4 Lecture 6: Feb 7

Last time

- Probability review

Today

- R basics
- Probability review
- Basic statistical concepts (PVR Chapter 1 - 2)
- Simple Linear Regression (JF Chapter 5)

Some common random variables

Discrete random variables

- $X \sim Bernoulli(p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

- $X \sim Binomial(n, p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim Geometric(p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda)$ (where $\lambda > 0$):

$$\Pr(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous random variables

- $X \sim Uniform(a, b)$ (where $a < b$):

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leqslant x \leqslant b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Exponential(\lambda)$ (where $\lambda > 0$):

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Normal(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The following table provides a summary of some of the properties of these distributions.

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$, for $0 \leqslant k \leqslant n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{x-1}$, for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\frac{\lambda^x}{x!}$, for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a} I(a \leqslant x \leqslant b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} I(x \geqslant 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Statistics: its objectives and scope (PVR Chapter 1)

We will use the word *statistics* in a broader sense:
Statistics refers to a body of scientific principles and methodologies that are useful for obtaining information about a phenomenon or a large collection of items. Statistical methods are techniques for using limited amounts of information to arrive at conclusions – called statistical inferences – about the phenomenon or the collection of items of interest.

Population and sample population

A *population* (sometimes referred to as a statistical population) is a collection (or aggregate) of measurements about which an inference is desired.

Example: An investigator is interested in evaluating the relationship between age, blood sugar level, and blood cholesterol level of insulin-dependent diabetics who are on a special experimental diet. The investigator wants to answer the following questions, among others:

1. How does the blood cholesterol level change with age and blood sugar level?
2. Are higher cholesterol levels associated with higher sugar levels?
3. Do older diabetics tend to have higher sugar and cholesterol levels?

What is the population of interest in this example?

Not that, in statistics, a measurement is one of the elements that form the population. In certain populations, each measurement may consist of several values. Populations in which each measurement

- is a single value are called *univariate* populations
- contains more than one value is called a *multivariate* population.

Sample and sample size

A *sample* consists of a finite number of measurements chosen from a population. The number of measurements in a sample is called the *sample size*.

Example: Answers to the questions about associations between the age, blood sugar level, and blood cholesterol level of diabetics can be based on measurements made on a sample of, say, $n = 40$ treated insulin-dependent diabetics. Such a sample is a collection of 40 measurements, each of which consists of three values: the age, blood sugar level, and blood cholesterol level of a treated patient.

Statistical components of a research study

A typical research study consists of three stages. The statistical techniques useful in these three stages are commonly known as *statistical methods in research*, and can be divided into three groups:

1. Methods for designing the research study

2. Methods for organizing and summarizing data
3. Methods for making inferences

In this course, we focus on the third stage that is to use the information in the samples to make conclusions about populations (i.e. making inferences). The key statistical issue in such inferences is their accuracy.

Example: Suppose that the average indoor radiation level in a sample of 15 homes built on reclaimed phosphate mine lands is 0.032 WL (working level is a historical unit of concentration of radioactive decay products of radon). Then 0.032 WL could be regarded as an estimate of the average indoor level in all homes built on reclaimed phosphate mine lands.

- How accurate is this estimate?
- Suppose there are a total of 4000 homes built on reclaimed lands, is our small sample representative of the whole population?
- If our sample could be regarded as representative of the population, it would be reasonable to expect that the difference between the estimated value of 0.032 WL and the true mean radiation level for all homes will be small, but how do we get/estimate the actual magnitude of this difference?

The natural question is whether it is possible to assess, with reasonable certainty, the magnitude of the error in our estimate. For example, can we say, with a reasonable degree of confidence, that the average level for the population of all homes will be within 0.001 WL of the average value calculated from sample homes?

Types of populations (PVR Chapter 2.2)

Statistical populations can be classified into categories depending upon the characteristics of the measurements contained in them.

- Univariate and multivariate populations
  - In a *univariate* population, each measurement consists of a single value
  - In a *multivariate* population, measurements consist of more than one value
- Real and conceptual populations
  - The population of 4000 indoor radon levels is a real population
  - The population of digestibility values for sheep fed June-harvested Pensacola Bahia grass is a conceptual population.
- Finite and infinite populations
  - A population may contain only a finite number of measurements, as in the case of the population of indoor radon levels of 4000 homes

  - A population may have infinitely many measurements, as in the case of a conceptual population of potential digestibility measurements, in which every value in the interval $[0\%, 100\%]$ is a possible value of a measurement in the population.

- Quantitative and qualitative populations
  - A measurement is said to be *quantitative* if its value can be interpreted on a natural and meaningful scale
  - A measurement is *qualitative* if its value serves the sole purpose of identifying an object or a characteristic. The value of a qualitative measurement has no numerical implications.

- Discrete and continuous populations
  - A population is said to be *discrete* if the distinct values of the measurements contained in it can be arranged in a sequence.
  - A continuous population consists of measurements that take all the values in one or more intervals of a real line.

Simple linear regression

Figure 4.1 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.

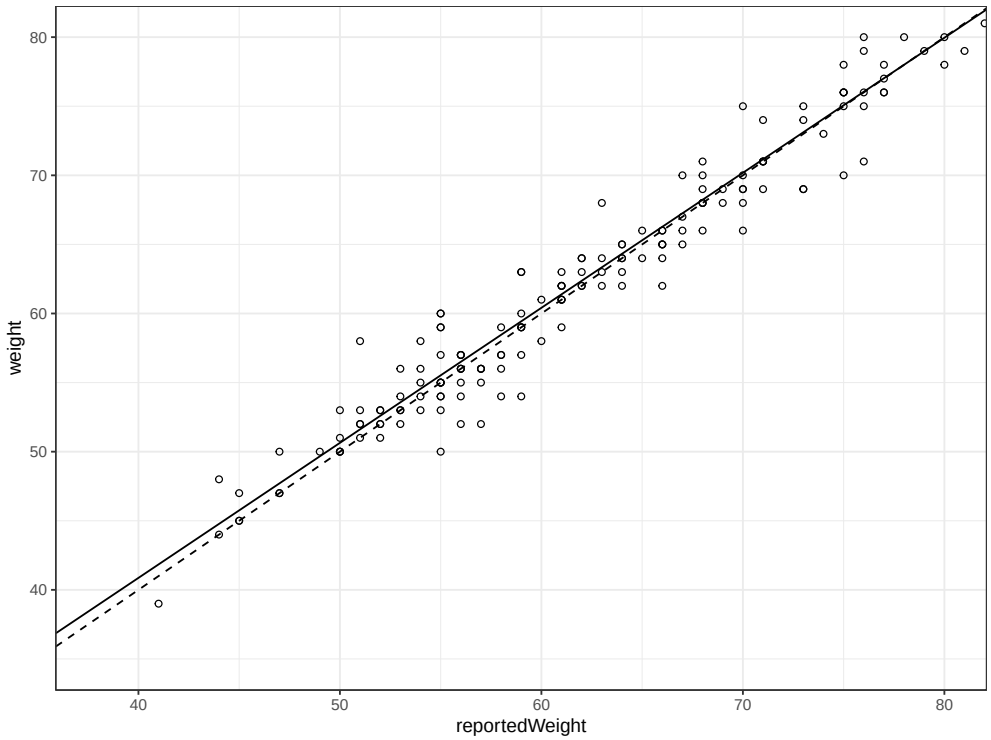

Figure 4.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$.

It's reasonable to assume that the relationship between measured and reported weight appears to be linear. Denote:

- measured weight by y_i: **response variable** or **dependent variable**
- reported weight by x_i: **predictor variable** or **independent variable**
- intercept: β_0
- slope: β_1
- residual/error term ϵ_i.

Then the simple linear regression model writes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

For given $(\hat{\beta}_0, \hat{\beta}_1)$ values, the *fitted value* or *predicted value* for observation i is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Therefore, the residual is

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

Fitting a linear model

Choose the "best" values for β_0, β_1 such that

$$SS[E] = \sum_1^n \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)^2 = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n \hat{\epsilon}_i^2$$

is minimized. These are **least squares** (LS) estimates:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \tfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

Definition: The line satisfying the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the linear regression of y on x which is also called the least squares line.

For Davis's data, we have

$$\begin{aligned} n &= 101 \\ \bar{y} &= \frac{5780}{101} = 57.228 \\ \bar{x} &= \frac{5731}{101} = 56.743 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 4435.9 \\ \sum (x_i - \bar{x})^2 &= 4539.3, \end{aligned}$$

so that

$$\begin{aligned} \hat{\beta}_1 &= \frac{4435.9}{4539.3} = 0.97722 \\ \hat{\beta}_0 &= 57.228 - 0.97722 \times 56.743 = 1.7776 \end{aligned}$$

Least squares estimates

The simple linear regression (SLR) model writes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The least squares estimates minimizes the sum of squared error (SSE) which is

$$SS[E] = \sum_1^n \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)^2 = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n \hat{\epsilon}_i^2.$$

The **least squares** (LS) estimates (in vector form):

$$\hat{\beta}_{ls} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \end{pmatrix}.$$

Definition: The line satisfying the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the linear regression of y on x which is also called the least squares line.

SLR Model in Matrix Form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Jargons

- $\mathbf{X}$ is called the *design matrix*
- $\boldsymbol{\beta}$ is the vector of parameters
- $\boldsymbol{\epsilon}$ is the error vector
- $\mathbf{Y}$ is the response vector.

The Design Matrix

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Vector of Parameters

$$\boldsymbol{\beta}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vector of Error terms

$$\boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vector of Responses

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Gramian Matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

Therefore, we have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Assume the Gramian matrix has full rank (which actually should be the case, why?), we want to show that

$$\hat{\boldsymbol{\beta}}_{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The inverse of the Gramian matrix is

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_i (x_i-\bar{x})^2}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

Now we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ls} =& (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=& \frac{1}{n\sum_i (x_i-\bar{x})^2}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}\begin{bmatrix} \mathbf{1}_n^T \\ \mathbf{x}^T \end{bmatrix}\mathbf{y} \\
=& \frac{1}{n\sum_i (x_i-\bar{x})^2}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}\begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} \\
=& \frac{1}{n\sum_i (x_i-\bar{x})^2}\begin{bmatrix} (\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i) \\ n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i) \end{bmatrix} \\
=& \begin{bmatrix} \bar{y} - \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\ \bar{x} \\ \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2} \end{bmatrix}
\end{aligned}$$

Some properties:

- (a) $\sum x_i \hat{\epsilon}_i = 0$.
- (b) $\sum \hat{y}_i \hat{\epsilon}_i = 0$ (HW1).

Proof: For (a), we look at

$$\begin{aligned}
&\mathbf{X}^T \hat{\boldsymbol{\epsilon}} \\
=&\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
=&\mathbf{X}^T[\mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\
=&\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=&\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Y} \\
=&\mathbf{0}
\end{aligned}$$

Other quantities in Matrix Form

Fitted values

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Hat matrix

$$\begin{aligned}
\hat{\mathbf{Y}} =&\mathbf{X}\hat{\boldsymbol{\beta}} \\
\hat{\mathbf{Y}} =&\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
\hat{\mathbf{Y}} =&\mathbf{H}\mathbf{Y}
\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called "hat matrix" because it turns $\mathbf{Y}$ into $\hat{\mathbf{Y}}$.

Davis's data example

For Davis's data, we have

$$\begin{aligned}
n &= 101 \\
\bar{y} &= \frac{5780}{101} = 57.228 \\
\bar{x} &= \frac{5731}{101} = 56.743 \\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= 4435.9 \\
\sum (x_i - \bar{x})^2 &= 4539.3,
\end{aligned}$$

so that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{4435.9}{4539.3} = 0.97722 \\
\hat{\beta}_0 &= 57.228 - 0.97722 \times 56.743 = 1.7776
\end{aligned}$$

Figure 4.2 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.

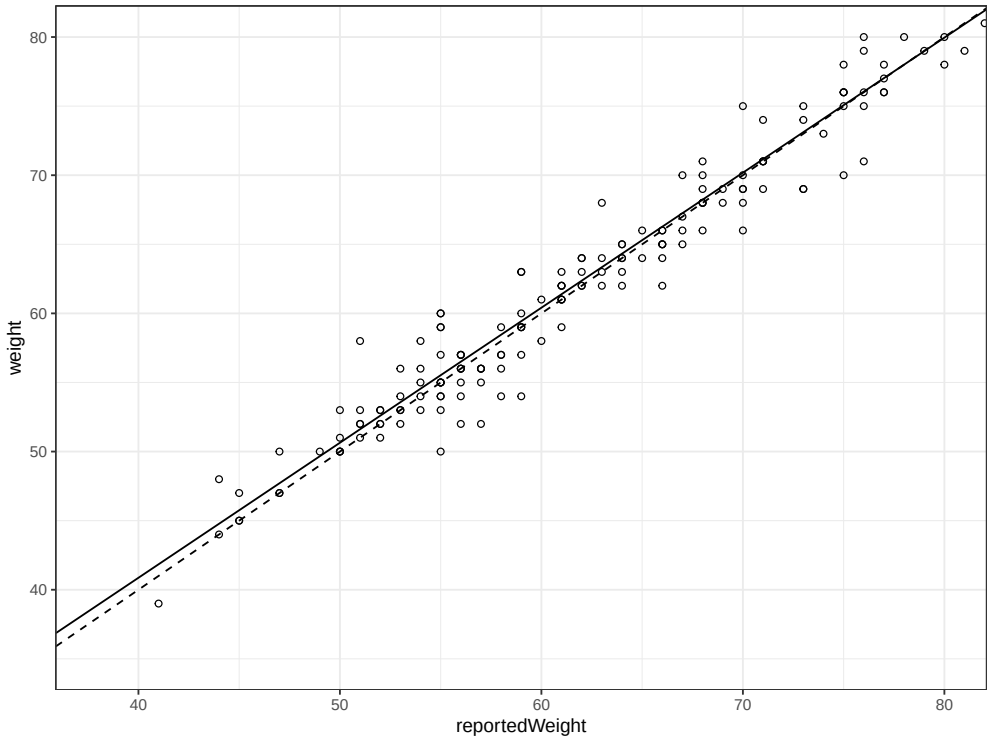

Figure 4.2: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$. The solid line gives the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.