Lecture 2 – Program

1. Introduction

2. Simultaneous distributions

3. Covariance and correlation

4. Conditional distributions

5. Prediction

Basic ideas

We will often consider two (or more) variables simultaneously.

Examples (B& S, page 15)

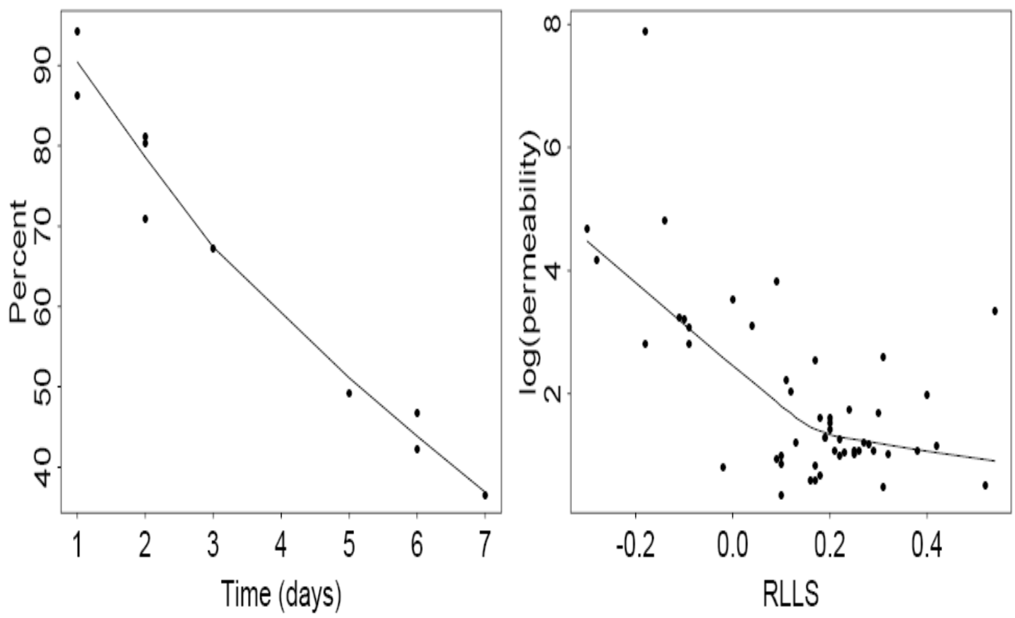


Figure 1: Scatter plots of serum gold concentrations (in percent of the original) against days after injection (left) and log(permeability) against RLLS (right).

There are two typical ways this is can be done:

(1) The data $(x_1, y_1), \ldots, (x_n, y_n)$ are considered as independent replications of a pair of random variables, (X, Y).

(2) The data are described by a linear regression model

$$y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Here $y_1, \ldots, y_n$ are the responses that are considered to be realizations of random variables, while $x_1, \ldots, x_n$ are considered to be fixed (i.e. non-random) and the ε_i's are random errors (noise)

Situation 1) occurs for *observational studies*, while situation 2) occur for *planned experiments* (where the values of the x_is are under the control of the experimenter).

In situation 1) we will often *condition* on the observed values of the x_i's, and analyse the data as if they are from situation 2)

In this lecture we focus on situation 1)

Joint or simultaneous distributions

The most common way to describe the simultaneous distribution of a pair of random variables (X, Y), is through their *simultaneous probability density*, $f(x, y)$

This is defined so that

$$P(\,(X, Y) \in A\,) = \int_A f(x, y)\, dx\, dy$$

The *marginal density* of X is obtained by integrating over all possible values of Y:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

and similarly for the marginal density $f_2(y)$ of Y.

If $f(x, y) = f_1(x)\, f_2(y)$, the random variables X and Y are *independent*.

Otherwise, they are *dependent*, which means that there is a relationship between X and Y, so that certain realizations of X tend to occur more often together with certain realizations of Y than others.

Covariance and correlation

The dependence between X and Y is often summarized by the *covariance*:

$$\gamma = \text{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)]$$

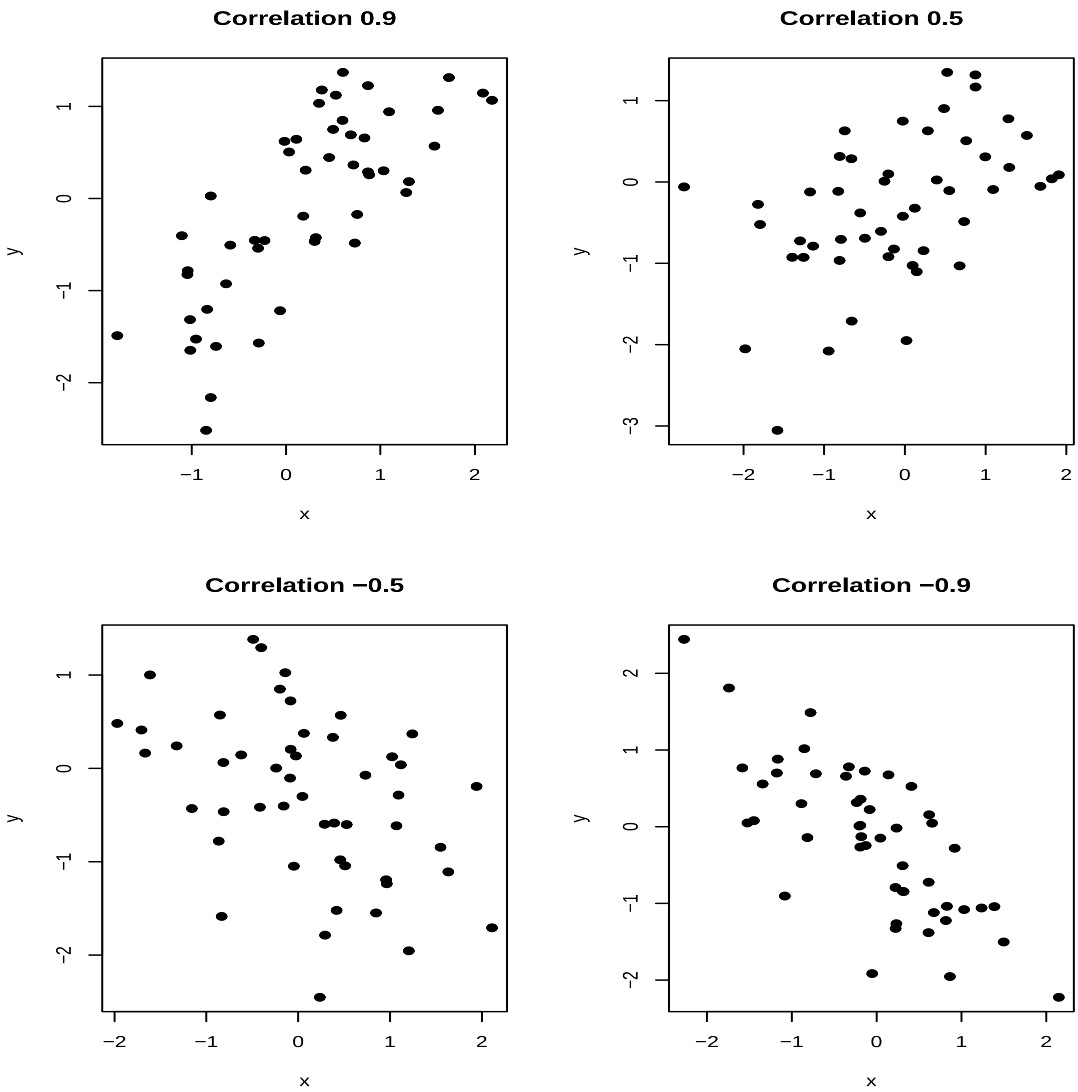
and the *correlation coefficient*:

$$\rho = \text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X)\,\text{sd}(Y)}$$

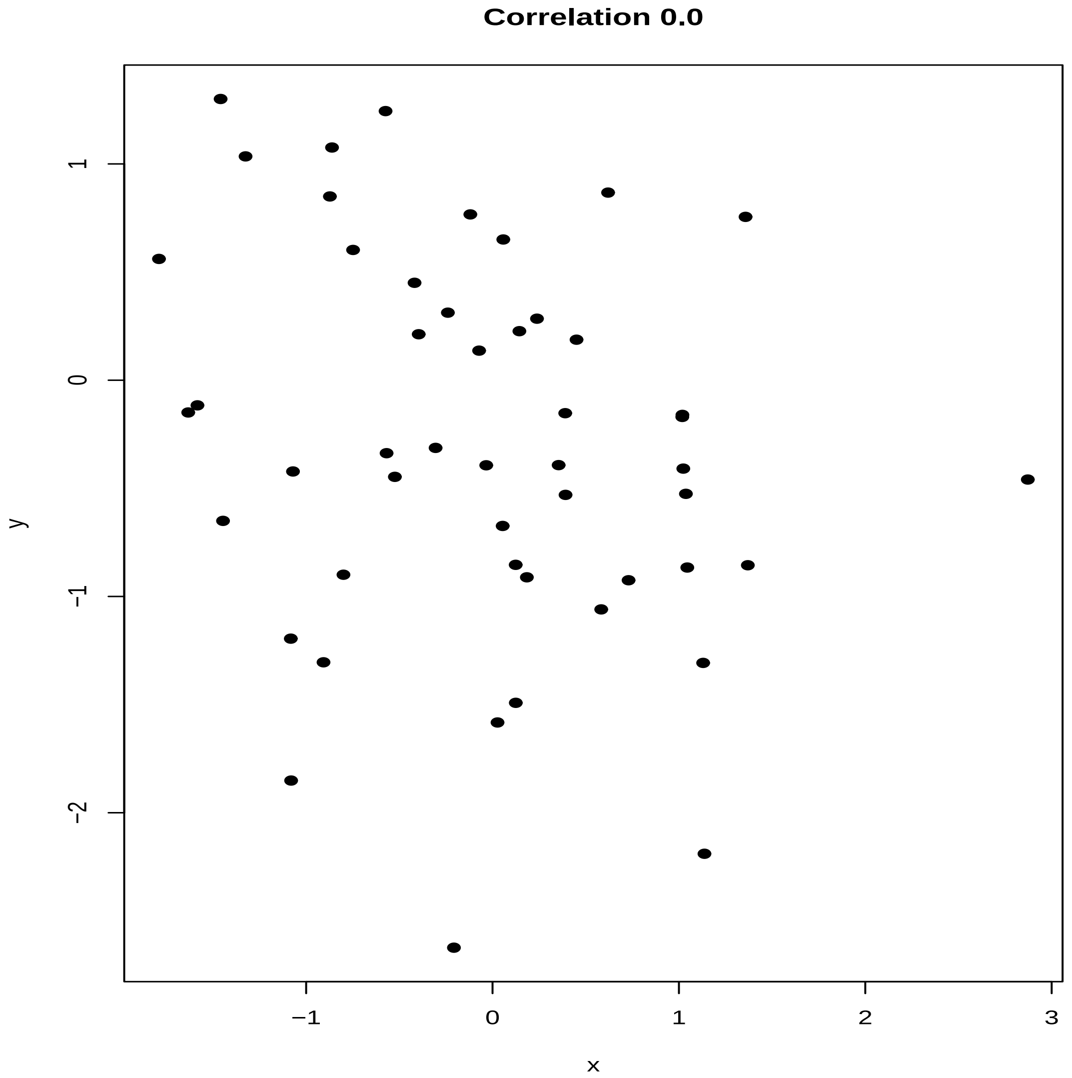
The following are important properties of the correlation coefficient.

- $\text{corr}(X, Y)$ takes values in the interval $[-1, 1]$
- $\text{corr}(X, Y)$ describes the *linear* relationship between Y and X.
- If X and Y are independent $\text{corr}(X, Y) = 0$, but not (necessarily) the other way around

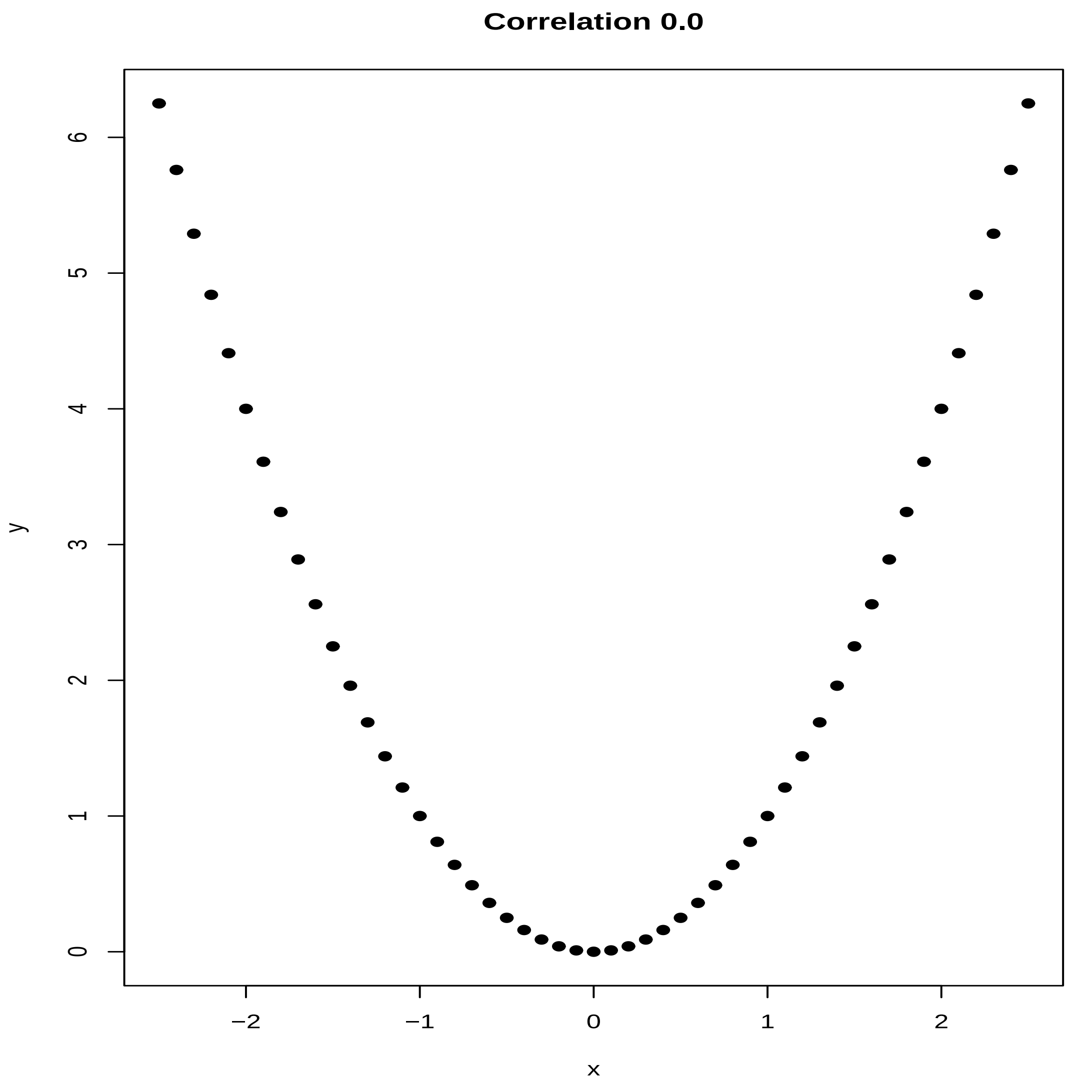
Correlation: correlated data



Correlation: uncorrelated data

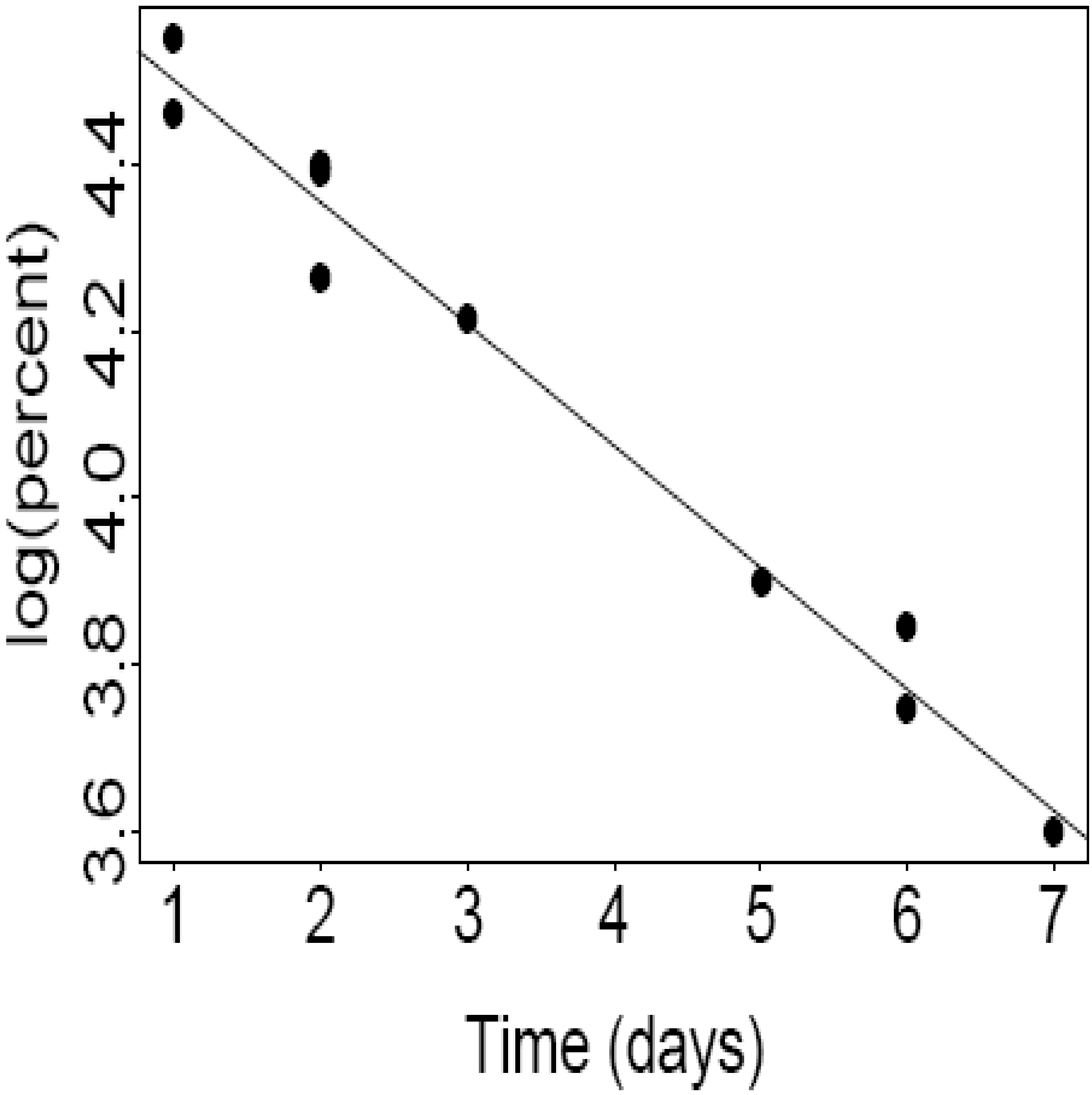


Correlation: uncorrelated data



Transformations

Sometimes a transformation may improve the linear relation



Sample versions of covariance and correlation

Data $(x_1, y_1), \dots, (x_n, y_n)$ are independent replicates of (X, Y).

Empirical analogues to the population concepts and basic results:

- Empirical covariance:

$$\widehat{\gamma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

- Empirical correlation coefficient:

$$\widehat{\rho} = \frac{\widehat{\gamma}}{s_{1n}\, s_{2n}}$$

- When n increases:

$$\widehat{\gamma} \to \gamma$$

$$\widehat{\rho} \to \rho$$

Conditional distributions

The *conditional density* of Y given $X = x$ is given by

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)}$$

If X and Y are independent, so that $f(x,y) = f_1(x)f_2(y)$, we see that $f_2(y|x) = f_2(y)$. This is reasonable, and corresponds to the fact that there are no information in a realization of X about the distribution of Y

Using the conditional density, one may find the conditional mean and the conditional variance:

- Conditional mean: $\mu_{2|x} = E(Y|x)$
- Conditional variance : $\sigma^2_{2|x} = \mathrm{Var}(Y|x)$

When (X,Y) is bivariate, normally distributed, $\mu_{2|x}$ is linear in x, and is known as the regression of Y on $X = x$ (cf. below).

Prediction

When X and Y are dependent, it is reasonable that knowledge of the value of X can be used to improve the prediction for the corresponding realization of Y.

Let $\widehat{Y}(x)$ be such a predictor. Then:

- $\widehat{Y}(x) - Y$ is the *prediction error*

- $\widehat{Y}_{opt}(x) = E(Y|x)$ minimizes $E[(\widehat{Y}(x) - Y)^2]$, the *mean squared prediction error*

- $E(Y|x)$ will often depend on unknown parameters, and it may be complicated to compute

Linear prediction

It is convenient to consider *linear predictors*, i.e. predictors of the form:

$$\widehat{Y}_{lin}(x) = a + bx$$

Minimizing $E[(a + bX - Y)^2]$ w.r.t. a and b yields:

$$b = \frac{\gamma}{\sigma_1^2} \quad \text{and} \quad a = \mu_2 - b\mu_1$$

The minimum is $E[(\widehat{Y}_{lin}(x) - Y)^2] = \sigma_2^2\,(1 - \rho^2)$.

Note that if ρ^2 increases, the mean squared error decreases.

Linear prediction, contd.

Without knowledge of the value of X, the best predictor is the unconditional mean of Y, i.e. $\tilde{Y}_0 = \mu_2$.

This has mean squared error $E[(\tilde{Y}_0 - Y)^2] = \sigma_2^2$.

Hence, a sensible measure of the quality of a prediction is the ratio

$$\frac{E[(\widehat{Y}_{lin}(x) - Y)^2]}{E[(\tilde{Y}_0 - Y)^2]} = 1 - \rho^2.$$

For judging a prediction, the squared correlation coefficient is the appropriate measure.

When a and b are unknown, we plug in the empirical counterparts:

$$\widehat{b} = \frac{\widehat{\gamma}}{\widehat{\sigma}_1^2} \quad \text{and} \quad \widehat{a} = \widehat{\mu}_2 - \widehat{b}\widehat{\mu}_1 = \bar{y} - \widehat{b}\bar{x}$$

The bivariate normal distribution

When (X, Y) is bivariate normal:

- The distribution is described by the five parameters $\mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2$ and ρ
- The marginal distributions of X and Y are normal, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$
- $\text{corr}(X, Y) = \rho$ and $\text{Cov}(X, Y) = \rho\,\sigma_1\,\sigma_2$
- The conditional distributions are normal
- $E(Y|x) = \mu_2 + \rho\,\frac{\sigma_2}{\sigma_1}\,(x - \mu_1)$
- $\text{Var}(Y|x) = \sigma_2^2\,(1 - \rho^2)$
- $b = \rho\,\frac{\sigma_2}{\sigma_1} = \frac{\gamma}{\sigma_1^2}$ and $a = \mu_2 - b\mu_1$