2014.4.17

第十八讲. 一般线性假设及单因素方差分析

回顾：部分回归系数的显著性检验

全模型: $Y_{n\times1}=X_{n\times p}\beta_{p\times1}+\varepsilon_{n\times1}=X_1\beta_{(1)}+X_2\beta_{(2)}+\varepsilon,\quad \varepsilon\sim N(\mathbf{0},\sigma^2 I_n)$

其中$X=(X_1,X_2)$，X_1第一列为$\mathbf{1}$，$\beta=\begin{pmatrix}\beta_{(1)}\\ \beta_{(2)}\end{pmatrix}$，$\beta_{(2)}$长度为$q$.

原假设H_0：$\beta_{(2)}=\mathbf{0}_{q\times1}$，

子模型：$Y=X_1\beta_{(1)}+\varepsilon$，$\varepsilon\sim N(\mathbf{0},\sigma^2 I_n)$（原假设下的模型）

$$F=\frac{\|\hat{Y}-\hat{Y}_0\|^2/q}{\|Y-\hat{Y}\|^2/(n-p)}=\frac{\|X_2^{\perp}\hat{\beta}_{(2)}\|^2/q}{\hat{\sigma}^2}$$

其中$\hat{Y}=P_X Y$，$\hat{Y}_0=P_{X_1}Y$，且$\hat{Y}-\hat{Y}_0=X_2^{\perp}\hat{\beta}_{(2)}$ 其中$X_2^{\perp}=X_2-P_{X_1}X_2$

定理：原假设下 $F\sim F_{q,n-p}$

证明1：利用 $F=\dfrac{\|X_2^{\perp}\hat{\beta}_{(2)}\|^2}{q\hat{\sigma}^2}$。

由$\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1})$，知$\hat{\beta}_{(2)}\sim N(\beta_{(2)},\sigma^2(X_2^{\perp}{}'X_2^{\perp})^{-1})$

H_0成立时，$\beta_{(2)}=0$，所以$A=\|X_2^{\perp}\hat{\beta}_{(2)}\|^2=\hat{\beta}_{(2)}'X_2^{\perp}{}'X_2^{\perp}\hat{\beta}_{(2)}\sim\sigma^2\chi_q^2$

另外，$B=(n-p)\hat{\sigma}^2\sim\sigma^2\chi_{n-p}^2$，且与$\hat{\beta}$独立，

所以$\dfrac{A/q}{B/(n-p)}=\dfrac{\|X_2^{\perp}\hat{\beta}_{(2)}\|^2}{q\hat{\sigma}^2}=F\sim F_{k,n-p}$

证明2：利用$F=\dfrac{n-p}{q}\times\dfrac{\|\hat{Y}-\hat{Y}_0\|^2}{\|Y-\hat{Y}\|^2}$。

(1) $Y-\hat{Y}=(I_n-P_X)Y=(I_n-P_X)\boldsymbol{\varepsilon}$

$\dfrac{(n-p)\hat{\sigma}^2}{\sigma^2}=\dfrac{\|Y-\hat{Y}\|^2}{\sigma^2}\sim\chi_{n-p}^2$(不论$H_0$成立与否).

(2) 原假设下$Y=X_1\beta_{(1)}+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2 I_n)$

$\Rightarrow\hat{Y}-\hat{Y}_0=P_{X_2^{\perp}}Y=P_{X_2^{\perp}}(X_1\beta_{(1)}+\boldsymbol{\varepsilon})=P_{X_2^{\perp}}\boldsymbol{\varepsilon}$

$\Rightarrow\|\hat{Y}-\hat{Y}_0\|^2=\boldsymbol{\varepsilon}'P_{X_2^{\perp}}\boldsymbol{\varepsilon}\sim\sigma^2\chi_q^2$

(3) 因为$(I_n-P_X)P_{X_2^{\perp}}=0$，所以$\|\hat{Y}-\hat{Y}_0\|^2$与$\|Y-\hat{Y}\|^2$独立，

$$\Rightarrow F=\frac{\|\hat{Y}-\hat{Y}_0\|^2/q}{\|Y-\hat{Y}\|^2/(n-p)}\sim F_{q,n-p}$$

ANOVA：F可以表达为比较模型的拟合值、残差、拟合优度

$$F=\frac{\|\hat{Y}-\hat{Y}_0\|^2/q}{\|Y-\hat{Y}\|^2/(n-p)}$$

$$F=\frac{n-p}{q}\times\frac{\mathrm{RSS}_0-\mathrm{RSS}}{\mathrm{RSS}}$$

$$F=\frac{n-p}{q}\times\frac{R^2-R_0^2}{1-R^2}$$

回归方程的显著性检验

$Y=\mathbf{1}\beta_0+Z\gamma+\varepsilon,\beta_0$是截距

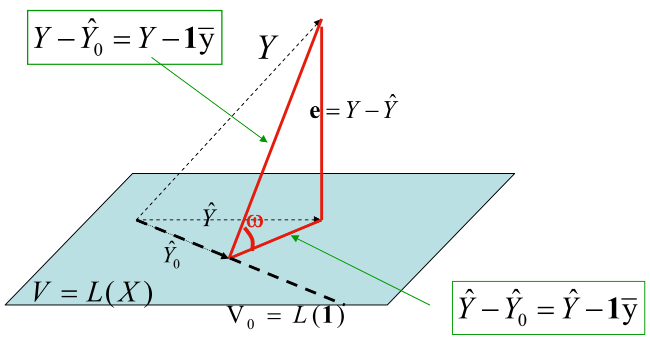
$H_0:\gamma=(\beta_1,\ldots,\beta_{p-1})'=0$

记$Z^{\perp}=Z-P_{\mathbf{1}}Z=Z-\mathbf{1}\bar{x}'$为自变量的中心化矩阵。

回归方程显著性的 F检验：

$$F=\frac{n-p}{p-1}\times\frac{\|\hat{Y}-\hat{Y}_0\|^2}{\|Y-\hat{Y}\|^2}=\frac{\|Z^{\perp}\hat{\gamma}\|^2}{(p-1)\hat{\sigma}^2}\sim_{H_0}F_{p-1,n-p}$$

其中$\hat{Y}_0=\mathbf{1}\bar{y}$，所以 $F=\frac{n-p}{p-1}\times\frac{R^2}{1-R^2}$



$$F=\frac{n-p}{p-1}\times\frac{\|\hat{Y}-\hat{Y}_0\|^2}{\|Y-\hat{Y}\|^2}=\frac{n-p}{p-1}\times\frac{R^2}{1-R^2}=\frac{n-p}{p-1}\times(ctg\omega)^2$$

$\cos^2\omega=\frac{\|\hat{Y}-\mathbf{1}\bar{y}\|^2}{\|Y-\mathbf{1}\bar{y}\|^2}=R^2,\omega=$中心化后Y与X平面的最大夹角

一般线性假设

一般线性假设通常表述为

$H_0:A\beta=\mathbf{0}_{q\times1}$，$q\le p,A$为$q\times p$已知矩阵(行满秩).

记均值向量 $\boldsymbol{\mu}=X\beta$，全模型可表示为$:Y=\boldsymbol{\mu}+\boldsymbol{\varepsilon},\boldsymbol{\mu}\in V=L(X)$
原假设下β处于q个约束之下，则$\boldsymbol{\mu}\in$某个$V_0\subset V,V_0$由A决定
(比如若$A=(0,\ I_q)$,则$A\beta=\beta_{(2)}=0,\boldsymbol{\mu}=X\beta=X_1\beta_{(1)}\in L(X_1)\subset V$).
记$\hat{Y}=P_V Y$，$\hat{Y}_0=P_{V_0}Y$

一般线性假设(H_0：$\boldsymbol{\mu}\in V_0$)的F - 检验为：

$$F=\frac{\|\hat{Y}-\hat{Y}_0\|^2/q}{\|Y-\hat{Y}\|^2/(n-p)}=\frac{(RSS_0-RSS)/q}{RSS/(n-p)}\sim_{H_0}F_{q,n-p}$$

原假设下的A,X决定的投影空间V_0是什么？

例1. $H_0: A\beta=0$, 其中 $A=\begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ & & & & \cdots & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$

$\Leftrightarrow H_0: \beta_1=\ldots=\beta_{p-1}$,

H_0成立时

$$Y=\mathbf{1}\beta_0+\mathbf{x}_1\beta_1+..+\mathbf{x}_{p-3}\beta_{p-3}+\mathbf{x}_{p-2}\beta_{p-2}+\mathbf{x}_{p-1}\beta_{p-1}+\varepsilon$$
$$=\mathbf{1}\beta_0+(\mathbf{x}_1+..+\mathbf{x}_{p-1})\gamma+\varepsilon$$

记 $X_0=(\mathbf{1},\ \mathbf{x}_1+\ldots+\mathbf{x}_{p-1})$，所以$V_0=L(X_0)$

例2. $A=(0\quad 1\quad 1\quad -2\quad 0\quad \ldots\quad 0)$, $H_0: A\beta=0, \Leftrightarrow H_0: \beta_1+\beta_{p-3}=2\beta_3$.

H_0成立时

$$Y=\mathbf{1}\beta_0+\mathbf{x}_1\beta_1+\mathbf{x}_2\beta_2+\mathbf{x}_3\beta_3+\mathbf{x}_4\beta_4+..+\varepsilon$$
$$=\mathbf{1}\beta_0+(\mathbf{x}_1+\mathbf{x}_3/2)\beta_1+(\mathbf{x}_2+\mathbf{x}_3/2)\beta_2+\mathbf{x}_4\beta_4+..+\varepsilon,$$

记 $X_0=(\mathbf{1}, \mathbf{x}_1+\mathbf{x}_3/2, \mathbf{x}_2+\mathbf{x}_3/2, \mathbf{x}_4,\ldots,\mathbf{x}_{p-1})$，所以$V_0=L(X_0)$

对于一般线性假设：

$H_0: A\beta=\mathbf{0}_{q\times 1}$, $q\le p$, A为$q\times p$已知矩阵(行满秩).

由$\boldsymbol{\mu}=X\beta \Rightarrow \beta=(X'X)^{-1}X'\boldsymbol{\mu}$，所以

$$A\beta=A(X'X)^{-1}X'\boldsymbol{\mu}=0 \Leftrightarrow \boldsymbol{\mu}\perp L(X(X'X)^{-1}A'),\quad \boldsymbol{\mu}\in L(X)$$

$$V_0=L(X(X'X)^{-1}A')^{\perp}\cap L(X)\Rightarrow P_{V_0}=P_X-P_{X(X'X)^{-1}A'},$$

$$\hat{Y}_0=\hat{Y}-X(X'X)^{-1}A'\left[A(X'X)^{-1}A'\right]^{-1}A(X'X)^{-1}X'Y$$

$$\Rightarrow \|\hat{Y}-\hat{Y}_0\|^2=\hat{\beta}'A'\left[A(X'X)^{-1}A'\right]^{-1}A\hat{\beta}\Rightarrow$$

$$F=\frac{n-p}{q}\frac{\|\hat{Y}-\hat{Y}_0\|^2}{\|Y-\hat{Y}\|^2}=\hat{\beta}'A'\left[A(X'X)^{-1}A'\right]^{-1}A\hat{\beta}\big/q\hat{\sigma}^2$$

注1: 事实上, $H_0: A\beta=0$的检验可直接构造得到：

H_0下，$A\hat{\beta}\sim N_q(0,\sigma^2A(X'X)^{-1}A')\Rightarrow \hat{\beta}'A'(A(X'X)^{-1}A')^{-1}A\hat{\beta}/\sigma^2\sim\chi^2_q$

度量了H_0成立的证据，plug-in $\hat{\sigma}^2$并除以q,即得到

$$F=\hat{\beta}'A'(A(X'X)^{-1}A')^{-1}A\hat{\beta}/q\hat{\sigma}^2\overset{H_0}{\sim}F_{q,n-p}$$

注2：如果原假设具有如下形式

$H_0: A\beta=\mathbf{c}_{q\times 1}$ ($\mathbf{c}$已知), A为$q\times p$已知矩阵(行满秩)

$$F=\frac{(A\hat{\beta}-c)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta}-c)}{q\hat{\sigma}^2}\sim_{H_0}F_{q,n-p}$$

比如，R输出中给出了$\hat{\beta}$及其标准差$se(\hat{\beta})$，以及$H0:\beta=0$的t-检验。如果检验$H0:\beta=\beta_0$(已知),只需计算$t=(\hat{\beta}-\beta_0)/se(\hat{\beta})$

单因素方差分析(one-way anova)

随机化控制试验：某因子变量有K个水平(处理, treatment),
对研究对象随机分配处理 (随机分组), 观察响应y,
考察因子变量有无效应.　数据和模型假设：

第k组$y_{k1},\ldots,y_{kn_k}\sim N(\mu_k,\sigma^2), k=1,\ldots,K$,各组独立，

$H0: \mu_1=\mu_2=\ldots=\mu_K$

$$
\begin{array}{c} \text{第一组} \\ \\ \\ \text{第二组} \\ \\ \\ \\ \text{第}K\text{组} \\ \\ \end{array}
\begin{pmatrix} y_{11}\\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{K1} \\ \vdots \\ y_{Kn_K}\end{pmatrix}
=
\begin{array}{c} \mathbf{x}_1 \quad \mathbf{x}_2 \ldots \quad \mathbf{x}_K \\ \begin{pmatrix} 1&0&\ldots&0\\ \vdots&\vdots&\ldots&0 \\ 1&0&\ldots&0\\ 0&1&\ldots&0\\ \vdots&\vdots&&\\ 0&1&\ldots&0\\ \vdots&\vdots&\vdots&\vdots\\ 0&0&0&1\\ \vdots&\vdots&\vdots&\vdots\\ 0&0&0&1 \end{pmatrix}\end{array}
\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_K\end{pmatrix} + \boldsymbol{\varepsilon}
$$

全模型$Y = \mathbf{x}_1\mu_1 + \mathbf{x}_2\mu_2 + \ldots + \mathbf{x}_K\mu_K + \varepsilon$

$\hat{Y} = (\bar{y}_{1\bullet},\ldots\bar{y}_{1\bullet},\bar{y}_{2\bullet},\ldots,\bar{y}_{2\bullet},\ldots,\bar{y}_{K\bullet},\ldots\bar{y}_{K\bullet}),$

其中$\bar{y}_{k\bullet} = (y_{k1} + \ldots y_{kn_k})/n_k$

原假设下$\mu_1 = \mu_2 = \ldots = \mu_K$,

$Y = (\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_K)\mu_1 + \varepsilon = \mathbf{1}\mu_1 + \varepsilon$

$\hat{Y}_0 = (\bar{y}_{\bullet\bullet},\ldots,\bar{y}_{\bullet\bullet}),\ \bar{y}_{\bullet\bullet} = \sum\sum y_{ij}/n,\ n = n_1 + ..+ n_K$

$$F = \frac{\|\hat{Y}-\hat{Y}_0\|^2/(K-1)}{\|Y-\hat{Y}\|^2/(n-K)} = \frac{\left(n_1(\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet})^2 + \ldots + n_K(\bar{y}_{K\bullet}-\bar{y}_{\bullet\bullet})^2\right)/(K-1)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2/(n-K)}$$

$$= \frac{SS_{\text{组间}}/(K-1)}{SS_{\text{组内}}/(n-K)} \boxed{\sim_{H_0} F_{K-1,n-K}}$$

事实上，通常的做法是直接进行方差（平方和）分解:

$$SS_{\text{总}} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{\bullet\bullet})^2 = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet}+\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2$$

$$= \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2 + \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2 \triangleq SS_{\text{组间}} + SS_{\text{组内}}$$

$$F = \frac{SS_{\text{组间}}/(K-1)}{SS_{\text{组内}}/(n-K)} \sim F_{K-1,n-K} \quad \text{(原假设下)}$$

称为单因素方差分析(one-way anova)

方差分析表:

来源	因子/分组	残差
平方和	$SS_{\text{组间}}$	$SS_{\text{组内}}$
自由度	$K-1$	$n-k$

AOV函数(Analysis of Variance)

> aov(y ~ group)

group 代表了每个观察所属组别，是因子变量。

```
> aov(TS~D, data=sleep1)
Call:
   aov(formula = TS ~ D, data = sleep1)

Terms:
                          D      Residuals
Sum of Squares    457.2556    752.4122
Deg. of Freedom          4          53

Residual standard error: 3.767818
Estimated effects may be unbalanced
4 observations deleted due to missingness

# F= (457.2556/4)/(752.4122/ 53)=8.052
```

← 方差分析表

非参数检验方法（不假设正态分布）

两样本: Wilcoxon秩和检验（或Mann-Whitney- Wilcoxon检验)

处理组：$y_1,\ldots,y_{n_1}$ iid $\sim f(x-\mu_1)$, f密度,但未知, μ_1为均值

对照组：$y_{n_1+1},\ldots,y_{n_1+n_2}$ iid $\sim f(x-\mu_2)$, μ_2为均值

$H0:\mu_1=\mu_2$

```
> x=c(12,22,8,5,30)
> rank(x)
[1] 3 4 2 1 5
```

y_i在所有$n=n_1+n_2$个观察值中的排名/秩记为$\mathrm{R_i}$

第一组的平均秩$\overline{\mathrm{R}}=(\mathrm{R}_1+\ldots+\mathrm{R}_{n_1})/n_1$

第二组的平均秩$\overline{\mathrm{R}}=(\mathrm{R}_{n_1+1}+\ldots+\mathrm{R}_{n_1+n_2})/n_2$

原假设$H0:\mu_1=\mu_2$下,两组的平均秩应该相差不大,

Wilcoxcon检验统计量为:

$\mathrm{W}=\overline{\mathrm{R}}_2-\overline{\mathrm{R}}_1$

如何评价W的显著性? 只要能否算出其p值.

原假设$H0:\mu_1=\mu_2$下,两组无差异,那么我们随机置换 $y_1,\ldots,y_{n_1},\ldots y_{n_1+n_2}$,即随机分组,置换后的数据重新计算平均秩的差值,记为W*, 如果H0成立,那么$\mathrm{W}^*\approx\mathrm{W}$

反复置换N次,所得平均秩之差为$\mathrm{W}_1^*,\ldots,\mathrm{W}_N^*$

W的p值： $\mathrm{p}=\{|\mathrm{W}_i^*|>|\mathrm{W}|\text{的个数}\}/\mathrm{N}$

```
> wilcox.test ( y ~group)
```


K样本: Kruskal-Wallis检验

组1：$y_1,\ldots,y_{n_1}$ iid $\sim f(x-\mu_1)$, f密度,但未知, μ_1为均值

组2：$y_{n_1+1},\ldots,y_{n_1+n_2}$ iid $\sim f(x-\mu_2)$,

...

组K：$y_{n-n_K+1},\ldots,y_n$ iid $\sim f(x-\mu_K)$,

$H0:\mu_1=\mu_2=\ldots=\mu_K$

y_i在所有$n=n_1+n_2$个观察值中的排名/秩记为$\mathrm{R_i}$

记第k组的平均秩$\overline{\mathrm{R}}_\mathrm{k}$,原假设下$\overline{\mathrm{R}}_1,\ldots,\overline{\mathrm{R}}_\mathrm{K}$应该差别不大.

```
> kruskal.test ( y ~ group)
```