回归分析

职晓阳 xyzhi@ynu.edu.cn

目录

0.1 基本概念

1822 年 2 月 16 日，F. 高尔顿出生于英国伯明翰。父亲塞缪尔·德丢·高尔顿是位银行家。外祖父伊拉斯谟斯·达尔文是一位诗人、医生、进化论理论家。1859 年高尔顿的表兄 C.R. 达尔文出版的《物种起源》引起了他对人类遗传的兴趣。他的科学兴趣很快转移到与生命有关的领域。他把达尔文关于围绕着群的平均值的偶发变异原理应用于人类研究，开拓了以个体差异为主题的实验心理学的新领域，并于 1869 年发表了专著《遗传的天才》。1883 年他出版了专著《人类才能及其发展的研究》，书中概述了自由联想和关于心理意象的问卷调查两项实验心理学上划时代的研究方法和成果。在这本书中，高尔顿还第一次提出了一个以人类的自觉选择来代替自然选择的社会计划，为此他还创造了“优生学”（Eugenics）这个词。从 1884 年起，高尔顿先是在国际卫生博览会、后是在南肯新顿博物馆开设了一个人类学测量室。在那里人们可以测量出自己的身高、体重、握力等多种生理指标。在六年时间里，该实验室共收集了 9337 位男女的详细资料，为人类个体差异研究提供了大量数据。1888 年，他开始对指纹研究产生兴趣，并于 1892 年在

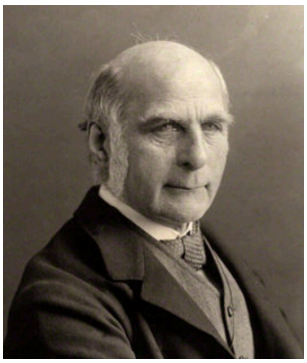

图 1: 弗朗西斯·高尔顿 (Francis Galton 1822-1911)

其专著《指纹学》中提出了指纹分类法。他的指纹编码法被苏格兰场使用采用，作为伯特隆测量系统的补充手段来建立犯人档案。

1886 年左右高尔顿观察了 1074 对父母其其一成年儿子的身高，以父母的平均身高 X 作自变量，孩子的身高 Y 作因变量，并将 (X,Y) 值标记在直角坐标系上，发现二者的关系近乎一条直线。总的趋势是 Y 随 X 的增加而增加，这并不让人觉得意外。然而随着高尔顿对所得数据进行深入分析，发现了某种有趣的现象。高尔顿计算出 1074 个 X 的算数均值为 68 英寸，而 Y 的均值为 69 英寸，子代升高正体上增加了 1 英寸。但是当高尔顿分段来观察数据后发现，当父母身高均值为 72 英寸时，子代均值只有 71 英寸，而当父母身高均值为 64 英寸时，子代均值只有 67 英寸。

他解释道：大自然有一种约束机制，使使人类身高分布保持某种稳定形态而不作两极分化。高尔顿引入 “回归” 一词来描述这一现象。

“回归” 虽源于此，但作为变量关系统计分析之 “回归”，其意义却与高尔顿的 “回归” 有明显的不同。

回归分析着重寻求变量之间近似的函数关系。

实际问题中存在着大量相互之间具有某种联系的变量，如身高与体重、降水量与作物产量等。当考量两个随机变量关系时，常冠以自变量 X 与因变量 Y 之名，虽然它们并无明显的因果关系。同时，现实的复杂性决定了，自变量并不能严格决定因变量。因为除了抽象为自变量的因素之外，还存在对自变量同样有影响，而不能对其进行分析研究的因素。

设在一个问题中有因变量 Y，及自变量 $X_1,\cdots,X_n$。理论上 Y 的值由两部分构成：一部分来自 $X_1,\cdots,X_n$ 的影响，这一部分常表示为自变量的

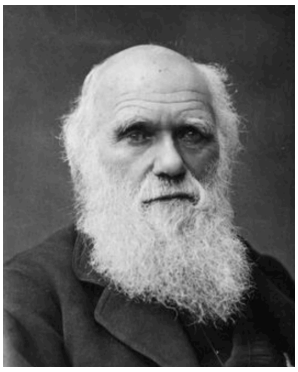

图 2: 查尔斯·罗伯特·达尔文 (Charles Robert Darwin 1809-1882)

函数 $f(X_1, \cdots, X_n)$；另一部分则有其它众多未及考虑的因素，包括随机因素的影响，常记为 ε。于是有模型

$$Y = f(X_1, \cdots, X_n) + \varepsilon$$

同时，要求 $E(\varepsilon) = 0$。模型中函数 $f(X_1, \cdots, X_n)$ 称为 Y 对 $X_1, \cdots, X_n$ 的*理论回归函数*，方程 $y = f(x_1, \cdots, x_n)$ 称为*理论回归方程*。

在实际问题中，理论回归函数一般总是未知的，回归分析的任务就在于根据 $X_1, \cdots, X_n$ 和 Y 的观测值，去估计这个函数，以及讨论与此有关的种种统计推断问题，包括假设检验和区间估计。通过数据对回归函数或方程中的参数进行估计，得到的称为*经验回归函数*和*经验回归方程*。回归分析的方法基本上取决于模型中的假定，即对回归函数 f 和随机误差 ε 的假定。

对于回归函数 f，一种情况是对 f 的数学形式无特殊的假定，称为*非参数回归*。另一种较为常见的情况，是假定 f 的数学形式是已知的，只是其中若干参数未知，称为*参数回归*。在诸多关于 f 的数学形式中，无论应用上的重要程度还是理论上发展的完善程度，当属*线性函数*的最高。

$$f(x_1, \cdots, x_n) = b_0 + b_1 x_1 + \cdots + b_n x_n$$

上式即为典型的 “线性回归” 方程。与线性回归函数相对应的是一大类统称为*非线性回归函数*。

对于随机误差 ε，除假定 $E(\varepsilon) = 0$，ε 的方差 σ^2 是回归模型中的一个重要参数，因为 $\sigma^2 = E(\varepsilon^2) = E[Y - f(X_1, \cdots, X_n)]^2$，$\sigma^2$ 越小，说明用 $f(X_1, \cdots, X_n)$ 逼近 Y 而产生的均方误差越小，回归的效果也就越好。随机

误差 ε 的方差 σ^2 取决于两方面因素：1. 对 Y 有重要影响的因素是否都通过自变量表达出来了。2. 回归函数的形式是否准确。在实际的回归分析中，常假定 ε 服从正态分布。如果没有这个假定，就需要使用大样本方法。

回归分析的实际应用价值是显而易见的。从应用的角度，其价值体现在以下四个方面：

- 对数据进行描述性地总结；
- 估计回归函数；
- 对数据的趋势进行预测；
- 对数据进行控制。

0.2 一元线性回归

回归函数为线性函数的情形，包括能转化为线性函数的，称为线性回归。

线性回归中只包含一个自变量 X 的情形，称为一元线性回归。其模型表达式为

$$Y = b_0 + b_1 X + \varepsilon$$

其中 b_0, b_1 为未知参数，b_0 称为常数项或截距，b_1 称为回归系数，ε 为随机误差，且 $E(\varepsilon) = 0, \quad 0 < Var(\varepsilon) = \sigma^2 < \infty$，误差方差 σ^2 未知。对这一对变量关系的追踪观察，得样本 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$，且根据假定的线性模型关系有方程

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \quad (i = 1, \cdots, n)$$

这里 ε_i 是第 i 此观察时随机误差的值，它是不能观察的 (如能观察，就可以归纳为一自变量了)。由于每次观察是独立的，所以随机误差 $\varepsilon_1, \cdots, \varepsilon_n$ 也是独立同分布的，且均值方差符合上面所提到的条件。上述方程联合随机误差的分布要求，称为*一元线性回归模型*。

自变量与因变量的算数均值分别记作 $\overline{X}$ 和 $\overline{Y}$，借此改写回归模型为

$$Y_i = \beta_0 + \beta_1 (X_i - \overline{X}) + \varepsilon_i \quad (i = 1, \cdots, n)$$

与原形的关系是

$$\beta_1 = b_1, \quad \beta_0 = b_0 + b_1 \overline{X}$$

这种改写又称为“中心化”，因为因子 $X_i - \overline{X}, (i = 1, \cdots, n)$ 求和后为 0。

0.2.1 参数估计

0.2.1.1 最小二乘法

如果用 $\hat{\beta}_0, \hat{\beta}_1$ 去分别估计参数 β_0, β_1。则回归函数的因变量 Y_i 则可以用 $\hat{\beta}_0 + \hat{\beta}_1(X_i - \overline{X})$ 来估计，记作 $\hat{Y}_i$。那么理论估计值 $\hat{Y}_i$ 与实际观测值 Y_i 之间就有了偏离值 $Y_i - \hat{Y}_i$。衡量偏离大小的一个合理的单一指标为它们的平方和，这和方差有类似的考虑，即通过平方去掉符号的影响。因此有

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - \overline{X})]^2$$

作为衡量偏离大小的表达式。对参数进行估计的目标是使偏离大小达到最小，所以利用多元函数求极值的方法，就需要对方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}[Y_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - \overline{X})] = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(X_i - \overline{X})[Y_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - \overline{X})] = 0 \end{cases}$$

求解得：

$$\hat{\beta}_0 = \overline{Y}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

是偏差最小的估计方法，称为最小二乘法。这个重要的方法一般归功于德国大数学家高斯在 1799-1809 年间的工作。最小二乘估计有如下良好的性质：

- $\hat{\beta}_0, \hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计

由 $Y_i = \beta_0 + \beta_1(X_i - \overline{X}) + \varepsilon_i$ 知 $E(Y_i) = \beta_0 + \beta_1(X_i - \overline{X})$

$$
\begin{aligned}
E(\hat{\beta}_0) &= E(\overline{Y}) \\
&= E(E(Y_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}[\beta_0 + \beta_1(X_i - \overline{X})] \\
&= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1(X_i - \overline{X}) \\
&= \beta_0 \\
E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})E(Y_i)}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})[\beta_0 + \beta_1(X_i - \overline{X})]}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^{n}[(X_i - \overline{X})\beta_0 + \beta_1(X_i - \overline{X})^2]}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})\beta_0 + \sum_{i=1}^{n}\beta_1(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \beta_1
\end{aligned}
$$

- $\hat{\beta}_0, \hat{\beta}_1$ 的方差分别为

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var(\overline{Y}) \\
&= Var(\frac{\sum_{i=1}^{n} Y_i}{n}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n} \\
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})Y_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right) \\
&= Var\left(\sum_{i=1}^{n}\frac{(X_i-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}Y_i\right) \\
&= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^2 Var(Y_i) \\
&= \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{[\sum_{i=1}^{n}(X_i-\overline{X})^2]^2} Var(Y_i) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}
\end{aligned}
$$

- $\hat{\beta}_0, \hat{\beta}_1$ 是 Y 的线性函数
- $\hat{\beta}_0, \hat{\beta}_1$ 的的协方差为 0

$$
\begin{aligned}
Cov(\hat{\beta}_0,\hat{\beta}_1) &= E(\hat{\beta}_0 - E\hat{\beta}_0)(\hat{\beta}_1 - E\hat{\beta}_1) \\
&= E\left(\frac{\sum_{j=1}^n Y_j}{n} - E(Y_j)\right)\left(\frac{\sum_{i=1}^n (X_i-\overline{X})Y_i}{\sum_{i=1}^n (X_i-\overline{X})^2} - \frac{\sum_{i=1}^n (X_i-\overline{X})E(Y_i)}{\sum_{i=1}^n (X_i-\overline{X})^2}\right) \\
&= E\left(\frac{\sum_{j=1}^n Y_j - nE(Y_j)}{n}\right)\left(\frac{\sum_{i=1}^n (X_i-\overline{X})[Y_i - E(Y_i)]}{\sum_{i=1}^n (X_i-\overline{X})^2}\right) \\
&= E\left(\frac{\sum_{j=1}^n [Y_j - E(Y_j)]}{n}\right)\left(\frac{\sum_{i=1}^n (X_i-\overline{X})[Y_i - \beta_0 + \beta_1(X_i-\overline{X})]}{\sum_{j=1}^n (X_j-\overline{X})^2}\right) \\
&= E\left(\frac{\sum_{j=1}^n [Y_j - \beta_0 + \beta_1(X_j-\overline{X})]}{n}\right)\left(\frac{\sum_{i=1}^n (X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^n (X_i-\overline{X})^2}\right) \\
&= E\left(\frac{\sum_{j=1}^n \varepsilon_j}{n}\right)\left(\frac{\sum_{i=1}^n (X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^n (X_i-\overline{X})^2}\right) \\
&= E\left(\frac{\sum_{j=1}^n \varepsilon_j}{n}\cdot\frac{\sum_{i=1}^n (X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^n (X_i-\overline{X})^2}\right) \\
&= \frac{1}{n\sum_{i=1}^n (X_i-\overline{X})^2} E\left(\sum_{j=1}^n \varepsilon_j \sum_{i=1}^n (X_i-\overline{X})\varepsilon_i\right) \\
&= \frac{1}{n\sum_{i=1}^n (X_i-\overline{X})^2} E\left(\underbrace{\cdots + (X_i-\overline{X})\varepsilon_i\varepsilon_j + \cdots}_{i\neq j} + \underbrace{(X_i-\overline{X})\varepsilon_i^2}_{i=j}\right) \\
&= \frac{1}{n\sum_{i=1}^n (X_i-\overline{X})^2} E\left(\underbrace{\cdots + (X_i-\overline{X})\varepsilon_i\varepsilon_j + \cdots}_{i\neq j} + \sum_{i=1}^n (X_i-\overline{X})\varepsilon_i^2\right) \\
&= \frac{1}{n\sum_{i=1}^n (X_i-\overline{X})^2} \sum_{i=1}^n (X_i-\overline{X})E(\varepsilon_i^2) \\
&= 0
\end{aligned}
$$

$\hat{\beta}_0,\hat{\beta}_1$ 不相关，凸显了“中心化”的好处。此外不相关并不意味着它们独立。但是如果 ε 服从正态分布，则 Y 也服从正态分布，$\hat{\beta}_0,\hat{\beta}_1$ 作为 Y 的线性函数，也服从正态分布。因此在这种情况下可推出它们独立。

将 $\hat{\beta}_0,\hat{\beta}_1$ 代入愿参数的变换函数，可得 $\hat{b}_0 = \hat{\beta}_0 - \hat{b}_1\overline{X} = \overline{Y} - \hat{b}_1\overline{X}$，$\hat{b}_1 = \hat{\beta}_1$

一元线性回归模型中除了 β_0,β_1 还有一个重要的参数，即随机误差的方差 σ^2。对 σ^2 的估计，需要借助一个新的概念——*残差*。因变量 Y 观测值与模型预测值 $\hat{Y}$ 之差

$$\delta_i = Y_i - \hat{Y}_i (i = 1,\cdots,n)$$

称为残差。当模型正确时，残差反映的就是预测的精度，与随机误差的大小，即误差方差的大小有关。因此残差可以提供一个 σ^2 的估计。

$$
\begin{aligned}
\delta_i &= Y_i - \hat{Y}_i \\
&= \beta_0 + \beta_1(X_i - \overline{X}) + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - \overline{X}) \\
&= \beta_0 - \hat{\beta}_0 + (X_i - \overline{X})(\beta_1 - \hat{\beta}_1) + \varepsilon_i \\
&= \beta_0 - \overline{Y} + (X_i - \overline{X})(\beta_1 - \hat{\beta}_1) + \varepsilon_i \\
&= \beta_0 - \frac{1}{n}\sum_{i=1}^{n}[\beta_0 + \beta_1(X_i - \overline{X}) + \varepsilon_i] + (X_i - \overline{X})(\beta_1 - \hat{\beta}_1) + \varepsilon_i \\
&= \beta_0 - \frac{1}{n}\sum_{i=1}^{n}\beta_0 - \frac{1}{n}\sum_{i=1}^{n}\beta_1(X_i - \overline{X}) - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + (X_i - \overline{X})(\beta_1 - \hat{\beta}_1) + \varepsilon_i \\
&= -\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + (X_i - \overline{X})\left(\beta_1 - \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) + \varepsilon_i \\
&= -\overline{\varepsilon} + (X_i - \overline{X})\left(\beta_1 - \frac{\sum_{i=1}^{n}(X_i - \overline{X})[\beta_0 + \beta_1(X_i - \overline{X}) + \varepsilon_i]}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) + \varepsilon_i \\
&= -\overline{\varepsilon} + (X_i - \overline{X})\left(\beta_1 - \frac{\sum_{i=1}^{n}[(X_i - \overline{X})\beta_0 + \beta_1(X_i - \overline{X})^2 + (X_i - \overline{X})\varepsilon_i]}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) + \varepsilon_i \\
&= -\overline{\varepsilon} + (X_i - \overline{X})\left(\beta_1 - \frac{\sum_{i=1}^{n}(X_i - \overline{X})\beta_0 + \sum_{i=1}^{n}\beta_1(X_i - \overline{X})^2 + \sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) + \varepsilon_i \\
&= -\overline{\varepsilon} + (X_i - \overline{X})(-\frac{\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}) + \varepsilon_i
\end{aligned}
$$

等式取平方，得

$$
\delta_i^2 = \left[\varepsilon_i - \overline{\varepsilon} - (X_i - \overline{X})\frac{\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^2
$$

$$
\delta_i^2 = (\varepsilon_i - \overline{\varepsilon})^2 + \left[(X_i - \overline{X})\frac{\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^2 - 2(\varepsilon_i - \overline{\varepsilon})(X_i - \overline{X})\frac{\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}
$$

对 $i=1,\cdots,n$ 求和, 得

$$\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2+\sum_{i=1}^{n}\left[(X_i-\overline{X})\frac{\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]^2$$
$$-2\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})(X_i-\overline{X})\frac{\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$
$$\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2+\sum_{i=1}^{n}\left[(X_i-\overline{X})\frac{\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]^2$$
$$-2[\sum_{i=1}^{n}\varepsilon_i(X_i-\overline{X})-\sum_{i=1}^{n}\overline{\varepsilon}(X_i-\overline{X})]\frac{\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$
$$\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2+\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{[\sum_{i=1}^{n}(X_i-\overline{X})^2]^2}$$
$$-2\sum_{i=1}^{n}\varepsilon_i(X_i-\overline{X})\frac{\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$
$$\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2+\frac{[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}-2\frac{[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$
$$\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2-\frac{[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

求期望，得

$$E(\sum_{i=1}^{n}\delta_i^2)=E\Big(\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2\Big)-E\Big(\frac{[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\Big)$$

$$E(\sum_{i=1}^{n}\delta_i^2)=E\Big((n-1)\frac{\sum_{i=1}^{n}(\varepsilon_i-\overline{\varepsilon})^2}{n-1}\Big)-E\Big(\frac{[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\Big)$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-1)\sigma^2-\frac{E\Big([\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]^2\Big)}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-1)\sigma^2-\frac{Var[\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i]+[E(\sum_{i=1}^{n}(X_i-\overline{X})\varepsilon_i)]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-1)\sigma^2-\frac{\sum_{i=1}^{n}(X_i-\overline{X})Var(\varepsilon_i)}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-1)\sigma^2-\frac{\sum_{i=1}^{n}(X_i-\overline{X})\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-1)\sigma^2-\sigma^2$$

$$E(\sum_{i=1}^{n}\delta_i^2)=(n-2)\sigma^2$$

所以 $\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\delta^2}{n-2}$ 是 σ^2 得无偏估计。此处，自由度为 $n-2$，比样本量小 2。这是因为有两个自由度用于未知参数 β_0,β_1 得估计了。

$\sum_{i=1}^{n}\delta^2$ 称为残差平方和，它有一个重要得性质，即当 ε 服从正态分布 $N(0,\sigma^2)$ 时，有 $\frac{\sum_{i=1}^{n}\delta^2}{\sigma^2}$ 服从自由度为 $n-2$ 的 χ^2 分布。

0.2.1.2 最大似然法

0.2.2 区间估计

0.2.2.1 回归系数的区间估计

将一元线性回归模型中关于随机误差 ε 的分布函数做进一步的假定，即 ε 服从正态分布 $N(0,\sigma^2)$。

对于 $\hat{\beta}_1$，已知它是 Y 的线性函数，有均值 β_1 和方差 $\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$，因此有

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}} \sim N(0,1)$$

但是由于 σ^2 未知，用估计 $\hat{\sigma}^2$ 代替，则正态分布将变为 t 分布

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}} \sim t_{n-2}$$

这个结果就可以用来估计 β_1 的置信区间，即

$$[\hat{\beta}_1 - \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} t_{n-2}(\frac{\alpha}{2}), \quad \hat{\beta}_1 + \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} t_{n-2}(\frac{\alpha}{2})]$$

类似地，对 β_0 得置信区间

$$[\hat{\beta}_0 - \frac{\hat{\sigma}}{\sqrt{n}} t_{n-2}(\frac{\alpha}{2}), \quad \hat{\beta}_0 + \frac{\hat{\sigma}}{\sqrt{n}} t_{n-2}(\frac{\alpha}{2})]$$

0.2.2.2 回归函数的区间估计

0.2.3 显著性检验

0.2.3.1 回归系数的 t 检验

针对回归系数 β_1，常见的检验问题是

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

接受零假设，表明回归函数为一常数，即 β_0，与 x 无关。也就是说所选定的自变量 X 对因变量 Y 无影响。对于该问题，有检验函数

$$\varphi : 当|\hat{\beta}_1 - \beta_1| \leq C时接受H_0, 当|\hat{\beta}_1 - \beta_1| > C时否定H_0$$

其功效函数为

$$\beta_\varphi(\beta_1) = P_{\beta_1}(|\hat{\beta}_1 - \beta_1| > C)$$

当用水平 α 来约束时,

$$\beta_\varphi(\beta_1) = P_{\beta_1}(|\hat{\beta}_1 - \beta_1| > C) \leq \alpha$$

相当于

$$P_\theta(|\hat{\beta}_1-\beta_1|\le C)\ge 1-\alpha$$

$$P_\theta(|\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n(X_i-\overline{X})^2}}}|\le\frac{C}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n(X_i-\overline{X})^2}}})\ge 1-\alpha$$

$$\mathrm{T}(\frac{\mathrm{C}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2}}})-\mathrm{T}(-\frac{\mathrm{C}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2}}})\ge 1-\alpha$$

$$2\mathrm{T}(\frac{\mathrm{C}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2}}})-1\ge 1-\alpha$$

$$\mathrm{T}(\frac{\mathrm{C}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2}}})\ge 1-\frac{\alpha}{2}$$

取等号时，得

$$C=\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n(X_i-\overline{X})^2}}\cdot t_{n-2}(\frac{\alpha}{2})$$

所以

$$\varphi:\text{当}|\hat{\beta}_1|\le\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n(X_i-\overline{X})^2}}\cdot t_{n-2}(\frac{\alpha}{2})\text{时接受}H_0,\text{否则否定}H_0$$

0.2.3.2 回归方程的 F 检验

0.2.3.3 相关系数的 t 检验

0.3 多元线性回归

线性回归中包含 p 个自变量 $X_1,\cdots,X_p$ 的情形，称为*多元线性回归*。因变量仍记作 Y，则模型的表达式为

$$Y=b_0+b_1X_1,+\cdots+b_pX_p+\varepsilon$$

其中 b_0 仍称为常数项或截距，b_k 称为 Y 对 X_k 的回归系数，或*偏回归系数*，ε 仍为随机误差。

现对 $X_1,\cdots,X_p$ 和 Y 进行试验观察，第 i 次观察所得数据分别记为 $X_{1i},\cdots,X_{pi}$ 和 Y_i，随机误差记为 ε_i，则得方程

$$Y_i=b_0+b_1X_{1i},+\cdots+b_pX_{pi}+\varepsilon_i\quad(i=1,\cdots,n).$$

这里假定 $\varepsilon_1, \cdots, \varepsilon_i$ 独立同分布，且 $E(\varepsilon_i) = 0, 0 < Var(\varepsilon_i) = \sigma^2 < \infty$，方差 σ^2 未知。

与一元线性回归模型一样，需要根据所得数据 $X_{1i}, \cdots, X_{pi}$ 和 Y_i，对参数 $b_0, \cdots, b_p$ 和误差方差 σ^2 进行估计，对因变量 Y 进行预测，以及有关的假设检验问题。方法与概念上多元回归模型与一元回归模型无甚差别，但在计算和理论方面，多元情形都要比一元情形复杂的多。

类似地，为方便计需要对每个自变量进行 “中心化”。第 k 个自变量 X_k 在 n 次试验中观察所得算数均值 $\overline{X_k} = (X_{k1} + \cdots + X_{kn})/n$，所以有

$$X_{ki}^* = X_{ki} - \overline{X_k} \quad (i = 1, \cdots, n;\ k = 1, \cdots, p)$$

则多元回归方程可改写为

$$Y_i = \beta_0 + \beta_1 X_{1i}^*, + \cdots + \beta_p X_{pi}^* + \varepsilon_i \quad (i = 1, \cdots, n).$$

其中

$$\beta_k = b_k\ (k = 1, \cdots, p); \quad \beta_0 = b_0 + b_1\overline{X_1} + \cdots + b_p\overline{X_p}$$

数学上处理多元问题时，采用高等代数中的矩阵和向量的记号最为方便。

0.3.1 参数估计

与一元的情形一样，参数 $\beta_0, \cdots, \beta_p$ 的估计分别计作 $\hat{\beta}_0, \cdots, \hat{\beta}_p$，令

$$Q(\hat{\beta}_0, \cdots, \hat{\beta}_p) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*]^2.$$

要确定当上式达到最小时的 $\hat{\beta}_0, \cdots, \hat{\beta}_p$ 值。需要令

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = -2\sum\limits_{i=1}^{n}\dfrac{\partial \hat{\beta}_0}{\partial \hat{\beta}_0}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}^*, - \cdots - \hat{\beta}_p X_{pi}^*] = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = -2\sum\limits_{i=1}^{n}\dfrac{\partial (\hat{\beta}_1 X_{1i}^*)}{\partial \hat{\beta}_0}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}^*, - \cdots - \hat{\beta}_p X_{pi}^*] = 0 \\ \cdots \\ \dfrac{\partial Q}{\partial \hat{\beta}_p} = -2\sum\limits_{i=1}^{n}\dfrac{\partial (\hat{\beta}_p X_{pi}^*)}{\partial \hat{\beta}_0}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}', - \cdots - \hat{\beta}_p X_{pi}^*] = 0 \end{cases}$$

整理后，首先得

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$

然后有

$$\begin{cases} \sum_{i=1}^{n} X_{1i}^*[(Y_i - \overline{Y}) - (\hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*)] = 0 \\ \sum_{i=1}^{n} X_{2i}^*[(Y_i - \overline{Y}) - (\hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*)] = 0 \\ \cdots \\ \sum_{i=1}^{n} X_{pi}^*[(Y_i - \overline{Y}) - (\hat{\beta}_1 X_{1i}', + \cdots + \hat{\beta}_p X_{pi}^*)] = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} X_{1i}^*(\hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*) = \sum_{i=1}^{n} X_{1i}^*(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_{1i} - \overline{X_1})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_{1i}^* Y_i \\ \sum_{i=1}^{n} X_{2i}^*(\hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*) = \sum_{i=1}^{n} X_{2i}^*(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_{2i} - \overline{X_2})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_{2i}^* Y_i \\ \cdots \\ \sum_{i=1}^{n} X_{pi}^*(\hat{\beta}_1 X_{1i}', + \cdots + \hat{\beta}_p X_{pi}^*) = \sum_{i=1}^{n} X_{pi}^*(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_{pi} - \overline{X_p})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_{pi}^* Y_i \end{cases}$$

下面即可引入矩阵和向量

$$\mathbf{X} = \begin{Bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{Bmatrix}, \quad \mathbf{Y} = \begin{Bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{Bmatrix}$$

$$\mathbf{L} = \begin{Bmatrix} \sum_{i=1}^{n} X_{1i}X_{1i} & \sum_{i=1}^{n} X_{1i}X_{2i} & \cdots & \sum_{i=1}^{n} X_{1i}X_{pi} \\ \sum_{i=1}^{n} X_{2i}X_{1i} & \sum_{i=1}^{n} X_{2i}X_{2i} & \cdots & \sum_{i=1}^{n} X_{2i}X_{pi} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ni}X_{1i} & \sum_{i=1}^{n} X_{ni}X_{2i} & \cdots & \sum_{i=1}^{n} X_{ni}X_{pi} \end{Bmatrix}, \quad \hat{\beta} = \begin{Bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{Bmatrix}$$

将方程组变为

$$\mathbf{L} \cdot \hat{\beta} = \mathbf{X} \cdot \mathbf{Y}$$

所以

$$\hat{\beta} = \mathbf{L^{-1}} \cdot \mathbf{X} \cdot \mathbf{Y}$$

其中 $\mathbf{L^{-1}}$ 是 $\mathbf{L}$ 的逆矩阵。$\mathbf{L^{-1}}$ 在回归分析中有很重要的地位。

误差方差 σ^2 的估计，仍如一元的情形，先定义多元线性回归模型的残差

$$\delta_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i}^*, + \cdots + \hat{\beta}_p X_{pi}^*) \quad (i = 1, \cdots, n)$$

可以证明残差平方和，可用于误差方差 σ^2 的无偏估计，即

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} \delta_i^2}{n - (p+1)}$$

当随机误差服从正态分布时，可以证明

$$\sum_{i=1}^{n} \delta_i^2 / \sigma^2$$

服从自由度为 $n - p - 1$ 的 χ^2 分布，因为已经有 $p + 1$ 自由度用于估计参数 $\beta_0, \beta_1, \cdots, \beta_p$。

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - \left(\hat{\beta}_1 \sum_{i=1}^{n} X_{1i} Y_i + \cdots + \hat{\beta}_p \sum_{i=1}^{n} X_{pi} Y_i \right)$$

上式右边括号内的各项，已在计算 $\hat{\beta}$ 时得到。

0.3.2 区间估计

0.3.2.1 回归系数的区间估计

0.3.2.2 回归函数的区间估计

0.3.3 假设检验

0.4 广义线性回归

0.4.1 基本概念

广义线性回归，顾名思义是线性回归的推广。为理解广义线性模型究竟在哪些方面进行了推广，首先需要再审视 (多元) 线性模型的两个重要假设：

1. Y 服从正态，或接近正态分布；
2. $E(Y) = \boldsymbol{\beta}^T \mathbf{X}$。

线性模型要求随机误差 ε 服从正态分布 $N(0, \sigma^2)$，由于 Y 与 $\mathbf{X}$ 存在线性关系,所以 Y 也服从正态分布,或接近正态分布。此外,对 $Y = \beta_0 + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon$ 求期望得 $E(Y)$，即模型预测值的期望。显然这两个重要假设的条件是苛刻的，很多实际问题是达不到如此假设前提的。那么为使线性回归模型具有更广泛的应用，就放松需要这两个假设条件限制：

1. Y 的分布属于指数分布族；

指数分布族实质上是对一类具有以下形式的概率密度函数或具有此类密度函数的分布的总括。

$$f(y, \theta) = h(y) e^{\eta(\theta) T(y) - a(\theta)}$$

其中，θ 是分布的自然参数或典范参数，$T(y)$ 叫做充分统计量，通常情况下 $T(y)=y$；$a(\theta)$ 是对数分配函数，而 a, h, T 一般都是给定的，随着 θ 的变化，会得到不同的分布。下面看两个具体的实例：

- 两点分布

已知 Bernoulli 两点分布 $B(p)$，p 为分布的均值，随着 p 的变化，可以得到不同的伯努利分布。

$$\begin{aligned} f(y, p) &= p^y (1-p)^{1-y} \\ &= e^{y \ln p} \cdot e^{(1-y) \ln (1-p)} \\ &= \exp\left(y \ln p + (1-y) \ln (1-p) \right) \\ &= \exp\left([\ln(\frac{p}{1-p})] y + \ln (1-p) \right) \end{aligned}$$

对应到指数分布族的标准形式，则有

$$\begin{aligned} h(y) &= 1 \\ T(y) &= y \\ \eta(\theta) &= \ln \frac{p}{1-p} \\ a(\theta) &= -\ln (1-p) \\ &= \ln(1 + e^{\eta(\theta)}) \end{aligned}$$

- 正态分布

已知高斯分布 $N(\mu, \sigma^2)$。其概率密度函数如下

$$\begin{aligned} f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(y-\mu)^2}{2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} y^2 \right) \cdot \exp\left(\frac{\mu}{\sigma^2} y - \frac{1}{2\sigma^2}\mu^2 \right) \end{aligned}$$

对应到指数分布族的标准形式，则有

$$\begin{aligned} h(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} y^2 \right) \\ T(y) &= y \\ \eta(\theta) &= \frac{\mu}{\sigma^2} \\ a(\theta) &= \frac{1}{2\sigma^2}\mu^2 \\ &= \frac{1}{2}[\eta(\theta)]^2 \sigma^2 \end{aligned}$$

其他属于指数族分布族的分布还包括：多项式分布，用来对多元分类问题进行建模；泊松分布，用来对计数过程进行建模，如网站的访客数量、商店的顾客数量等；Gamma 分布和指数分布，用来对时间间隔进行建模，如等车时间等；Beta 分布和 Dirichlet 分布，用于概率分布；Wishart 分布，用于协方差矩阵分布。

显然线性模型中关于 Y 服从正态分布的假设，是包含在该推广条件的。如此推广的意义在于使得广义线性模型可以处理离散型数据，特别是 $(0, 1)$ 的分类数据。

2. $g(E(Y)) = \boldsymbol{\beta}^T \mathbf{X}$，$g$ 为一严格单调，充分光滑的已知函数。

函数 g，称为*连接函数* (非线性的)。此时，因变量 Y 不再与自变量 $\mathbf{X}$ 发生直接联系，存在线性关系的是 $\mathbf{X}$ 和 $g(E(Y))$。进而有 $E(Y) = g^{-1}(\boldsymbol{\beta}^T \mathbf{X})$。自变量 $\mathbf{X}$ 经线性变换后得 $\boldsymbol{\beta}^T \mathbf{X}$，再经过函数 g^{-1} 的转换才和自变量 Y 产生联系。按照指数分布族的公式形式，*连接函数*g 即 $\eta(\theta)$。广义线性模型中，如 g^{-1} 的简化如 $y = x$，同时限制 Y 服从正态分布，则模型将简化为线性回归模型。

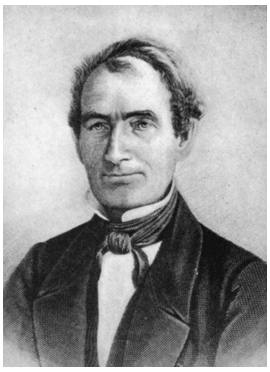

图 3: 皮埃尔·弗朗索瓦·韦吕勒 (Pierre François Verhulst 1804-1849)

0.4.2 Logit 回归

如将 Y 的分布限定为 Bernoulli 两点分布，有均值 p，按照两点分布的指数分布族形式，有

$$g(E(Y)) = g(p) = \ln \frac{p}{1-p} = \boldsymbol{\beta}^T \mathbf{X}$$

所以有

$$E(Y) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{X}}}$$

形如 $f(x) = \frac{1}{1+e^{-x}}$ 函数称为 Sigmoid 函数，或 Logistic 逻辑函数。最早 Sigmoid 函数是皮埃尔·弗朗索瓦·韦吕勒在 1844 或 1845 年在研究它与人口增长的关系时命名的。1838-1847 年间，P.F. 韦吕勒在 A. 凯特勒的指导下，通过调整指数增长模型，将最早的 Sigmoid 函数引入到了用于人口增长的数学模型研究中。Sigmoid 函数之所以叫 sigmoid，是因为函数的图像很想一个字母 S。这个函数是一个很有意思的函数，从图像上我们可以观察到一些直观的特性：函数的取值在 0-1 之间，且在 0.5 处为中心对称，并且越靠近 $x = 0$ 的取值斜率越大。

Sigmoid 函数和正态分布函数的积分形式的形状非常类似 (图5)。

0.4.3 柏松回归

如将 Y 的分布限定为柏松分布，有均值 λ。将柏松分布律改写为指数分布族形式有

图 4: 阿道夫·凯特勒 (Adolphe Quetelet 1796-1874)

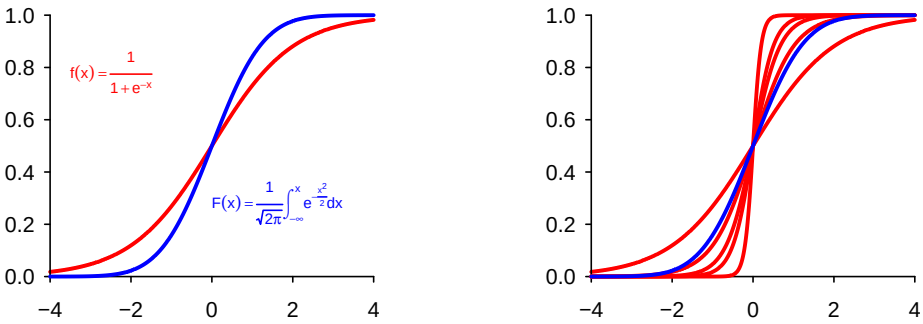

图 5: Sigmoid 函数与标准正态概率分布函数

$$\begin{aligned} f(y,\lambda) &= e^{-\lambda}\frac{\lambda^y}{y!} \\ &= \frac{1}{y^2}e^{-\lambda}e^{\ln\lambda^y} \\ &= \frac{1}{y^2}e^{y\ln\lambda-\lambda} \end{aligned}$$

所以

$$\begin{aligned} h(y) &= \frac{1}{y^2} \\ T(y) &= y \\ \eta(\theta) &= \ln\lambda \\ a(\theta) &= \lambda \\ &= e^{\eta(\theta)} \end{aligned}$$

所以

$$g(E(Y)) = g(\lambda) = \ln\lambda = \boldsymbol{\beta}^T\mathbf{X}$$

因此

$$E(Y) = e^{\boldsymbol{\beta}^T\mathbf{X}}$$

0.4.4 多项回归

0.5 回归模型优化

0.5.1 回归诊断

0.5.2 逐步回归

0.6 非线性回归

0.6.1 可化线性回归的曲线回归

0.6.2 多项式回归

0.6.3 非线性模型