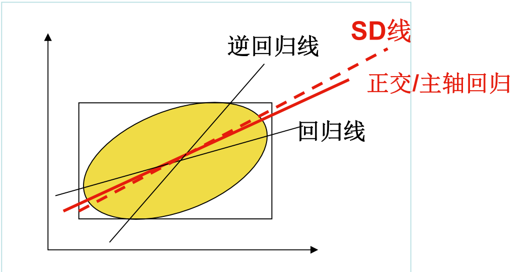
第十二讲. 特殊话题



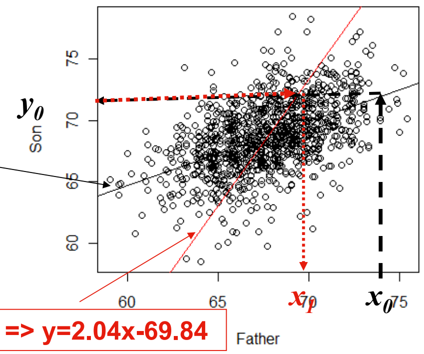
Outline

1. 正交/主轴回归及其它

- 哪个是 *y*（*响应*）,哪个是 *x*（*自变量*）?
  - 在身高-体重数据中，可以从身高预测体重，也可从体重预测身高；
  - Pearson父子身高数据中，我们可以从父亲身高预测儿子身高，也可以从儿子身高判断父亲身高。
  - 两个变量对称、平等。
- 通常的回归是在给定自变量条件下进行的，为了考虑自变量的随机性 (比如自变量带误差模型，error-in-variable model), 可使用对称回归：即平等对待x, y。
- 对称回归的估计方法不再用通常的LS (OLS: ordinary LS), 而是采用Total least squares, 正交回归是其特殊情况。

例1. Pearson父子身高数据

y~x:
y=33.89+0.514*x



x~y (逆回归，inverse regression)
x=34.125+0.4886*y

=> y=2.04x-69.84

对给定的 x_0, 由第一个方程得预测 y_0,
对给定的 y_0, 由第二个方程(红线)得预测 x_1
一般地，$x_1 \neq x_0$ (不对称!)

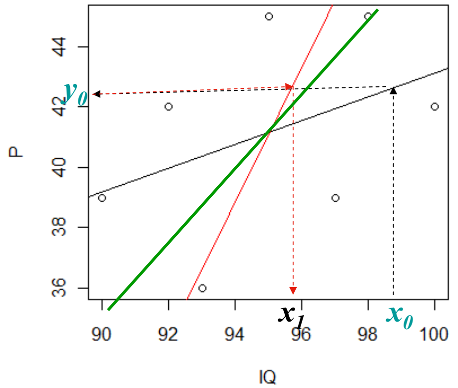
例2：8个小孩的IQ和学习成绩P (performance)

IQ =c(90, 92, 93, 95, 97, 98,100);

P = c(39, 42, 36, 45, 39, 45, 42);

回归　lm(P~IQ)：P=3.6+0.4*IQ

逆回归 lm(IQ~P)：IQ=76.5+0.5*P　　=> P = -163 + 2*IQ



正交/主轴/Deming回归：　P=-40.5+0.86*IQ, IQ=47.1+1.16*P

对称回归分析

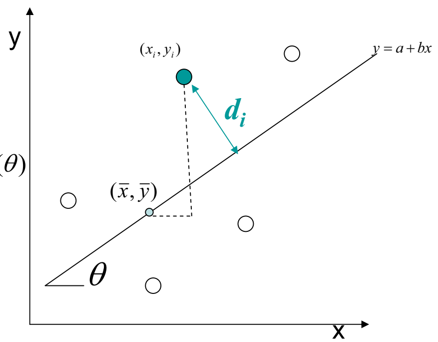
- 如果难以确定x, y中哪个是响应变量, 如何建立两者之间的函数关系?
- 如果x, y地位对等(对称)，y~x以及 x~y 都不合理。应该使用对称回归方法，包括
  - Pearson 正交回归(orthogonal reg)，也称作主轴回归(major-axis regression)，或 Deming 回归, 是Total least square 的特殊情况。
  - reduced major reg (SD 线),
  - double regression 或 bisector reg

(1). 正交回归/ major-axis regression

$b = \tan(\theta)$

$d_i = [y_i - \bar y - (x_i - \bar x)\tan(\theta)]\cos(\theta)$

$= (y_i - \bar y)\cos(\theta) - (x_i - \bar x)\sin(\theta)$



Major - axis Regression :

$$\min \sum d_i^2 = \min \sum \left((y_i - \bar y)\cos(\theta) - (x_i - \bar x)\sin(\theta)\right)^2$$

$$\text{对}\theta\text{求导得}: \tan(2\theta) = \frac{2s_{xy}}{s_{xx} - s_{yy}} = \frac{2b}{1-b^2}$$

$$\Rightarrow \hat b_{ma} = \frac{s_{xx} - s_{yy} + \sqrt{(s_{xx} - s_{yy})^2 + 4s_{xy}^2}}{2s_{xy}}$$

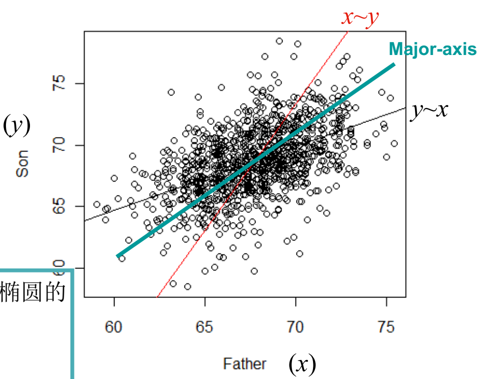
记$\hat b_1 = s_{xy}/s_{xx}$为$y \sim x$的斜率估计

记$\hat b_2 = s_{yy}/s_{xy}$为$x \sim y$的斜率估计

则$\hat b_1 \hat b_2 = s_{yy}/s_{xx} \ge 0$,　$\hat b_1/\hat b_2 = r^2 \le 1$

$$\hat b_{ma} == \frac{(\hat b_2 - 1/\hat b_1) + \text{sgn}(r)\sqrt{4 + (\hat b_2 - 1/\hat b_1)^2}}{2}$$

介于$\hat b_1, \hat b_2$之间



$\hat\theta = \arctan(\hat b)$为如下二元正态分布等概率椭圆的主轴$(major-axis)$/主成分方向：

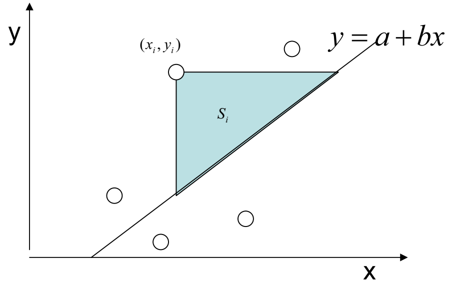
$$\frac{x^2}{s_{xx}} - 2r\frac{xy}{\sqrt{s_{xx}s_{yy}}} + \frac{y^2}{s_{yy}} = c$$

(2).Reduced major axis regression： the SD line

$\min\limits_{a,b}\sum S_i \Rightarrow$

$\hat{b}_{RMA} = \mathrm{sgn}(r)\sqrt{\hat{b}_1\hat{b}_2} = \mathrm{sgn}(r)\sqrt{\frac{s_{yy}}{s_{xx}}}$

直线方程为SD Line：$y-\bar{y} = \mathrm{sgn}(r)\sqrt{\frac{s_{yy}}{s_{xx}}}(x-\bar{x})$

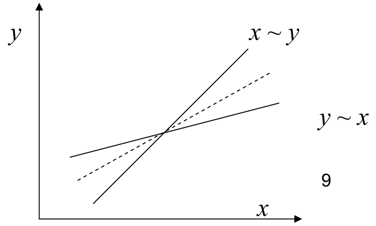


(3).Bisector regression (double regression):

平分y~x, x~y回归直线的夹角

$y \leftrightarrow x$(bisector regression)

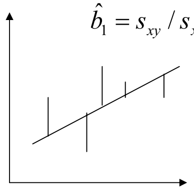
$$\hat{b}_{bisect} = \frac{\hat{b}_1\hat{b}_2 - 1 + \sqrt{(1+\hat{b}_1^2)(1+\hat{b}_2^2)}}{\hat{b}_1+\hat{b}_2}$$



总结一下：

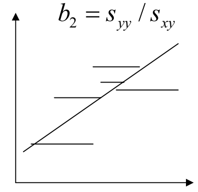
$y\sim x$

$\hat{b}_1 = s_{xy}/s_{xx}$



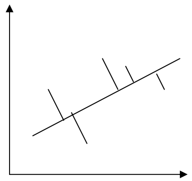
$x\sim y$

$\hat{b}_2 = s_{yy}/s_{xy}$



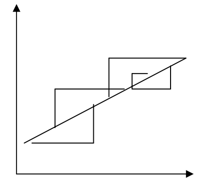
$y \leftrightarrow x$(major-axis regression)

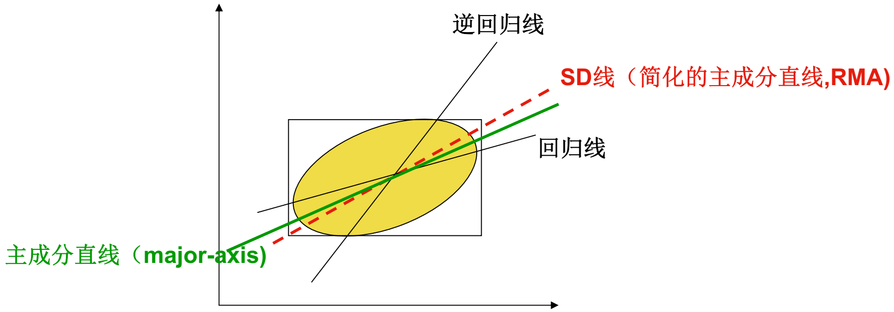
$$\hat{b}_{major-axis} = \frac{(\hat{b}_2 - 1/\hat{b}_1) + \mathrm{sgn}(r)\sqrt{4+(\hat{b}_2-1/\hat{b}_1)^2}}{2}$$



$y \leftrightarrow x$(reduced major-axis regression)

$\hat{b}_{RMA} = \mathrm{sgn}(r)\sqrt{\hat{b}_1\hat{b}_2} = \mathrm{sgn}(r)\sqrt{\frac{s_{yy}}{s_{xx}}}$





回归：$\min E\left(y-(a+bx)\,|\,x\right)^2$

逆回归：$\min E\left(x-(c+dy)\,|\,y\right)^2, x=c+dy \Rightarrow y = x/d - c/d$

Major-axis：$\max \mathrm{E}\,\|(x,y) - \mathrm{P}\left((x,y)\,|\,\mathrm{u}\right)\|^2 \Rightarrow$ 主成分直线

```
> Library(MethComp)
> Deming(x,y)  #正交/主轴/Deming回归
```


2. 中心化

$(x_i, y_i), i=1,2,\ldots,n$满足模型：　$y_i = a + bx_i + \varepsilon_i$

重新改写为：　$y_i = a + b\bar{x} + b(x_i-\bar{x}) + \varepsilon_i$

记$\alpha = a + b\bar{x}, x_i^{(c)} = x_i - \bar{x}$(中心化），模型为

$$y_i = \alpha + bx_i^{(c)} + \varepsilon_i \qquad (*)$$

注意$\sum x_i^{(c)} = 0$,即$\bar{x}^{(c)} = 0$.这将使得LS的求解以及其它计算变得容易

对于模型(*),LS估计为

$$\hat{\alpha} = \bar{y},\quad \hat{b} = s_{x^{(c)}y}/s_{x^{(c)}x^{(c)}} = \frac{\sum x_i^{(c)}y_i}{\sum (x_i^{(c)})^2} = \frac{\sum (x_i-\bar{x})y_i}{\sum (x_i-\bar{x})^2}$$

$$\mathrm{var}(\hat{\alpha})=\sigma^2/n,\quad \mathrm{var}(\hat{b})=\sigma^2/\sum(x_i^{(c)})^2$$

$$\mathrm{cov}(\hat{\alpha},\hat{b})\propto \mathrm{cov}(\sum x_i^{(c)}y_i,\sum y_i)=\sum x_i^{(c)}\sigma^2=0\,!$$

最后由于$\alpha=a+b\bar{x}$，原始模型中截距项的估计为：

$$\hat{a}=\hat{\alpha}-\hat{b}\bar{x}=\bar{y}-\hat{b}\bar{x}$$

总结如下：

原模型： $\mathbf{y}=\mathbf{1}a+\mathbf{x}b+\boldsymbol{\varepsilon}$

改写为新模型： $\mathbf{y}=\mathbf{1}(a+b\bar{x})+(\mathbf{x}-\mathbf{1}\bar{x})b+\boldsymbol{\varepsilon}=\mathbf{1}\alpha+\mathbf{x}^{\perp}b+\boldsymbol{\varepsilon}$

其中$\mathbf{x}^{\perp}=\mathbf{x}-\mathbf{1}\bar{x}$（中心化）， $\alpha=a+b\bar{x}$

新模型中$\mathbf{1}\perp\mathbf{x}^{\perp}$，参数的LS估计容易求得：

$$\hat{\alpha}=\bar{y},\quad \hat{b}=\frac{\mathbf{x}^{\perp\prime}\mathbf{y}}{\mathbf{x}^{\perp\prime}\mathbf{x}^{\perp}}=\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$$

3. 过原点的回归

回归模型中，截距项为自变量为0时的响应变量的均值，可以称为*baseline*. 有时我们已知截距项为0,

称为过原点的回归模型：

$$y=bx+\varepsilon,\varepsilon\sim(0,\sigma^2),\varepsilon\perp x\quad \text{(单个自变量情形)}$$

比如研究弹簧伸长长度y,与悬挂物重量关系时，可以假设截距项为0。

模型： $y=bx+\varepsilon,\varepsilon\sim(0,\sigma^2),\varepsilon\perp x$

最小二乘：

$$\min\sum(y_i-bx_i)^2$$

求导得到正则方程：

$$\sum x_i(y_i-bx_i)=0$$

$$\Rightarrow LS\text{估计：}\ \hat{b}=\frac{\sum x_iy_i}{\sum x_i^2}$$

方差公式：

$$\mathrm{var}(\hat{b})=\frac{\sigma^2}{\sum x_i^2}$$

$$sd(\hat{b})=\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

```
> lm(y~x-1) # R命令
```


拟合值和残差：

拟合值：$\hat{y}_i=\hat{b}x_i$

残差：$e_i=y_i-\hat{y}_i=y_i-\hat{b}x_i$

残差平方和：$RSS=SS_e=S_{ee}=\sum e_i^2$

$\hat{\sigma}^2=RSS/(n-1)$

正交分解：

由正则方程知：$\sum x_ie_i=0,\Rightarrow\sum\hat{y}_ie_i=0$, 有

$$SS_y=\sum y_i^2=\sum\hat{y}_i^2+\sum e_i^2\ :=\mathrm{SS}_{\hat{y}}+\mathrm{SS}_e$$

或$\mathrm{SS}_{总}=\mathrm{SS}_{回}+\mathrm{RSS}$

复相关系数平方：

$$\mathrm{R}^2 = \frac{\mathrm{SS}_{\hat{y}}}{SS_{\mathrm{y}}}（或\mathrm{R}^2 = \frac{\mathrm{SS}_{回}}{SS_{总}}）$$

容易验证：

$$\mathrm{R}^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{b}^2 \sum x_i^2}{\sum y_i^2} = \frac{\left(\sum x_i y_i\right)^2}{\sum y_i^2 \sum x_i^2} = r_{xy}^2 = r_{\hat{y}y}^2$$

t - 检验：

正态假设下检验$H0: b = 0$

$$t = \frac{\hat{b}}{sd(\hat{b})} = \frac{\sum x_i y_i}{\hat{\sigma}\sqrt{\sum x_i^2}} = \sqrt{n-1}\,\frac{r}{\sqrt{1-r^2}} \sim_{H0} t_{n-1}$$