《统计方法与机器学习》Note. 2023.12.
——统计方法部分.

(基础知识回顾)主要是数学基础,概统(多元)内容.

1.向量和矩阵微分

$x \in R^n \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad f(x): R^n \to R \qquad \frac{\partial f}{\partial x} = \nabla_x f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$

$A \in R^{m\times n} \quad f: R^{m\times n} \to R \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \frac{\partial f}{\partial A} = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial a_{m1}} & \cdots & \frac{\partial f}{\partial a_{mn}} \end{bmatrix}$

$\frac{\partial (C_1 f(x) + C_2 g(x))}{\partial x} = C_1 \frac{\partial f}{\partial x} + C_2 \frac{\partial g}{\partial x} \qquad \frac{\partial (f(x)\cdot g(x))}{\partial x} = \frac{\partial f}{\partial x}\cdot g + \frac{\partial g}{\partial x}\cdot f$

$\frac{\partial f}{\partial x^T} = (\frac{\partial f}{\partial x})^T \qquad dx = \begin{bmatrix} dx_1 \\ \vdots \\ dx_n \end{bmatrix}$

$\boxed{\begin{array}{l} f\text{是个标量的话 } f = Trf \\ df = d(Trf) = Tr(df) \end{array}}$

$d(AB) = (dA)B + A(dB) \qquad d(A+B) = dA + dB \qquad d(A^T) = (dA)^T$

A可以不是标量,而是向量/矩阵,微分和迹运算也可交换 $d(Tr(A)) = Tr(dA)$

迹的性质: $Tr(A+B) = Tr(A) + Tr(B) \qquad Tr(AB) = Tr(BA) \qquad Tr(A) = Tr(A^T)$

$Tr(A_1 A_2 \cdots A_n) = \underbrace{Tr(A_n A_1 \cdots A_{n-1}) = Tr(A_{n-1} A_n A_1 \cdots A_{n-2})}_{\text{这个要满足矩阵乘法性质}}$

$Tr(A^T B) = \sum_i \sum_j A_{ij} B_{ij}$

$df(x_1 \cdots x_n) = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \cdots + \frac{\partial f}{\partial x_n} dx_n$ ← 多元函数

$\boxed{\text{迹微分法}}$ $f: R^{m\times n} \to R \quad A \in R^{m\times n}$ 无关量抛开!

$df = Tr((\frac{\partial f}{\partial A})^T dA)$

做一个练习. $f(x) = x^T A x \longrightarrow x \in R^n \quad A \in R^{n\times n} \quad \frac{\partial f}{\partial x} = ? \quad \frac{\partial f}{\partial A} = ?$

$df = dTr(f) = Tr(df) = Tr(dx^T \cdot Ax + x^T(dA\cdot x + A dx))$

$= Tr(dx^T \cdot Ax) + Tr(x^T A dx)$

$= Tr(x^T A^T dx + x^T A dx) = Tr((x^T A^T + x^T A) dx)$

故 $\frac{\partial f}{\partial x} = (x^T A^T + x^T A)^T = Ax + A^T x \qquad df = Tr(x^T dA \cdot x)$

$= Tr(x x^T dA)$

$\frac{\partial f}{\partial A} = (x x^T)^T = x x^T$

F范数 $\|A\|_F^2 = Tr(A^T A)$

行列式 $\frac{\partial |A|}{\partial A} = (A^*)^T \quad A^*$为A的伴随阵

$A^* \cdot A = A \cdot A^* = |A| \cdot I$

若A可逆, $A = |A| \cdot A^{-1}$

若不是实值函数了.

$f: R^m \to R^n \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$

$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}$

转化为一堆实值函数

X可逆 $dX^{-1} = ? \qquad X \cdot X^{-1} = I$

$(dX) \cdot X^{-1} + X\, d(X^{-1}) = 0$

$dX^{-1} = -X^{-1}(dX)(X^{-1})$

$$\frac{\partial f}{\partial x^T}=\begin{bmatrix}\frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n}\\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_n}\end{bmatrix}$$

↓ 该方向为 f

→ 该方向为 x

2、概率论中那些进阶点的知识（当然也有基础的）

(x, y)二维随机变量　$cov(X,Y)=E\big((X-EX)(Y-EY)\big)$　$cov(x,x)=var(x)$

$cov(X,Y)=E(XY)-E(X)\cdot E(Y)$　若 x, Y 独立 则 $cov(X,Y)=0$

$var(X\pm Y)=var(x)+var(Y)\pm 2cov(X,Y)$

$corr(X,Y)=\dfrac{cov(X,Y)}{\sqrt{var(x)\cdot var(Y)}}$　若 $cov(X,Y)=0$ 则 X, Y 不一定独立 但一定不相关

设 x 是 n 维随机变量 $X=\begin{bmatrix}X_1\\ \vdots\\ X_n\end{bmatrix}$　$EX=\begin{bmatrix}EX_1\\ \vdots\\ EX_n\end{bmatrix}$

$$cov(X)=E\big(\underset{n\times 1}{(X-EX)}\underset{1\times n}{(X-EX)^T}\big)=\begin{bmatrix}Var(x_1) & cov(x_1,x_2) & \cdots & cov(x_1,x_n)\\ cov(x_2,x_1) & var(x_2) & \cdots & cov(x_2,x_n)\\ \vdots & \vdots & & \vdots\\ cov(x_n,x_1) & cov(x_n,x_2) & \cdots & Var(x_n)\end{bmatrix}$$

这个矩阵是 X 的方差-协方差矩阵　性质：非负定，对称

n 维随机变量 $X=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$　联合密度函数 $P(x_1,x_2\cdots x_n)$

$$\begin{cases}y_1=g_1(x_1,x_2\cdots x_n)\\ \vdots\\ y_n=g_n(x_1,x_2\cdots x_n)\end{cases}$$ 有连续偏导 且 $$\begin{cases}x_1=h_1(y_1\cdots y_n)\\ \vdots\\ x_n=h_n(y_1\cdots y_n)\end{cases}$$ 唯一 (逆唯一)

$$J=\left|\frac{\partial x}{\partial y}\right|=\begin{vmatrix}\frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1}\\ \vdots & & \vdots\\ \frac{\partial x_1}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n}\end{vmatrix}$$ → 行列式

$P_Y(y_1,y_2\cdots y_n)=P_X(h_1,h_2,\cdots h_n)\cdot|J|$

若 $X\sim N_n(\vec{\mu},\Sigma)$ 对于任意常数矩阵 $A_{m\times n}$，有 $Y=AX\sim N_m(A\mu, A\Sigma A^T)$

m 维常向量　$y=Ax+b\sim N_m(A\mu+b, A\Sigma A^T)$

$X=\begin{bmatrix}X_1\\ \vdots\\ X_n\end{bmatrix}$　$\mu=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_n\end{bmatrix}=\begin{bmatrix}E(x_1)\\ \vdots\\ E(x_n)\end{bmatrix}$　$\Sigma=\begin{bmatrix}\delta_{11} & \cdots & \delta_{1n}\\ \vdots & & \vdots\\ \delta_{n1} & \cdots & \delta_{nn}\end{bmatrix}$　每个分量 $x_i\sim N(\mu_i,\delta_{ii})$

$X=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$ $y=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$ $A\in R_{m\times n}$ $B\in R_{k\times n}$ m 维常向量 c.

$E(AX+C)=AE(X)+C$　$Var(AX+C)=A\,Var(X)\cdot A^T$

$Cov(AX,By)=A\,Cov(X,y)\,B^T$

$Cov(X,X)=Var(X)$　方差-协方差矩阵.

3、相关统计基本知识（尤其关于数理统计）

$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$　$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$.　$X\sim N(\mu,\sigma^2)$　$\frac{X-\mu}{\sigma}\sim N(0,1)$　$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$

① 卡方分布 $X_1,X_2\cdots X_n\sim N(0,1)$ 皆来自于 $N(0,1)$

且独立同分布　$\chi^2=X_1^2+X_2^2+\cdots+X_n^2\sim\chi^2(n)\rightarrow E(\chi^2)=n$.　$Var(\chi^2)=2n$.

$\chi^2(n)=Ga(\frac{n}{2},\frac{1}{2})$　$\chi^2(1)=Ga(\frac{1}{2},\frac{1}{2})$

② F 分布　$X_1\sim\chi^2(m)$　$X_2\sim\chi^2(n)$　X_1,X_2 独立.　$F=\frac{X_1/m}{X_2/n}\sim F(m,n)$

$F_\alpha(m,n)=\frac{1}{F_{1-\alpha}(n,m)}$

③ t 分布　X_1,X_2 独立　$X_1\sim N(0,1)$　$X_2\sim\chi^2(n)$　$t=\frac{X_1}{\sqrt{\frac{X_2}{n}}}\sim t(n)$

正态分布下的样本.　$X_1,X_2\cdots X_n$ from $N(\mu,\sigma^2)$

样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$　$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$.　$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$

$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2(n-1)$　$t=\frac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t(n-1)$

$\bar{X}$ 与 S^2 独立.

Δ

引入 设 $X_1,X_2\cdots X_m$ 来自 $N(\mu_1,\sigma_1^2)$　$y_1,y_2\cdots y_n$ from $N(\mu_2,\sigma_2^2)$

两样本相互独立.　$\cancel{S_x^2}$　$\frac{(m-1)S_x^2}{\sigma_1^2}\sim\chi^2(m-1)$　$\frac{(n-1)S_y^2}{\sigma_2^2}\sim\chi^2(n-1)$

$F=\frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2}\sim F(m-1,n-1)$　若 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时

$S_w^2=\frac{(m-1)S_x^2+(n-1)S_y^2}{m+n-2}$.　然后可得　$\frac{(\bar{X}-\bar{y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(n+m-2)$

这里省了些步骤.

"假设检验"这里你只要这样想：假设条件是成立的，然后在此基础之上看偏的程度大不大.
(H_0)

$rank(X)\geq rank(X^TX)$

$\cancel{H=X(X^TX)^{-1}X^TX}$　$H=X(X^TX)^{-1}X^T$ 的性质：

H 是 n 阶对称阵　$H=H^2$.

$Tr(H)=P+1\longrightarrow Tr(H)=Tr(X(X^TX)^{-1}X^T)$
$=Tr((X^TX)^{-1}X^TX)=Tr(I_{P+1})$

3

若 $X \sim N_n(\mu, \delta_1^2 I_n)$ 若 C 是个对称矩阵且 $rank(C)=r$，$C^2=C$ $(r \le n)$

那么 $\frac{X^TCX}{\delta_1^2} \sim \chi^2(r, \delta_2)$ $\delta_2 = \frac{1}{\delta_1^2}\mu^TC\mu$

A 为 n 阶对称阵
B 为 $m \times n$ 阵 $\Leftrightarrow$ BX 和 X^TAX 相独立
$BA = 0_{m\times n}$

A, B 为 n 阶对称矩阵 $\Leftrightarrow$ X^TAX 与 X^TBX 相独立
$AB = 0_{n \times n}$

$X \sim N_n(\mu, \Sigma)$ 给定 $m \times n$ 常矩阵 A
和 m 维常向量 b

$$y = AX + b \sim N_m(A\mu + b, A\Sigma A^T)$$

正态假设下，$y \sim N(X\beta, \delta^2 I_n)$
则 $\hat{\beta} \sim N(\beta, \delta^2(X^TX)^{-1})$
(因为 $\hat{\beta}$ 与 y 的线性组合)

$$\Sigma = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1p} \\ \delta_{21} & \delta_{22} & & \\ & & \cdots & \delta_{pp} \end{bmatrix}$$

其中每个分量 X_j
$X_j \sim N(\mu_j, \delta_{jj})$
也满足正态分布

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

n 阶方阵 A. A^* 伴随阵 $= \begin{bmatrix} A_{11} & \cdots & A_{n1} \\ A_{1n} & A_{2n} \cdots & A_{nn} \end{bmatrix}$ $A^{-1} = \frac{A^*}{|A|}$

A 是 n 阶实对称矩阵 <u>特征值分解</u> $A = V\Lambda V^T$
now;

$\Lambda = diag[\lambda_1 \cdots \lambda_n]$
$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$

$V = [\vec{V}_1 \cdots \vec{V}_n]$
$\vec{V}_i$ 是 λ_i 所对特征向量

$|A| = \prod_{i=1}^{n} \lambda_i$

假设 A 对称
X 是 p 维随机向量
$\mu = E(X)$ $\Sigma = Var(X)$
那么 $E(X^TAX) = Tr(A\Sigma) + \mu^TA\mu$

- 方差分析
- 线性回归
- 变量选择
- 多重共线性
- 聚类

统计方法几大块内容.

方差分析 Note. 2023.12.7.

她今年只讲了单因子方差分析，没有讲多因子~.

二样本独立t检验

比较两个方差相等的独立正态分布均值是否相等.

样本1: $x_{11}, x_{12}, \dots, x_{1m}$ ← 相互独立

样本2: $x_{21}, x_{22}, \dots, x_{2n}$

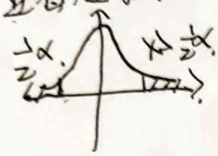
样本1: $N(\mu_1, \sigma^2)$ 2: $N(\mu_2, \sigma^2)$

$H_0: \mu_1=\mu_2$ $H_1: \mu_1 \neq \mu_2$

检验统计量 $t=\dfrac{\bar{x}_1-\bar{x}_2-(\quad)}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$ H_0成立,为0了

其中 $S_w^2=\dfrac{(m-1)S_x^2+(n-1)S_y^2}{m+n-2}$

显著性水平α. 拒绝域



$|t| \geq t_{1-\frac{1}{2}\alpha}(m+n-2)$

p值 $p=2P(t>|t_0|)$

这个t_0为代入 $\bar{x}, \bar{y}, s_w$ 算出来的那个t值.

若 $p<\alpha$, 说明落入拒绝域要拒绝; $p \geq \alpha$, 不拒绝接收.

单因子方差分析模型

响应/因变量 y

因子 a 你可以认为有多少个组(多少个总体,比如上面的例子 a=2)

重复数(每组)样本量一定，都假设为 m. 样本总量 n=am.

水平	响应变量(观测到的)	总和	均值
1	$y_{11} \cdots y_{1m}$	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	$y_{21} \cdots y_{2m}$	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
⋮			⋮
a	$\cdots\cdots$	$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

$y_{i\cdot}=\sum_{j=1}^{m} y_{ij}$

$\bar{y}_{i\cdot}=\dfrac{y_{i\cdot}}{m}$

$\boxed{\text{均值模型}}$ $y_{ij}=\mu_i+\varepsilon_{ij}$

$i=1\cdots a \quad j=1\cdots m$

ε_{ij} 随机误差 $E(\varepsilon_{ij})=0$

$E(y_{ij})=\mu_i$ 第i个水平的均值.

$\boxed{\text{效应模型}}$ $y_{ij}=\mu+\underbrace{\alpha_i}_{\text{效应}}+\varepsilon_{ij}$

$i=1\cdots a \quad j=1\cdots m$

参数数量多了,

约束 $\sum_{i=1}^{a}\alpha_i=0$

$\mu=\frac{1}{a}\sum_{i=1}^{a}\mu_i$ 总体均值.

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

不同水平响应变量方差是一致的.

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

假设检验的问题是什么？

(均值模型) $H_0: \mu_1 = \cdots = \mu_a$ VS H_1 $\exists i \neq j, i,j \in [1,a]$ 使得 $\mu_i \neq \mu_j$

(效应~) $H_0: \alpha_1 = \alpha_2 = \cdots \alpha_a = 0$ VS $H_1: \cdots$

两种模型等价 怎样检验它们？

平方和分解公式 $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{..})^2$

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

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

其中 $2\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) = 2\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..}) \cdot \boxed{\left(\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.})\right)}$

然后 $\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.}) = \sum_{j=1}^{m} y_{ij} - m\sum_{j=1}^{m}\bar{y}_{i.} = 0$ 故 $\boxed{} = 0$

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

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

SSA 有不同水平引发的数据差异，还有随机误差的影响

SS_A 组间偏差平方和

组内偏差平方和 SS_E (同一水平)

给定了一组数据，SST 不变。~~原~~ 若 H_0 成立，SSA 应该不太大

SSA 很大，说明偏的很厉害 拒绝H_0。

检验统计量构造 $\frac{SSA}{SSE}$

结论

$\frac{SS_E}{\delta^2} \sim \chi^2(n-a)$

H_0 成立时 $\frac{SS_A}{\delta^2} \sim \chi^2(a-1)$

$\boxed{SS_A 与 SS_E 独立}$

取一个水平 $x_1 \cdots x_m \overset{iid}{\sim} N(\mu, \delta^2)$

$\frac{(m-1)}{\delta^2} S_x^2 \sim \chi^2(m-1)$ 然后 χ^2 有可加性

有 a 个水平 故一堆 $\frac{\sum_{i}^{a}\sum_{j}^{m}}{\delta^2} = \frac{SS_E}{\delta^2} = \cancel{\frac{S_E}{\delta^2}}$

$\sim \chi^2(am-a)$

$= \chi^2(n-a)$

这就记下来了。

还要验证(效应模型)

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

$$= \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{m}(\varepsilon_{ij} - \underbrace{\bar{\varepsilon}_{i.}}_{均值})^2}_{样本方差 \cdot (m-1)}$$

$\varepsilon_{ij} \sim N(0, \delta^2)$

$\bar{\varepsilon}_{i.} = \frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij}$ 可认为是 $\varepsilon_{i1} \cdots \varepsilon_{im}$ 的样本均值，再用结论。

$$SS_A = m\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= m\sum_{i=1}^{a}\left(\frac{1}{m}\sum_{j=1}^{m}(\mu + \alpha_i + \varepsilon_{ij}) - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}(\mu + \alpha_i + \varepsilon_{ij})\right)^2$$

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

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

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

$\varepsilon_{ij} \sim N(0, \delta^2)$ 相互独立

交叉项 $E\left(2m\sum_{i=1}^{a}\alpha_i(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})\right)=2m\sum_{i=1}^{a}\alpha_i E(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})=0$

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

$\bar{\varepsilon}_{i\cdot}\sim N(0,\frac{\sigma^2}{m})$ $\bar{\varepsilon}_{\cdot\cdot}\sim N(0,\frac{\sigma^2}{n})$

若 H_0 成立 $\alpha_1=\alpha_2=\cdots\alpha_a=0$ 故 $\frac{SS_A}{\sigma^2}\sim\chi^2(a-1)$

$$\underbrace{E(SS_A)=(a-1)\sigma^2+m\sum_{i=1}^{a}\alpha_i^2}_{\text{无论}H_0\text{是否成立}}$$

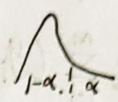
SS_A 与 SS_E 独立 $SS_A\perp SS_E$

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

可以是 $\bar{\varepsilon}_{1\cdot},\cdots\bar{\varepsilon}_{a\cdot}$ 的函数

$\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot})^2$ 与 $\bar{\varepsilon}_{i\cdot}$ 相互独立，且…… 故 SS_A 与 SS_E 独立

检验统计量 $F_A=\dfrac{\frac{SS_A}{a-1}}{\frac{SS_E}{n-a}}\sim F(a-1,n-a)$



若 $F_A\geq F_{1-\alpha}(a-1,n-a)$ 拒绝 H_0

方差分析表

				F值
因子A	SS_A	$a-1$	$MS_A=\frac{SS_A}{a-1}$	$F_A=\frac{MS_A}{MS_E}$
误差E	SS_E	$n-a$	$MS_E=\frac{SS_E}{n-a}$	
总和	SS_T	$n-1$		

P值 $P_A=P(F\geq F_A)$ 若 $P_A<\alpha$，拒绝 H_0

参数估计是什么？就是 $\mu,\alpha_1,\alpha_2\cdots\alpha_a,\sigma^2$？

点估计 极大似然估计结果 $\hat{\mu}=\bar{y}_{\cdot\cdot}$ $\hat{\alpha}_i=\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}$ （其实 $\hat{\mu}_i=\bar{y}_{i\cdot}$
$\hat{\mu}_i=\hat{\mu}+\hat{\alpha}_i$）

$\hat{\sigma}^2_{MLE}=\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{SS_E}{n}$ 这样就无偏了

$E(SS_E)=(n-a)\sigma^2$

$\frac{SS_E}{\sigma^2}\sim\chi^2(n-a)$ 故常用 $\hat{\sigma}^2=\frac{SS_E}{n-a}=MS_E$

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

$\frac{SS_E}{\sigma^2}\sim\chi^2(n-a)$ $\bar{y}_{1\cdot},\bar{y}_{2\cdot},\cdots\bar{y}_{a\cdot}$ 与 SS_E 独立

$$t=\frac{\frac{\sqrt{m}(\bar{y}_{i\cdot}-\mu_i)}{\sigma}}{\sqrt{\frac{SS_E}{\sigma^2(n-a)}}}=\frac{\sqrt{m}(\bar{y}_{i\cdot}-\mu_i)}{\sqrt{\frac{SS_E}{n-a}}}\sim t(n-a)$$

区间估计 μ_i 的 $1-\alpha$ 置信区间

$$\left[\bar{y}_{i\cdot}-\boxed{t_{1-\frac{1}{2}\alpha}(n-a)}\hat{\sigma},\ \bar{y}_{i\cdot}+\boxed{t_{1-\frac{1}{2}\alpha}(n-a)}\cdot\hat{\sigma}\right]$$ $\hat{\sigma}^2=\frac{SS_E}{n-a}$

单因子方差分析模型中，经检验A是显著的（F检验拒绝原假设）

有充分理由，认为因子A的各水平中至少存在一对水平的均值不相等
但这不认定所有水平的均值都不相等

我们想知道哪些水平的均值不相等　　一对水平(i,j)　　$\mu_i-\mu_j$的区间估计？

$\bar{y}_{i\cdot}\sim N(\mu_i,\frac{\sigma^2}{m})$　$\bar{y}_{j\cdot}\sim N(\mu_j,\frac{\sigma^2}{m})$

$\bar{y}_{i\cdot}$和$\bar{y}_{j\cdot}$相独立　　$\bar{y}_{i\cdot}-\bar{y}_{j\cdot}\sim N(\mu_i-\mu_j,\frac{2\sigma^2}{m})$

但这个分布σ^2未知，用$\hat{\sigma}^2=\frac{SSE}{n-a}$代替！但不能简单替换！

替换方法，配凑

$$\frac{(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})-(\mu_i-\mu_j)}{\sigma\sqrt{\frac{2}{m}}}\sim N(0,1)$$　　$\frac{SS_E}{\sigma^2}\sim\chi^2(n-a)$

$$\frac{1}{\sigma}\sqrt{\hat{\sigma}^2}\Leftarrow\sqrt{\frac{SS_E}{\sigma^2(n-a)}}\sim\sqrt{\frac{\chi^2(n-a)}{n-a}}$$

即 $$\frac{(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})-(\mu_i-\mu_j)}{\sqrt{\frac{2}{m}}\cdot\hat{\sigma}}\sim t(n-a)$$

置信区间 $(\mu_i-\mu_j)$　$(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})\pm\sqrt{\frac{2}{m}}\hat{\sigma}\,t_{1-\frac{1}{2}\alpha}(n-a)$

若这个区间覆盖0，说明μ_i和μ_j无明显差异。

多重比较　因子A有a种水平　$C_a^2=\frac{a(a-1)}{2}$对

(Bonferroni方法) 得到的置信区间保守，精度差，但easy。

将$t_{1-\frac{1}{2}\alpha}(n-a)$调整为$t_{1-\frac{1}{a(a-1)}\alpha}(n-a)$

(~~☆~~ Tukey方法)　$H_0^{ij}:\mu_i=\mu_j$　$1\le i<j\le a$

H_0^{ij}成立时$|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|$不应太大，太大就拒绝$H_0^{ij}$

拒绝域的形式　$W=\bigcup_{1\le i<j\le a}\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\ge C_{ij}\}$　$P(W)=\alpha$

求C_a^2个C_{ij}太难了！对问题进行一些简化

各水平重复次数(m)相等，基于对称性，假设C_{ij}统一，记为C

$$P(W)=P\Big(\bigcup_{1\le i<j\le a}\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\ge c\}\Big)=1-P\Big(\bigcap_{1\le i<j\le a}\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|<c\}\Big)$$

$$=1-P\Big(\max_{1\le i<j\le a}|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|<c\Big)=P\Big(\underbrace{\max_{1\le i<j\le a}|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|}_{\text{找最大区间}}\ge c\Big)$$

$$P(\max_{1\le i<j\le a}|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\ge c) = P\left(\max_{1\le i<j\le a}\left|\frac{(\bar{y}_{i\cdot}-\mu)-(\bar{y}_{j\cdot}-\mu)}{\frac{\hat{\delta}}{\sqrt{m}}}\right| \ge \frac{c}{\frac{\hat{\delta}}{\sqrt{m}}}\right)$$

$$= P\Big(\underbrace{\max_i \sqrt{m}\frac{\bar{y}_{i\cdot}-\mu}{\hat{\delta}} - \min_i \sqrt{m}\frac{\bar{y}_{i\cdot}-\mu}{\hat{\delta}}}_{\text{找极差}} \ge \frac{c\sqrt{m}}{\hat{\delta}}\Big)$$

$q(a,df) =$ 那个极差　由于 $\sqrt{m}\frac{\bar{y}_{i\cdot}-\mu}{\hat{\delta}} \sim t(n-a)$

故 $q(a,df)$ 指 a个 独立 同分布 自由度为 df 的 t分布 的随机变量的极差

↳ t化极差统计量　$q(a,n-a)$

算法：水平数目 a，t分布的自由度 df，重复次数 N 给定，

↓然后，去求 N 个观测到的 t化极差统计量

for $n=1\to N$

从 $N(0,1)$ 产生 a 个随机数 $x_1, x_2 \cdots x_a$

$x_1 \cdots x_a$ sort，$\to x_{max}$ x_{min}

再从 df 自由度 χ^2 分布产生随机数 y

$q_n = \frac{x_{max}-x_{min}}{\sqrt{y}}\cdot\sqrt{df}$

....

$P(W)=\alpha$

$P(W) = P\left(q(a,df) \ge \sqrt{m}\frac{c}{\hat{\delta}}\right)$

可得 $\boxed{q_{1-\alpha}(a,df)\frac{\hat{\delta}}{\sqrt{m}} = c}$

注意这里为 $\hat{\delta} = \sqrt{\frac{SSE}{n-a}}$

知道了 c 值，若 $|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}| \ge c$，拒绝 H_0^{ij}，否则 ...

第1块就结束了，开始第2块：多元线性回归（今年没讲一元）

线性回归模型：$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$

y 响应变量　x 自变量　$\begin{bmatrix}\beta_0\\ \vdots\\ \beta_p\end{bmatrix}$ 未知参数 $p+1$ 维，

ε 随机误差

$E(\varepsilon)=0$

$Var(\varepsilon)=\delta^2$.

n组观测数据

$$y=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}\quad X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}\quad \beta=\begin{bmatrix}\beta_0\\ \vdots\\ \beta_p\end{bmatrix}\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix}$$

假定

$rank(X)=p+1<n$

(自变量之间不相关)

样本量 $n>$ 自变量个数 $(p+1)$

X 满秩！

$y=X\beta+\varepsilon$.　β 如何估计？

$E(\varepsilon_i)=0$　$i=1,2\cdots n$ 无系统误差

$cov(\varepsilon_i,\varepsilon_j)=\begin{cases}\delta^2 & i=j\\ 0 & i\ne j\end{cases}$　$\varepsilon\sim N(\vec{0},\delta^2I_n)$

$\varepsilon_i\sim N(0,\delta^2)$　$\varepsilon_1,\cdots\varepsilon_n$ 相互独立.　$E(y)=X\beta$　$Var(y)=\delta^2I_n$.

最小二乘估计. 离差. $\underset{观测}{y_i} - \underset{估计}{x_i^T\beta}$ 离差最小化. $\hat{\beta}_{LS} = \underset{\beta}{\arg\min} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$.

$\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = ||y - X\beta||^2 = (y - X\beta)^T(y - X\beta) = \beta^T X^T X\beta - 2\beta^T X^T y + y^T y$

$||x||^2 = x^T x$ 求导可得 $\hat{\beta}_{LS} = (X^TX)^{-1}X^Ty$，这就是最小二乘估计的值.

但是这个能直接拿来用吗？No! $(X^TX)^{-1}$ 必须存在，且 $|X^TX| \neq 0$

$X \in R^{n\times(p+1)}$ X^TX 为 $p+1$ 阶满秩阵. $rank(X^TX) = p+1$. 那么 $rank(X) \geq rank(X^TX) = p+1$

X 为 $n\times(p+1)$ 阵，故 $n \geq p+1$

[样本量大于参数量]

回归值/拟合值. $\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{=H \text{ 帽子矩阵}}y = Hy$

~~参~~残差？实际与估计之差 $e = y - \hat{y} = (I - H)y$

$\hat{y}^Te = (Hy)^T(I-H)y = y^TH^T(I-H)y = 0$ 回归值 $\hat{y}$ 与 e 垂直

方差-协方差矩阵？

$Var(e) = cov(e,e) = (I-H)cov(y,y)(I-H)^T$

$= (I-H)\cdot\sigma^2 I_n(I-H)^T = \sigma^2(I-H)(I-H)^T$

$(I-H)^T = (I-H)$ $(I-H)^2 = I - H$ 故该式 $= \sigma^2(I-H)$

方差 σ^2 的估计？ $\hat{\sigma}^2_{LS} = \frac{1}{n-p-1}(e^Te) = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$.

why? 因为 $I-H$ 中 I 为 $n\times n$ 阵. $rank(I-H) = Tr(I-H)$

$= Tr(I) - Tr(H)$

$= n - (p+1) = n - p - 1$

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

$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ 联合密度函数为 $f(y;\beta,\sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2 I_n|^{\frac{1}{2}}} \cdot \exp\{-\frac{1}{2}(y - X\beta)^T(\sigma^2 I_n)^{-1}(y - X\beta)\}$

似然函数为 $(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\}$

--- 极大似然估计 $\hat{\beta}_{ML} = (X^TX)^{-1}X^Ty = \hat{\beta}_{LS} = \hat{\beta}$ 记作.

但是 $\hat{\sigma}^2_{ML} = \frac{1}{n}(e^Te)$ $\hat{\sigma}^2_{ML}$ 有偏，但 $\hat{\sigma}^2_{LS}$ 无偏

☆ 对 $\hat{\beta} = (X^TX)^{-1}X^Ty$ 而言. $E(\hat{\beta}) = \beta$. $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

$E(\hat{\beta}) = E((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^TE(y) = (X^TX)^{-1}X^TE(X\beta + \varepsilon)$

$= (X^TX)^{-1}X^T(X\beta + E(\varepsilon)) = (X^TX)^{-1}X^TX\beta = \beta$

$$Var(\hat\beta) = Var((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^T \cdot Var(y)\cdot X(X^TX)^{-1}$$
$$= (X^TX)^{-1}X^T\cdot \sigma^2 I \cdot X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

☆ 还有些性质，如 $\hat\beta$ 和 e 线性不相关

$Cov(\hat\beta, e) = 0$ （正态分布假定下，$\hat\beta$ 与 e 独立，此时 $\hat\beta$ 与 $SS_E = e^Te$ 也独立）

$$Cov(\hat\beta, e) = cov((X^TX)^{-1}X^Ty, \underbrace{\cancel{(X^TX)^{-1}X^T}}_{(I-H)} y) = \sigma^2\cdot (X^TX)^{-1}X^T\cdot(I-H)$$
$$= \sigma^2\cdot((X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T) = \sigma^2\cdot 0 = 0$$

中心化和标准化，请大致清楚推证方法，具体细节不用管。

原模型 $\hat y_i = x_i^T\beta + \varepsilon = \beta_0 + \sum_{j=1} \beta_j \cdot x_{ij} + \varepsilon$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \beta_0 \to \beta_{intercept}, \quad \begin{bmatrix}\beta_1 \\ \vdots \\ \beta_p\end{bmatrix} \to \beta_{slope}$$

$\hat y_i = \hat\beta_{intercept} + x^T\hat\beta_{slope}$

原始 $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $X = [1_n \ \ X_0]$ $X_0 = [x_1 \cdots x_p]$

中心化 $y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix}$ $X^* = [1_n \ \ X_C]$ $X_C = [x_1^* \cdots x_p^*]$

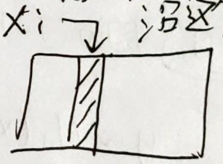
☆ $X_C = (I_n - H_{1n})X_0$

$H_{1n} = 1_n(1_n^T1_n)^{-1}1_n^T$

$\hat\beta_{c,intercept} = 0$

$\hat\beta_{c,slope} = \hat\beta_{slope}$

x_i 沿这个方向进行



$y_i^* = y_i - \bar y$

$\bar y = \frac{1}{n}\sum_{i=1}^n y_i$

$x_{ij}^* = x_{ij} - \bar x_j$

$\bar x_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$

$1_n^T(I_n - H_{1n}) = 0$!

标准化

也是沿这个方向进行

$x_{ij}^{**} = \frac{x_{ij}^*}{\sqrt{L_{jj}}}$ L_{jj}是 x_j 的离差平方和

$L_{jj} = \sum_{i=1}^n (x_{ij} - \bar x_j)^2$

$\bar x_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$

$y_i^{**} = \frac{y_i^*}{\sqrt{L_{yy}}}$

$y^{**} = \frac{1}{\sqrt{L_{yy}}} y^*$ $X_S = X_C L = (I - H_{1n})X_0 L$ $L = \begin{bmatrix} \frac{1}{\sqrt{L_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{L_{pp}}} \end{bmatrix}$

$\hat\beta_{s,intercept} = 0$

$\hat\beta_{s,slope} = \frac{1}{\sqrt{L_{yy}}} L^{-1}\hat\beta_{c,slope} = \frac{1}{\sqrt{L_{yy}}} L^{-1}\hat\beta_{slope}$

分量 $\hat\beta_{sj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat\beta_{cj}$ ☆ $= \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat\beta_j$

$\boxed{\text{上回归}}$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} n & 1_n^T X_0 \\ X_0^T 1_n & X_0^T X_0 \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T \\ X_0^T \end{bmatrix} y$$

$X = [1_n \ \ X_0]$

$X_0 = [x_1 \cdots x_p]$

$$X^T X = \begin{bmatrix} 1_n^T \\ X_0^T \end{bmatrix} [1_n \ \ X_0] = \begin{bmatrix} 1_n^T 1_n & 1_n^T X_0 \\ X_0^T 1_n & X_0^T X_0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n} + \frac{1}{n^2} 1_n^T X_0 A_0 X_0^T 1_n & -\frac{1}{n} 1_n^T X_0 A_0 \\ -\frac{1}{n} A_0 X_0^T 1_n & A_0 \end{bmatrix} \begin{bmatrix} 1_n^T \\ X_0^T \end{bmatrix} y$$

$$= \begin{bmatrix} \frac{1}{n} 1_n^T + \frac{1}{n^2} 1_n^T X_0 A_0 X_0^T 1_n 1_n^T - \frac{1}{n} 1_n^T X_0 A_0 X_0^T \\ -\frac{1}{n} A_0 X_0^T 1_n 1_n^T + A_0 X^T \end{bmatrix} y$$

其中 $A_0 = (X_0^T X_0 - \frac{1}{n} X_0^T 1_n 1_n^T X_0)^{-1}$

然后中心化，走一波类似的操作，X_c 换 X_0 $A_c = (X_c^T X_c - \frac{1}{n} X_c^T 1_n 1_n^T X_c)^{-1} = A_0$

$\beta_{intercept} = (\frac{1}{n} 1_n^T + \frac{1}{n^2} 1_n^T X_0 A_0 X_0^T 1_n 1_n^T - \frac{1}{n} 1_n^T X_0 A_0 X^T) y$

$\beta_{slope} = (-\frac{1}{n} A_0 X_0^T 1_n 1_n^T + A_0 X_0^T) y$ …… 可证…… 具体过程skip.

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$, $\varepsilon \sim N(0, \delta^2)$

这个模型中，y 与 $x_1 \cdots x_p$ 是否存在明显线性关系？ → 检验回归方程（F检验）

若存在，$\beta_0 \cdots \beta_p$ 这些回归系数显著性如何？ → t检验

F检验 看回归方程 → H_0 成立 y与 $x_1 \cdots x_p$ 间用线性模型刻画不合理

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ VS H_1: 存在 $\beta_j \neq 0$ $(j = 1, \cdots p)$

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SS_E}$$

SS_T 离差平方和，SS_E 偏差平方和

正态假设下 $\frac{SS_E}{\delta^2} \sim \chi^2(n-p-1)$ $SS_E = e^T e$. SS_E 与 SS_R 相独立 ↓ 证明略看课件

$y \sim N(X\beta, \delta^2 I_n)$ 若 H_0 成立 $\frac{SS_R}{\delta^2} \sim \chi^2(p)$ 检验统计量 $F_0 = \dfrac{\frac{SS_R}{p}}{\frac{SS_E}{n-p-1}}$

$SS_E = e^T e = y^T (I - H) y$

$y \sim N(X\beta, \delta^2 I_n)$ rank$(I-H) = n-p-1$

$\frac{1}{\delta^2} [E(y)]^T (I-H) E(y) = 0$

$$\boxed{\begin{aligned} SS_R &= y^T H (I - H_{1n}) H y \\ &= y^T (H - H_{1n}) y \\ H H_{1n} &= H_{1n} = H_{1n} H \end{aligned}}$$

故 $\frac{SS_E}{\delta^2} \sim \chi^2(n-p-1)$（看P4的结论） ①若 $F_0 > F_{1-\alpha}(p, n-p-1)$ 拒绝 H_0 ②若 $p_0 = P(F \geq F_0) < \alpha$，拒绝 H_0

来源	平方和	自由度	均方	F值
回归	SS_R	p	$\frac{SS_R}{p}$	$F=\frac{\frac{SS_R}{p}}{\frac{SS_E}{n-p-1}}$
误差	SS_E	$n-p-1$	$\frac{SS_E}{n-p-1}$	
总和	SS_T	$n-1$		

$SS_T=y^T(I_n-H_{1n})y$

$SS_R=y^T(H-H_{1n})y$

$SS_E=y^T(I_n-H)y$

t检验 检验的问题：哪个β_j无意义？

$H_{0j}:\beta_j=0 \quad j=1,2\cdots p$ vs $H_{1j}:\exists j$ 使得$\beta_j\neq 0$

$\hat\beta_j\sim N(\beta_j, c_{jj}\sigma^2)$ 检验统计量 $t_j=\frac{\hat\beta_j}{\sqrt{c_{jj}}\,\hat\sigma}$

其中$\hat\sigma^2=\frac{1}{n-p-1}SS_E$

H_{0j}成立时，$\beta_j=0$ $t_j\sim t(n-p-1)$ 若$|t_j|\geq t_{1-\frac{1}{2}\alpha}(n-p-1)$拒绝$H_{0j}$

认为β_j显著不为0

一元线性回归中，t和F检验等价

多元不等价。

拟合优度 → 度量回归方程对样本观测值的拟合程度。

样本决定系数 $R^2=\frac{SS_R}{SS_T}=1-\frac{SS_E}{SS_T}\in[0,1]$ R^2接近1说明回归拟合效果好。

复相关系数 $R=\sqrt{R^2}$ 衡量作为一个整体的$x_1\cdots x_p$与y线性关系

相关系数衡量的是单个随机变量x_j与y线性关系

给定$x_0=[1\ x_{01}\cdots x_{0p}]^T$ 我们关心的是 $y_0=x_0^T\beta+\varepsilon_0$

基本假定$E(\varepsilon_0)=0$ $Var(\varepsilon_0)=\sigma^2$. $E(y_0)=x_0^T\beta$ $Var(y_0)=\sigma^2$.

点预测$\hat y_0=$ ~~$x_0\hat\beta$~~ $x_0^T\hat\beta$ 如何判断$\hat y_0$的好坏？ 我们的指标 $E(\hat y_0-y_0)^2$.

$E(\hat y_0-y_0)^2=Var(\hat y_0)+Bias^2(\hat y_0)+Var(\varepsilon_0)$ (中间省了许多步)

$\hat y_0$是y_0无偏预测 $E(\hat y_0)=E(y_0)$

$Bias^2(\hat y_0)=0$

$\hat y_0$是y的线性函数，故$\hat y_0$是y_0线性预测

在y_0所有线性，无偏预测中，

$\hat y_0$是否best？

常向量

$\hat\beta$是β的LS估计 对任意c来说 $c^T\hat\beta$ 是$c^T\beta$的最小方差线性无偏估计

y_0的一切线性无偏预测中，$\hat y_0$的方差最小！

$\varepsilon=\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix}\sim N_n(\vec 0,\sigma^2 I_n)$ $\quad\varepsilon_0\sim N(0,\sigma^2)$

$\hat y_0=x_0^T\hat\beta$ 的期望？ $x_0^T\beta$

方差？ $\sigma^2x_0^T(X^TX)^{-1}x_0$

$\hat y_0-y_0\sim N(0,\sigma^2(1+x_0^T(X^TX)^{-1}x_0))$

$\hat y_0-y_0$与$\hat\sigma^2$相互独立。

原因：

$y\sim N_n(X\beta,\sigma^2I_n)$

$\hat\beta\sim N_m(\beta,\sigma^2(X^TX)^{-1})$

$\hat y_0\sim N(x_0^T\beta,\sigma^2x_0^T(X^TX)^{-1}x_0)$ ☆

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

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-p-1)$$

$$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2\left(1 + x_0^T (X^T X)^{-1} x_0\right)\right)$$

$$\frac{\hat{y}_0 - y_0}{\boxed{\hat{\sigma}} \cdot \sqrt{\boxed{1} + x_0^T (X^T X)^{-1} x_0}} \sim t(n-p-1)$$

这个 $\hat{\sigma}^2 = \frac{1}{n-p-1} SS_E$

有1

y_0 的 $1-\alpha$ 置信区间：$\hat{y}_0 \pm t_{1-\frac{1}{2}\alpha}(n-p-1) \cdot \hat{\sigma} \cdot \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$

$E(y_0)$ 的 $1-\alpha$ 置信区间 $\hat{y}_0 \pm t_{1-\frac{1}{2}\alpha}(n-p-1) \cdot \hat{\sigma} \cdot \sqrt{x_0^T (X^T X)^{-1} x_0}$

没有1

马上就要开始接下来的重要内容了，我插一句题外话

就是 11在11.7说GMM推导过程不作为理论作业，但不妨碍期末考，

变量选择：其实就研究欠拟合和过拟合现象

欠拟合：模型预测偏差大，拟合不足，丢失了重要变量

过~：容纳了过多的不重要变量，泛化能力差，过度拟合

自变量选择方法：全子集回归 p个自变量 1个一组（$C_p^1 = p$）

2个一组（C_p^2）

$$\sum_{k=1}^{p} C_p^k = 2^p$$

p个自变量纳入模型：全模型 $y = X_p \beta_p + \varepsilon$

$$\vec{\beta}_p = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

从p个自变量中选p_1个（$p_1 < p$）

↳ 选模型

为了简化 就认为 $x_1 \cdots x_{p_1}$ 是 $x_1 \cdots x_p$ 的前 p_1 个

$$y = X_{p_1} \beta_{p_1} + \varepsilon$$

$$X_p = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad X_{p_1} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p_1} \\ 1 & x_{21} & \cdots & x_{2p_1} \\ \vdots & & & \\ 1 & \cdots & & x_{np_1} \end{bmatrix}$$

if 全模型正确，却使用了选模型，欠拟合。

if 选模型 true，却 use 全~，过拟合。

全模型时 参数估计 $\hat{\beta}_p = (X_p^T X_p)^{-1} X_p^T y$

$$\hat{\sigma}_p^2 = \frac{SS_E^p}{n-p-1}$$

$$SS_E^p = (y - \hat{y}_p)^T (y - \hat{y}_p)$$

在 $x_{p,0}$ 时预测值 $\hat{y}_0 = x_{p,0}^T \hat{\beta}_p$

选模型与之类似，把p用p_1换即可

假设全模型 true. $y = X_p \beta_p + \varepsilon = [X_{p_1} \ Z] \begin{bmatrix} \beta_{p_1} \\ \gamma \end{bmatrix} + \varepsilon$

$$= X_{p_1} \beta_{p_1} + Z\gamma + \varepsilon$$

却使用选模型（$y = X_{p_1}\beta_{p_1} + \varepsilon$）

则丢失信息

$\mathrm{rank}(X_p) > \mathrm{rank}(X_{p_1})$

$\gamma \neq 0_{p-p_1}$

因为我们错误地选了选模型，所以我们使用的参数估计

$\hat{\beta}_{p_1} = (X_{p_1}^T X_{p_1})^{-1} X_{p_1}^T y$ $\quad \hat{\sigma}_{p_1}^2 = \frac{1}{n-p_1-1} SS_E^{p_1}$ $\quad$ 预测值 $\hat{y}_0 = x_{p_1}^T \hat{\beta}_{p_1}$

$E(\hat{\beta}_{p_1}) = (X_{p_1}^T X_{p_1})^{-1} X_{p_1}^T E(y) = \boxed{\quad} E(X_{p_1}\beta_{p_1} + Z\gamma + \varepsilon)$

??? why

因为真实的还是全模型

$= \boxed{\quad} (X_{p_1}\beta_{p_1} + Z\gamma + E(\varepsilon))$

$= \boxed{\quad} (X_{p_1}\beta_{p_1} + Z\gamma) = \beta_{p_1} + (X_{p_1}^T X_{p_1})^{-1} X_{p_1}^T Z\gamma$

由于 $\gamma \neq 0$ $\hat{\beta}_{p_1}$ 有偏 <u>但假如</u> $X_{p_1}^T Z = 0$，$\hat{\beta}_{p_1}$ 就无偏了！

↳ 通常认为有偏！

$SS_E^{p} = y^T(I - H_{X_p})y \quad SS_E^{p_1} = y^T(I - H_{X_{p_1}})y$

$H_{X_p} = X_p (X_p^T X_p)^{-1} X_p^T = [X_{p_1} \;\; Z] \begin{bmatrix} X_{p_1}^T X_{p_1} & X_{p_1}^T Z \\ Z^T X_{p_1} & Z^T Z \end{bmatrix}^{-1} \begin{bmatrix} X_{p_1}^T \\ Z^T \end{bmatrix}$

$H_{X_p} = \cdots\cdots = \cancel{H_{X_{p_1}} + N_{X_{p_1}}} \ \& \cdots$ $H_{X_{p_1}} +$ 一大串 $\Rightarrow$ 故 $SS_E^{p_1} = SS_E^{p} +$ 一大串

<u>一阵推导，skip了</u>

见slide.

$E(SS_E^{p_1}) = E(SS_E^{p}) + E(\text{一大串}) = \cdots\cdots$

$= (n - p_1 - 1)\sigma^2 + \gamma^T Z^T N_{X_{p_1}} Z\gamma$

$E(\hat{\sigma}_{p_1}^2) = \frac{1}{n-p_1-1} E(SS_E^{p_1}) = \sigma^2 + \frac{\cdots}{n-p_1-1} > \sigma^2$

(因为 $\text{rank}(X_p) > \text{rank}(X_{p_1})$)

可以看到 $\hat{\sigma}_{p_1}^2$ 有偏！

又是一大堆推导(skip了) 结论 全模型true 但错用选模型时 预测值 $\text{var}(\hat{y}_0)$ 比真实值 $\text{var}(\hat{y}_{0,\text{全}})$ 更小！

过拟合推导也类似，skip了！

☆ 大结论	欠拟合	过拟合
点估计	$\hat{\beta}_{p_1}$ 有偏 $\hat{\sigma}_{p_1}^2$ 有偏	$\hat{\beta}_p$ 无偏 $\hat{\sigma}_p^2$ 无偏
预测	有偏预测 ($\hat{y}_0$ 有偏) $\text{var}(\hat{y}_0)$ 更小	无偏预测 ($\hat{y}_0$ 无偏) $\text{var}(\hat{y}_0)$ 更大

对于 $(\beta_{p_1}, 0^T)^T$！

自变量数量适可而止！

但有可能好预测误差更小。

无明显偏差 但会引起较大波动。

自变量选择准则 p个变量都两种情况 [选 / 不选] 2^p种可能

SSE 和 R^2 常用衡量模型拟合数据的好坏
残差平方和 决定系数
它们是否能用于选择自变量? ans: No!

$SSE_E^{p_1} = y^T(I_n - H_{X p_1})y$ $SSE_E^{p_1+1}$ 关系: $SSE_E^{p_1+1} \le SSE_E^{p_1}$
自变量个数↑,残差平方和↓.

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ SST 不变 (给定dataset) 故 关系 $R^2_{p_1+1} \ge R^2_{p_1}$
自变量↑ R^2↑

reason

那我们如何作为变量选择的metrix?

① 修正 $R^2 \longrightarrow \tilde{R}_p^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$ $\tilde{R}^2 \le R^2$.

$\tilde{R}^2$ 随自变量↑并不一定↑. 因为有惩罚项.

δ^2 无偏估计 $\hat{\delta}^2 = \frac{1}{n-p-1} SSE$.

② 赤池信息量 AIC = $-2\ln$(模型极大似然) + 2(模型独立参数个数)

线性模型中 $AIC = -2\ln\left((2\pi e\frac{SSE}{n})^{-\frac{n}{2}}\right) + 2(p+1)$

why? $\propto n\ln\left(\frac{SSE}{n}\right) + 2(p+1)$

取AIC最小者! 就是最优

③ 贝叶斯信息量 BIC = $-2\ln$(模型极大似然) + $\ln(n)$(模型独立参数个数)

线性模型 $BIC = n\ln\left(\frac{SSE}{n}\right) + \ln(n)(p+1)$

取BIC最小者! 就是最优.

④ 马洛斯统计量. 用预测的角度选自变量.

$\hookrightarrow C_p = \frac{SSE^p}{\hat{\delta}^2} - n + 2(p+1) = (n-p-1)\frac{SSE^{p_1}}{SSE^p} - n + 2(p_1+1)$

选 C_p 最小的子集.

先确定准则(如AIC) → 前进法 → 缺点: 某变量可能开始显著之后不显著 ↓ 这是弊端

还可以逐步回归

再从最小的模型开始(从 $x_1 \cdots x_p$ 中确定 x 放入model)

基本思想 有进有出

再确定一个 x_i $i=2\cdots p$ 放入model 以此类推.

后退法 → 类似前进法, 不过是在剔元素. → 缺点: 最初计算量很大, 出去了就回不来了.

逐步回归：自变量一个一个引入，引入时，对已选入的变量逐个确定

若之前引入的自变量因当前自变量引入而导致模型不再优化，将其剔除

反复执行直到加入其他任何一个自变量，模型并不会更优，stop.

或剔除

（多重共线性）

最小二乘估计 $\hat{\beta}=(X^TX)^{-1}X^Ty$ 若 $x_1\cdots x_p$ 完全线性相关 $|X^TX|=0$

这样无法最小二乘估计 $\hat{\beta}$　相关程度很高时，$|X^TX|\approx 0$

$\hat{\beta}$ 可得 但 $var(\hat{\beta})=\delta^2(X^TX)^{-1}$ 精度很低

自变量之间完全不相关情形很少见.　比如就两个自变量 $\beta_0\ \beta_1\ \beta_2$

若有多个自变量，若自变量间相关性弱　若 β_1 和 β_2 相关性↑

↳ 认为符合多元线性回归模型假设　$Var(\hat{\beta}_1)$↑ $Var(\hat{\beta}_2)$↑

不然，认为不符合

可能有过多虚拟变量 $x_m=I$(男性) $x_f=I$(女性)　$x_f=1-x_m$

若J分类，至多可设置J-1个虚拟变量.

（某一变量由其他变量计算而成

度量收入与压力关系，收入 [个人收入 / 家庭~] $\xrightarrow{\text{主成分分析}}$ 合并为收入 并以此为自变量.

多重共线性诊断
- 方差扩大因子法
- 特征值判定法
- 直观判定法

方差扩大因子.　$X=[x_1\cdots x_p]$ 做标准化，记为 X_s

相关系数矩阵 $X_s^TX_s=r$　$C=r^{-1}=(X_s^TX_s)^{-1}$

C 的主对角线元素 $VIF_j=c_{jj}$ 为自变量 x_j 的方差扩大因子

在对 y 和 $x_1\cdots x_p$ 均进行标准化后

$\hat{\beta}_{sj}=\frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j$

使用标准化　未使用标准化

这里 我们仅考虑对 $x_1\cdots x_p$ 标准化　$\hat{\beta}_{sj}=\sqrt{L_{jj}}\hat{\beta}_j$

$Var(\hat{\beta}_j)=\frac{c_{jj}}{L_{jj}}\delta^2$　c_{jj}↑　$L_{jj}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$

x_j 所对应的 $\hat{\beta}_j$ 的 var 也↑

我们若将 x_j 作为因变量，与其他 $p-1$ 个自变量建多元线性回归模型，

R_j^2 为其复决定系数 $\boxed{C_{jj} = \frac{1}{1-R_j^2}}$ 作业题，

R_j^2 越大 C_{jj}↑，多重共线性较大，

↳ 说明 x_j 与其余自变量线性相关程度也越大

取个 C_{VIF}（比如 5, 10, 100）若 $VIF_j < C_{VIF}$，认为 x_j 与其余自变量不存在多重~，否则~

如何度量整个矩阵多重共线性？

$\overline{VIF} = \frac{1}{p}\sum_{j=1}^{p} VIF_j$ $\overline{VIF}$ 大时 标存在…

样本小时 $\overline{VIF}$ 易很大 → 因此要结合样本量讨论

特征值判定法 X 经过标准化，

$|X^TX| \approx 0$ 时 X^TX 至少有一个特征值近似于 0

X^TX 特征值为 $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ 取 $k_j = \sqrt{\frac{\lambda_1}{\lambda_j}}$ 为条件数 $j=1\cdots p$

最小特征值 λ_p 对应的条件数 $k_p = \sqrt{\frac{\lambda_1}{\lambda_p}}$ 若接近无穷（很大），X 存在多重共线性

否则……，X 不存在…

若 $0 < k_p < C_K$ 时，X 不存在… 否则 X 存在 C_K 常取 10, 100, 1000

直观判定法 增/删一个自变量时，其他自变量回归系数估计值/显著性发生较大变化，

↳ 存在严重~

若定性分析认为重要自变量在回归方程中从未通过显著性检验

↳ 初步认为，存在严重…

若与因变量之间简单相关系数绝对值较大的自变量在回归方程中未通过显著性检验

还有些 若…，判定…，此处不多说了，见 slides!

消除多重共线性 (1). 删不重要变量 (2) 增样本量 n. (3) 改进经典最小二乘估计

↙ 岭回归 ↘ 主成分回归

假定 X 标准化了，y 未标准化，

岭回归 改进一下传统最小二乘估计 → $\begin{cases} \hat{\beta} = (X^TX)^{-1}X^Ty \\ Var(\hat{\beta}) = \sigma^2(X^TX)^{-1} \end{cases}$

↓ 用 $X^TX + kI$ ($k>0$) 代替 X^TX

k 较小时 $X^TX + kI \approx X^TX$ 且 $X^TX + kI$ 避免奇异，

$\hat{\beta}(k)=(X^TX+kI)^{-1}X^Ty$ 为 β 的岭回归估计，k: 岭参数

X 已经标准化 X^TX 就是自变量样本的相关系数矩阵

k 不唯一确定！

性质① $\hat{\beta}(k)$ 是 β 的有偏估计

$$E(\hat{\beta}(k))=E((X^TX+kI)^{-1}X^Ty)=(X^TX+kI)^{-1}X^TE(y)$$
$$=(X^TX+kI)^{-1}X^TX\beta\neq\beta$$

性质② $\hat{\beta}(k)$ 与 $\hat{\beta}(0)$（最小二乘估计）什么关系？

$$\hat{\beta}(k)=(X^TX+kI)^{-1}X^Ty=(X^TX+kI)^{-1}(X^TX)(X^TX)^{-1}X^Ty$$
$$=(X^TX+kI)^{-1}(X^TX)\hat{\beta}(0)$$

看似 $\hat{\beta}(k)$ 是 $\hat{\beta}$ 的一种线性变换 但实际 k 也由数据确定

（$\hat{\beta}(k)$ 并非 $\hat{\beta}$ 的线性变换

也并非 y 的线性函数）

性质③ ~~岭回归~~ $\hat{\beta}(k)$ 是否比 $\hat{\beta}(0)$ 更优？

利用均方误差作为指标

$\exists k>0$ 使得 $MSE(\hat{\beta}(k))\leq MSE(\hat{\beta}(0))$ 证明见slides，略！

$$MSE(\hat{\beta}(k))=E((\hat{\beta}(k)-\beta)^T(\hat{\beta}(k)-\beta))$$

X^TX 特征值 $\lambda_1\geq\lambda_2\geq\cdots\lambda_p>0$ 特征向量 $v_1\cdots v_p$

性质
$$\begin{cases}(X^TX)v_j=\lambda_jv_j \quad j=1\cdots p\\(X^TX+kI)v_j=(X^TX)v_j+kv_j=\lambda_jv_j+kv_j=(\lambda_j+k)v_j\\\text{故}\ (X^TX+kI)^{-1}v_j=\dfrac{1}{\lambda_j+k}v_j\end{cases}$$

k 的选取与 δ^2，β 都有关；无法找一个一一对应的 k 使得 $MSE(\hat{\beta}(k))_{min}$

另一角度看岭回归 最小二乘估计：$\hat{\beta}=\arg\min_{\beta}(y-X\beta)^T(y-X\beta)$

无约束条件

$\downarrow$

岭回归估计（加上了惩罚项）

$\hat{\beta}(k)=\arg\min_{\beta}(y-X\beta)^T(y-X\beta)+k\beta^T\beta$

等价于 $\min(y-X\beta)^T(y-X\beta)$

s.t. $\beta^T\beta\leq S$

① 若最小二乘估计 $\hat{\beta}^T\hat{\beta}\leq S$

$\hat{\beta}$ 就是我们想要的

② 若 $\hat{\beta}^T\hat{\beta}>S$

$\hat{\beta}(k)^T\hat{\beta}(k)\leq S<\hat{\beta}^T\hat{\beta}$

性质④ 对 $\forall k>0$. $\|\beta\| \neq 0$ 总有 $\|\hat\beta(k)\| < \|\hat\beta\|$

☆ X^TX 特征值 $\lambda_1 \cdots \lambda_p$ $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$. $V_1 \cdots V_p$ 特征向量

$X^TX = V^T \Lambda V$. $y = X\beta + \varepsilon = XV^TV\beta + \varepsilon = Z\alpha + \varepsilon$.

$\alpha = V\beta$ 可得. $\beta = V^{-1}\alpha = V^T\alpha$.

$\hat\alpha = \underbrace{(Z^TZ)^{-1}Z^Ty}_{\text{why???}} = (VX^TXV^T)^{-1}Z^Ty = (VV^T\Lambda VV^T)^{-1}Z^Ty$
$= \Lambda^{-1}Z^Ty$

~~$\hat\beta = V$~~ $\hat\beta = (X^TX)^{-1}X^Ty = V^T\Lambda^{-1}VX^Ty = V^T\hat\alpha$.

$\hat\alpha(k) = (\Lambda + kI)^{-1}Z^Ty$ $\hat\beta(k) = V^T\hat\alpha(k)$

故 $\|\hat\beta(k)\| = \|\hat\alpha(k)\| = \|(\Lambda + kI)^{-1}\Lambda\hat\alpha\| < \|\hat\alpha\| = \|\hat\beta\|$

乘正交阵不改变范数

$\hat\beta(k)$ ~~$\|\hat\beta(k)\|$~~ 是 $\hat\beta$ 向原点的压缩 $MSE(\hat\beta) = E((\hat\beta - \beta)^T(\hat\beta - \beta))$
$= E(\hat\beta^T\hat\beta) - \beta^T\beta$
$= E\|\hat\beta\|^2 - \|\beta\|^2$.

故 $E\|\hat\beta\|^2 = \|\beta\|^2 + MSE(\hat\beta)$
$= \|\beta\|^2 + \delta^2 \sum_{j=1}^{p} \lambda_j^{-1}$

X 出现多重共线性时, $\delta^2 \sum_{j=1}^{p} \lambda_j^{-1}$ 较大, 对其就该做压缩!

k 的选择 ① 岭迹法 取一个 k, $\hat\beta(k) = (X^TX + kI)^{-1}X^Ty$ 的第 j 个分量 $\hat\beta_j(k)$

$k \in [0, +\infty)$ 时. $\underbrace{k - \hat\beta_j(k)}_{\text{有除!}}$ 的图像为岭迹!

看什么时候各岭迹基本稳定.

② 方差扩大因子法. $Var(\hat\beta(k)) = \cancel{\delta^2(X^TX)}$
$\delta^2(X^TX + kI)^{-1}X^TX(X^TX + kI)^{-1}$
$= \delta^2 C(k)$

$C(k)$ 的对角线元素 $C_{jj}(k)$ 为岭估计方差扩大因子.

$k\uparrow$ $C_{jj}(k)\downarrow$. 选择合适的 k 使所有 $C_{jj}(k) \leq CVIF$
: 确定 k!

③ Horel-Kennad 公式 / Mcdorard-Garaneau 公式
HK 公式
MG 公式 → 使得 $\|\hat\beta\|^2 - \|\hat\beta(k)\|^2 \approx \hat\delta^2 \sum_{j=1}^{p} \lambda_j^{-1}$

$k_{HK} = \dfrac{\hat\delta^2}{\max_j \hat\alpha_j^2}$. $\alpha = V\beta$??? 这里要看看!

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

主成分回归。

主成分回归 = 主成分分析 + 回归分析

→ 研究 y 与 $X=[x_1,x_2 \cdots x_p]^T$ 之间关系

即 $y = X^T\beta + \varepsilon$

用 k 个自变量 的
线性变换(主成分)
代替原本的 p 个自变量 $(k<p)$

如何求主成分？ 如何利用主成分估计 β？ 主成分回归 $\hat{\beta}_{PC}$ 与 $\hat{\beta}$ 什么关系？

自变量个数太多 → 我们用较少的综合变量代替原本那么多的自变量
而且这 k 个综合变量还能尽可能 cover 较多原本变量信息

$X=\begin{bmatrix}x_1\\ \vdots \\ x_p\end{bmatrix}$ p 维随机向量 $E(x)=\mu$ $Var(x)=\Sigma$ 方差-协方差阵

$z_1 = a_1^T X = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p$

$z_2 = a_2^T X = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p$

$\cdots$

$z_p = a_p^T X = \cdots$

我感觉这里下标有问题

这问题!

$Var(z_i) = a_i^T \Sigma a_i$ $Cov(z_i, z_j) = a_i^T \Sigma a_j$

我们希望用 z_1 来代替原本的 p 个变量 $x_1 \cdots x_p$

新变量 z_1 尽可能多反映原来 p 个变量信息

我们如何度量信息？ 用方差 $Var(z_1)$ 大

- $Var(z_1) = a_1^T \Sigma a_1$ 对 a_1 作出一些限制 $a_1^T a_1 = 1$

若存在满足以上约束的 a_1 使 $Var(z_1)$ max 则 z_1 为第1主成分。

…进一步得第2主成分，第3主成分，…

有个需要 alert 的，在第1主成分中体现的信息不能出现于第2,3…主成分中

$Cov(z_2, z_1) = 0 = a_2^T \Sigma a_1 = 0$ ……

- $X=\begin{bmatrix}x_1\\ \vdots \\ x_p\end{bmatrix}$ 为 p 维~ $a_i = \begin{bmatrix}a_{1i}\\ \vdots \\ a_{pi}\end{bmatrix}$ 若존 z_i

$z_i = a_i^T X$ 是 X 的线性组合

① 且 $a_i^T a_i = 1$ ② $i>1$ 时 $a_i^T \Sigma a_j = 0$ $(i \neq j)$

③ $Var(z_i) = \max Var(a_i^T x)$
条件 $a_i^T a_i = 1$
$a_i^T \Sigma a_j = 0$ $j=1 \cdots i-1$

则 $z_i = a_i^T X$ 为 X 的第 i 主成分

这只是性质，具体如何求？

若 $E(x)=0$ $Var(x)=\Sigma$

第1主成分求法 $a_1 = \begin{bmatrix}a_{11}\\ \vdots \\ a_{p1}\end{bmatrix}$

$a_1 = \arg\max_a a^T X$

s.t. $a^T a = 1$

$L(a) = Var(a^T x) - \lambda(a^T a - 1)$
$= a^T \Sigma a - \lambda(a^T a - 1)$

$$\begin{cases} \frac{\partial L}{\partial a} = 0 = 2(\Sigma - \lambda I)a \\ \frac{\partial L}{\partial \lambda} = 0 = a^T a - 1 \end{cases}$$ 可得 $|\Sigma - \lambda I| = 0$ (由于 $a \neq \vec{0}$) → 故求第1主成分等价于求Σ的特征值和特征向量.

之后的主成分求法也类似

Σ的特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

$a_1 \cdots a_p$ 为相应的单位正交特征向量

X的第i主成分为 $Z_i = a_i^T X \quad i = 1, 2 \cdots p$.

Σ 主成分总是优先方差较大的变量，有时结果不太合理.

↓ 为了消除量纲不同带来的不合理结果，自变量要标准化

$E(x_i) = \mu_i \quad Var(x_i) = \delta_i^2. \quad x_i^* = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} = \frac{x_i - \mu_i}{\delta_i}$

X^* 的 Σ^* 就是原随机向量X的相关阵 corr(x)

→ 由此求主成分

由 corr(x) 可角定的主成分为

$$Z^* = \begin{bmatrix} Z_1^* \\ \vdots \\ Z_p^* \end{bmatrix} \quad var(Z^*) = \Lambda^* = diag\{\lambda_1^* \cdots \lambda_p^*\}$$

$$\sum_{i=1}^{p} \lambda_i^* = p$$

☆ 之前讨论的问题实际中Σ未知，要通过样本估计呢！

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_n^T \end{bmatrix}$$

Σ的估计 $S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T = (S_{kl})_{p \times p}$

样本协方差

样本相关阵 $R = (r_{kl})_{p \times p} \quad r_{kl} = \frac{S_{kl}}{\sqrt{S_{kk} S_{ll}}}$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

标准化后 X X^*

$$X^* = \begin{bmatrix} x_{11}^* & \cdots & x_{1p}^* \\ \vdots & & \vdots \\ x_{n1}^* & \cdots & x_{np}^* \end{bmatrix} \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{L_{jj}}}$$

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

$(X^*)^T X^*$ 可等价于 R.

$\bar{x}$ 按列平均

$$S_{kl} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$$

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$$

以上都不太有用感觉!! 写到了课件 P105

主成分回归我是真不太懂，所以笔记2页先空了，回头再补吧！

聚类

人类对个体特征进行归纳，相似个体归为一类，以类的特征代替个体信息

聚类：识别从属总体的个体 / 识别异常个体

↳ 形成整体认知

↳ 核心问题：如何定义个体间相似性？如何确定类别数？如何选个体特征？如何评价聚类的结果？

无标签数据集　无监督学习

样本 $x_1 \cdots x_n$　$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- 按行聚类：X中集群发现相似个体
- 按列聚类：对X进行降维

两点距离越小，说明相似度越大

曼哈顿 $\sum_{i=1}^{p} |x_{1i} - x_{2i}|$

欧氏 $\|x_1 - x_2\|_2 = \sqrt{\sum_{i=1}^{p} (x_{1i} - x_{2i})^2}$

切比雪夫 $\max_i |x_{1i} - x_{2i}|$

平方欧氏 $\|x_1 - x_2\|_2^2$

闵氏 $\left(\sum_{i=1}^{p} (x_{1i} - x_{2i})^q\right)^{\frac{1}{q}}$

兰氏 $\sum_{i=1}^{p} \frac{|x_{1i} - x_{2i}|}{|x_{1i}| + |x_{2i}|}$

广义欧氏 $\sqrt{(x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2)}$

余弦相似度 $\cos\theta = \frac{a^T b}{|a||b|}$

皮尔逊相关系数 $r = \frac{\sum_{i=1}^{p} (x_{1i} - \bar{x}_1)\cdot(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{p} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{p} (x_{2i} - \bar{x}_2)^2}}$

$\overset{\mathrm{corr}}{\cancel{\mathrm{cov}}}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x) \cdot \mathrm{var}(y)}}$

一堆距离……

点间距离　$s_j = s_j(x_{kj}, x_{lj})$　x_k 和 x_l 在第 j 个特征之间的相似性

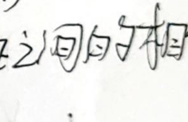
相似度定义　两个观测值 x_k 和 x_l 间相似度

$$s(x_k, x_l) = \frac{\sum_{i=1}^{p} s_i(x_{ki}, x_{li})\,\delta(x_{ki}, x_{li}) \cdot w_i}{\sum_{i=1}^{p} \delta(x_{ki}, x_{li})\, w_i}$$

$$\delta(x_{ki}, x_{li}) = \begin{cases} 0 & x_{ki} \text{ 或 } x_{li} \text{ 中存在缺失观测} \\ 1 & \text{其他} \end{cases}$$

w_i 是权重

类间距离（关联准则）两类中距离最短 / 最长 / 重心



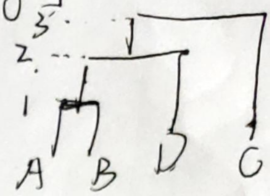
LW公式回答！

层次聚类

- 自下而上　所有样本各自分到一类中，之后将类间距离最近的两类关联起来，建一个新的类，反复，直到所有样本聚合到一个类中
- 自上而下　所有样本归入一个大类，之后在类中距离最远的样本记为两个新类，基于它们将未进行聚类的点逐一与这2个中心点比距离，并放入；之后在更小的类中执行相同操作。

自下而上—层次聚类, 4个点

$$\begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1 & 3 & 2 \\ 1 & 0 & 5 & 6 \\ 3 & 5 & 0 & 4 \\ 2 & 6 & 4 & 0 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \begin{array}{ccc} A,B & C & D \end{array} \\ \begin{bmatrix} 0 & 3 & 2 \\ 3 & 0 & 4 \\ 2 & 4 & 0 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \begin{array}{cc} A,B,D & C \end{array} \\ \begin{bmatrix} 0 & 3 \\ 3 & 0 \end{bmatrix} \end{array}$$



A,B距离min,聚起来

再选A,B,D.

K均值 kmeans 计算快 事先确定聚类数目K

将n个样本划入K类中 $C_i \cap C_j = \emptyset$ $\bigcup_{i=1}^{K} C_i = X$

要找一个最优划分,使类内距离足够小,类间距离足够大.

以平方欧氏距离最通用. 损失函数 $W(C_1 \cdots C_K)$

$$= \sum_{i=1}^{K} \sum_{j \in C_i} \|x_j - m_i\|^2.$$

$\arg\min W(C_1, C_2 \cdots C_K)$

其中m_i为第i类的重心

NP-hard问题,迭代法解

每次迭代2个步骤: ① 确定k个类的重心 m_i

将样本逐一分配到其最近的中心所对类中,得聚类结果

② 更新各类的样本均值(重心) 重复①②直至W变化不大(收敛)

各类数据集非凸,又倒收敛

或聚类结果不变 也可认为收敛

混合高斯模型 GMM 核心: 分布的假定

假定第i个样本x_i来自于第j类正态分布 $N_p(\mu_j, \Sigma_j)$

$$f(x_i) = (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \mu_j)^T \cancel{\Sigma} \Sigma_j^{-1} (x_i - \mu_j)\}.$$

超参数还是K,聚类数目.

$$\delta_{ij} = \begin{cases} 1 & x_i \sim N_p(\mu_j, \Sigma_j) \\ 0 & x_i \nsim N_p(\mu_j, \Sigma_j) \end{cases}$$

$\delta_i = \begin{bmatrix} \delta_{i1} \\ \vdots \\ \delta_{iK} \end{bmatrix}$ 独立同分布 多维分布 $M(1, \pi_1 \cdots \pi_K)$

$f(\delta_i) = \prod_{j=1}^{K} (\pi_j)^{\delta_{ij}}$ $\pi_j = P(\delta_{ij} = 1)$ $i = 1 \cdots n$

给定δ_i后...

$f(x_i | \delta_i) = \prod_{j=1}^{K} \left((2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\{\cdots\}\right)^{\delta_{ij}}$

联合密度函数 $\prod_{i=1}^{n} f(x_i, \delta_i) = \prod_{i=1}^{n} f(\delta_i) \cdot f(x_i | \delta_i)$

$$= \prod_{i=1}^{n} \prod_{j=1}^{K} \left((\pi_j) \cdot (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\{\cdots\}\right)^{\delta_{ij}}$$

要估未知参数 $(\pi_1 \cdots \pi_K, \mu_1 \cdots \mu_K, \Sigma_1 \cdots \Sigma_K)$

现在只有样本 $x_1\cdots x_n$ 无法观测 $\delta_1\cdots\delta_n$ 无法直接估计未知参数.

x_i 的密度函数

$$f(x_i)=f(x_i|\delta_i)f(\delta_i)=\sum_{j=1}^{K}\pi_j(2\pi)^{-\frac{p}{2}}|\Sigma_j|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\}$$

高斯混合模型了 EM算法未解 证明skip了,太长了. 看课件吧!

对于24那个联合密度函数 1吗,再……

DBSCAN 基于密度的聚类算法 要判断两个样本属于同一类 那么在这两个样本附近,能找个属于同一类别的样本.

给定 $\varepsilon>0$ 和最少样本个数MinPts

若 $N_\varepsilon(k)=\{x_L\in X \mid d(x_k,x_L)\le\varepsilon\}$ → 称为 x_k 的 ε 邻域

若 ~~$|N_\varepsilon(\varepsilon)|$~~ $|N_\varepsilon(k)|\ge MinPts$, 称 x_k 为核心点

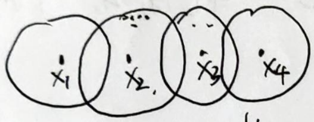
若 $x_L\in N_\varepsilon(k)$ 且 $|N_\varepsilon(k)|\ge MinPts$, x_L 可以从 x_k 直接密度可达

$\exists x_i$: x_i 使得 x_k 和 x_L 均从 x_i 密度可达, 则称 x_k 和 x_L 密度连接.

密度可达: 一堆点 $x_k\cdots x_u$ 使得 x_{i+1} 可从 x_i 直接密度可达 则 x_u 可从 x_1 密度可达

x_k 为隐含核心点

x_u, x_L 不一定.



x_2, x_3 为核心

[x_4 可由 x_2 密度可达] 找这个图就好了.

密度连接 之从 x_i x_j 中间找一点 x_u 使 x_i 和 x_j 均可从 x_u 密度可达

有点抽象.

噪声: 分了若干类 发现有的点没找到任何类!

DBSCAN algorithm:

$X=(x_1\cdots x_n)^T$ ε, MinPts给定

初始化: 类别序号 $k=0$

标记所有样本点为未访问过 $V[i]=0$ $i=0,1\cdots n$

```
for 从X中随机选一个未访问过的样本点xi {
    ~~if xL 未访问~~  xi访问过了 V[i]=1;
    计算 xi 的 ε 邻域 Nε(i)
    if xi 不是核心点 ; xi是噪声
    else {
        k++;  G(i)表示 xi 所有密度可达的点
        for xj ∈ G(i) {
            if(xj 未访问) vis[j]=1 将 xj 归入k类
        }
    }
}
```

聚类有效性评价 $\left[\begin{array}{l}\text{外部聚类有效性} \\ \text{内部}\sim\end{array}\right]$ 区别:是否用外部的信息

n 个样本 $x_1 \cdots x_n$ 分类结果 $C_1 \cdots C_K$, 可有如下矩阵:

假设真实的标签是 $P_1 \cdots P_{K'}$

$$\begin{array}{c|ccc|c} & P_1 & \cdots & P_{K'} & \\ \hline C_1 & n_{11} & \cdots & n_{1K'} & n_{1\cdot} \\ \vdots & \vdots & & & \\ C_K & n_{K1} & \cdots & n_{KK'} & n_{K\cdot} \\ \hline & n_{\cdot 1} & \cdots & n_{\cdot K'} & n \end{array}$$

$$P_{ij} = \frac{n_{ij}}{n} \qquad P_{i\cdot} = P_i = \frac{n_{i\cdot}}{n}$$

$$P_j = \frac{n_{\cdot j}}{n}$$

熵 $E = -\sum_i P_i \left(\sum_j \frac{P_{ij}}{P_i} \ln \frac{P_{ij}}{P_i} \right)$

纯度 $P = \sum_i P_i \left(\max_j \frac{P_{ij}}{P_i} \right)$

内部聚类有效性. (1) 紧密度 (同一类内不同个体之间紧密关联的度量)

低方差 $\Leftrightarrow$ 紧密度好. 很多紧密度定义依赖于距离

(2) 区分度 (不同类间区别程度的度量)

两类中心的距离 $\cdots$ 密度也行.

均方标准差 $RMSSTD = \left(\frac{\sum_{j=1}^{K} \sum_{i \in C_j} \|x_i - m_j\|^2}{P \sum_{j=1}^{K} (n_K - 1)} \right)^{\frac{1}{2}}$ 轮廓法, CH 指数 略!

R方 $RS = 1 - \frac{\sum_{j=1}^{K} \sum_{i \in C_j} \|x_i - m_j\|^2}{\sum_{i=1}^{n} \|x_i - \bar{x}\|^2}$

stat_ml 作业题回顾 2024.1

① 方差分析那里的先跳过，问好看

$I-H$ 幂等阵 $(I-H)^2 = I-H$ $\quad rank(I-H) = Tr(I-H) = Tr(I) - Tr(H) = n-(p+1) = n-p-1$

② 多元线性回归model中，y_i 的回归值 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

X 为满秩矩阵，证明 $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$

因为 $H = H^T$

$y - \hat{y} = y - X\hat{\beta} = y - Hy = (I-H)y$

而 $1^T(y-\hat{y}) = 1^T(I-H)y = (1^T - 1^T H)y = (1^T - (H\cdot\vec{1})^T)y$

然后 $H\cdot\vec{1} = X(X^TX)^{-1}X^T\vec{1}$ $\quad$ X 的第1列一堆1 ($\vec{1}$) 故 ~~$X\cdot 1 = 1$~~

$X\cdot\begin{bmatrix}1\\0\\ \vdots\\0\end{bmatrix} = \vec{1}$

$H\cdot\vec{1} = X(X^TX)^{-1}X^T\vec{1} = X(X^TX)^{-1}X^TX\alpha = X\vec{\alpha} = \vec{1}$

全消了！ (取 $\alpha = \begin{bmatrix}1\\0\\ \vdots\\0\end{bmatrix}$)

故 $1^T(y-\hat{y}) = (1^T - (H\cdot 1)^T)y$

$= (1^T - 1^T)y = \vec{0}$ 原式得证

③ $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ $\longrightarrow$ 得最小二乘估计为 $\hat{\beta}$

若我们对 $y_1, y_2 \cdots y_n$ 中心化，对 $x_{1j}\ x_{2j} \cdots x_{nj}$ $(j = 1, 2 \cdots p)$ 标准化

$\hookrightarrow$ 得最小二乘估计 $\tilde{\beta}$ $\quad$ $\hat{\beta}$ 和 $\tilde{\beta}$ 有何关系？ $E(\tilde{\beta}) = ?$ $Var(\tilde{\beta}) = ?$

原模型 ✱ $y = X\beta + \varepsilon$ $\quad y = \begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$ $\quad X = [\vec{1}_n\ \ X_0]$ $\quad X_0 = [\vec{x}_1 \cdots \vec{x}_p]$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \left[\begin{array}{c|c} n & 1_n^T X_0 \\ \hline X_0^T 1_n & X_0^T X_0\end{array}\right]^{-1}\begin{bmatrix}1_n^T\\ X_0^T\end{bmatrix}y$$

$H_{1n} = \vec{1}_n(1_n^T 1_n)^{-1}1_n^T$

中心化 y 得 $\vec{y^*} = y - H_{1n}y = (I - H_{1n})y$

标准化 $x_{1j}, x_{2j} \cdots x_{nj}$ $\quad$ 得 $X_S = X_C L = (I - H_{1n})X_0 L$

$L = diag\{\frac{1}{\sqrt{L_{11}}}, \cdots \frac{1}{\sqrt{L_{pp}}}\}$

标准化 $y^{**} = \frac{1}{\sqrt{L_{yy}}}y^* = \frac{1}{\sqrt{L_{yy}}}(I - H_{1n})y$

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

$L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$$\cancel{\hat{\beta}_{s,slope} =} \quad \tilde{\beta} = \left[\begin{array}{c|c} n & 1_n^T X_S \\ \hline X_S^T 1_n & X_S^T X_S\end{array}\right]^{-1}\begin{bmatrix}1_n^T\\ X_S^T\end{bmatrix}y^*$$

$$= \begin{bmatrix} \frac{1}{n}1_n^T + \frac{1}{n^2}1_n^T X_S A_1 X_S^T 1_n 1_n^T - \frac{1}{n}1_n^T X_S A_1 X_S^T \\ -\frac{1}{n}A_1 X_S^T 1_n 1_n^T + A_1 X_S^T\end{bmatrix}y^*$$

$$A_1 = [X_S^T X_S - \frac{1}{n}X_S^T 1_n 1_n^T X_S]^{-1}$$

其中 $1_n^T X_s = 1_n^T X_c \cdot L = 1_n^T (I - H_{1n}) X_0 L = 0_{1\times n}$ $\quad 1_n^T H_{1n} = 1_n^T \cdot 1_n \cdot (1_n^T 1_n)^{-1} 1_n^T = 1_n^T$

故 $\frac{1}{n^2} 1_n^T X_s A_1 X_s^T 1_n 1_n^T = 0_{1\times n}$ $\quad X_s^T 1_n = 0_{n\times 1}$

$-\frac{1}{n} 1_n^T X_s A_1 X_s^T = 0_{1\times n}$

故 $\tilde\beta = \begin{bmatrix}\frac{1}{n} 1_n^T \\ A_1 X_s^T - \cancel{\frac{1}{n} A_1}\end{bmatrix} \cancel{(I-)} y^* = \begin{bmatrix}\frac{1}{n} 1_n^T \\ A_1 X_s^T\end{bmatrix}(I - H_{1n}) y = \begin{bmatrix}0 \\ A_1 X_s^T (I - H_{1n}) y\end{bmatrix}$

同理 $A_1 = [X_s^T X_s - \frac{1}{n} X_s^T 1_n 1_n^T X_s]^{-1} = [X_s^T X_s]^{-1}$

故 $A_1 X_s^T (I - H_{1n}) y = (X_s^T X_s)^{-1} X_s^T (I - H_{1n}) y = (X_s^T X_s)^{-1} X_s^T \cdot y^*$

$= \sqrt{L_{yy}} \cdot (X_s^T X_s)^{-1} X_s^T y^{**}$

统一转化为中心化 好看些

正常来说，全面中心化之后解为

$(X_s^T X_s)^{-1} X_s^T y^{**}$

$\hat\beta_{s,slope} = \frac{1}{\sqrt{L_{yy}}} L^{-1} \hat\beta_{slope}$

又对比可知，题中所求 $\hat\beta_{new,slope} = A_1 X_s^T (I - H_{1n}) y = \sqrt{L_{yy}} \cdot (X_s^T X_s)^{-1} \cdot X_s^T y^{**}$

$= \sqrt{L_{yy}} \cdot \hat\beta_{s,slope} = L^{-1} \hat\beta_{slope}$

$\tilde\beta = \begin{bmatrix}0 \\ L^{-1}\hat\beta_{slope}\end{bmatrix}$ $\quad \hat\beta = \begin{bmatrix}\hat\beta_{intercept} \\ \hat\beta_{slope}\end{bmatrix}$ $\quad \tilde\beta = \begin{bmatrix}0_{1\times 1} & 0_{1\times p} \\ 0_{p\times 1} & L^{-1}\end{bmatrix} \cdot \hat\beta$

这便是关系！

$E(\hat\beta) = \beta$ $\quad T = \begin{bmatrix}0 & 0 \\ 0 & L^{-1}\end{bmatrix}$

$E(\tilde\beta) = T\beta$ $\quad Var(\hat\beta) = \sigma^2 (X^T X)$ $\quad Var(\tilde\beta) = \sigma^2 T (X^T X) T^T$

④. 证明：单因子方差分析模型可看作一种多元线性回归 model 具体证明先不写了

⑤、X_s 表示标准化后的特征 $X_s = (x_1, x_2 \cdots x_p)$ $\quad X_s$ 相关系数阵为 $X_s^T X_s$

逆矩阵为 $(X_s^T X_s)^{-1} = C = \{c_{ij}\}$ 我们将 x_j 作因变量，而将剩余特征作自变量，建立多元线性回归 model：$x_j = \alpha_1^j x_1 + \cdots \alpha_{j-1}^j x_{j-1} + \alpha_{j+1}^j x_{j+1} + \cdots + \alpha_p^j x_p$

该模型复决定系数 R_j^2 $\quad$ 证明：$c_{jj} = \frac{1}{1 - R_j^2}$

$SS_R^j = \sum_{i=1}^n (\hat x_{ji} - \bar x_j)^2$ 数据被标准化，则 $1^T x_i = 0$ $\quad x_i^T x_i = 0$ $\quad (i = 1, 2 \cdots p)$

$= \sum_{i=1}^n \hat x_{ji}^2 = \hat x_j^T \cdot \hat x_j$ $\quad$ 其中 $\hat x_j = X_{sj} \hat\beta$

$= X_{sj} (X_{sj}^T X_{sj})^{-1} X_{sj}^T x_j$

故 $SS_R^j = \hat x_j^T \hat x_j = \cdots$ $\quad$ ↳ 因为这里 x_j 为因变量

$= x_j^T \cdot X_{sj} (X_{sj}^T X_{sj})^{-1} X_{sj}^T x_j$ $\quad X_{sj} = (x_1\ x_2 \cdots x_{j-1}, x_{j+1} \cdots x_p)$

$$SS_T^j = \sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2 = \sum_{k=1}^{n} x_{jk}^2 = 1$$

故 $$R_j^2 = \frac{SS_R^j}{SS_T^j} = x_j^T x_{sj} (x_{sj}^T x_{sj})^{-1} x_{sj}^T x_j$$

$$X_S = [\, x_j \;\; x_{sj} \,] A_j \qquad A_j = \begin{bmatrix} & 1 & & \\ 1 & & \ddots \; 0 & \\ & & & \ddots \end{bmatrix} \qquad A_j^{-1} = A_j^T$$

把I_p的第j列换到第1列
第1至j-1列都往后推一列

$$X_S A_j = [\, x_j \;\; x_{sj} \,]$$

$$X_S = [\, x_j \;\; x_{sj} \,] A_j^T$$

$$C = (X_S^T X_S)^{-1} = \left[A_j \begin{bmatrix} x_j^T \\ x_{sj}^T \end{bmatrix} [\, x_j \;\; x_{sj} \,] A_j^T \right]^{-1} = A_j \begin{bmatrix} 1 & x_j^T x_{sj} \\ x_{sj}^T x_j & x_{sj}^T x_{sj} \end{bmatrix}^{-1} A_j^T$$

A_j左乘意味着…　A_j^T右乘意味着…

$A_j \begin{bmatrix} & \\ & \end{bmatrix}^{-1} A_j^T \rightarrow$ 代表的是取中间矩阵第一行第1个元素！
根据分块矩阵的逆
原式得证

$$c_{jj} = \frac{1}{1 - x_j^T x_{sj} (x_{sj}^T x_{sj})^{-1} x_{sj}^T x_j}$$

但我感觉这个方法 too stupid.

(6) $\hat{\beta} = (X^T X)^{-1} X^T y \quad MSE(\hat{\beta}) = E(\beta^T - \beta)(\beta^T - \beta) = ?$

$$MSE(\hat{\beta}) = E\left((\hat{\beta} - \beta)^T (\hat{\beta} - \beta)\right) = E\left((\hat{\beta} - E(\hat{\beta}))^T (\hat{\beta} - E(\hat{\beta}))\right)$$

$\hookrightarrow E(\hat{\beta}) = \beta$

$$= Tr(cov(\hat{\beta})) = Tr(\sigma^2 (X^T X)^{-1}) = \sigma^2 Tr((X^T X)^{-1})$$

$\boxed{\text{统计方法实验内容review}}$

$\bar{y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{m}) \qquad \frac{\bar{y}_{i\cdot} - \mu_i}{\sqrt{\sigma^2 \frac{1}{m}}} = \frac{\bar{y}_{i\cdot} - \mu_i}{\sigma \sqrt{\frac{1}{m}}} \sim N(0,1)$

$\frac{SSE}{\sigma^2} \sim \chi^2(n-a) \qquad$ 故 $\frac{\bar{y}_{i\cdot} - \mu_i}{\sigma \sqrt{\frac{1}{m}}} \cdot \frac{1}{\sqrt{\frac{SSE}{\sigma^2 (n-a)}}} = \frac{\sqrt{m}(\bar{y}_{i\cdot} - \mu_i)}{\hat{\sigma}} \sim t(n-a)$

这也是Tukey方法的核心

dqw上课主要讲统计学习方法(李航)第一篇：监督学习 2023.12.

所以我复习主要围绕这个展开，

用到的Note结构是这样的：先放了一堆例题，再自己亲自去学那本书对应部分并留Note，我结合其中的一些过于精简的部分加一些内容

数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

由此来确定一个朴素贝叶斯分类器，并确定 $x=(2,S)^T$ 的类标记 y

取值集合 $A_1=\{1,2,3\}$ $A_2=\{S,M,L\}$ $Y\in C=\{1,-1\}$ 为类标记

我的理解，要求 $\cancel{P(X=x|Y}$ $P(X=x,Y=c_k)\to$ 先求 $P(Y=c_k)$ 再求 $P(X=x|Y=c_k)$

然后再乘一下，找 $P(X=x,Y=c_k)$ max 的 c_k 为分类

$\cancel{=P(X^{(1)}=\dots,X^{(2)}=\dots)}$

$\cancel{=\prod_{i=1}^{n}P(X^{(i)}=x_i)}$

$=P(X^{(1)}=\dots,X^{(n)}=\dots|Y=c_k)$

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

(先不做拉普拉斯平滑)

$P(Y=-1)=\frac{2}{5}$ $P(Y=1)=\frac{3}{5}$

$P(X^{(1)}=1|Y=-1)=\frac{1}{2}$ $P(X^{(1)}=1|Y=1)=\frac{2}{9}$

$P(X^{(1)}=2|Y=-1)=\frac{1}{3}$ $P(X^{(1)}=2|Y=1)=\frac{1}{3}$

$P(X^{(1)}=3|Y=-1)=\frac{1}{6}$ $P(X^{(1)}=3|Y=1)=\frac{4}{9}$

$P(X^{(2)}=S|Y=-1)=\frac{1}{2}$ $P(X^{(2)}=S|Y=1)=\frac{1}{9}$

$P(X^{(2)}=M|Y=-1)=\frac{1}{3}$ $P(X^{(2)}=M|Y=1)=\frac{4}{9}$

$P(X^{(2)}=L|Y=-1)=\frac{1}{6}$ $P(X^{(2)}=L|Y=1)=\frac{4}{9}$

然后求 $P(X=x,Y=1)=P(X=x|Y=1)\cdot P(Y=1)$

$=P(Y=1)\cdot P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)$

$=\frac{1}{45}$

$P(X=x,Y=-1)=P(X=x|Y=-1)\cdot P(Y=-1)=P(Y=-1)\cdot P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$

后面那个更大，故 $Y=-1$ 这便是分类结果 $=\frac{1}{15}$

No.
Date.
再用拉普拉斯平滑 $\lambda=1$ 二分类问题 $K=2$.

$$P(Y=1)=\frac{9+1}{15+2\times1}=\frac{9}{17}\quad P(Y=-1)=\frac{6+1}{15+2\times1}=\frac{7}{17}$$

$$P(X^{(1)}=1\mid Y=1)=\frac{2+1}{9+3}$$

→ $X^{(1)}=1$ 且 $Y=1$ 的数量 $\boxed{Y是二分类}$

→ 这时拉普拉斯平滑要加这个

主语是 $X^{(1)}$!

→ $X^{(1)}$ 是3分类; 分子一直都+的是 λ!

↓ $Y=1$ 的数量

终于搞明白这个了, 这之后的过程类似.

我先来大致写一下SVM的基本理论，再去看看"核技巧"。

No. Date. 2023 12 23

训练集 $\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$ $y_i=\begin{cases}+1 & 正例\\ -1 & 负\end{cases}$

SVM要解决：通过间隔最大化(或等价形式)求得分离超平面 $w\cdot x+b=0$ 点积

相应的分类决策函数 $f(x)=\mathrm{sign}(w\cdot x+b)$

超平面关于(x_i,y_i)的函数间隔 $\hat{\gamma}_i=y_i(w\cdot x_i+b)$

关于整个训练集的函数间隔 $\hat{\gamma}=\min\limits_{i=1\cdots N}\hat{\gamma}_i$

几何间隔 → 分别对应乘 $\frac{1}{||w||}$ 就行 → 比如 $\gamma_i=\frac{\hat{\gamma}_i}{||w||}$ 一般取 $||w||=1$ 约束 (不必要)

最大间隔分离超平面 $\max\limits_{w,b}\gamma$

$s.t.\ y_i\left(\frac{w\cdot x_i}{||w||}+\frac{b}{||w||}\right)\geq\gamma \quad i=1,2\cdots n$

线性可分支持向量机学习的最优化问题：

$\min\limits_{w,b}\frac{1}{2}||w||^2$

$s.t.\ y_i(wx_i+b)\geq 1 \quad i=1,2\cdots n$ → 得最优解 w^*, b^*

最优化问题复习 $\min\limits_{w} f(w)$

$s.t.\ g_i(w)\leq 0 \quad i=1,2\cdots k$

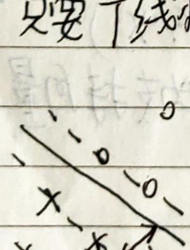
$h_i(w)=0 \quad i=1,2\cdots L$

拉格朗日 $L(w,\vec{\alpha},\vec{\beta})=f(w)+\sum\limits_{i=1}^{k}g_i(w)\alpha_i+\sum\limits_{i=1}^{L}h_i(w)\cdot\beta_i$

其中 $\alpha_i\geq 0$ $(i=1\cdots k)$

最大间隔分离超平面唯一 只要了线性就行

线性可分时 训练集的样本点 与分离超平面距离最近的 称为支持向量 (即 $y_i(wx_i+b)-1=0$ 成立)



例题 正例点 $(3,3)$, $(4,3)$ 负例点 $(1,1)$ 求分离超平面？

$\min\frac{1}{2}(w_1^2+w_2^2)$

$s.t.\ 3w_1+3w_2+b\geq 1$

$4w_1+3w_2+b\geq 1$

$-w_1-w_2-b\geq 1$

得 $w_1=\frac{1}{2}=w_2$ $b=-2$ → 故所求为 $\frac{1}{2}x^{(1)}+\frac{1}{2}x^{(2)}-2=0$

(背面有对偶算法)

P1

No.

Date.

$\to$ 拉~乘子

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

$$\left[\begin{array}{l}\min\frac{1}{2}\|w\|^2\\ s.t.\ y_i(wx_i+b)\ge 1\\ -y_i(wx_i+b)\le -1\end{array}\right.$$

对偶问题 $\to$ 极大极小问题：

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

1、$\min\limits_{w,b} L(w,b,\alpha)=?$

$$\nabla_w L(w,b,\alpha)=w-\sum_{i=1}^n\alpha_i y_i x_i$$

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

$$\Rightarrow\left[\begin{array}{l}w=\sum_{i=1}^n\alpha_i y_i x_i\\ \sum_{i=1}^n\alpha_i y_i=0\end{array}\right.$$

再代入 $L(w,b,\alpha)$（消去 w,b）得 $\min\limits_{w,b} L(w,b,\alpha)$

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

两者点乘

2、$\max\limits_{\alpha}(\cdots)$

$$\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\alpha_i\alpha_j y_i y_j(x_i x_j)+\sum_{i=1}^n\alpha_i$$

再换为min即可

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

这个可得 α^*（对偶问题最优解）

原始问题最优解 $w^*=\sum_{i=1}^n\alpha_i^* y_i x_i$

选个 $\alpha_j^*>0$ $\quad b^*=y_j-\sum_{i=1}^n\alpha_i^* y_i(x_i\cdot x_j)$

这个j是任意选，只要 $\alpha_j^*>0$ 即可

支持向量：训练集中对应于 $\alpha_i^*>0$ 的 (x_i,y_i) 即为支持向量

例、正例点 $x_1=(3,3)^T$ $x_2=(4,3)^T$ $x_3=(1,1)^T$

对偶形式：$\min\limits_{\alpha}\frac{1}{2}\sum_{i=1}^3\sum_{j=1}^3\alpha_i\alpha_j y_i y_j(x_i x_j)-\sum_{i=1}^3\alpha_i$

$$=\min_{\alpha}\ 9\alpha_1^2+\frac{25}{2}\alpha_2^2+\alpha_3^2+21\alpha_1\alpha_2-6\alpha_1\alpha_3-7\alpha_2\alpha_3-\alpha_1-\alpha_2-\alpha_3$$

$$s.t.\ \alpha_1+\alpha_2-\alpha_3=0$$

$$\alpha_i\ge 0\quad i=1,2,3.$$

$\alpha_3 = \alpha_1 + \alpha_2$ 代入式中 $S(\alpha_1,\alpha_2) = 4\alpha_1^2 + \frac{13}{2}\alpha_2^2 + 10\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2$,

对 α_1,α_2 求偏导,令其为0 得 $\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix} = \begin{bmatrix}\frac{3}{2}\\ -1\end{bmatrix}$ 但这不行

↳ 边界上取极小值

$\alpha_1 = 0$ $S_{min} = -\frac{2}{13}$, $\alpha_2 = 0$ $S_{min} = -\frac{1}{4}$

故 $\alpha_1^* = \frac{1}{4}$ $\alpha_2^* = 0$ $\alpha_3^* = \frac{1}{4}$ $\xrightarrow{代入}$ 故 x_1 和 x_3 为支持向量(大于0)

$$W^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i = \frac{1}{4}\cdot 1 \begin{bmatrix}3\\3\end{bmatrix} + 0\cdot 1\begin{bmatrix}4\\3\end{bmatrix} + \frac{1}{4}(-1)\begin{bmatrix}1\\1\end{bmatrix}$$

$$= \begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\end{bmatrix} \quad W_1^* = \frac{1}{2} = W_2^*$$

比如选 $\alpha_1^* = \frac{1}{4}$, ~~求~~ $b^* = 1 - \sum_{i=1}^{3} \alpha_i^* y_i (x_i \cdot x_1)$

$$= 1 - \left(\frac{1}{4}\cdot 18 + 0 + \frac{1}{4}\cdot(-1)\cdot 6\right) = -2$$

分离超平面 $\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$

分离决策函数 $f(x) = sign\left(\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2\right)$

去年期末 正类 ~~(1,1)~~ $(1,1)$ $(2,2)$, $(2,0)$ 负类 $(0,0)$ $(1,0)$ $(0,1)$

用对偶方法求~~换~~解?

对偶问题 $\min\limits_{\alpha} \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} \alpha_i\alpha_j \cdot y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{6}\alpha_i$ 它别忘了!

s.t. $\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 - \alpha_6 = 0$

$\alpha_i \geq 0, i = 1, 2, \ldots 6$

即 $\min\limits_{\alpha} \alpha_1^2 + 4\alpha_2^2 + 2\alpha_3^2 + \frac{1}{2}\alpha_5^2 + \frac{1}{2}\alpha_6^2 + 4\alpha_1\alpha_2 + 2\alpha_1\alpha_3$

$+ 4\alpha_2\alpha_3 - \alpha_1\alpha_5 - \alpha_1\alpha_6 - 2\alpha_2\alpha_5 - 2\alpha_2\alpha_6 - 2\alpha_3\alpha_5 - \sum_{i=1}^{6}\alpha_i$

求导 对 α_1: $2\alpha_1 + 4\alpha_2 + 2\alpha_3 - \alpha_6 ~~-\alpha~~ - 1 = 0$

对 α_2: $8\alpha_2 + 4\alpha_1 + 4\alpha_3 - 2\alpha_5 - 2\alpha_6 - 1 = 0$

... 不对 反正就这样么算,运算量大!

因为 w, b 在给定数据集时一定一模一样,你笨办法先搞个答案

列一堆试了,与答案就行)

P3

No.
Date.

$+1: (1,1)\ (2,2)\ (2,0) \quad -1: (0,0)\ (1,0)\ (0,1)$

朱妯法求故. $\min\limits_{w} \frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2+w_2^2)$

s.t. $w_1+w_2+b \cancel{\geq 0} \geq 1 \qquad b \leq -1$

$2w_1+2w_2+b \geq 1 \qquad b+w_1 \leq -1$

$2w_1+b \geq 1 \qquad w_2+b \leq -1$

这个怎样搞? 有 $\begin{cases} w_1+w_2+b \geq 1 \\ 2w_1+b \geq 1 \end{cases}$ 可想到 $w_1 = w_2$, $\min w_1^2$

$\begin{cases} 2w_1+b \geq 1 \\ 4w_1+b \geq 1 \\ b \leq -1 \\ b+w_1 \leq -1 \end{cases} \Rightarrow \begin{matrix} w_1 \geq 2 \\ b \leq -1 \\ \cdots \end{matrix} \quad \begin{matrix} \min w_1^2 \quad w_1 = 2 = w_2 \\ b \text{ 此时要为 } -3 \\ \text{故} \cdots \end{matrix}$

P4.

看不见

状态 → 状态

隐马尔模型练习Note

No.

Date 24.1.2

$A = \begin{bmatrix} \\ \\ \end{bmatrix}_{n\times n}$ a_{ij} 表示从状态i到状态j的概率

$B = \begin{bmatrix} \\ \\ \end{bmatrix}_{n\times m}$ b_{ij} 表示从状态i到观测j的P

$\pi(\underbrace{\quad,\quad}_{共N个元素})$ 表示初始各状态的P

状态 → 观测

看得见

两个假设 ① 状态只与t−1时有关 ② 观测只与当前状态有关

(*) 概率calc问题：给定A, B, π，计算 $P(O|\lambda)$，$\lambda=(A,B,\pi)$　O是观测 I是状态

~~前向算法~~

$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{5} & \frac{3}{10} \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$ $B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$ $\pi = (0.2\ 0.4\ 0.4)^T$

$O = (红, 白, 红)$

① 前向算法 $\alpha_t(i) = P(o_1, o_2 \cdots o_t, i_t = q_i \mid \lambda)$

↳ t时刻状态为i的概率

前向

i状态直接转到观测

初始 $\alpha_1(i) = \pi_i \cdot b_i(o_1)$　$i=1,2\cdots N$

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

j状态先到i状态，再从i状态到观测

$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

② 后向算法 $\beta_t(i) = P(o_{t+1}, \cdots o_T \mid i_t = q_i, \lambda)$

初始：$\beta_T(i) = 1$　$i=1,2\cdots N$

转移：$\beta_t(i) = \left[\sum_{j=1}^{N} \boxed{a_{ij}} \cdot \boxed{b_j(o_{t+1})} * \beta_{t+1}(j)\right]$

注意是 a_{ij}！　生成观测　概率

$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i)$　从初始到观测再依概率组合

前向后向写成统一形式

$P(O|\lambda) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j)$　$t=1,2\cdots T-1$

No.

Date. 前向概率练习例题

$$A=\begin{bmatrix} \frac{1}{2} & 0.2 & 0.3 \\ 0.3 & \frac{1}{2} & 0.2 \\ 0.2 & 0.3 & \frac{1}{2} \end{bmatrix} \quad B=\begin{array}{c} \begin{array}{cc} 红 & 白 \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix} \end{array}$$

$\pi=(0.2\ \ 0.4\ \ 0.4)^T$

$O=$ ~~(红,白,红)~~ (红,白)

求 $P(O|\lambda)$

初始时刻 $\alpha_1(1)=\pi_1\cdot b_1(o_1)=0.2\times0.5=0.1$

$\alpha_1(2)=\pi_2 b_2(o_1)=0.4\times0.4=0.16$

$\alpha_1(3)=\pi_3\cdot b_3(o_1)=0.4\times0.7=0.28$

转移 $\alpha_2(1)=\left(\sum_{i=1}^{3}\alpha_1(i)a_{i1}\right)\cdot b_1(o_2)=\cdots=0.077$

$\alpha_2(2)=\left(\sum_{i=1}^{3}\alpha_1(i)a_{i2}\right)\cdot b_2(o_2)=\cdots 0.1104$

$\alpha_2(3)=\left(\sum_{i=1}^{3}\alpha_1(i)a_{i3}\right)\cdot b_3(o_2)=\cdots=0.0606$

$P(O|\lambda)=\sum_{i=1}^{3}\alpha_2(i)=0.248$

若用后向概率算呢？

初始 $\beta_2(1)=\beta_2(2)=\beta_3(3)=1$

转移 $\beta_1(1)=\sum_{j=1}^{3}\left(a_{1j}\cdot b_j(o_2)\cdot\beta_2(j)\right)=0.46$

$\beta_1(2)=\sum_{j=1}^{3}\left(a_{2j}\cdot b_j(o_2)\cdot\beta_2(j)\right)=0.51$

$\beta_1(3)=\sum_{j=1}^{3}\left(a_{3j}\cdot b_j(o_2)\cdot\beta_2(j)\right)=0.43$

$P(O|\lambda)=\sum_{i=1}^{3}\beta_1(i)\cdot\pi_i\cdot b_i(o_1)=0.248$

当然最关键还是那道作业题！

$$A=\begin{bmatrix} \frac{1}{2} & 0.1 & 0.4 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} \quad B=\begin{array}{c} \begin{array}{cc} 红 & 白 \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix} \end{array} \quad \pi=(0.2,0.3,0.5)^T$$

设 $T=8$ $O=$(红,白,红,红,白,红,白,白) 用前向后向概率

算 $P(i_4=q_3|O,\lambda)$

在4时刻为3状态

前向算法计算前向概率

$\alpha_1(1)=0.2\times0.5=0.1$

$\alpha_1(2)=0.3\times0.4=0.12$

$\alpha_1(3)=0.5\times\cancel{0.2}\,0.7=0.35$

$$\alpha_2(1)=\cancel{\left(\sum_{j=1}^{3}\alpha_1(j)\,a_{j1}\right)}\left(\sum_{j=1}^{3}\alpha_1(j)\cdot a_{j1}\right)\cdot b_1(O_2)$$

$$=(0.1\times0.5+0.12\times0.3+0.35\times0.2)\times0.5=0.078$$

$$\cancel{\alpha_3(1)}\ \alpha_2(2)=\left(\sum_{j=1}^{3}\alpha_1(j)\cdot a_{j2}\right)\cdot b_2(O_2)=0.084$$

$$\alpha_2(3)=\left(\sum_{j=1}^{3}\alpha_1(j)\cdot a_{j3}\right)\cdot b_3(O_2)=0.0822$$

$$\alpha_3(1)=\left(\sum_{j=1}^{3}\alpha_2(j)\cdot a_{j1}\right)\cdot b_1(O_3)=\cancel{0.08064}\ 0.04032$$

$$\alpha_3(2)=\left(\sum_{j=1}^{3}\alpha_2(j)\cdot a_{j2}\right)\cdot b_2(O_3)=0.026496$$

$$\alpha_3(3)=\left(\sum_{j=1}^{3}\alpha_2(j)\cdot a_{j3}\right)\cdot b_3(O_3)=0.068124$$

$\dots\ \alpha_4(1)=0.0208668\quad \alpha_4(2)=0.01236192\quad \alpha_4(3)=0.0436112$

后向算法计算后向概率：$\beta_8(1)=\beta_8(2)=\beta_8(3)=1$

$$\beta_7(1)=\sum_{j=1}^{3}\left(a_{1j}\cdot\cancel{b_j(O_7)}\,\beta_8(j)\cdot b_j(O_8)\right)=0.43$$

$$\beta_7(2)=\sum_{j=1}^{3}\left(a_{2j}\cdot\beta_8(j)\cdot b_j(O_8)\right)=0.51$$

$$\beta_7(3)=\sum_{j=1}^{3}\left(a_{3j}\cdot\beta_8(j)\cdot b_j(O_8)\right)=0.4$$

$$\beta_6(1)=\left(\sum_{j=1}^{3}a_{1j}\cdot\beta_7(j)\cdot b_j(O_7)\right)=0.1861$$

$$\beta_6(2)=\left(\sum_{j=1}^{3}a_{2j}\,\beta_7(j)\,b_j(O_7)\right)=0.2415$$

$$\beta_6(3)=\left(\sum_{j=1}^{3}a_{3j}\,\beta_7(j)\,b_j(O_7)\right)=0.1762$$

$\dots\ \beta_5(1)=0.105521\quad \beta_5(2)=0.100883\quad \beta_5(3)=0.111934$

$\dots\ \beta_4(1)=0.04586531\quad \beta_4(2)=0.05276409\quad \beta_4(3)=0.04277618$

$$P(i_4=q_3\mid O,\lambda)=\frac{\alpha_4(3)\cdot\beta_4(3)}{\alpha_4(1)\beta_4(1)+\alpha_4(2)\beta_4(2)+\alpha_4(3)\beta_4(3)}\approx0.5368638876$$

第4步为状态3

EM算法

中间的相关定理全Skip了，关键是会用它！ θ要估的参数

Y：观测到的随机变量的数据 Z：隐变量的数据 (Y,Z)完全数据

~~$P(Y|\theta)$为要估计的~~ $P(Y|\theta)$ Y在给定θ下的PDF

Y,Z联合 $P(Y,Z|\theta)$ log：$\log P(Y,Z|\theta)$

θ初始为$\theta^{(0)}$ 随机取 $\rightarrow Q(\theta,\theta^{(i)})=E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}]$

$$\downarrow$$

$$\theta^{(i+1)}=\arg\max_{\theta} Q(\theta,\theta^{(i)})$$

E步

期望

关于隐变量！

但是$E(\ )$里是 $\log(Y,Z|\theta)$！

$$Q(\theta,\theta^{(i)})=E_Z(\quad)$$

$$=\sum_Z \log P(Y,Z|\theta^{(i)})P(Z|Y,\theta^{(i)})$$

在给定$Y,\theta^{(i)}$前提下，对Z的条件概率的期望！

以GMM为例，来理解一下EM

$\hookrightarrow$ nl说：虽然课上不细讲，但不妨碍她考，我还要自己证明遍

GMM核心：假定来自一个类的样本均服从同一正态分布

第k个类 $N_p(\mu_k,\Sigma_k)$ $k=1,2\cdots K$ 超参

$$f(x_i)=(2\pi)^{-\frac{1}{2}p}|\Sigma_k|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\}$$

x_i为p维向量 $i=1,2\cdots n$

<u>如果我们知道</u> 第i个样本来自于第k个高斯分布总体

假设

$$\delta_{ik}=\begin{cases}1 & x_i\text{来自}N_k\\ 0 & \text{其他}\end{cases}$$

$\delta_i=[\delta_{i1}\cdots\delta_{iK}]^T$ 独立同分布，服从多维分布$M(1,\pi_1,\cdots\pi_K)$

$$\pi_k=P(\delta_{ik}=1)\qquad \sum_{i=1}^{K}\pi_i=1$$

δ_i的PDF

$$f(\delta_i)=\prod_{k=1}^{K}(\pi_k)^{\delta_{ik}}$$

在给定 δ_i 后，x_i 的 PDF

$$f(x_i\mid\delta_i)=\prod_{k=1}^{K}\left((2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\}\right)^{\delta_{ik}}$$

完全数据 (x_i,δ_i) 的联合 PDF：

δ_i 隐变量

$$L(\theta)=\prod_{i=1}^{n}f(x_i,\delta_i)=\prod_{i=1}^{n}f(\delta_i)f(x_i\mid\delta_i)$$

(x_i,δ_i) 的 $\mathcal{L}$

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

这个也是 $\theta=(\pi_1\cdots\pi_K,\mu_1,\cdots\mu_K,\Sigma_1\cdots\Sigma_K)$ 的似然函数.

$$f(x_i)=\sum_{k=1}^{K}\pi_k f(x_i;\mu_k,\Sigma_k)$$

$$\ln L(\theta)=\ln\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k(2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\}\right)^{\delta_{ik}}\right)$$

只量和 θ 有关.

$$\underset{\text{近似}}{\propto}-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\cdot\left((x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)+\ln|\Sigma_k|\right)$$

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

E 步：目标求 $E_{\delta_i}(Q_0(\theta)\mid x_i,\theta^{(t)})$

δ_i 一大堆，但拆开来看就一个个数. 设 δ_{ik} 期望为 π_{ik}^*

则 ~~$Q_0(\theta)$~~ 所求 $Q(\theta)=-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^*(\quad)+\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^*\ln(\pi_k)$

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

个体的类别期望?

其中 $\phi(x_i;\mu_k,\Sigma_k)$ 为第 k 个…

M步：让 $Q(\theta)$ 关于 θ 最大化.

$\rightarrow Q_1(\theta)$ 与 μ_k, Σ_k 有关

$$Q(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^{*}\left((x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)+\ln|\Sigma_k|\right)$$

要分开

$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^{*}\ln(\pi_k) \rightarrow Q_2(\theta) \rightarrow \text{与 } \pi_k \text{ 有关.}$$

$$\left[\begin{aligned}&\frac{\partial Q_1(\theta)}{\partial \mu_k} \propto \sum_{i=1}^{n}\pi_{ik}^{*}\Sigma_k^{-1}(x_i-\mu_k)\\ &\frac{\partial Q_1(\theta)}{\partial \Sigma_k} \propto \sum_{i=1}^{n}\pi_{ik}^{*}\left(-\Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}+\Sigma_k^{-1}\right)\end{aligned}\right.$$

由此 $\hat{\mu}_k = \dfrac{\sum_{i=1}^{n}\pi_{ik}^{*}\cdot x_i}{\sum_{i=1}^{n}\pi_{ik}^{*}}$ $\quad \hat{\Sigma}_k = \dfrac{\sum_{i=1}^{n}\pi_{ik}^{*}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\pi_{ik}^{*}}$

$Q_2(\theta)$ 求解用到拉格朗日乘子法 (因为 $\sum_{k=1}^{K}\pi_k = 1$) 可得 $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\pi_{ik}^{*}$

$$\phi(x_i;\mu_k,\Sigma_k) = (2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}$$

2024.1.2.

之前我有一页Note，但当天打折到了，就是用EM推证GMM

虽然自己时间确实有限，但GMM实在太重要了，我还是再写一下吧！

什么是EM algo？观测到的数据；未观测到的~

$P(Y|\theta)=\sum_z P(z|\theta)P(Y|z,\theta)$ Y观测到的，Z未观测到的

$\hat{\theta}=\arg\max_\theta \log P(Y|\theta)$ $\theta=(\pi,p,q)$ 参数

输入了

联合分布 $P(Y,z|\theta)$ 条件分布 $P(Y|$ $P(z|Y,\theta)$

E步 求给定Y和当前参数估计 $\theta^{(i)}$ 下隐变量Z的条件概率分布

$Q(\theta,\theta^{(i)})=E_z(\log P(Y,z|\theta)|Y,\theta^{(i)})$

对隐变量 完全数据

$=\boxed{\sum_z \log P(Y,z|\theta)P(z|Y,\theta^{(i)})}$

M步 求使 $Q(\theta,\theta^{(i)})$ max的 θ

$\theta^{(i+1)}=\arg\max Q(\theta,\theta^{(i)})$

重复E、M，直至收敛 后面是GMM证明过程

高斯混合模型：$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$ $\begin{cases}\alpha_k\geq 0\\ \sum_{k=1}^{K}\alpha_k=1\end{cases}$

$\phi(y|\theta_k)$ 是高斯分布PDF：$\theta_k=(\mu_k,\delta_k^2)$

$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\delta_k}\exp\{-\frac{1}{2\delta_k^2}(y-\mu_k)^2\}$ 第k个模型

GMM：$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$ $\theta=(\alpha_1,\alpha_2\cdots\alpha_K;\theta_1\cdots\theta_K)$

观测数据 $y_1\cdots y_N$ 是怎样产生的？

依概率 α_k 选择第k个 $\phi(y|\theta_k)$，然后产生 y_j

y_j 已知，但 y_j 来自第k个模型却未知 → 隐变量，用 γ_{jk} 表示

$\gamma_{jk}=\begin{cases}1 & y_j\text{来自第k个分模型}\\ 0 & \text{其他}\end{cases}$

完全数据 $(y_1,\gamma_{11}\cdots\gamma_{1K},y_2,\gamma_{21}\cdots\gamma_{2K}\cdots)$

超重点 GMM的证明

完全数据 $(y_1, \gamma_{11}\ \gamma_{12} \cdots \gamma_{1K}, y_2, \gamma_{21} \cdots \gamma_{2K}, \cdots)$

$$\gamma_{jk} = \begin{cases} 1 & \text{表示 } y_j \text{ 来自模型 } k \text{ 的} \\ 0 & \text{其他} \end{cases}$$

完全数据的似然函数

$$P(y,\gamma|\theta) = \prod_{j=1}^{N} P(y_j, \gamma_{j1} \cdots \gamma_{jK} | \theta)$$

$$= \prod_{j=1}^{N} \prod_{k=1}^{K} [\alpha_k \phi(y_j|\theta_k)]^{\gamma_{jk}}$$

来自 也要遵循概率

$$= \cancel{\prod_{k=1}^{K} \alpha_k^{\gamma_{jk}}} \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} [\phi(y_j|\theta_k)]^{\gamma_{jk}}$$

$$n_k = \sum_{j=1}^{N} \gamma_{jk} \qquad \sum_{k=1}^{K} n_k = N$$

$$= \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} \left[\frac{1}{\sqrt{2\pi}\delta_k} \exp\left(-\frac{(y_j-\mu_k)^2}{2\delta_k^2}\right)\right]^{\gamma_{jk}}$$

$$\log P(y,\gamma|\theta) = \sum_{k=1}^{K} \left\{ n_k \log\alpha_k + \sum_{j=1}^{N} \gamma_{jk} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\delta_k - \frac{1}{2\delta_k^2}(y_j-\mu_k)^2\right]\right\}$$

$$E: Q(\theta, \theta^{(i)}) = E_\gamma(\log(y,\gamma|\theta) | y, \theta^{(i)})$$ → 对 γ 的期望

$$= \sum_{k=1}^{K} \left[\sum_{j=1}^{N} E(\gamma_{jk})\right] \log\alpha_k + \sum_{k=1}^{K} \sum_{j=1}^{N} E(\gamma_{jk}) \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\delta_k - \frac{1}{2\delta_k^2}(y_j-\mu_k)^2\right]$$

$$E(\gamma_{jk}|y,\theta^{(i)}) = ? = \hat{\gamma}_{jk} = P(\gamma_{jk}=1|y,\theta^{(i)}) \cdot 1 + P(\gamma_{jk}=0|y,\theta^{(i)}) \cdot 0 = P(\gamma_{jk}=1|y,\theta^{(i)})$$

$$= \frac{P(\gamma_{jk}=1, y_j|\theta^{(i)})}{\sum_{k=1}^{K} P(\gamma_{jk}=1, y_j|\theta^{(i)})} \xrightarrow{\text{条件概率}} = \frac{P(y_j|\gamma_{jk}=1,\theta^{(i)}) P(\gamma_{jk}=1|\theta^{(i)})}{\sum_{k=1}^{K} P(y_j|\gamma_{jk}=1,\theta^{(i)}) \cdot P(\gamma_{jk}=1|\theta^{(i)})}$$

$$= \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j|\theta_k)}$$

再将 $\hat{\gamma}_{jk}$ 代入，可得

$Q(\theta,\theta^{(i)}) = \cdots$

EM, M步看 Note.P80

华东师范大学期末试卷（A）

2021-2022 学年第一学期

课程名称：统计方法与机器学习

学生姓名：________　　学　　号：________

专　　业：________　　年级/班级：________

课程性质：专业必修课

一	二	三	四	五	六	总分	阅卷人签名

一、（本题共 20 分）

表 1 是一个不完整的双因素方差分析表。

表 1 不完整的双因素方差分析表

来源	自由度	平方和	均方	F统计量	p 值
因素 A	/	/	0.0833	0.05	0.952
因素 B	/	96.333	96.333	57.80	<0.001
交互效应 AB	2	12.167	6.0833	3.65	/
误差	6	10	/		
汇总	11	118.667			

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

$\bar{y}_{i.}$

请根据表 1 回答以下问题：

1.（2 分） 因素 A 的平方和SS_A是 0.167 。

2.（2 分） 因素 A 的自由度为 2 。

3.（2 分） 在实验中，因素 B 的水平数为 2 。

4.（2 分） 均方误差为 1.667 。

这什么意思啊？

5.（2 分） 在这个实验中，每种组合的重复次数为 2 。

6.（5 分） 如何计算交互效应的p值？并给定显著性水平$\alpha = 0.05$时，简述如何判断交互效应的显著性。

7.（5 分） 证明：在双因素方差分析中，$SS_T = SS_A + SS_B + SS_{AB} + SS_E$ 。

1 层次聚类适用范围最广
但边界模糊时效果较差
2 kmeans 收敛慢，且依赖初始中心点的选取，适合组内距离小的 data.

二、（本题共 20 分）

3 GMM.
更适合各类别尺寸不同，聚类间有相关关系的情况

现有一个数据集，其中包含 400 条观测，每条观测有 1 个因变量y以及 20 个中心化后的特征$x_1, x_2, \cdots, x_{20}$。前 5 行数据如图 2 所示。

4 DBSCAN
基于密度.
不太适合稀疏数据集。而且超参数，如minpts影响还挺大.

	y	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
0	-1.24	0.31	-0.95	-0.99	-0.58	-0.47	0.70	-0.88	1.05	0.03	...	-1.42	-0.85	-0.37	-0.52	-0.19	-1.75	0.78	-0.78	0.23	0.05
1	-0.72	-0.04	-2.58	1.32	-0.75	-0.92	1.43	-1.85	-0.83	1.36	...	0.31	-0.01	-0.49	-1.20	0.08	-0.83	-0.98	2.89	-0.79	-0.82
2	6.40	-0.89	0.91	-0.07	0.14	1.31	0.60	0.34	-0.96	1.67	...	1.58	0.44	1.80	-0.04	1.65	-0.06	-1.01	-0.87	-0.41	-1.03
3	1.10	1.81	0.20	-0.70	-1.03	0.72	-0.83	1.56	-0.03	-0.28	...	0.61	0.01	-1.39	-0.78	-1.20	-2.09	-0.70	-0.73	-1.96	0.48
4	2.33	0.15	-0.27	-0.82	-0.21	0.42	-0.15	-0.04	0.80	2.55	...	-0.44	-0.47	-1.08	-2.31	0.87	-0.62	1.10	1.02	1.26	0.58

图 2 前 5 条数据的示意图

取显著性水平 $\alpha = 0.05$，现回答以下问题：

1.（5 分）同学 A 想构建利用X_1来预测y，从而构建了一个一元线性回归模型。请根据图 3 中 Python 运行的结果，写出一元线性回归模型，并从一个角度阐述该模型是否显著。 P太小 到拒绝域，故显著.

评判拟合程度.

$\hat{y} = 0.6954 + 1.6034x$

$P < \alpha$ 显著

Dep. Variable:	y	R-squared:	0.127
Model:	OLS	Adj. R-squared:	0.125
Method:	Least Squares	F-statistic:	58.04
Date:	Fri, 02 Dec 2022	Prob (F-statistic):	1.90e-13
Time:	15:55:38	Log-Likelihood:	-1120.5
No. Observations:	400	AIC:	2245.
Df Residuals:	398	BIC:	2253.
Df Model:	1		
Covariance Type:	nonrobust		

F检验统计量

$F = \frac{SSR / p}{SSE / (n-p-1)}$

偏差平方 ~~Var~~ Biass
方差 Var
均方误差 MSE

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	$\boxed{0.6954}$	0.200	3.482	0.001	0.303	1.088
X1	$\boxed{1.6034}$	0.210	7.618	0.000	1.190	2.017

MSE = Biass + Var.
MSE 先↓后↑
~~方差较~~ 变量少时候
方差较小 偏差平方大
欠拟合

实验题
外部 聚类有效性 纯度
内部 决定系数 R^2

Omnibus:	2.449	Durbin-Watson:	2.120
Prob(Omnibus):	0.294	Jarque-Bera (JB):	2.200
Skew:	-0.093	Prob(JB):	0.333
Kurtosis:	2.688	Cond. No.	1.05

变量多，偏差平方较小
但是方差较大，过拟合
检验图 里面是 $k - \beta(k)$

图 3 Python 的运行结果（一个特征）

2.（5 分）根据图 3 中 Python 运行的结果，请给出当X_1的取值为 0.5 时，y的点预测。同时，阐述如何计算其$1-\alpha$的预测区间。

3.（5 分）同学 B 将特征X_1和X_2同时纳入线性回归模型，并利用 Python 得到结果，如图 4 所示。将图 3 和图 4 进行比较，发现在线性回归模型中R^2从 0.127 提升到了 0.320，即结果为$R^2_{mod_1} = 0.127 \le R^2_{mod_2} = 0.320$。请问这个结论是否普

$\hat{y}_0 = x_0^T (X^T X)^{-1} X^T y = x_0^T \hat{\beta}$

$\hat{y}_0 \sim N(x_0^T\beta, \delta^2 x_0^T (X^T X)^{-1} x_0)$

$\hat{\beta} \sim N(\beta, \delta^2 (X^T X)^{-1})$

$y_0 \sim N(x_0^T\beta, \delta^2)$

$\hat{y}_0 - y_0 \sim N(\cdots, \cdots)$

问题其实在说：X变多了，R^2变大这件事普遍存在吗？
No! X_1, X_2存在线性关系时，如X_1为男性 X_2表示女性 $X_1+X_2=1$
在X_1的基础上增加X_2不会提升R^2

ymy写错了其实

$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$
SST不变 $R_{P+1}^2 \geq R_P^2$
∴具体推证略

遍存在？如果是，请证明它；如果不是，请举出反例。

Dep. Variable:	y	R-squared:	0.320
Model:	OLS	Adj. R-squared:	0.317
Method:	Least Squares	F-statistic:	93.54
Date:	Fri, 02 Dec 2022	Prob (F-statistic):	5.19e-34
Time:	17:04:30	Log-Likelihood:	-1070.5
No. Observations:	400	AIC:	2147.
Df Residuals:	397	BIC:	2159.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6500	0.177	3.682	0.000	0.303	0.997
X1	1.4421	0.187	7.728	0.000	1.075	1.809
X2	1.8526	0.174	10.618	0.000	1.510	2.196

Omnibus:	2.166	Durbin-Watson:	2.004
Prob(Omnibus):	0.339	Jarque-Bera (JB):	2.246
Skew:	-0.169	Prob(JB):	0.325
Kurtosis:	2.859	Cond. No.	1.12

图 4 Python 的运行结果（两个特征）

4.（5 分）经验所知，R^2越大表明特征的拟合效果越好。于是，同学 C 逐一将特征放入线性回归模型中。具体方案是，第一个模型的特征是X_1；第二个模型的特征是X_1和X_2；第三个模型的特征是X_1, X_2和X_3，以此类推。结果发现R^2的数值如表 1 所示。

表 1 20 个模型中不同特征维度下的R^2值

维度	1	2	3	4	5	6	7	8	9	10
R^2	0.127	0.320	0.495	0.568	0.637	0.707	0.779	0.841	0.902	0.948
维度	11	12	13	14	15	16	17	18	19	20
R^2	0.949	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.952

请问，R^2是否适合作为模型选择的指标？并请说明理由。如果不是，请给出一个改进方案。ymy：不适合

因为特征越多，通常会使R^2提升
这种方法未考虑模型的复杂程度
容易过拟合。改进方案：
使用AIC或BIC作模型选择的指标
或 $\tilde{R}^2=1-\frac{n-1}{n-p-1}(1-R^2)$

线性SVM $\min_{w,b} L(w,b)=\frac{1}{2}\|w\|^2$

三、（本题共 10 分）

请阐述一下，如何诊断出数据中存在多重共线性？（提示：只需要提供一种完整的方案）。

感知机 $\min_{w,b} L(w,b)=-\sum_{x_i \in M} y_i(wx_i+b)$

四、（本题共 15 分）

比较感知机和线性 SVM 的损失函数。

感知机：$L(w,b)=-\sum_{x_i \in M} y_i(wx_i+b)$

线性SVM损失函数？？？

对偶问题那个问题 min ◯ 吗？

五、（本题共 10 分） $P(x,y)$ $P(y|x)$

1.（5 分）解释生成式模型和判别式模型，并分析二者的不同点；

2. 列出三种判别式模型（3 分）和两种生成式模型（2 分）。

SVM 感知机 决策树、 隐马尔可夫，朴素贝叶斯

HMM 隐马尔可夫模型 ∈ 生成式，

六、（本题共 25 分）

考虑利用线性支持向量机对如下两类可分数据进行分类：

+1：(1,1), (2,2), (2,0)

-1：(0,0), (1,0), (0,1)

1.（8 分）在图中做出这 6 个训练点，构造具有最优超平面和最优间隔的权重向量；

2.（4 分）哪些是支撑向量？

3.（13 分）通过寻找拉格朗日乘子来构造在对偶空间的解，并将它与第一小问中的结果比较。

1、$\min \frac{1}{2}\|w\|^2=\frac{1}{2}(w_1^2+w_2^2)$

s.t. $w_1+w_2+b\geq 1$

$$\begin{cases} 2w_1+2w_2+b\geq 1 \\ 2w_1+b\geq 1 \\ -b\geq 1 \\ -(w_1+b)\geq 1 \\ -(w_2+b)\geq 1 \end{cases}$$

解得 $w_1=2$

$w_2=2$

$b=-3$

故最优超平面 $2x^{(1)}+2x^{(2)}-3=0$

权重向量 $(w^T,b)^T=(2,2,-3)$

2、取 $f(x_i)=y_i(wx_i+b)$

$f(x_1)=1$ $f(x_2)=5$ $f(x_3)=1$

$f(x_4)=3$ $f(x_5)=1$ $f(x_6)=1$

故 (1,1), (1,0), (0,1) (2,0) 为…

3、$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6}\alpha_i\alpha_j y_i y_j (x_i x_j)-\sum_{i=1}^{6}\alpha_i$

s.t. $\boxed{\sum_{i=1}^{6}\alpha_i y_i=0}$ ☆

可得对偶问题 $\min_{\alpha} \alpha_1^2+4\alpha_2^2+2\alpha_3^2+\frac{1}{2}\alpha_5^2+\frac{1}{2}\alpha_6^2$

$+4\alpha_1\alpha_2+2\alpha_1\alpha_3-\alpha_1\alpha_5-\alpha_1\alpha_6$

$+4\alpha_2\alpha_3-2\alpha_2\alpha_5-2\alpha_2\alpha_6$

$-2\alpha_3\alpha_5-(\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5+\alpha_6)$

s.t. $\alpha_1+\alpha_2+\alpha_3-\alpha_4-\alpha_5-\alpha_6=0$

$\alpha_i\geq 0$

$w^*=\sum_{i=1}^{6}\alpha_i^* y_i x_i$ $b^*=y_j-\sum_{i=1}^{6}\alpha_i^* y_i(x_i x_j)$

Stat-ml 最后一晚边默写边梳理

2024.1.11. 晚

期末：2024.1.12, 13:00-15:00

我各种资料有限参考了一下

σ^2 一样.

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

$\varepsilon_{ij} \sim N(0, \sigma^2)$

效应模型 $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $i=1,2\cdots a$ $j=1,2\cdots m$.

约束 $\sum_{i=1}^{a} \alpha_i = 0$.

要检验的问题 $H_0: \mu_1 = \mu_2 = \cdots \mu_a$ VS $H_1: \cdots$ 方法：F检验.

$\frac{SS_E}{\sigma^2} \sim \chi^2(n-a)$ $\frac{SS_A}{\sigma^2} \overset{H_0}{\sim} \chi^2(a-1)$ 则 $F_0 = \frac{\frac{SS_A}{a-1}}{\frac{SS_E}{n-a}} \overset{H_0}{\sim} F(a-1, n-a)$

$P<\alpha$ 拒绝!

或 $F_0 > F_{1-\alpha}(a-1, n-a)$

$E(SS_A) = \sigma^2 (a-1) + \boxed{\sum_{i=1}^{a} \alpha_i^2}$

$SS_A + SS_E = SS_T$

$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$

$SS_A \perp SS_E$

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

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

参数估计 $\hat{\mu} = \bar{y}_{\cdot\cdot}$

$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$

$\hat{\sigma}^2 = \frac{SS_E}{n-a}$

$\bar{y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{m})$ $\bar{y}_{j\cdot} \sim N(\mu_j, \frac{\sigma^2}{m})$

有一对排手 (i,j)

$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \sim N(\mu_i - \mu_j, \frac{2\sigma^2}{m})$

构造t分布 $t_0 = \frac{\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sqrt{\frac{2}{m}}}}{\sqrt{\frac{SS_E}{n-a}}} \sim t(n-a)$

现在就知道了 $\mu_i - \mu_j$ 的 $1-\alpha$ 置信区间

多重比较任务

~~$\bar{y}_{i\cdot} - y$~~ $|\mu_i - \mu_j|$ 不能太大.

Tukey $P(W) = P(\bigcup \{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| \geq c\})$

拒绝

不断转化, 直到拆分

关键在于构造 t 检验统计量.

$\bar{y}_{i\cdot}$ 分布, 怎样构造?

$\bar{y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{m})$ 构造1个即可, 不用两个!

都检验统计量了!

假设 $\mu_1 = \mu_2 = \cdots = \mu_a$ (H_0)

$\frac{\frac{\bar{y}_{i\cdot} - \mu_i}{\sigma\sqrt{\frac{1}{m}}}}{\sqrt{\frac{SS_E}{\sigma^2(n-a)}}} = \frac{\sqrt{m}(\bar{y}_{i\cdot} - \mu_i)}{\hat{\sigma}} = \frac{\sqrt{m}(\bar{y}_{i\cdot} - \mu)}{\hat{\sigma}} \sim t(n-a)$

临界值 get!

t化极差 $t(n-a)$ 的 $\max - \min \leq \cdots$

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$

$E(\varepsilon_i) = 0$ $var(\varepsilon_i) = \sigma^2$.

$\cdots$

$y = X\beta + \varepsilon$ $E(y) = X\beta$ $Var(y) = \sigma^2 I_n$

$\hat{\beta} = (X^T X)^{-1} X^T y$ $E(\hat{\beta}) = \beta$

$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

拟合值 $\hat{y} = X\hat{\beta} = Hy$

$H = H^2$ $Tr(H) = p+1$

~~$e = \hat{y} - y$~~ $e = y - \hat{y}$

$= (I - H) y$

$y^T e = 0$ ~~$Var(e) =$~~

$\frac{\partial (x^T B x)}{\partial x} = \cancel{x+x} (B + B^T) x$

$\boxed{n \geq p+1}$

样本量比模型参数大

$Var(e) = Var((I-H)y) = cov((I-H)y, (I-H)y) = (I-H)\,Cov(y,y)(I-H)^T$

$= (I-H)\cdot \delta^2 I_n (I-H) = \delta^2(I-H)^2 = \delta^2(I-H)$

☆ $cov(AX, BY) = \cancel{A\,cov(A,B)}\ A\,cov(X,Y)B^T$ $Var(AX+C) = A\,Var(X)A^T$

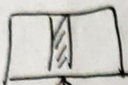
$cov(\hat\beta, e) = 0$ 线性不相关 $\hat\beta$ 与 $SS_E = e^Te$ 独立

$SS_T = \sum_{i=1}^{n}(y_i - \bar y)^2$ $SS_R = \sum_{i=1}^{n}(\hat y_i - \bar y)^2$ $SS_E = \sum_{i=1}^{n}(\hat y_i - y_i)^2$

$SS_T = SS_R + SS_E$

$\cancel{SS_R \sim}$ $\frac{SS_R}{\delta^2} \sim \chi^2(p)$ $\frac{SS_E}{\delta^2} \sim \chi^2(n-p-1)$

$1_n^T(I-H_1) = 0$

中心化  沿此方向 $X_C = (I-H_1)X_0$

$\hat\beta_{C,intercept} = 0$ $\hat\beta_{C,slope} = \hat\beta_{0,slope}$

$L_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar x_j)^2$ $L_{yy} = \sum_{i=1}^{n}(y_i - \bar y)^2$ $\boxed{X_S = X_C L}$

$L = diag\{\frac{1}{\sqrt{L_{11}}}, \cdots \frac{1}{\sqrt{L_{pp}}}\}$

$\hat\beta_{S,intercept} = 0$

$\hat\beta_{S,slope} = \frac{1}{\sqrt{L_{yy}}}L^{-1}\hat\beta_{C,slope}$

$\hat\beta_{Sj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat\beta_{Cj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat\beta_j$

第j分量

t检验 → 回归系数

F检验：模型合理否？(方程) $H_0: \beta_1 = \beta_2 = \cdots = 0$

$F_0 = \frac{SS_R/p}{SS_E/(n-p-1)} \overset{H_0}{\sim} F(p, n-p-1)$

拟合值 $\hat y_i = x_i^T\hat\beta$ $\hat e_i = y_i - \hat y_i$ 偏差

t检验：剔除次要的自变量 $H_{0j}: \beta_j = 0$ vs H_{1j} …

$\hat\beta \sim N(\beta, \delta^2(X^TX)^{-1})$

☆ $X \sim N(\mu, \Sigma)$ $AX+b \sim N(A\mu+b, A\Sigma A^T)$

则 $\hat\beta_j \sim N(\beta_j, \cancel{\delta^2 C}\ \delta^2 c_{jj})$

检验统计量 $t_j = \frac{\hat\beta_j \frac{1}{\delta\sqrt{c_{jj}}}}{\sqrt{\frac{SS_E}{\delta^2(n-p-1)}}} = \frac{\hat\beta_j}{\hat\delta\sqrt{c_{jj}}}$

$\sim t(n-p-1)$ 此处 $\hat\delta^2 = \frac{SS_E}{n-p-1}$ $R^2 = \frac{SS_R}{SS_T}$

$\hat\beta_j$ 已求得了，β_j 我们未知，但能判定

$R^2 \to 1$ 效果好

复相关系数 $R = \sqrt{R^2}$

整体的 $x_1 \cdots x_p$ 与 y 线性关系

$\hat y_0 = x_0^T\hat\beta$ $E(\hat y_0) = E(y_0)$ $y_0 \sim N(x_0^T\beta, \delta^2)$

$E(y_0) = x_0^T\beta$

☆ y_0 的线性无偏预测中，$\hat y_0$ 最小

$\hat y_0 \sim N(x_0^T\beta, \delta^2 x_0^T(X^TX)^{-1}x_0)$

相关系数：单个…线性关系

$\hat\beta \sim N(\beta, \delta^2(X^TX)^{-1})$

故 $\hat y_0 - y_0 \sim N(0, \delta^2(1 + x_0^T(X^TX)^{-1}x_0))$

~~$\beta \sim N($~~

两凑t分布(用 $\frac{SS_E}{n-p-1}$ 替换) 得 $\boxed{y_0}$ 的置信区间，$E(y_0)$ 区间就是 $\sqrt{0+\square}$

y_0 是…$\sqrt{1+\square}$ 类似的！

[欠拟合：$\hat{\beta}$, $\hat{\sigma}^2$ 有偏估计；有偏预测；预测方差小
过拟合：$\hat{\beta}$, $\hat{\sigma}^2$ 无偏~；无偏预测；预测方差大]

SS_E, R^2 不能用于变量选择，自变量个数↑ SS_E↓ $SS_E^{q+1} \le SS_E^{q}$

$R^2_{q+1} \ge R_q$ 修正 R^2 $\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$ 惩罚项

$AIC = -2\ln$(模型最大似然) $+ 2\cdot$(模型独立参数数)

$BIC = -2\ln$(模型最大似然) $+ \ln(n)\cdot$(...)

自变量很多 前进法/后退法/逐步回归

自变量间相关性弱 → 很好，符合理想

相关性强 → 违背模型假设，多重共线性

设置太多虚拟变量

$X_m = I$(男性) $X_f = I$(女性)

$X_m = 1 - X_f$ X_f 完全可由 X_m 完全线性表示

没必要把两个虚拟变量同时纳入 model

个人收入、家庭收入 } 收入 PCA合并

$X = (x_1, x_2 \cdots x_p)$ 标准化 X_s

$(X_s^T X_s)^{-1} = C = (C_{ij})_{p \times p}$

$C_{jj} = VIF_j$ 方差扩大因子 $\hat{\beta}_j$ 的

$\hat{\beta}_{sj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}} \hat{\beta}_j$ (若对 X, y 都标准化)

仅仅把 X 标准化 $Var(\hat{\beta}_j) = \frac{C_{jj}}{L_{jj}} \sigma^2$

$C_{jj} = \frac{1}{1-R_j^2}$ R_j^2：以 X_j 为因变量与其余 $p-1$ 个自变量建多元~

R_j^2↑ VIF_j↑ 多重共线性严重

$VIF = \frac{1}{p}\sum_{j=1}^{p} VIF_j$

$X_s^T X_s$ 存在一个特征值近似为0，则 X 的列向量间必存在多重共线性

特征值法 → $X^T X$ 特征值 $\lambda_1 \ge \cdots \lambda_p$

条件数 $k_j = \sqrt{\frac{\lambda_1}{\lambda_j}}$ k_j 很大，超过阈值 → 多重~

消除多重共线性方法 { 删自变量(谨慎)；增大样本量；改进最小二乘估计(岭回归，PCA) }

岭回归 $\hat{\beta}(k) = (X^T X + kI)^{-1} X^T y$ $\hat{\beta}(k)$ 是 β 有偏估计

$\hat{\beta}(k)$ 是 $\hat{\beta}$ 的一种线性变换(本质这不对)

☆ 均方误差 $MSE(\hat{\beta}(k)) \le MSE(\hat{\beta}(0))$

L_2 范数 对 $\forall k > 0$ $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$

$MSE(\hat{\beta}(k)) = E((\hat{\beta}(k) - \beta)^T(\hat{\beta}(k) - \beta))$

A 对称 $E(x^T A x) = Tr(A\Sigma) + \mu^T A \mu$

$\Sigma = Var(x)$ $\mu = E(x)$

另一角度看岭回归估计：$\min (y - X\beta)^T(y - X\beta)$ s.t. $\beta^T\beta \le s$

岭迹 $k - \hat{\beta}_j(k)$ (第 j 个分量) → 选岭迹大致稳定的

岭参数 k 的选择 → 方差扩大因子 $C_{jj} > C_{VIF}$ (有多重…)

$Var(\hat{\beta}(k)) = \sigma^2 (X^T X + kI)^{-1} X^T X (X^T X + kI)^{-1} = \sigma^2 C(k)$

$C(k)$ 的对角线元素 $C_{jj}(k)$ → 新的 VIF k↑ $C_{jj}(k)$↓

选 k 使所有 $C_{jj}(k) \le C_{VIF}$ 确定 k

主成分估计 有偏估计 $MSE(\hat{\beta}_{PC}) < MSE(\hat{\beta})$

层次聚类、kmeans、GMM、DBSCAN 具体内容

外部聚类有效性：使用外部信息；内部～：不使用

纯度 $P = \sum_i p_i (\max_j \frac{p_{ij}}{p_i})$　熵 $E = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \ln \frac{p_{ij}}{p_i})$

内部～：~~方差~~ 紧密度 → 同一类内不同个体间紧密关联的度量

方差 ⇒ 低方差紧密度小

区分度 → 不同类间区别程度的度量　密度　对应 RS、RMSSTD?

CH指数 → 最大化，可找到最优聚类数目 K.　GMM 具体看那几页！

（类内差异小，类间差异大）

$\alpha_1(i) = \pi(i) \cdot b_i(o_1)$　~~$\alpha_{t+1}(i)=$~~ $\alpha_{t+1}(i) = (\sum_{j=1}^{n} \alpha_t(j) \cdot a_{ji}) \cdot b_i(o_{t+1})$　$P(O|\lambda) = \sum_{i=1}^{n} \alpha_T(i)$

$\beta_T(i) = 1$　$\beta_t(i) = (\sum_{j=1}^{n} \beta_{t+1}(j) \cdot a_{ij} \cdot b_j(o_{t+1}))$　$P(O|\lambda) = \sum_{i=1}^{n} \pi_i b_i(o_1) \beta_1(i)$