describing the "center" of the distribution of all possible values. The population mean is also referred to as the **expected value** or **expectation** of Y.

Recall that if we have a **random sample** of observations on a random variable Y, say $Y_1, \ldots, Y_n$, then the **sample mean** is just the average of these:

$$\overline{Y} = n^{-1} \sum_{j=1}^{n} Y_j.$$

For example, if $Y =$ rat weight, and we were to obtain a random sample of $n = 50$ rats and weigh each, then $\overline{Y}$ represents the average we would obtain.

- The sample mean is a natural **estimator** for the **population mean** of the probability distribution from which the random sample was drawn.

The **population variance** may be thought of as measuring the spread of all possible values that may be observed, based on the squared deviations of each value from the "center" of the distribution of all possible values. More formally, variance is based on averaging squared deviations across the population, which is represented using the expectation operator, and is given by

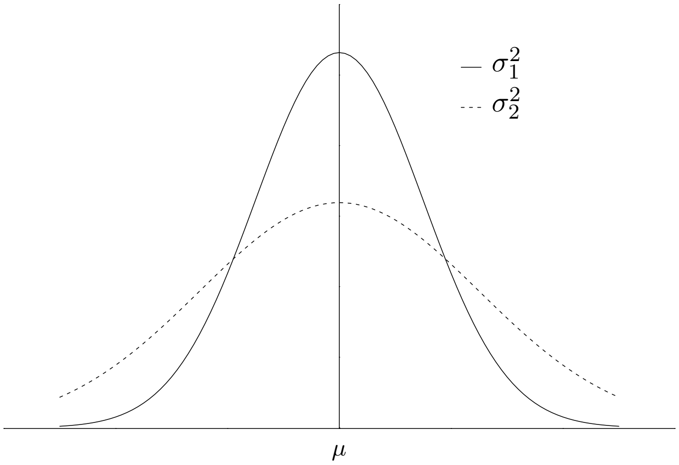
$$\text{var}(Y) = E\{(Y-\mu)^2\}, \quad \mu = E(Y). \tag{3.2}$$

(3.2) shows the interpretation of variance as an average of squared deviations from the mean across the population, taking into account that some values are more likely (occur with higher probability) than others.

- The use of squared deviations takes into account magnitude of the distance from the "center" but not direction, so is attempting to measure only "spread" (in either direction).

The symbol "σ^2" is often used generically to represent population variance. Figure 1 shows two normal distributions with the same mean but different variances $\sigma_1^2 < \sigma_2^2$, illustrating how variance describes the "spread" of possible values.

Figure 1: *Normal distributions with mean μ but different variances*



Variance is on the scale of the response, squared. A measure of spread that is on the same scale as the response is the **population standard deviation**, defined as $\sqrt{\text{var}(Y)}$. The symbol σ is often used.

Recall that for a random sample as above, the **sample variance** is (almost) the average of the squared deviations of each observation Y_j from the sample mean $\overline{Y}$.

$$S^2 = (n-1)^{-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2.$$

- The sample variance is used as an **estimator** for population variance. Division by $(n-1)$ rather than n is used so that the estimator is **unbiased**, i.e estimates the true population variance well even if the sample size n is small.

- The **sample standard deviation** is just the square root of the sample variance, often represented by the symbol S.

GENERAL FACTS: If b is a fixed scalar and Y is a random variable, then

- $E(bY) = bE(Y) = b\mu$; i.e. all values in the average are just multiplied by b. Also, $E(Y + b) = E(Y) + b$; adding a constant to each value in the population will just shift the average by this same amount.
- $\text{var}(bY) = E\{(bY - b\mu)^2\} = b^2\text{var}(Y)$; i.e. all values in the average are just multiplied by b^2. Also, $\text{var}(Y + b) = \text{var}(Y)$; adding a constant to each value in the population does not affect how they vary about the mean (which is also shifted by this amount).

SOURCES OF VARIATION: We now consider why the values of a characteristic that we might observe **vary**. Consider again the rat weight example.

- *Biological variation.* It is well-known that biological entities are different; although living things of the same type tend to be similar in their characteristics, they are not exactly the same (except perhaps in the case of genetically-identical clones). Thus, even if we focus on rats of the same strain, age, and gender, we expect variation in the possible weights of such rats that we might observe due to inherent, natural **biological variation**.

  Let Y represent the weight of a randomly chosen rat, with probability distribution having mean μ. If all rats were biologically identical, then the population variance of Y would be equal to 0, and we would expect all rats to have exactly weight μ. Of course, because rat weights vary as a consequence of biological factors, the variance is > 0, and thus the weight of a randomly chosen rat is not equal to μ but rather **deviates** from μ by some positive or negative amount. From this view, we might think of Y as being represented by

  $$Y = \mu + b, \tag{3.3}$$

  where b is a random variable, with population mean $E(b) = 0$ and variance $\text{var}(b) = \sigma_b^2$, say.

  Here, Y is "decomposed" into its mean value (a **systematic** component) and a **random deviation** b that represents by how much a rat weight might deviate from the mean rat weight due to inherent biological factors.

  (3.3) is a simple **statistical model** that emphasizes that we believe rat weights we might see vary because of biological phenomena. Note that (3.3) implies that $E(Y) = \mu$ and $\text{var}(Y) = \sigma_b^2$.

- *Measurement error.* We have discussed rat weight as though, once we have a rat in hand, we may know its weight exactly. However, a scale usually must be used. Ideally, a scale should register the true weight of an item each time it is weighed, but, because such devices are imperfect, measurements on the same item may vary time after time. The amount by which the measurement differs from the truth may be thought of as an **error**; i.e. a deviation up or down from the true value that could be observed with a "perfect" device. A "fair" or **unbiased** device does not systematically register high or low most of the time; rather, the errors may go in either direction with no pattern.

  Thus, if we only have an unbiased scale on which to weigh rats, a rat weight we might observe reflects not only the true weight of the rat, which varies across rats, but also the error in taking the measurement. We might think of a random variable e, say, that represents the error that might contaminate a measurement of rat weight, taking on possible values in a hypothetical "population" of all such errors the scale might commit.

  We still believe rat weights vary due to biological variation, but what we see is also subject to measurement error. It thus makes sense to revise our thinking of what Y represents, and think of Y = "**measured** weight of a randomly chosen rat." The population of all possible values Y could take on is all possible values of rat weight we might measure; i.e., all values consisting of a true weight of a rat from the population of all rats contaminated by a measurement error from the population of all possible such errors.

  With this thinking, it is natural to represent Y as

  $$Y = \mu + b + e = \mu + \epsilon, \tag{3.4}$$

  where b is as in (3.3). e is the deviation due to measurement error, with $E(e) = 0$ and $\text{var}(e) = \sigma_e^2$, representing an unbiased but imprecise scale.

  In (3.4), $\epsilon = b + e$ represents the **aggregate** deviation due to the effects of **both** biological variation and measurement error. Here, $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2 = \sigma_b^2 + \sigma_e^2$, so that $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2$ according to the model (3.4). Here, σ^2 reflects the "spread" of measured rat weights and depends on both the spread in true rat weights **and** the spread in errors that could be committed in measuring them.

There are still further sources of variation that we could consider; we defer discussion to later in the course. For now, the important message is that, in considering statistical models, it is critical to be aware of different **sources of variation** that cause observations to vary. This is especially important with longitudinal data, as we will see.

We now consider these concepts in the context of a familiar statistical model.

SIMPLE LINEAR REGRESSION: Consider the simple linear regression model. At each fixed value $x_1, \ldots, x_n$, we observe a corresponding random variable Y_j, $j = 1, \ldots, n$. For example, suppose that the x_j are doses of a drug. For each x_j, a rat is randomly chosen and given this dose. The associated response for the jth rat (given dose x_j) may be represented by Y_j.

The simple linear regression model as usually stated is

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

where ϵ_j is a random variable with mean 0 and variance σ^2; that is $E(\epsilon_j) = 0, \ \ \text{var}(\epsilon_j) = \sigma^2$. Thus, $E(Y_j) = \beta_0 + \beta_1 x_j$ and $\text{var}(Y_j) = \sigma^2$.

This model says that, ideally, at each x_j, the response of interest, Y_j, should be exactly equal to the fixed value $\beta_0 + \beta_1 x_j$, the **mean** of Y_j. However, because of factors like (i) biological variation and (ii) measurement error, the values we might see at x_j vary. In the model, ϵ_j represents the deviation from $\beta_0 + \beta_1 x_j$ that might occur because of the aggregate effect of these sources of variation.

If Y_j is a continuous random variable, it is often the case that the **normal distribution** is a reasonable probability model for the population of ϵ_j values; that is,

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2).$$

This says that the total effect of all sources of variation is to create deviations from the mean of Y_j that may be equally likely in either direction as dictated by the **symmetric** normal probability distribution.

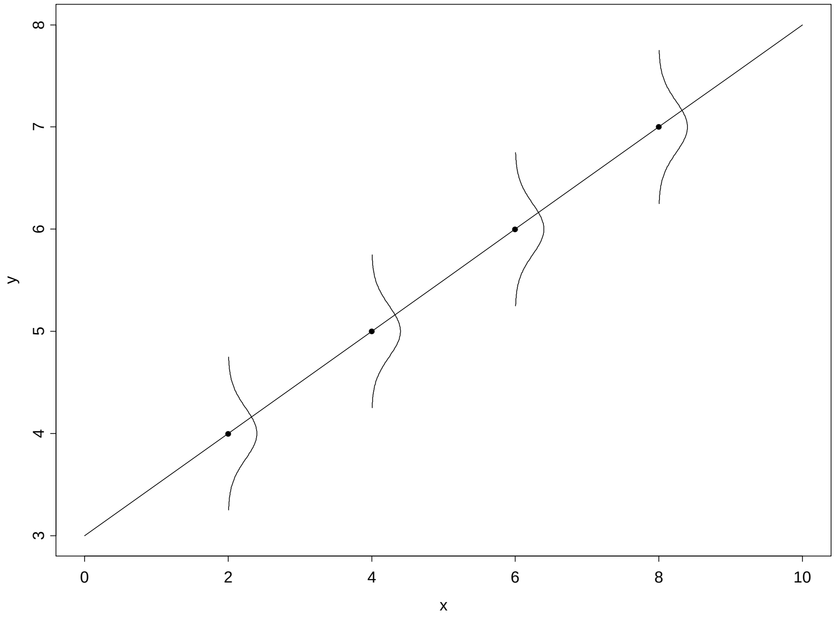
Under this assumption, we have that the population of observations we might see at a particular x_j is also normal and centered at $\beta_0 + \beta_1 x_j$; i.e.

$$Y_j \sim \mathcal{N}(\beta_0 + \beta_1 x_j, \, \sigma^2).$$

- This model says that the chance of seeing Y_j values above or below the mean $\beta_0 + \beta_1 x_j$ is the same (symmetry).
- This is an especially good model when the **predominant** source of variation (represented by the ϵ_j) is due to a measuring device.
- It may or may not be such a good model when the predominant source of variation is due to biological phenomena (more later in the course!).

The model thus says that, at each x_j, there is a population of possible Y_j values we might see, with mean $\beta_0 + \beta_1 x_j$ and variance σ^2. We can represent this pictorially by considering Figure 2.

Figure 2: *Simple linear regression*



"ERROR": An unfortunate convention in the literature is that the ϵ_j are referred to as **errors**, which causes some people to believe that they represent **solely** deviation due to measurement error. We prefer the term **deviation** to emphasize that Y_j values may deviate from $\beta_0 + \beta_1 x_j$ due to the combined effects of **several** sources (but not limited to measurement error).

INDEPENDENCE: An important assumption for simple linear regression and, indeed, more general problems, is that the random variables Y_j, or equivalently, the ϵ_j, are **independent**.

(Statistical) independence is a formal statistical concept with an important practical interpretation. In particular, in our simple linear regression model, this says that the way in which Y_j at x_j takes on its values is completely **unrelated** to the way in which $Y_{j'}$ observed at another position $x_{j'}$ takes on its values. This is certainly a reasonable assumption in many situations.

- In our example, where x_j are doses of a drug, each given to a different rat, there is no reason to believe that responses from different rats should be related in any way. Thus, the way in which Y_j values turn out at different x_j would be totally unrelated.

The consequence of independence is that we may think of data on an **observation-by-observation** basis; because the behavior of each observation is unrelated to that of others, we may talk about each one in its own right, without reference to the others.

Although this way of thinking may be relevant for regression problems where the data were collected according to a scheme like that in the example above, as we will see, it may not be relevant for longitudinal data.

3.2 Random vectors

As we have already mentioned, when several observations are taken on the **same** unit, it will be convenient, and in fact, necessary, to talk about them together. We thus must extend our way of thinking about random variables and probability distributions.

RANDOM VECTOR: A random vector is a vector whose elements are random variables. Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

be a $(n \times 1)$ random vector.

- Each element of $\boldsymbol{Y}$, Y_j, $j = 1, \ldots, n$, is a random variable with its own mean, variance, and probability distribution; e.g.

  $$E(Y_j) = \mu_j, \quad \text{var}(y_j) = E\{(Y_j - \mu_j)^2\} = \sigma_j^2.$$

  We might furthermore have that Y_j is normally distributed; i.e.

  $$Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2).$$

- Thus, if we talk about a particular element of $\boldsymbol{Y}$ **in its own right**, we may speak in terms of its particular probability distribution, mean, and variance.
- Probability distributions for single random variables are often referred to as **univariate**, because they refer only to how one (scalar) random variable takes on its values.

JOINT VARIATION: However, if we think of the elements of $\boldsymbol{Y}$ together, we must consider the fact that they come together in a group, so that there might be **relationships** among them. Specifically, if we think of $\boldsymbol{Y}$ as containing possible observations on the same unit at times indexed by j, there is reason to expect that the value observed at one time and that observed at another time may turn out the way they do in a "common" fashion. For example,

- If $\boldsymbol{Y}$ consists of the heights of a pine seedling measured on each of n consecutive days, we might expect a "large" value one day to be followed by a "large" value the next day.
- If $\boldsymbol{Y}$ consists of the lengths of baby rats in a litter of size n from a particular mother, we might expect all the babies in a litter to be "large" or "small" relative to babies from other litters.

This suggests that if observations can be naturally thought to arise together, then they may not be legitimately viewed as **independent**, but rather **related** somehow.

- In particular, they may be thought to **vary together**, or **covary**.
- This suggests that we need to think of how they take on values **jointly**.

JOINT PROBABILITY DISTRIBUTION: Just as we think of a probability distribution for a random variable as describing the frequency with which the variable may take on values, we may think of a **joint** probability distribution that describes the frequency with which an entire set of random variables takes on values **together**. Such a distribution is referred to as **multivariate** for obvious reasons. We will consider the specific case of the **multivariate normal distribution** shortly.

We may thus think of any two random variables in $\boldsymbol{Y}$, Y_j and Y_k, say, as having a joint probability distribution that describes how they take on values together.

COVARIANCE: A measure of how two random variable vary together is the **covariance**. Formally, suppose Y_j and Y_k are two random variables that vary together. Each of them has its own probability distribution with means μ_j and μ_k, respectively, which is relevant when we think of them separately. They also have a joint probability distribution, which is relevant when we think of them together. Then we define the **covariance** between Y_j and Y_k as

$$\text{cov}(Y_j, Y_k) = E\{(Y_j - \mu_j)(Y_k - \mu_k)\}. \tag{3.5}$$

Here, the expectation operator denotes average over all possible pairs of values Y_j and Y_k may take on together according to their joint probability distribution.

Inspection of (3.5) shows

- Covariance is defined as the average across all possible values that Y_j and Y_k may take on jointly of the product of the deviations of Y_j and Y_k from their respective means.
- Thus note that if "large" values ("larger" than their means) of Y_j and Y_k tend to happen **together** (and thus "small" values of Y_j and Y_k tend to happen together), then the two deviations $(Y_j - \mu_j)$ and $(Y_k - \mu_k)$ will tend to be **positive** together and **negative** together, so that the product

  $$(Y_j - \mu_j)(Y_k - \mu_k) \tag{3.6}$$

  will tend to be positive for most of the pairs of values in the population. Thus, the average in (3.5) will likely be positive.
- Conversely, if "large" values of Y_j tend to happen coincidently with "small" values of Y_k and vice versa, then the deviation $(Y_j - \mu_j)$ will tend to be positive when $(Y_k - \mu_k)$ tends to be negative, and vice versa. Thus the product (3.6) will tend to be negative for most of the pairs of values in the population. Thus, the average in (3.5) will likely be negative.
- Moreover, if in truth Y_j and Y_k are **unrelated**, so that "large" Y_j are likely to happen with "small" Y_k **and** "large" Y_k and vice versa, then we would expect the deviations $(Y_j - \mu_j)$ and $(Y_k - \mu_k)$ to be positive and negative in no real systematic way. Thus, (3.6) may be negative or positive with no special tendency, and the average in (3.5) would likely be zero.

Thus, the quantity of **covariance** defined in (3.5) makes intuitive sense as a measure of how "associated" values of Y_j are with values of Y_k.

- In the last bullet above, Y_j and Y_k are **unrelated**, and we argued that $\text{cov}(Y_j, Y_k) = 0$. In fact, formally, if Y_j and Y_k are statistically independent, then it follows that $\text{cov}(Y_j, Y_k) = 0$.
- Note that $\text{cov}(Y_j, Y_k) = \text{cov}(Y_k, Y_j)$.
- Fact: the covariance of a random variable Y_j and **itself**,

  $$\text{cov}(Y_j, Y_j) = E\{(Y_j - \mu_j)(Y_j - \mu_j)\} = \text{var}(Y_j) = \sigma_j^2.$$
- Fact: If we have two random variables, Y_j and Y_k, then

  $$\text{var}(Y_j + Y_k) = \text{var}(Y_j) + \text{var}(Y_k) + 2\text{cov}(Y_j, Y_k).$$

> That is, the variance of the population consisting of all possible values of the sum $Y_j + Y_k$ is the sum of the variances for each population, adjusted by how "associated" the two values are. Note that if Y_j and Y_k are independent, $\text{var}(Y_j + Y_k) = \text{var}(Y_j) + \text{var}(Y_k)$.

We now see how all of this information is summarized.

EXPECTATION OF A RANDOM VECTOR: For an entire n-dimensional vector random $\boldsymbol{Y}$, we summarize the means for each element in a vector

$$\boldsymbol{\mu} = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}.$$

We define the expected value or mean of $\boldsymbol{Y}$ as

$$E(\boldsymbol{Y}) = \boldsymbol{\mu};$$

the expectation operation is applied to each element in the vector $\boldsymbol{Y}$, yielding the vector $\boldsymbol{\mu}$ of means.

RANDOM MATRIX: A random matrix is simply a matrix whose elements are random variables; we will see a specific example of importance to us in a moment. Formally, if $\boldsymbol{\mathcal{Y}}$ is a $(r \times c)$ matrix with element Y_{jk}, each a random variable, then each element has an expectation, $E(Y_{jk}) = \mu_{jk}$, say. Then the expected value or mean of $\boldsymbol{\mathcal{Y}}$ is defined as the corresponding matrix of means; i.e.

$$E(\boldsymbol{\mathcal{Y}}) = \begin{pmatrix} E(Y_{11}) & E(Y_{12}) & \cdots & E(Y_{1c}) \\ \vdots & \vdots & \vdots & \vdots \\ E(Y_{r1}) & E(Y_{r2}) & \cdots & E(Y_{rc}) \end{pmatrix}.$$

COVARIANCE MATRIX: We now see how this concept is used to summarize information on covariance among the elements of a random vector. Note that

$$(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})' = \begin{pmatrix} (Y_1-\mu_1)^2 & (Y_1-\mu_1)(Y_2-\mu_2) & \cdots & (Y_1-\mu_1)(Y_n-\mu_n) \\ (Y_2-\mu_2)(Y_1-\mu_1) & (Y_2-\mu_2)^2 & \cdots & (Y_2-\mu_2)(Y_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (Y_n-\mu_n)(Y_1-\mu_1) & (Y_n-\mu_n)(Y_2-\mu_2) & \cdots & (Y_n-\mu_n)^2 \end{pmatrix},$$

which is a random matrix.

Note then that

$$
\begin{aligned}
E\{(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})'\} &= \begin{pmatrix} E(Y_1-\mu_1)^2 & E(Y_1-\mu_1)(Y_2-\mu_2) & \cdots & E(Y_1-\mu_1)(Y_n-\mu_n) \\ E(Y_2-\mu_2)(Y_1-\mu_1) & E(Y_2-\mu_2)^2 & \cdots & E(Y_2-\mu_2)(Y_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(Y_n-\mu_n)(Y_1-\mu_1) & E(Y_n-\mu_n)(Y_2-\mu_2) & \cdots & E(Y_n-\mu_n)^2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} = \boldsymbol{\Sigma},
\end{aligned}
$$

say, where for $j,k=1,\ldots,n$, $\text{var}(Y_j)=\sigma_j^2$ and we define

$$\text{cov}(Y_j,Y_k)=\sigma_{jk}.$$

The matrix $\boldsymbol{\Sigma}$ is called the **covariance matrix** or **variance-covariance matrix** of $\boldsymbol{Y}$.

- Note that $\sigma_{jk}=\sigma_{kj}$, so that $\boldsymbol{\Sigma}$ is a **symmetric**, **square** matrix.
- We will write succinctly $\text{var}(\boldsymbol{Y})=\boldsymbol{\Sigma}$ to state that the random vector $\boldsymbol{Y}$ has covariance matrix $\boldsymbol{\Sigma}$.

JOINT PROBABILITY DISTRIBUTION: It follows that, if we consider the joint probability distribution describing how the entire set of elements of $\boldsymbol{Y}$ take on values together, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the features of this distribution characterizing "center" and "spread **and** association."

- $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are referred to as the **population mean** and **population covariance** (matrix) for the population of data vectors represented by the joint probability distribution.
- The symbols $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are often used generically to represent population mean and covariance, as above.

CORRELATION: It is informative to separate the information on "spread" contained in variances σ_j^2 from that describing "association." Thus, we define a particular measure of association that takes into account the fact that different elements of $\boldsymbol{Y}$ may vary differently on their own.

The **population correlation coefficient** between Y_j and Y_k is defined as

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_j^2}\sqrt{\sigma_k^2}}.$$

Of course, $\sigma_j = \sqrt{\sigma_j^2}$ is the population standard deviation of Y_j, on the same scale of measurement as Y_j, and similarly for Y_k.

- ρ_{jk} scales the information on association in the covariance in accordance with the magnitude of variation in each random variable, creating a "unitless" measure. Thus, it allows one to think of the associations among variables measured on different scales.
- $\rho_{jk} = \rho_{kj}$.
- Note that if $\sigma_{jk} = \sigma_j \sigma_k$, then $\rho_{jk} = 1$. Intuitively, if this is true, it says that the ways Y_j and Y_k vary separately is identical to how they vary together, so that if we know one, we know the other. Thus, a correlation of 1 indicates that the two random variables are "perfectly positively associated." Similarly, if $\sigma_{jk} = -\sigma_j \sigma_k$, then $\rho_{jk} = -1$ and by the same reasoning they are "perfectly negatively associated."
- Clearly, $\rho_{jj} = 1$, so a random variable is perfectly positively correlated with itself.
- It may be shown that correlations must satisfy $-1 \leq \rho_{jk} \leq 1$.
- If $\sigma_{jk} = 0$ then $\rho_{jk} = 0$, so if Y_j and Y_k are independent, then they have 0 correlation.

CORRELATION MATRIX: It is customary to summarize the information on correlations in a matrix: The **correlation matrix $\boldsymbol{\Gamma}$** is defined as

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}.$$

For now, we use the symbol $\boldsymbol{\Gamma}$ to denote the correlation matrix of a random vector.

ALTERNATIVE REPRESENTATION OF COVARIANCE MATRIX: Note that knowledge of the variances $\sigma_1^2, \ldots, \sigma_n^2$ and the correlation matrix $\mathbf{\Gamma}$ is equivalent to knowledge of $\mathbf{\Sigma}$, and vice versa. It is often easier to think of associations among random variables on the unitless correlation scale than in terms of covariance; thus, it is often convenient to write the covariance matrix another way that presents the correlations explicitly.

Define the "standard deviation" matrix

$$\boldsymbol{T}^{1/2} = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{pmatrix}.$$

The "1/2" reminds us that this is a diagonal matrix with the square roots of the variances on the diagonal. Then it may be verified that (try it)

$$\boldsymbol{T}^{1/2}\mathbf{\Gamma}\boldsymbol{T}^{1/2} = \mathbf{\Sigma}. \tag{3.7}$$

The representation (3.7) will prove convenient when we wish to discuss associations implied by models for longitudinal data in terms of correlations. Moreover, it is useful to appreciate (3.7), as it allows calculations involving $\mathbf{\Sigma}$ that we will see later to be implemented easily on a computer.

GENERAL FACTS: As we will see later, we will often be interested in **linear combinations** of the elements of a random vector $\boldsymbol{Y}$; that is, functions of the form

$$c_1 Y_1 + \cdots c_n Y_n,$$

which may be written succinctly as $\boldsymbol{c}'\boldsymbol{Y}$, where $\boldsymbol{c}$ is the column vector

$$\boldsymbol{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}.$$

- Note that $\boldsymbol{c}'\boldsymbol{Y}$ is a **scalar** quantity.

It is possible using facts on the multiplication random variables by scalars (see above) and the definitions of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ to show that

$$E(\boldsymbol{c}'\boldsymbol{Y}) = \boldsymbol{c}'\boldsymbol{\mu} \quad \text{var}(\boldsymbol{c}'\boldsymbol{Y}) = \boldsymbol{c}'\mathbf{\Sigma}\boldsymbol{c}.$$

(Try to verify these.)

More generally, if we have a set of q such linear combinations defined by vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, we may summarize them all in a matrix whose rows are the $\boldsymbol{c}_k'$; i.e.

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{q1} & \cdots & c_{qn} \end{pmatrix}.$$

Then $\boldsymbol{CY}$ is a $(q \times 1)$ random vector. For example, if we consider the simple linear model in matrix notation, we noted earlier that if $\boldsymbol{Y}$ is the random vector consisting of the observations, then the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

which is such a linear combination. It may be shown using the above that

$$E(\boldsymbol{CY}) = \boldsymbol{C\mu} \quad \mathrm{var}(\boldsymbol{CY}) = \boldsymbol{C\Sigma C}'.$$

Finally, the results above may be generalized. If $\boldsymbol{A}$ is a $(q \times 1)$ vector, then

- $E(\boldsymbol{CY} + \boldsymbol{a}) = \boldsymbol{C\mu} + \boldsymbol{a}$.
- $\mathrm{var}(\boldsymbol{CY} + \boldsymbol{a}) = \boldsymbol{C\Sigma C}'$.
- We will make extensive use of this result.
- It is important to recognize that there is nothing mysterious about these results – they merely represent a streamlined way of summarizing information on operations performed on all elements of a random vector succinctly. For example, the first result on $E(\boldsymbol{CY} + \boldsymbol{a})$ just summarizes what the expected value of several different combinations of the elements of $\boldsymbol{Y}$ is, where each is shifted by a constant (the corresponding element in $\boldsymbol{a}$). Operationally, the results follow from applying the above definitions and matrix operations.

3.3 The multivariate normal distribution

A fundamental theme in much of statistical methodology is that the **normal probability distribution** is a reasonable model for the population of possible values taken on by many random variables of interest. In particular, the normal distribution is often (but not always) a good approximation to the true probability distribution for a random variable y when the random variable is **continuous**. Later in the course, we will discuss other probability distributions that are better approximations when the random variable of interest is **continuous** or **discrete**.

If we have a random vector $\boldsymbol{Y}$ with elements that are continuous random variables, then, it is natural to consider the normal distribution as a **probability model** for each element Y_j. However, as we have discussed, we are likely to be concerned about **associations** among the elements of $\boldsymbol{Y}$. Thus, it does not suffice to describe each of the elements Y_j separately; rather, we seek a probability model that describes their **joint** behavior. As we have noted, such probability distributions are called **multivariate** for obvious reasons.

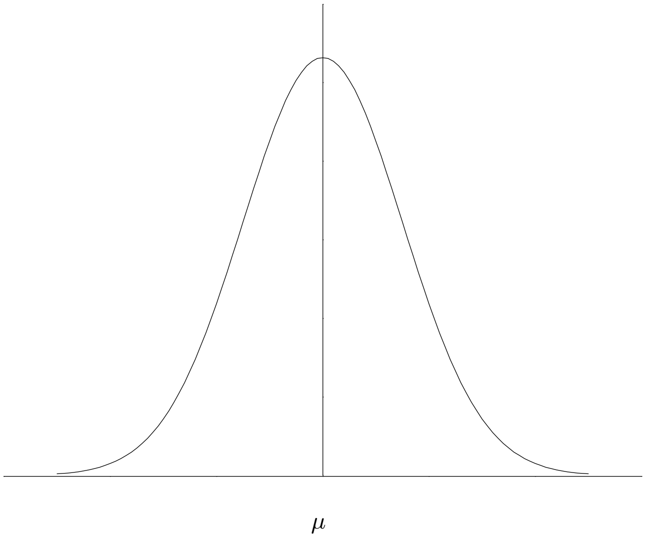
The **multivariate normal distribution** is the extension of the normal distribution of a single random variable to a random vector composed of elements that are each normally distributed. Through its form, it naturally takes into account correlation among the elements of $\boldsymbol{Y}$; moreover, it gives a basis for a way of thinking about an extension of "least squares" that is relevant when observations are not independent but rather are correlated.

NORMAL PROBABILITY DENSITY: Recall that, for a random variable y, the normal distribution has **probability density function**

$$f(y) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-(y-\mu)^2/(2\sigma^2)\right\}. \tag{3.8}$$

This function has the shape shown in Figure 3. The shape will vary in terms of "center" and "spread" according to the values of the population mean μ and variance σ^2 (e.g. recall Figure 1).

Figure 3: *Normal density function with mean μ.*



Several features are evident from the form of (3.8):

- The form of the function is determined by μ and σ^2. Thus, if we know the population mean and variance of a random variable Y, and we know it is normally distributed, we know everything about the probabilities associated with values of Y, because we then know the function (3.8) completely.
- The form of (3.8) depends critically on the term
$$-\frac{(y-\mu)^2}{\sigma^2} = (y-\mu)(\sigma^2)^{-1}(y-\mu). \tag{3.9}$$
Note that this term depends on the **squared deviation** $(y-\mu)^2$.
- The deviation is **standardized** by the standard deviation σ, which has the same units as y, so that it is put on a unitless basis.
- This standardized deviation has the interpretation of a **distance** measure – it measures how far y is from μ, and then puts the result on a unitless basis relative to the "spread" about μ expected.
- Thus, the normal distribution and methods such as **least squares**, which depends on minimizing a sum of squared deviations, have an intimate connection. We will use this connection to motivate the interpretation of the form of multivariate normal distribution informally now. Later in the course, we will be more formal about this connection.

SIMPLE LINEAR REGRESSION: For now, to appreciate this form and its extension, consider the method of least squares for fitting a simple linear regression. (The same considerations apply to multiple linear regression, which will be discussed later in this chapter.) As before, at each fixed value $x_1, \ldots, x_n$, there is a corresponding random variable Y_j, $j = 1, \ldots, n$, which is assumed to arise from

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)'$$

The further assumption is that Y_j are each normally distributed with means $\mu_j = \beta_0 + \beta_1 x_j$ and variance σ^2.

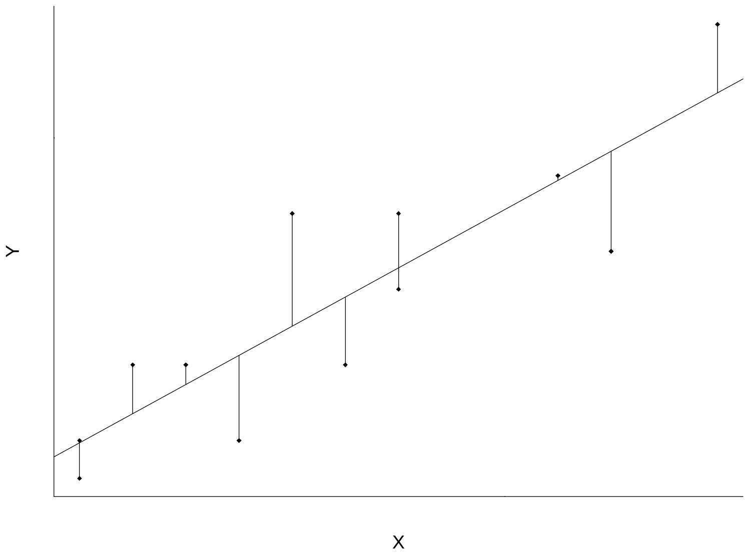
- Thus, each $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$, so that they have different means but the **same** variance.
- Furthermore, the Y_j are assumed to be **independent**.

The method of least squares is to minimize in $\boldsymbol{\beta}$ the sum of squared deviations $\sum_{j=1}^{n}(Y_j - \mu_j)^2$ which is the same as minimizing

$$\sum_{j=1}^{n}(Y_j - \mu_j)^2/\sigma^2 \tag{3.10}$$

as σ^2 is just a constant. Pictorially, realizations of such deviations are shown in Figure 4.

Figure 4: *Deviations from the mean in simple linear regression*



IMPORTANT POINTS:

- Each deviation gets "equal weight" in (3.10) – all are "weighted" by the same constant, σ^2.
- This makes sense – if each Y_j has the **same** variance, then each is subject to the same magnitude of variation, so the information on the population at x_j provided by Y_j is of "equal quality." Thus, information from all Y_j is treated as equally valuable in determining $\boldsymbol{\beta}$.
- The deviations corresponding to each observation are **summed**, so that each contributes to (3.10) in its own right, **unrelated** to the contributions of any others.
- (3.10) is like an overall distance measure of Y_j values from their means μ_j (put on a unitless basis relative to the "spread" expected for any Y_j).

MULTIVARIATE NORMAL PROBABILITY DENSITY: The joint probability distribution that is the extension of (3.8) to a $(n \times 1)$ random vector $\boldsymbol{Y}$, each of whose components are normally distributed (but possibly **associated**), is given by

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2 \right\} \tag{3.11}$$

- (3.11) describes the probabilities with which the random variable $\boldsymbol{Y}$ takes on values **jointly** in its n elements.

- The form of (3.11) is determined by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, as in the univariate case, if we know the mean vector and covariance matrix of a random vector $\boldsymbol{Y}$, and we know each of its elements are normally distributed, then we know everything about the joint probabilities associated with values $\boldsymbol{y}$ of $\boldsymbol{Y}$.

- By analogy to (3.9), the form of $f(\boldsymbol{y})$ depends critically on the term

$$(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}). \tag{3.12}$$

Note that this is a **quadratic form**, so it is a scalar function of the elements of $(\boldsymbol{y}-\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}^{-1}$. Specifically, if we refer to the elements of $\boldsymbol{\Sigma}^{-1}$ as σ^{jk}, i.e.

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma^{11} & \cdots & \sigma^{1n} \\ \vdots & \ddots & \vdots \\ \sigma^{n1} & \cdots & \sigma^{nn} \end{pmatrix},$$

then we may write

$$(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = \sum_{j=1}^{n}\sum_{k=1}^{n} \sigma^{jk}(y_j-\mu_j)(y_k-\mu_k). \tag{3.13}$$

Of course, the elements σ^{jk} will be complicated functions of the elements σ_j^2, σ_{jk} of $\boldsymbol{\Sigma}$, i.e. the variances of the Y_j and the covariances among them.

- This term thus depends on not only the **squared deviations** $(y_j-\mu_j)^2$ for each element in $\boldsymbol{y}$ (which arise in the double sum when $j = k$), but also on the **crossproducts** $(y_j-\mu_j)(y_k-\mu_k)$. Each contribution of these squares and crossproducts is being "standardized" somehow by values σ^{jk} that somehow involve the variances and covariances.

- Thus, although it is quite complicated, one gets the suspicion that (3.13) has an interpretation, albeit more complex, as a **distance measure**, just as in the univariate case.

BIVARIATE NORMAL DISTRIBUTION: To gain insight into this suspicion, and to get a better understanding of the multivariate distribution, it is instructive to consider the special case $n = 2$, the simplest example of a multivariate normal distribution (hence the name **bivariate**).

Here,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Using the inversion formula for a (2×2) matrix given in Chapter 2,

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

We also have that the **correlation** between Y_1 and Y_2 is given by

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2}.$$

Using these results, it is an algebraic exercise to show that (try it!)

$$(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \frac{1}{1-\rho_{12}^2}\left\{\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho_{12}\frac{(y_1-\mu_1)}{\sigma_1}\frac{(y_2-\mu_2)}{\sigma_2}\right\}. \tag{3.14}$$

Compare this expression to the general one (3.13).

Inspection of (3.14) shows that the quadratic form involves two components:

- The sum of standardized squared deviations

$$\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}.$$

  This sum alone is in the spirit of the sum of squared deviations in least squares, with the difference that each deviation is now **weighted** in accordance with its variance. This makes sense – because the variances of Y_1 and Y_2 differ, information on the population of Y_1 values is of "different quality" than that on the population of Y_2 values. If variance is "large," the quality of information is poorer; thus, the larger the variance, the smaller the "weight," so that information of "higher quality" receives more weight in the overall measure. Indeed, then, this is like a "distance measure," where each contribution receives an appropriate weight.

- In addition, there is an "extra" term that makes (3.14) have a different form than just a sum of weighted squared deviations:
$$-2\rho_{12}\frac{(y_1-\mu_1)}{\sigma_1}\frac{(y_2-\mu_2)}{\sigma_2}.$$
This term depends on the **crossproduct**, where each deviation is again weighted in accordance with its variance. This term modifies the "distance measure" in a way that is connected with the **association** between Y_1 and Y_2 through their crossproduct and their **correlation** ρ_{12}. Note that the larger this correlation in magnitude (either positive or negative), the more we modify the usual sum of squared deviations.

- Note that the entire quadratic form also involves the multiplicative factor $1/(1-\rho_{12}^2)$, which is greater than 1 if $|\rho_{12}|>0$. This factor scales the overall distance measure in accordance with the magnitude of the association.

INTERPRETATION: Based on the above observations, we have the following practical interpretation of (3.14):

- (3.14) is an overall measure of **distance** of the value $\boldsymbol{y}$ of $\boldsymbol{Y}$ from its mean $\boldsymbol{\mu}$.

- It contains the usual distance measure, a sum of appropriately weighted squared deviations.

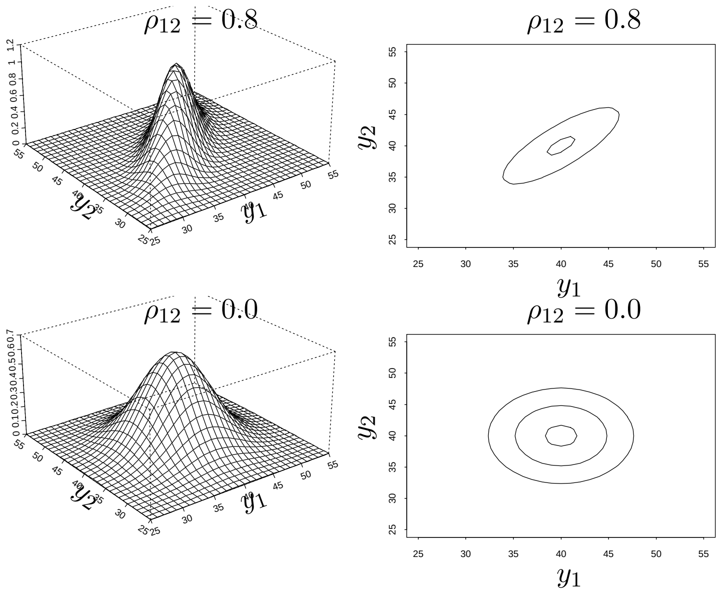
- However, if Y_1 and Y_2 are **positively correlated**, $\rho_{12}>0$, it is likely that the crossproduct $(Y_1-\mu_1)(Y_2-\mu_2)$ is positive. The measure of distance is thus reduced (we subtract off a positive quantity). This makes sense – if Y_1 and Y_2 are positively correlated, knowing one tells us a lot about the other. Thus, we won't have to "travel as far" to get from Y_1 to μ_1 and Y_2 to μ_2.

- Similarly, if Y_1 and Y_2 are **negatively correlated**, $\rho_{12}<0$, it is likely that the crossproduct $(Y_1-\mu_1)(Y_2-\mu_2)$ is negative. The measure of distance is again reduced (we subtract off a positive quantity). Again, if Y_1 and Y_2 are negatively correlated, knowing one still tells us a lot about the other (in the opposite direction).

- Note that if $\rho_{12}=0$, which says that there is **no association** between values taken on by Y_1 and Y_2, then the usual distance measure is not modified – there is "nothing to be gained" in traveling from Y_1 to μ_1 by knowing Y_2, and vice versa.

This interpretation may be more greatly appreciated by examining pictures of the bivariate normal density for different values of the correlation ρ_{12}. Note that the density is now an entire **surface** in 3 dimensions rather than just a curve in the plane, because account is taken of all possible **pairs** of values of Y_1 and Y_2. Figure 5 shows a the bivariate density function with $\mu_1 = 40$, $\mu_2 = 40$, $\sigma_1^2 = 5$, $\sigma_2^2 = 5$ for $\rho_{12} = 0.8$ and $\rho_{12} = 0.0$.

Figure 5: *Bivariate normal distributions with different correlations*



- The two panels in each row are the surface and a "bird's-eye" view for the 2 ρ_{12} values.

- For $\rho_{12} = 0.8$, a case of strong positive correlation, note that the picture is "tilted" at a 45 degree angle and is quite narrow. This reflects the implication of positive correlation – values of Y_1 and Y_2 are highly associated. Thus, the "overall distance" of a pair (Y_1, Y_2) from the "center" $\boldsymbol{\mu}$ is constrained by this association.

- For $\rho_{12} = 0$, Y_1 and Y_2 are not at all associated. Note now that the picture is not "tilted" – for a given value of Y_1, Y_2 can be "anything" within the relevant range of values for each. The "overall" distance of a pair (Y_1, Y_2) from the "center" $\boldsymbol{\mu}$ is not constrained by anything.

INDEPENDENCE: Note that if Y_1 and Y_2 are independent, then $\rho_{12} = 0$. In this case, the second term in the exponent of (3.14) disappears, and the entire quadratic form reduces to

$$\frac{(y_1 - \mu_1)^2}{\sigma_1^2} + \frac{(y_2 - \mu_2)^2}{\sigma_2^2}.$$

This is just the usual sum of weighted squared deviations.

EXTENSION: As you can imagine, these same concepts carry over to higher dimensions $n > 2$ in an analogous fashion; although the mechanics are more difficult, the ideas and implications are the same.

- In general, the quadratic form $(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ is a distance measure taking into account associations among the elements of $\boldsymbol{Y}$, $Y_1, \ldots, Y_n$, in the sense described above.

- When the Y_j are all mutually independent, the quadratic form will reduce to a weighted sum of squared deviations, as observed in particular for the bivariate case. It is actually possible to see this directly.

  If Y_j are independent, then all the correlations $\rho_{jk} = 0$, as are the covariances σ_{jk}, and it follows that $\boldsymbol{\Sigma}$ is a **diagonal** matrix. Thus, if

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

  then

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{pmatrix},$$

  so that (verify)

$$(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \sum_{j=1}^{n} (y_j - \mu_j)^2/\sigma_j^2.$$

  Note also that, as $\boldsymbol{\Sigma}$ is diagonal, we have

$$|\boldsymbol{\Sigma}| = \sigma_1^2 \sigma_2^2 \cdots \sigma_n^2.$$

  Thus, $f(\boldsymbol{y})$ becomes

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{1/2}\sigma_1} \exp\{-(y_1 - \mu_1)^2/(2\sigma_1^2)\} \cdots \frac{1}{(2\pi)^{1/2}\sigma_n} \exp\{-(y_n - \mu_n)^2/(2\sigma_n^2)\}; \qquad (3.15)$$

  $f(\boldsymbol{y})$ reduces to the product of individual normal densities. This is a defining characteristic of **statistical independence**; thus, we see that if $Y_1, \ldots, Y_n$ are each normally distributed and uncorrelated, they are independent. Of course, this independence assumption forms the basis for the usual method of least squares.

SIMPLE LINEAR REGRESSION, CONTINUED: We now apply the above concepts to extension of usual least squares. We have seen that estimation of $\boldsymbol{\beta}$ is based on minimizing an appropriate distance measure. For classical least squares under the assumptions of

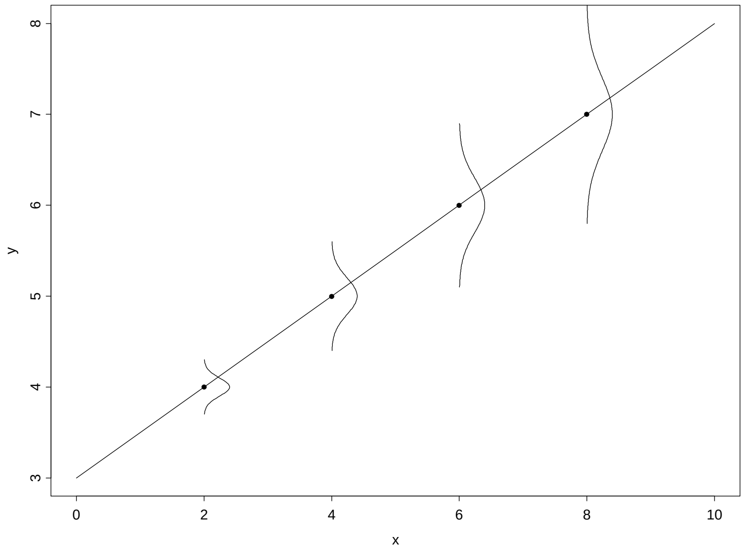
(i) constant variance

(ii) independence

the distance measure to be minimized is a sum of squared deviations, where each receives the same weight.

- Consider relaxation of (i); i.e. suppose we believe that $Y_1, \ldots, Y_n$ were each normally distributed and uncorrelated (which implies independent or totally unrelated), but that $\text{var}(Y_j)$ is not the same at each x_j. This situation is represented pictorially in Figure 6.

  Figure 6: *Simple linear regression with nonconstant variance*



  Under these conditions, we believe that the joint probability density of $\boldsymbol{Y}$ is given by (3.15), so we would want to obtain the estimator for $\boldsymbol{\beta}$ that minimizes the overall distance measure associated with this, the one that takes the fact that there are different variances, and hence different "quality" of information, at each x_j; i.e. the weighted sum of squared deviations

$$\sum_{j=1}^{n}(Y_j - \mu_j)^2/\sigma_j^2.$$

  Estimation of $\boldsymbol{\beta}$ in linear regression based on minimization of this distance measure is often called **weighted least squares** for obvious reasons.

(Note that, to actually carry this out in practice, we would need to know the values of each σ_j^2, which is unnecessary when all the σ_j^2 are the same. We will take up this issue later.)

- Consider relaxation both of (i) and (ii); we believe that $Y_1, \ldots, Y_n$ are each normally distributed but correlated with possibly different variances at each x_j. In this case, we believe that $\boldsymbol{y}$ follows a general multivariate normal distribution. Thus, we would want to base estimation of $\boldsymbol{\beta}$ on the overall distance measure associated with this probability density, which takes both these features into account; i.e. we would minimize the **quadratic form**

  $$(\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}).$$

  Estimation of $\boldsymbol{\beta}$ in linear regression based on such a general distance measure is also sometimes called **weighted least squares**, where it is understood that the "weighting" also involves information on correlations (through terms involving crossproducts).

  (Again, to carry this out in practice, we would need to know the entire matrix $\boldsymbol{\Sigma}$; more later.)

NOTATION: In general, we will use the following notation. If $\boldsymbol{Y}$ is a $(n \times 1)$ random vector with a multivariate normal distribution, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, we will write this as

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- The subscript n reminds us that the distribution is n-variate
- We may at times omit the subscript in places where the dimension is obvious.

PROPERTIES:

- If $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then if we have a linear combination of $\boldsymbol{Y}$, $\boldsymbol{CY}$, where $\boldsymbol{C}$ is $(q \times n)$, then $\boldsymbol{CY} \sim \mathcal{N}_n(\boldsymbol{C\mu}, \boldsymbol{C\Sigma C}')$.
- If also $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\tau}, \boldsymbol{\Gamma})$ and is independent of $\boldsymbol{Y}$, then $\boldsymbol{Z} + \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu} + \boldsymbol{\tau}, \boldsymbol{\Sigma} + \boldsymbol{\Gamma})$ (as long as $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ are nonsingular).
- We will use these two facts alone and together.

3.4 Multiple linear regression

So far, we have illustrated the usefulness of matrix notation and some key points in the context of the problem of simple linear regression, which we have referred to informally throughout our discussion. Now that we have discussed the multivariate normal distribution, it is worthwhile to review formally the usual multiple linear regression model, of which the simple linear regression model is a special case, and summarize what we have discussed from the broader perspective we have developed in terms of this model in one place. This will prove useful later, when we consider more complex models for longitudinal data.

SITUATION: The situation of the general multiple linear regression model is as follows.

- We have responses $Y_1, \ldots, Y_n$, the jth of which is to be taken at a setting of k **covariates** (also called predictors or independent variables) $x_{j1}, x_{j2}, \ldots, x_{jk}$.

- For example, an experiment may be conducted involving n men. Each man spends 30 minutes walking on a treadmill, and at the end of this period, $Y =$ his oxygen intake rate (ml/kg/min) is measured. Also recorded are $x_1 =$ age (years), $x_2 =$ weight (kg) $x_3 =$ heart rate while resting (beats/min), and $x_4 =$ oxygen rate while resting (ml/kg/min). Thus, for the jth man, we have response

$$Y_j = \text{ oxygen intake rate after 30 min}$$

  and his covariate values $x_{j1}, \ldots, x_{j4}$.

  The objective is to develop a **statistical model** that represents oxygen intake rate after 30 minutes on the treadmill as a function of the covariates. One possible use for the model may be to get a sense of how oxygen rates after 30 minutes might be for men with certain baseline characteristics (age, weight, resting physiology) in order to develop guidelines for an exercise program.

- A standard model under such conditions is to assume that each covariate affects the response in a linear fashion. Specifically, if there are k covariates ($k = 4$ above), then we assume

$$Y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} + \epsilon_j, \quad \mu_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk}. \tag{3.16}$$

  Here, ϵ_j is a random deviation with mean 0 and variance σ^2 that characterizes how the observations on Y_j deviate from the mean value μ_j due to the **aggregate effects** of relevant **sources of variation**.

- More formally, under this model, we believe that there is a population of all possible Y_j values that could be seen for, in the case of our example, men with the particular covariate values $x_{j1}, \ldots, x_{jk}$. This population is thought to have mean μ_j given above. ϵ_j reflects how such an observation might deviate from this mean.

- The model itself has a particular interpretation. It says that if the value of one of the covariates, x_k, say, is increased by one unit, then the value of the mean increases by the amount β_k.

- The usual assumption is that at any setting of the covariates, the population of possible Y_j values is well-represented by a **normal distribution** with mean μ_j and variance σ^2. Note that the variance σ^2 is the **same** regardless of the covariate setting. More formally, we may state this as

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2) \text{ or equivalently } Y_j \sim \mathcal{N}(\mu_j, \sigma^2).$$

- Furthermore, it is usually assumed that the Y_j are **independent**. This would certainly make sense in our example – we would expect that if the men were completely unrelated (chosen **at random** from the population of all men of interest), then there should be no reason to expect that the response observed for any one man would have anything to do with that observed for another.

- The model is usually represented in matrix terms: letting the row vector $\boldsymbol{x}_j' = (1, x_{j1}, \ldots, x_{jk})$, the model is written

$$Y_j = \boldsymbol{x}_j'\boldsymbol{\beta} + \epsilon_j, \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad (p \times 1),$$

where $p = k + 1$ is the dimension of $\boldsymbol{\beta}$, so that the $(n \times p)$ **design matrix** $\boldsymbol{X}$ has rows $\boldsymbol{x}_j'$.

- Thus, thinking of the data as the **random vector $\boldsymbol{Y}$**, we may summarize the assumptions of **normality**, **independence**, and **constant variance** succinctly. We may think of $\boldsymbol{Y}$ $(n \times 1)$ as having a multivariate normal distribution with mean $\boldsymbol{X\beta}$. Because the elements of $\boldsymbol{Y}$ are assumed independent, all covariances among the Y_j are 0, and the covariance matrix of $\boldsymbol{Y}$ is **diagonal**. Moreover, with constant variance σ^2, the variance is the same for each Y_j. Thus, the covariance matrix is given by

$$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \boldsymbol{I},$$

  where $\boldsymbol{I}$ is a $(n \times n)$ identity matrix.

  We thus may write

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}).$$

- Note that the simple linear regression model is a special case of this with $k = 1$. The only real difference is in the complexity of the assumed model for the mean of the population of Y_j values; for the general multiple linear regression model, this depends on k covariates. The simple linear regression case is instructive because we are able to depict things graphically with ease; for example, we may plot the relationship in a simple x-y plane. For the general model, this is not possible, but in principle the issues are the same.

LEAST SQUARES ESTIMATION: The goal of an analysis of data of this form under assumption of the multiple linear regression model (3.16) is to estimate the **regression parameter $\boldsymbol{\beta}$** using the data in order to characterize the relationship.

Under the usual assumptions discussed above, i.e.

- Y_j (and equivalently ϵ_j) are normally distributed with variance σ^2 for all j
- Y_j (and equivalently ϵ_j) are independent

the usual estimator for $\boldsymbol{\beta}$ is found by minimizing the sum of squared deviations

$$\sum_{j=1}^{n} (Y_j - \beta_0 - x_{j1}\beta_1 - \cdots - x_{jp}\beta_k)^2.$$

In matrix terms, the sum of squared deviations may be written

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}). \tag{3.17}$$

In these terms, the sum of squared deviations may be seen to be just a quadratic form.

- Note that we may write these equivalently as

  $$\sum_{j=1}^{n}(Y_j-\beta_0-x_{j1}\beta_1-\cdots-x_{jk}\beta_k)^2/\sigma^2,$$

  and

  $$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{I}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})/\sigma^2;$$

  because σ^2 does not involve $\boldsymbol{\beta}$, we may equally well talk about minimizing these quantities. Of course, as we have previously discussed, this shows that all observations are getting "equal weight" in determining $\boldsymbol{\beta}$, which is sensible if we believe that the populations of all values of Y_j at any covariate setting are equally variable (same σ^2). We now see that we are minimizing the distance measure associated with a multivariate normal distribution where all of the Y_j are mutually independent with the same variance (all covariances/correlations = 0).

- Minimizing (3.17) means that we are trying to find the value of $\boldsymbol{\beta}$ that minimizes the **distance** between responses and the means; by doing so, we are attributing as much of the overall differences among the Y_j that we have seen to the fact that they arise from different settings of $\boldsymbol{x}_j$, and as little as possible to random variation.

- Because the quadratic form (3.17) is just a scalar function of the p elements of $\boldsymbol{\beta}$, it is possible to use calculus to determine that values of these p elements that minimize the quadratic form. Formally, one would take the derivatives of (3.17) with respect to each of $\beta_0, \beta_1, \ldots, \beta_k$ and set these p expressions equal to zero. These p expressions represent a system of equations that may be solved to obtain the solution, the **estimator** $\widehat{\boldsymbol{\beta}}$.

- The set of p simultaneous equations that arise from taking derivatives of (3.17), expressed in matrix notation, is
$$-2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\beta = \mathbf{0} \text{ or } \boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$
We wish to solve for $\boldsymbol{\beta}$. Note that $\boldsymbol{X}'\boldsymbol{X}$ is a **square** matrix $(p \times p)$ and $\boldsymbol{X}'\boldsymbol{y}$ is a $(p \times 1)$ vector. Recall in Chapter 2 we saw how to solve a set of simultaneous equations like this; thus, we may invoke that procedure to solve
$$\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$
as long as the **inverse** of $\boldsymbol{X}'\boldsymbol{X}$ **exists.**

- Assuming this is the case, from Chapter 2, we know that $\boldsymbol{X}'\boldsymbol{X}$ will be **of full rank** (rank = number of rows and columns = p) if $\boldsymbol{X}$ has rank p. We also know from Chapter 2 that if a square matrix is of full rank, it is **nonsingular**, so its **inverse exists**. Thus, assuming $\boldsymbol{X}$ is of full rank, we have that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, and we may premultiply both sides by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ to obtain
$$\begin{aligned}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

- Thus, the **least squares estimator** for $\boldsymbol{\beta}$ is given by
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.18}$$

- Computation for general p is not feasible by hand, of course; particularly nasty is the inversion of the matrix $\boldsymbol{X}'\boldsymbol{X}$. Software for multiple regression analysis includes routines for inverting a matrix of any dimension; thus, estimation of $\boldsymbol{\beta}$ by least squares for a general multiple linear regression model is best carried out in this fashion.

ESTIMATION OF σ^2: It is often of interest to estimate σ^2, the assumed common variance. The usual estimator is

$$\widehat{\sigma}^2 = (n-p)^{-1}\sum_{j=1}^{n}(Y_j - \boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})^2 = (n-p)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

- This makes intuitive sense. Each squared deviation $(Y_j - \boldsymbol{x}_j'\boldsymbol{\beta})^2$ contains information about the "spread" of values of Y_j at $\boldsymbol{x}_j$. As we assume that this spread is the same for all $\boldsymbol{x}_j$, a natural approach to estimating its magnitude, represented by the variance σ^2, would be to **pool** this information across all n deviations. Because we don't know $\boldsymbol{\beta}$, we replace it by the estimator $\widehat{\boldsymbol{\beta}}$.

- We will see a more formal rationale later.

SAMPLING DISTRIBUTION: When we estimate a **parameter** (like $\boldsymbol{\beta}$ or σ^2) that describes a population by an **estimator** (like $\widehat{\boldsymbol{\beta}}$ or $\widehat{\sigma}^2$), the estimator is some function of the responses, $\boldsymbol{Y}$ here. Thus, the quality of the estimator, i.e. how reliable it is, depends on the variation inherent in the responses and how much data on the responses we have.

- If we consider every possible set of data we might have ended up with of size n, each one of these would give rise to a value of the estimator. We may think then of the **population** of all possible values of the estimator we might have ended up with.

- We would hope that the **mean** of this population would be equal to the **true value** of the parameter we are trying to estimate. This property is called **unbiasedness**.

- We would also hope that the **variability** in this population isn't too large.

- If the values vary **a lot** across all possible data sets, then the estimator is not very reliable. Indeed, we ended up with a particular data set, which yielded a particular estimate; however, had we ended up with another data set, we might have ended up with quite a different estimate.

- If on the other hand these values vary **little** across all possible data sets, then the estimator is reliable. Had we ended up with another set of data, we would have ended up with an estimate that is quite similar to the one we have.

Thus, it is of interest to characterize the population of all possible values of an estimator. Because the estimator depends on the response, the properties of this population will depend on those of $\boldsymbol{Y}$. More formally, we may think of the **probability distribution** of the estimator, describing how it takes on all its possible values. This probability distribution will be connected with that of the $\boldsymbol{Y}$.

A probability distribution that characterizes the population of all possible values of an estimator is called a **sampling distribution**.

To understand the nature of the sampling distribution of $\widehat{\boldsymbol{\beta}}$, we thus consider the probability distribution of

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{3.19}$$

which is a **linear combination** of the elements of $\boldsymbol{Y}$. We may thus apply earlier facts to derive mathematically the sampling distribution.

- We may determine the mean of this distribution by applying the expectation operator to the expression (3.19); this represents averaging across all possible values of the expression (which follow from all possible values of $\boldsymbol{Y}$). Now $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ under the usual assumptions, thus $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$. Thus, using the results in section 3.2,

  $$E(\widehat{\boldsymbol{\beta}}) = E\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

  showing that $\widehat{\boldsymbol{\beta}}$ under our assumptions is an **unbiased** estimator of $\boldsymbol{\beta}$.

- We may also determine the variance of this distribution. Formally, this would mean applying the expectation operator to

  $$\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\beta}\}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\beta}\}';$$

  i.e. finding the covariance matrix of (3.19). Rather than doing this directly, it is simpler to exploit the results in section 3.2, which yield

  $$\begin{aligned}\text{var}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{var}(\boldsymbol{Y})\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

  Note that the variability of the population of all possible values of $\widehat{\boldsymbol{\beta}}$ depends directly on σ^2, the variation in the response. It also depends on n, the sample size, because $\boldsymbol{X}$ is of dimension $(n \times p)$.

- In fact, we may say more – because under our assumptions $\boldsymbol{Y}$ has a multivariate normal distribution, it follows that the probability distribution of all possible values of $\widehat{\boldsymbol{\beta}}$ is multivariate normal with this mean and covariance matrix; i.e.

  $$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\}.$$

This result is used to obtain estimated **standard errors** for the components of $\widehat{\boldsymbol{\beta}}$; i.e. estimates of the standard deviation of the sampling distributions of each component of $\widehat{\boldsymbol{\beta}}$.

- In practice, σ^2 is unknown, thus, it is replaced with the estimate $\widehat{\sigma}^2$.
- The estimated standard error of the kth element of $\widehat{\boldsymbol{\beta}}$ is then the square root of the kth diagonal element of $\widehat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

It is also possible to derive a sampling distribution for $\widehat{\sigma}^2$. For now, we will note that it is possible to show that $\widehat{\sigma}^2$ is an **unbiased** estimator of σ^2. That is, it may be shown that

$$E\{(n-p)^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\} = \sigma^2.$$

This may be shown by the following steps:

- First, it may be demonstrated that (try it!)

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y}'\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{Y}\end{aligned}$$

  We have just expressed the original quadratic form in a different way, which is still a quadratic form.

- Fact: It may be shown that if $\boldsymbol{Y}$ is any random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ that for any square matrix $\boldsymbol{A}$,

$$E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$

  Applying this to our problem, we have $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, and $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$. Thus, using results in Chapter 2,

$$\begin{aligned}E(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \mathrm{tr}[\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\sigma^2\boldsymbol{I}] + \boldsymbol{\beta}'\boldsymbol{X}'\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{X}\boldsymbol{\beta} \\ &= \sigma^2\mathrm{tr}\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\} + \boldsymbol{\beta}'\boldsymbol{X}'\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

  Thus, to find $E(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$, we must evaluate each term.

- We also have: If $\boldsymbol{X}$ is any $(n \times p)$ matrix of full rank, writing $\boldsymbol{I}_q$ to emphasize the dimension of the identity matrix of dimension q, then

$$\text{tr}\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\} = \text{tr}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\} = \text{tr}(\boldsymbol{I}_p) = p,$$

  so that

$$\text{tr}\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\} = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{I}_p) = n - p.$$

  Furthermore,

$$\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}.$$

  Applying these to the above expression, we obtain

$$E(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \sigma^2(n-p) + 0 = \sigma^2(n-p).$$

  Thus, we have $E\{(n-p)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\} = \sigma^2$, as desired.

EXTENSION: The discussion above focused on the usual multiple linear regression model, where it is assumed that

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}).$$

In some situations, although it may be reasonable to think that the population of possible values of Y_j at $\boldsymbol{x}_j$ might be normally distributed, the assumptions of constant variance and independence may not be realistic.

- For example, recall the treadmill example, where Y_j was oxygen intake rate after 20 minutes on the treadmill for man j with covariates (age, weight, baseline characteristics) $\boldsymbol{x}_j$. Now each Y_j was measured on a different man, so the assumption of independence among the Y_j seems realistic.
- However, the assumption of constant variance may be suspect. Young men in their 20s will all tend to be relatively fit, simply by virtue of their age, so we might expect their rates of oxygen intake to not vary too much. Older men in their 50s and beyond, on the other hand, might be quite variable in their fitness – some may have exercised regularly, while others may be quite sedentary. Thus, we might expect oxygen intake rates for older men to be more variable than for younger men. More formally, we might expect the distributions of possible values of Y_j at different settings of $\boldsymbol{x}_j$ to exhibit different **variances** as the ages of men differ.

- Recall the pine seedling example. Suppose the seedling is planted and its height is measured on each of n consecutive days. Here, Y_j would be the height measured at time x_j, say, where x_j is the time measured in days from planting. We might model the mean of Y_j as a function of x_j, e.g.

  $$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j,$$

  a quadratic function of time. After n days, we have the vector $\boldsymbol{Y}$. As discussed earlier, however, it may not be realistic to think that the elements of $\boldsymbol{Y}$ are all mutually independent. In fact, we do not expect the height to follow the "smooth" quadratic trend; rather, it "fluctuates" about it; e.g. the seedling may undergo "growth spurts" or "dormant periods" along the way. Thus, we would expect to see a "large" value of Y on one day followed by a "large" value the next day. Thus, the elements of Y_j **covary** (are **correlated**).

In these situations, we still wish to consider a multiple linear regression model; however, the standard assumptions do not apply. More formally, we may still believe that each Y_j follows a normal distribution, so that $\boldsymbol{Y}$ is multivariate normal, but the assumption that

$$\text{var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$$

for some constant σ^2 is no longer relevant. Rather, we think that

$$\text{var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$$

for some covariance matrix $\boldsymbol{\Sigma}$ that summarizes the variances of each Y_j and the covariances thought to exist among them. Under these conditions, we would rather assume

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \boldsymbol{\Sigma}).$$

Clearly, the usual method of least squares, discussed above, is inappropriate for estimating $\boldsymbol{\beta}$; it minimizes an inappropriate distance criterion.

WEIGHTED LEAST SQUARES: The appropriate distance condition is

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}). \tag{3.20}$$

Ideally, we would rather estimate $\boldsymbol{\beta}$ by minimizing (3.20), because it takes appropriate account of the possibly different variances and the covariances among elements of $\boldsymbol{Y}$.

- In the constant variance/independence situation, recall that σ^2, the assumed common variance, is not involved in estimation of $\boldsymbol{\beta}$.

- In addition, if σ^2 is unknown, as is usually the case in practice, we saw that an intuitively appealing, unbiased estimator $\widehat{\sigma}^2$ may be derived, which is based on "pooling" information on the common σ^2.
- Here, however, with possibly different variances for different Y_j, and different covariances among different pairs (Y_j, Y_k), things seem much more difficult! As we will see momentarily, estimation of $\boldsymbol{\beta}$ by minimizing (3.20) will now involve $\boldsymbol{\Sigma}$, which further complicates matters.
- We will delay discussion of the issue of how to **estimate $\boldsymbol{\Sigma}$** in the event that it is unknown until we talk about longitudinal data from several individuals later.

For now, assume that $\boldsymbol{\Sigma}$ is **known**, which is clearly unrealistic in practice, to gain insight into the principle of minimizing (3.20).

- Analogous to the simpler case of constant variance/independence, to determine the value $\widehat{\boldsymbol{\beta}}$ that minimizes (3.20), one may use calculus to derive a set of p simultaneous equations to solve, which turn out to be

$$-2\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0},$$

  which leads to the solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}. \tag{3.21}$$

  $\widehat{\boldsymbol{\beta}}$ in (3.21) is often called the **weighted least squares** estimator.

- Note that $\widehat{\boldsymbol{\beta}}$ is still a **linear function** of the elements of $\boldsymbol{Y}$.
- Thus, it is straightforward to derive its sampling distribution. $\widehat{\boldsymbol{\beta}}$ is unbiased, as

$$E(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

$$\text{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}.$$

- Furthermore, because $\boldsymbol{Y}$ is multivariate normal, we have

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\{\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\}.$$

- Thus, if we knew $\boldsymbol{\Sigma}$, we would be able to construct estimated standard errors for elements of $\widehat{\boldsymbol{\beta}}$, etc.

The notion of weighted least squares will play a major role in our subsequent development of methods for longitudinal data. We will revisit it and tackle the issue of how to estimate $\boldsymbol{\Sigma}$ later.