生物统计学

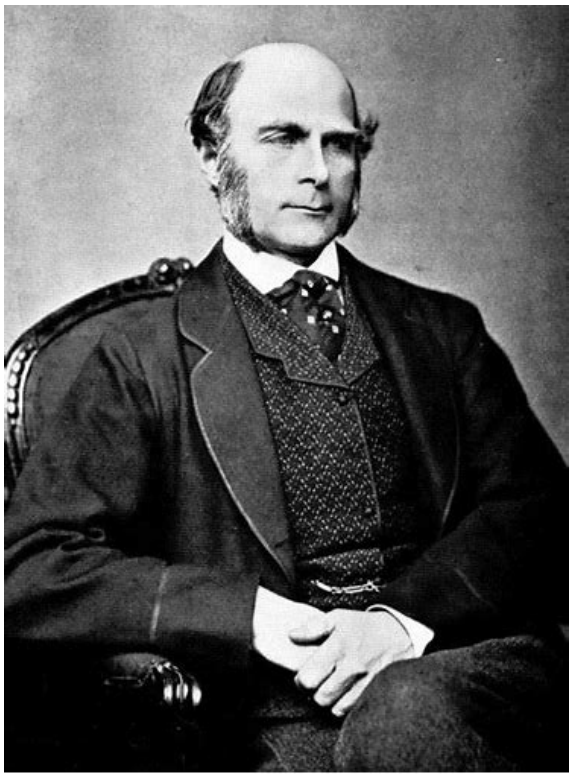
第十章 回归分析

云南大学 生命科学学院

會澤百家　至公天下

1. “回归”的故事
2. 回归与相关的基本概念
3. 线性回归分析
4. 非线性回归分析
5. 回归分析与方差分析的关系

10.1 “回归”的故事



Francis Galton (1822-1911)

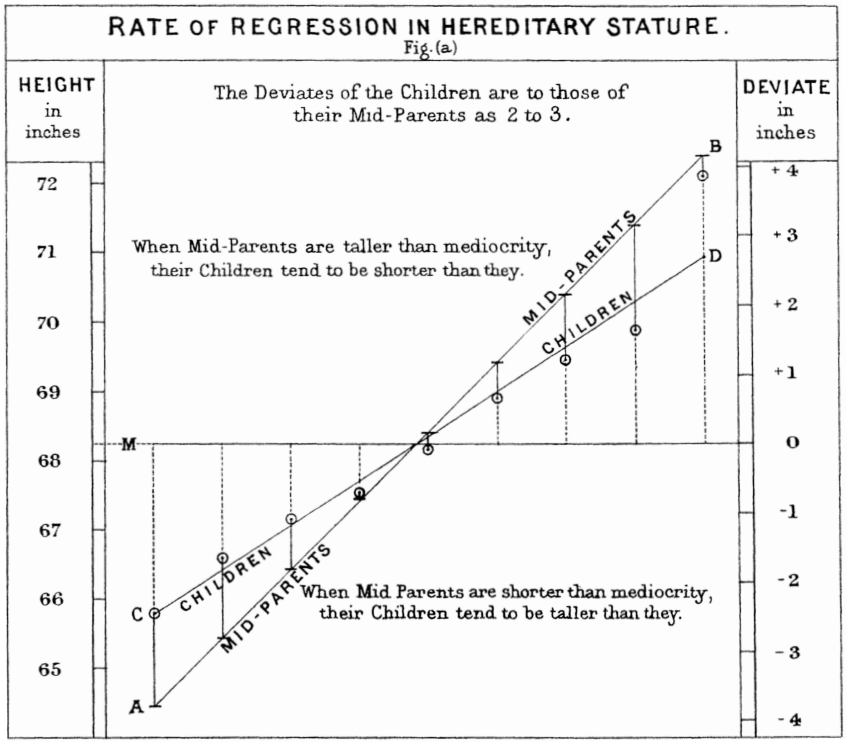


图 **10.1** Galton 绘制的子代身高（中位数，Y 轴）与父母平均身高（X 轴）的回归关系图（引自：Galton F. 1886. Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15:246–263.）

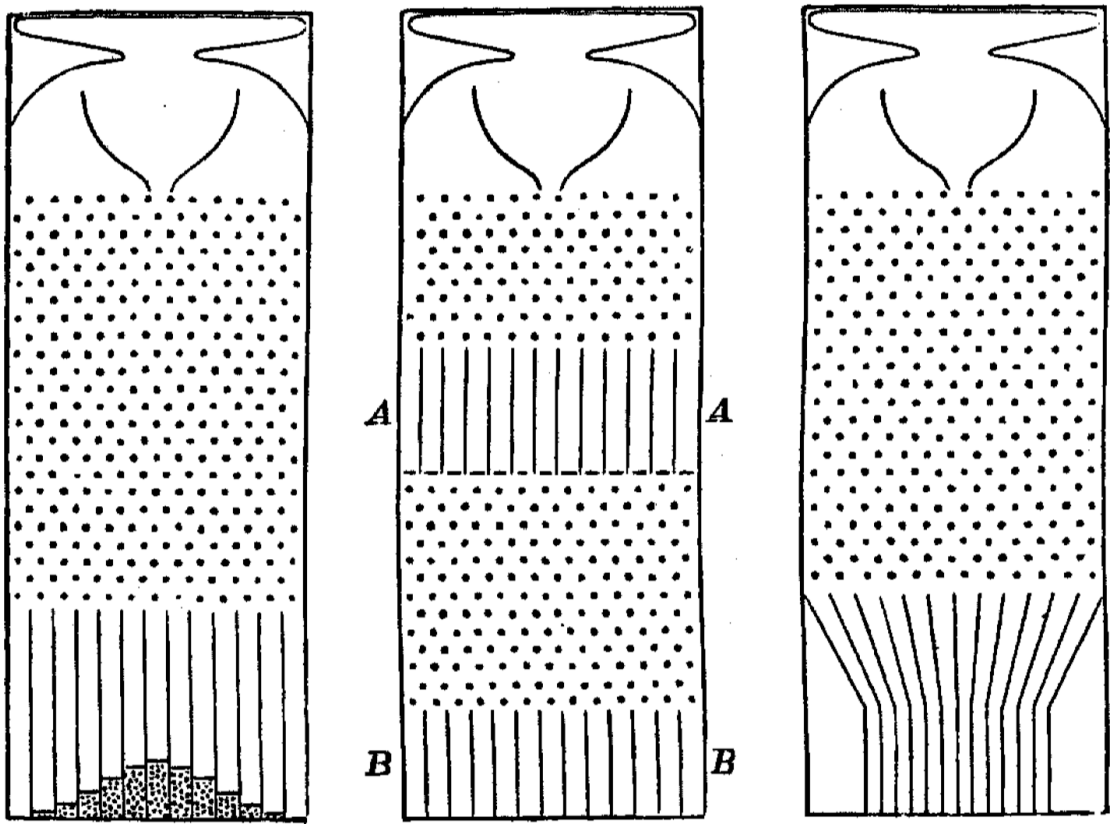


图 **10.2** Galton 梅花机示意图

（引自：Natural Inheritance. By Francis Galton, F.R.S. London: Macmillan and Co., 1889.）

10.1 “回归”的故事

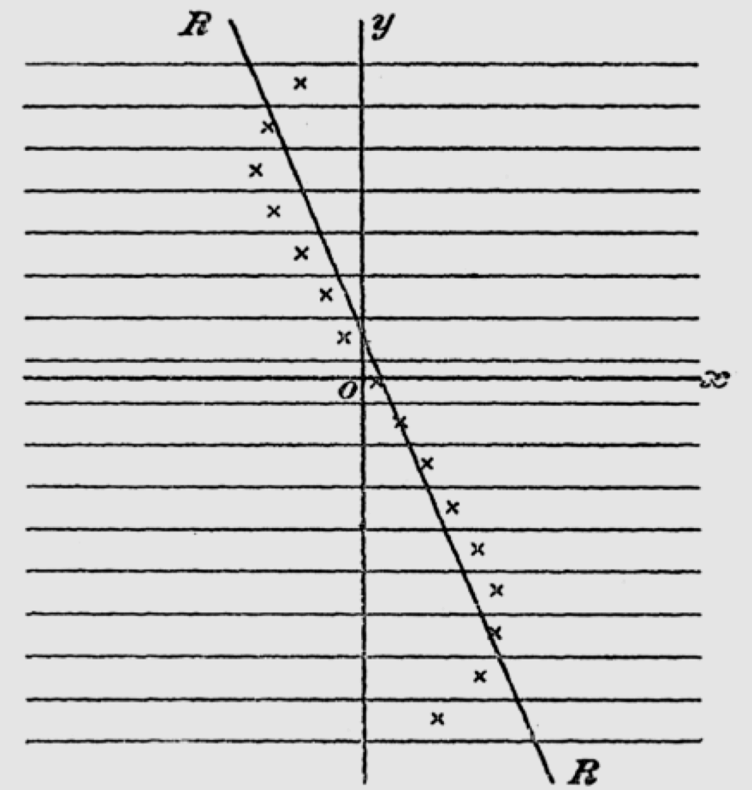


图 **10.3** 三维频率曲面上 y 值和平均 x 值的关系图（引自：Yule GU. 1897. On the significance of Bravais' formulæ for regression, &c., in the case of skew correlation. Proc. R. Soc. Lond., 60:477–489.）


GEORGE UDNY YULE

George U. Yule (1871-1951)

统计学上把这种变量的相互关系，称为协变关系(covariant relation)，具有协变关系的变量称为协变量(covariate)。

- 因果关系指一个变量（因变量，dependent variable）的变化受另一个变量或几个变量（自变量，independent variable）的制约。
- 平行关系指两个以上变量之间共同受其他因素的影响。

10.2 回归与相关的基本概念

统计学上把这种变量的相互关系，称为协变关系(covariant relation)，具有协变关系的变量称为协变量(covariate)。

- 因果关系指一个变量（因变量，dependent variable）的变化受另一个变量或几个变量（自变量，independent variable）的制约。
- 平行关系指两个以上变量之间共同受其他因素的影响。

研究协变关系，统计学上有回归分析(regression analysis) 和相关分析(correlation analysis) 两类方法。

- 根据回归的数学模型不同，可分为线性回归分析(linear regression analysis)和非线性回归分析(nonlinear regression analysis) 两类。
- 根据变量数量的不同，可分为简单相关分析(simple correlation analysis)和复相关分析(multiple correlation analysis)

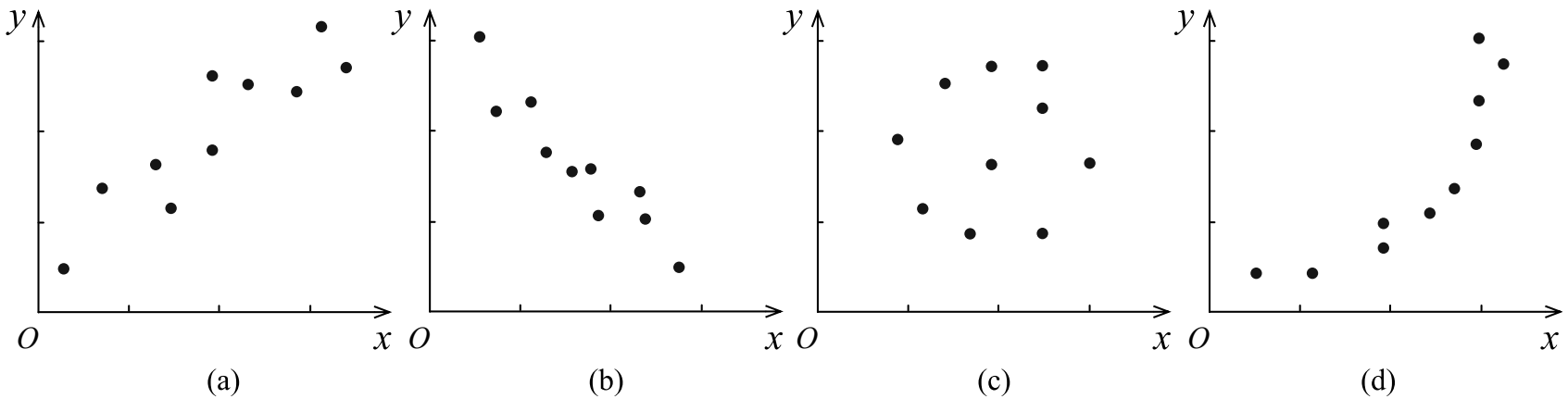


图 **10.4** x 和 y 的关系

线性回归的数学模型为

$$y = \mu_y + \beta(x - \mu_x) + \varepsilon \tag{10.1}$$

线性回归的数学模型为

$$y = \mu_y + \beta(x - \mu_x) + \varepsilon \tag{10.1}$$

$$y = (\mu_y - \beta\mu_x) + \beta x + \varepsilon \tag{10.2}$$

线性回归的数学模型为

$$y = \mu_y + \beta(x - \mu_x) + \varepsilon \tag{10.1}$$

$$y = (\mu_y - \beta\mu_x) + \beta x + \varepsilon \tag{10.2}$$

令 $\alpha = \mu_y - \beta\mu_x$，则线性回归的数学模型可改写为

$$y = \alpha + \beta x + \varepsilon \tag{10.3}$$

$$y = \alpha + \beta x + \varepsilon \tag{10.3}$$

该数学模型各部分的意义如下：

$$y = \alpha + \beta x + \varepsilon \tag{10.3}$$

该数学模型各部分的意义如下：

- α，是回归直线在纵坐标轴上的截距，故称作总体回归截距 (regression intercept)。

$$y = \alpha + \beta x + \varepsilon \tag{10.3}$$

该数学模型各部分的意义如下：

- α，是回归直线在纵坐标轴上的截距，故称作总体回归截距 (regression intercept)。
- βx，是因变量 y 的变化中，由 y 和 x 的线性回归关系决定的部分，也就是可用 x 估计的部分。β 称为总体回归系数(regression coefficient)，是回归直线的斜率(slope)。

$$y = \alpha + \beta x + \varepsilon \tag{10.3}$$

该数学模型各部分的意义如下：

- α，是回归直线在纵坐标轴上的截距，故称作总体回归截距 (regression intercept)。
- βx，是因变量 y 的变化中，由 y 和 x 的线性回归关系决定的部分，也就是可用 x 估计的部分。β 称为总体回归系数(regression coefficient)，是回归直线的斜率(slope)。
- ε，又称回归估计误差(errors of regression) 或残差(residual)。它表示 y 的变化中由 x 引起的以外，其它所有未被纳入该模型的部分。

基于线性回归模型的回归分析，应符合以下基本假定：

1. 自变量 x 是没有误差的固定变量，至少和因变量 y 相比，x 的误差可以忽略不计。而 y 是典型的随机变量，有随机误差。
2. 因 y 是随机变量，那么任意一个 x 实际上都对应于一个 y 总体。该 y 总体服从正态分布，有条件平均数 (conditional mean) $\mu_{y|x} = \alpha + \beta x$，且方差 $\sigma^2_{y|x}$ 不受 x 影响。
3. 随机误差 ε 相互独立，且服从正态分布 $N(0, \sigma_e^2)$。

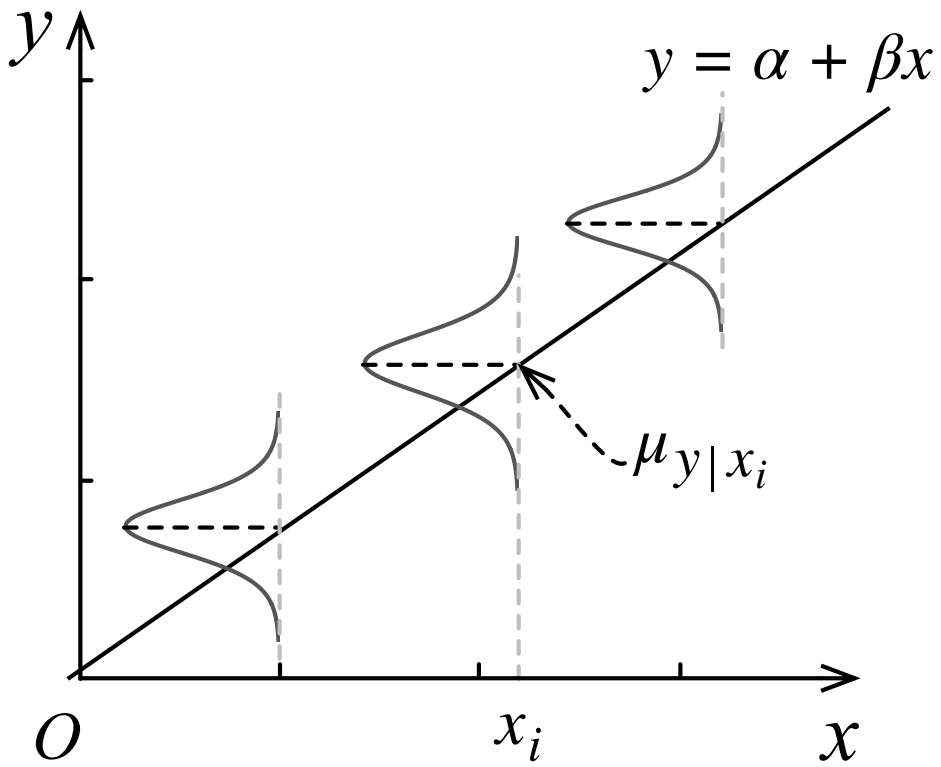


图 **10.5** 线性回归模型示意图

所谓回归方程，就是利用样本数据资料估计线性回归模型中的各项参数，并代入模型公式所得到的线性方程，即

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = a + bx \tag{10.6}$$

所谓回归方程，就是利用样本数据资料估计线性回归模型中的各项参数，并代入模型公式所得到的线性方程，即

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = a + bx \tag{10.6}$$

参数估计的方法即最小二乘法(the method of least squares)

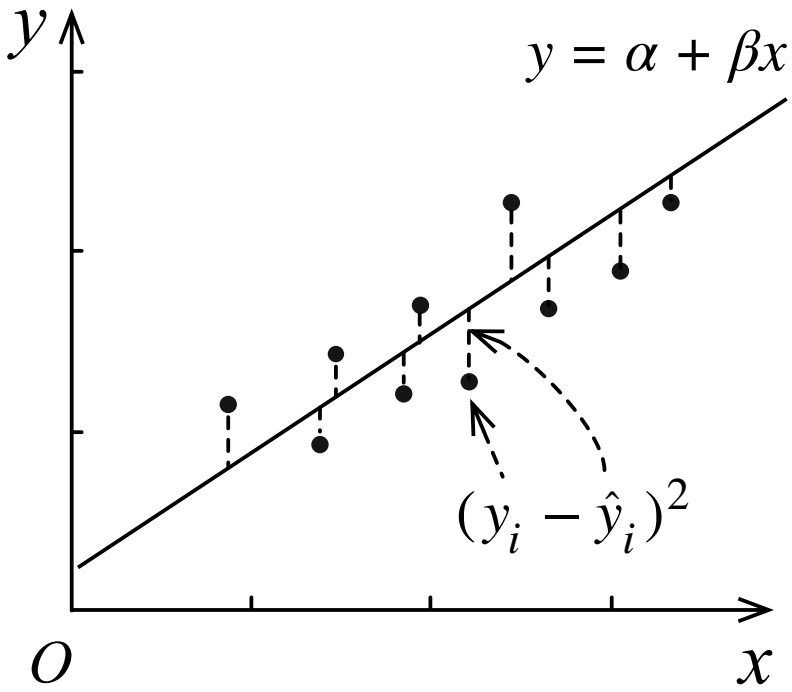


图 **10.6** 线性回归模型与观测值偏差

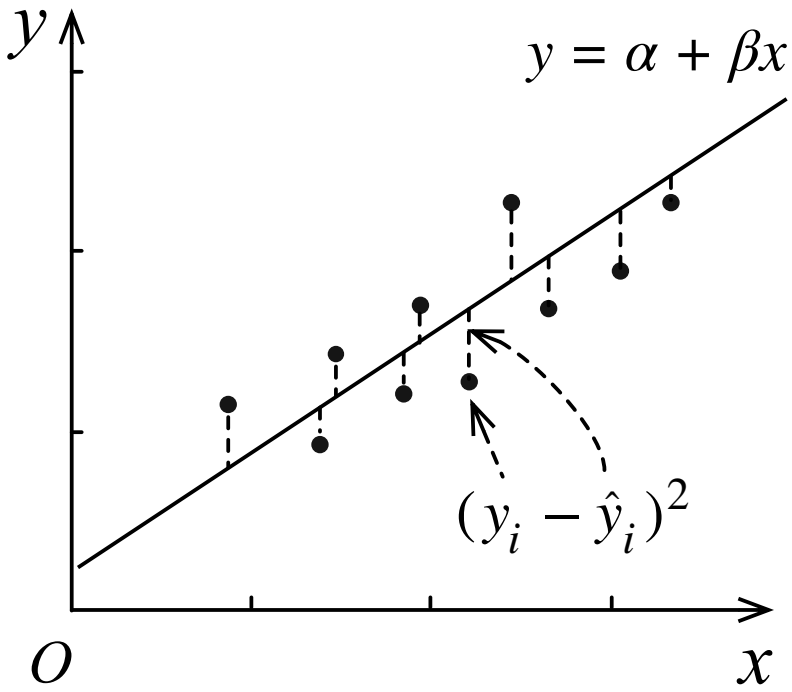


图 **10.6** 线性回归模型与观测值偏差

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{10.7}$$

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{cases} \tag{10.10}$$

- a 为样本回归截距，是回归直线与纵坐标轴交点的纵坐标，总体回归截距α 的无偏估计值；
- b 为样本回归系数，是回归直线的斜率，总体回归系数β 的无偏估计值。

b 的分子部分是 x 的离均差与 y 的离均差的乘积和，简称乘积和(sum of products)，记作 SP_{xy}；分母部分是 x 的离均差平方和，记作 SS_x。

- 回归系数的 t 检验
- 回归方程的 F 检验

零假设H_0：回归系数 $\beta = 0$；备择假设H_1：回归系数 $\beta \neq 0$

零假设H_0：回归系数 $\beta = 0$；备择假设H_1：回归系数 $\beta \neq 0$

$$t = \frac{b - \beta}{\frac{s_e}{\sqrt{SS_x}}} \sim t(n-2) \tag{1}$$

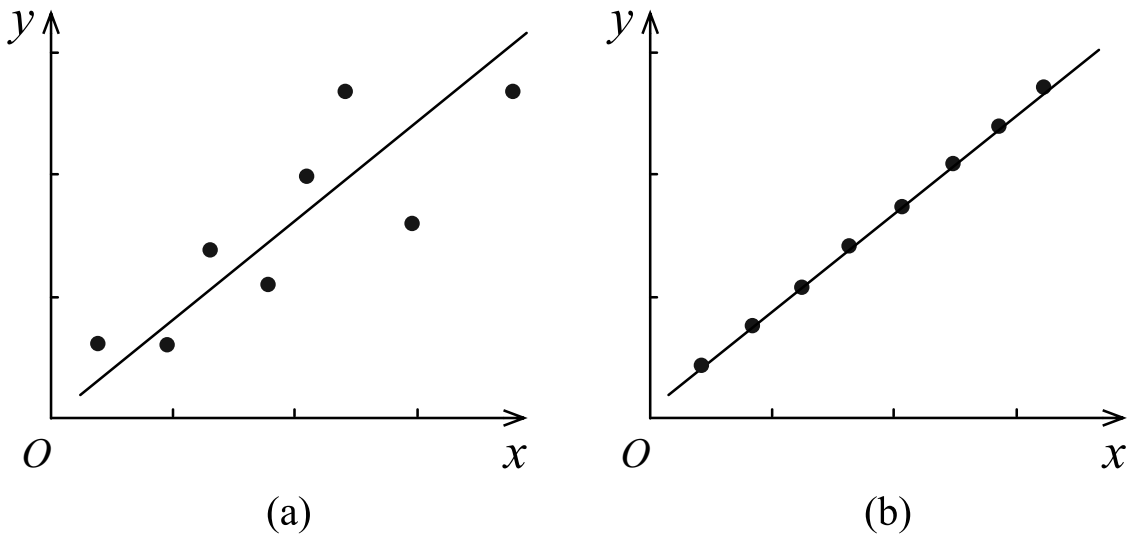


图 **10.7** 两种不同的线性回归效果

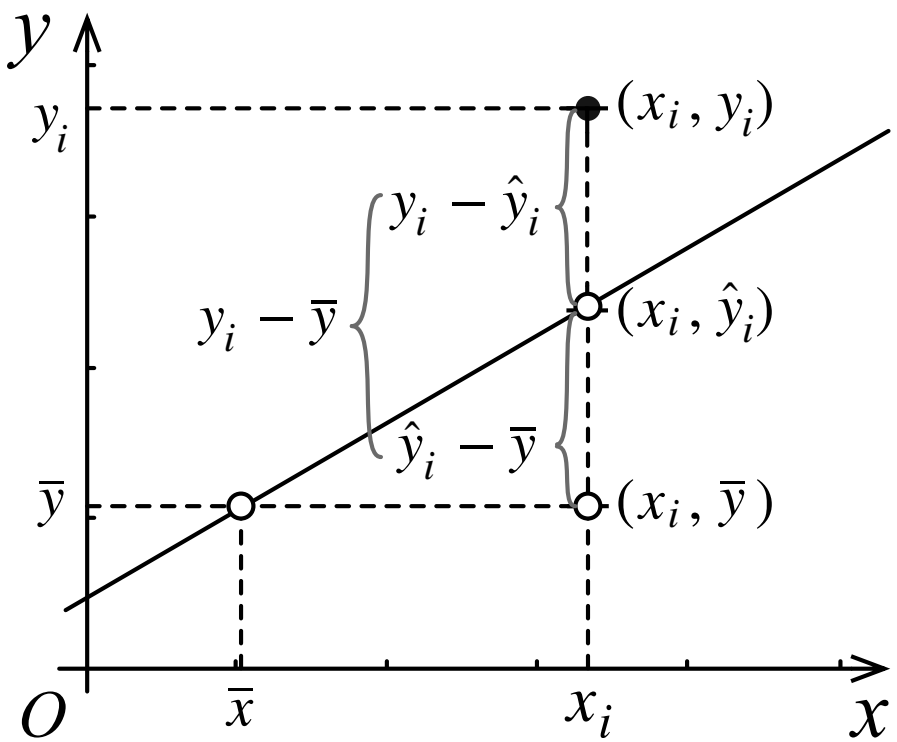


图 **10.8** 因变量离均差 $(y - \bar{y})$ 的分解

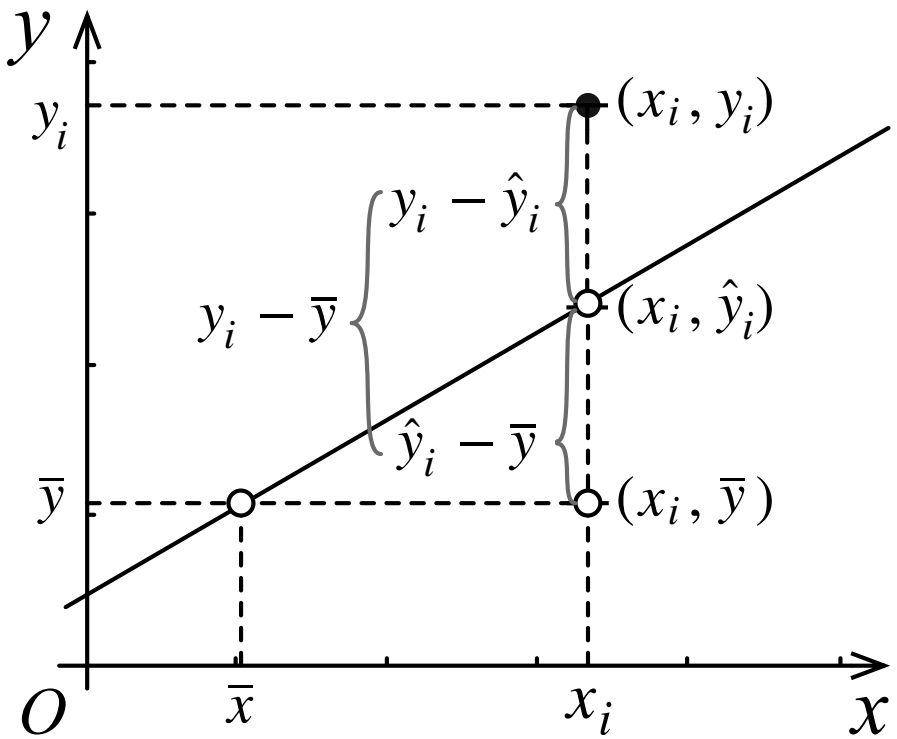


图 **10.8** 因变量离均差 $(y - \bar{y})$ 的分解

回归效果可通过 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 反映，该平方和越小，回归效果越好。

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10.24}$$

SS_y 可分解为以下两部分：

- $\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ 是由 x 的变异引起的 y 变异的平方和，称为回归平方和 (regression sum of square)，记作 SS_r。
- $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 是误差引起的变异的平方和，称为离回归平方和或残差平方和 (residual sum of square)，记作 SS_e。

$$SS_y = SS_r + SS_e \tag{10.25}$$

$$SS_y = SS_r + SS_e \tag{10.25}$$

- 总自由度：$n-1$；
- 残差自由度：$n-1-1=n-2$；
- 回归方差自由度：$n-1-(n-2)=1$。

按照 F 统计量的定义，s_r^2 除以回归的总体方差 σ_r^2 得 F 统计量的分子部分，s_e^2 除以残差的总体方差 σ_e^2 得 F 统计量的分母部分，即

$$F = \frac{\frac{s_r^2}{\sigma_r^2}}{\frac{s_e^2}{\sigma_e^2}} = \frac{s_r^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_r^2} \tag{10.27}$$

服从自由度为 1 和 $n-2$ 的 F 分布。

按照 F 统计量的定义，s_r^2 除以回归的总体方差 σ_r^2 得 F 统计量的分子部分，s_e^2 除以残差的总体方差 σ_e^2 得 F 统计量的分母部分，即

$$F = \frac{\frac{s_r^2}{\sigma_r^2}}{\frac{s_e^2}{\sigma_e^2}} = \frac{s_r^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_r^2} \tag{10.27}$$

服从自由度为 1 和 $n-2$ 的 F 分布。

$$F_c = \frac{s_r^2}{s_e^2} = \frac{(n-2)SS_r}{SS_e} \tag{10.28}$$

回归系数的 t 检验和回归方程的 F 检验可能得到不同的结果吗?

回归系数的 t 检验和回归方程的 F 检验可能得到不同的结果吗?

$$t^2 = \left(\frac{b}{\frac{s_e}{\sqrt{SS_x}}} \right)^2 = \frac{b^2 SS_x}{s_e^2} = \frac{SS_r}{\frac{SS_e}{n-2}} = F \tag{10.29}$$

建立回归方程的过程称为拟合(fitting)。

统计上，我们可以用决定系数 (coefficient of determination) 来定量拟合度，记作

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SS_r}{SS_y}, \quad 0 \le R^2 \le 1 \tag{10.45}$$

建立回归方程的过程称为拟合(fitting)。

统计上，我们可以用决定系数 (coefficient of determination) 来定量拟合度，记作

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SS_r}{SS_y}, \quad 0 \le R^2 \le 1 \tag{10.45}$$

离回归平方和也可以评价回归关系。有公式

$$s_e = \sqrt{\frac{SS_e}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{10.46}$$

又称离回归标准误，表示回归估计值 $\hat{y}$ 与实际观测值 y 的偏差程度。

例 (10.1)

土壤氮含量对植物的生长有重要影响，适当增加氮含量可以促进植物生长。为探讨土壤氮含量（g/kg）与某种牧草干重（100 $\mathrm{g/m^2}$）的关系，研究人员选择了 7 个不同氮含量处理水平的试验田，这些试验田的其他条件基本一致，等量播撒牧草种子半年后测定各试验田中牧草植株干重（数据见 `nitrogenGrass` 数据集），试做回归分析。

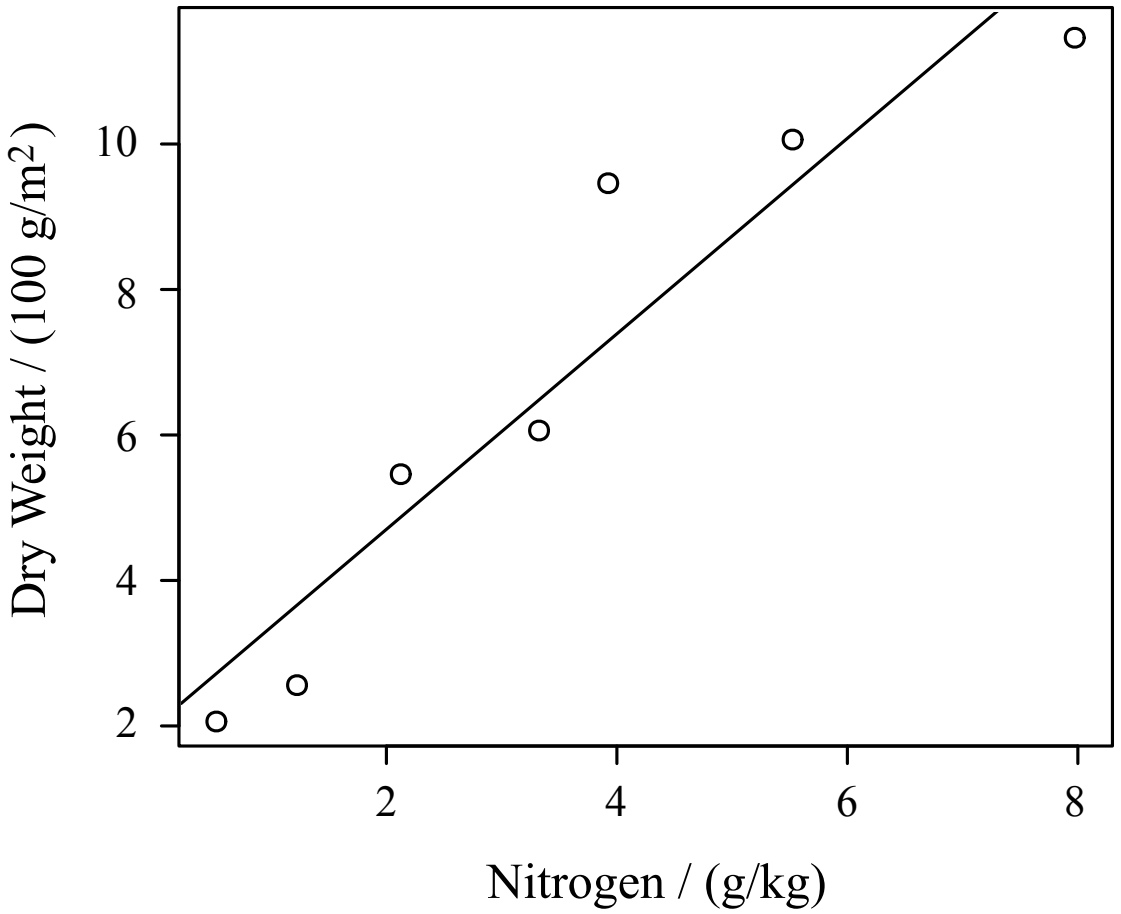


图 **10.10** 土壤氮含量与牧草干重的回归分析

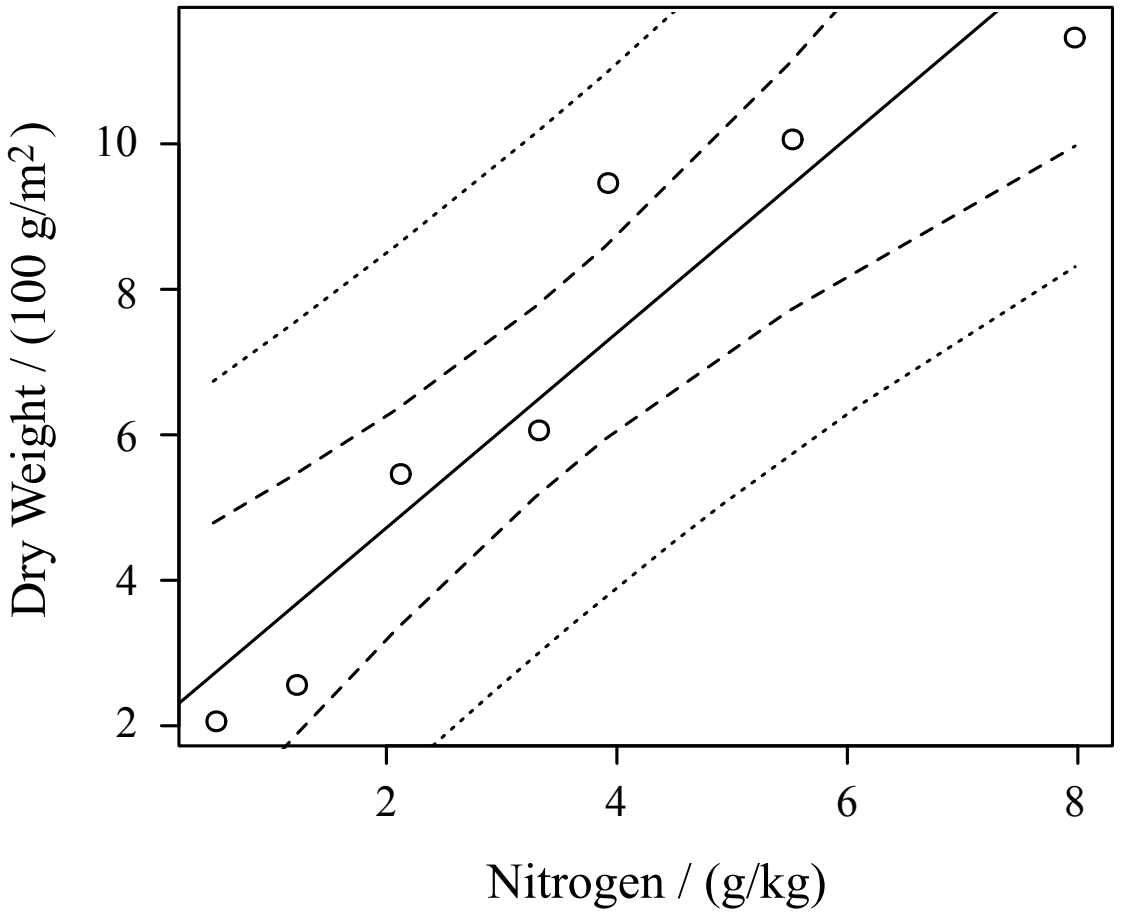


图 **10.11** 土壤氮含量与牧草干重的回归分析和 95% 置信带

例 (10.2)

R 自带数据包 **datasets** 中 **anscombe** 数据集记录了统计学家 Francis Anscombe 在 1973 年发表的一篇文章[a]所用的 4 组数据。每组数据包含 11 个数据点，试对每组数据分别做回归分析。

[a] Anscombe F. 1973. Graphs in Statistical Analysis. The American Statistician, 27(1):17–21.

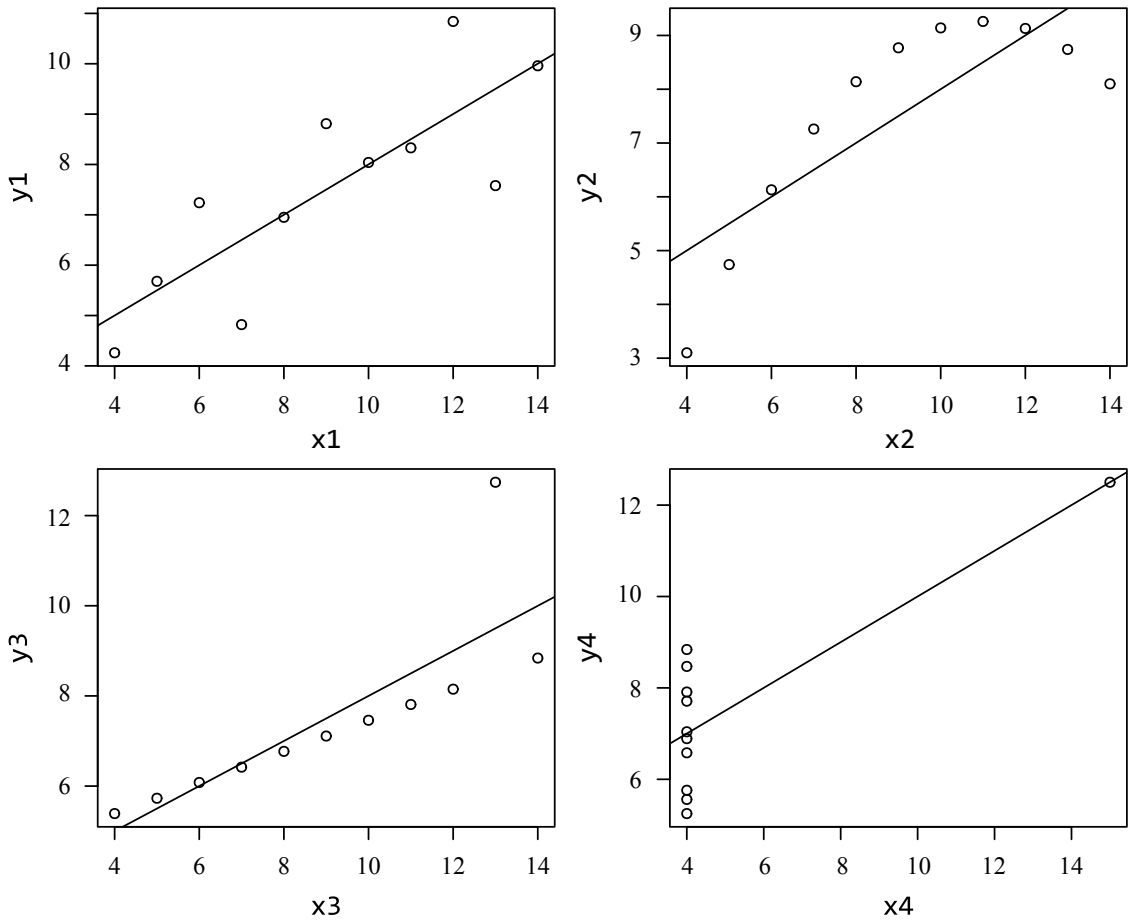


图 **10.12** Anscombe 数据的回归分析

	Estimate	Std. Error	t value	Pr(>|t|)
x1	0.5000909	0.1179055	4.241455	0.002169629
x2	0.5000000	0.1179637	4.238590	0.002178816
x3	0.4997273	0.1178777	4.239372	0.002176305
x4	0.4999091	0.1178189	4.243028	0.002164602

1. 回归分析要有实际意义
2. 回归变量的确定
3. 观测值要尽可能的多
4. 回归效果须检验
5. 预测和外推要谨慎

10.5 回归分析与方差分析的关系

以单因素方差分析为例，方差分析的数学模型为

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{10.51}$$

回归分析的数学模型为

$$y = \alpha + \beta x + \varepsilon \tag{10.52}$$

它们都是线性模型，所以 R 使用同样的函数 `lm()` 来处理它们。

10.5 回归分析与方差分析的关系

方差分析中的总离均差平方和可分解为

$$SS = SS_t + SS_e \tag{10.53}$$

回归分析中 y 的总离均差平方和可分解为

$$SS_y = SS_r + SS_e \tag{10.54}$$

它们都可用F 检验来完成显著性检验。

10.5 回归分析与方差分析的关系

回归分析与方差分析的不同之处包括：

- 分析目的不同

  回归分析侧重于两变量间的数量关系，而方差分析侧重于试验因素内部各水平的差异。

- 数据类型不同

  方差分析中处理因素是离散的分类变量；回归分析中，处理因素（即自变量 x）是连续型变量。

本章小结

1. “回归”的故事
2. 回归与相关的基本概念
3. 线性回归分析
   - 回归的数学模型与基本假定
   - 回归方程及其性质
   - 回归的显著性检验
   - 回归的区间估计
   - 回归方程的评价
   - 回归分析实例
   - 回归分析应注意的问题
4. 非线性回归分析
5. 回归分析与方差分析的关系