Memo No.
Date / /

单因子方差分析

1. 均值模型 $y_{ij} = \mu_i + \varepsilon_{ij}$. $\begin{cases} i=1,\cdots,a \\ j=1,\cdots,m \end{cases}$

$E(\varepsilon_{ij}) = 0$.

效应模型 $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. $\begin{cases} i=1,\cdots,a \\ j=1,\cdots,m \end{cases}$

约束：$\sum_{i=1}^{a} \alpha_i = 0$

假定：$\varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

* 因为只考虑了一个因子 α_i, 所以称为单因子方差分析模型

2. $y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$.

3. 假设检验. $H_0: \mu_1 = \mu_2 = \cdots = \mu_a$.

H_1: 存在 $\mu_i \neq \mu_j$.

或 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$.

H_1: 存在 $\alpha_i \neq 0$.

本质上等价（使用了均值/效应模型）

4. 偏差平方和 $SST = \sum_{i=1}^{a}\sum_{j=1}^{m} (y_{ij} - \bar{y}_{..})^2$

$= \sum_{i=1}^{a}\sum_{j=1}^{m} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{m} (y_{ij} - \bar{y}_{i.})^2$

$= SSA + SSE$.

Memo No.
Date / /

考虑平方项展开，$SST=\sum_{i=1}^{a}\sum_{j=1}^{m}\left((\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2+(y_{ij}-\bar{y}_{i\cdot})^2+2(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})\cdot(y_{ij}-\bar{y}_{i\cdot})\right)$

其中交叉项为零，因为 $\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})=0$.

5. $\frac{SSE}{\sigma^2}\sim\chi^2(n-a)$

$$SSE=\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2$$
$$=\sum_{i=1}^{a}\sum_{j=1}^{m}\left(\mu+\alpha_i+\varepsilon_{ij}-\frac{1}{m}\sum_{k=1}^{m}\mu+\alpha_i+\varepsilon_{ik}\right)^2$$
$$=\sum_{i=1}^{a}\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot})^2$$

因为 $\varepsilon_{ij}\sim N(0,\sigma^2)$，且 ε_{ij} 互相独立。

$\sum_{j=1}^{m}\left(\frac{\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot}}{\sigma}\right)^2$ 可以看成 m 个标准正态分布的平方和，且互不相关。

由卡方分布定义，$\sum_{j=1}^{m}\left(\frac{\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot}}{\sigma}\right)^2\sim\chi^2(m-1)$

再由卡方分布可加性，$SSE/\sigma^2\sim\chi^2(a(m-1))$

$a(m-1)=am-a=n-a$

所以 $\frac{SSE}{\sigma^2}\sim\chi^2(n-a)$

Memo No.
Date / /

6. $E(SSA) = (a-1)\sigma^2 + m\sum_{i=1}^{a}\alpha_i^2$.

原假设H_0成立时. $\frac{SSA}{\sigma^2} \sim \chi^2(a-1)$

$SSA = \sum_{i=1}^{a}\sum_{j=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = m\sum_{i=1}^{a}\left(\frac{1}{m}\sum_{j=1}^{m}(\mu+\alpha_i+\varepsilon_{ij}) - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}(\mu+\alpha_i+\varepsilon_{ij})\right)^2 \overset{(\sum_{i=1}^{a}\alpha_i=0)}{=\!=} m\cdot\sum_{i=1}^{a}(\alpha_i+\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2$

$= m\cdot\sum_{i=1}^{a}\alpha_i^2 + m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2 + 2m\cdot\sum_{i=1}^{a}\alpha_i\cdot(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})$.

其中交叉项期望 $E\left(2m\sum_{i=1}^{a}\alpha_i\cdot(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})\right) =$

$2m\sum_{i=1}^{a}\alpha_i E(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}) = 0$, $\bar{\varepsilon}_{i\cdot} \sim N(0, \frac{\sigma^2}{m})$.

其中 $\sum_{i=1}^{a}m\left(\frac{\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}}{\sigma}\right)^2$ 可以视为a个互不相关的标准正态分布的平方和, 即 $\sum_{i=1}^{a}\left(m\frac{\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}}{\sigma}\right)^2 \sim \chi^2(a-1)$

于是 $E\left(m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2\right) = \sigma^2\cdot(a-1)$.

所以 $E(SSA) = (a-1)\sigma^2 + m\sum_{i=1}^{a}\alpha_i^2$.

当H_0成立时. $SSA = m\sum_{i=1}^{a}(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2$.

所以 $\frac{SSA}{\sigma^2} \sim \chi^2(a-1)$.

Memo No.
Date / /

7. SSA ⊥ SSE.

$$SSA = m\sum_{i=1}^{a}(\alpha_i + \overline{\varepsilon_{i\cdot}} - \overline{\varepsilon_{\cdot\cdot}})^2.$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{m}(\varepsilon_{ij} - \overline{\varepsilon_{i\cdot}})^2$$

由于SSA 可以看成 $\overline{\varepsilon_{1\cdot}}, \overline{\varepsilon_{2\cdot}}, \cdots, \overline{\varepsilon_{a\cdot}}$ 的函数.

且 $\sum_{j=1}^{m}(\varepsilon_{ij} - \overline{\varepsilon_{i\cdot}})^2$ 与 $\overline{\varepsilon_{i\cdot}}$ 相互独立. 且 ε_{ij} 相互独立

可得 SSA 与 SSE 相互独立

8. 检验统计量 $F_A = \frac{SSA/(a-1)}{SSE/(n-a)} \sim F(a-1, n-a)$

9. 若 $F_A \geq F_{1-\alpha}(a-1, n-a)$ 则拒绝原假设

10. 方差分析表.

11. 若使用p值判断. $p_A = P(F \geq F_A)$

12. 参数估计.
$$\begin{cases} \hat{\mu} = \bar{y}_{\cdot\cdot}, \\ \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \ i=1,2,\cdots,a. \\ \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = \frac{SSE}{n} \end{cases}$$

但 $E(SSE)=\sigma^2(n-a)$，即 $\hat{\sigma}^2_{MLE}$ 不是无偏估计.

通常用 $\hat{\sigma}^2=\frac{SSE}{n-a}=MSE$.

采用极大似然法. 由于 $y_{ij}\overset{独立}{\sim}N(\mu+\alpha_i,\sigma^2)$.

$$L(\alpha_1,\cdots,\alpha_a,\sigma^2)=\prod_{i=1}^{a}\prod_{j=1}^{m}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(y_{ij}-\mu-\alpha_i)^2}{2\sigma^2}}$$

其对数似然函数. $l(\alpha_1,\cdots,\alpha_a,\sigma^2)=\ln L(\alpha_1,\cdots,\alpha_a,\sigma^2)$

$$=-\frac{n}{2}\cdot\ln\cancel{\sqrt{2\pi\sigma^2}}(2\pi\sigma^2)-\sum_{i=1}^{a}\sum_{j=1}^{m}\frac{(y_{ij}-\mu-\alpha_i)^2}{2\sigma^2}$$

对各个参数求偏导:

$$\begin{cases}\frac{\partial l}{\partial\mu}=\frac{1}{\sigma^2}\cdot\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)=0.\\ \frac{\partial l}{\partial\alpha_i}=\frac{1}{\sigma^2}\cdot\sum_{i=1}^{m}(y_{ij}-\mu-\alpha_i)=0.\\ \frac{\partial l}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\mu-\alpha_i)^2=0.\end{cases}$$

考虑效应模型约束 $\sum_{i=1}^{a}\alpha_i=0$.

得上述参数估计结果.

Memo No.
Date / /

12. 因子A的第i个水平的均值 μ_i 的 $1-\alpha$ 置信区间.

为 $[\bar{y}_{i\cdot} - t_{1-\alpha/2}(n-a)\hat{\sigma}, \bar{y}_{i\cdot} + t_{1-\alpha/2}(n-a)\hat{\sigma}]$

其中 $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{SSE/(n-a)}$

由于 $\bar{y}_{i\cdot} = \frac{1}{m}\cdot\sum_{j=1}^{m} y_{ij} = \mu + \alpha_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu+\alpha_i, \sigma^2 m^{-1})$

$$\frac{\sqrt{m}\cdot(\bar{y}_i - \mu_i)}{\sqrt{SSE/(n-a)}} \sim t(n-a)$$

$$\frac{(\bar{y}_i - \mu_i)/\sqrt{\frac{\sigma^2}{m}}}{\sqrt{\frac{SSE}{\sigma^2}\cdot/(n-a)}} \overset{?}{=} \frac{\sqrt{m}\cdot(\bar{y}_i - \mu_i)}{\sqrt{SSE/(n-a)}} \sim t(n-a)$$

13. 水平 $a=2$ 时, 单因子方差分析与二样本独立t检验等价

由t分布和F分布的关系, $F(1,n) = t^2(n)$.

$t^2 = t^2_{1-\alpha}\cdot(2m-2)$ 即 $F = F_{1-\alpha}(1, 2m-2)$.

因此两种检验方法等价

Memo No.
Date / /

双因子方差分析.

1. 模型一般形式. $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

$\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$.

$s.t. \sum_{i=1}^{a} \alpha_i = 0.\ \sum_{j=1}^{b} \beta_j = 0.\ \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$

2. $SST = SS_A + SS_B + SS_{AB} + SS_E$.

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{m} (y_{ijk} - \bar{y}_{\cdots})^2$$

$$SSA = bm \cdot \sum_{i=1}^{a} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2.$$

$$SSB = am \cdot \sum_{j=1}^{b} (\bar{y}_{\cdot j \cdot} - \bar{y}_{\cdots})^2.$$

$$SS_{AB} = m \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j \cdot} + \bar{y}_{\cdots})^2.$$

$$SS_E = ab \sum_{k=1}^{m} (y_{ijk} - \bar{y}_{ij\cdot})^2.$$

通过展开后计算交叉项，可以发现交叉项(6项)全部为0.

3. $E(MS_A) = E\left(\frac{SSA}{a-1}\right)$. $E(MS_B) = E\left(\frac{SS_B}{b-1}\right)$

$E(MS_{AB}) = E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right)$ $E(MS_E) = E\left(\frac{SSE}{ab(m-1)}\right)$

Memo No.
Date / /

4. 双因子方差分析检验.

由 $SS_A \sim \chi^2(a-1)$. $SS_B \sim \chi^2(b-1)$

$SS_{AB} \sim \chi^2((a-1)(b-1))$ $SS_E \sim \chi^2(ab(m-1))$

得. $\frac{MSA}{MSE} \sim F(a-1, ab(m-1))$.

$\frac{MSB}{MSE} \sim F(b-1, ab(m-1))$

$\frac{MSAB}{MSE} \sim F((a-1)(b-1), ab(m-1))$.

由此确定相应F分布的临界值.

一元线性回归.

1.模型 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

$\cancel{\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)}$ $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$.

2.参数估计. 经验回归方程: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

3.最小二乘估计. 偏差平方和 $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - E(y_i))^2$.

$= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

得 $\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = L_{xy} \cdot L_{xx}^{-1} \end{cases}$

记 $\sum_{i=1}^{n}(x_i-\bar{x}) \cdot (y_i-\bar{y}) = L_{xy}$. $\sum_{i=1}^{n}(x_i-\bar{x})^2 = L_{xx}$.

对 β_0, β_1 分别求偏导.

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0.$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0.$$

整理后得到 $\begin{cases} n\beta_0 + n\bar{x}\beta_1 = n\bar{y}. \\ n\bar{x}\beta_0 + \sum_{i=1}^{n} x_i^2 \beta_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$

Memo No.

Date / /

另外，一阶导为零，所求的 $\hat{\beta}_0, \hat{\beta}_1$ 实际上是稳定点，为了判断是否为最小点，需要求二阶偏导并有导数>0。

4. 最大似然估计

$y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$. $f_i(y_i)=\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y_i-(\beta_0-\beta_1 x_i)]^2}{2\sigma^2}}$

似然函数 $L(\beta_0,\beta_1,\sigma^2)=\prod_{i=1}^{n} f_i(y_i)$

$\ln L(\beta_0,\beta_1,\sigma^2)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$

对 β_0, β_1 求偏导，得到的估计系数与最小二乘一致。

5. $\hat{\sigma}_{ML}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$.

更常用 σ^2 的无偏估计 $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

6. 最大似然估计和最小二乘估计的区别：

最大似然估计是在 $\varepsilon_i \sim N(0,\sigma^2)$ 的正态分布假设下求得的，但最小二乘估计对分布假设没有要求。

7. 在正态分布假定下 $(y_i \sim N(\beta_0+\beta_1 x, \sigma^2))$

$\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}})\sigma^2)$. $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$

$Cov(\hat{\beta}_0, \hat{\beta}_1)=-\frac{\bar{x}}{L_{xx}}\sigma^2$.

Memo No.

Date / /

将 $\hat{\beta}_0, \hat{\beta}_1$ 写成 $y_1, y_2, \cdots, y_n$ 的线性组合形式。

$$l_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{y}.$$

发现 $\sum_{i=1}^{n} x_i-\bar{x} = 0$. 即 $l_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})y_i$

所以 $\hat{\beta}_1 = l_{xy}\cdot l_{xx}^{-1} = l_{xx}^{-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})y_i = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{l_{xx}}\right)y_i.$

$$\hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} = \bar{y} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})\bar{x}}{l_{xx}}y_i.$$

$$= \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{l_{xx}}\right)y_i$$

由 $y_1, y_2, \cdots, y_n$ 的正态假定，以及正态分布可加性，可知 $\hat{\beta}_0, \hat{\beta}_1$ 也符合正态分布，只需要确认它们的期望和方差即可得到分布情况。

$$E(\hat{\beta}_1) = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{l_{xx}}\cdot E(\beta_0+\beta_1 x_i+\varepsilon_i).$$

$$= \sum_{i=1}^{n}\frac{(x_i-\bar{x})x_i}{l_{xx}}\cdot\beta_1 = \beta_1.$$

(利用 y_i 独立性)

$$Var(\hat{\beta}_1) = \cancel{\left[\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{l_{xx}}\right)\right]^2\cdot Var}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{l_{xx}}\right)^2\cdot Var(y_i).$$

$$= \frac{\sigma^2}{l_{xx}}.$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{l_{xx}}\right)\cdot(\beta_0+\beta_1 x_i) = \beta_0-(\beta_1\bar{x}-\beta_1\bar{x}) = \beta_0.$$

$$Var(\hat{\beta}_0) = \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}\cdot(x_i-\bar{x})}{l_{xx}}\right)^2\cdot Var(y_i) = \sigma^2\cdot\left(\frac{\bar{x}^2}{l_{xx}}+\frac{1}{n}\right)$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right)y_i, \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{l_{xx}}y_i\right)$$

由于 $Cov(A+B, C+D) = Cov(A,C) + Cov(A,D) + Cov(B,C) + Cov(B,D)$.

且 y_i 之间独立，$Cov(y_i, y_j) = 0\ (i \neq j)$

且 $Cov(aX, bY) = ab \cdot Cov(X, Y)$

由以上性质，有 $Cov(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right)\left(\frac{x_i-\bar{x}}{l_{xx}}\right)\cdot\sigma^2$

$$= -\frac{\bar{x}}{l_{xx}}\sigma^2$$

8. 显著性检验，

检验问题 $H_0: \beta_1 = 0$. vs $H_1: \beta_1 \neq 0$

9. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，$e_i = y_i - \hat{y}_i$.

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

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

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

$SST = SSR + SSE$

Memo No.
Date / /

10. $SSR=\hat{\beta}_1^2 l_{xx}$, $E(SSR)=\sigma^2+\beta_1^2 l_{xx}$.

$E(SSE)=(n-2)\sigma^2$.

$$SSR=\sum_{i=1}^n(\hat{y}_i-\bar{y})^2=\sum_{i=1}^n(\hat{\beta}_1\cdot(x_i-\bar{x}))^2=\hat{\beta}_1^2\cdot l_{xx}.$$

$$E(SSR)=l_{xx}\cdot E(\hat{\beta}_1^2)=l_{xx}(Var(\hat{\beta}_1)+E(\hat{\beta}_1)^2).$$

$$=l_{xx}\cdot\left(\frac{\sigma^2}{l_{xx}}+\beta_1^2\right)=\sigma^2+\beta_1^2 l_{xx}.$$

$$SSE=\sum_{i=1}^n(y_i-\hat{y}_i)^2=\sum_{i=1}^n(\beta_0+\beta_1x_i+\varepsilon_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

$$=\sum_{i=1}^n(\beta_0-\hat{\beta}_0)^2+\sum_{i=1}^n(\beta_1-\hat{\beta}_1)^2x_i^2+\sum_{i=1}^n2(\beta_0-\hat{\beta}_0)(\beta_1-\hat{\beta}_1)x_i$$

$$+\sum_{i=1}^n\varepsilon_i^2+2\sum_{i=1}^n(\beta_0-\hat{\beta}_0)\varepsilon_i+2\sum_{i=1}^n(\beta_1-\hat{\beta}_1)\varepsilon_i x_i$$

$$E(SSE)=n\cdot Var(\hat{\beta}_0)+Var(\hat{\beta}_1)\cdot\sum_{i=1}^n x_i^2+Cov(\beta_0,\beta_1)\cdot\sum_{i=1}^n 2x_i$$

$$+\sum_{i=1}^n Var(\varepsilon_i)-2\sum_{i=1}^n E(\hat{\beta}_0\varepsilon_i)-2\sum_{i=1}^n E(\hat{\beta}_1\varepsilon_i x_i)$$

考虑 $\sum_{i=1}^n E(\hat{\beta}_1\varepsilon_i x_i)=\sum_{i=1}^n E\left(\frac{\sum(x_i-\bar{x})x_i}{l_{xx}}y_i\cdot\varepsilon_i\right)$

$$=\sum_{i=1}^n E\left(\sum\frac{(x_i-\bar{x})x_i}{l_{xx}}\cdot(\beta_0+\beta_1x_i+\varepsilon_i)\cdot\varepsilon_i\right).$$

由 $\varepsilon_1,\cdots,\varepsilon_n$ 之间相互独立($E(\varepsilon_i\varepsilon_j)=E(\varepsilon_i)E(\varepsilon_j)\cdot i\neq j$).

且 $E(\varepsilon_i)=0$.

有 $\sum_{i=1}^n E(\hat{\beta}_0\varepsilon_i)=E(\varepsilon_i^2)\cdot\sum_{i=1}^n\frac{(x_i-\bar{x})x_i}{l_{xx}}=\sigma^2$.

Memo No.
Date / /

$$\sum_{i=1}^{n} E(\hat{\beta}_0 \varepsilon_i) = \sum_{i=1}^{n} E\left(\left(\frac{1}{n} - \frac{\bar{x}\cdot(x_i-\bar{x})}{l_{xx}}\right)\cdot(\beta_0+\beta_1 x_i+\varepsilon_i)\varepsilon_i\right).$$

$$= \sum_{i=1}^{n} E\left[\left(\frac{1}{n} - \frac{\bar{x}\cdot(x_i-\bar{x})}{l_{xx}}\right)\cdot\varepsilon_i^2\right] = \sigma^2$$

代入 $E(SSE)$ 中：

$$E(SSE) = n\cdot\sigma^2\cdot\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right) + \frac{\sigma^2}{l_{xx}}\sum_{i=1}^{n} x_i^2 + -\frac{\bar{x}}{l_{xx}}\sigma^2\sum_{i=1}^{n} 2x_i$$

$$+n\sigma^2 - 2\sigma^2 - 2\sigma^2$$

$$= \sigma^2\cdot\left(1+\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{l_{xx}} + n-4\right) = (n-2)\sigma^2$$

11. $SSE/\sigma^2 \sim \chi^2(n-2)$.

若 H_0 成立. $SSR/\sigma^2 \sim \chi^2(1)$

SSR 与 SSE，$\bar{y}$ 独立

$F_0 = \frac{SSR}{SSE/(n-2)}$，拒绝域 $F_0 \geq F_{1-\alpha}(1, n-2)$

通过构造正交矩阵证明.

12. t检验统计量 $t_0 = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \sim t(n-2)$

其中 $\hat{\sigma} = \sqrt{SSE/(n-2)}$. 拒绝域为 $W = \{|t_0| > t_{1-\alpha/2}(n-2)\}$.

由于 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$，$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$，$H_0$ 为真时 $\beta_1 = 0$

所以 $\frac{\hat{\beta}_1/\sqrt{\frac{\sigma^2}{l_{xx}}}}{\sqrt{\frac{SSE}{\sigma^2}/(n-2)}} = \frac{\hat{\beta}_1}{\sqrt{SSE/n-2}\cdot\sqrt{l_{xx}}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \sim t(n-2)$

Memo No.
Date / /

13. $r = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \dfrac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}$.

14. $r^2 = l^2_{xy} \cdot \dfrac{1}{l_{xx} \cdot l_{yy}} = \hat{\beta}_1^2 \dfrac{l_{xx}}{l_{yy}} = \dfrac{SSR}{SST}$.

15. $|r|$ 是 F_0 的严格单调函数. 可以从 F 分布的分位数得到相关系数检验的临界值. $r_{1-\alpha/2}(n-2)$.

$$r^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \frac{SSR/(SSE/(n-2))}{SSR/(SSE/(n-2)) + (n-2)} = \frac{F_0}{F_0+(n-2)}$$

16. 估计区间中 $\delta = t_{1-\alpha/2}(n-2)\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{xx}}}$

预测区间中 $\delta = t_{1-\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n} + \frac{(x_0-\bar{x})^2}{l_{xx}}}$

线性回归分析（矩阵版）

1. 模型：$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon$

假定 $E(\varepsilon)=0$，$Var(\varepsilon)=\sigma^2$

即 $y=X\beta+\varepsilon$

2. 假设 (1) X：是确定性变量. $rank(X)=p+1<n$.

（样本量应大于自变量个数，X满秩）

(2). $E(\varepsilon_i)=0$，$\cancel{Va}\ Cov(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, & i=j\\ 0, & i\neq j\end{cases}$

（高斯马尔可夫条件）

(3). $\begin{cases}\varepsilon_i\sim N(0,\sigma^2)\\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立.}\end{cases}$

即 $\varepsilon\sim N(0,\sigma^2I_n)$

故 $E(y)=X\beta$，$Var(y)=\sigma^2I_n$

3. 最小二乘估计. $Q(\beta)=\sum_{i=1}^{n}(y_i-x_i'\beta)^2$

$$=\|y-X\beta\|^2$$

$$=(y-X\beta)'(y-X\beta)$$

$$=\beta'X'X\beta-2\beta'X'y+y'y$$

~~分别~~对β求偏导. $\frac{\partial Q}{\partial \beta} = 2X'X\beta - 2X'y = 0$.

$\hat{\beta}_{LS} = (X'X)^{-1}X'y$.

4. $\hat{y} = X\hat{\beta}_{LS} = X\cdot(X'X)^{-1}X'y$.

记$X\cdot(X'X)^{-1}X' = H$.为帽子矩阵.有$\hat{y} = Hy$.

5. $H = X(X'X)^{-1}X'$,是n阶对称矩阵. $H = H'$

是幂等矩阵. $H = H^2$

H的迹为p+1. $tr(H) = p+1$.

$$H' = (X(X'X)^{-1}X')' = X\cdot(X'X)^{-1}X' = H.$$

$$H^2 = X\cdot(X'X)^{-1}\cdot X'\cdot X\cdot(X'X)^{-1}\cdot X' = X\cdot(X'X)^{-1}\cdot(X'X)\cdot(X'X)^{-1}\cdot X'$$
$$= X\cdot(X'X)^{-1}X' = H.$$

$$tr(H) = tr(X(X'X)^{-1}X') = tr((X'X)^{-1}\cdot(X'X)) = tr(I_{p+1}) = p+1$$

6. 定义残差 $e = y - \hat{y} = y - Hy = (I-H)y$.

$\hat{y}$与e垂直. $\hat{y}'\cdot e = (Hy)'\cdot(I-H)y = 0$.

Memo No.
Date / /

7. $Var(e) = \sigma^2 \cdot (I-H)$.

$$Var(e) = Cov(e, e) = Cov((I-H)y, (I-H)y)$$
$$= (I-H) Cov(y, y) \cdot (I-H)' \quad (Cov(Ax, By) = A\,Cov(x,y) B')$$
$$= \sigma^2 \cdot (I-H) I_n (I-H)'$$
$$= \sigma^2 \cdot (I-H)$$

8. 最大似然估计. $y \sim N_n(X\beta, \sigma^2 I_n)$

联合密度函数 $f(y; \beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} \cdot |\sigma^2 I_n|^{-\frac{1}{2}} \cdot$
$\exp\{-\frac{1}{2}(y-X\beta)' \cdot |\sigma^2 I_n|^{-1} \cdot (y-X\beta)\}$.

对数似然函数 $\ln L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2)$
$-\frac{1}{2\sigma^2} \cdot (y-X\beta)'(y-X\beta)$

关于 β 和 σ^2 求偏导
$$\begin{cases} \hat{\beta}_{ML} = (X'X)^{-1}X'y \\ \hat{\sigma}^2_{ML} = \frac{1}{n} \cdot (y-X\hat{\beta}_{ML})'(y-X\hat{\beta}_{ML}) = \frac{1}{n}e'e \end{cases}$$

9. $E(\hat{\beta}) = \beta$. $Var(\hat{\beta}) = \sigma^2 \cdot (X'X)^{-1}$

证

$$E(\hat{\beta}) = E((X'X)^{-1}X' \cdot (X\beta + \varepsilon)) = \beta + (X'X)^{-1}X'E(\varepsilon) = \beta.$$
$$Var(\hat{\beta}) = Var((X'X)^{-1}X' \cdot y) = (X'X)^{-1}X' \cdot Var(y) \cdot X \cdot (X'X)^{-1}$$
$$= \sigma^2 \cdot (X'X)^{-1}$$

Memo No.
Date / /

10. $Cov(\hat{\beta}, e) = 0.$

$$Cov\left((X'X)^{-1}X'y,\ (I-H)y\right) = (X'X)^{-1}X'\cdot Var(y)\cdot(I-H)'$$
$$= \sigma^2\cdot(X'X)^{-1}\cdot X'\cdot\left(I - X\cdot(X'X)^{-1}X'\right) = 0$$

11. 中心化相关.

(1). 原始数据集: $\begin{cases} y = (y_1, \cdots, y_n)' \\ X = (1_n, X_0),\ X_0 = (x_1, \cdots, x_p). \end{cases}$

最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'y.$

$$= \begin{pmatrix} n & 1_n'X_0 \\ X_0'1_n & X_0'X_0 \end{pmatrix}^{-1}\cdot\begin{pmatrix} 1_n' \\ X_0' \end{pmatrix} y.$$

分块矩阵求逆

$$= \begin{pmatrix} n^{-1}1_n' + n^{-2}1_n'X_0A_0X_0'1_n1_n' - n^{-1}1_n'X_0A_0X_0' \\ -n^{-1}A_0X_0'1_n1_n' + A_0X_0' \end{pmatrix}\cdot y$$

其中 $A_0 = (X_0'X_0 - n^{-1}X_0'1_n1_n'X_0)^{-1}.$

中心化过程: $\begin{cases} x^*_{ij} = x_{ij} - \bar{x}_j,\ \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}. \\ y^*_i = y_i - \bar{y},\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \end{cases}$

令 $\begin{cases} y^* = (y_1^*, \cdots, y_n^*)' \\ X_c = (x_1^*, \cdots, x_p^*),\ x_j^* = (x_{1j}^*, \cdots, x_{nj}^*)' \\ X^* = (1_n, X_c). \end{cases}$

Memo No.
Date / /

中心化后的最小二乘估计.

$$\hat{\beta}_c = \begin{pmatrix} n^{-1}1_n' + n^{-2}1_n'X_cA_cX_c'1_n1_n' - n^{-1}1_n'X_cA_cX_c \\ -n^{-1}A_cX_c'1_n1_n' + A_cX_c' \end{pmatrix} y^*$$

其中 $A_c = (X_c'X_c - n^{-1}X_c'1_n1_n'X_c)^{-1}$.

(2). $H_{1n} = 1_n(1_n'1_n)^{-1}1_n'$ 对称幂等. $(I_n - H_{1n})$ 同理

(3). $X_c = (I_n - H_{1n})X_0$.

(4). $1_n'(I_n - H_{1n}) = 1_n' - 1_n' \cdot 1_n \cdot (1_n'1_n)^{-1}1_n' = 0$.

(5) $A_0 = A_0$.

$$A_c = (X_c'X_c - n^{-1}X_c'1_n1_n' \cdot X_c)^{-1}$$

$$= (X_0' \cdot (I_n - H_{1n}) \cdot X_0 - n^{-1} \cdot X_0' \cdot (I_n - H_{1n}) \cdot 1_n \cdot 1_n'(I_n - H_{1n})X_0)^{-1}$$

$$= (X_0'(I_n - H_{1n})X_0)^{-1} = A_0$$

(6). $\hat{\beta}_{c,intercept} = 0$, $\hat{\beta}_{c,slope} = \hat{\beta}_{slope}$.

Memo No.
Date / /

$$\hat{\beta}_{c,intercept} = (n^{-1}\cdot 1_n' + n^{-2}1_n'X_c\cdot A_cX_c'1_n1_n' - n^{-1}1_n'X_cA_cX_c')y^*$$

$$= n^{-1}\cdot 1_n' y^*$$

$$= n^{-1}\cdot 1_n'(I_n - H_{1n})y = 0.$$

$$\hat{\beta}_{c,slope} = (-n^{-1}A_cX_c'1_n1_n' + A_cX_c')y^*.$$

$$= A_c\cdot X_c'\cdot(I_n - H_{1n})y.$$

$$= A_0\cdot X_0'\cdot(I_n - H_{1n})y.$$

12. 标准化. $\hat{\beta}_{s,intercept} = 0.$ $\hat{\beta}_{s.slope} = \frac{1}{\sqrt{L_{yy}}}\cdot L^{-1}\cdot\hat{\beta}_{c.slope}.$

其中. $L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $L = diag\{\frac{1}{\sqrt{L_{11}}}, \cdots, \frac{1}{\sqrt{L_{pp}}}\}$. $L_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

每一个分量 $\hat{\beta}_{sj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_{cj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j.$ $j = 1, 2, \cdots, p.$

Memo No.
Date / /

13. F检验 $F_0 = \frac{SSR/p}{SSE/(n-p-1)} \overset{H_0}{\sim} F(p, n-p-1)$

拒绝域 $F_0 > F_{1-\alpha}(p, n-p-1)$

14. t检验 检验单个自变量的显著性

检验统计量 $t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$，$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$

拒绝域. $|t_j| \geq t_{1-\alpha/2}(n-p-1)$

正态假设下. y服从正态分布，$\hat{\beta} = (X'X)^{-1}X'y$ 是y的线性组合，因此 $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$

用 c_{jj} 表示 $(X'X)^{-1}$ 中第 $(j+1)$ 行. $(j+1)$ 列元素.

有 $\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$

$$\therefore \frac{\hat{\beta}_j / \sqrt{c_{jj}\sigma^2}}{\sqrt{\frac{SSE}{\cancel{(n-p)\sigma^2}}(n-p-1)}} = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\cdot\hat{\sigma}} \sim t(n-p-1)$$

15. 复相关系数. $R^2 = \frac{SSR}{SST}$.

16. $\hat{y}_0 \sim N(x_0'\beta, \sigma^2 x_0'(x'x)^{-1}x_0)$.

$y_0 \sim N(x_0'\beta, \sigma^2)$.

$\hat{y}_0 - y_0 \sim N(0, \sigma^2(1+x_0'(x'x)^{-1}x_0))$.

$\hat{y}_0 - y_0$ 与 $\hat{\sigma}^2$ 相互独立. $\frac{SSE}{\sigma^2} \sim \chi^2(n-p-1)$.

所以. $\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1+x_0'(x'x)^{-1}x_0}} \sim t_{n-p-1}$.

预测区间. $\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}\sqrt{1+x_0'(x'x)^{-1}x_0}$.

置信区间. $\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}\sqrt{x_0'(x'x)^{-1}x_0}$.

17. 为什么单因子方差分析模型可以看作一种多元线性回归.

方差分析 $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.

$\sum_{i=1}^{a} \alpha_i = 0$. 即 $\alpha_a = -\sum_{i=1}^{a-1} \alpha_i$.

~~多元线性回归.~~ 设计矩阵 $X = \begin{pmatrix} 1_m & & & \\ & 1_m & & \\ & & \ddots & \\ & & & 1_m \\ & -1_{m(a-1)} & & \end{pmatrix}$, $\beta = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{a-1} \end{pmatrix}$

~~$y = X\beta + \varepsilon$~~.

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

多元线性回归 构造设计矩阵.

$$X = (1_n, L), \quad L = \begin{pmatrix} \mathrm{diag}\{1_m, \cdots, 1_m\} \\ -1 \end{pmatrix}.$$

$$\cancel{Q = \sum y_{ij} \quad \hat{y} = X\beta + \varepsilon.} \quad \cancel{y_{ij} = X\beta} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\cancel{Q = \sum (y_{ij} - \mu - \alpha_i)^2.}$$

$$\cancel{\frac{\partial Q}{\partial \alpha_i} = \sum_{i=1}^{n} -2(y_{ij} - \mu - \alpha_i) = 0.}$$

$$\cancel{\therefore \hat{\alpha}_i = \bar{y}_i - \mu.}$$

$y = X\beta + \varepsilon$. 形式与方差分析一致.

$$\begin{cases} \hat{\mu} = \bar{y}_{\cdot\cdot} \\ \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \end{cases}$$

~~检~~ 方差分析假设检验 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$.

多元线性回归显著性检验: $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$.

检验统计量 $F_0 = \dfrac{SSR/(a-1)}{SSE/(n-a)}$

$/ F_0 = \dfrac{SSA/(a-1)}{SSE/(n-a)}$

参数估计. 假设检验等价

补充. 多重比较.

1. Bonferroni 方法. 将 $t_{1-\alpha/2}(n-a)$ 调整到 $t_{1-\alpha/(a(a-1))}(n-a)$

于是 $P(\bigcap_{i=1}^{a(a-1)/2} A_i) \geq 1 - a(a-1)/2 \cdot \frac{\alpha}{a(a-1)/2} = 1-\alpha$.

2. Tukey 方法. $W = \bigcup_{1\leq i<i'\leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c_{ii'}\}$.

假设 $c_{ii'}$ 相等. 记为 c.

令 $q(a,df) = \max_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}} - \min_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}}$.

$\frac{\bar{y}_i - \mu}{\hat{\sigma}/\sqrt{m}} \sim t(n-a)$

~~由 $P(W)$~~ $\Rightarrow c = q_{1-\alpha}(a,df)\hat{\sigma}/\sqrt{m}$.

若 $|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c$. 则认为水平 i 与 i' 存在显著差异.

$P(W) = P(\bigcup_{1\leq i<i'\leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c\}) = 1 - P(\bigcap_{1\leq i<i'\leq a} \{|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| < c\})$.

$= P(\max_{1\leq i<i'\leq a} |\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| \geq c)$

$= P(\max_i \bar{y}_{i\cdot} - \min_i \bar{y}_{i\cdot} \geq c)$

$= P\left(\max_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}} - \min_i \frac{\bar{y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}} \geq \frac{c}{\hat{\sigma}/\sqrt{m}}\right)$.

在原假设 $\mu_1=\mu_2=\cdots=\mu_a=\mu$ 成立时

$\bar{y}_{i\cdot} \sim N(\mu, \frac{\sigma^2}{m})$，$\hat{\sigma}^2 = \cancel{\frac{SSE}{\ }}\ \frac{SSE}{n-a}$. $\frac{SSE}{\sigma^2} \sim t(n-a)$

$$\therefore \frac{\bar{y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}} = \frac{(\bar{y}_{i\cdot}-\mu)/\sqrt{\frac{\sigma^2}{m}}}{\sqrt{\frac{SSE}{\sigma^2}/n-a}} \sim t(n-a)$$

3. 蒙特卡洛法计算 $q(a,df) = \max\limits_i \frac{\bar{y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}} - \min\limits_i \frac{\bar{y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}$

1. 生成a个服从 $N(0,1)$ 的随机数 $x_1,\cdots,x_a$.
2. 排序后令 X_{max} 为最大值. X_{min} 为最小值.
3. 从自由度为df的 χ^2 分布生成一个随机数y.
4. 计算 $q_n=(X_{max}-X_{min})/\sqrt{y/df}$.
5. ~~重复1-4步N次，取 $q_{1-\alpha}$~~

可通过 $q_\alpha(a,df)$ 表示 $q(a,df)$ 的α分位数.

即 $c = q_{1-\alpha}(a,df)\,\hat{\sigma}/\sqrt{m}$.

Memo No.
Date / /

变量选择

1. 考虑过少的自变量 — 欠拟合

波动(方差)小. 偏差大.(不是无偏估计)

2. 考虑过多的自变量, — 过拟合.

波动(方差)大. 偏差小(无偏估计)

3. $SS_E^{p+1} \leq SS_E^{p_1}$ 随自变量增加. 残差平方和减少

$R^2_{p_1+1} \geq R^2_{p_1}$. 随自变量增加. 决定系数增加.

所以 SSE. R^2 不能用于选择自变量.

修正: $\tilde{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$

$\hat{\sigma}^2 = \frac{1}{n-p-1} SSE$.

4. AIC = −2ln(模型最大似然) + 2(模型独立参数个数)

线性模型中. $AIC = n \cdot \ln(2\pi) + n \ln\left(\frac{SSE}{n}\right) + n + 2(p+2)$

$\propto n \cdot \ln(SSE/n) + 2(p+1)$

5. BIC = −2ln(模型最大似然) + ln(n)(模型独立参数个数)

$\propto n \cdot \ln(SSE/n) + \ln(n)(p+1)$

Memo No.
Date / /

模型最大似然. 多元正态分布概率密度函数.

$$f(x_1,\cdots,x_n)=\frac{1}{\sqrt{(2\pi)^n\cdot|\Sigma|}}\cdot exp\cdot\left\{-\frac{1}{2}\cdot(x-\mu)'\cdot\Sigma^{-1}\cdot(x-\mu)\right\}$$

此处 $\Sigma=\sigma^2 I_n$.

$$\therefore f(x_1,\cdots,x_n)=(2\pi)^{-\frac{n}{2}}\cdot(\sigma^2)^{-\frac{n}{2}}\cdot exp\cdot\left\{-\frac{e'e}{2\sigma^2}\right\}.$$

对数似然 $\ln f(x_1,\cdots,x_n)=-\frac{n}{2}\cdot\ln(2\pi)-\frac{n}{2}\cdot\ln(\sigma^2)-\frac{e'e}{2\sigma^2}$

由 $\hat{\sigma}^2_{ML}=\frac{e'e}{n}$ 代入.

$$\ln f(x_1,\cdots,x_n)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\left(\frac{SSE}{n}\right)-\frac{n}{2}$$

6. 逐步回归.

前进法: 没有剔除机会, 不全面.

后退法: 计算量大. 剔除后没有进入机会.

Memo No.
Date / /

多重共线性

1. 当自变量完全线性相关，$|X'X|=0$，无法得到 $\hat{\beta}$。

相关性很高时 $|X'X|\approx 0$，由于 $Var(\hat{\beta})=\sigma^2(X'X)^{-1}$，

导致精度很低。

即，随着自变量间相关性增加，$\hat{\beta}$ 的方差会变大。

2. 方差扩大因子。$C=(c_{ij})=(X_s'X_s)^{-1}$。

$VIF_j=c_{jj}$，$\underline{Var(\hat{\beta}_j)=\frac{c_{jj}}{l_{jj}}\sigma^2}$（所以得名）。

$$Var(\hat{\beta}_s)=\sigma^2\cdot(X_s'\cdot X_s)^{-1}.$$

$$\Rightarrow Var(\hat{\beta}_{s,j})=\sigma^2\cdot c_{jj}.$$

由于此处仅对自变量标准化 $\hat{\beta}_{s,j}=\sqrt{l_{jj}}\,\hat{\beta}_j$。

$\therefore \hat{\beta}_j=\frac{\hat{\beta}_{s,j}}{\sqrt{l_{jj}}}$，$Var(\hat{\beta}_j)=\frac{c_{jj}}{l_{jj}}\sigma^2$

$\left(l_{jj}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\right)$。

3. $c_{jj}=\frac{1}{1-R_j^2}$，R_j^2 表示将 x_j 作为因变量，其余 $p-1$ 个自变量建立多元线性回归模型，所表示的复决定系数。

4. *特征值判定法（条件数）

$X'X$ 的特征值分别为 $\lambda_1 \geq \cdots \geq \lambda_p$.

称 $k_j = \sqrt{\frac{\lambda_1}{\lambda_j}}, j=1,2,\cdots,p$ 为特征值 λ_j 的条件数

5. 岭回归：用 $X'X+kI$，$k>0$ 代替 $X'X$

即 $\hat{\beta}(k) = (X'X+kI)^{-1} \cdot X'y$.

6. $E(\hat{\beta}(k)) \neq \beta$，即岭回归是有偏估计.

$$E[(X'X+kI)^{-1}X'y] = (X'X+kI)^{-1}X' \cdot E(y)$$
$$= (X'X+kI)^{-1}X' \cdot X\beta \neq \beta.$$

7. 若 k 与 y 无关，则 $\hat{\beta}(k) = (X'X+kI)^{-1} \cdot X'y$

$= (X'X+kI)^{-1}(X'X) \cdot (X'X)^{-1} \cdot X'y$

$= (X'X+kI)^{-1} \cdot (X'X) \cdot \hat{\beta}$.

即 $\hat{\beta}(k)$ 是 $\hat{\beta}$ 的线性变换，是 y 的线性函数.

8. $MSE(\hat{\beta}(k)) \leq MSE(\hat{\beta}(0))$

Memo No.
Date / /

9. 岭回归等价于.

$$\min \ (y-X\beta)'(y-X\beta)$$
$$s.t. \ \beta'\beta \le S.$$

即最小化带有 L_2 正则项的离差平方和的解.

$$\hat{\beta}(k) = \arg\min_{\beta} (y-X\beta)'\cdot(y-X\beta) + \lambda\beta'\beta.$$

10. $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$

11. 岭参数选择

(1) $k_{HK} = \frac{\hat{\sigma}^2}{\max_j \hat{\alpha}_j^2}$

(2) 选择 k 使得. $\|\hat{\beta}(k)\|^2 \approx \|\hat{\beta}\|^2 - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$

12. 主成分分析, $z_i = a_i'x$. $a_i'a_i = 1$.

$a_i'\Sigma a_j = 0 \quad (i > 1)$.

$Var(z_i) = \max Var(a_i'x)$.

~~即 max a~~

13. 求第一主成分等价于求Σ特征值和特征向量.

$$\max \; Var(a_1'X) = a_1'\Sigma a_1$$

$$s.t. \quad a_1'a_1 = 1$$

拉格朗日乘子法. $L(a_1) = a_1'\Sigma a_1 - \lambda(a_1'a_1 - 1)$

$\frac{\partial L}{\partial a_1} = 2(\Sigma - \lambda I)a_1 = 0$

$\frac{\partial L}{\partial \lambda} = a_1'a_1 - 1 = 0$

$\because a_1 \neq 0$. $\therefore |\Sigma - \lambda I| = 0$.

等价于求Σ的特征值和特征向量问题.

14. Σ的特征值为 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$,

单位正交特征向量 $a_1, a_2, \cdots, a_p$.

则 $Z_i = a_i'X$, $i = 1, 2, \cdots, p$

15. 为了消除量纲, 可以对X标准化,

即 $x_i^* = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} = \frac{x_i - \mu_i}{\sigma_i}$

此时的协方差矩阵 Σ^* 是原变量的相关阵 $Corr(X)$.

用 $Corr(X)$ 来求主成分.

16. 通过样本估计方差-协方差矩阵 Σ.

$$S = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' \overset{def}{=} (s_{kl})_{p\times p}.$$

17. $Z = X\cdot V'$.

$$y = X\beta+\varepsilon = \underline{X\cdot V'}\cdot\underline{V\beta}+\varepsilon = Z\cdot\alpha+\varepsilon.$$

当 $\lambda_{k+1},\cdots,\lambda_p$ 均近似为0时，$Z_{k+1},\cdots,Z_p$ 近似为0.

$$y = (Z_1, Z_2)\cdot(\alpha_1', \alpha_2')'+\varepsilon$$
$$= Z_1\alpha_1+Z_2\alpha_2+\varepsilon$$
$$\approx Z_1\alpha_1+\varepsilon$$

最小二乘估计 $\hat{\alpha}_1 = (Z_1'Z_1)^{-1}\cdot Z_1'y$

$$= \Lambda_1^{-1}\cdot Z_1'y.$$

$(Z'Z = V\cdot X'X\cdot V' \Leftarrow X'X = V'\Lambda V)$.

$$\Lambda = \begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix}$$

$\because\ \beta = V'\alpha$

$\therefore\ \hat{\beta}_{pc} = V\cdot\begin{pmatrix}\hat{\alpha}_1\\ 0\end{pmatrix} = V_1'\cdot\Lambda_1^{-1}\cdot Z_1'y.$

18. $\hat{\beta}_{PC} = V_1' V_1 \hat{\beta}$; $E(\hat{\beta}_{PC}) = V_1' V_1 \beta$. 有偏.

$$\hat{\beta}_{PC} = V_1' \Lambda_1^{-1} Z_1' y$$
$$= V_1' \Lambda_1^{-1} \cdot \cancel{Z_1' X} V_1 X' \cdot y$$
$$= V_1' \Lambda_1^{-1} \cdot X'X \cdot (X'X)^{-1} \cdot X' y$$
$$= V_1' \Lambda_1^{-1} \cdot X'X \cdot \hat{\beta}$$
$$= V_1' \Lambda_1^{-1} \cdot V' \Lambda X \cdot \hat{\beta}$$
$$= V_1' \Lambda_1^{-1} \cdot V_1 \cdot (V_1', V_2') \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \cdot \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \hat{\beta}$$
$$= V_1' V_1 \hat{\beta}.$$

19. $MSE(\hat{\beta}_{PC}) < MSE(\hat{\beta})$

~~存在合适的k~~.

若X存在多重共线性，存在k使得

20. $\|\hat{\beta}_{PC}\|^2 \le \|\hat{\beta}\|^2$ 是压缩估计

Memo No.
Date / /

聚类

1. GMM 混合高斯模型

假定第 i 个样本 x_i 来自于第 k 类正态分布 $N_p(\mu_k, \Sigma_k)$

x_i 的密度函数 $f(x_i) = \frac{1}{\sqrt{(2\pi)^p \cdot |\Sigma_k|}} \cdot \exp\left\{-\frac{1}{2}\cdot(x_i-\mu_k)'\Sigma_k^{-1}\cdot(x_i-\mu_k)\right\}$

构造 $\delta_{ik} = \begin{cases} 0. & \text{第 i 个样本 } x_i \text{ 不属于 } k. \\ 1. & \text{第 i 个样本 } x_i \text{ 属于 } k. \end{cases}$

$\pi_k = P(\delta_{ik}=1)$，满足 $0<\pi_k<1$. $\sum_{i=1}^{K}\pi_k = 1$.

δ_i 是独立同分布的随机向量.

$\delta_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{ik})'$ 的密度函数为.

$f(\delta_i) = \prod_{i=1}^{K}(\pi_k)^{\delta_{ik}}$, $i=1,2,\cdots,n$.

给定 δ_i 后，x_i 的密度函数.

$f(x_i|\delta_i) = \prod_{i=1}^{K}\left((2\pi)^{-p/2}\cdot|\Sigma_k|^{-1/2}\exp\left\{-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}\cdot(x_i-\mu_k)\right\}\right)^{\delta_{ik}}$

样本 $\{x_i, \delta_i\}$ 的联合密度函数. 即似然函数.

$\prod_{i=1}^{n} f(x_i, \delta_i) = \prod_{i=1}^{n} f(\delta_i)\cdot f(x_i|\delta_i)$

$= \prod_{i=1}^{n}\left(\pi_k\cdot(2\pi)^{-p/2}|\Sigma_k|^{-1/2}\exp\left\{-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}\cdot(x_i-\mu_k)\right\}\right)^{\delta_{ik}}$

$\prod^{K}$?

实际无法观测 δ_i.

$$f(x_i)=\sum_{k=1}^{K}\pi_k\cdot(2\pi)^{-p/2}|\Sigma_k|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}(x_i-\mu_k)\}.$$

采用EM算法.

$$l(\theta)=\ln L(\theta)$$

$$\propto -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}((x_i-\mu_k)'\Sigma_k^{-1}(x_i-\mu_k)+\ln|\Sigma_k|)$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\ln(\pi_k)=Q_0(\theta)$$

E步: 将潜变量 δ_{ik} 的期望 π^*_{ik} 代入 $Q_0(\theta)$.

$$\pi^*_{ik}=E(\delta_{ik}|x_i)=P(\delta_{ik}=1|x_i)=\frac{\pi_k\cdot\phi(x_i;\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\cdot\phi(x_i;\mu_k,\Sigma_k)}$$

其中 $\phi(x_i;\mu_k,\Sigma_k)=(2\pi)^{-p/2}\cdot|\Sigma_k|^{-\frac{1}{2}}\cdot\exp\{-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}\cdot(x_i-\mu_k)\}$.

$$Q_0(\theta)=Q_1(\theta)+Q_2(\theta).$$

$Q_1(\theta)$ 只与 $\{\mu_k,\Sigma_k\}$ 有关, $Q_2(\theta)$ 只与 $\{\pi_k\}$ 有关.

分别确定最大点.

$$\begin{cases}\mu_k=\dfrac{\sum_{i=1}^{n}\pi^*_{ik}x_i}{\sum_{i=1}^{n}\pi^*_{ik}}\\[2ex]\Sigma_k=\dfrac{\sum_{i=1}^{n}\pi^*_{ik}\cdot(x_i-\mu_k)\cdot(x_i-\mu_k)'}{\sum_{i=1}^{n}\pi^*_{ik}}\end{cases}$$

Memo No.
Date / /

拉格朗日乘子法 $\left(\sum_{k=1}^{K}\pi_k=1\right)$

$$Q_2^*(\theta)=\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^*\ln(\pi_k)-\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

得 $\pi_k=\frac{1}{n}\cdot\sum_{i=1}^{n}\pi_{ik}^*$

感知机：

$$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b).$$

（M为误分类点集合）

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$$

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

梯度

更新：梯度反方向。$w \leftarrow w + \eta y_i x_i$

$b \leftarrow b + \eta y_i$

学习算法 (1) 取初值 $w_0 = 0$，$b_0 = 0$.

(2) 选取数据 (x_i, y_i).

(3) 若 $y_i(w \cdot x_i + b) \le 0$（误分类点）.

更新：$w \leftarrow w + \eta y_i x_i$，

$b \leftarrow b + \eta y_i$.

(4) 转至(2)，直至训练集中没有误分类点

对偶形式：w是$y_i x_i$的线性组合，b是y_i的线性组合.

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i，\quad b = \sum_{i=1}^{N} \alpha_i y_i$$

模型：$f(x) = \text{sign}\left(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b\right)$

更新：$\alpha_i \leftarrow \alpha_i + \eta$，$b \leftarrow b + \eta y_i$

Memo No.
Date / /

朴素贝叶斯.

贝叶斯：$P(Y=c_k|X=x) = \dfrac{P(X=x|Y=c_k)\cdot P(Y=c_k)}{\sum_k P(X=x|Y=c_k)\cdot P(Y=c_k)}$

"朴素"：$P(X=x|Y=c_k) = P(X^{(1)}=x^{(1)}, \cdots, X^{(n)}=x^{(n)}|Y=c_k)$

$= \prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k)$.

其中. 先验概率 $P(Y=c_k) = \dfrac{\sum_{i=1}^{N} I(y_i=c_k)}{N}$

条件概率 $P(X^{(j)}=a_{jl}|Y=c_k) = \dfrac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)}{\sum_{i=1}^{N} I(y=c_k)}$.

拉普拉斯平滑：

计算 先验概率. $P(Y=c_k) = \dfrac{\sum_{i=1}^{N} I(y_i=c_k)+\lambda}{N+K\lambda}$

(K为Y的类别数)

条件概率 $P(X^{(j)}=a_{jl}|Y=c_k) = \dfrac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)+\lambda}{\sum_{i=1}^{N} I(y_i=c_k)+S_j\lambda}$.

(S_j为$X^{(j)}$的不同取值数).

SVM:

原目标: 最大间隔分类超平面.

$$\max_{w,b} \gamma$$

$$s.t. \quad y_i \frac{(wx_i+b)}{\|w\|^2} \geq \gamma$$

$\Downarrow$ (几何间隔和函数间隔的关系).

$$\max_{w,b} \frac{\hat{\gamma}}{\|w\|^2}$$

$$s.t. \quad y_i \cdot (w \cdot x_i + b) \geq \hat{\gamma}$$

$\Downarrow$

$$\max_{w,b} \frac{1}{\|w\|}$$

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1$$

$\Downarrow$

$$\min \frac{1}{2}\|w\|^2$$

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1$$

拉格朗日对偶.

$$L(\alpha, \beta, x) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i$$

$$\min_x \theta_p(x) = \min_x \max_{\alpha,\beta:\alpha_i \geq 0} L(x, \alpha, \beta)$$

KKT条件 $\Downarrow$

$$\max_{\alpha} \min_{w,b} L(w, b, \alpha)$$

Memo No.
Date / /

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \Rightarrow \quad \begin{cases} w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0. \end{cases}$$

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i y_i = 0.$$

回代 $L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n}\alpha_i$

对偶问题：
$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n}\alpha_i$$
$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0.$$
$$\alpha_i \geq 0.$$

计算分离超平面. 目标函数 $s(\alpha_1, \cdots, \alpha_{n-1})$.

计算 $s(\alpha_1, \cdots, \alpha_{n-1})$ 对 $\alpha_1, \cdots, \alpha_{n-1}$ 的偏导数并令其为0.

$\alpha_i \neq 0$ 对应的向量为支持向量.

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad b^* = y_i - w x_i = y_i - \sum_{j=1}^{n} \alpha_j^* y_j (x_j \cdot x_i)$$

(任取 $\alpha_i^* > 0$) 即 i 对应的是支持向量.

软间隔最大化：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i.$$
$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0.$$

Memo No.
Date / /

对应拉格朗日函数 $L(w,b,\xi,\alpha,\mu)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i$

$-\sum_{i=1}^{n}\alpha_i(y_i(wx_i+b)-1+\xi_i)-\sum_{i=1}^{n}\mu_i\xi_i$

同样,对拉格朗日函数分别求关于 w、b、ξ_i 偏导,

令导数为0.

$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$

$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0.$

$\frac{\partial L}{\partial \xi_i}=C-\alpha_i-\mu_i=0 \Rightarrow \mu_i=C-\alpha_i.$

回代, $L(w,b,\xi,\alpha,\mu)=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{n}\alpha_i$

对偶问题. $\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{n}\alpha_i.$

$s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0.$

$0\le\alpha_i\le C.$

Memo No.
Date / /

隐马尔可夫模型

模型 $\lambda=(A, B, \pi)$

状态转移矩阵 A、状态观测矩阵 B、初始状态概率 π.

前向算法. 初值 $\alpha_1(i)=\pi_i b_i(o_1)$.

递推. $\alpha_{t+1}(i)=\left[\sum_{j=1}^{N}\alpha_t(j)A_{ji}\right]b_i(o_{t+1})$.

后向算法. 初值. $\beta_T(i)=1$.

递推. $\beta_t(i)=\sum_{j=1}^{N}A_{ij}\cdot b_j(o_{t+1})\cdot\beta_{t+1}(j)$.

概率计算. $P(i_t=q_i|O,\lambda)=\frac{P(i_t=q_i, O|\lambda)}{P(O|\lambda)}$

$=\frac{\alpha_t(q_i)\cdot\beta_t(q_i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$

维特比算法. $\delta_i(j)$表示时刻i,状态为j的最大概率.

初值. $\delta_1(j)=\pi(j)b_j(o_1)$.

递推. $\delta_i(j)=\max_k\{\delta_{i-1}(k)\cdot A_{kj}\}\cdot b_j(o_i)$. 记录 $\psi_i(j)=\arg\max_k$.

$i^*_n=\arg\max_j\delta_n(j)$.

回溯. $\psi_t(i^*_t)=i^*_{t-1}$.