

MR



§0. MR 介绍

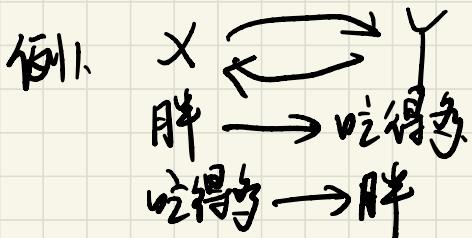
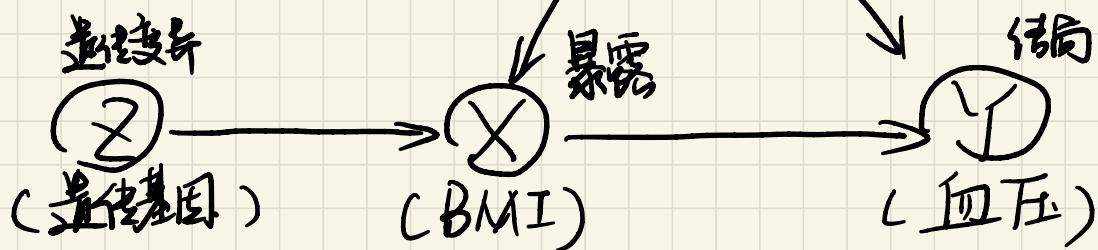
计量经济学

- { 1. 流行病学
- 2. 工具变量 (Instrumental Variable)
- 3. 遗传流行病学

孟德尔随机化

混杂因素, U (蛋白质的表现)

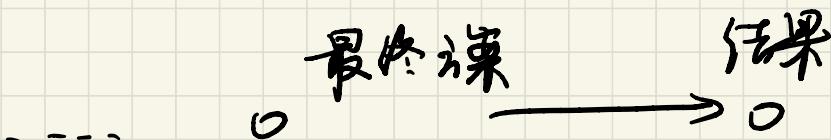
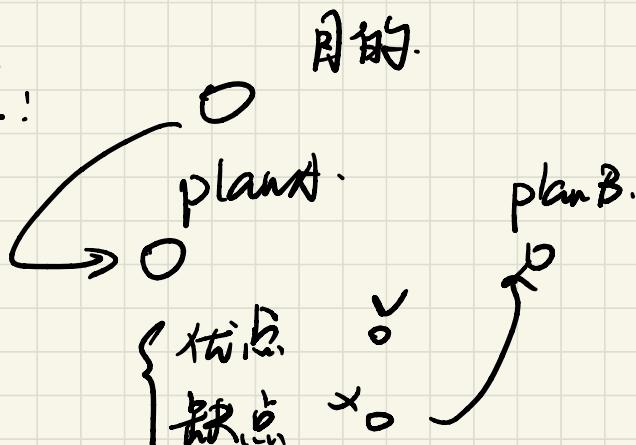
内生性.



孟德尔随机化(工具)

★ 明确两个表现之间的因果关系

步骤：



(为什么最终是MR)

权衡？

优点？

不足

§1. 线性回归和逻辑回归

1-1. 线性回归 (预测问题) $X \rightarrow Y?$

1.1.1 一元线性回归 (模型) (Linear Regression)

买书(本) \rightarrow 价格

1

3

2

3

2

6

J

3

9

n

?

4

?

12?

蛋白质量

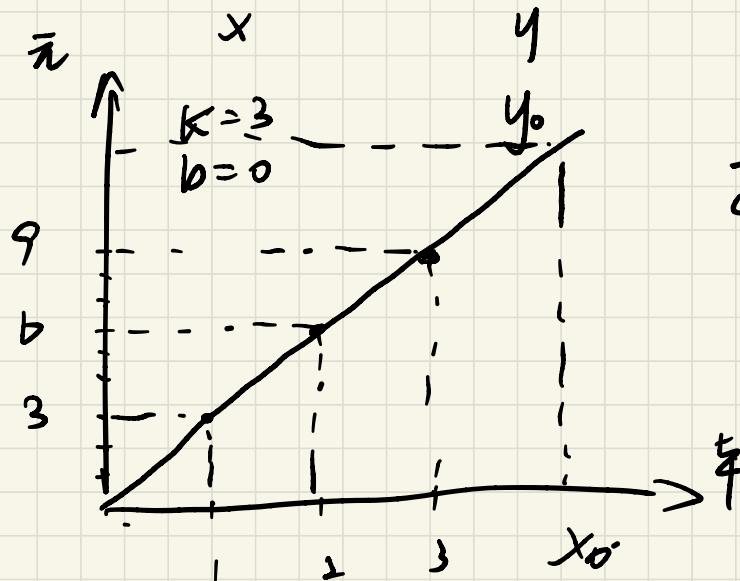
mass/l

x_0

血红

mass/g

y_0 ?



$$y = kx^1 \text{ 函数}$$
$$cy = kx + b$$

x^2

$$y = ax^2 + bx + c$$

x 不是 x^1

向量 y
 \uparrow
 BMI x

$$y = kx + b$$

$$x = x_0, y = ?$$

因果关系 $\xrightarrow{?}$ 预测

简单理解为：假如预测成立
 就有因果关系

例：书 $(1, 3) (2, 6) (3, 9)$

$$y = \underline{3x} + 0, (4, 12) \text{ 满足} \rightarrow \text{有因果关系}$$

\downarrow

$(\underline{4}, \underline{15})$ } 因果关系无
 $(\underline{4}, \underline{12})$
 $(\underline{4}, \underline{13})$

1.1.2 多元线性回归

$$\text{一元 } y = b_0 + b_1 x_1$$

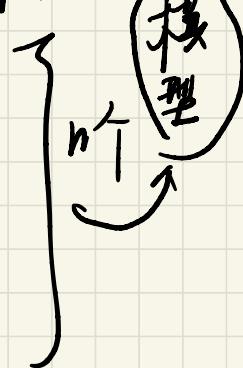
$$\text{多元 } y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

例：降雨的影响因素 $\begin{array}{l} x_1 \text{ 湿度} \\ x_2 \text{ 风力} \\ x_3 \text{ 速度} \end{array} \rightarrow y \text{ 降雨概率}$

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

预测 (训练集)

$$\begin{aligned} x'_1, x'_2, x'_3 &\rightarrow y' \\ x''_1, x''_2, x''_3 &\rightarrow y'' \\ &\vdots \\ x^n_1, x^n_2, x^n_3 &\rightarrow y^n \end{aligned}$$

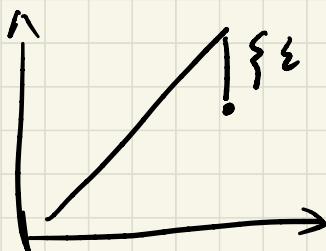


(测试集)

$$x^m_1, x^m_2, x^m_3$$

$$| y^m - y^m |$$

$$= \sum (\text{残差})$$

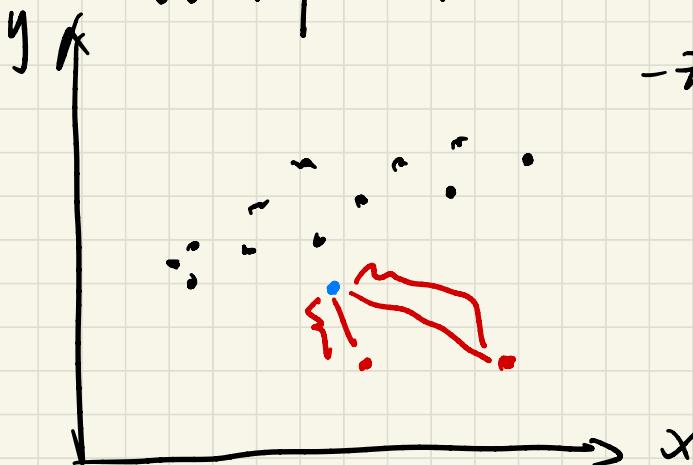


$$\begin{array}{c} \varepsilon = 10 \\ \varepsilon = 10^4 \end{array}$$

$$\begin{array}{c} \checkmark \\ \times \end{array}$$

1.1.3 最小二乘

怎样得到系数 $b_0, b_1, b_2, \dots, b_n$



$$-\bar{x} \quad y = b_0 + b_1 x$$

暴力(Brute-force)



优化空间

$$BF: \begin{cases} b_0=0, b_1=0 \\ b_0=1, b_1=1 \\ \vdots \end{cases}$$

缺点：搜索空间无限大

x

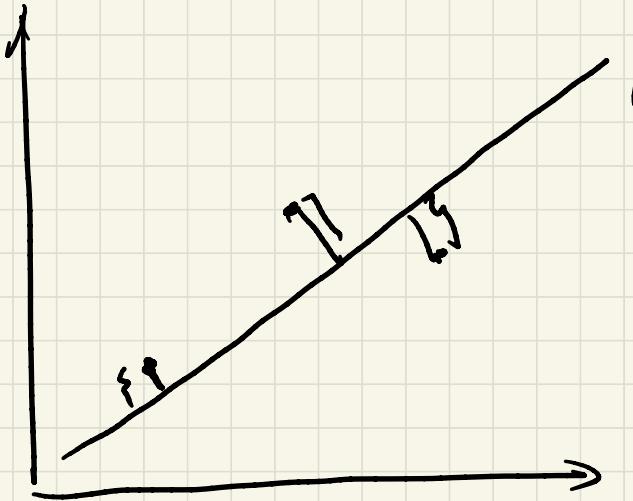


key-sight: 直线可能靠近点 (侧面图)



(欧氏) 距离

二维平面上 $(x_0 - x_1)^2 + (y_0 - y_1)^2$



$(x_1, y_1), \dots, (x_n, y_n)$

对 (x_i, y_i) 到直线 l_0
的距离 d_i

$$\text{最小化} \left(\sum_{i=1}^n d_i \right) \min$$

最后得到的是

满足一条直线

擴展到多元 $(x_1, x_2, x_3, \dots, x_n) \rightarrow y$

兩點 $D(X, X_i)$ 超平面 hypersphere

(重複) $y = b_0 + b_1 x$

多點下.

$$\left\{ \begin{array}{l} X_1 = (x'_1, x'_2, x'_3, \dots, x'_n, y') \\ \vdots \\ X_n = (x^n_1, x^n_2, \dots, x^n_n, y^n) \end{array} \right.$$

$$\left\{ \begin{array}{l} y = b_0 + b_1 x_1 + \dots + b_n x_n \\ \text{滿足 } \min \sum_{i=1}^n D(y, X_i) \end{array} \right.$$

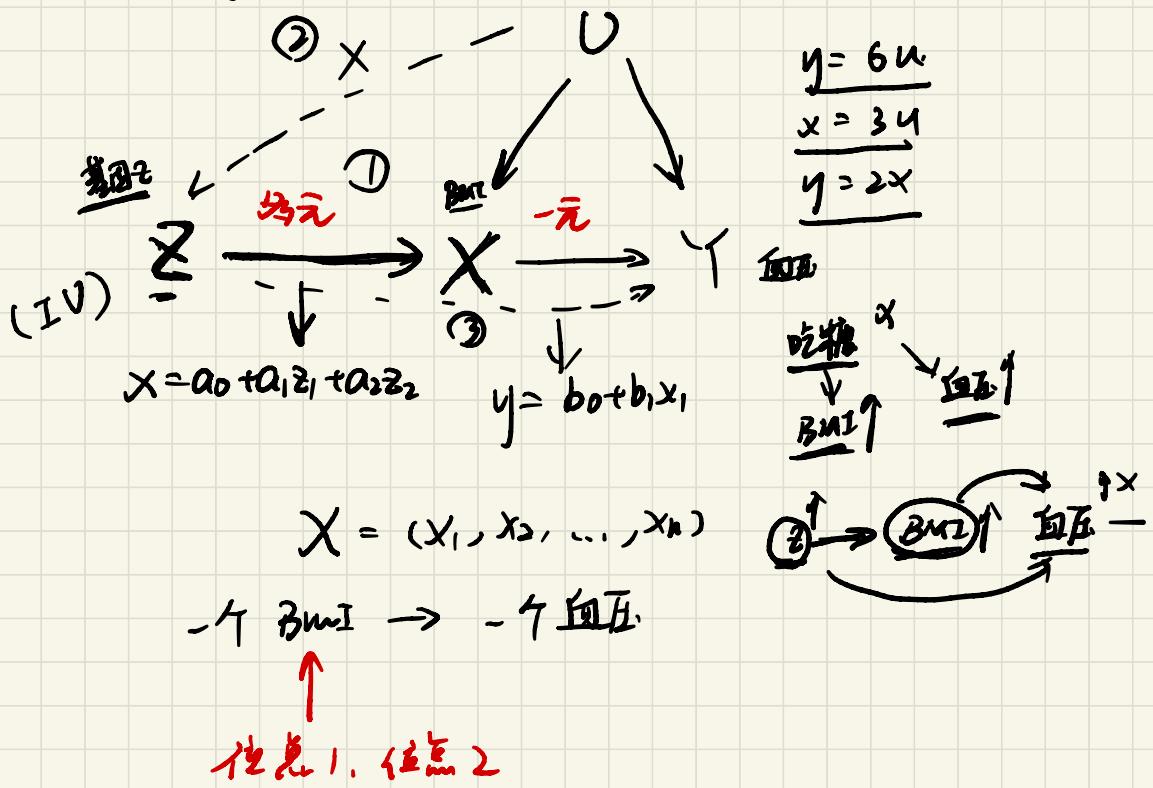
$$\boxed{B = (X^T X)^{-1} X^T C}$$

(大部分)

1.1.4. 回头看MR.

① 有浅性关系 (确定性) $X \rightarrow Y$
 ↓
 ② 有因果关系)

多元线性回归 + 最小二乘.



$X \rightarrow Y$ 线性关系有混杂因素的影响

因此 $X \rightarrow Y$ 的线性关系 $\neq X \rightarrow Y$ 的因果关系

通过选择 $IV(Z)$, 去掉混杂因素

此时, $Z \rightarrow X$, $X \rightarrow Y$ 的线性关系

即为因果关系

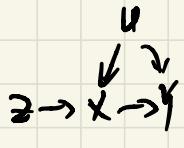
- ① 相关性: 有 $X \rightarrow Y$ 和 $Z \rightarrow X$.
- ② 独立性: 没有 $U \rightarrow Z$
- ③ 排斥性: 没有 $Z \rightarrow Y$

① 验证方法

- ① 目的: 因果 ($X \rightarrow Y$) \xrightarrow{BUT} MR 原理
- ② 线性 $\stackrel{?}{=} \text{因果}$ (U ?) $Z \rightarrow X, X \rightarrow Y$.
- ③ IV² (3大假设), 此时, 线性 = 因果.
- ④ 求线性 (线性回归和最小二乘)

例

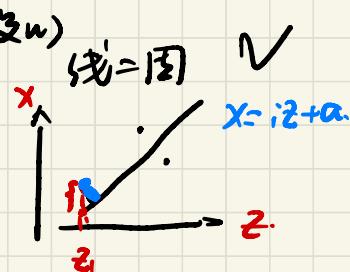
X BMI.	X_1, X_2, \dots	训练集
Y 血压	y_1, y_2	预测集
Z 血糖	z_1, z_2	x'
\underline{Z} 基础代谢	$\underline{z}_1, \underline{z}_2$	y'



① 线 → 因 $y = kx + b$ x 因 z .

② TSLS. i) $x = iz + a$ (线 \rightarrow 因) \checkmark

$iz_1 + a = x_1 \neq x_1$



ii) $\underline{y} = j\underline{x}' + c$

③ 驱动: 已知 \underline{z}'' . $x'' = iz'' + a$

$y'' = jx'' + c$

得到血压(预测值) \underline{y}''

\underline{y}''

(因为预测的值)

$|\underline{y}'' - \underline{y}'| < \varepsilon = 10^{-6}$

∴ 有线性关系

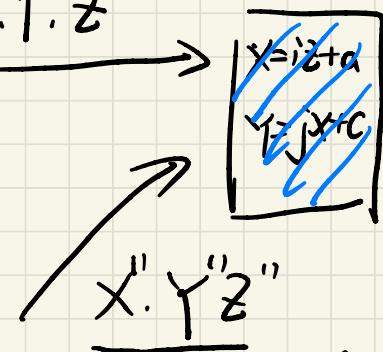
∴ 有因果关系

预测 \underline{y} 真实 \underline{y}

MR (R 实现)

(训练)

X, Y, Z



\rightarrow

\sum
显著水平.

$$\left| \begin{array}{l} |y'' - y| \\ (y'' - y)^2 \\ \vdots \\ | \end{array} \right|^3$$