ST3131 Simple Regression: Description

Semester 2 2023/2024

If printing, do DOUBLE-SIDED, each side TWO slides.

Introduction

- Let x and y be explanatory and response variables on n subjects. Subject i has values x_i and y_i, for $i = 1, \ldots, n$. The means and variances are

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \mathrm{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

  The standard deviations (SD) are $s_x = \sqrt{\mathrm{var}(x)}$ and $s_y = \sqrt{\mathrm{var}(y)}$.

- The goal is to summarise the data, and to predict y from x.

Topics

(I) Summarising two variables

(II) Regression line of y on x

(III) Predicting y from x

(IV) Analysis of variance

(V) Geometry of regression

(I) Karl Pearson's dataset

Pearson measured the heights of $n = 1,078$ father-son pairs in inches. His analysis revealed the phenomenon of regression.

We will use R to explore the data set and to motivate several concepts, which apply to any bivariate data set.

- Read `Pearsonheights.csv` into R. Make a scatter diagram.
- Let x and y be the height of fathers and sons respectively. Calculate the mean and SD to 2 decimal places.

$\overline{x} =$ $\quad s_x =$ $\quad \overline{y} =$ $\quad s_y =$

(I) Summarising scatter diagram

The centre and spread of a variable are indicated by the mean and SD.
Rule of thumb: (i) About 68% of data are within 1 SD of the mean. (ii) About 95% of data are within 2SD's of the mean.

- Add the point $(\overline{x}, \overline{y})$ to the scatter diagram. How to interpret the rule of thumb for x and y in terms of the diagram? How good is the rule?

- Calculate the p_x, the proportion of points with x values above $\overline{x}$. Similarly, calculate p_y.

- True or false? The proportion of points (x, y) with $x > \overline{x}$ and $y > \overline{y}$ is roughly $p_x p_y$.

(I) Correlation

- x and y are positively correlated: taller fathers tend to have taller sons.
- The correlation between x and y is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where the covariance between x and y is

$$s_{xy} = \mathrm{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

We assume that not all $x_1, \ldots, x_n$ are equal, i.e., $s_x > 0$, and likewise $s_y > 0$. Otherwise, r_{xy} is not defined.

(I) General formulae

Define

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

Let a, b, c be constants.

- $\text{var}(x) = \overline{x^2} - \overline{x}^2$, $\text{var}(y) = \overline{y^2} - \overline{y}^2$.
- $\text{cov}(x, y) = \overline{xy} - \overline{x}\,\overline{y}$. If $\text{cov}(x, y) = 0$, x and y are uncorrelated.
- $\text{cov}(ax + b, y) = a\,\text{cov}(x, y)$.
- $\text{var}(ax + by + c) = a^2\,\text{var}(x) + b^2\,\text{var}(y) + 2ab\,\text{cov}(x, y)$

Our abbreviations: $\text{var}(x) = s_x^2$, $\text{var}(y) = s_y^2$, $\text{cov}(x, y) = s_{xy}$.

(I) Change of scale

x: data variable. $a \neq 0$, b constants.

- Define $w = ax + b$, i.e., for each i, $w_i = ax_i + b$.

  This is called a change of scale. An example is the conversion from the Fahrenheit temperature scale to Celsius.

  $\overline{w} = a\overline{x} + b$. $\text{var}(w) = a^2 \, \text{var}(x)$.

- Define the standardised version of x:

$$x^* = \frac{x - \overline{x}}{s_x}$$

  What are the mean and SD of x^*?

(I) General properties of correlation

- The correlation is symmetric: $r_{yx} = r_{xy}$. Hence we just use r.
- Always, $-1 \leq r \leq 1$.
- If $r = \pm 1$, all points lie on a straight line.
- Let $w = ax + b$, where $a > 0$, b are constants. Then $r_{wy} = r$.
- r only measures linear association.

Concluding (I): A bivariate data set with variables x and y can be summarised by five numbers: the means $\overline{x}$ and $\overline{y}$, the SD's s_x and s_y, and the correlation r.

Topics

(I) Summarising two variables

(II) Regression line of y on x

(III) Predicting y from x

(IV) Analysis of variance

(V) Geometry of regression

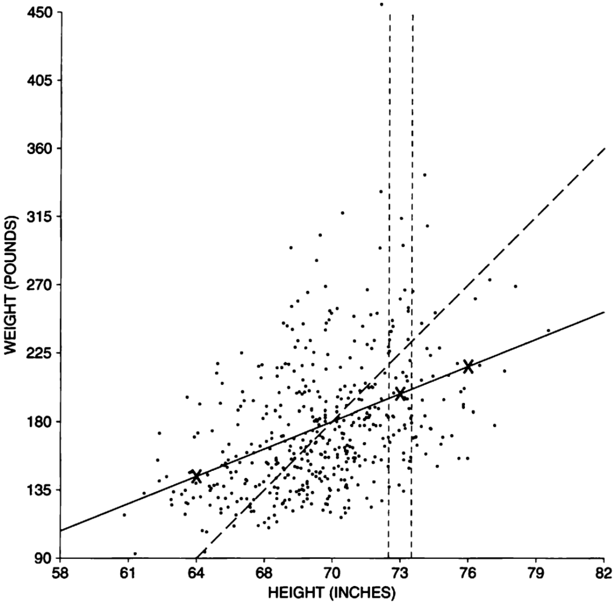
(II) Conditional distribution of y given x

- Find the range of x in the Pearson data set.
- For each integer h, calculate m_h, the mean of y values of the points with x values lying in the interval $[h - 0.5, h + 0.5)$.
- Add the points (h, m_h) to the scatter diagram. What does the graph of conditional means look like?
- For each integer h, calculate s_h, the SD of y values of the points with x values lying $[h - 0.5, h + 0.5)$. What do you observe about the conditional SD's?

(II) Nice scatter diagrams

- If a scatter diagram is shaped like an ellipse, then the conditional means are roughly on a straight line (the regression line), and it is homoschedastic: the conditional SD of y is roughly constant across x. The distribution is nicely summarised by five numbers: the means, the SDs, and the correlation.
- Fact: If x and y have a bivariate normal distribution, then their scatter diagram is like an ellipse.

(II) A pretty nice scatter diagram



FPP p159.

(II) Regression line of y on x

- The regression line of y on x is given by

$$y = mx + c$$

$$m = r\,\frac{s_y}{s_x}, \qquad c = \overline{y} - m\overline{x}$$

- Add this line to the scatter diagram. What do you observe about the regression line?

- Alternative expressions for the equation and gradient:

$$y = \overline{y} + m(x - \overline{x}), \qquad m = \frac{s_{xy}}{s_x^2}$$

(II) Interpreting m and c

- It seems natural to think m is the increase in y if we increase x by 1 for any point in the scatter diagram. This is a causal interpretation. Does it make sense for the Pearson data set?
- Consider two sets of points, the first with x values around x^*, and the second with x values around $x^* + 1$. Then the difference between the mean y values in the second group to the first group is around m. This is the meaning of:

  "Associated with a 1-unit increase in x, there is an increase in y of about m units."
- c is roughly the mean y values of those points with $x = 0$. Does it make sense for the Pearson data set?

(II) Regression effect

The equation of the regression line can be written

$$\frac{y - \overline{y}}{s_y} = r\,\frac{x - \overline{x}}{s_x}$$

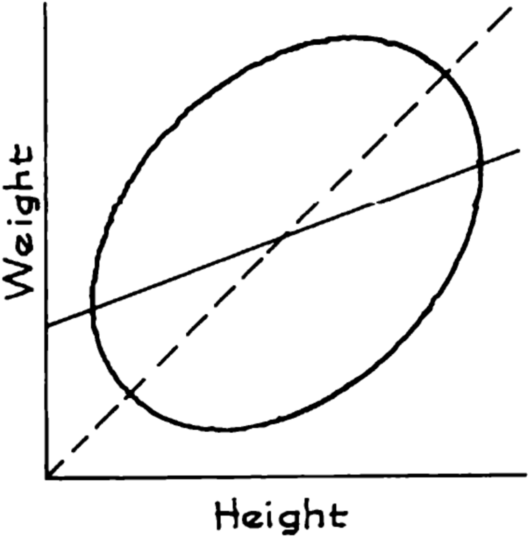
- Take fathers who are 1 SD above average. Their sons' mean height is roughly how many SDs above average?
- Take fathers who are 2 SDs below average. Their sons' mean height is roughly how many SDs below average?

Pearson's "regression to mediocrity" happens whenever the scatter diagram is like an ellipse:

Associated with 1 SD increase in x, there is an increase in y of not 1 SD, but _____ SD's.

(II) SD line vs regression line

From Freedman, Pisani and Purves page 174.



(II) Less nice scatter diagrams

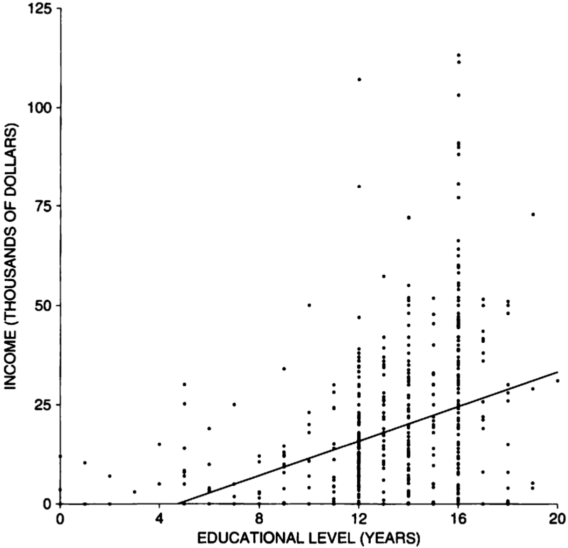
- Non-linear association. The the regression line does not track the conditional means so well.
- Heteroschedastic: conditional SD of y varies with x.

In these cases, the five numbers may not summarise the distribution adequately.

Sometimes the data can be transformed to make the scatter diagram nicer.

Even if a scatter diagram is not nice, the regression line is still the least square line.

(II) A heteroschedastic scatter diagram



FPP p192.

Topics

(I) Summarising two variables

(II) Regression line of y on x

(III) Predicting y from x

(IV) Analysis of variance

(V) Geometry of regression

(III) Predicting y from x

- Given a new subject with (x, y) values (x^*, y^*), the regression line can be used to predict y^* as

$$\overline{y} + m(x^* - \overline{x})$$

  The prediction error is

$$y^* - (\overline{y} + m(x^* - \overline{x})) = y^* - \overline{y} - m(x^* - \overline{x})$$

  Thus, "actual $=$ predicted $+$ error".
- The general idea, not restricted to the regression line, is that if prediction errors are rather small, then there is less need to measure y^*. This is clearly advantageous.

This hope is behind all applications in predictive analytics.

(III) “Internal” prediction

Getting a new subject takes work. It is simpler to gauge the prediction power of the regression line on the same data.

- For subject i, the predicted value is

$$\hat{y}_i = \overline{y} + m(x_i - \overline{x})$$

- The prediction error or residual is

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

Hence,

$$\begin{array}{ccccc} y_i & = & \hat{y}_i & + & e_i \\ \text{actual} & = & \text{predicted} & + & \text{error} \end{array}$$

(III) Least square line

Imagine predicting y using $ax + b$, the line with gradient a and y-intercept b. Then the errors are

$$y_i - (ax_i + b)$$

What choice of a and b minimises the mean squared error?

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Answer: $a = m = r\,\frac{s_y}{s_x}, b = c = \overline{y} - m\overline{x}$: the regression line.
This is why it is also called the least square line.

(III) Calculus: Normal equations

Goal: find a and b that minimise

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Set the two partial derivatives to 0, then solve for a and b.

(III) Algebraic approach

Regression line prediction errors $e_i = y_i - mx_i - c$.

For the line $y = ax + b$, errors are $f_i = y_i - ax_i - b$.

Goal: show that

$$\frac{1}{n}\sum_{i=1}^{n} f_i^2 \geq \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

The only way equality holds is $a = m$ and $b = c$. I.e., the regression line is the least square line.

Key starting result: $f_i = e_i + (m-a)x_i + (\overline{y} - m\overline{x} - b)$, or

$$f = e + (m-a)x + (\overline{y} - m\overline{x} - b)$$

(III) General facts about regression residuals

$e = y - mx - c$, i.e., $e_i = y_i - mx_i - c$.

1. $$\sum_{i=1}^{n} e_i = 0$$

   Hence $\overline{e} = 0$.

2. $$\sum_{i=1}^{n} x_i e_i = 0$$

   Hence $\overline{xe} = 0$.

3. Consequently, $\mathrm{cov}(x, e) = 0$.

(III) Algebraic proof (continued)

From $f = e + (m-a)x + (\overline{y} - m\overline{x} - b)$, we get

$$\begin{aligned} \mathrm{var}(f) &= \mathrm{var}(e) + (m-a)^2 \mathrm{var}(x) + 2(m-a)\mathrm{cov}(e,x) \\ &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 + (m-a)^2 s_x^2 \end{aligned}$$

Since $\mathrm{var}(f) = \frac{1}{n}\sum_{i=1}^{n} f_i^2 - (\overline{y} - a\overline{x} - b)^2$,

$$\frac{1}{n}\sum_{i=1}^{n} f_i^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 + (m-a)^2 s_x^2 + (\overline{y} - a\overline{x} - b)^2$$

Hence for any other line, $\frac{1}{n}\sum_{i=1}^{n} f_i^2 \geq \frac{1}{n}\sum_{i=1}^{n} e_i^2$, and equality implies $a = m$ and $b = c$.

(III) On prediction

The LS line minimises the MSE of prediction errors, among all straight lines.

- Since these predictions are internal, the MSE may not apply to new subjects.
- A better gauge of LS line's predictive power is to collect data on new subjects, by calculating the MSE of the new prediction errors.
- While waiting for new subjects, one reasonable way is to separate the existing subjects into two sets. The LS line based on the first set (training data) is used to make predictions for the second set (testing data).

Topics

(I) Summarising two variables

(II) Regression line of y on x

(III) Predicting y from x

(IV) Analysis of variance

(V) Geometry of regression

(IV) Analysis of variance

- Back to regression:
$$y = \hat{y} + e$$
Variation in y from two sources: along the line, in $\hat{y}$, and around the line, in e.
- Show that $\text{cov}(\hat{y}, e) = 0$.
- Consequently,
$$\text{var}(y) = \text{var}(\hat{y}) + \text{var}(e)$$

$\text{var}(y)$ has been separated into two components.

(IV) Regression RMSE

- Show that $\text{var}(\hat{y}) = r^2 s_y^2$.
- Hence
$$\text{var}(e) = \text{var}(y) - \text{var}(\hat{y}) = (1 - r^2)s_y^2$$
- Since $\overline{e} = 0$,
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{1 - r^2}\,s_y$$

  This is the root-mean-square error of the regression line. It quantifies the mean size of the prediction errors.

(IV) A picture

Sketch a triangle with sides of length

$$s_y \qquad |r|\, s_y \qquad \sqrt{1-r^2}\, s_y$$

Is there a right angle? Deduce that $r^2 \leq 1$.

(IV) More on RMSE

Focus on points with x values in an interval of width 1. Consider the RMS of their regression prediction errors.

- If a scatter diagram is homoschedastic, the above RMS is roughly equal to the RMSE $= \sqrt{1 - r^2} s_y$, for any interval.
- what can you say if it is heteroschedastic?

(IV) Explained variance

Call

$$\begin{aligned}\text{var}(y) &: \text{ total variance}\\ \text{var}(\hat{y}) &: \text{ explained variance}\\ \text{var}(e) &: \text{ unexplained or residual variance}\end{aligned}$$

Since $\text{var}(\hat{y}) = r^2\text{var}(y)$, it is often said that

> r^2 is the fraction of variance explained by the regression.

The exact meaning is unclear, because variance is not a good measure of spread, unlike SD.

(IV) Example with $r = 0.95$

- $r^2 \approx 0.9$, so we might be pleased to announce

  "Around 90% of the variance in y has been explained by the regression on x."

- The variance of the residuals is about 10% of the total variance. But their SD is $\sqrt{1 - r^2} \approx 30\%$ of SD(y). The spread in the residuals is only 30% of the spread in y, not 10% as one might expect from the statement.

(IV) The R command `lm`

- Run these lines, then match numbers in `regr` with symbols.

```
regr = lm(y ~ x)
regr
```

```
regr$coe
regr$fit
regr$res
attributes(regr)
```

- Match numbers from below with symbols.

```
anova(regr)
```

Topics

(I) Summarising two variables

(II) Regression line of y on x

(III) Predicting y from x

(IV) Analysis of variance

(V) Geometry of regression

(V) Points in space

- The scatter diagram presents data as n points with coordinates $(x_1, y_1), \ldots, (x_n, y_n)$. It is useful to think of $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ as points in an n-dimensional space called the Euclidean space.
- By Pythagoras' Theorem, the distance between x and y, $d(x, y)$, is $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ in the plane, and $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$ in space.
- In n dimensions, $d(x, y)$ is *defined* as $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
- We can also see x and y as vectors going from the origin to the points. Then the length of vector x is $|x| = \sqrt{\sum_{i=1}^{n} x_i^2}$, which is also the distance from the origin to point x: $d(0, x)$.

(V) Dot product

The dot product between x and y is

$$x \cdot y = \sum_{i=1}^{n} x_i y_i$$

Some general properties:

- $x \cdot y = y \cdot x$.
- For vectors x, y, z and numbers a, b, c,

$$(ax + by + c) \cdot z = a\, x \cdot z + b\, y \cdot z + c \sum_{i=1}^{n} z_i$$

- $|x| = \sqrt{x \cdot x}$.

(V) Orthogonality

- Pythagoras' Theorem: in the plane or space, x and y are perpendicular if and only if $x \cdot y = 0$.
- Definition: In n dimensions, x and y are orthogonal if $x \cdot y = 0$, denoted $x \perp y$.
- The distance squared between x and y, $d(x, y)^2$, is

$$\begin{aligned}(x-y)\cdot(x-y) &= x\cdot x - x\cdot y - y\cdot x + y\cdot y \\ &= x\cdot x + y\cdot y - 2x\cdot y\end{aligned}$$

Hence $|x-y|^2 = |x|^2 + |y|^2$ exactly when $x \perp y$. This is the high-dimensional Pythagoras' Theorem.

(V) Orthogonal projection

Let x and y be non-zero vectors. The orthogonal projection of y on x is the vector κx, where κ is a constant such that $y - \kappa x \perp x$. This generalises "dropping a perpendicular" for $n = 2, 3$.

(V) Regression and projection

Suppose $s_y, s_x > 0$. Define

$$v = y - \overline{y} = (y_1 - \overline{y}, \ldots, y_n - \overline{y}) \qquad u = x - \overline{x} = (x_1 - \overline{x}, \ldots, x_n - \overline{x})$$

Then v and u are non-zero.

- Fact: The orthogonal projection of v on u is mu, where $m = r \frac{s_y}{s_x}$.

Indeed, $v - mu = (y - \overline{y}) - m(x - \overline{x}) = y - \hat{y} = e$, which is orthogonal to $x - \overline{x}$.

(V) A similar triangle

The triangle lies in n dimensions.

Length of	Formula
$y - \overline{y}$	$\lvert y - \overline{y} \rvert = \sqrt{n}\,\mathrm{SD}(y)$
$\hat{y} - \overline{y}$	$\lvert \hat{y} - \overline{y} \rvert = \sqrt{n}\,\mathrm{SD}(\hat{y})$
$e = y - \hat{y}$	$\lvert e \rvert = \sqrt{n}\,\mathrm{SD}(e)$

Pythagoras' Theorem: $|y - \overline{y}|^2 = |\hat{y} - \overline{y}|^2 + |e|^2$ is the geometric interpretation of the analysis of variance: $\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(e)$.

Summary

- The regression line of y on x has slope $m = r\, s_y / s_x$ and y-intercept $c = \overline{y} - m\overline{x}$.
- If the scatter diagram is shaped like an ellipse, the conditional means of y roughly fall on the regression line, and the conditional SD is more or less constant (homoschedastic). In this case, the regression effect is often obtained: If $r > 0$, among points which are z s_x's above $\overline{x}$, their mean y values is above $\overline{y}$ by not z, but only rz, s_y's.
- The regression line is the unique least square line.
- Let the predicted values and residuals be $\hat{y} = \overline{y} + m(x - \overline{x})$ and $e = y - \hat{y}$. The analysis of variance says

$$\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(e)$$

  which has a geometric interpretation in n dimensions.
- The regression "explains" a proportion r^2 of the variation in y. Better: the RMSE of regression is $\sqrt{1 - r^2}s_y$.

Random variable

- A random variable (RV) is a mechanism that generates unpredictable numbers, but with a long-run stability.
- An RV is usually denoted by a capital letter, say X. Its realisation is denoted by a small letter, like x. X is random, x is fixed.
- Any x is unpredictable, but with many realisations, their histogram gets close to the distribution of X (the Law of Large Numbers).
- The distribution of X is centred at the expectation $\mathrm{E}(X)$ and its spread is indicated by the standard deviation $\mathrm{SD}(X) = \sqrt{\mathrm{var}(X)}$, the square root of the variance.

Expectation and variance

X, Y random variables, a, b, c constants.
Let $\mu_X = \mathrm{E}(X)$, $\mu_Y = \mathrm{E}(Y)$.

$$\begin{aligned} \mathrm{var}(X) &:= \mathrm{E}[(X-\mu_X)^2] \\ \mathrm{cov}(X,Y) &:= \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \end{aligned}$$

- $\mathrm{E}(aX + bY + c) = a\,\mu_X + b\,\mu_Y + c$.
- $\mathrm{var}(X) = \mathrm{E}(X^2) - \mu_X^2$.
- $\mathrm{var}(aX + bY + c) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) + 2ab\,\mathrm{cov}(X,Y)$.
- If X and Y are independent, $\mathrm{cov}(X,Y) = 0$.
- If X and Y are uncorrelated normal RV's, then they are independent.

Normal distribution

$X \sim \mathrm{N}(\mu, \sigma^2)$ has density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}$$

$\mathrm{E}(X) = \mu$, $\mathrm{var}(X) = \sigma^2$.

- If $a \neq 0$ and b are constants, $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$.
- $Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$ is a standard normal RV.
  $\Pr(-1 \leq Z \leq 1) \approx 0.68$, $\Pr(-2 \leq Z \leq 2) \approx 0.95$.
- If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y \sim \mathrm{N}(\nu, \tau^2)$ are independent, then

$$X + Y \sim \mathrm{N}(\mu + \nu, \sigma^2 + \tau^2)$$

Related distributions

Definitions of chi-square, t and F distributions:

- Let $Z_1, \ldots, Z_k \sim$ N(0,1) be independent. $\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$.
- Let $Z \sim$ N(0,1) and $V \sim \chi_k^2$ be independent.

$$\frac{Z}{\sqrt{V/k}} \sim t_k$$

- Let $W \sim \chi_h^2$ and $V \sim \chi_k^2$ be independent.

$$\frac{W/h}{V/k} \sim F_{h,k}$$

Matrix products (1)

- Let A be a $m \times n$ matrix, and let its (i, j)-entry (at row i and column j) be a_{ij}.
- A', the transpose of A, is the $n \times m$ matrix with (i, j)-entry equal to a_{ji}. If $A' = A$, A is symmetric.
- Let B be a $n \times p$ matrix. Then $C = AB$ is the $m \times p$ matrix with

  $$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \qquad 1 \leq i \leq m, 1 \leq j \leq p$$

  $C' = B'A'$.
- What is the size of $A'A$? What are its entries in terms of a_{ij}?
- What is the size of AA'? What are its entries in terms of a_{ij}?

Matrix products (2)

- An $n \times 1$ matrix is a column vector, denoted by a small letter and often written as a transpose. If

$$a = (a_1, \ldots, a_n)', \qquad b = (b_1, \ldots, b_n)'$$

what are $a'b$, ab', $a'a$ and aa'?

- Write the data variables x and y as column vectors.
  1. Express $x \cdot y$ as a matrix product, in two ways.
  2. Express the distance between the points x and y, and the length of the vector x, as matrix products.

Random vectors (1)

Let $Y_1, \ldots, Y_n$ be RV's.

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

is a random vector.

- The expectation is

$$\mu_Y = \mathrm{E}(Y) = \begin{bmatrix} \mathrm{E}(Y_1) & \cdots & \mathrm{E}(Y_n) \end{bmatrix}'$$

- The variance is

$$\mathrm{var}(Y) = \mathrm{E}\{(Y - \mu_Y)(Y - \mu_Y)'\}$$

What is the size of $\mathrm{var}(Y)$, and what is the (i,j)-entry?

Random vectors (2)

Y: $n \times 1$ random vector, A: $m \times n$ constant matrix, b: $m \times 1$ constant vector. Let $W = AY + b$.

- $$E(W) = A\mu_Y + b$$
- $$\mathrm{var}(W) = A\,\mathrm{var}(Y)A'$$

What are the entries of $\mathrm{var}(W)$ in terms of σ_{ij}?

Fact: Suppose $Y_1, \ldots, Y_n$ are independent normal RV's, and none of A's rows consists of only zeros.
Then W has a multivariate normal distribution. In particular, each W_i is normally distributed with expectation being the i-entry of $E(W)$ and variance being the (i, i)-th entry of $\mathrm{var}(W)$.