Marketing Analytics

消费者分析

R 软件

我们课程将用到 R 语言。请按照下面的连接下载并安装 R 和 RStudio 两款软件。

- 请在 这里下载 R
- 请在 这里下载 RStudio

注意安装路径必须为纯英文，否则某些功能无法正常使用！

课程网站

你可以在这里查到我们全部的课程资料：

https://ximarketing.github.io/class/ConsumerAnalytics/index.html

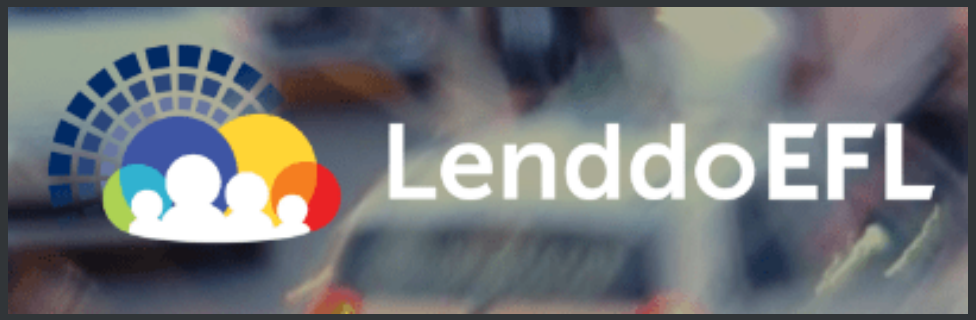
Question:

When talking about marketing analytics,
what first comes to your mind?

当我们说消费者分析的时候，你最先想到的是什么？

Lenddo, a Singaporean based start-up, helps financial institutions collect your social network data, but why?

Lenddo，一家总部位于新加坡的初创公司，帮助金融机构收集你的社交网络数据，但这有什么用呢?



Wal-Mart's Shopycat-Gift Recommendation

Wal-mart's Shopycat app will help you buy the ideal gift for your friend during the holiday buying rush. Walmart's Shopycat recommends gifts for friends based on the social data extracted from their Facebook profiles. The app also provides links to the Walmart products so that users can easily purchase the product.

沃尔玛的 Shopycat 智能礼品推荐

沃尔玛的 Shopycat 是一款应用，能够直接从你的 Facebook 账号中获取和分析你的社交信息，并且根据你的社交信息向你推荐不同的节日礼物。这个APP甚至会直接给你提供产品链接帮助你更方便的完成购物。

Dentsu's Data Driven Advertising

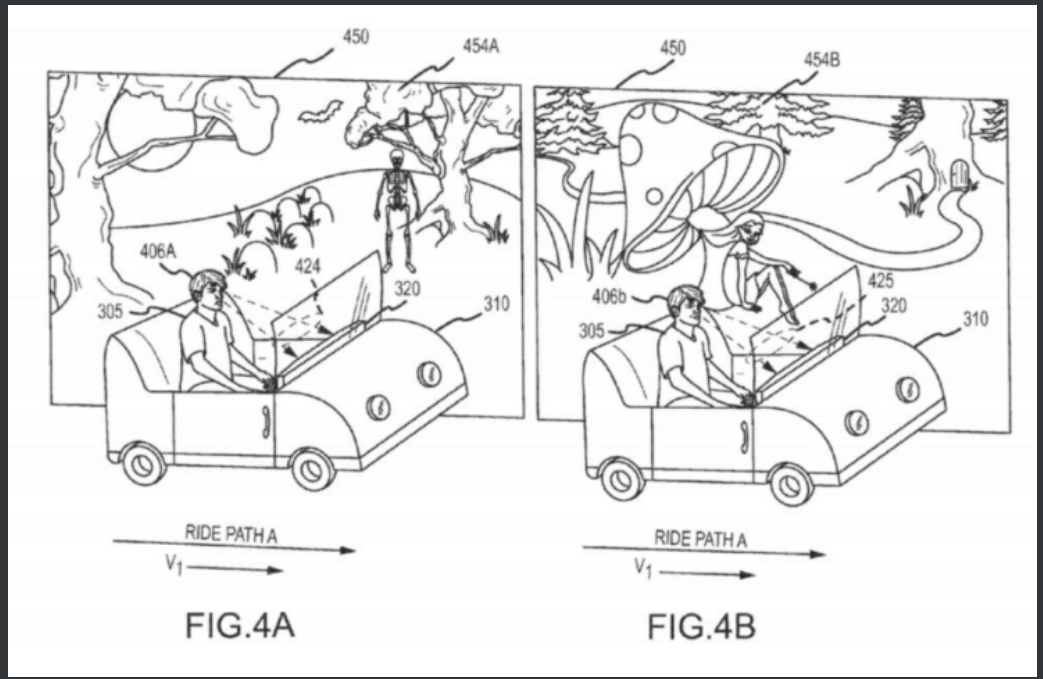
Dentsu is one of the largest and most influential advertising agencies in the world, founded in 1901 and headquartered in Tokyo, Japan. In 2019, Dentsu partnered with Cloudian, Intel, and other companies on the "DeepAd project." This initiative aimed to track and capture car images on Tokyo roads and deliver dynamic advertisements on digital billboards as a result.

电通的数据驱动广告

电通是世界上最大和最有影响力的广告公司之一，成立于1901年，总部位于日本东京。2019年，电通与Cloudian，英特尔等公司合作进行了“DeepAd项目”。该倡议旨在在东京的道路上跟踪和捕捉汽车图像，并在数字广告牌上投放动态广告。

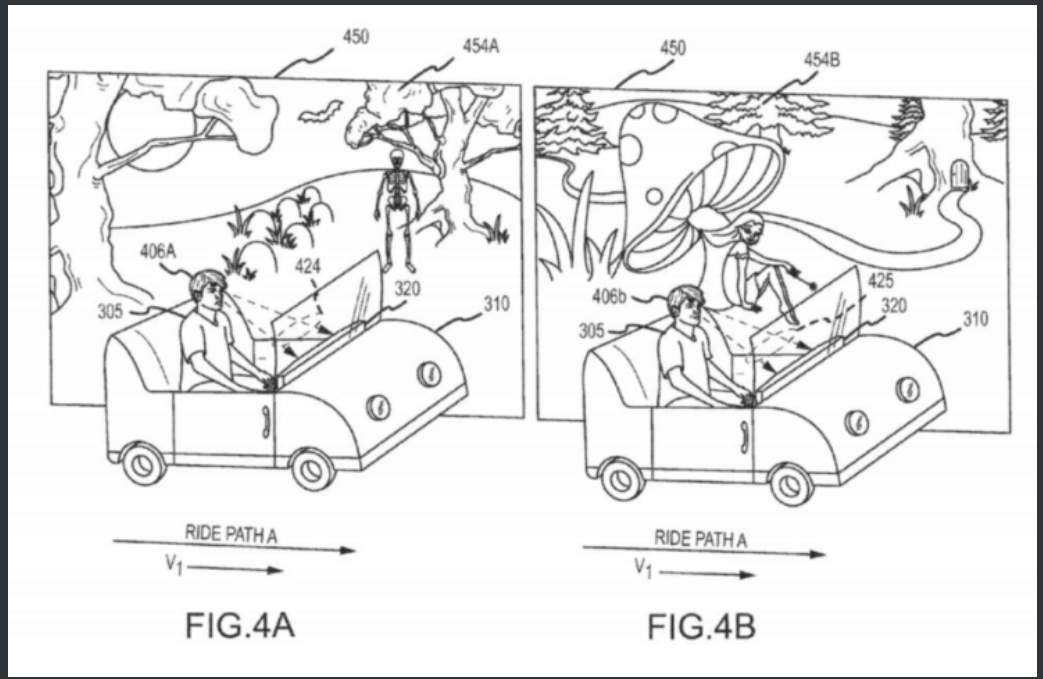
Disney's Data Driven User Experience

In 2017, Disney filed a patent and here is a figure from the patent. Do you know what Disney is doing?



迪士尼的数据驱动用户体验

2017年，迪士尼申请了一项专利，并这是专利中的一个图示。您知道迪士尼在做什么吗?




ALLTID
24
ÖPPET
Mighty
McMuffin
& Co

“顺手买一件”


上午9:20
确认订单
肌肉小王子旗舰店
【22袋肉】肌肉小王子鸡胸肉健身代餐即食鸡肉速食轻食品低脂零食
¥109.00
x1
可享先用后付
购买数量
1
服务
此商品性质不支持7天退货
配送方式
普通配送
快递 免邮
配送时间
现在下单，预计7月12日送达
隔日达
店铺优惠
省30元:组合优惠
- ¥ 30.00
订单备注
选填,请先和商家协商一致
共1件 小计: ¥79.00
顺手买一件
肌肉小王子豆浆粉无蔗糖添加七彩豆浆独立袋装代餐早餐冲饮速溶
口味:七彩豆浆粉1盒7袋*28克【第
现价 ¥9.90 价格 ¥49
共1件，合计: ¥79.00
提交订单


配送方式
普通配送
快递 免邮
运费险
退换货可赔付10元
¥ 2.98
请关注保险公司信息，并阅读《保险条款和重要说明》，本模块由蚂蚁保保险代理有限公司管理。
店铺优惠
省16元:组合优惠
- ¥ 16.00
开具发票
本次不开具发票
订单备注
选填,请先和商家协商一致
共2件 小计: ¥147.90
顺手买一件
全能猫砂
uah有哈猫砂混合猫砂膨润土豆腐猫砂无香无尘抑菌除臭7L
香味:全能混合猫砂*1包;
现价 ¥18.80 价格 ¥39.90

Instructor:

Xi Li (xili@hku.hk)

Professor of Marketing, Innovation and Information Management

Director, Asia Case Research Centre

Associate Director, Institute of Digital Economy and Innovation

Ph.D., Management, University of Toronto

M.Phil., Operations Research, HKUST

B.E., Computer Science, Tsinghua University

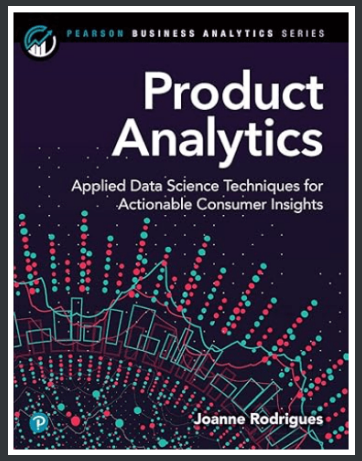
授课教师：

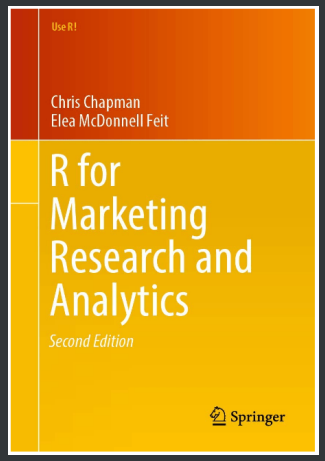
李曦 (xili@hku.hk)
香港大学市场学教授，创新与信息管理学教授
亚洲案例研究中心主任
数字经济与创新中心副主任

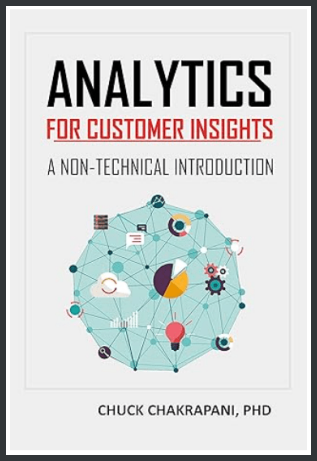
多伦多大学管理学博士
香港科技大学工业工程与物流管理哲学硕士
清华大学计算机科学与技术学士

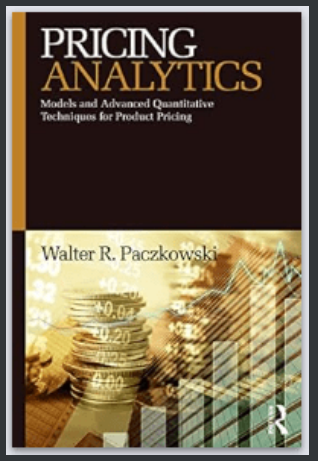
We do not have any textbooks. Our content is partly inspired by the following books:

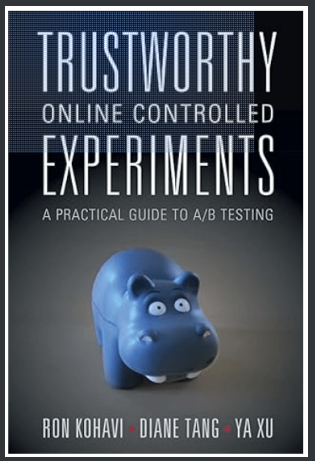
这是我们课程的一些参考读物 [供有兴趣的同学参考]











Class Overview

课程概览

What is the brand of the HKU president's car?

猜一猜: 香港大学校长开什么车?

Discrete Choice Modeling 离散选择模型

We are familiar with linear regression, which allows you to use your independent variable X to predict your dependent variable Y, i.e., $Y = \alpha + \beta X$. In linear regression, the output, Y, is a real number, e.g., $Y = 1.319$.

线性回归中，我们的方程是$Y = \alpha + \beta X$，其中X是自变量，Y是因变量。这里，Y必须是一个实数，例如 $Y = 1.319$.

Discrete Choice Modeling

But when consumer makes choices, the choices are often discrete: It can be purchase or nonpurchase, it can be the brand that you choose (CCB vs. BOC vs. ICBC vs. ABC).

Linear regression does not work here! We can use discrete choice models to model consumers' choices.

离散选择模型

但现实中，消费的选择往往是离散的。在买东西时，消费者选择可能是买或者不买。消费者的选择也可能是某一个品牌(建设银行 vs 中国银行 vs 工商银行 vs 农业银行)。

线性回归对于这种问题无能为力：我们的因变量是一个个离散的数据。

Discrete Choice Modeling

We introduce three discrete choice models:

- Logistic regression: The dependent variable is either 1 or 0 (e.g., purchase vs. non-purchase).
- Multinomial logit model (MNL): The choice depends on the choice maker's characteristics (e.g., age, gender).
- Conditional logit model: The choice depends on the alternatives' characteristics.

离散选择模型

我们介绍三种离散选择模型:

- 逻辑回归: 被解释变量可以取两个值，即1或0 (例如购买和非购买两个选项).
- Multinomial logit model (MNL): 多个选择，选择本身是基于消费者的个人特征 (如年龄，性别，工作信息).
- Conditional logit model: 多个选择，选择本身是基于选项的特征(例如价格，品牌，质量等)

离散选择模型

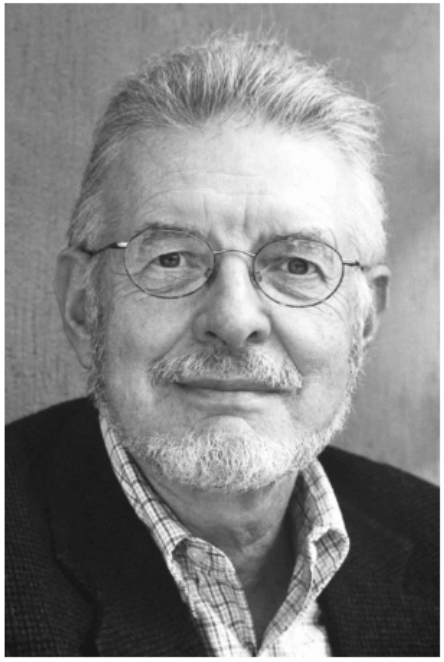


Photo from the Nobel Foundation archive.

Daniel L. McFadden
The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2000

Born: 29 July 1937, Raleigh, NC, USA

Affiliation at the time of the award: University of California, Berkeley, CA, USA

Prize motivation: "for his development of theory and methods for analyzing discrete choice"

Prize share: 1/2

Causality and AB Testing

With a bit knowledge of statistics, we can often find that two things are correlated. But this does not necessarily mean one thing has caused another. A famous example is the correlation between chocolate consumption and winning Nobel prize.

因果关系与AB测试

通过使用基本的数据分析，我们很容易发现两个事情是相关的。但是相关并不意味着一件事导致了另一件事。一个著名的例子是巧克力消费与诺贝尔奖之间的相关性。

因果关系与AB测试

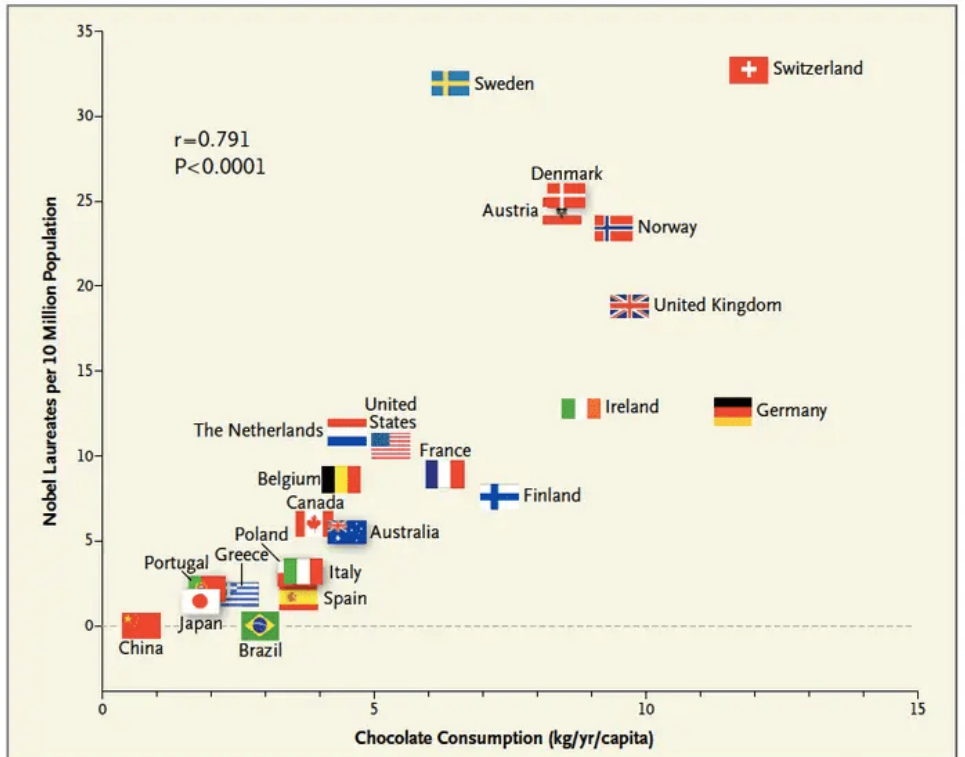
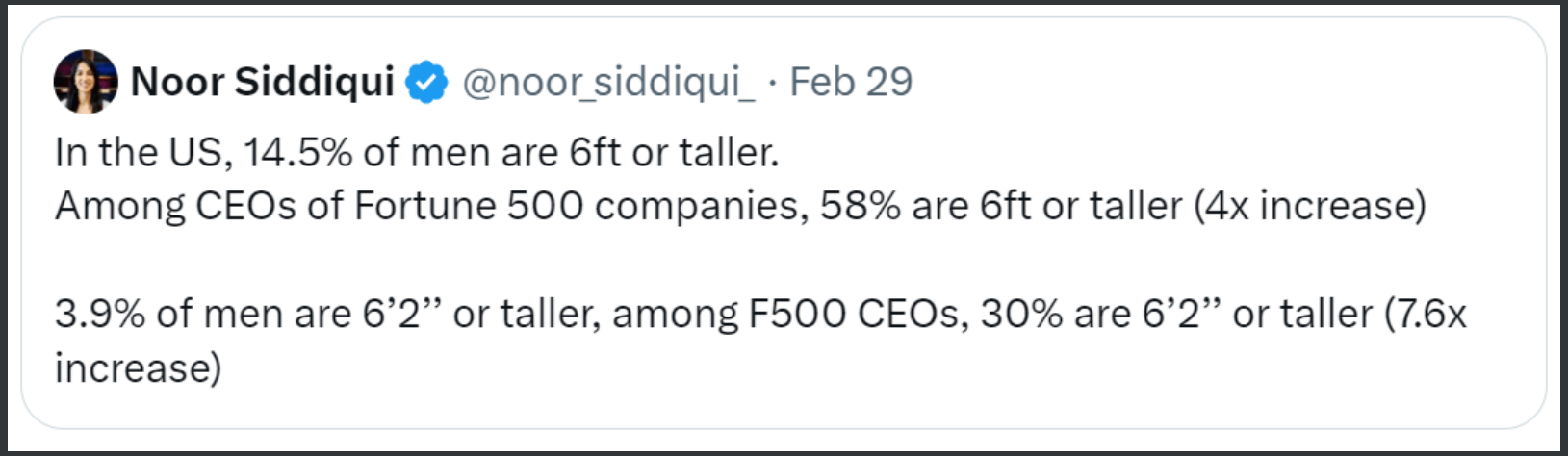


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

因果关系与AB测试



《财富》世界500强的CEO中，有58%的身高超过1.82米，30%的超过1.88米。身高越高越容易当CEO?

Causality and AB Testing

How to identify causality? We can establish AB tests which are common used in big tech firms nowadays.

We will learn

- How to compare clicks/purchases in AB tests;
- How to compare profits/revenues in AB tests;
- How to compare distributions in AB tests.

因果关系与AB测试

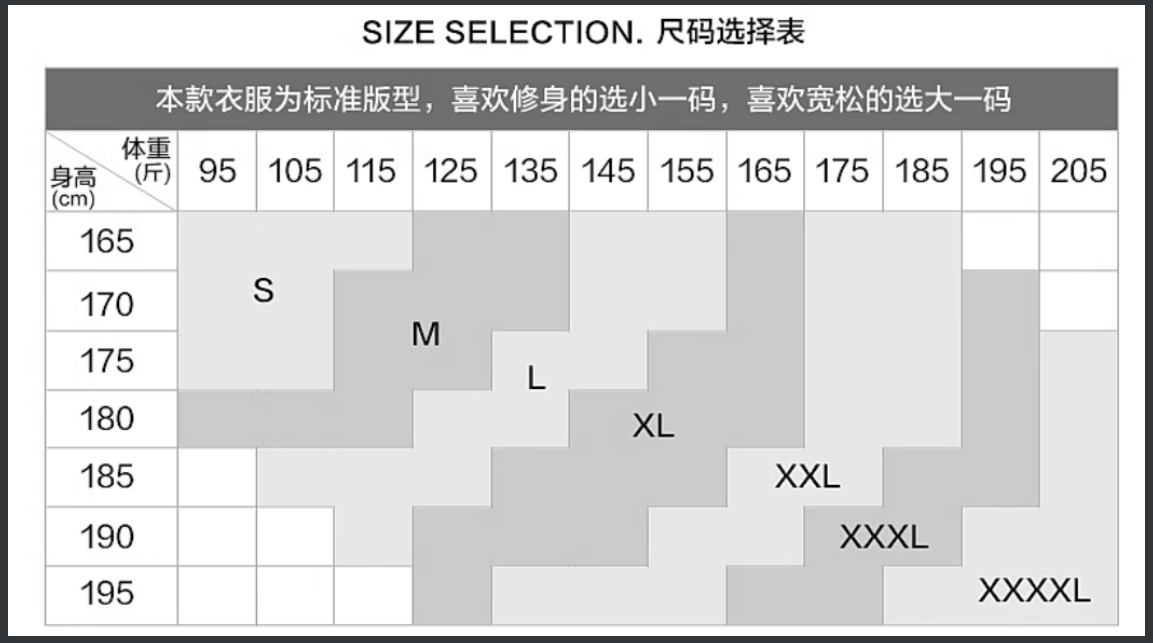
如何建立因果关系？ 一个常见的方法是使用 AB 测试，这一方法在科技企业越来越流行。

我们将学到:

- 怎么在 AB 测试中分析用户点击的不同。
- 怎么在 AB 测试中分析利润或者收入的不同。
- 怎么在 AB 测试中分析统计分布的不同。

Segmentation

How to determine sizes for your clothing?



消费者细分

服饰商应该如何消费者设计尺码表?

SIZE SELECTION. 尺码选择表

本款衣服为标准版型，喜欢修身的选小一码，喜欢宽松的选大一码

身高(cm) \ 体重(斤)	95	105	115	125	135	145	155	165	175	185	195	205
165	S	S	S	M	M							
170	S	S	S	M	M							
175				M	L	L	L					
180						XL	XL	XL				
185								XXL	XXL	XXL		
190									XXXL	XXXL	XXXL	
195											XXXXL	XXXXL

Segmentation

How to choose locations for your store?

消费者细分

建设银行是如何为网点选址的?

Recommendation System

Will Alice like Movie 3?

	Movie 1	**Movie 2**	**Movie 3**	**Movie 4**
Alice	4	4		1
Bob		2	2	3
Carol	1	5	3	
Dennis	3		4	1
Emma	5	2	1	4
Flora	3	1		5

推荐系统

Alice 会喜欢 3 号电影吗?

	Movie 1	**Movie 2**	**Movie 3**	**Movie 4**
Alice	4	4		1
Bob		2	2	3
Carol	1	5	3	
Dennis	3		4	1
Emma	5	2	1	4
Flora	3	1		5

Market Basket Analysis

Which products do consumers tend to purchase together?

When they buy beers, should we also recommend them some potato chips as well?

购物篮分析

哪些产品总是同时被购买?

当消费者买啤酒的时候，我们是不是应该建议他们买点薯条?

他们买了你推荐的基金之后，还应该推荐点什么?

Pricing

How to set price for this new electric surfboard?

定价

这个新的电动滑板应该卖多少钱?

Schedule 课程计划

1. Introduction and R Basics 课程简介与R语言基础
2. Discrete Choice Models 离散选择模型
3. Causality and AB Testing 相关性和AB测试
4. Segmentation 消费者细分
5. Market Basket Analysis and Pricing Analytics
   购物篮分析和定价
6. Recommendation Systems 推荐系统

Question	Method
How do consumers make choices among different alternatives?	discrete choice models
Does something really cause another? Which of the strategies work best?	A/B testing
How do I segment the market? How many types of consumers do I have?	Segmentation
Which products should be recommended to my consumers?	Recommendation system
What are the underlying relationship between different products? Which products do consumers buy together?	Market basket analysis
How much will buyers pay for my product? What is my optimal price?	Price analytics

问题	方法
消费者是如何在不同产品中进行选择的?	离散分析
这些因素到底有没有相互影响? 哪个设计最有效?	AB 测试
我应该如何将消费者分类? 我的消费者有哪几种?	消费者细分
我应该推荐什么给消费者?	推荐系统
不同的产品之间有哪些联系? 哪些产品往往被一起购买?	购物篮分析
消费者愿意花多少钱买我的产品? 我应该卖多少钱?	定价分析

From Physics to the Science of Marketing

从物理学到市场科学

The R Software

Our class uses R for teaching. Please install R and RStuido on your laptop and bring it with you for the next class. You can

- Download R here
- Download RStudio here

Note that your installation path should not contain any non-English letters, otherwise you will be unable to use some functions.

安装路径必须为纯英文，否则某些功能无法正常使用！

R 软件

我们课程将用到 R 语言。请按照下面的连接下载并安装 R 和 RStudio 两款软件。

- 请在 这里下载 R
- 请在 这里下载 RStudio

注意安装路径必须为纯英文，否则某些功能无法正常使用！

为什么选择 R

- R 包含丰富的统计库，让数据分析变得简单。
- R 拥有一个巨大的用户社区。
- 作为一门程序语言，R 对学习者比较友好，基本操作比较简单。
- R 是完全免费的。

A Review of Regression
回顾：线性回归

Regression

Imagine that you want to examine how income changes with Age, then you can consider the following regression:

$$\text{Income}_i = \alpha + \beta \times \text{Age}_i$$

where α and β are parameters to be estimated.

回归分析

假设我们想知道收入是如何随着年龄变化的，我们可以考虑下面简单的线性回归：

$$\text{Income}_i = \alpha + \beta \times \text{Age}_i$$

其中 α 和 β 是未知的参数。

Regression 回归

```r
file = "https://ximarketing.github.io/class/teachingfiles/r-exercise.txt"
mydata <- read.table(file, header = TRUE)
result <- lm(Income ~ Age, data = mydata)
summary(result)
```

Regression 回归

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -74588      10950  -6.812 5.68e-10 ***
Age             4097        332  12.341  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\text{Income}_i = -74,588 + 4,097 \times \text{Age}_i$$

Regression 回归

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -74588      10950  -6.812 5.68e-10 ***
Age             4097        332  12.341  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stars denote statistical significance. When there are three stars, for example, it means the significance is smaller than 0.1%. Usually, when the significance is smaller than 5% (at least one star here), we say the result is statistically significant.

Regression 回归

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -74588      10950  -6.812 5.68e-10 ***
Age             4097        332  12.341  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

星号表示统计显著性。例如，当有三个星号时，意味着显著性小于0.1%。通常情况下，当显著性小于5%（至少有一个星号时），我们称结果具有统计学显著性。

Fixed Effects

Consider another regression: You want to analyze how the deposit of a consumer changes with his or her job (professional, self-employed, manufacturing etc.) However, job is not a number, how can you run a regression?

固定效应

考虑另一种回归分析：你想要分析用户的存款如何随着其职业（如专业人士、自雇、制造业等）而变化。然而，职业并不是一个数字，如何进行回归分析呢?

Fixed Effects

```r
data = read.csv("https://ximarketing.github.io/data/fixed_effects.csv")
head(data)
result = lm(Deposit ~ Age + factor(Job), data = data)
summary(result)
```

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 64380.22     854.43   75.35   <2e-16 ***
Age                           322.25      15.75   20.45   <2e-16 ***
factor(Job)Professional     27754.26     627.44   44.23   <2e-16 ***
factor(Job)Self-employed   -12077.59     631.16  -19.14   <2e-16 ***
```

Fixed Effects

The regression output suggests that

$$\text{Deposit} = 64380 + 322 \times \text{Age} + \begin{cases} 0 & \text{if job is manufacuring,} \\ 27754 & \text{if job is professional,} \\ -12077 & \text{if job is self-employed.} \end{cases}$$

Here, manufacturing is treated as a benchmark and we compare other jobs against this benchmark.

固定效应

上述回归的结果可以这样解释：

$$\text{Deposit} = 64380 + 322 \times \text{Age} + \begin{cases} 0 & \text{if job is manufacuring,} \\ 27754 & \text{if job is professional,} \\ -12077 & \text{if job is self-employed.} \end{cases}$$

这里，我们把制造业作为一个基准职业，把其他职业同这一基准进行比较。

Why do we set a benchmark?

Why manufacturing is set as a benchmark? This is because R adopts alphabetical order, and “manufacturing” is before “professional” and “self-employed.” However, you can change your benchmark as well:

```
data$Job = relevel(factor(data$Job), ref = "Professional")
result = lm(Deposit ~ Age + factor(Job), data = data)
summary(result)
```

为什么需要一个基准?

那为什么选择制造业为基准职业呢? 这是因为 R 语言采用字母顺序，而制造业 “manufacturing” 排在自雇 “self-employed” 和职业人士 “professional”的前面。当然，你可以随时更换你的基准:

```
data$Job = relevel(factor(data$Job), ref = "Professional")
result = lm(Deposit ~ Age + factor(Job), data = data)
summary(result)
```

课程成绩

- 考勤成绩：25分，包含按时上线学习和看视频回放；缺勤一次扣2.5分，即上课5次为22.5分，以此类推
- 平时表现：25分，可以通过上课期间的互动或者课后通过调查问卷思考题得分；10次或以上为满分，每少1次扣1分，即9次为24分，以此类推
- 开卷考试：50分，为50道单项选择题，每道题1分

课后讨论问题：

我们今天介绍了消费者分析的一些方法。你能在你的工作中用到这些方法吗？它们可以解决哪些问题？