概率统计 B
第七 + 章 期末复习

原著：陈家鼎、刘婉如、汪仁官
制作：李东风，邓明华

2025 年春季

本章目录

本节目录

总体

- 把所研究的对象的全体称为**总体**；
- 把总体中每一个基本单位称为**个体**，主要关心每个个体的某一特性值在总体中的分布情况，
- 因为只关心总体的某个特性值，所以把总体认作是其特性值的随机变量 X，也可以将总体理解成待估计的概率分布。

样本

- 在一个总体 X 中，随机抽取的 n 个个体 $X_1, X_2, \ldots, X_n$ 称为总体 X 的一个容量为 n 的 **样本**，也称 n 为**样本量**。
- 由于 $X_1, X_2, \ldots, X_n$ 是从总体 X 随机抽取出来的可能结果，可以看成是 n 个随机变量。通常采用无放回抽取，使得 $X_1, \cdots, X_n$ **独立同分布**，称其为**简单随机样本**，记作 $i.i.d.$
- 在一次抽取之后，又可以看成是 n 个具体的数值，称为**样本值**, 在使用这个意义时记为小写的 $x_1, x_2, \ldots, x_n$。
- 样本具有二义性：研究其分布特性时是随机变量，具体计算数值时是数值。

经验分布函数

- 问题：给定样本值 $x_1, x_2, \ldots, x_n$，如何估计分布函数 $F(x)$？
- 注意到 $F(x) = P(X \le x)$, 而概率可以用频率来估计；
- **定义（经验分布函数）** 设 $X_1, X_2, \ldots, X_n$ 是 X 的样本，称 x 的函数

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

为 X 的经验分布函数，其中 $I(\cdot)$ 为示性函数。

- **强相合性**：根据大数定律和强大数定律，对固定的 x，只要 n 相当大，$F_n(x)$ 与 $F(x)$ 很接近。原因是 $I(X_i < x) \sim Ber(F(x))$, $E(I(X_i < x)) = F(x)$.
- **一致强相合性 (Glivenko-Cantelli)**：

$$D_n = \sup_x |F_n(x) - F(x)|, \qquad P\left(\lim_{n\to\infty} D_n = 0\right) = 1$$

直方图法

- 直方图法用阶梯函数估计密度函数。
- 把样本 $x_1, x_2, \ldots, x_n$ 从小到大排列为次序统计量 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 后，把数轴分成 m 个小区间，在每个小区间中用

$$\frac{\text{落入小区间的样本点个数}}{n} \cdot \frac{1}{\text{小区间长度}}$$

估计该小区间的密度值 $p(x)$。

直方图估计的理论依据

- 设 $x_1, x_2, \ldots, x_n$ 为来自密度为 $p(x)$ 的总体的样本。
- 用 $R_n(a, b)$ 表示落入区间 $(a, b]$ 的样本点个数。
- 若 $(a, b]$ 很短，可认为 $x \in (a, b]$ 时 $p(x)$ 近似为常数，于是

$$P(a < X \leq b) = \int_a^b p(x)\,dx \approx p(x)(b-a)$$

- 用频率 $R_n(a, b)/n$ 估计概率 $P(a < X \leq b)$，有

$$\begin{aligned} \frac{R_n(a,b)}{n} &\approx p(x)(b-a) \\ p(x) &\approx \frac{R_n(a,b)}{n} \cdot \frac{1}{b-a}, \quad x \in (a, b] \end{aligned}$$

直方图估计的相合性

- 若密度函数 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，对某个 $\delta > 0$，

$$\int_{-\infty}^{\infty} |x|^{\delta} p(x)\, dx$$

收敛，又小区间长度 h_n 满足

$$\lim_{n\to\infty} h_n = 0$$

$$h_n \geq \frac{(\ln n)^2}{n}$$

- 则有

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |p_n(x) - p(x)|\right) = 1.$$

(一致强相合)

Rosenblatt 估计

- 为了估计 $p(x)$，若 $p(x)$ 连续，可根据

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

- 其中 $F(x)$ 可以用经验分布函数 $F_n(x)$ 估计。有

$$\hat{p}_n(x) = \frac{1}{2h}\left[F_n(x+h) - F_n(x-h)\right], \quad x \in (-\infty, \infty)$$

- 这叫做 Rosenblatt 密度估计。
- 注意到

$$\begin{aligned} F_n(x+h) - F_n(x-h) &= \frac{R_n(x-h, x+h)}{n} \\ \hat{p}_n(x) &= \frac{R_n(x-h, x+h)}{n}\frac{1}{2h} \end{aligned}$$

- 所以 Rosenblatt 估计和直方图估计类似。

核密度估计

- Rosenblatt 估计采用了 x 邻域 $[x-h, x+h]$ 内的样本点数，x 邻域内样本点越多，密度估计越大。
- 推广：采纳样本点时，不一刀切，而是离 x 越近的样本点加权越大。
- **定义：** 设 $K(x)$ 是非负函数且 $\int_{-\infty}^{\infty} K(x)\,dx = 1$, 则称 $K(x)$ 是核函数。此时称

$$\tilde{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

  为 $p(x)$ 的核估计。
- 核函数一般选为偶函数，且在正半轴单调下降（类似于正态分布曲线）。

核估计的相合性

- 当 n 无限增大且 $h = h_n$ 无限减小时，核估计 $\tilde{p}(x)$ 与密度 $p(x)$ 无限接近。
- **一致强相合性：** 若 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，且

$$\lim_{n\to\infty} h_n = 0$$

$$\sum_{n=1}^{\infty} \exp\left\{-rnh_n^2\right\} < \infty \quad (\forall r > 0)$$

又核函数 $K(x)$ 为有界变差函数，则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |\tilde{p}_n(x) - p(x)| = 0\right) = 1.$$

- 在核函数和 h 选取合适时核估计比直方图估计精度更高。

最近邻估计

- 核估计是 x 附近的样本点越多则密度估计值越大。
- 最近邻估计是固定 x 附近需要有的样本点数，令邻域区间长度可变。
- 取自然数 $K(n)$（n 为样本量），令

$$
\begin{aligned}
a_n(x) &= \min\{t : t > 0, R_n(x-t, x+t) \ge K(n)\} \\
p_n^*(x) &= \frac{K(n)}{n} \frac{1}{2a_n(x)}
\end{aligned}
$$

最近邻估计的相合性

- 适当条件下 $n \to \infty$ 时 $p_n^*(x)$ 与 $p(x)$ 可以任意接近。
- **一致强相合性：** 若 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，且

$$\lim_{n\to\infty} \frac{K(n)}{n} = 0, \quad \lim_{n\to\infty} \frac{K(n)}{\ln n} = \infty$$

则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |p_n^*(x) - p(x)| = 0\right) = 1.$$

最大似然估计

- 给定样本值 $x_1, x_2, \ldots, x_n$ 后，令

$$L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{n} p(x_i; \theta_1, \theta_2, \ldots, \theta_m)$$

  称为样本 $x_1, x_2, \ldots, x_n$ 的**似然函数**

- **定义：** 如果 $L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$ 在 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 达到最大值，则称 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 为参数 $(\theta_1, \theta_2, \ldots, \theta_m)$ 的**最大似然估计** (Maximum likelihood estimation, MLE)。
- 在相当一般的条件下，最大似然估计有如下优良性质：
  1. 相合性： n 充分大时最大似然估计结果与参数真值之间可以无限接近。
  2. 有效性： 一定意义下没有比最大似然估计更精确的估计。
  3. 渐近正态性： n 充分大时最大似然估计近似服从正态分布。

常用分布参数的最大似然估计

- 指数分布 $p(x;\lambda)=\lambda e^{-\lambda x},\ x>0,\lambda>0$, 其最大似然估计

$$\hat{\lambda}=\frac{1}{\bar{X}}$$

- 正态分布 $N(\mu,\sigma^2)$,

$$\hat{\mu}=\bar{X},\quad \hat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

- Poisson 分布 $P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$

$$\hat{\lambda}=\bar{X}$$

- 均匀分布 $U(a,b)$ 的最大似然估计为

$$\hat{a}=X_{(1)},\quad \hat{b}=X_{(n)}$$

期望和方差的点估计

- 期望 $E(X)$ 的无偏估计为**样本均值**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 方差 $Var(X)$ 的无偏估计为**样本方差**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

矩估计法

- 设随机变量 X 的分布密度是 $p(x;\theta_1,\theta_2,\ldots,\theta_m)$，如果 μ_k 是 X 的 k 阶矩 $(k=1,2,\ldots)$ 存在，显然 μ_k 是 $\theta_1,\theta_2,\ldots,\theta_m$ 的函数：

$$\mu_k = E\left(X^k\right) = g_k(\theta_1,\theta_2,\ldots,\theta_m)$$

- 基于样本可以得到样本各阶矩 $\hat{\mu}_k = \sum_i^n X_i^k$, 从而可以构造方程组

$$\begin{aligned}
\mu_1 &= g_1(\theta_1,\theta_2,\ldots,\theta_m) = \hat{\mu}_1 \\
\mu_2 &= g_2(\theta_1,\theta_2,\ldots,\theta_m) = \hat{\mu}_2 \\
&\ldots\ldots\ldots\ldots \\
\mu_m &= g_m(\theta_1,\theta_2,\ldots,\theta_m) = \hat{\mu}_m
\end{aligned}$$

方程的解 $\hat{\theta}_1,\cdots,\hat{\theta}_m$ 称之为参数的**矩估计** 。

点估计的无偏性和相合性

- 参数 θ 的估计 $\hat{\theta}$ 称为**无偏的**，如果

$$E(\hat{\theta}) = \theta$$

- **定义：** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的**相合估计**，若对任意 $\varepsilon > 0$，

$$\lim_{n\to\infty} P\left(|\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| > \varepsilon\right) = 0.$$

- **定义：** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的**强相合估计**，若

$$P\left(\lim_{n\to\infty} |\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| = 0\right) = 1.$$

点估计的有效性

- **定义：** 若 $\varphi_1(X_1,X_2,\ldots,X_n)$ 和 $\varphi_2(X_1,X_2,\ldots,X_n)$ 都是 $g(\theta)$ 的估计量，满足

$$\begin{aligned}&E_\theta\left[\varphi_1(X_1,X_2,\ldots,X_n)-g(\theta)\right]^2\\ \leq &E_\theta\left[\varphi_2(X_1,X_2,\ldots,X_n)-g(\theta)\right]^2 \quad (\forall\theta\in\Theta)\end{aligned}$$

且存在 θ_0 使上式中小于号成立，则称 φ_1 比 φ_2 **有效**。

- **定义：** 如果 $\varphi(X_1,X_2,\ldots,X_n)$ 是 $g(\theta)$ 的无偏估计量，而且对于 $g(\theta)$ 的任一无偏估计量 $\psi(X_1,X_2,\ldots,X_n)$ 都有

$$D\left(\varphi(X_1,X_2,\ldots,X_n)\right)\leq D\left(\psi(X_1,X_2,\ldots,X_n)\right) \quad (\forall\theta\in\Theta)$$

则称 $\varphi(X_1,X_2,\ldots,X_n)$ 为 $g(\theta)$ 的**最小方差无偏估计**。

置信区间

- **置信区间** 随机区间

$$P_\theta(\theta \in [L(X), U(X)]) \geq 1-\alpha$$

称为参数 θ 的 $1-\alpha$ **置信区间**，称这样的置信区间的**置信水平**为 $1-\alpha$。

- 理解：如果做 K 次抽样（每次抽 n 个样品），则从平均的意义讲，有 $[K(1-\alpha)]$ 次使得区间 $[L(X), U(X)]$ 包含真值 θ。
- 注意：在计算出置信区间后，我们不能说 $E(X)$ 属于这个区间的概率是 $(1-\alpha)$。因为一个计算出来的区间或者包含 $E(X)$, 或者不包含 $E(X)$。
- 样本量 n 越大，置信区间越短。置信度越高，计算的置信区间越长。

正态分布置信区间

- 对于均值 μ, 当方差 σ^2 已知时，$1-\alpha$ 置信区间可取

$$\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

  其中 $\bar{X}$ 为样本均值，S 为样本标准差。
- 对于均值 μ, 当方差 σ^2 已知时，$1-\alpha$ 置信区间可取

$$\left[\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

  其中 $\bar{X}$ 为样本均值，S 为样本标准差。
- 对于方差 σ^2,$1-\alpha$ 置信区间为

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

- 对于两参数 (μ,σ^2), 应用 Bonferroni 不等式可得其 $1-\alpha$ 置信区间

$$\left[\bar{X}-t_{\alpha/4}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{\alpha/4}(n-1)\frac{S}{\sqrt{n}}\right]\cap\left[\frac{(n-1)S^2}{\chi^2_{\alpha/4}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/4}(n-1)}\right]$$

非正态分布的情形

- 如果 X 不是服从正态分布，根据中心极限定理，当 n 充分大时

$$\eta = \frac{\bar{X} - E(X)}{\sqrt{\frac{\mathrm{Var}(X)}{n}}}$$

近似服从标准正态分布。

- 所以 $E(X)$ 的置信度为 $1-\alpha$ 的置信区间仍可用公式

$$\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\mathrm{Var}(X)}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\mathrm{Var}(X)}{n}}\right]$$

计算。

本节目录

假设检验的一般提法

- 一般来讲, 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, θ 是总体 X 的未知参数，人们希望对参数 θ 进行某种推断，可以用 Θ_0 或者 Θ_1 加以区分，其中 Θ_0, Θ_1 是互不相交的参数集合。
- 对于假设

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

检验法 R 可以被事件 R 完全确定，事件 R 发生时拒绝 H_0，称 R 为拒绝域（这里我们将拒绝域与检验方法等同看待）。
- **定义：** 设 α 是 $(0, 1)$ 中的常数. 如果对一切的 $\theta \in \Theta_0$, 有

$$P_\theta(X \in R) \le \alpha,$$

就称拒绝域 R 的检验水平或显著性水平是 α, 实际中是通过检验水平来寻找拒绝域 R.

错误概率

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	Correct rejection	Type I error
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

- 第一类错误率： 当 $\theta \in \Theta_0$ 时

$$\text{第一类错误率} = P_\theta(X \in R)$$

- 第二类错误率： 当 $\theta \in \Theta_1$ 时

$$\text{第二类错误率} = P_\theta(X \in R^c) = 1 - P_\theta(X \in R)$$

功效函数

- 因为

$$P_\theta(X \in R) = \begin{cases} \text{第一类错误率} & \theta \in \Theta_0; \\ 1 - \text{第二类错误率} & \theta \in \Theta_0^c; \end{cases}$$

- 故定义功效函数 (Power function)为

$$\beta(\theta) = P_\theta(X \in R), \forall \theta \in \Theta_0 \cup \Theta_0^c$$

- 理想的情形是 $\beta(\theta)$ 在 $\theta \in \Theta_0$ 尽量接近于零，而 $\beta(\theta)$ 在 $\theta \in \Theta_0^c$ 尽量接近 1. 也就是两类错误率尽量小。

单边检验 p-值

- **定义**：对于单边检验 $H_0: \theta \in \Theta_0$，统计量为 $\phi(X_1, \cdots, X_n)$，则样本 $x_1, \cdots, x_n$ 对于的 p-值为

$$p(x_1, \cdots, x_n) = \sup_{\theta \in \Theta_0} P_\theta(\phi(X_1, \cdots, X_n) > \phi(x_1, \cdots, x_n))$$

- **引理**：设对给定的 $\alpha \in (0,1)$，如果恰有一个 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) = \alpha$$

则 $\varphi(x_1, x_2, \ldots, x_n) > \lambda \Longleftrightarrow p(x_1, x_2, \ldots, x_n) < \alpha.$

- **引理**：设对给定的 $\alpha \in (0,1)$，有 λ 满足

$$\begin{aligned} &\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) \leq \alpha \\ < &\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \geq \lambda) \end{aligned}$$

则 $\varphi(x_1, x_2, \ldots, x_n) > \lambda \Longleftrightarrow p(x_1, x_2, \ldots, x_n) \leq \alpha$

双边检验 p-值

- **定义：** 对于双边检验 $H_0: \theta \in \Theta_0$，统计量为 $\phi(X_1, \cdots, X_n)$，若存在某个参考值 λ_0 介于双侧检验临界值 λ_1,λ_2 之间，即
$\lambda_1 \le \lambda_0 \le \lambda_2$, 则样本 $x_1, \cdots, x_n$ 对于的 p-值为

$$
\begin{aligned}
p(x_1, \cdots, x_n) &= \min\{\sup_{\theta \in \Theta_0} 2P_\theta(\phi(X_1, \cdots, X_n) < \phi(x_1, \cdots, x_n)), 1\} \\
&\quad \text{当}\phi(x_1, \cdots, x_n) \le \lambda_0\text{时} \\
p(x_1, \cdots, x_n) &= \min\{\sup_{\theta \in \Theta_0} 2P_\theta(\phi(X_1, \cdots, X_n) > \phi(x_1, \cdots, x_n)), 1\} \\
&\quad \text{当}\phi(x_1, \cdots, x_n) > \lambda_0\text{时}
\end{aligned}
$$

- 上式可以简化为

$$
\begin{aligned}
& p(x_1, \cdots, x_n) \\
= & \sup_{\theta \in \Theta_0} 2\min\{P_\theta(\phi(X_1, \cdots, X_n) < \phi(x_1, \cdots, x_n)), \\
& \qquad P_\theta(\phi(X_1, \cdots, X_n) > \phi(x_1, \cdots, x_n))\}
\end{aligned}
$$

双边检验 p-值

- **引理：** 设对给定的 $\alpha \in (0,1)$，有唯一的 λ_1 和 λ_2 满足

$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) < \lambda_1) = \frac{\alpha}{2}$$
$$\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda_2) = \frac{\alpha}{2}$$

则

$$\varphi(x_1, x_2, \ldots, x_n) < \lambda_1 \text{ 或} \varphi(x_1, x_2, \ldots, x_n) > \lambda_2$$
$$\Longleftrightarrow p(x_1, x_2, \ldots, x_n) < \alpha.$$

双边检验 p-值

- **引理：** 设对给定的 $\alpha \in (0,1)$，有 λ_1 和 λ_2 满足

$$\begin{aligned}\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)<\lambda_1) &= \frac{\alpha}{2} \\ &< \sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\le\lambda_1) \\ \sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)>\lambda_2) &= \frac{\alpha}{2} \\ &< \sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\ge\lambda_2)\end{aligned}$$

则

$$\begin{aligned}&\varphi(x_1,x_2,\ldots,x_n)<\lambda_1 \text{ 或} \varphi(x_1,x_2,\ldots,x_n)>\lambda_2 \\ \iff & p(x_1,x_2,\ldots,x_n)\le\alpha.\end{aligned}$$

单样本均值检验 (方差已知)

- 若 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, 其中 σ^2 已知。
- 容易验证 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- 标准化得到到统计量

$$Z = \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

- Z 服从标准正态分布 $N(0,1)$, 即 $Z \sim N(0,1)$

Z 检验方法

- 检验问题

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}|\overline{x} - \mu_0|}{\sigma} \geq z_{\frac{\alpha}{2}}\}$$

- 检验问题

$$H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}(\overline{x} - \mu_0)}{\sigma} \geq z_{\alpha}\}$$

- 检验问题

$$H_0 : \mu \geq \mu_0 \leftrightarrow H_1 : \mu < \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}(\overline{x} - \mu_0)}{\sigma} \leq -z_{\alpha}\}$$

单样本均值检验 (方差未知)

- 若 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, 其中 σ^2 未知。
- 因为 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 其中 S^2 为样本方差。可以证明 $\overline{X}$ 和 S^2 独立。
- 于是定义统计量

$$T = \frac{(\overline{X} - \mu)}{\frac{S}{\sqrt{n}}}$$

- T 服从自由度 $n-1$ 的 t 分布，即 $T \sim t(n-1)$.

T 检验方法

- 检验问题

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}|\overline{x} - \mu_0|}{S} \geq t_{\frac{\alpha}{2}}(n-1)\}$$

- 检验问题

$$H_0 : \mu \leq \mu_0 \leftrightarrow H_1 : \mu > \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}(\overline{x} - \mu_0)}{S} \geq t_{\alpha}(n-1)\}$$

- 检验问题

$$H_0 : \mu \geq \mu_0 \leftrightarrow H_1 : \mu < \mu_0$$

- 拒绝域

$$R = \{x : \frac{\sqrt{n}(\overline{x} - \mu_0)}{S} \leq -t_{\alpha}(n-1)\}$$

单样本方差检验（已知均值）

- 若 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, 其中 μ 已知。
- 定义统计量
$$W_1 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2}$$
在均值已知时， W_1 是方差 σ^2 的极大似然估计。
- 由第 2 章抽样分布结论，在 H_0 下， $W_1 \sim \chi^2(n)$

单样本方差检验（均值未知）

- 若 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, 其中 μ 未知。
- 定义统计量
$$W_2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$
- 由第 2 章抽样分布结论，$W_2 \sim \chi^2(n-1)$。
- 下面以均值未知情形为例给出方差的检验方法，通常称之为卡方检验。

方差的卡方检验

- 双边检验

$$H_0 : \sigma^2 = \sigma_0^2 \leftrightarrow H_1 : \sigma^2 \neq \sigma_0^2$$

- 拒绝域

$$\left\{ \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1) \right\} \cup \left\{ \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\frac{\alpha}{2}}(n-1) \right\}$$

- 单边检验

$$H_0 : \sigma^2 \leq \sigma_0^2 \leftrightarrow H_1 : \sigma^2 > \sigma_0^2$$

- 拒绝域

$$\left\{ \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha}(n-1) \right\}$$

- 单边检验

$$H_0 : \sigma^2 \geq \sigma_0^2 \leftrightarrow H_1 : \sigma^2 < \sigma_0^2$$

- 拒绝域

$$\left\{ \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n-1) \right\}$$

两样本均值检验

- 若 $X_1, \cdots, X_m \sim N(\mu_x, \sigma^2)$, $Y_1, \cdots, Y_n \sim N(\mu_y, \sigma^2)$, 其中 σ^2 未知。
- 检验统计量

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

- 其中 $S_p^2 = \frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}$ 称之为混合样本方差 (Pooled sample variance).
- 可以证明 $T \sim t(m+n-2)$。

两样本 T 检验

- 双边检验 $H_0: \mu_x = \mu_y \Leftrightarrow H_1: \mu_x \neq \mu_y$
- 拒绝域

$$\left\{|T| > t_{\frac{\alpha}{2}}(m+n-2)\right\}$$

- 单边检验 $H_0: \mu_x \geq \mu_y \Leftrightarrow H_1: \mu_x < \mu_y$
- 拒绝域

$$\{T < -t_\alpha(m+n-2)\}$$

- 单边检验 $H_0: \mu_x \leq \mu_y \Leftrightarrow H_1: \mu_x > \mu_y$
- 拒绝域

$$\{T > t_\alpha(m+n-2)\}$$

两样本方差检验

- 若 $X_1, \cdots, X_m \sim N(\mu_x, \sigma_x^2)$, $Y_1, \cdots, Y_n \sim N(\mu_y, \sigma_y^2)$, 其中 μ_x, μ_y 未知。
- 检验统计量
$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2},$$
- 其中 $S_x^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \overline{X})^2$,$S_y^2 = \frac{1}{m-1}\sum_{j=1}^n (Y_i - \overline{Y})^2$ 为各自的样本方差;
- 可以证明 $F \sim F(m-1, n-1)$.

两样本 F 检验

- 双边检验 $H_0 : \sigma_x^2 = \sigma_y^2 \Leftrightarrow H_1 : \sigma_x^2 \neq \sigma_y^2$
- 拒绝域

$$\left\{F > F_{\frac{\alpha}{2}}(m-1, n-1)\right\} \cup \left\{F < F_{1-\frac{\alpha}{2}}(m-1, n-1)\right\}$$

- 单边检验 $H_0 : \sigma_x^2 \geq \sigma_y^2 \Leftrightarrow H_1 : \sigma_x^2 < \sigma_y^2$
- 拒绝域

$$\{F < F_{1-\alpha}(m-1, n-1)\}$$

- 单边检验 $H_0 : \sigma_x^2 \leq \sigma_y^2 \Leftrightarrow H_1 : \sigma_x^2 > \sigma_y^2$
- 拒绝域

$$\{F > F_{\alpha}(m-1, n-1)\}$$

正态均值检验

	H_0	H_1	方差σ^2 已知	方差σ^2 未知
单样本	$\mu = \mu_0$	$\mu \neq \mu_0$	单样本Z检验	单样本T检验
	$\mu = \mu_0$	$\mu > \mu_0$		
	$\mu = \mu_0$	$\mu < \mu_0$		
	$\mu \leq \mu_0$	$\mu > \mu_0$		
	$\mu \geq \mu_0$	$\mu < \mu_0$		
两样本	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y \neq \delta_0$	两样本Z检验	两样本T检验
	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y > \delta_0$		
	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y < \delta_0$		
	$\mu_X - \mu_Y \leq \delta_0$	$\mu_X - \mu_Y > \delta_0$		
	$\mu_X - \mu_Y \geq \delta_0$	$\mu_X - \mu_Y < \delta_0$		
成对样本		$\mu_X - \mu_Y \neq \delta_0$	成对样本 Z检验	成对样本 T检验
	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y > \delta_0$		
		$\mu_X - \mu_Y < \delta_0$		
	$\mu_X - \mu_Y \leq \delta_0$	$\mu_X - \mu_Y > \delta_0$		
	$\mu_X - \mu_Y \geq \delta_0$	$\mu_X - \mu_Y < \delta_0$		

正态方差检验

	H_0	H_1	均值μ 已知	均值μ未知
	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$		
	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$		
单样本	$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	卡方检验	卡方检验
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		
	$\sigma_X^2 / \sigma_Y^2 = \lambda_0$	$\sigma_X^2 / \sigma_Y^2 \neq \lambda_0$		
	$\sigma_X^2 / \sigma_Y^2 = \lambda_0$	$\sigma_X^2 / \sigma_Y^2 > \lambda_0$		
两样本	$\sigma_X^2 / \sigma_Y^2 = \lambda_0$	$\sigma_X^2 / \sigma_Y^2 < \lambda_0$	F检验	F检验
	$\sigma_X^2 / \sigma_Y^2 \leq \lambda_0$	$\sigma_X^2 / \sigma_Y^2 > \lambda_0$		
	$\sigma_X^2 / \sigma_Y^2 \geq \lambda_0$	$\sigma_X^2 / \sigma_Y^2 < \lambda_0$		

比率的假设检验

- 设 $X_1, X_2, \cdots, X_n \sim \text{Ber}(p)$, $0 < p < 1$ 是未知参数，p 就是“比率”，如成功率、失败率、有效率等，我们需要对 p 的取值进行检验，进行单样本检验

$$
\begin{aligned}
H_0 : p = p_0, &\quad v.s \quad H_1 : p \neq p_0 \\
H_0 : p = p_0, &\quad v.s \quad H_1 : p \neq p_0 \\
H_0 : p = p_0, &\quad v.s \quad H_1 : p \neq p_0
\end{aligned}
$$

  小样本时采用二项精确检验，大样本时可以用正态近似进行检验。
- 有时候要比较两个比率，即 $X_1, \cdots, X_n \sim \text{Ber}(p_1)$, $Y_1, \cdots, Y_m \sim \text{Ber}(p_2)$, 需要比较 p_1 与 P_2 的大小

$$
\begin{aligned}
H_0 : p_1 = p_2, &\quad v.s \quad H_1 : p_1 \neq p_2 \\
H_0 : p_1 < p_2, &\quad v.s \quad H_1 : p_1 > p_2 \\
H_0 : p_1 > p_2, &\quad v.s \quad H_1 : p_1 < p_2
\end{aligned}
$$

  小样本时可以用 Fisheries 精确检验，大样本时采用正态近似进行检验。

比率检验

	H_0	H_1	精确检验	近似检验
单样本	$p = p_0$	$p \neq p_0$	二项精确检验	正态近似 卡方近似
	$p = p_0$	$p > p_0$		
	$p = p_0$	$p < p_0$		
	$p \leq p_0$	$p > p_0$		
	$p \geq p_0$	$p < p_0$		
两样本	$p_X = p_Y$	$p_X \neq p_Y$	Fisher精确检验	正态近似 卡方近似
	$p_X = p_Y$	$p_X > p_Y$		
	$p_X = p_Y$	$p_X < p_Y$		
	$p_X \leq p_Y$	$p_X > p_Y$		
	$p_X \geq p_Y$	$p_X < p_Y$		
多样本	比率全相等	比率不全相等		卡方检验
	独立	不独立		

本节目录

一元线性回归的条件正态模型

- 一元回归模型

$$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, \cdots, n.$$

- 假设有
  (1). 对任意 i 有 $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_i^2 < +\infty$, 对任意 $i \neq j$, $cov(\epsilon_i, \epsilon_j) = 0$.
  (2). ϵ_i 独立，都服从正态分布；
  (3). Y_i 独立，$i = 1, \cdots, n$.
  (4). x_i 是固定的 (不是随机变量)，$i = 1, \cdots, n$.
- 因此

$$Y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

$$EY_i = \alpha + \beta x_i$$

$$Var(Y_i) = \sigma^2$$

一元线性回归的最小二乘解

- 考虑最简单的线性回归问题
- 残差平方和 (Residual Sum of Squares, RSS)

$$RSS = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$$

- 使得 RSS 最小的参数称为**最小二乘解**

$$\hat{\beta} = \frac{l_{xy}}{l_{xx}}, \hat{\alpha} = \overline{y} - \hat{b}\overline{x}$$

其中

$$l_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad l_{xY} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}), \quad l_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- 计算得到，估计 $\hat{\alpha}$,$\hat{\beta}$ 的协方差

$$Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \overline{x}}{l_{xx}}$$

平方和分解

- 平方和分解公式

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$l_{YY}=U+Q$$

其中等式左边称为总离差平方和 (Total sum of squares,SST), 右边前者称为回归平方和 (Explained sum of squares, ESS), 后者称为残差平方和 (Residual Sum of Squares, RSS).

- 对于回归平方和有

$$\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2=\sum_{i=1}^{n}[(\hat{\alpha}+\hat{\beta}x_i)-\overline{Y}]^2$$
$$=\sum_{i=1}^{n}[\overline{Y}-\hat{\beta}\overline{x}+\hat{\beta}x_i-\overline{Y}]^2=\sum_{i=1}^{n}[\hat{\beta}(x_i-\overline{x})]^2$$
$$=\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{l_{xY}^2}{l_{xx}}$$

统计量分布

- 如果令

$$S^2 = Q/(n-2) = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}\hat{\varepsilon}_i^2$$

形式上类似于正态分布总体的样本方差，有时候也称之为条件正态分布的样本方差。

- **定理**：在条件正态模型下，
  (1). $(\hat{\alpha}, \hat{\beta})$ 与 S^2 相互独立；
  (2). 残差 Q 服从自由度为 $n-2$ 的卡方分布，$Q/\sigma^2 \sim \chi^2_{n-2}$.
- 对于回归平方和

$$U = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(\bar{Y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{Y})^2 = \hat{\beta}^2 l_{xx}$$

- 可以证明：$U/\sigma^2 \sim \chi^2(1)$，$F = \frac{U}{Q/(n-2)} \sim F(1, n-2)$.
- 于是 $T = \sqrt{F} = \hat{\beta}\sqrt{\frac{l_{xx}}{Q/n-2}} \sim t_{n-2}$. T 统计量可以用于对斜率的检验，等价于利用 F 统计量进行检验。

简单线性回归的方差分析表

- 考虑回归直线斜率的检验问题

$$H_0 : \beta = 0 \leftrightarrow H_1 : \beta \neq 0$$

误差来源	自由度	平方和	均方和	F统计量	P值
回归系数(斜率)	1	SS(Reg.) $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MS(Reg.) $\frac{S_{xy}^2}{S_{xx}}$	$F = \frac{\text{MS(Reg.)}}{\text{MS(Res.)}}$	$1 - F_{1,n-2}(F)$
残差	$n-2$	SS(Res.) RSS $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MS(Res.) $\frac{\text{RSS}}{n-2}$		
合计	$n-1$	SST $\sum_{i=1}^{n}(y_i - \bar{y})^2$			

样本相关系数

- 设 U, V 是两个随机变量，其相关系数定义为

$$\rho = \frac{\mathrm{Cov}(U, V)}{\sqrt{\mathrm{Var}(U)\mathrm{Var}(V)}}$$

- 在线性回归中虽然 x 是非随机的变量，但也可以定义样本相关系数为

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{l_{xY}}{\sqrt{l_{xx} l_{YY}}}$$

- 易见

$$\begin{aligned} R^2 = & \frac{l_{xy}^2}{l_{xx} l_{yy}} = \frac{l_{xy}^2}{l_{xx}^2} \cdot \frac{l_{xx}}{l_{yy}} \\ = & \frac{\hat{b}^2 l_{xx}}{l_{yy}} = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}} \end{aligned}$$

决定系数

- 样本决定系数 (Coefficient of Determination)表示的是回归关系已经解释的因变量变异在其总变异中所占的比率.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{l_{xy}^2}{l_{xx}l_{yy}}$$
$$= \frac{\text{回归平方和 ESS}}{\text{总偏差平方和 SST}}$$

- 显然有

$$0 \leq R^2 \leq 1$$
$$R^2 = 1: \text{所有样本点在拟合的直线上}$$
$$R^2 \approx 0: \text{所有样本点远离拟合直线}$$

https://baike.baidu.com/item/可决系数

预测点的样本分布

- 可以证明预测点 Y_0 满足，

$$Y_0-(\hat{\alpha}+\hat{\beta}x_0)\sim N\left(0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}\right]\right)$$

- 标准化得到

$$\frac{Y_0-\hat{\alpha}-\hat{\beta}x_0}{\sqrt{\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}\right]}}\sim N(0,1)$$

$$\frac{(n-2)S^2}{\sigma^2}\sim\chi^2_{n-2}$$

- 由此得到 T 统计量

$$\frac{Y_0-\hat{\alpha}-\hat{\beta}x_0}{\sqrt{S^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}\right]}}\sim T_{n-2}$$

预测区间

- **定义：** 对应一个未知的随机变量 Y, 其基于观测数据 X 的$1-\alpha$ 预测区间是一个随机区间 $[L(X), U(X)]$, 使得对任意参数 θ, 都有

$$P_\theta(Y \in [L(X), U(X)]) \geq 1-\alpha$$

- 预测区间和置信区间的差别：置信区间是针对固定的参数的覆盖，而预测区间针对的是随机变量。

预测区间

- 注意到，

$$\frac{Y_0-(\hat{\alpha}+\hat{\beta}x_0)}{\sqrt{S^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}\right]}}\sim T_{n-2}$$

- 于是得到 Y_0 的 $1-\rho$ 置信区间为

$$\hat{\alpha}+\hat{\beta}x_0-t_{n-2,\rho/2}S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}\leq Y_0$$

$$\leq\hat{\alpha}+\hat{\beta}x_0+t_{n-2,\rho/2}S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}$$

多元线性回归模型

- 设因变量 Y 与自变量 $x_1, x_2, \ldots, x_k$ 有关系式

$$Y_i = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon_i$$

- 其中自变量 $x_1, x_2, \ldots, x_p$ 是非随机的变量，$\varepsilon_i \sim N(0, \sigma^2)$ 相互独立。
- 有 n 组数据

$$\begin{array}{c}
(Y_1; x_{11}, x_{12}, \ldots, x_{1p}) \\
(Y_2; x_{21}, x_{22}, \ldots, x_{2p}) \\
\ldots\ldots\ldots \\
(Y_n; x_{n1}, x_{n2}, \ldots, x_{np})
\end{array}$$

- 首先需要从数据中估计参数 $b_0, b_1, \cdots, b_p$ 和参数 σ^2.

最小二乘估计

- 称使得残差平方和

$$
\begin{aligned}
&Q(b_0, b_1, \ldots, b_k) \\
=&\sum_{t=1}^{n}[y_t - (b_0 + b_1 x_{t1} + b_2 x_{t2} + \cdots + b_k x_{tk})]^2
\end{aligned}
$$

达到最小值的点 $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ 为参数 $b_0, b_1, \ldots, b_k$ 的**最小二乘估计**。

- 可以证明：$\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p$ 为如下的正规方程的解：

$$
\begin{aligned}
&l_{11} b_1 + l_{12} b_2 + \cdots + l_{1p} b_p = l_{1Y} \\
&l_{21} b_1 + l_{12} b_2 + \cdots + l_{1p} b_p = l_{2Y} \\
&\cdots\cdots\cdots\cdots\cdots \\
&l_{p1} b_1 + l_{p2} b_2 + \cdots + l_{pp} b_p = l_{pY} \\
&b_0 = \bar{Y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \cdots - b_p \bar{x}_p
\end{aligned}
$$

其中

$$
\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t, \quad \bar{x}_j = \frac{1}{n}\sum_{t=1}^{n} x_{tj}, j = 1, \cdots, p.
$$

$$
l_{ij} = l_{ji} = \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j), i, j = 1, \cdots, p
$$

$$
l_{iY} = \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(Y_t - \bar{Y}), i = 1, \cdots, p
$$

平方和分解和 σ^2 的无偏估计

- 平方和分解公式：总偏差平方和分解为回归平方和与残差平方和之和。

$$l_{YY}=\sum_{t=1}^{n}(Y_t-\bar{Y})^2=Q+U$$

$$Q=\sum_{t=1}^{n}(Y_t-\hat{Y}_t)^2$$

$$U=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2=\hat{b}_1 l_{1Y}+\hat{b}_2 l_{2Y}+\cdots+\hat{b}_p l_{pY}$$

$$\hat{Y}_t=\hat{b}_0+\hat{b}_1 x_{t1}+\hat{b}_2 x_{t2}+\cdots+\hat{b}_p x_{tp},\quad t=1,2,\ldots,n$$

- 可以证明 $Q/\sigma^2\sim\chi^2(n-p-1)$，令

$$S^2=\hat{\sigma}^2=\frac{1}{n-p-1}Q$$

- 从而

$$E(S^2)=E\left(\frac{1}{n-p-1}Q\right)=\sigma^2$$

即为 σ^2 的无偏估计，也称为回归模型的样本方差。

相关性检验

- 检验自变量与因变量之间线性相关关系是否成立

$$H_0: b_1 = b_2 = \cdots = b_p = 0, \quad H_1: \text{至少有一个系数非零}$$

- 检验统计量为

$$F = \frac{U/p}{Q/(n-p-1)}$$

- 可以证明，在 H_0 下 $F \sim \mathrm{F}(p, n-p-1)$。
- 给定检验水平 α 后，当且仅当 $F > F_\alpha(p, n-p-1)$ 时拒绝 H_0，其中 $F_\alpha(p, n-p-1)$ 为 F 分布的 α 上分位点。
- 检验的 p 值为

$$p = P(F > v)$$

其中 v 为观察数据下统计量 F 的取值，F 为服从 $\mathrm{F}(p, n-p-1)$ 分布的随机变量，当且仅当 p 值小于 α 时拒绝 H_0。

偏回归平方和

- 在平方和分解中，回归平方和 U 代表了所有 p 个自变量的作用。
- 为了研究某个自变量 x_i 的贡献，从原来的数据中建立 Y 对抽掉 x_i 的 $p-1$ 个变量的回归，得到得到一个回归平方和 $U_{(i)}$，一定有 $U_{(i)} \le U$。
- 称

$$u_i = U - U_{(i)}$$

为 $x_1, x_2, \ldots, x_p$ 中 x_i 的**偏回归平方和**。
- 注意偏回归平方和都是在一个变量集合的前提下讨论的。
- u_i 的计算不需要真的重新拟合回归模型，而是有公式

$$u_i = \frac{\hat{b}_i^2}{v_{ii}}$$

其中 v_{ii} 为 $L = (l_{ij})_{p\times p}$ 的逆矩阵的第 i 个主对角线元素。

单个自变量的检验

- 考虑假设检验问题：$H_0: b_i = 0, \quad H_1: b_i \neq 0$，
- 检验统计量

$$F_i = \frac{u_i}{S^2}$$

- 在 H_0 下 $F_i \sim \mathrm{F}(1, n-p-1)$。
- 水平 α 的拒绝域 $R = \{F_i > F_\alpha(1, n-p-1)\}$, 其中 $F_\alpha(1, n-p-1)$ 为自由度 $1, n-p-1$ 的 F 分布 α 上分位点。
- 设 F_i 的观察值为 v，则 p-值为

$$p = P(F > v)$$

其中 F 为 $\mathrm{F}(1, n-p-1)$ 分布随机变量。当 p-值小于 α 拒绝 H_0.

逻辑斯蒂回归模型

- 设因变量和自变量间的关系为

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\sum_{i=1}^{p}\beta_i x_i$$

其中 $\beta_0,\beta_1,\ldots,\beta_k$ 是常数，这时称二分类变量 Y 与自变量 $x_1,x_2,\ldots,x_k$ 的关系符合逻辑斯蒂回归模型。

- 该模型等同于

$$P(Y=1|x_1,x_2,\ldots,x_p)=\frac{\exp(\beta_0+\sum_{i=1}^{p}\beta_i x_i)}{1+\exp(\beta_0+\sum_{i=1}^{p}\beta_i x_i)}$$

逻辑斯蒂回归参数估计

- 模型中的常数 $\beta_0, \beta_1, \ldots, \beta_k$ 通常是未知的，需要从数据中估计，与前面的回归模型不同，这个模型中没有方差项。
- 下面只考虑 $p = 1$，即只有一个自变量的情形，用 x 表示 x_1

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x$$

- 令 $p(x) = P(Y = 1|x)$，则

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- 参数估计可以用最大似然法和最小二乘法。

最大似然估计

- 设数据为 $(x_i, y_i), i = 1, 2, \ldots, n$。
- 则

$$P(Y = y_i | x_i) = [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}$$

- 观测值 $(x_i, y_i), i = 1, 2, \ldots, n$ 对应的似然函数为

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}$$

- 对数似然函数为

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0 + \beta_1 x_i})$$

- 令一阶偏导数都等于零的似然方程组

$$\sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right) = 0$$
$$\sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right) x_i = 0$$

- 若 $(\hat{\beta}_0, \hat{\beta}_1)$ 是似然方程组的根且 $x_1, x_2, \ldots, x_n$ 不全相等，则似然方程组的根是惟一的，而且 $(\hat{\beta}_0, \hat{\beta}_1)$ 是 $L(\beta_0, \beta_1)$ 的最大值点从而是模型参数的最大似然估计。
- 可以证明 $\ln L(\beta_0, \beta_1)$ 是二元严格凹函数。
- 似然方程组有时无解，如所有 y_i 都等于 1 时。

加权最小二乘估计

- 数据有特殊要求。
- 设 $x = x_i$ 时共有 n_i 次观测，n_i 较大，其中事件 $\{Y = 1\}$ 发生了 γ_i 次 $(i = 1, 2, \ldots, m)$（$x_1, x_2, \ldots, x_m$）两两不同）。
- 用

$$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$$

作为 $\ln \frac{p(x_i)}{1-p(x_i)}$ 的估计值 $(i = 1, 2, \ldots, m)$。
- 令

$$\nu_i = \frac{(n_i + 1)(n_i + 2))}{n_i(\gamma_i + 1)(n_i - \gamma_i + 1)} \quad (i = 1, 2, \ldots, m) \tag{3.5}$$

$$\tilde{Q}(\beta_0, \beta_1) = \sum_{i=1}^{m} \frac{1}{\nu_i}(z_i - \beta_0 - \beta_1 x_i)^2$$

- 使 $\tilde{Q}(\beta_0, \beta_1)$ 达到最小值的 $\tilde{\beta}_0, \tilde{\beta}_1$ 称为 β_0, β_1 的加权最小二乘估计。
- 可以证明加权最小二乘估计存在且惟一。
- 令两个一阶偏导数都等于零的方程组

$$\beta_0 \sum_{i=1}^{m} \frac{1}{\nu_i} + \beta_1 \sum_{i=1}^{m} \frac{x_i}{\nu_i} = \sum_{i=1}^{m} \frac{z_i}{\nu_i}$$

$$\beta_0 \sum_{i=1}^{m} \frac{x_i}{\nu_i} + \beta_1 \sum_{i=1}^{m} \frac{x_i^2}{\nu_i} = \sum_{i=1}^{m} \frac{x_i z_i}{\nu_i}$$

- 记

$$l_1=\sum_{i=1}^{m}\frac{1}{\nu_i}, \qquad l_2=\sum_{i=1}^{m}\frac{x_i}{\nu_i}$$

$$l_3=\sum_{i=1}^{m}\frac{x_i^2}{\nu_i} \qquad l_4=\sum_{i=1}^{m}\frac{x_iz_i}{\nu_i}$$

$$l_5=\sum_{i=1}^{m}\frac{z_i}{\nu_i}$$

- 则

$$\tilde{\beta}_0=\frac{l_5l_3-l_2l_4}{l_1l_3-l_2^2} \tag{3.6}$$

$$\tilde{\beta}_1=\frac{l_1l_4-l_2l_5}{l_1l_3-l_2^2} \tag{3.7}$$

加权最小二乘法的理由

- 应该用 $\frac{\gamma_i}{n_i-\gamma_i}$ 作为 $\frac{p(x_i)}{1-p(x_i)}$ 的估计，为避免分子和分母出现零，做连续型修正变成 $\frac{\gamma_i+0.5}{n_i-\gamma_i+0.5}$。
- 可以证明，

$$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$$

近似服从正态分布

$$\mathrm{N}\left(\ln \frac{p(x_i)}{1-p(x_i)}, \frac{1}{n_i p(x_i)[1-p(x_i)]}\right)$$

- 利用 (3.2)，有

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ i = 1, 2, \ldots, m$$

其中 ε 近似服从 $\mathrm{N}(0, \nu_i)$。

- 令

$$\tilde{\varepsilon}_i = \frac{1}{\sqrt{\nu_i}} \varepsilon_i$$

- 则

$$\frac{1}{\sqrt{\nu_i}} z_i = \frac{1}{\sqrt{\nu_i}} (\beta_0 + \beta_1 x_i) + \tilde{\varepsilon}_i, i = 1, 2, \ldots, n$$

其中 $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_n$ 的方差相等，仿照最小二乘法思想令

$$\sum_{i=1}^{m} \left[\frac{1}{\sqrt{\nu_i}} z_i - \frac{1}{\sqrt{\nu_i}} (\beta_0 + \beta_1 x_i) \right]^2$$

达到最小，即 $\tilde{Q}(\beta_0, \beta_1)$ 达到最小。