Chapter 0 Basic Prerequisite Knowledge and Introduction

1 Statistical analysis of one variable

1.1 (Random) Statistical observations

Suppose we observe n subjects from a **population**. One variable, Y, is measured for each subject and the values (called n observations) are

$$Y_1, \quad Y_2, \quad ..., \quad Y_n.$$

- Sample mean:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- sample variance

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  also denoted by $s^2(Y)$.

  Simple facts:

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0;$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

- Standard deviation

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

  also denoted by $s(Y)$.

  Example: the observations of heights, denoted by H: 1.84, 1.67, 1.68, 1.42, 1.54, 1.59, 1.60, 1.74, 1.83, 1.65, 1.51, 1.80, 1.64, 1.80, 1.62, 1.67, 1.67, 1.69, 1.74, 1.73

Then, we have the sample mean

$$\bar{H} = \frac{1}{20}\{1.84 + 1.67 + ... + 1.73\} = 1.6715$$

and sample variance

$$s_H^2 = \frac{1}{20-1}\{(1.84-1.6715)^2 + (1.67-1.6715)^2 + ... + (1.73-1.6715)^2\} = 0.01169763$$

standard deviation

$$s_H = \sqrt{s_H^2} = 0.1081556$$

- Estimation of population parameters, $EY = \mu_Y$ (or μ) and $Var(Y) = \sigma_Y^2$ (or σ^2, or $\sigma^2(Y)$). They can be estimated as

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = S_Y^2, \qquad \hat{\sigma} = S_Y$$

or

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/n, \quad \hat{\sigma} = \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/n}.$$

Example For the above data, we have

$$\hat{\mu}_H = 1.6715, \qquad \hat{\sigma}_H = 0.1081556$$

- Distribution of the observations: Histogram.

For the above data, its histogram is shown below, suggesting that they are normally distributed

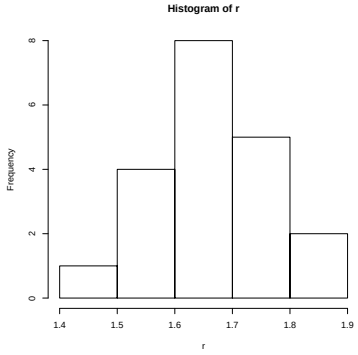


Figure 1:

Please check the following histograms. Are they normally distributed?

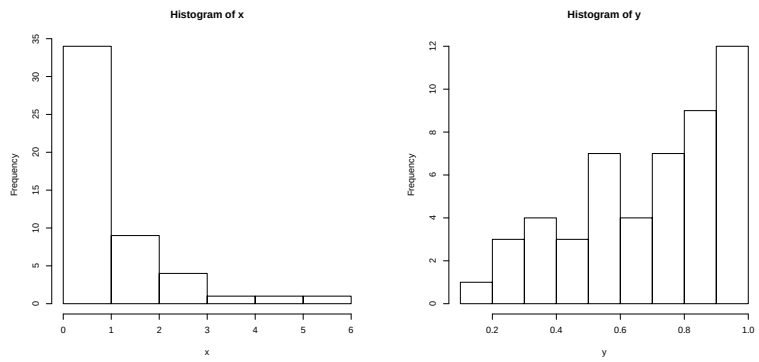


Figure 2:

- Hypothesis testing: Suppose $Y_1, ..., Y_n$ are samples from $N(\mu, \sigma^2)$. We can test, for example, $H_0 : \mu = \mu_0$.

  – If σ is known, then we use the Z-statistic. Under H_0

  $$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

  at significant level $\alpha = 0.05$ (say),

  $$\text{we accept } H_0, \text{ if } |Z| \leq 1.96$$

  $$\text{we reject } H_0, \text{ if } |Z| > 1.96$$

  – If σ is unknown, then we use the T-statistic. Under H_0

  $$T = \frac{\bar{Y} - \mu_0}{S_Y/\sqrt{n}} \sim t(n-1)$$

  at significant level α,

  $$\text{we accept } H_0, \text{ if } |T| \leq t_{1-\alpha/2}(n-1)$$

  $$\text{we reject } H_0, \text{ if } |T| > t_{1-\alpha/2}(n-1)$$

  Example Suppose we need to test $H_0 : \mu_H = 1.65$ based on the above data at significant level $\alpha = 0.01$. Calculate

  $$T = \frac{\bar{H} - 1.65}{0.1082/\sqrt{20}} = 0.8886$$

  Since $|T| < t_{0.995}(19) = 2.861$, we accept H_0.

2 population and random variable

- The distribution of the population and the distribution of Y.
- mean $E(Y)$, and variance $Var(Y)$ (or $\sigma^2(Y)$)

  simple fact: $Var(Y) = E\{(Y - E(Y))^2\} = EY^2 - (E(Y))^2$
- α-quantile with $0 < \alpha < 1$ for Y: q_α

$$P(Y \leq q_\alpha) = \alpha$$

2.1 Statistical distributions

- Normal: $X \sim N(\mu, \sigma^2)$, where μ and σ are two parameters. Then $(X-\mu)/\sigma \sim N(0,1)$ (standard normal), with p.d.f.

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

  See Figure 3. (*how to find the quantile (or critical values) in the statistical table?*)

  We have

$$P(X \leq -1.96) = 0.025, \qquad P(X \geq 1.96) = 0.025,$$

  and thus

$$P(|X| < 1.96) = 0.95$$

- Student distribution (or t-distribution): $t(v)$, where $v \geq 1$ is a parameter [called the number of degrees of freedom]. The p.d.f. is

$$f(x) = \frac{\Gamma\{(v+1)/2\}}{\sqrt{v\pi}\Gamma(v/2)}(1 + x^2/v)^{-(v+1)/2}$$

  See Figure 4.

  If $X \sim t(v)$, *how to find quantile (or critical value) q such that* $P(|X| > q) = \alpha$

  For example, if $X \sim t(2)$, then $P(|X| > 4.303) = 0.05$.

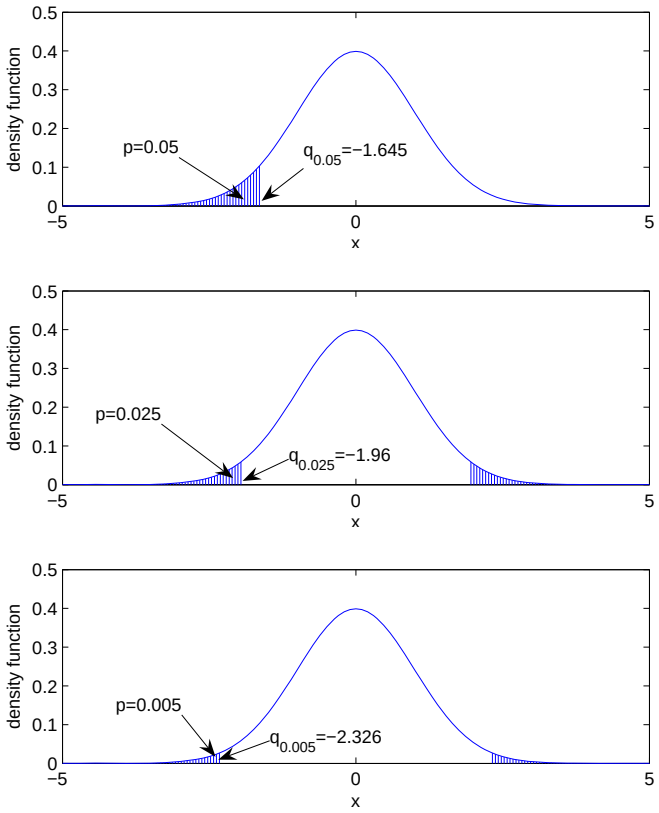


Figure 3: Normal distribution and quantiles (also known as critical values at significant level α)

- χ^2 distribution: $\chi^2(v)$, where v is a parameter [called the number of degrees of freedom]. See Figure 5.

  Example $X \sim \chi^2(10)$, then $P(X > 18.31) = 0.05$ (*How to find the quantiles (or critical value) in the statistical table?*)

- F distribution: $F(v_1, v_2)$, where $v_1, v_2 > 0$ are two parameters [called the numbers of degrees of freedom]. See Figure 6.

  Example $X \sim F(4, 10)$, then $P(X > 3.48) = 0.05$ (*How to find the quantiles (critical value) in the statistical table?*)

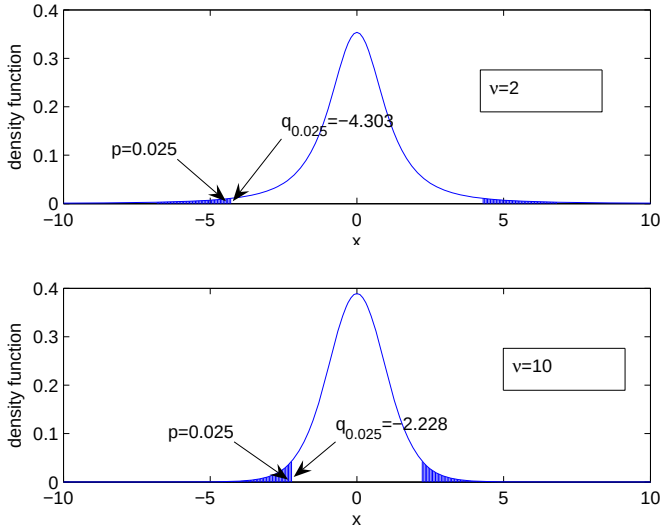


Figure 4: shapes of t-distribitions

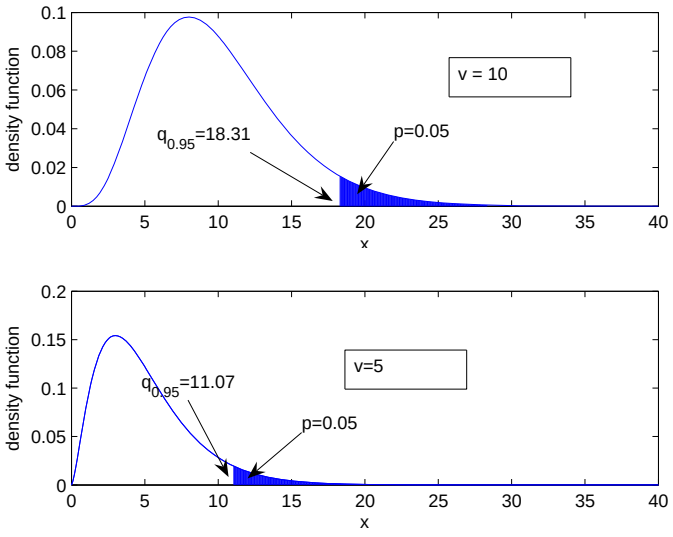


Figure 5: the shape of the density function for χ^2-distribution

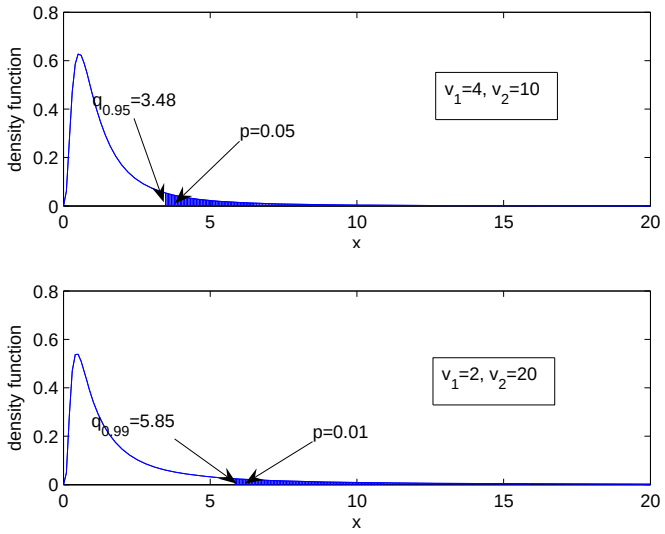


Figure 6: shapes of the density functions for F distribution

3 Statistical analysis for two variables

Suppose we observe n subjects from a population, TWO variables are measured for each subject. We have n observations

$$(X_1, Y_1), \quad (X_2, Y_2), \quad ..., \quad (X_n, Y_n)$$

from the population, denoted by (X, Y).

Besides the statistical analysis of each variable separately (see above), we are also interested in the relationship between X and Y

- (Sample) covariance

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Simple facts

$$\sum_{i=1}^{n}(X_i - \bar{X})\bar{Y} = 0; \quad \sum_{i=1}^{n}(Y_i - \bar{Y})\bar{X} = 0;$$
$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

- (Sample) correlation coefficient

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{S_X S_Y}$$

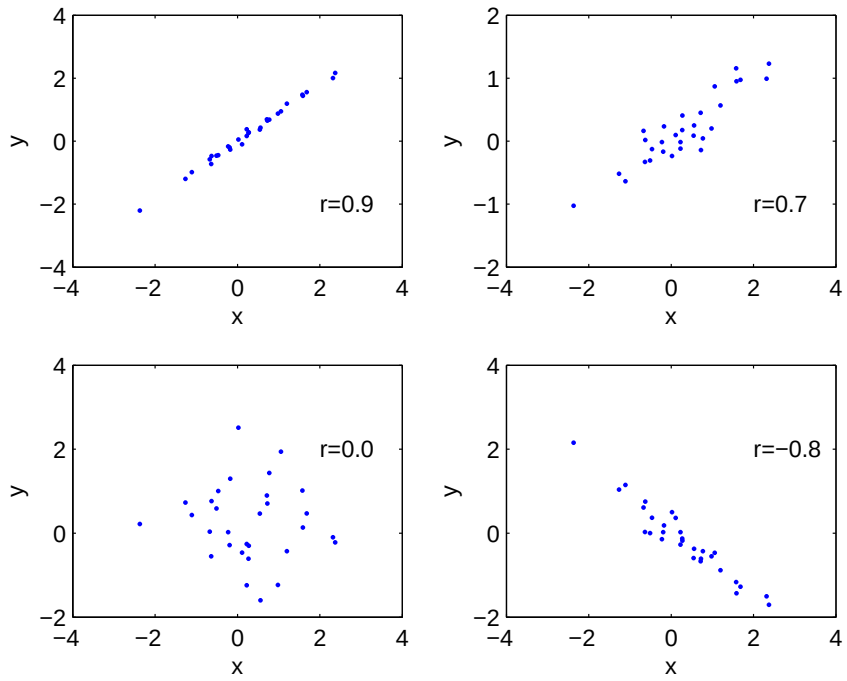
Some basic facts

- ♠ $-1 \leq r_{XY} \leq 1$.
- ♠ $r_{XY} = 0$, there is no linear correlation between X and Y
- ♠ $r_{XY} > 0$, there is positive linear correlation between X and Y
- ♠ $r_{XY} < 0$, there is negative linear correlation between X and Y
- ♠ $r_{XY} = \pm 1$, there is a constant c such that $Y_i = cX_i$

Example Suppose the observations for people's height (H) and weight (W) are: (1.84, 91.31) (1.67, 88.63) (1.68, 83.94) (1.42, 75.55) (1.54, 79.57) (1.59, 82.68) (1.60 80.41) (1.74, 82.42) (1.83, 92.21) (1.65, 79.63) (1.51, 71.15) (1.80 95.24) (1.64, 77.38) (1.80 91.67) (1.62, 79.57) (1.67, 80.64) (1.67, 87.26) (1.69, 89.52) (1.74, 93.50) (1.73, 88.57) we have $S_H = 0.1082; S_W = 6.6527$ and $S_{HW} = 0.6077$

$$r_{HW} = 0.8442.$$

Scatter plot of two variables and correlation coefficients



Discussion on linear Correlation

- "two variables have linear correlation" does not mean that they are causally related. Often a third variable, a lurking variable, that is not included in the analysis is responsible (causes) for the first two variables. A lurking variable is a variable that loiters in the background and affects both of the original variables
- the correlation coefficient can only detect the linear relationship, it may fail to detect the nonlinear relationships.

  Example

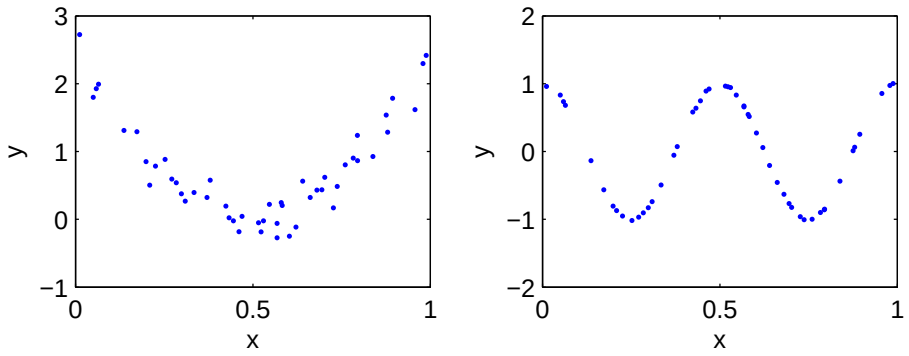


Figure 7: there is strong relationship between Y and X, but the linear correlation coefficient is 0 (there is no *overall* trend between Y and X)

- Population covariance

$$Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

The population correlation coefficient is defined as

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Simple facts

$$Cov(X,X) = Var(X), \qquad Cov(X,Y) = Cov(Y,X)$$

- Estimation of the correlation coefficient (by sample correlation coefficient)

$$\hat{\rho}_{XY} = r_{XY}.$$

4 Other relationship

- (deterministic) functional relationship [not discussed in this module]. For the (two) variables, X and Y, we hope to predict one variable based on the other(s). A functional (mathematical) relation allow us to make accurate/exact prediction.

  Example: for a circle, the circumference Y and its diameter X has a deterministic functional relation

  $$Y = \pi X$$

  and its area Z has a relation with X as

  $$Y = \frac{1}{4}\pi X^2$$

  See Figure 8.

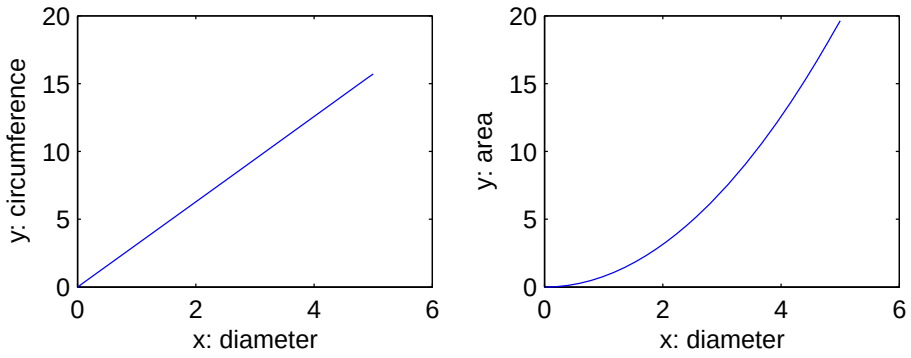


Figure 8: two examples of deterministic functional relationships

- **Regressive relationship (regression analysis)**. However, for most statistical problems, we cannot predict the "true" value because of random effect. We can only predict the "expected" value, i.e. $E(Y) = f(X)$. A simple case is

  $$E(Y) = a + bX$$

  called (simple) linear regression model.

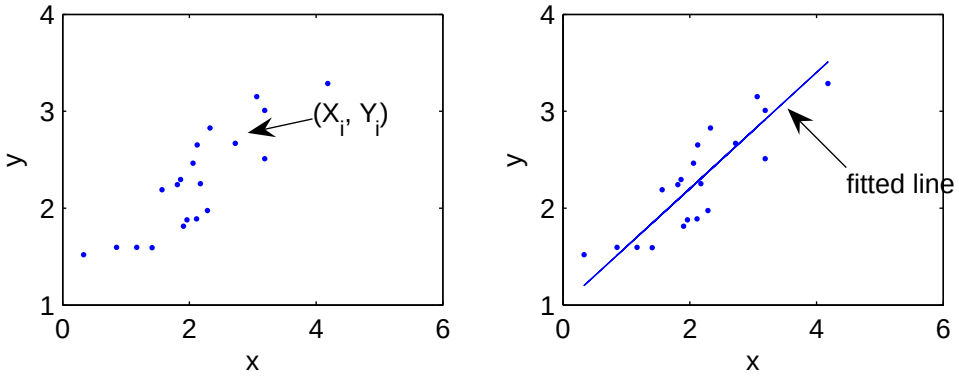


Figure 9: An example of linear regression model

♠ If (X, Y) have joint normal distribution, then the their relation can be modeled by

$$E(Y) = a + bX$$

or

$$Y = a + bX + \varepsilon \tag{1}$$

where ε is independent of X. Model (1) is also called linear regression model. The model can also be written as

$$Y_1 = a + bX_1 + \varepsilon_1$$

$$Y_2 = a + bX_2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = a + bX_n + \varepsilon_n$$

♠ Why do we call the model "regression"? The response variable Y tends to "revert" or "regress" to the mean of Y.

why linear regression is popular? and why it is widely used in practice?

* Liner regression relationship is easy to investigate and is stable
* If the joint distribution is normal, then their relationship is linear.
* Liner regression relationship is a good approximation, especially locally.

♠ Nonlinear regression

$$Y_i = f(X_i) + \varepsilon$$

See Figure 11.

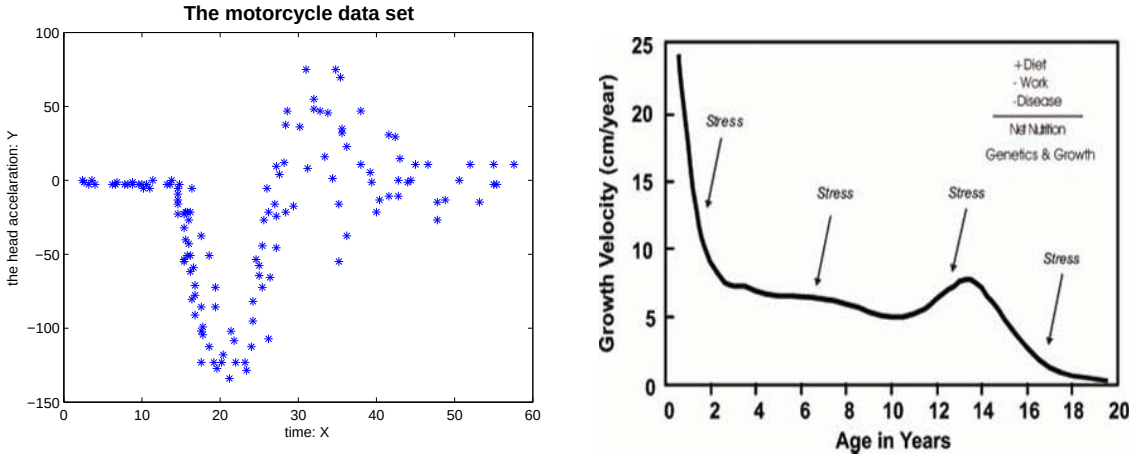


Figure 11: Two examples showing the nonlinear regression relationships

5 More about Expectation, variance and covariance

Suppose $a_0, a_1, ..., a_k$ are constants, and $\varepsilon_1, ..., \varepsilon_k$ are random variables, then

- $$E(a_0 + a_1\varepsilon_1 + ... + a_k\varepsilon_k) = a_0 + a_1E(\varepsilon_1) + ... + a_kE(\varepsilon_k)$$

- $$Var(a_0 + a_1\varepsilon_1 + ... + a_k\varepsilon_k) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_ia_jCov(\varepsilon_i, \varepsilon_j)$$

- $$Cov(a_1\varepsilon_1 + ... + a_k\varepsilon_k, b_1\xi_1 + ... + b_\ell\xi_\ell) = \sum_{i=1}^{k}\sum_{j=1}^{\ell} a_ib_jCov(\varepsilon_i, \xi_j)$$

  where $b_1, ..., b_\ell$ are constants and $\xi_1, ..., \xi_\ell$ are random variables.

- if $\varepsilon_1, ..., \varepsilon_k$ are mutually independent, then

$$Var(a_0 + a_1\varepsilon_1 + ... + a_k\varepsilon_k) = \sum_{i=1}^{k} a_i^2 Var(\varepsilon_i)$$

- if $\varepsilon_1, ..., \varepsilon_k$ are IID (independent and identically distributed), then

$$Var(\bar{\varepsilon}) = \frac{1}{k} Var(\varepsilon_1)$$

  where $\bar{\varepsilon} = (\varepsilon_1 + ... + \varepsilon_k)/k$

- Suppose $\xi_i \sim N(\mu_i, \sigma_i^2), i = 1, ..., n$ are independent and that $a_0, a_1, ..., a_n$ are constants. Then

$$a_0 + a_1\xi_1 + ... + a_n\xi_n \sim N(\tilde{\mu}, \tilde{\sigma}^2)$$

  where $\tilde{\mu} = a_0 + a_1\mu_1 + ... + a_n\mu_n$ and $\tilde{\sigma}^2 = a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$