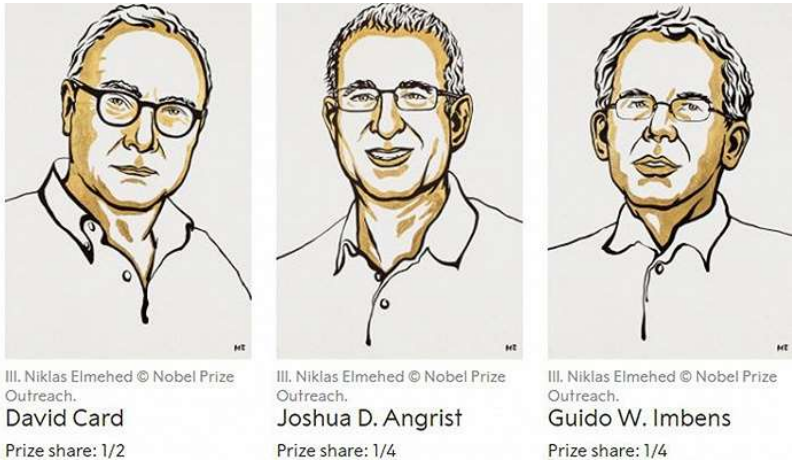
课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第七讲 简单线性回归模型（续）

2022.10.14



“LS估计的方差”的估计

Plug-in

$\mathrm{var}(\hat{b}\,|\,\mathbf{x}) = \dfrac{\sigma^2}{s_{xx}}$中将$\sigma^2$的估计代入(plug-in)得：

$$\widehat{\mathrm{var}}(\hat{b}\,|\,\mathbf{x}) = \frac{\hat{\sigma}^2}{s_{xx}},$$

标准差：$se(\hat{b}) = \sqrt{\widehat{\mathrm{var}}(\hat{b}\,|\,\mathbf{x})} = \hat{\sigma}/\sqrt{s_{xx}}$,

截距项LS估计$\hat{a}$的方差估计类似得到（一般不关心）。

Wald检验

Wald检验方法是构造检验统计量常用的方法之一（另外两种常用方法是似然比检验，Score检验）。一般地，若参数θ的估计为$\hat{\theta}$,其标准差为$se(\hat{\theta})$,则$H_0:\theta=0$的Wald检验统计量定义为：

$$W = \hat{\theta}/se(\hat{\theta})。$$

对于简单回归中的原假设$H_0:b=0$, Wald检验统计量：

$$T = \frac{\hat{b}}{se(\hat{b})} = \frac{\sqrt{s_{xx}}\,\hat{b}}{\hat{\sigma}}$$

正态模型下的统计推断

命题5. 假设模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_1, ..., \varepsilon_n \ \text{iid} \sim N(0, \sigma^2)$,则

(1) $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$

(2) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, 且$\hat{\sigma}^2$与$(\hat{a}, \hat{b})$独立

(3) $\dfrac{\sqrt{s_{xx}}(\hat{b} - b)}{\hat{\sigma}} \sim t_{n-2}$

证明:(1) $y_i \mid x_i \sim N(a + bx_i, \sigma^2) \Rightarrow$ 给定$\mathbf{x}$时，$\hat{b} = \dfrac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \sim N(b, \sigma^2 / s_{xx})$,

所以 $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \mid_{\mathbf{x}} \sim N(0,1)$与x无关，所以无条件地$\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$。

(2) 由引理1，$RSS = s_{\varepsilon\varepsilon} - s^2_{x\varepsilon}/s_{xx} = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - \left(\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \right)^2 / s_{xx}$

$$= \sum_{i=1}^{n} \varepsilon_i^2 - n\bar{\varepsilon}^2 - \left(\sum_{i=1}^{n} \left(\frac{x_i - x}{\sqrt{s_{xx}}} \right) \varepsilon_i \right)^2 \hat{=} \| \boldsymbol{\varepsilon} \|^2 - (\mathbf{u}^{\mathsf{T}} \boldsymbol{\varepsilon})^2 - (\mathbf{v}^{\mathsf{T}} \boldsymbol{\varepsilon})^2$$

其中$\mathbf{u}^{\mathsf{T}} = (1/\sqrt{n}, ..., 1/\sqrt{n}), \mathbf{v}^{\mathsf{T}} = \left((x_1 - \bar{x})/\sqrt{s_{xx}}, ..., (x_n - \bar{x})/\sqrt{s_{xx}} \right)$,模长1,$\mathbf{u} \perp \mathbf{v}$。

由第二讲引理1(2)即可得证。

第二讲引理1（2）没有证明，所以下面给出本命题的证明细节

令$\boldsymbol{\varepsilon}=(\varepsilon_1,...,\varepsilon_n)^T\sim N(0,\sigma^2 I_n)$,令$\mathbf{z}=A\boldsymbol{\varepsilon}\sim N(0,\sigma^2 I_n)$,
其中A是一个正交矩阵,其第一二行分别为$\mathbf{u}^{\mathsf{T}},\mathbf{v}^{\mathsf{T}}$:

$$A=\begin{pmatrix}\mathbf{u}^{\mathsf{T}}\\ \mathbf{v}^{\mathsf{T}}\\ *\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}}\\ \frac{x_1-\bar{x}}{\sqrt{s_{xx}}} & \frac{x_2-\bar{x}}{\sqrt{s_{xx}}} & \cdots & \frac{x_n-\bar{x}}{\sqrt{s_{xx}}}\\ * & * & \cdots & *\end{pmatrix}$$

则$z_1=\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon}=\sqrt{n}\bar{\varepsilon}$, $z_2=\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon}=\sum_{i=1}^{n}(x_i-x)\varepsilon_i/\sqrt{s_{xx}}$。因为$\|\mathbf{z}\|^2=\|\boldsymbol{\varepsilon}\|^2$（$A$正交阵），

所以 $RSS=\|\boldsymbol{\varepsilon}\|^2-(\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon})^2-(\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon})^2=\|\mathbf{z}\|^2-z_1^{\,2}-z_2^{\,2}=\sum_{i=3}^{n}z_i^{\,2}$

$(n-2)\hat{\sigma}^2/\sigma^2=RSS/\sigma^2=\sum_{i=3}^{n}z_i^{\,2}/\sigma^2\sim\chi^2_{n-2}$,且与$z_1,z_2$独立.

而$\hat{b}=b+\frac{\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2}=b+z_2/\sqrt{s_{xx}}$, $\hat{a}=\bar{y}-\hat{b}\bar{x}=a+b\bar{x}+\bar{\varepsilon}-\hat{b}\bar{x}$

$=a+z_1/\sqrt{n}-z_2\bar{x}/\sqrt{s_{xx}}$,仅与$z_1=\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon},z_2=\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon}$ 有关，所以$\hat{\sigma}^2$与$(\hat{a},\hat{b})$独立。

(3). 由（1,2）知：$\mathrm{A}=\frac{\sqrt{s_{xx}}(\hat{b}-b)}{\sigma}\sim N(0,1)$, $\mathrm{B}=\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-2}$,

且两者独立，由t分布的定义$\frac{\mathrm{A}}{\sqrt{\mathrm{B}/(n-2)}}=\frac{\sqrt{s_{xx}}(\hat{b}-b)}{\hat{\sigma}}\sim t_{n-2}$。

假设检验与置信区间

回归系数的Wald检验

原假设$H_0: b = 0$, Wald检验统计量

$$t = \hat{b} / se(\hat{b}) = \sqrt{s_{xx}}\hat{b} / \hat{\sigma} \overset{H_0}{\sim} t_{n-2}$$

检验准则：$|t| \geq t_{n-2}(\alpha/2)$时拒绝原假设。

Wald检验即相关性检验

命题6. $t = \sqrt{s_{xx}}\hat{b} / \hat{\sigma} = \sqrt{n-2}\dfrac{r}{\sqrt{1-r^2}}$。

验证：由 $\hat{b} = s_{xy} / s_{xx}$，$\hat{\sigma}^2 = \dfrac{1}{n-2}RSS = \dfrac{1}{n-2}\left(s_{yy} - s_{xy}^2 / s_{xx}\right)$,

$$t = \frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}} = \frac{s_{xy} / \sqrt{s_{xx}}}{\sqrt{\dfrac{1}{n-2}\left(s_{yy} - s_{xy}^2 / s_{xx}\right)}} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$ 。

置信区间

b的$(1-\alpha)100\%$置信区间：$\left[\hat{b} \mp \dfrac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(\alpha/2)\right]$

两样本t检验

两样本*t*-检验是回归系数显著性检验的特殊情形

$$y_1,\ldots,y_{n_1} \ \text{iid} \sim N(\mu_1,\sigma^2) \quad \leftarrow x_1,\ldots,x_{n_1}=1$$
$$y_{n_1},\ldots,y_{n_1+n_2} \ \text{iid} \sim N(\mu_2,\sigma^2) \quad \leftarrow x_{n_1+1},\ldots,x_{n_1+n_2}=0$$

给两组样本分别赋予标号$x_i=0,1$,两样本问题写成线性模型：

$$y_i = a + bx_i + \varepsilon_i \quad (a=\mu_2, b=\mu_1-\mu_2), \quad \varepsilon_i \ \text{iid} \sim N(0,\sigma^2)$$
$$H_0: b=0 \Leftrightarrow \mu_1=\mu_2$$

容易验证该模型 $H_0: b=0$的检验统计量 $t=\dfrac{\hat{b}}{\sqrt{\hat{\sigma}^2/s_{xx}}}$

等于两样本 t检验统计量 $\dfrac{\bar{y}_1-\bar{y}_2}{\sqrt{(n_1^{-1}+n_2^{-1})s^2}}$

总结：简单模型的参数及其LS估计

模型(总体)：$y = a + bx + \varepsilon$,

$\varepsilon \sim (0, \sigma^2), \varepsilon$与$x$独立

模型(样本)：$y_i = a + bx_i + \varepsilon_i$,

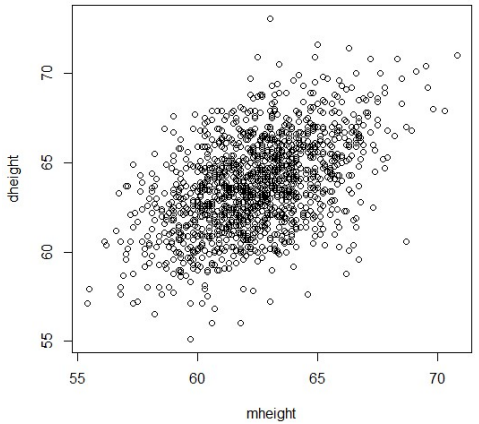
$\varepsilon_i \sim (0, \sigma^2), \varepsilon_i$与$x_i$独立

参数	估计
(1) $b = \mathrm{cov}(x, y) / \mathrm{var}(x) = \rho_{xy}\sigma_y / \sigma_x$ (2) $a = \mu_y - b\mu_x$, (3) $\sigma^2 = (1 - \rho_{xy}^2)\sigma_y^2$	(1) $\hat{b} = s_{xy} / s_{xx} = r_{xy}s_y / s_x$ (2) $\hat{a} = \bar{y} - \hat{b}\bar{x}$ (3) $\hat{\sigma}^2 = (1 - r_{xy}^2)s_y^2 \times (n-1)/(n-2)$
回归函数：$a + bx$ 误差：$\varepsilon = y - (a + bx)$ ε与x独立	拟合值：$\hat{y}_i = \hat{a} + \hat{b}x_i$ 残差$e_i = y_i - (\hat{a} + \hat{b}x_i)$ $(e_1, \ldots, e_n)^{\mathsf{T}} \perp (x_1, \ldots, x_n)^{\mathsf{T}}$
$R^2 = \dfrac{\mathrm{var}(a + bx)}{\mathrm{var}(y)} = \rho_{xy}^2$	$R^2 = \dfrac{SS_{回}}{SS_{总}} = \dfrac{s_{\hat{y}\hat{y}}}{s_{yy}} = r^2$

例1. K.Pearson收集了1375对母女身高数据(单位：英寸)，如右图。前3行数据如下：

	mheight	dheight
1	59.7	55.1
2	58.2	56.5
3	60.6	56.0

假设简单线性模型：
dheight = a+ b × mheigh + ε



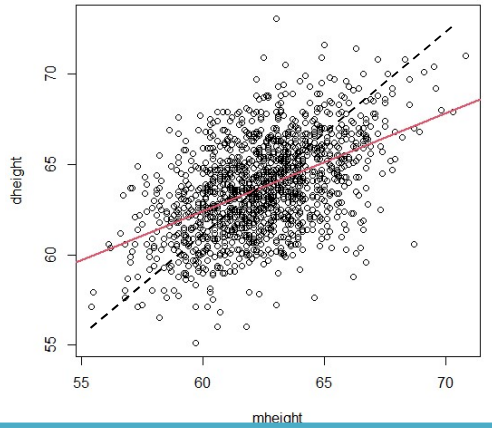
R 函数 lm, ~ 符号左边为响应，右边为自变量

```
>  myfit <- lm(dheight ~ mheight, data=Heights)

Coefficients:
(Intercept)     mheight
   29.9174       0.5417
```

$\hat{a}$ $\hat{b}$

```
> plot(Heights)   #散点图
> abline(myfit)   #添加拟合回归直线
```



注意回归效应：若母亲身高80，则女儿身高的期望(预测)为
$\hat{a}$ + 80$\hat{b}$=73 < 80

t value=Estimate ÷ Std.Error
=第1列除以第2列

结果汇总:

> summary(myfit)

Coefficients: LS估计 标准差 t检验 (前两列之比) p值

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.92	1.62	18.47	<2e-16 ***
mheight	0.54	0.026	20.77	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$\hat{b} = 0.54,$

$se(\hat{b}) = 0.026$

$t = \frac{\hat{b}}{se(\hat{b})} = \frac{0.54}{0.026} = 20.77$

$p值 < 2e-16$

Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-squared: 0.2408, Adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF, p-value: < 2.2e-16

$\hat{\sigma} = 2.266,$

$R^2 = 0.2408$（决定系数，R中称为Multiple R - squared

附：对称回归

哪个是响应？回归与逆回归

被预测的是响应。在身高-体重数据中，可以从身高x预测体重y，也可从体重y预测身高x（如果前者称为回归，后者称为逆回归）.
如果x, y地位对等，则直线描述/逼近称为对称回归。

回归: $y = a + bx + \varepsilon$, 以x_i预测y_i的误差（y-轴方向）

$$\varepsilon_i = y_i - a - bx_i$$

极小化 $\Sigma\varepsilon_i^2 = \Sigma(y_i - a - bx_i)^2 \Rightarrow \hat{b} = s_{xy}/s_{xx}$

逆回归: $x = c + dy + \delta$, 以y_i预测x_i的误差（x-轴方向）

$$\delta_i = x_i - c - dy_i$$

极小化 $\Sigma\delta_i^2 = \Sigma(x_i - c - dy_i)^2 \Rightarrow \hat{d} = s_{xy}/s_{yy}$

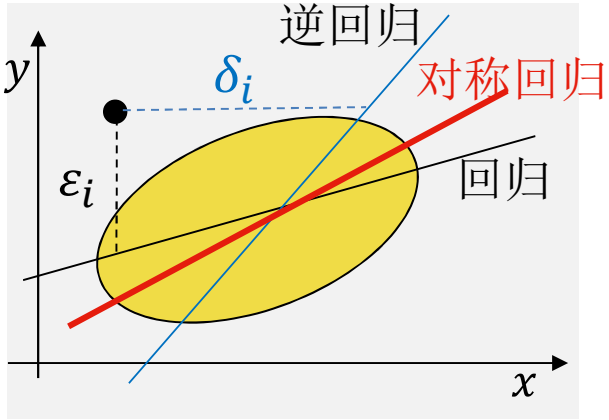
$\Rightarrow x$系数$\hat{b}_{逆} = 1/\hat{d} = s_{yy}/s_{xy}$

由回归和逆回归得到估计$\hat{b}, \hat{b}_{逆}$满足

$$\hat{b}\hat{b}_{逆} = s_{yy}/s_{xx} \geq 0, \quad \hat{b}/\hat{b}_{逆} = r^2 \leq 1$$

所以$\hat{b}, \hat{b}_{逆}$符号相同，但$|\hat{b}| \leq |\hat{b}_{逆}|$。

对称回归在回归与逆回归之间。



x, y 对称

有些时候两个变量对称、平等，我们并不希望用一个变量去预测/描述另外一个变量, 而是单纯描述两者之间的关系，可用相关系数进行描述，也可使用对称回归。

自变量x带误差/随机性

通常的回归分析是在给定自变量条件下进行的，可认为自变量是常数、非随机的。在研究自变量与响应变量的回归关系时，如果需要考虑自变量的随机性，称为自变量误差模型（error-in-variable model), 此时可使用对称回归。

例如，如果我们希望研究智力 (Int) 与成绩(P)的关系，建立模型

$$P = a + b \times Int + \varepsilon$$

智力Int不可完全测量， IQ 作为Int的测量带有误差：

$$IQ = Int + \delta$$

Total Least Sqaures

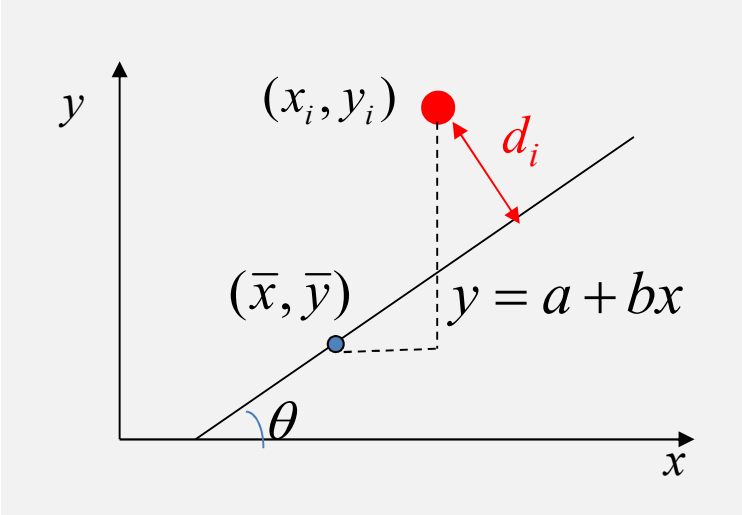
对称回归的估计方法不再用通常的LS方法，一般采用Total least squares，常用的有正交回归、约简的主轴回归等（参见后面几页 – 只需关注图示）

正交回归/主轴回归/Deming回归

数据$(x_i, y_i), i = 1, \ldots, n$

目标：求对称回归直线 $y = a + bx$

主轴回归： $\min \sum d_i^2$，其中d_i为(x_i, y_i)到$y = a + bx$的垂直距离



$b = \tan(\theta)$

$d_i = (y_i - \bar{y})\cos(\theta) - (x_i - \bar{x})\sin(\theta),$

$\sum d_i^2$对θ求导得

$$\frac{2s_{xy}}{s_{xx} - s_{yy}} = \tan(2\theta) = \frac{2b}{1 - b^2}$$

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

$\hat{\theta} = \arctan(\hat{b}_{ma})$为二元正态分布等概率椭圆

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

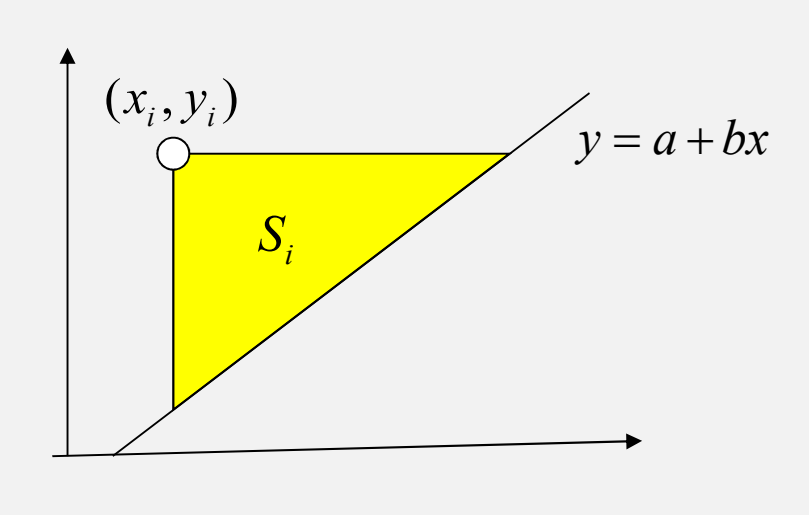
的主轴方向。

约简的主轴回归：SD线

Reduced major - axis regression

目标： $\min\limits_{a,b} \sum S_i$,

S_i为数据点(x_i, y_i)与直线$y = a + bx$之间的三角形面积

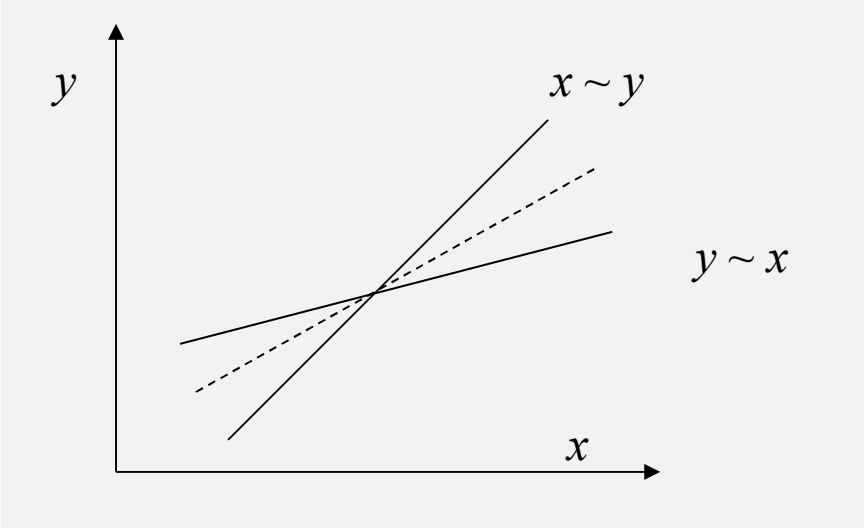


$\hat{b}_{RMA} = \text{sgn}(r)\sqrt{\frac{s_{yy}}{s_{xx}}}$，直线方程为SD Line：$y - \bar{y} = \text{sgn}(r)\sqrt{\frac{s_{yy}}{s_{xx}}}(x - \bar{x})$

$$\frac{y - \bar{y}}{\sqrt{s_{yy}}} = \text{sgn}(r)\left(\frac{x - \bar{x}}{\sqrt{s_{xx}}}\right)$$

Bisector regression (double regression)

平分回归和逆回归之间的夹角



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

$\hat{b}_1, \hat{b}_2$分别是回归和逆回归得到的x的系数的估计。

回归效应

我们仅考察简单模型中的回归效应。假设简单线性模型

$y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2)$, ε与x独立，由命题2:

回归直线

记$\mu_x = E(x), \mu_y = E(y), \sigma_x{}^2 = \mathrm{var}(x) = \Sigma_{xx}, \sigma_y{}^2 = \mathrm{var}(y) = \Sigma_{yy}$，$\rho = \rho_{xy}$，则

$\Sigma_{xy} = \mathrm{cov}(x, y) = \rho\sigma_x\sigma_y$,

(1) $b = \dfrac{\Sigma_{xy}}{\Sigma_{xx}} = \rho\dfrac{\sigma_y}{\sigma_x}$, $a = \mu_y - b\mu_x$,

(3) $\sigma^2 = (1-\rho^2)\sigma_y^2$

回归直线$y = a + bx = \mu_y + \rho\dfrac{\sigma_y}{\sigma_x}(x - \mu_x) \Leftrightarrow$ $\dfrac{y-\mu_y}{\sigma_y} = \rho \times \dfrac{x-\mu_x}{\sigma_x}$

SD线

正相关时的SD线: $y = \mu_y + \dfrac{\sigma_y}{\sigma_x}(x - \mu_x)$ 或 $\dfrac{y-\mu_y}{\sigma_y} = \dfrac{x-\mu_x}{\sigma_x}$,

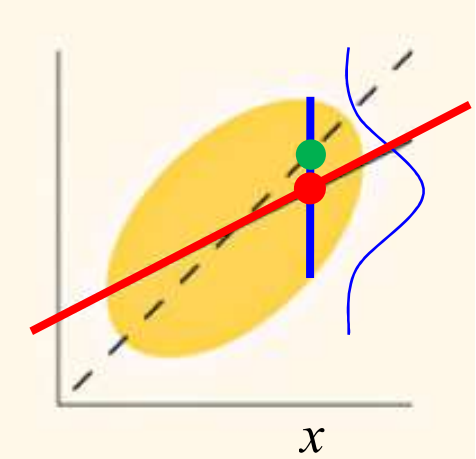
该直线上，x偏离中心一个单位σ_x时，y偏离其中心一个单位σ_y。

回归效应

当$x > \mu_x$时，$y = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) < \mu_y + \frac{\sigma_y}{\sigma_x}(x-\mu_x)$

当$x < \mu_x$时，$y = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) > \mu_y + \frac{\sigma_y}{\sigma_x}(x-\mu_x)$

相比于SD线 $y = \mu_y + \frac{\sigma_y}{\sigma_x}(x-\mu_x)$，回归直线在两端有向中心回归的趋势。



直观解释：

对给定的$x > \mu_x$,假设变量y服从图中所示的正态分布

该正态分布的中心(红点)：$E(y \mid x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$

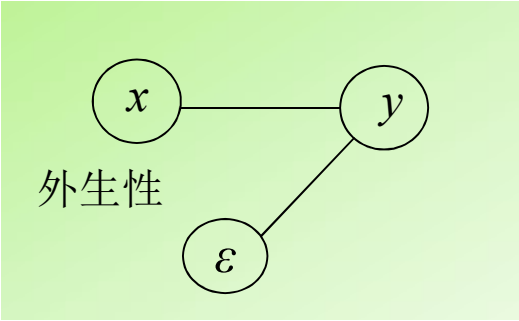
它是蓝色线段的中心，在虚线下方。

附: 工具变量法 (Instrumental Variable method)

工具变量法试图在线性模型中误差与自变量不独立的情况下，求解回归系数的无偏估计, 发现因果 (2021诺贝尔经济奖J. Agrist, G.Imbens, D.Card）。

外生变量

如果线性模型 $y = a + bx + \varepsilon$ 中 x 不是研究对象本身固有而是外界随机赋予的，称x是外生的（exogenous），此时x与ε独立。外生性是推断x和y之间因果关系的关键条件。



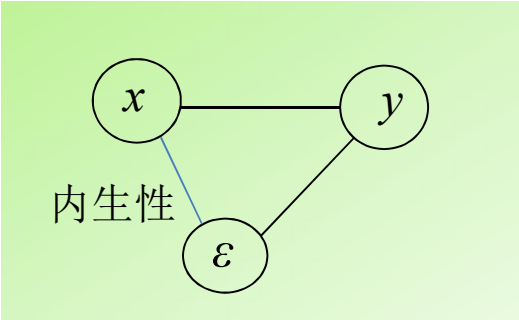
例如，第1讲霍乱案例中，比较两个自来水公司客户的问题可表述成：

$$y = a + bx + \varepsilon$$

其中y为霍乱状态，x为饮用水来源（是否污染）， Snow 医生论证了x是外生变量（随机取值,自然实验）。

内生变量

观察研究中，模型 $y = a + bx + \varepsilon$ 中x是研究对象本身固有的，一般与ε不独立，称x是内生的 (endogenous)，如何推断x和y之间因果关系？



线性模型以及LS方法都在某种意义下实践了 y关于x的去相关化：

$\varepsilon = y - (a + bx)$与$x$不相关

能否实现y与x的去相依化，即发现函数f使得

$\varepsilon = y - f(x)$与x独立？

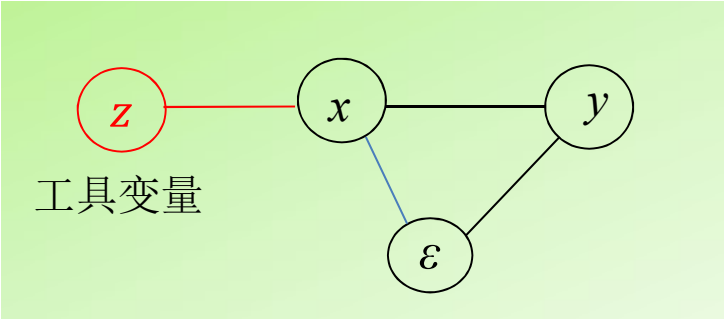
几乎不可能(?)

当x是内生变量时，可以在回归模型中控制所有与x, y都有关系的干扰因素，（非常困难）。计量经济学家发明了工具变量法（Wright, 1928），其关键在于发现恰当的自然试验(natural experiment)，当然，这也是非常困难的。

工具变量

假设存在一个自然试验(natural experiment)产生的外生变量z，即z与ε独立，但与x存在某种程度的相似或相关，称之为工具变量。

工具变量法: 在研究y与x的关系时，以$\hat{x} = f(z)$替代x。



工具变量z满足条件：

(1) z与ε独立；

(2) z与x相关；

与霍乱案例不同的是，工具变量法处理x不是外生的情况，但假设存在与x有关的另外一个外生变量z。

工具变量估计

假设模型$y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2)$，$\varepsilon$和$x$相关。

假设z是一个工具变量，满足条件：(1) z与ε独立；(2) z与x相关；

因为z与ε独立 $\Rightarrow 0 = \mathrm{cov}(\varepsilon, z) = \mathrm{cov}(y - a - bx - \varepsilon, z) = \mathrm{cov}(y, z) - b\,\mathrm{cov}(x, z)$

$\Rightarrow b = \mathrm{cov}(y, z) / \mathrm{cov}(x, z) = \Sigma_{yz} / \Sigma_{xz}$

因此基于样本$(y_i, x_i, z_i), i = 1, \ldots, n$得到矩估计（IVLS）

$$\hat{b}_{IV} = \frac{s_{yz}}{s_{xz}} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

性质：IVLS估计是渐近无偏的

$E(\hat{\mathbf{b}}_{IV} \mid Z) \to \mathbf{b}$ (当$n \to \infty$)

IVLS的两步估计法(2SLS)

1. $\mathrm{lm}(x \sim z) \Rightarrow$ 拟合值$\hat{x} = s_{xz} s_{zz}^{-1} z$
2. $\mathrm{lm}(y \sim \hat{x}) \Rightarrow$ LS估计 $\hat{b}_{IV} = s_{\hat{x}y} s_{\hat{x}\hat{x}}^{-1} = s_{yz} / s_{xz}$

拟合值$\hat{x}$仅与z有关，$x^{\perp} = x - \hat{x}$与z不相关

$y = a + bx + \varepsilon = a + b\hat{x} + (bx^{\perp} + \varepsilon) = a + b\hat{x} + \tilde{\varepsilon}$, $\tilde{\varepsilon} = bx^{\perp} + \varepsilon$与$\hat{x}$不相关

例2. 2021诺贝尔经济奖获得者Agrist 研究了教育程度（x，受教育时长，月）是否与收入（y）存在因果关系(Angrist and Krueger, 1991) 。考虑线性模型

$$y = a + bx + \varepsilon,$$

与收入有关的其它因素（比如能力，家庭因素，努力程度等）无法全部测量或精确测量到，我们将它们放到 ε 中。显然，这些因素与自变量x有关，因此基于上述简单模型无法正确地（无偏）估计x的效应。

Angrist and Krueger 注意到小学入学时所有儿童年龄最大有1年的差距（当年12月达到6岁即可入学)

出生季节	1 (10-12月)	2 (1-3月)	3 (4-6月)	4 (7-9月)
入学年龄	5 ¾ -6	6 -6 ¼	6 ¼ - 6 ½	6 ½ - 6 ¾

而义务教育法规定青少年在16岁生日之前必须在校学习，所以对于16岁生日当天离开学校的那些学生而言，其受教育时长是由其生日 z 决定的，所以生日与x有关，但与ε无关（天然试验）, 所以 z 可作为工具变量

$$y = earnings$$
$$x = years\ of\ education$$
$$z = 出生季节$$

例3. Angrist (1990) 研究了服兵役对生活质量的影响。

$y =$ 生活质量$(life\ quality)$，
$x =$ 服过兵役与否，
$\varepsilon =$ 其它与y有关的因素

志愿征兵制下，入伍与否是个人的选择，因此x是内生的。

1970's 越战期间，美国推行了基于“抽签”的强制征兵制度（draft lottery）。每个19-26岁符合兵役条件的男性被分配了一个随机号码RSN（1-365）。征兵前国家公布一个数字T，小于T的人将被列入强制入伍候选，再进行体检等其它程序（1970，71，72年分别T=195，125，95）。这是一个自然试验，变量 $z = 1_{(RNS<T)}$是工具变量，它与兵役状态x相关，但与研究对象独立。

参考文献：

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*. June, 80:3, pp. 313–36.

Angrist, Joshua D. and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*. November, 106:4, pp. 979–1014.

Angrist, J. D., and Krueger, A. B. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *The Journal of Economic Perspectives* (15:4), pp. 69-85.

其它工具变量的例子(Angrsit and Krueger 2001)

Natural and Randomized Experiments

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*	*Reference*
		1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling-Sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)

工具变量估计：两步估计-观点1

前面我们得到了 IVLS估计，注意到有下述表示，

$$\hat{b}_{IV} = \frac{s_{yz}}{s_{xz}} = \frac{s_{yz}/s_{zz}}{s_{xz}/s_{zz}},$$

IVLS估计可由下述两步LS方法得到：

$$(1)\ lm(x \sim z):\ x = c + dz + \varepsilon^{(1)} \Rightarrow LS估计 \hat{d} = s_{xz}/s_{zz}$$
$$(2)\ lm(y \sim z):\ y = e + fz + \varepsilon^{(2)} \Rightarrow LS估计 \hat{f} = s_{yz}/s_{zz} \qquad \Rightarrow \hat{b}_{IV} = \hat{f}/\hat{d}.$$

工具变量估计：两步估计-观点2

注意到IVLS估计也可表示如下

$$\hat{b}_{IV}=\frac{s_{yz}}{s_{xz}}=\frac{s_{yz}}{s_{zz}\left(s_{xz}/s_{zz}\right)}=\frac{s_{yz}}{s_{zz}\hat{d}}=\frac{\Sigma(\hat{d}\,z_i-\hat{d}\,\bar{z})(y_i-\bar{y})}{\Sigma(\hat{d}\,z_i-\hat{d}\,\bar{z})^2}=\frac{s_{yw}}{s_{ww}}$$

$$\text{其中}\hat{d}=s_{xz}/s_{zz},\quad w_i=\hat{d}\,z_i$$

所以，IVLS估计看作是下述两步LS估计：

$$(1)\ lm(x\sim z):\ x=c+dz+\varepsilon^{(1)}\ \Rightarrow\ LS\text{估计}\hat{d}=s_{xz}/s_{zz}\Rightarrow w_i=\hat{d}z_i$$

$$(2)\ lm(y\sim w):\ y=e+fw+\varepsilon^{(2)}\Rightarrow LS\text{估计}\hat{f}=s_{yw}/s_{ww}=\hat{b}_{IV}$$

这是前面（P20）的两步估计法，该观点反映了工具变量法的本质：
第一步用z把x表示出来，得到的拟合值w只与z有关，w与ε独立。

常见的内生性原因

(1) 丢失变量(Omitted variable,回归方程中没有控制相关变量)

假设正确模型为

$$y = a + bx + cz + \delta, \delta \perp x$$（$\perp$ 代表独立，下同）

其中 z 与 x 相关。如果我们没有测量,或者没有在上述模型中控制z，那么工作模型 $y = a + bx + \varepsilon$中 $\varepsilon = cz + \delta$ 与 x 不独立。

(2) 因果颠倒(reverse causation,响应变量是自变量的原因)

假设正确模型为 $y = a + bx + \varepsilon$,$\varepsilon \perp x$;

但实际操作中工作模型取为: $x = c + dy + \delta$,

从正确模型我们知道 $x = -a/b - y/b - \varepsilon/b$, 所以工作模型中 $\delta = -\varepsilon/b$与y有关。

(3) 自变量带误差模型 (Error in Variable，EV模型)

正确模型：$y = a + bx_0 + \varepsilon_0$，$x_0 \perp \varepsilon_0$. 假设对于$x_0$的测量有误差，即我们只能测量到: $x = x_0 + \delta$,其中$x_0 \perp \delta$.则

$$y = a + b(x - \delta) + \varepsilon_0 = a + bx + (\varepsilon_0 - b\delta) \hat{=} a + bx + \varepsilon$$

显然x与$\varepsilon = \varepsilon_0 - b\delta$相关.