

CS7323 从数据学习因果关系

第二次作业

1. 给定随机变量 X 和 Y ，他们的相关系数定义如下：

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

有条件期望如下：

$$E(Y|X) = E(Y) + \rho_{XY} * \sigma_Y \frac{X - E(X)}{\sigma_X}.$$

(x_i, y_i) 是随机变量 X, Y 的一组取值，并总有如下线性关系：

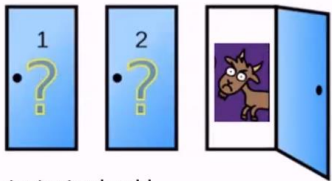
$$y_i = \beta_1 x_i + \beta_0 + \epsilon_i.$$

问：为什么 X 和 Y 的相关系数 $\rho_{XY} = 0$ 时， $\beta_1 = 0$ 。

2. 三门问题如图一所示。请用图一中右下角的因果图进行解释：为什么参赛者选择换门之后，获胜概率将提升？同时结合该例子，使用 Rubin 的 potential outcome framework 解释伯克森悖论（Berkson's paradox）。

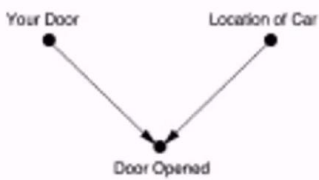
MONTY HALL PROBLEM

in September 1990, a writer named Marilyn vos Savant wrote a column in Parade magazine:
“you’re given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, whoknows what’s behind the doors, opens another door, say #3, which has a goat. He says to you, ‘Do you want to pick door #2?’ Is it to your advantage to switch your choice of doors?”



vos Savant argued that contestants should switch doors.

Door 1	Door 2	Door 3	Outcome If You Switch	Outcome If You Stay
Auto	Goat	Goat	Lose	Win
Goat	Auto	Goat	Win	Lose
Goat	Goat	Auto	Win	Lose



Causal diagram for *Let's Make a Deal*

图一. 三门问题。

3. （选做）在核酸检测中，有时会出现假阴性和假阳性的情况。（1）请用统计的角度说明为什么会出现假阴性和假阳性。（2）说明为什么将假阳性的比例控制得较低，其代价是更高的假阴性。为什么低假阳性和低假阴性二者是矛盾的？