第四章 回归分析与线性模型

第一节 引言

- 统计学中非常重要的一个领域，研究两个或多个变量间的关系。
- 相对简单，研究的较充分。
- 实用性强，应用前景广阔。
- 不断得到各种类型的推广，新的研究活跃。
- “回归”（Regression）一词最早由十九世纪英国科学家Galton（高尔顿）在研究遗传规律时使用。父子身高/智商?
- “线性模型”指用线性组合描述变量之间的近似关系。

第一节 引言

变量间的关系可以是确定性（$I = U/R$，$S = \frac{1}{2}gt^2$）的或相关性的（体重/身高，儿童识字个数/鞋码）。随机变量Y可能由多个变量x_i确定，即

$$Y = f(x_1, \cdots, x_k)$$

但k可能很大，有些x_i无法观测到。将其中感兴趣的、较重要的、能够处理的变量找出来，无妨设得到关系式

$$Y = g(x_1, \cdots, x_p) + \varepsilon \approx g(x_1, \cdots, x_p)$$

即将不关心、不重要、无法处理部分的影响表示为随机误差ε。

如果用数学上最简单的线性函数去近似表示$g(x_1, \cdots, x_p)$，并仍用ε表示新的随机误差，则有线性模型

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

● 有时可以将非线性问题转化为线性模型，例如$Y \approx Ae^{-B/x}$；有时可以用x的多项式近似表示Y，得线性模型。

第一节 引言

● 理论方面用途：科学问题研究，尤其是复杂问题研究，给出近似关系式，剔除非相关因素，描述因果关系，逼近真理。

● 应用方面用途：很丰富。主要利用以往经验及相关数据进行预测。

● 股市预测。

● 学术界/业界的不同关注点。

● 与大数据关系（big data, huge data, massive data） ……

◇ 重要变量
◇ 模型方法
◇ 计算手段

第一节 引言

常用名称等:

- 因变量（响应变量，Response）Y：被表示的量，本课程中一维。
- 自变量（协变量，Covariate）x：表示Y的量，一个或多个。
- x与Y的关系：有时是因果的（腐蚀时间与腐蚀深度），有时是描述的（汽车里程与油耗），有时是未知的（新科学问题）。
- 一般假设x是非随机的，Y是随机的， Y的随机性来自于误差ε。
- 任务：① 判断哪些x与Y相关（检验），相关性强弱；② 找出Y与相关x的近似关系式（估计）；③ 预测；④ 控制。
- 研究实际问题，往往需要补充相关学科的一些基础知识。

第二节 一元线性回归

一元线性回归就是只有一个自变量，模型为：

$$Y = a + bx + \varepsilon$$

其中a、b为未知参数（回归系数），ε是随机误差。设数据为$(x_1, y_1), \cdots, (x_n, y_n)$，则

$$\begin{cases} y_1 = a + bx_1 + \varepsilon_1 \\ y_2 = a + bx_2 + \varepsilon_2 \\ \quad\cdots \qquad \cdots \\ y_n = a + bx_n + \varepsilon_n \end{cases}$$

对模型的假设为：

① $E(\varepsilon_i) = 0$

② $Var(\varepsilon_i) = \sigma^2$

③ $\varepsilon_1, \cdots, \varepsilon_n$相互独立，

故$Cov(\varepsilon_i, \varepsilon_j) = 0 (i \neq j)$

④ $\varepsilon_i \sim N(0, \sigma^2)$

①′ $E(Y_i) = a + bx_i$

②′ $Var(Y_i) = \sigma^2$

③′ $Y_1, \cdots, Y_n$相互独立，

故$Cov(Y_i, Y_j) = 0 (i \neq j)$

④′ $Y_i \sim N(a + bx_i, \sigma^2)$

第二节 一元线性回归

零均值的假设①是最弱的，永远有，其它假设视情况可适当放宽，如当只做估计时，可以不要假设④；有时③可减弱为不相关。有时方差的假设②也可以放宽或改变，例如对模型$Y = bx + \varepsilon$（油耗/汽车里程）。

一、最小二乘估计（Least Square Estimate, LSE）

我们想找一条直线“最好”地拟合数据构成的散点图，希望直线与所有散点的“距离”最近。如何定义“距离”？当“距离”的平方采用散点到直线的竖直距离（非垂直距离）平方和时，就得到最小二乘估计。

故由模型，定义

$$Q(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

使$Q(a, b)$达到最小的$\hat{a}$、$\hat{b}$，就称为a、b的LSE。因为$Q(a, b)$是a、b的二次多项式，易知在很宽的条件下，最小值存在唯一。利用微积分得

第二节 一元线性回归

$$\begin{cases}\dfrac{\partial Q(a,\,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial Q(a,\,b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0\end{cases}$$

其（唯一，条件：x_i不全相等）解为

$$\begin{cases}\hat{a} = \bar{y} - \hat{b}\bar{x} \quad （回归直线总经过(\bar{x},\bar{y})） \\ \hat{b} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2\end{cases}$$

求导是最直接的方法。对此二次多项式，也可以用配方的方法，请自行推导。

二、检验线性关系

在“一”中可以发现，任意两个变量（体重/身高，体重/智商）x、y都可由最小二乘法建立回归方程

第二节 一元线性回归

$$\hat{y} = \hat{a} + \hat{b}x$$

但x、y是否真的相关？设

$$H_0: b = 0$$

若由数据，不否定H_0，则线性回归无意义。

● 因为$\hat{b}$是随机变量Y_i的线性组合，故也是随机变量，若是连续型，则它取0的概率为0。

● $|\hat{b}|$的大小不能说明H_0是否成立，它与x、y的单位选取有关。

● 应考虑$|\hat{b}|$相对于随机误差的大小。

● H_0成立，仅说明x、y线性不相关，不说明无其他关系，如$y = x^2 + \varepsilon$。

第二节 一元线性回归

平方和分解公式：定义

残差：$e_i = y_i - \hat{y}_i$；

偏差平方和：$l_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$；

回归平方和：$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$；

残差平方和：$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$。

引理1. $l_{yy} = U + Q$。

证明（典型方法）：

$$
\begin{aligned}
l_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= Q + U + 2I
\end{aligned}
$$

第二节 一元线性回归

而

$$
\begin{aligned}
I &= \sum \left(y_i - \hat{a} - \hat{b}x_i\right)\left(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\bar{x}\right) \\
&= \sum \left[y_i - \bar{y} + \hat{b}\bar{x} - \hat{b}x_i\right]\left[\hat{b}(x_i - \bar{x})\right] \\
&= \hat{b}\sum \left[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})\right](x_i - \bar{x}) \\
&= \hat{b}\left[\sum (y_i - \bar{y})(x_i - \bar{x}) - \hat{b}\sum (x_i - \bar{x})^2\right] = 0
\end{aligned}
$$

偏差平方和l_{yy}（已经中心化）表示y的总的波动，回归平方和U表示l_{yy}中可以由x的变化解释的部分，而残差平方和Q刻画了除x外纯随机误差引起的变化。易知$Q = Q(\hat{a}, \hat{b})$。

第二节 一元线性回归

可以证明（多元回归时证），若误差$\varepsilon_i \sim N(0, \sigma^2)$，则在$H_0: b = 0$成立时，有

$$F = :\frac{U}{Q/(n-2)} \left(= \frac{\frac{U}{\sigma^2}/1}{\frac{Q}{\sigma^2}/(n-2)} \right)$$

服从$F(1, n-2)$。给定检验水平α，查表可得临界值λ，当$\{F > \lambda\}$时否定H_0。此时我们习惯上说“线性回归（不）显著”。

● 实际计算：记$l_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$，$l_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$，则易知$\hat{b} = l_{xy}/l_{xx}$，$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\bar{x}\right)^2 = \hat{b}^2 l_{xx} = \hat{b} l_{xy} = l_{xy}^2/l_{xx}$；而仍有$Q = l_{yy} - U$。

● 可以证明，$\hat{a}$、$\hat{b}$分别是a、b的无偏估计；当假设$\varepsilon_i \sim N(0, \sigma^2)$成立时，它们也是其MLE，此时$\sigma^2$的MLE为$Q(\hat{a}, \hat{b})/n$，而其无偏估计为$Q/(n-2)$。

第二节 一元线性回归

三、相关系数

对比概率论中相关系数的定义，我们令

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

易知$|r|\leq 1$。r表示x与y线性相关程度的强弱：$|r|$越接近1，则x与y线性相关程度越大。由下式，

$$r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}}=\frac{U}{l_{yy}}=\frac{U}{U+Q}=1-\frac{Q}{l_{yy}}$$

即回归平方和占偏差平方和的比例。还可得到

$$F=\frac{U}{Q/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}$$

故F与r^2一一对应（单调增）。r有更强的直观意义。

第二节 一元线性回归

四、预测

有了回归方程（已通过检验），当给定自变量新的值$x = x_0$时，相应的$Y_0 = ?$此时模型为

$$Y_0 = a + bx_0 + \varepsilon_0$$

由回归方程，自然得到Y_0的点估计（无偏，也是$E(Y_0)$的点估计）

$$\hat{Y}_0 = \hat{a} + \hat{b}x_0$$

那么Y_0的置信区间呢？在正态性假设下，构造随机变量

$$T = \frac{Y_0 - \hat{Y}_0}{\sqrt{dQ/(n-2)}} = \frac{Y_0 - \hat{a} - \hat{b}x_0}{\sqrt{dQ/(n-2)}} \sim t(n-2)$$

（多元回归时证明），查表可得临界值λ及Y_0的置信区间。上式中，

$$d = 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}$$

● 对Y_0预测的误差来自两部分：① $\hat{a}$、$\hat{b}$与其真值a、b的误差；② Y_0与$a + bx_0$间的误差ε_0。

第二节 一元线性回归

● 当x_0变化时，置信区间上、下限的轨迹构成双曲线。

五、控制

是预测的反问题。

1. 若想因变量取值在Y_0附近， x_0应取何值？$(Y_0-\hat{a})/\hat{b}$。
2. 若想使Y_0取值在区间$[A, B]$内，给定$1-\alpha$，问x_0应在什么范围？

利用“四”的结果，当$x=x_0$时， Y_0的$1-\alpha$水平的置信区间为：

$$\left[\hat{Y}_0-\lambda\sqrt{dQ/(n-2)}, \hat{Y}_0+\lambda\sqrt{dQ/(n-2)}\right]$$

对于$\hat{b}>0$的情形（不妨设），由不等式$\hat{Y}_0-\lambda\sqrt{dQ/(n-2)}\geq A$解出$x_0\geq C_1$，由不等式$\hat{Y}_0+\lambda\sqrt{dQ/(n-2)}\leq B$解出$x_0\leq C_2$，则当$x_0\in[C_1, C_2]$时， Y_0以不少于$1-\alpha$的概率落在区间$[A, B]$内（保守的）。

第二节 一元线性回归

若区间$[A, B]$太窄（和/或Q太大时），可能会有$C_1 > C_2$，此时问题无解。

六、一元齐次线性回归

有些时候，由背景知识知，截距项（常数项）$a = 0$（例如x为汽车行驶里程，Y为耗油量），此时若其他假设仍成立，则得到一元齐次线性回归模型：

$$y = bx + \varepsilon$$

其中ε为零均值的随机误差项。

若数据仍为$(x_1, y_1), \cdots, (x_n, y_n)$，则仿照一元线性回归，令

$$Q(b) = \sum_{i=1}^{n} (y_i - bx_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

求导得

$$Q'(b) = -2 \sum_{i=1}^{n} (y_i - bx_i)\, x_i = 0$$

第二节 一元线性回归

因此最小二乘估计为

$$\hat{b} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

如何检验假设$H_0: b = 0$?

设$\varepsilon_1, \cdots, \varepsilon_n$独立同分布$N(0, \sigma^2)$，记$X = (x_1, \cdots, x_n)'$，$Y = (y_1, \cdots, y_n)'$，取$A$为正交矩阵，其第一行取为$\frac{1}{\sqrt{\sum x_i^2}}(x_1, \cdots, x_n)$，则

$$Z = (z_1, \cdots, z_n)' = AY$$

是正交变换，而$\hat{b} = \frac{1}{\sqrt{\sum x_i^2}} z_1$。因此

$$\| Y \|^2 = Y'Y = Z'Z = \left(Y - \hat{b}X + \hat{b}X\right)'\left(Y - \hat{b}X + \hat{b}X\right)$$
$$= \left(Y - \hat{b}X\right)'\left(Y - \hat{b}X\right) + \hat{b}^2 X'X + 2B$$

显然，$\left(Y - \hat{b}X\right)'\left(Y - \hat{b}X\right) = \| Y - \hat{b}X \|^2 = Q(\hat{b}) = Q$。

第二节 一元线性回归

而交叉项为

$$B=\left(Y-\hat{b}X\right)'\cdot\hat{b}X=Y'\hat{b}X-\hat{b}^2X'X=\frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}-\frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}=0$$

---------- end 20240401

所以，我们有

$$\sum_{i=1}^{n}z_i^2=\sum_{i=1}^{n}y_i^2=\|Y\|^2=\hat{b}^2\|X\|^2+\|Y-\hat{b}X\|^2=z_1^2+\|Y-\hat{b}X\|^2$$

即$\|Y-\hat{b}X\|^2=\sum_{i=2}^{n}z_i^2$。

当H_0成立时，$Y\sim N(0,\sigma^2I_n)$，故$Z\sim N(0,\sigma^2I_n)$，可推出z_1与$\|Y-\hat{b}X\|^2$相互独立，且$z_1\sim N(0,\sigma^2)$，

$$\frac{1}{\sigma^2}\|Y-\hat{b}X\|^2\sim\chi^2(n-1)$$

因此，检验统计量为

$$F=\frac{(z_1/\sigma)^2}{\frac{1}{\sigma^2}\|Y-\hat{b}X\|^2\Big/(n-1)}=\frac{\hat{b}^2\sum_{i=1}^{n}x_i^2}{Q/(n-1)}\sim F(1,n-1)$$

第二节 一元线性回归

而交叉项为

$$B=\left(Y-\hat{b}X\right)'\cdot\hat{b}X=Y'\hat{b}X-\hat{b}^2X'X=\frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}-\frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}=0$$

所以，我们有

$$\sum_{i=1}^{n}z_i^2=\sum_{i=1}^{n}y_i^2=\|Y\|^2=\hat{b}^2\|X\|^2+\|Y-\hat{b}X\|^2=z_1^2+\|Y-\hat{b}X\|^2$$

即$\|Y-\hat{b}X\|^2=\sum_{i=2}^{n}z_i^2$。

当H_0成立时，$Y\sim N(0,\sigma^2I_n)$，故$Z\sim N(0,\sigma^2I_n)$，可推出z_1与$\|Y-\hat{b}X\|^2$相互独立，且$z_1\sim N(0,\sigma^2)$，

$$\frac{1}{\sigma^2}\|Y-\hat{b}X\|^2\sim\chi^2(n-1)$$

因此，检验统计量为

$$F=\frac{(z_1/\sigma)^2}{\frac{1}{\sigma^2}\|Y-\hat{b}X\|^2\Big/(n-1)}=\frac{\hat{b}^2\sum_{i=1}^{n}x_i^2}{Q/(n-1)}\sim F(1,n-1)$$

此结论也可通过利用本章第四节定理4.1得到。

第二节 一元线性回归

● 最小二乘估计是将平方和作为“距离”的平方。若采用其他距离，则可得到其他估计。例如若使绝对值之和最小，则得到最小一乘估计（亦称最小绝对偏差估计），它比LSE稳健。

例1.（教材P149例2.1）合成纤维强度……。

解：由表2.1中数据，$n=24$，$\sum x_i=127.5$，$\sum y_i=113.1$，$\sum x_i^2=829.61$，$\sum y_i^2=650.93$，$\sum x_i y_i=731.60$，所以

$$l_{xx}=\sum x_i^2-\frac{1}{n}\left(\sum x_i\right)^2\approx 152.27$$

$$l_{xy}=\sum x_i y_i-\frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)\approx 130.76$$

$$l_{yy}=\sum y_i^2-\frac{1}{n}\left(\sum y_i\right)^2\approx 117.95$$

第二节 一元线性回归

因此，$\hat{b} = l_{xy}/l_{xx} \approx 0.859$，$\hat{a} = \bar{y} - \hat{b}\bar{x} \approx 0.15$，经验方程为

$$\hat{Y} = 0.15 + 0.859x$$

下面检验$H_0: b = 0$。经计算，$U = l_{xy}^2/l_{xx} \approx 112.29$， $Q = l_{yy} - U \approx 5.66$，因此，$F = \frac{112.29}{5.66/22} \approx 436.35$。查表可知，$1 - 0.01$分位点为$7.95$，$1 - 0.0005$分位点为$16.7$，故否定$H_0$，即$Y$与$x$显著相关。

此时，相关系数$r = {l_{xy}}/{\sqrt{l_{xx}l_{yy}}} \approx 0.9757$，非常接近1。

若预测$x_0 = 7.0$时的强度Y_0，其点估计为$\hat{Y}_0 = 0.15 + 0.859 \times 7 = 6.163$。其置信区间为$\left[\hat{Y}_0 - \lambda\sqrt{dQ/(n-2)}, \hat{Y}_0 + \lambda\sqrt{dQ/(n-2)}\right]$，查表得（$df = 22$，$\alpha = 0.05$）$\lambda = 2.074$，而$d \approx 1.06$，故置信区间为$[5.080, 7.246]$。

第二节 一元线性回归

若希望Y以不小于95%的概率落在$[3.0, 6.0]$内，解方程

$$0.15 + 0.859C_1 - 2.074\sqrt{dQ/(n-2)} = 3.0$$

d中含有C_1，但因$(C_1 - \bar{x})^2/l_{xx} \approx 0$，近似计算，取$d \approx 1 + 1/n$，解得$C_1 \approx 4.57$。同理，解方程

$$0.15 + 0.859C_2 + 2.074\sqrt{dQ/(n-2)} = 6.0$$

得到$C_2 \approx 5.56$。故应取$x \in [4.57, 5.56]$。

例2. 一元回归应用实例：北京地区语文高考作文成绩监控方法（北京大学孙山泽教授，见北京大学学报自然科学版，1996，第32卷，第1期，P1）

● 在一般实际应用中，由于问题多数较复杂，一元回归很少直接应用。

● 即使问题中仅有一个自变量x，如果与Y的关系是非线性的，也可以用x的多项式逼近Y（多元回归）。

第三节 线性模型的参数估计

一、线性模型

设数据为$\{y_i;\ x_{i1},\ \cdots,\ x_{ip},\ i=1,\ \cdots,\ n\}$，其中$y_i$为第$i$次观测中因变量的值，$x_{i1},\ \cdots,\ x_{ip}$为第1，…，第$p$个自变量的值。用线性关系式描述：

$$Y=\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon$$

拟合数据，就得到线性模型

$$\begin{cases} y_1=\beta_1 x_{11}+\cdots+\beta_p x_{1p}+e_1 \\ \cdots\cdots \\ y_n=\beta_1 x_{n1}+\cdots+\beta_p x_{np}+e_n \end{cases} \tag{1}$$

其中$\beta_1,\ \cdots,\ \beta_p$是未知参数，$e_1,\ \cdots,\ e_n$为随机误差。

在做理论分析时，习惯上不设常数项，因为总可以形式地令$x_1\equiv 1$，故β_1即可为常数项。我们设自变量$x_1,\ \cdots,\ x_p$是非随机的，而误差$e_1,\ \cdots,\ e_n$（从而因变量$Y_1,\ \cdots,\ Y_n$）为随机的。利用矩阵语言，记

$$Y=\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1},\ X=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times p},\ \beta=\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{p\times 1},\ e=\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}_{n\times 1}$$

第三节 线性模型的参数估计

则线性模型（1）可表示为

$$Y = X\beta + e \qquad (1)'$$

我们需要估计（至少）p个参数，因此要求$n \geq p$。（当代有“小n大p”问题）

下面根据研究目的的不同，对误差e给出两种假设。

假设（Assumption）A:

$$E(e) = 0, \quad Cov(e, e) = E(e \cdot e') = \sigma^2 I_n$$

即不假设正态，不假设相互独立，但假设零均值，不相关，同方差。

假设（Assumption）B:

$$e \sim N(0, \sigma^2 I_n)$$

即e_i间相互独立，均服从$N(0, \sigma^2)$。

● 还可有更弱的假设，如e_i的方差不同。

第三节 线性模型的参数估计

二、最小二乘估计

如何估计$\beta_1, \cdots, \beta_p$，使得拟合的好？按一元回归思想，应使$e_1, \cdots, e_n$整体上最小。令

$$Q(\beta) = \sum_{i=1}^{n} e_i^2 = e' \cdot e = \sum_{i=1}^{n} \left[y_i - \left(\beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right) \right]^2$$
$$= (Y - X\beta)'(Y - X\beta) = \| Y - X\beta \|^2$$

定义1. 称$\hat{\beta}$为β的最小二乘估计，若

$$Q(\hat{\beta}) \leq Q(\beta) \qquad (\forall \beta \in R^p)$$

定理3.1. （1）最小二乘估计（LSE）一定存在。

（2） $\hat{\beta}$是LSE的充要条件是它满足正规方程$X'X\beta = X'Y$。

（3）LSE唯一的充要条件是X满秩，此时$\hat{\beta} = (X'X)^{-1}X'Y$。

第三节 线性模型的参数估计

证明：（1）记X的p个列向量张成的R^n的线性子空间（其维数$\leq p$，为X的秩）为$\mu(X)$，记$\xi = Proj(Y) \in \mu(X)$为向量Y到$\mu(X)$的投影，则$Y = \xi + (Y - \xi)$，ξ与$Y - \xi$正交，$Y - \xi$是Y到$\mu(X)$的最短距离，因此，

$$\| Y - \xi \| \leq \| Y - X\beta \|, \qquad (\forall \beta \in R^p)$$

因为$\xi \in \mu(X)$，故存在（可能不唯一）$\hat{\beta}$，使得$\xi = X\hat{\beta}$，即$\hat{\beta}$是一个LSE。

（2）$X'X\hat{\beta} = X'Y \Leftrightarrow X'\left(Y - X\hat{\beta}\right) = 0 \Leftrightarrow Y - X\hat{\beta}$与$X$的所有列向量正交$\Leftrightarrow Y - X\hat{\beta}$与$\mu(X)$正交 $\Leftrightarrow \| Y - X\hat{\beta} \|$达到最小值。

（3）充分性显然，下证必要性，反证：若X不满秩，则存在$a \neq 0$，$a \in R^p$，使得$Xa = 0$。设$\hat{\beta}$是一个LSE，则

$$X'X\left(\hat{\beta} + a\right) = X'X\hat{\beta} + X'Xa = X'Y$$

即$\hat{\beta} + a$也是一个LSE，矛盾，证毕。

第三节 线性模型的参数估计

形式上，

$$Q(\beta) = (Y - X\beta)'(Y - X\beta) = (Y' - \beta'X')(Y - X\beta)$$
$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

是β的二次型。

对$Q(\beta)$求偏导，并令$\nabla Q(\beta) = 0$，得正规方程

$$X'X\beta = X'Y$$

将其代入$Q(\beta)$，得

$$Q(\hat{\beta}) = Y'Y - \hat{\beta}'X'Y = Y'Y - Y'X\hat{\beta}$$

特别地，当X满秩时，$X'X$是$p \times p$正定矩阵，故而

$$Q(\hat{\beta}) = Y'Y - Y'X(X'X)^{-1}X'Y = Y'[I_n - X(X'X)^{-1}X']Y =: Y'AY$$

一元回归中，我们介绍过，LSE无偏，$Q/(n-2) = Q(\hat{a}, \hat{b})/(n-2)$是$\sigma^2$的无偏估计。将其扩展，得到下面一般性结论。

第三节 线性模型的参数估计

定理3.2. 设X满秩，且假定A成立，则

(1) $E(\hat{\beta}) = \beta$,

(2) $Cov(\hat{\beta}, \hat{\beta}) = \sigma^2 (X'X)^{-1}$,

(3) $EQ(\hat{\beta}) = (n-p)\sigma^2$。

证明：

(1) $$E(\hat{\beta}) = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'EY$$
$$= (X'X)^{-1}X'(X\beta + 0) = \beta$$

(2) $$Cov(\hat{\beta}, \hat{\beta}) = Cov\left((X'X)^{-1}X'(X\beta + e), (X'X)^{-1}X'(X\beta + e)\right)$$
$$= Cov\left((X'X)^{-1}X'e, (X'X)^{-1}X'e\right)$$
$$= E[(X'X)^{-1}X'e \cdot e'X(X'X)^{-1}]$$
$$= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

第三节 线性模型的参数估计

(3) $EQ(\hat{\beta}) = E(Y'AY)$ （记$A = I_n - X(X'X)^{-1}X'$）

$= Etr(Y'AY)$ （$Y'AY$是数，即1×1矩阵）

$= Etr(AY \cdot Y')$ （因为$tr(AB) = tr(BA)$）

$= tr\big(A \cdot E(YY')\big)$

注意到$YY' = X\beta\beta'X' + e\beta'X' + X\beta e' + e \cdot e'$，故而

$$E(YY') = X\beta\beta'X' + 0 + 0 + \sigma^2 I_n = \sigma^2 I_n + X\beta\beta'X'$$

又因为$AX = \big(I_n - X(X'X)^{-1}X'\big)X = X - X = 0$，故

$$A \cdot E(YY') = \sigma^2 A + AX\beta\beta'X' = \sigma^2 A$$

所以$EQ(\hat{\beta}) = \sigma^2 tr(A) = \sigma^2\big[tr(I_n) - tr\big(X(X'X)^{-1}X'\big)\big]$

$= \sigma^2\big[n - tr\big((X'X)^{-1}X' \cdot X\big)\big]$ （因为$tr(AB) = tr(BA)$）

$= \sigma^2\big[n - tr\big(I_p\big)\big] = \sigma^2(n-p)$。 证毕。

此时，σ^2的无偏估计为：$\hat{\sigma}^2 = Q(\hat{\beta})/(n-p)$。

第三节 线性模型的参数估计

例1.（第二节例1，合成纤维强度，教材P149例2.1 ）

解：形式地引入变量$x^{(1)} \equiv 1$，令$x^{(2)} = x$，则$p = 2$，$n = 24$。

$$X = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.0 \\ \vdots & \vdots \\ 1 & 10.0 \end{pmatrix}$$ 满秩，故$X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} 24 & 127.5 \\ 127.5 & 829.61 \end{pmatrix}$

$X'Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} 113.1 \\ 731.6 \end{pmatrix}$，$|X'X| = n\sum x_i^2 - n^2\overline{x}^2 = nl_{xx}$，所以

$$(X'X)^{-1} = \frac{1}{|X'X|}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{1}{l_{xx}}\begin{pmatrix} \frac{\sum x_i^2}{n} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} \approx \begin{pmatrix} 0.227 & -0.0349 \\ -0.0349 & 0.0066 \end{pmatrix}$$

因此$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y = \frac{1}{l_{xx}}\begin{pmatrix} \frac{\sum x_i^2}{n} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \cdots$

$$= \begin{pmatrix} \overline{y} - \overline{x}\, l_{xy}/l_{xx} \\ l_{xy}/l_{xx} \end{pmatrix} \approx \begin{pmatrix} 0.15 \\ 0.859 \end{pmatrix}。$$

第三节 线性模型的参数估计

● 求解$\hat{\beta}$等价于解一个p元一次方程组。当$p=2$时，用不用矩阵区别不大，当p较大时，矩阵有优势。

● 正规方程组$X'X\beta=X'Y$总有解，但当$|X'X|=0$（即X不满秩）时，解不唯一。

例2. 两个物体的重量分别为β_1、β_2，一起放在天平上称量，因有测量误差（设其服从正态分布），n次的结果分别为$y_1,\cdots,y_n$，试估计β_1、β_2。

解：从直观上看，β_1、β_2不可估。引入形式变量$x_1=x_2\equiv 1$，则由线性模型，有

$$\begin{cases} y_1=\beta_1 x_1+\beta_2 x_2+e_1 \\ \quad\cdots\quad\cdots \\ y_n=\beta_1 x_1+\beta_2 x_2+e_n \end{cases}$$

其中$e\sim N(0,\sigma^2 I)$，而

第三节 线性模型的参数估计

$X=\begin{pmatrix}1 & 1\\ \vdots & \vdots\\ 1 & 1\end{pmatrix}$不满秩，$X'X=\begin{pmatrix}n & n\\ n & n\end{pmatrix}$不可逆，$X'Y=\begin{pmatrix}\sum y_i\\ \sum y_i\end{pmatrix}$，故任何满足$X'X\beta=X'Y$，即满足$\hat{\beta}_1+\hat{\beta}_2=\frac{1}{n}\sum y_i=\bar{y}$的$\hat{\beta}$都是$\beta$的LSE。

● 上例中，β_1（同理β_2）不存在无偏估计。反证，若$\varphi(y_1,\cdots,y_n)$是，则

$E\varphi(Y_1,\cdots,Y_n)=\beta_1$，而因为$Y_i\sim N(\beta_1+\beta_2,\sigma^2)$，故对任意的$\beta_1$、$\beta_2$，有

$$g(\beta_1,\beta_2)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\varphi(y_1,\cdots,y_n)\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(y_i-\beta_1-\beta_2)^2}dy_1\cdots dy_n=\beta_1$$

注意到$g(0,0)=g(1,-1)$，推出$0=1$，矛盾。

● （不）可识别性：称模型（或参数）是不可识别的，若Θ中存在$\theta_1\neq\theta_2$，使得

$$P_{\theta_1}=P_{\theta_2}$$

第三节 线性模型的参数估计

一般情况下，X不满秩是因为某些自变量线性相关，即某个自变量可由其他几个自变量的线性表示。如……。解决方法为删除一些自变量。当自变量间可完全决定，但非线性时，则不必删除。

例3. （教材P167，例3.2）

解：首先对t中心化得到x，并将T 变换为y。

由散点图，用x的二次函数拟合。此时$n = 11$，

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} = \begin{pmatrix} 1 & -5 & 25 \\ 1 & -4 & 16 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 25 \end{pmatrix}$$

满秩，$\sum_{i=1}^{11} y_i = 106$，$\sum_{i=1}^{n} x_i y_i = 20$，$\sum_{i=1}^{n} x_i^2 y_i = 688$，

因此

第三节 线性模型的参数估计

$$X'X=\begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}=\begin{pmatrix} 11 & 0 & 110 \\ 0 & 110 & 0 \\ 110 & 0 & 1958 \end{pmatrix}$$

而$|X'X|=\cdots=1038180$，故

$$(X'X)^{-1}=\frac{1}{1038180}\begin{pmatrix} 215380 & 0 & -12100 \\ 0 & 9438 & 0 \\ -12100 & 0 & 1210 \end{pmatrix}$$，而

$$X'Y=\begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{pmatrix}=\begin{pmatrix} 106 \\ 20 \\ 688 \end{pmatrix},$$

所以

第三节 线性模型的参数估计

$$\hat{\beta}=\frac{1}{1038180}\begin{pmatrix}14505480\\188760\\-450120\end{pmatrix}\approx\begin{pmatrix}13.9721\\0.1818\\-0.4336\end{pmatrix}。$$

$Q=Q(\hat{\beta})=Y'Y-Y'X\hat{\beta}=\cdots\approx 1188-1186.3776=1.6382$，故 $\hat{\sigma}^2=Q/(n-3)\approx 0.2048$。利用定理3.2中的（2）可知，$\beta_1$、$\beta_2$和$\beta_3$的估计的方差分别为0.0425、0.0019和0.00024。经验方程为

$$\hat{y}=13.9721+0.1818x-0.4336x^2$$

或（代回）

$$T=98.727+0.1077t-0.001734t^2$$

----------end 4月6日

三、线性可估性

由本节前面例2知，β_1、β_2不可估，但$\beta_1+\beta_2$可估。又如，三个物体称重，前两个总在一起，第三个可分开，则β_3可估，β_1、β_2不可估。可见，整体可估性存在问题时，可能部分参数可估，或一些参数的线性组合可估。

第三节 线性模型的参数估计

设$c = (c_1, \cdots, c_p)' \in R^p$为任一$p$维向量。

定义2. 称$c'\beta$是（线性）可估的，若存在$a = (a_1, \cdots, a_n)' \in R^n$，使得$E(a'Y) = c'\beta$对任意的$\beta \in R^p$成立。

当X满秩时，显然对任意的$c \in R^p$，$c'\beta$可估，此时$a' = c'(X'X)^{-1}X'$。当X不满秩时，仅对某些$c \in R^p$，$c'\beta$可估。

定理3.3. 设假定A成立，则$c'\beta$可估的充要条件为$c \in \mu(X')$，即c'可表示为X的行向量的线性组合。

证明：$c \in \mu(X') \Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ c' = a'X$

$\Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ (a'X - c')\beta = 0\ \ (\forall \beta \in R^p)$

$\Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ E(a'Y) = a'X\beta = c'\beta\ \ (\forall \beta \in R^p)$

$\Leftrightarrow$ $c'\beta$可估，证毕。

第三节 线性模型的参数估计

当X满秩时，$\mu(X')=R^p$，故任给$c\in R^p$， $c'\beta$可估。

● 对例2，所有X的行向量均为$(1,1)$，故$\mu(X')$维数为一，所有可估的组合为$k(\beta_1+\beta_2)$（$k\in R^1$）。

若$c'\beta$可估，如何估计？当X满秩时，自然用$c'\hat{\beta}$。若不满秩，则$\hat{\beta}$不唯一。

---------- end 20240415

定理3.4.（高斯-马尔科夫）设假定A成立， $c'\beta$线性可估，则任取$\hat{\beta}$为β的一个LSE，$c'\hat{\beta}$必为$c'\beta$的唯一的最小方差线性无偏估计。

证明：因为$c'\beta$可估，设$a'Y$是其任一无偏估计，令

$$a^*=Proj_{\mu(X)}(a)\in\mu(X),\qquad \tilde{a}=a-a^*,$$

则$a^*\in\mu(X)$，$\tilde{a}'X=0$，即$\tilde{a}$与X的所有列向量正交。因为$a'Y=a^{*\prime}Y+\tilde{a}'Y$，取期望得$c'\beta=E[a^{*\prime}Y]+\tilde{a}'X\beta=E[a^{*\prime}Y]$，故$a^{*\prime}Y$也是$c'\beta$的一个无偏估计。

第三节 线性模型的参数估计

$$\hat{\beta}=\frac{1}{1038180}\begin{pmatrix}14505480\\188760\\-450120\end{pmatrix}\approx\begin{pmatrix}13.9721\\0.1818\\-0.4336\end{pmatrix}。$$

$Q=Q(\hat{\beta})=Y'Y-Y'X\hat{\beta}=\cdots\approx 1188-1186.3776=1.6382$，故 $\hat{\sigma}^2=Q/(n-3)\approx 0.2048$。利用定理3.2中的（2）可知，$\beta_1$、$\beta_2$和$\beta_3$的估计的方差分别为0.0425、0.0019和0.00024。经验方程为

$$\hat{y}=13.9721+0.1818x-0.4336x^2$$

或（代回）

$$T=98.727+0.1077t-0.001734t^2$$

三、线性可估性

由本节前面例2知，β_1、β_2不可估，但$\beta_1+\beta_2$可估。又如，三个物体称重，前两个总在一起，第三个可分开，则β_3及$\beta_1+\beta_2$可估，β_1、β_2不可估。可见，整体可估性存在问题时，可能部分参数可估，或一些参数的线性组合可估。

第三节 线性模型的参数估计

设$c = (c_1, \cdots, c_p)' \in R^p$为任一$p$维向量。

定义2. 称$c'\beta$是（线性）可估的，若存在$a = (a_1, \cdots, a_n)' \in R^n$，使得$E(a'Y) = c'\beta$对任意的$\beta \in R^p$成立。

当X满秩时，显然对任意的$c \in R^p$，$c'\beta$可估，此时$a' = c'(X'X)^{-1}X'$。当X不满秩时，仅对某些$c \in R^p$，$c'\beta$可估。

定理3.3. 设假定A成立，则$c'\beta$可估的充要条件为$c \in \mu(X')$，即c'可表示为X的行向量的线性组合。

证明：$c \in \mu(X') \Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ c' = a'X$

$\Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ (a'X - c')\beta = 0\ \ (\forall \beta \in R^p)$

$\Leftrightarrow$ 存在$a \in R^n$, $s.t.\ \ E(a'Y) = a'X\beta = c'\beta\ \ (\forall \beta \in R^p)$

$\Leftrightarrow$ $c'\beta$可估，证毕。

第三节 线性模型的参数估计

当X满秩时，$\mu(X')=R^p$，故任给$c\in R^p$， $c'\beta$可估。

● 对例2，所有X的行向量均为$(1,1)$，故$\mu(X')$维数为一，所有可估的组合为$k(\beta_1+\beta_2)$（$k\in R^1$）。

若$c'\beta$可估，如何估计？当X满秩时，自然用$c'\hat{\beta}$。若不满秩，则$\hat{\beta}$不唯一。

定理3.4.（高斯-马尔科夫）设假定A成立， $c'\beta$线性可估，则任取$\hat{\beta}$为β的一个LSE，$c'\hat{\beta}$必为$c'\beta$的唯一的最小方差线性无偏估计。

证明：因为$c'\beta$可估，设$a'Y$是其任一无偏估计，令

$$a^*=Proj_{\mu(X)}(a)\in\mu(X),\qquad \tilde{a}=a-a^*,$$

则$a^*\in\mu(X)$，$\tilde{a}'X=0$，即$\tilde{a}$与X的所有列向量正交。因为$a'Y=a^{*\prime}Y+\tilde{a}'Y$，取期望得$c'\beta=E[a^{*\prime}Y]+\tilde{a}'X\beta=E[a^{*\prime}Y]$，故$a^{*\prime}Y$也是$c'\beta$的一个无偏估计。

第三节 线性模型的参数估计

设$b'Y$是$c'\beta$的另一个无偏估计，则同理可有b^*和$\tilde{b}$，且$b^{*\prime}Y$也是。故

$$E\left[a^{*\prime}Y - b^{*\prime}Y\right] = (a^* - b^*)'X\beta = 0 \qquad \forall \beta \in R^p$$

因此$(a^* - b^*)'X = 0$，但$a^* - b^* \in \mu(X)$，故而$a^* = b^*$。

此时

$$Var(b'Y) = b'cov(Y, Y)b = b'\sigma^2 Ib = \sigma^2 \| b \|^2$$
$$= \sigma^2\left(\| b^* \|^2 + \| \tilde{b} \|^2\right) = \sigma^2\left(\| a^* \|^2 + \| \tilde{b} \|^2\right) \geq \sigma^2 \| a^* \|^2 = Var\left(a^{*\prime}Y\right)$$

且等号成立当且仅当$\| \tilde{b} \| = 0$，即$b = a^*$。因此，$a^{*\prime}Y$是$c'\beta$的唯一的最小方差线性无偏估计。

最后证明$a^{*\prime}Y = c'\hat{\beta}$。任取$\beta$的LSE$\hat{\beta}$，由定理3.1的证明，$X\hat{\beta} = Proj(Y)$，故$Y - X\hat{\beta} \perp \mu(X)$，因此有$a^{*\prime}\left(Y - X\hat{\beta}\right) = 0$，即$a^{*\prime}Y = a^{*\prime}X\hat{\beta}$，因为$E\left(a^{*\prime}Y\right) = a^{*\prime}X\beta = c'\beta$对任意$\beta \in R^p$成立，所以$a^{*\prime}X = c'$，故$a^{*\prime}Y = a^{*\prime}X\hat{\beta} = c'\hat{\beta}$，证毕。

第三节 线性模型的参数估计

定义3. 若$\widehat{\beta}$是β的LSE，则称$c'\widehat{\beta}$（well-defined）为$c'\beta$的LSE。

● 可以证明，若假定B成立， $c'\beta$可估，则$c'\widehat{\beta}$是$c'\beta$的一致最小方差无偏估计。

定理3.5. 设假定A成立，X的秩为r，则$EQ(\widehat{\beta}) = (n-r)\sigma^2$，即$\widehat{\sigma}^2 = Q(\widehat{\beta})/(n-r)$是$\sigma^2$的无偏估计。

证明：取$\xi_1, \cdots, \xi_r$为$\mu(X)$的一组标准正交基，添加$\xi_{r+1}, , \cdots, \xi_n$使之成为$R^n$的一组标准正交基。记

$$V_1 = (\xi_1 \cdots \xi_r)_{n\times r},\ V_2 = (\xi_{r+1} \cdots \xi_n)_{n\times(n-r)}$$

$$V = (V_1 V_2)_{n\times n}$$

则V为正交矩阵。令$Z = V'Y$，则有

$$EZ = V'X\beta = \begin{pmatrix} V_1' \\ V_2' \end{pmatrix} X\beta = \begin{pmatrix} V_1'X\beta \\ 0 \end{pmatrix}$$

第三节 线性模型的参数估计

$$cov(Z, Z) = \sigma^2 V'V = \sigma^2 I$$

即Y经正交变换后仍不相关。因此，

$$Y = VZ = (\xi_1 \cdots \xi_n)\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \sum_{i=1}^{r} Z_i\xi_i + \sum_{i=r+1}^{n} Z_i\xi_i$$

记$\xi = \sum_{i=1}^{r} Z_i\xi_i$，$\eta = \sum_{i=r+1}^{n} Z_i\xi_i$，则$\xi \in \mu(X)$，$\eta \perp \mu(X)$，$Y = \xi + \eta$是一个正交分解，故

$$Q(\hat{\beta}) = \| Y - \xi \|^2 = \| \eta \|^2 = \sum_{i=r+1}^{n} Z_i^2$$

注意当$i > r$时，$EZ_i = 0$，$var(Z_i) = \sigma^2$，故

$$EQ(\hat{\beta}) = \sum_{i=r+1}^{n} EZ_i^2 = \sum_{i=r+1}^{n} [var(Z_i) + (EZ_i)^2] = (n-r)\sigma^2$$

证毕。

- 当假定B成立时，可以证明，$\hat{\sigma}^2$是σ^2的（一致）最小方差无偏估计。

- 当X满秩时，结论与定理3.2中的（3）一致。

第三节 线性模型的参数估计

四、带约束的线性模型的参数估计

背景为参数之间存在已知的线性联系，例如天文测量三个天体围成的三角形的角度，对三个角独立测量，均有测量误差，因此一般$y_1 + y_2 + y_3 \neq \pi$，但我们已知$\theta_1 + \theta_2 + \theta_3 = \pi$，称平滑问题。又如测量铁球的体积和重量，比重已知，故也有线性关系。

一般约束条件：线性模型如前，但要求β满足线性方程组

$$H\beta = r_0$$

其中H为$s \times p$矩阵， H的秩为 $s < p$，r_0为s维向量。

定义4. 记$\Theta_0 = \{\beta \in R^p : H\beta = r_0\}$，若$\hat{\beta}$满足：① $\hat{\beta} \in \Theta_0$，② $\| Y - X\hat{\beta} \|^2 = \min_{\beta \in \Theta_0} \| Y - X\beta \|^2$，则称$\hat{\beta}$是$\beta$的带约束条件$H\beta = r_0$的LSE，记为$\hat{\beta} = \hat{\beta}_H$。

上述天文测量的例子中，$\beta = (\theta_1, \theta_2, \theta_3)'$，$H = (1, 1, 1)$，$r_0 = \pi$。

第三节 线性模型的参数估计

我们自然要求约束条件不相互矛盾，即$H\beta = r_0$有解，如不能同时出现约束条件$\theta_1 + \theta_2 + \theta_3 = \pi$和$\theta_1 + \theta_2 + \theta_3 = 2\pi$，且不重复，即$s$个约束条件不线性相关，即$H$的秩为$s$。

在上述要求满足时，带约束条件的LSE总存在，这可以在下面求法中直接得到结论。

1. 消去多余参数法

例4.（平滑问题）天文测量，三个角度的测量值分别为y_1、y_2、y_3，模型为

$$\begin{cases} y_1 = \theta_1 + e_1 \\ y_2 = \theta_2 + e_2 \\ y_3 = \theta_3 + e_3 \end{cases}$$

第三节 线性模型的参数估计

其中e_1、e_2、e_3相互独立，$Ee_i=0$，$Ee_i^2=\sigma^2$，约束条件为

$$\theta_1+\theta_2+\theta_3=\pi$$

即$(1,1,1)(\theta_1,\theta_2,\theta_3)'=\pi$。求$\theta_1$、$\theta_2$、$\theta_3$的满足约束条件的LSE。

解：因为$\theta_1=\pi-\theta_2-\theta_3$，代入模型得

$$\begin{cases} y_1=\pi-\theta_2-\theta_3+e_1 \\ y_2=\theta_2+e_2 \\ y_3=\theta_3+e_3 \end{cases}$$

此时对参数θ_2、θ_3无约束条件，新模型为：

$$\widetilde{Y}=\widetilde{X}\begin{pmatrix}\theta_2\\ \theta_3\end{pmatrix}+e$$

其中$\widetilde{Y}=(y_1-\pi,y_2,y_3)'$，$\widetilde{X}=\begin{pmatrix}-1 & -1\\ 1 & 0\\ 0 & 1\end{pmatrix}$，因此，

第三节 线性模型的参数估计

$$\tilde{X}'\tilde{X} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$(\tilde{X}'\tilde{X})^{-1}\tilde{X}' = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

故

$$\begin{pmatrix} \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix} = \begin{pmatrix} y_2 + \frac{\pi}{3} - \frac{1}{3}(y_1 + y_2 + y_3) \\ y_3 + \frac{\pi}{3} - \frac{1}{3}(y_1 + y_2 + y_3) \end{pmatrix}$$

再代回约束条件，得

$$\hat{\theta}_1 = y_1 + \frac{\pi}{3} - \frac{1}{3}(y_1 + y_2 + y_3)$$

● 此时，易知$E\hat{\theta}_i = \theta_i$，$var(\hat{\theta}_i) = \frac{2}{3}\sigma^2 < \sigma^2 = var(y_i)$。我们不仅使得估计满足了约束条件，还减小了估计的方差（实际上估计角度仅测两个就足够，测三个相当于增加了一个样本）。

第三节 线性模型的参数估计

一般的方法形式上如下：

① 设H的前s列线性无关，则$H=(H_1H_2)$，H_1为$s\times s$可逆方阵。

② 记$\beta=\begin{pmatrix}\beta^{(1)}\\ \beta^{(2)}\end{pmatrix}$，则当$\beta\in\Theta_0$时，约束条件变为$H\beta=H_1\beta^{(1)}+H_2\beta^{(2)}=r_0$，即$\beta^{(1)}=H_1^{-1}r_0-H_1^{-1}H_2\beta^{(2)}$。

③ 代入原线性模型，得新的线性模型$\tilde{Y}=\tilde{X}\beta^{(2)}+e$，其中$\tilde{Y}=Y-X\begin{pmatrix}H_1^{-1}\\ 0\end{pmatrix}$，$\tilde{X}=X\begin{pmatrix}-H_1^{-1}H_2\\ I\end{pmatrix}$。

④ 对新的线性模型（无约束）求$\beta^{(2)}$的LSE，代回得$\beta^{(1)}$的LSE，即得β的带约束条件的LSE。

●β的带约束条件的LSE可能不唯一。

第三节 线性模型的参数估计

2. 拉格朗日乘子法

数学上著名的拉格朗日乘子法在此问题上可直接应用。

定理3.6. $\widehat{\beta}$是约束条件下LSE的充要条件是，存在$\lambda \in R^s$使得$\widehat{\beta}$满足

$$\begin{cases} X'X\widehat{\beta} - H'\lambda = X'Y \\ H\widehat{\beta} = r_0 \end{cases}$$

证明：“$\Rightarrow$”（必要性）：显然$H\widehat{\beta} = r_0$。

取t为任一实数，b为垂直于H的所有s个行向量的p维向量，即

$$b \perp \mu(H')$$

令$\beta = \widehat{\beta} - tb$，则

$$H\beta = H\widehat{\beta} - tHb = H\widehat{\beta} = r_0$$

即β满足约束条件。

第三节 线性模型的参数估计

由于$\hat{\beta}$是带约束的LSE，故

$$\| Y - X\hat{\beta} \|^2 \leq \| Y - X\beta \|^2 = \| Y - X\hat{\beta} + X(\hat{\beta} - \beta) \|^2$$

即

$$\| Y - X\beta \|^2 - \| Y - X\hat{\beta} \|^2 = \| X(\hat{\beta} - \beta) \|^2 + 2(Y - X\hat{\beta})' X(\hat{\beta} - \beta)$$
$$= t^2 \| Xb \|^2 + 2t(Y - X\hat{\beta})' Xb \geq 0$$

由t的任意性，$\Longrightarrow$

$$(Y - X\hat{\beta})' Xb = 0$$

即

$$(X'Y - X'X\hat{\beta})' b = 0$$

再由$b \perp \mu(H')$的任意性，$\Longrightarrow$

$$X'Y - X'X\hat{\beta} \in \mu(H')$$

故存在$\lambda \in R^s$，使得

$$X'X\hat{\beta} - X'Y = H'\lambda$$

必要性得证。

第三节 线性模型的参数估计

“⇐”（充分性）：任给β满足约束条件$H\beta = r_0$，

$$
\begin{aligned}
\| Y - X\beta \|^2 &= \| \left(Y - X\widehat{\beta}\right) + X\left(\widehat{\beta} - \beta\right) \|^2 \\
&= \| Y - X\widehat{\beta} \|^2 + \| X\left(\widehat{\beta} - \beta\right) \|^2 + 2\left(Y - X\widehat{\beta}\right)' X\left(\widehat{\beta} - \beta\right) \\
&= \| Y - X\widehat{\beta} \|^2 + \| X\left(\widehat{\beta} - \beta\right) \|^2 + 2\left(X'Y - X'X\widehat{\beta}\right)' \left(\widehat{\beta} - \beta\right) \\
&= \| Y - X\widehat{\beta} \|^2 + \| X\left(\widehat{\beta} - \beta\right) \|^2 + 2(-\lambda' H)\left(\widehat{\beta} - \beta\right) \\
&= \| Y - X\widehat{\beta} \|^2 + \| X\left(\widehat{\beta} - \beta\right) \|^2 - 2\lambda'(r_0 - r_0) \\
&= \| Y - X\widehat{\beta} \|^2 + \| X\left(\widehat{\beta} - \beta\right) \|^2 \geq \| Y - X\widehat{\beta} \|^2
\end{aligned}
$$

故$\widehat{\beta}$是约束条件下的LSE。证毕。

对平滑问题，若用拉格朗日乘子法解，则$X = I_3$，$X'X\widehat{\beta} = I_3\widehat{\beta} = \widehat{\beta}$，$X'Y = Y$，$H'\lambda = (1,\ 1,\ 1)'\lambda = (\lambda, \lambda, \lambda)'$，由定理3.6，

第三节 线性模型的参数估计

$$\begin{cases}\hat{\theta}_1-\lambda=y_1\\ \hat{\theta}_2-\lambda=y_2\\ \hat{\theta}_3-\lambda=y_3\\ \hat{\theta}_1+\hat{\theta}_2+\hat{\theta}_3=\pi\end{cases}$$

得到$\hat{\theta}_i=y_i+\frac{\pi}{3}-\frac{1}{3}(y_1+y_2+y_3)$，$i=1,2,3$。

五、关于LSE的讨论

● 缺点1，X不满秩，或接近不满秩，但因观测误差，$|X'X|\approx 0$。解决方法：① 除去多余自变量，使X满秩（检验）；② 利用岭回归估计 $\tilde{\beta}=(X'X+\lambda I)^{-1}X'Y$。

●不稳健：使用其他方法，如最小一乘估计，M估计等。

第四节 线性模型的假设检验

一、参数线性相关性的检验

参数线性相关，即存在非零的向量$h \in R^p$，使得$h'\beta = 0$。由于可能同时检验多个线性关系，一般表示为：

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

其中H为$s \times p$矩阵。因为此后要做检验（及构造置信区间），我们假设假定B成立。

H的形式可以有多种，表示各种可能的假设。但最典型的情况如下：

H的第一行是$(0, 1, 0, \cdots)$，则H_0表示待检验的假设是：协变量x_2的回归系数为0， x_2可从模型中删去。

第四节 线性模型的假设检验

若H的第一行是$(0,\ 1,\ -1,\ 0,\ \cdots)$，则$H_0$表示待检验的假设是：协变量$x_2$、$x_2$的回归系数相等 …

记$W=\mu(X)$，则其维数为r，又记$W_0=\{\eta=X\beta\in W,\ H\beta=0\}\subset W$，设其秩为$0<q<r$，再记$\Theta=\{(\beta,\ \sigma^2):\ \beta\in R^p,\sigma^2>0\}$，$\Theta_0=\{(\beta,\ \sigma^2):\ \beta\in R^p,H\beta=0,\ \sigma^2>0\}$，又记$\hat{\beta}$、$\hat{\beta}_0$分别为$\beta$的无约束条件下和约束条件$H\beta=0$下的LSE（可以不唯一），则易知，

$$\hat{\xi}=Proj_W(Y)=X\hat{\beta}$$
$$\hat{\xi}_0=Proj_{W_0}(Y)=X\hat{\beta}_0$$

利用广义似然比检验法，似然函数为

$$L(Y,\ \beta,\ \sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\|Y-X\beta\|^2}$$

第四节 线性模型的假设检验

易知L在Θ与Θ_0上的最大值分别在$\left(\hat{\beta},\ \frac{1}{n}\| Y-X\hat{\beta}\|^2\right)=\left(\hat{\beta},\ \frac{1}{n}\| Y-\hat{\xi}\|^2\right)$与$\left(\hat{\beta}_0,\frac{1}{n}\| Y-X\hat{\beta}_0\|^2\right)=\left(\hat{\beta}_0,\frac{1}{n}\| Y-\hat{\xi}_0\|^2\right)$上达到，代入后得到广义似然比为：

$$\lambda=\frac{sup_{\Theta}L\left(Y,\ \beta,\ \sigma^2\right)}{sup_{\Theta_0}L\left(Y,\ \beta,\ \sigma^2\right)}=\cdots=\left(\frac{\| Y-\hat{\xi}_0\|^2}{\| Y-\hat{\xi}\|^2}\right)^{\frac{n}{2}}$$

注意到

$$Y-\hat{\xi}_0=Y-X\hat{\beta}_0=Y-\hat{\xi}+\hat{\xi}-\hat{\xi}_0=Y-X\hat{\beta}+X\hat{\beta}-X\hat{\beta}_0$$

$$Y-X\hat{\beta}\perp\mu(X),\ \ X\hat{\beta}-X\hat{\beta}_0\in\mu(X)$$

故

$$\| Y-\hat{\xi}_0\|^2=\| Y-\hat{\xi}\|^2+\| \hat{\xi}-\hat{\xi}_0\|^2$$

第四节 线性模型的假设检验

又记

$$F=\frac{\|\hat{\xi}-\hat{\xi}_0\|^2/(r-q)}{\|Y-\hat{\xi}\|^2/(n-r)}=\frac{\|X\hat{\beta}-X\hat{\beta}_0\|^2/(r-q)}{\|Y-X\hat{\beta}\|^2/(n-r)}$$

则

$$\lambda=\left(1+\frac{\|\hat{\xi}-\hat{\xi}_0\|^2}{\|Y-\hat{\xi}\|^2}\right)^{\frac{n}{2}}=\left(1+\frac{(r-q)}{(n-r)}F\right)^{\frac{n}{2}}$$

是F的单调增函数，故广义似然比检验为

$$\varphi_0=\begin{cases}1 & \text{当}F>C\\ 0 & \text{否则}\end{cases}$$

如何确定常数C？下面有定理。

定理4.1. 设假定B成立，则当H_0成立时， F的分布为$F(r-q,\ n-r)$。

证明： （典型方法，多维正态做正交变换）

第四节 线性模型的假设检验

取$\xi_1, \cdots, \xi_q$为W_0的一组标准正交基，添加$\xi_{q+1}, \cdots, \xi_r$使之成为W的一组标准正交基，再添加$\xi_{r+1}, \cdots, \xi_n$成为R^n的一组标准正交基。记

$$V_1 = (\xi_1 \cdots \xi_q)_{n\times q} \qquad V_2 = (\xi_{q+1} \cdots \xi_n)_{n\times(n-q)}$$

$$V = (V_1 V_2)_{n\times n}$$

则V为正交矩阵。做正交变换

$$Z = (Z_1, \cdots, Z_n)' = V'Y$$

则

$$Y = VZ = \sum_{i=1}^{n} Z_i \xi_i$$

故而

$$\hat{\xi}_0 = \sum_{i=1}^{q} Z_i \xi_i$$

$$\hat{\xi} = \sum_{i=1}^{r} Z_i \xi_i$$

$$Y - \hat{\xi} = \sum_{i=r+1}^{n} Z_i \xi_i$$

所以，

第四节 线性模型的假设检验

$$F=\frac{\|\sum_{i=q+1}^{r}Z_i\xi_i\|^2/(r-q)}{\|\sum_{i=r+1}^{n}Z_i\xi_i\|^2/(n-r)}=\frac{\sum_{i=q+1}^{r}Z_i^2/(r-q)}{\sum_{i=r+1}^{n}Z_i^2/(n-r)}$$

因为$Z=V'Y$，故

$$Z\sim N(V'\xi,\ \sigma^2V'V)=N(V'\xi,\ \sigma^2I_n)$$

又因为

$$V'\xi=\begin{pmatrix}V_1'\\V_2'\end{pmatrix}\xi=\begin{pmatrix}V_1'\xi\\0\end{pmatrix}$$

（因为当H_0成立时，$\xi\in W_0$，与$\xi_{q+1},\ ,\cdots,\ \xi_n$正交），

所以当H_0成立时，

$$\frac{1}{\sigma^2}\sum_{i=q+1}^{r}Z_i^2\sim\chi^2\ （r-q）$$

$$\frac{1}{\sigma^2}\sum_{i=r+1}^{n}Z_i^2\sim\chi^2\ （n-r）$$

且相互独立，故$F\sim F(r-q,\ n-r)$。证毕。

第四节 线性模型的假设检验

若$\hat{\beta}$为β的LSE，则$\hat{\xi}=X\hat{\beta}$，$Y-\hat{\xi}=Y-X\hat{\beta}$与$\hat{\xi}=X\hat{\beta}$独立。

称$\hat{e}=Y-X\hat{\beta}$为残差，称

$$Q=\|Y-X\hat{\beta}\|^2=\|\hat{e}\|^2=\hat{e}'\hat{e}=\sum_{i=1}^{n}\left\{y_i-\left(x_{i1}\hat{\beta}_1+\cdots+x_{ip}\hat{\beta}_p\right)\right\}^2$$

为残差平方和。

推论. $X\hat{\beta}$与$\hat{e}$相互独立，从而$X\hat{\beta}$与Q相互独立，且$Q/\sigma^2\sim\chi^2(n-r)$。

证明：从定理4.1证明可知，无论H_0成立与否，总有

$$Z\sim N\left(V'\xi,\ \sigma^2 I_n\right)$$

故独立性得证。又总有$\xi_{r+1},\cdots,\xi_n$与$\hat{\xi}$正交，，故$Z_{r+1},\cdots,Z_n$独立同分布$N(0,\ \sigma^2)$，因此

$$Q/\sigma^2=\frac{1}{\sigma^2}\sum_{i=r+1}^{n}Z_i^2\sim\chi^2(n-r)$$

证毕。

第四节 线性模型的假设检验

● 不论H_0是否成立，F的分母总是σ^2的无偏估计。

二、参数线性组合的检验与置信区间

在“一”中，我们讨论了检验β的某些线性组合是否为0的方法。如何检验β的某个线性组合是否为某已知常数（不一定为0）？特别地，能否检验某β_i为某已知的值？如何构造β_i的置信区间？上面的“个”可否改为“些”？这都对应什么样的实际问题？下面进行一般性的讨论。

定理4.2. 设$\hat{\beta}$是β的（一个）LSE，$c'\beta$线性可估，则$c'\hat{\beta}$与Q相互独立。

证明：如前，取$\xi_1, \cdots, \xi_r$为W的一组标准正交基，添加$\xi_{r+1}, \cdots, \xi_n$成为R^n的一组标准正交基。记V_1、V_2、V如前，做正交变换

$$Z = V'Y$$

则

$$Z \sim N(V'X\beta, \sigma^2 I)$$

第四节 线性模型的假设检验

因为$c'\beta$线性可估，故存在$a \in \mu(X)$，使得$c'\hat{\beta} = a'Y$。因此

$$c'\hat{\beta} = a'VZ = a'(V_1\ V_2)Z = (a'V_1\ 0)Z = a'V_1(Z_1, \cdots, Z_r)'$$

又因为

$$Q = \hat{e}'\hat{e} = \sum_{i=r+1}^{n} Z_i^2$$

且$Z_1, \cdots, Z_n$相互独立，故$c'\hat{\beta}$与Q相互独立。证毕。

推论. 如果X满秩，则$\hat{\beta}$与Q相互独立。

证明：当X满秩时，对任意的$c \in R^p$， $c'\beta$线性可估。由上述证明，$c'\hat{\beta}$可以表示成$Z_1, \cdots, Z_r$的线性组合。再由c的任意性，可知$\hat{\beta}$与Q相互独立。

此时，

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$
$$Q/\sigma^2 \sim \chi^2(n-p)$$

---------- end 20240418

第四节 线性模型的假设检验

因为$c'\beta$线性可估，故存在$a \in \mu(X)$，使得$c'\hat{\beta} = a'Y$。因此

$$c'\hat{\beta} = a'VZ = a'(V_1\ V_2)Z = (a'V_1\ 0)Z = a'V_1(Z_1, \cdots, Z_r)'$$

又因为

$$Q = \hat{e}'\hat{e} = \sum_{i=r+1}^{n} Z_i^2$$

且$Z_1, \cdots, Z_n$相互独立，故$c'\hat{\beta}$与Q相互独立。证毕。

推论. 如果X满秩，则$\hat{\beta}$与Q相互独立。

证明：当X满秩时，对任意的$c \in R^p$， $c'\beta$线性可估。由上述证明，$c'\hat{\beta}$可以表示成$Z_1, \cdots, Z_r$的线性组合。再由c的任意性，可知$\hat{\beta}$与Q相互独立。

此时，

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$
$$Q/\sigma^2 \sim \chi^2(n-p)$$

第四节 线性模型的假设检验

● 在试图构造$c'\beta$的置信区间时，自然想到利用$c'\hat{\beta}-c'\beta=c'(\hat{\beta}-\beta)$，它服从零均值正态分布，但其方差含未知参数$\sigma^2$，以及与$c$相关的常数。我们下面应该

① 用适当的估计代替σ^2；

② 找出与c相关的常数。

● 当X满秩时，存在$a\in\mu(X)$，使得$c'\hat{\beta}=a'Y$，且$a=X(X'X)^{-1}c$。

定义1. 设$a\in\mu(X)$，$a'Y$是$c'\beta$的无偏估计，则称a为c的伴随元。

● 在条件$a\in\mu(X)$下，由定理3.4的证明， a存在唯一，且$a'Y=c'\hat{\beta}$，即定义是完备的。

第四节 线性模型的假设检验

定理4.3. 设$c \neq 0$使得$c'\beta$可估，记$\hat{\sigma}^2 = Q/(n-r)$， a为c的伴随元，则

$$\frac{c'(\hat{\beta}-\beta)}{\hat{\sigma}\parallel a \parallel} \sim t(n-r)$$

其中$t(k)$表示k个自由度的t分布。（注意：$c'\beta = E(Y_0)$）

证明：$c'\hat{\beta} = a'Y$是正态分布线性组合，故仍服从正态分布。

$$E\left(c'\hat{\beta}\right) = c'\beta$$

$$var\left(c'\hat{\beta}\right) = var(a'Y) = a'cov(Y, Y)a = a'\sigma^2 Ia = \parallel a \parallel^2 \sigma^2$$

所以$c'\hat{\beta} \sim N(c'\beta, \parallel a \parallel^2 \sigma^2)$，即

$$\eta = \frac{c'(\hat{\beta}-\beta)}{\parallel a \parallel \sigma} \sim N(0, 1)$$

又η与Q相互独立，$Q/\sigma^2 \sim \chi^2(n-r)$，故

$$\frac{c'(\hat{\beta}-\beta)}{\hat{\sigma}\parallel a \parallel} = \frac{\eta/1}{\sqrt{Q/(n-r)\sigma^2}} \sim t(n-r)$$

证毕。

第四节 线性模型的假设检验

● 当X满秩时，$a = X(X'X)^{-1}c$，故

$$\| a \|^2 = a'a = c'(X'X)^{-1}X' \cdot X(X'X)^{-1}c = c'(X'X)^{-1}c$$

故定理4.3的结论变为

$$\frac{c'(\hat{\beta} - \beta)}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}} \sim t(n-p)$$

● 利用定理4.3，我们可以对可估的$c'\beta$做假设检验

$$H_0: c'\beta = r_0 \qquad (r_0\text{已知})$$

此时，检验统计量为

$$\frac{c'\hat{\beta} - r_0}{\hat{\sigma}\,\| a \|} \sim t(n-r)$$

● 利用定理4.3，我们还可以求可估的$c'\beta$的置信区间。皆为典型方法。

● Q: 定理4.1与定理4.3的关系与区别？F分布↔ t分布；H_0中多维↔一维；H_0中$= 0 \leftrightarrow = r_0$可$\neq 0$。当问题重合时，两方法是否一样？哪个好？

第四节 线性模型的假设检验

定理4.4. 设$c = x_0 = \left(x_1^{(0)}, \cdots, x_p^{(0)}\right)'$使得$x_0'\beta = c'\beta$可估，$Y_0 = c'\beta + e_0$，$e_0 \sim N(0, \sigma^2)$与$e$相互独立，$\hat{\sigma}^2 = Q/(n-r)$，$a$为$c$的伴随元，则

$$T = \frac{c'\hat{\beta} - Y_0}{\hat{\sigma}\sqrt{\| a \|^2 + 1}} \sim t(n-r)$$

证明：由定理4.3的证明，$c'\hat{\beta} \sim N(c'\beta, \| a \|^2\sigma^2)$，故

$$c'\hat{\beta} - Y_0 = c'\hat{\beta} - c'\beta - e_0 \sim N\left(0, \left(\| a \|^2 + 1\right)\sigma^2\right)$$

此时仍有

$$Q/\sigma^2 \sim \chi^2（n-r)$$

且与分子相互独立，故结论成立，证毕。

- 此定理可用于预测。

- 即使X不满秩，即$\hat{\beta}$不唯一，只要$x_0'\beta$可估，仍可预测$Y_0 = x_0'\beta + e_0$。

第四节 线性模型的假设检验

● 一元线性回归的假设检验：$H_0: b = 0$。若$x_1, \cdots, x_n$不全相等，则矩阵X满秩。

$$W_0 = \{\eta = X\beta,\ H\beta = 0,\ \beta = (a,\ b) \in R^2\} = \{(a, \cdots, a)',\ a \in R^1\}$$

$$W = \{\eta = X\beta,\ \beta \in R^2\}$$

$$r = p = 2,\ \ q = 1$$

$$\hat{\xi} = X\hat{\beta},\quad \hat{\xi}_0 = (\bar{y},\ \cdots,\ \bar{y})'$$

均为n维列向量。故而，

$$\| Y - \hat{\xi} \|^2 = Q$$

仍为残差平方和，

$$\| \hat{\xi} - \hat{\xi}_0 \|^2 = \| X\hat{\beta} - \hat{\xi}_0 \|^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = U$$

为回归平方和。所以由定理4.1，

$$F = \frac{U/(2-1)}{Q/(n-2)} \sim F(1,\ n-2)$$

第四节 线性模型的假设检验

● 一元线性回归的预测：当$x = x_0$，$\hat{Y}_0 = \hat{a} + \hat{b}x_0$。

此时$c' = (1, x_0)$，$r = p = 2$，由$a = X(X'X)^{-1}c$，不难计算，

$$\| a \|^2 = a' \cdot a = c'(X'X)^{-1}X' \cdot X(X'X)^{-1}c$$

$$= c'(X'X)^{-1}c = \frac{1}{|X'X|}(1, x_0)\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}\begin{pmatrix} 1 \\ x_0 \end{pmatrix}$$

$$= \cdots = \frac{\sum (x_i - x_0)^2}{nl_{XX}} = \frac{l_{XX} + \sum (x_0 - \bar{x})^2}{nl_{XX}} = \frac{1}{n} + \frac{\sum (x_0 - \bar{x})^2}{nl_{XX}}$$

故由定理4.4，记$d = 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}$，则

$$T = \frac{\hat{Y}_0 - Y_0}{\sqrt{dQ/(n-2)}} \sim t(n-2)$$

为一元回归结果。

第五节 回归分析

研究目的是根据数据探讨变量间的相关关系，而非科学原理探索，是经验性的。

本课程仅讨论一个响应变量的情形，自变量可有多个，称多元回归。

● “回归”由十九世纪英国遗传学家Galton开始使用，本意是，对给定的自变量，相应的响应变量期望向其均值“回归”。例如，高个父亲的儿子，平均身高大于群体平均，但小于其父亲；……

● 但这不仅是生物遗传问题，而更是统计问题！见下例。

● 以夫妻智商为例，设平均智商均为100，方差相等，相关系数大于0（例如0.5），散点图为（男x，女Y）：……，回归直线是否椭圆的长轴?

第五节 回归分析

● 注意：我们最小化的目标是散点到回归直线的竖直距离平方和。几何上看，回归直线不是椭圆长轴，而是……（草图）。

● 以此预测（如你智商$x=140$），你未来妻子的智商Y的预测结果如何？期望为120？

● 如果将妻子智商作为x，丈夫智商作为Y？

● 概率论分析：

若(X, Y)二维正态，则$(Y-\mu_2)|_{X=x} \sim N(\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \left(1-\rho^2\right)\sigma_2^2)$，例如$\mu_1=\mu_2=100$，$\sigma_1^2=\sigma_2^2=10^2$，$\rho=0.5$，则$Y|_{X=140} \sim N(120, 75)$。

第五节 回归分析

● 在线性模型$Y = X\beta + \varepsilon$中，回归使得Y“收敛”，ε使得Y“发散”。

● 回归应用：应慎重。不能随意将“相关”解释为“因果”，例如吸烟/肺癌（D. R. Cox）。

● 潜在变量：能否观测，能否加入模型，替代变量?

设数据为

$$\begin{matrix} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ & \cdots\cdots & & \\ y_n & x_{n1} & \cdots & x_{np} \end{matrix}$$

第五节 回归分析

一、回归系数估计

下面总假设X满秩，则LSE存在唯一，为

$$\hat{\beta} = (X'X)^{-1}X'Y$$

此时的回归方程为

$$\hat{Y} = X\hat{\beta}$$

从分析的角度，$\hat{\beta}$也可通过求（多元）二次函数

$$Q(\beta) = (Y - X\beta)'(Y - X\beta)$$

的最小值得到，即得到正规方程

$$(X'X)\beta = X'Y$$

若假设$Ee = 0$，$cov(e, e) = \sigma^2 I$，则$\hat{\beta}$是β的无偏估计；

若进一步假设$e \sim N(0, \sigma^2 I)$，则$\hat{\beta}$是β的MLE。

第五节 回归分析

● 常数项问题：（在需要时）总可以引入形式变量$x_{\cdot 1} \equiv 1$，则相应的β_1即为常数项。此时真正的自变量个数为$p-1$。

● σ^2的无偏估计为

$$\hat{\sigma}^2 = Q(\hat{\beta})/(n-p) = Q/(n-p)$$

注意分母应为n减去回归系数的总个数。教材中为$(n-p-1)$，因其常数项单列，实际含p个真实自变量，共$p+1$个回归系数。

二、检验

1. 回归系数的检验

有些自变量实际上与Y不相关，或相关性不显著，应考虑从回归模型中删除。其主要目的为：

第五节 回归分析

① 简化模型与计算，防止过拟合（例如，当用多项式拟合数据时，若数据满足一些一般性条件，则只要多项式次数足够高，残差可为0）；

② 得到对模型更好的解释（interpretation）；

③ 提高预测精度，构造置信区间时不出现过小的$(n-p)$。

一般地，可同时对回归系数中某k个皆为0的假设进行检验（$1 \le k < p$），即$H_0: \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_k} = 0$，其中$1 \le i_1 < \cdots < i_k \le p$，此时，

$$H = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & & 0 \\ & & \cdots & & & & \cdots & \\ 0 & & & \cdots & & 0 & 1 \quad \cdots & 0 \end{pmatrix}$$

其中……。假设检验问题为：

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

第五节 回归分析

由定理4.1，$F=\frac{\|\hat{\xi}-\hat{\xi}_0\|^2/(r-q)}{\|Y-\hat{\xi}\|^2/(n-r)}\sim F(r-q,\ n-r)$，

注意此时$r=p$，$q=p-k$（W_0的维数），$\|Y-\hat{\xi}\|^2=Q$，$\|\hat{\xi}-\hat{\xi}_0\|^2$?

命题5.1. $\|\hat{\xi}-\hat{\xi}_0\|^2=\hat{\beta}'H'[H(X'X)^{-1}H']^{-1}H\hat{\beta}$。

证明：设H_0成立下，即约束条件$H\beta=0$下的LSE为$\tilde{\beta}$，则

$\hat{\xi}_0=X\tilde{\beta}$，$\hat{\xi}=X\hat{\beta}$，$\hat{\xi}-\hat{\xi}_0=X(\hat{\beta}-\tilde{\beta})$。

由定理3.6（拉格朗日乘子法），存在$\lambda\in R^k$，使得

$$\begin{cases}X'X\tilde{\beta}-H'\lambda=X'Y\\ H\tilde{\beta}=0\end{cases}$$

故$X'X\tilde{\beta}-H'\lambda=X'Y=X'X\hat{\beta}$，即

$$X'X(\tilde{\beta}-\hat{\beta})=H'\lambda\implies(\tilde{\beta}-\hat{\beta})=(X'X)^{-1}H'\lambda$$

第五节 回归分析

左乘H，并注意到$H\tilde{\beta}=0$，故而

$$-H\hat{\beta}=H(X'X)^{-1}H'\lambda$$

因此，$\lambda=-\left[H(X'X)^{-1}H'\right]^{-1}H\hat{\beta}$，所以

$$\begin{aligned}\|\hat{\xi}-\hat{\xi}_0\|^2&=\left(\tilde{\beta}-\hat{\beta}\right)'X'X\left(\tilde{\beta}-\hat{\beta}\right)\\&=\left(\tilde{\beta}-\hat{\beta}\right)'H'\lambda\\&=-\hat{\beta}'H'\lambda\\&=\hat{\beta}'H'\left[H(X'X)^{-1}H'\right]^{-1}H\hat{\beta}\end{aligned}$$

证毕。

由此命题，检验方法为：

① 对于给定的检验水平α，查表（$F(k,n-p)$）得临界值C；

② 计算检验统计量F的值；

③ 当$F\le C$时不否定H_0，则自变量$(x_{i_1},\cdots,x_{i_k})$可以从回归模型中删除；

④ 当$F>C$时否定H_0，此时……（因为H（即H_0）有多种取法，后面讲）。

第五节 回归分析

实际应用中，$\| \hat{\xi} - \hat{\xi}_0 \|^2$可以使用另一种更直观的计算方法：

$$\| \hat{\xi} - \hat{\xi}_0 \|^2 = U - \tilde{U} = \| Y - \hat{\xi}_0 \|^2 - \| Y - \hat{\xi} \|^2 = \tilde{Q} - Q$$

Q是已知的残差平方和， 即检验统计量F 的分母，$\tilde{Q} = \| Y - \hat{\xi}_0 \|^2$也是残差平方和，但对应的模型为：

响应变量Y对自变量$\{x_1, \cdots, x_p\} - \{x_{i_1}, \cdots, x_{i_k}\}$的线性回归。

计算方法：对上述模型拟合，得到新残差平方和$\tilde{Q}$。

解释： Q是Y的总变差中，自变量$\{x_1, \cdots, x_p\}$解释不了的随机误差部分。

$\tilde{Q}$是Y的总变差中，自变量$\{x_1, \cdots, x_p\} - \{x_{i_1}, \cdots, x_{i_k}\}$解释不了的随机误差部分。（总有$\tilde{Q} \geq Q$）

所以$\tilde{Q} - Q$是Y的总变差中，已有自变量$\{x_1, \cdots, x_p\} - \{x_{i_1}, \cdots, x_{i_k}\}$后，再添加$\{x_{i_1}, \cdots, x_{i_k}\}$，能够解释的部分。

第五节 回归分析

若$\widetilde{Q}-Q$相对较小，说明当已有$\{x_1,\cdots,x_p\}-\{x_{i_1},\cdots,x_{i_k}\}$时，再添加$\{x_{i_1},\cdots,x_{i_k}\}$，并不能对$Y$的变化给出显著更好的解释，故认为$\{x_{i_1},\cdots,x_{i_k}\}$可以删除。

已有非常重要，例如研究体重↔进食热量、进食重量等关系时，已考虑进食热量，则进食重量可删除；未考虑进食热量，则进食重量不可删除。

“相对”指相对于$Q/(n-p)$，即检验统计量F的分母。是否“较小”？与临界值比较。

当自变量个数p较大时，$\{x_{i_1},\cdots,x_{i_k}\}$的可能选择非常多。且各种检验结果可能不一致。

实用中，可使用如下建模方法。

---------- end 20240425

第五节 回归分析

若$\widetilde{Q}-Q$相对较小，说明当已有$\{x_1,\cdots,x_p\}-\{x_{i_1},\cdots,x_{i_k}\}$时，再添加$\{x_{i_1},\cdots,x_{i_k}\}$，并不能对$Y$的变化给出显著更好的解释，故认为$\{x_{i_1},\cdots,x_{i_k}\}$可以删除。

已有非常重要，例如研究体重↔进食热量、进食重量等关系时，已考虑进食热量，则进食重量可删除；未考虑进食热量，则进食重量不可删除。

“相对”指相对于$Q/(n-p)$，即检验统计量F的分母。是否“较小”？与临界值比较。

当自变量个数p较大时，$\{x_{i_1},\cdots,x_{i_k}\}$的可能选择非常多。且各种检验结果可能不一致。

实用中，可使用如下建模方法。

第五节 回归分析

第一步：选取检验水平$0 < \alpha < 1$，先对所有（真实的）自变量$\{x_2, \cdots, x_p\}$拟合回归模型$Y = X\beta + \varepsilon$。此时我们默认$x_1 \equiv 1$为常数项。

第二步：之后取$k = 1$，对所有（真实的）变量x_j，分别做假设检验$H_{0j}: \beta_j = 0$，并由此得到（$p-1$个）检验统计量$F_2, \cdots, F_p$。

第三步：在$F_2, \cdots, F_p$中选取最小的一个，与$F(1, n-p)$的临界值比较，若不显著，则将相应自变量从回归模型中删除。

第四步：重复以上步骤（此时少了一个自变量），直至无自变量可删除。

● 三点需要注意：

① 此方法先删除的可能是“显著性”较好，但可被其它变量替代的变量。

第五节 回归分析

② 删除某自变量后，新的回归方程必须重新拟合，所有回归系数的估计都可能出现变化。这是因为自变量之间往往线性相关。

③ 不是一次删除多个变量，而是每次仅删除一个最不显著的。

● 教材中第六节*详细介绍了“逐步回归”，与上过程相反，且更复杂，即从零个变量的模型开始，重复“添加/删除”过程。（实用中问题很多…，可替代性）

● 还有一种检验

$$H_0: \beta_2 = \cdots = \beta_p = 0$$

若不否定H_0，则此项回归分析终止。

建模结束后，不妨仍设自变量个数为p，线性回归模型为$Y = X\beta + \varepsilon$，拟合的回归方程为

$$\hat{Y} = \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p$$

第五节 回归分析

例1. 用下列数据拟合模型$Y=b_0+b_1x_1+b_2x_2+\varepsilon$，并检验显著性。

$\sum_{i=1}^{50}y_i=50,\quad \sum_{i=1}^{50}x_{1i}=50,\quad \sum_{i=1}^{50}x_{2i}=100,$

$\sum_{i=1}^{50}y_i^2=100,\quad \sum_{i=1}^{50}x_{1i}^2=75,\quad \sum_{i=1}^{50}x_{2i}^2=250,$

$\sum_{i=1}^{50}x_{1i}y_i=80,\quad \sum_{i=1}^{50}x_{2i}y_i=120,\quad \sum_{i=1}^{50}x_{1i}x_{2i}=125。$

解：$n=50,\ X'X=\begin{pmatrix}50&50&100\\50&75&125\\100&125&250\end{pmatrix}=25\begin{pmatrix}2&2&4\\2&3&5\\4&5&10\end{pmatrix},$

所以$(X'X)^{-1}=\frac{1}{25}\begin{pmatrix}5/2&0&-1\\0&2&-1\\-1&-1&1\end{pmatrix},\ X'Y=\begin{pmatrix}50\\80\\120\end{pmatrix}。$

因此，$\hat{\beta}=\begin{pmatrix}\hat{b}_0\\\hat{b}_1\\\hat{b}_2\end{pmatrix}=\frac{1}{25}\begin{pmatrix}5/2&0&-1\\0&2&-1\\-1&-1&1\end{pmatrix}\begin{pmatrix}50\\80\\120\end{pmatrix}=\begin{pmatrix}0.2\\1.6\\-0.4\end{pmatrix},$

故回归方程为：$\hat{Y}=0.2+1.6x_1-0.4x_2。$

第五节 回归分析

下面做检验。

（一）、$H_0: b_1 = b_2 = 0$

$$Q = Y'Y - Y'X(X'X)^{-1}X'Y = 100 - (50 \quad 80 \quad 120)\begin{pmatrix} 0.2 \\ 1.6 \\ -0.4 \end{pmatrix} = 100 - 90 = 10$$

$$l_{YY} = \sum_{i=1}^{50}(Y_i - \overline{Y})^2 = \sum_{i=1}^{50} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{50} y_i\right)^2 = 100 - \frac{50^2}{50} = 50$$

所以$\| \hat{\xi} - \hat{\xi}_0 \|^2 = U = l_{YY} - Q = 40$（此处也可利用命题5.1），相应的平方和分解公式为$l_{YY} = 50 = U + Q = 40 + 10$。故而

$F = \frac{40/2}{10/47} = 94$，而$F(2, 47) < 3.23$，否定$H_0$，整体回归显著。

（二）、$H_0': b_2 = 0$

对模型$Y = b_0 + b_1 x_1 + \varepsilon$进行拟合，得到（一元回归）

$\hat{b}_1' = \cdots = 1.2$，$\hat{b}_0' = \cdots = -0.2$，故回归方程为

$$\hat{Y} = -0.2 + 1.2x_1$$

第五节 回归分析

此时易知，新的回归平方和为$U_1 = \cdots = 36$，新的平方和分解公式为：

$$l_{YY} = 50 = U_1 + Q_1 = \cdots = 36 + 14$$

故如前所求，$\| \hat{\xi} - \hat{\xi}_0 \|^2 = Q_1 - Q = 14 - 10 = 4$。因此，

$F' = \frac{4/1}{10/47} = 18.8$，故否定$H_0'$，即$x_2$不能删除。

（三）、$\widetilde{H}_0: b_1 = 0$

对模型$Y = b_0 + b_2 x_2 + \varepsilon$进行拟合，得到（一元）回归方程为

$$\hat{Y} = 0.2 + 0.4 x_2$$

此时的回归平方和为$U_2 = \cdots = 8$，新的平方和分解公式为：

$$l_{YY} = 50 = U_2 + Q_2 = \cdots = 8 + 42$$

故而，$\| \hat{\xi} - \hat{\xi}_0 \|^2 = Q_2 - Q = 42 - 10 = 32$。因此，

$\widetilde{F} = \frac{32/1}{10/47} = 150.4$，故否定$\widetilde{H}_0$，即$x_1$不能删除。

最后，回归方程为：$\hat{Y} = 0.2 + 1.6 x_1 - 0.4 x_2$。

第五节 回归分析

● 三个平方和分解公式中，$\{x_1, x_2\}$可以解释$l_{YY} = 50$中的40；$\{x_1\}$可以解释$l_{YY} = 50$中的36；$\{x_2\}$可以解释$l_{YY} = 50$中的8。注意到$36 + 8 > 40$。

● 每次（试图）删除自变量时，回归系数都可能变化。

● 因为未删除任何变量，检验统计量的分母始终为$10/47$。如果能删除，则下一轮的检验中，检验统计量的分母也会变化。

● x_1, x_2间的相关性可以导致，当试图删除x_1时，x_2系数的正负号都出现了变化。x_2对Y的影响是正面/负面？（设想Y为体重，x_1为进食热量，x_2为进食重量，若数据中，x_2大的多进食蔬菜，当x_1存在时，x_2的系数可能为负：能否说多吃有助减肥？）

● 请大家自行利用命题5.1完成上述检验（有挑战性）。

第五节 回归分析

2*. 线性性检验与残差分析

线性模型的一些假设可能是错误的，包括：模型不是线性的；误差不服从正态分布，等等。

- 线性性检验的目的及与回归系数为0检验的关系：4个图（画图）。

- 当线性性检验通过后，才能检验回归系数是否为0。

如果自变量X存在一些重复，则可以对线性性进行检验：

设X共有h种不同的取值$X_{(i)}$（组合），其中第i种取值时Y_i的期望为μ_i，

则H_0: $\mu_i = X_{(i)}\beta\ (i = 1, \cdots, h)$。

第五节 回归分析

记$X = X_{(i)}$时，对Y共有n_i次观测$Y_{ij}\ (j = 1, \cdots, n_i)$，又记$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$，则可以证明，

$$Q = \sum_{i=1}^{h}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{\mu}_i\right)^2 = \cdots$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 + \sum_{i=1}^{h} n_i(\overline{Y}_{i\cdot} - \hat{\mu}_i)^2 =: Q_1 + Q_2$$

其中$\hat{\mu}_i$为按照线性模型对μ_i的拟合值（画图）。

Q_1：真实随机误差的描述，与（线性）模型是否成立无关；

Q_2：数据偏离线性模型程度的刻画。

当H_0成立时，可以证明，

$$F = \frac{Q_2/(h-p)}{Q_1/(n-h)} \sim F(h-p,\ n-h)$$

查表得临界值C，当$F > C$时否定H_0。

第五节 回归分析

定义：设$\hat{\beta}$是线性回归模型$Y = X\beta + \varepsilon$的LSE，记$\hat{Y} = X\hat{\beta}$，则称$\hat{e}_i = Y_i - \hat{Y}_i$为残差。

对$\hat{e} = (\hat{e}_1, \cdots, \hat{e}_n)'$的分析称残差分析，主要内容包括是否存在异常值、正态性假设是否正确、线性假设是否成立、方差相等假设是否正确、残差间是否相关等。

图示方法（常对一个自变量）：对某个x_i画出残差散点图，目测上述假设是否成立。例如：……

各种检验法较多，典型的如“学生化残差”方法：

记p_{ii}为正定矩阵$X(X'X)^{-1}X'$主对角线上的第i个元素，令$\hat{\sigma} = \sqrt{Q/(n-p)}$，

$$\gamma_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-p_{ii}}} \qquad (i = 1, \cdots, n)$$

第五节 回归分析

称上述γ_i为学生化残差。

可以证明，若n较大，$\gamma_1, \cdots, \gamma_n$近似地相互独立，近似服从标准正态分布。

设H_0: $\varepsilon_1, \cdots, \varepsilon_n$相互独立，共同分布为$N(0, \sigma^2)$。计算学生化残差，如果某个学生化残差过大，或其绝对值大于2的个数太多（如大于5%），则否定H_0。

例2. （P203，例5.5）

● 上述方法还可用于找出“异常值”（outlier）。如例2中的z_5。

第五节 回归分析

三、预测

当回归方程$\widehat{Y} = X\widehat{\beta}$已建立，并已通过各种检验后，可利用其进行预测。

设$x_0' = c' = (x_{01}, \cdots, x_{0p})$已知，问相应的$Y_0$ =?

记$Y_0 = \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon_0$，其中$\varepsilon_0$仍服从模型假设，即$E\varepsilon_0 = 0$，$\varepsilon_0$与$\varepsilon_1, \cdots, \varepsilon_n$相互独立，$\varepsilon_0 \sim N(0, \sigma^2)$。

Y_0的点估计（预测）：$\widehat{Y}_0 = \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p} = x_0'\widehat{\beta}$。由高斯-马尔科夫定理，它是$EY_0 = \beta_1 x_{01} + \cdots + \beta_p x_{0p} = c'\beta$的最小方差线性无偏估计。

$EY_0 = c'\beta$的置信区间：

第五节 回归分析

$$t=\frac{c'\hat{\beta}-c'\beta}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}}\sim t(n-p)$$

即$\frac{\hat{Y}_0-EY_0}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}}$服从$n-p$个自由度的$t$分布。给定置信水平$1-\alpha$，查表得临界值$\lambda$，则$EY_0$的水平为$1-\alpha$的置信区间为：

$$\left[\hat{Y}_0-\lambda\hat{\sigma}\sqrt{c'(X'X)^{-1}c},\hat{Y}_0+\lambda\hat{\sigma}\sqrt{c'(X'X)^{-1}c}\right]$$

最后，我们求Y_0的置信区间。注意到$Y_0=EY_0+\varepsilon_0$，

$$\hat{Y}_0-Y_0=\hat{Y}_0-EY_0-\varepsilon_0=c'\hat{\beta}-c'\beta-\varepsilon_0$$

仍为正态分布，与$\hat{Y}_0-EY_0$相比仅多出$-\varepsilon_0$一项，且注意到$-\varepsilon_0$与$(Y_1,\cdots,Y_n)$相互独立，故与$\hat{\sigma}^2$独立，因此$\hat{Y}_0-Y_0$与$\hat{\sigma}^2$独立。

由定理4.4，易知：

第五节 回归分析

$$T=\frac{\hat{Y}_0-Y_0}{\hat{\sigma}\sqrt{1+c'(X'X)^{-1}c}}\sim t(n-p)$$

其中分母根号中多“1”是因为分子多了$-\varepsilon_0$。

查表得临界值λ，则Y_0的水平为$1-\alpha$的置信区间为：

$$\left[\hat{Y}_0-\lambda\hat{\sigma}\sqrt{1+c'(X'X)^{-1}c},\hat{Y}_0+\lambda\hat{\sigma}\sqrt{1+c'(X'X)^{-1}c}\right]$$

四、控制

提法仍为找自变量X的取值范围，使相应的Y以不小于$1-\alpha$的概率落在$[A,\ B]$中。

因X多元，控制范围的表示一般较复杂。

常附加对X的其他要求（因成本等），成为优化问题。

第五节 回归分析

例3. （P192，例5.2）。

五、二值响应变量回归

设Y仅取0或1（成功或失败，称成败型数据），其取1的概率与X有关。

1. Logistic回归模型

$$P(Y=1|X)=logit(X\beta)=\frac{exp(X\beta)}{1+exp(X\beta)}$$

$$P(Y=0|X)=1-logit(X\beta)=\frac{1}{1+exp(X\beta)}$$

$$log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)=X\beta$$

即优比（odds ratio，$p/(1-p)$）的对数为自变量的线性函数。

第五节 回归分析

参数估计方法：常求MLE，似然函数为（记$P(Y=1|X=x_i)=p(x_i)$，相应的响应变量为y_i）

$$L(\beta)=\prod_{i=1}^{n} p(x_i)^{y_i}\left(1-p(x_i)\right)^{1-y_i}$$

故取对数，求导，得估计方程（组）。

一定条件下，MLE存在唯一。

2. Probit回归模型

$$P(Y=1|X)=\Phi(X\beta)$$
$$P(Y=0|X)=1-\Phi(X\beta)$$

其中$\Phi(\cdot)$为标准正态分布函数。

类似地，可以写出似然函数，并得到MLE。

第五节 回归分析

称$logit(\cdot)$和$\Phi(\cdot)$为联系函数（link function），均为$(-\infty, +\infty)$至$(0, 1)$的单调函数（刻画概率）。

两模型均已有较成熟的估计、检验、计算等方法及软件。

六、一个例子

例4. 研究1500米男子世界纪录，数据为自1896年首届现代奥林匹克运动会以来，至1983年8月，所有得到确认的世界纪录及其被创造的时间。自变量为时间t，响应变量为纪录Y_t。

解：首先，利用一元线性回归模型，得到

$$\hat{Y}_t = 4.123 - 0.0073t$$

第五节 回归分析

检验结果为“显著”。但问题为：

1. 残差图显示，数据不符合线性模型，模型应改进；
2. 第一个数据为异常值，原因……，应删除；
3. 不能用于预测：Y_t不可能取负值。

利用上述分析，及相关专业知识，我们可以提出非线性模型如下：

$$Y_t = L + ae^{-bt}$$

参数估计的方法超出范围（不可能化为线性模型），结果为

$$\widehat{Y}_t = 2.986 + 1.157e^{-0.010t}$$

如果去掉第一个数据（outlier），结果为

第五节 回归分析

$$\widehat{Y}_t = 3.095 + 1.053e^{-0.0102t}$$

此经验公式可用于解释、预测等:

- 人类1500米男子极限时间为3分5.7秒;
- 到2083年，世界纪录可望达到3分15秒（现为3分26秒，平均每百米13.733秒）；
- 也可以问（控制），哪年可以打破3分10秒的纪录?

第五章 试验设计与方差分析

第一节 全面试验的方差分析

试验设计是数理统计的一个重要方向，研究各种因素（自变量）对响应变量的影响，以及如何通过经济合理的设计方案，找到响应变量结果最优化的因素组合。

方差分析是研究在不同总体（因素不同水平）下，对样本方差的来源进行分解、分析、比较，从而确定各因素对响应变量影响的大小。

● 与线性回归同样的：均讨论响应变量与自变量（因素）的关系。

● 与线性回归不同的：

1. 不做线性假设；
2. 因素离散化，可以是定性变量；
3. 目标为优化，而非实际或近似关系。
4. 数据常为试验数据，而非现场数据。

第一节 全面试验的方差分析

设Y与自变量（因素）$X_1, \cdots, X_p$的关系为

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

● 在此，我们不关心f的形式，也不假设其（近似）为线性函数，而是在每个因素X_i仅取有限个值时，利用统计方法找到其（某个）组合，使得Y的期望达到最大（最小）。

● 也希望判断$X_1, \cdots, X_p$中，哪些对Y有显著影响，哪些没有。

● 当$p = 2$时，格子点方法。

● 一般$X_1, \cdots, X_p$的所有取值可能太多，只能安排部分试验，重复试验?

若X_i有s_i个水平，则共有$s_1 \times \cdots \times s_p$个试验，若还安排重复试验?

----------end 20240509

● 如何安排部分试验？如何分析实验数据，是试验设计的重点问题。