

# 论文课 LESSON4

科研论文的数据的采集和分析 (二)

( 刘佳妮, [joeyliu1997@163.com](mailto:joeyliu1997@163.com) )

# 论文课的计划

8-2: 科研论文写作的基本规范 ( 科研论文的结构 )

8-6: 如何选择研究课题、写好文献综述 ?

8-9: 科研论文数据的采集和分析 ( 一 )

**8-15: 科研论文数据的采集和分析 ( 二 )**



- 回归分析的基本原理与实操

- 如何利用回归模型来刻画解释变量对于被解释变量的意义?

- 回归模型背后的原理是什么? 如何判断回归模型的好坏?

- 如何实现每一步操作? 如何解读回归结果?

8-20: 如何写好学术论文的每个部分、论文的投稿、修改和发表

# 上节课的内容

## 8-9: 科研论文数据的采集和分析（一）

### - 实证分析的理论知识：

- 什么是实证分析？使用数据对现实世界的一般问题展开定量分析的研究方法
- 从使用的数据类型出发，实证分析可以分成哪些类？
  - 基于实验数据的实证研究
  - 基于观测数据的实证研究（本课程的重点）
- 实证研究的基本范式：被解释变量、解释变量、主效应
- 总体与样本的关系；变量与数据的关系：统计数据、统计变量、理论变量

### - 优质数据集有哪些特征？如何判断我们想要的数据是否可行、是否易得？

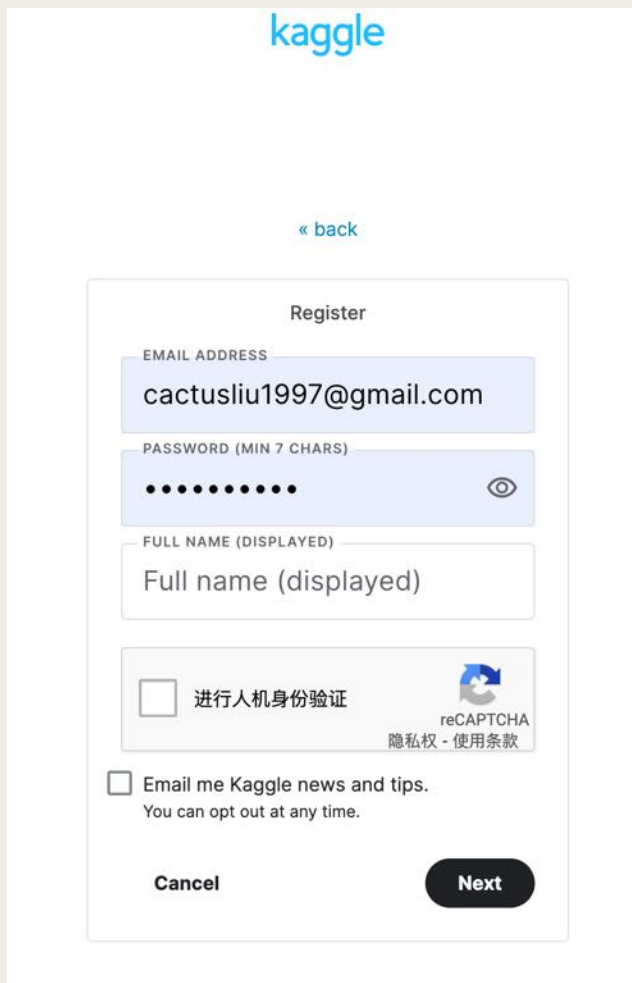
- 可获得、信息丰富、与个体有关、有听众感兴趣的指标、独特可以带来新鲜洞见

### - 数据采集：

- 学会爬虫，是开展科研工作的必经之路吗？
- 我们可以通过哪些方式、按照何种流程，来完成研究数据的采集？

# 上节课的作业

1. Kaggle的注册问题：可以正常浏览数据，但无法注册账户、下载数据集

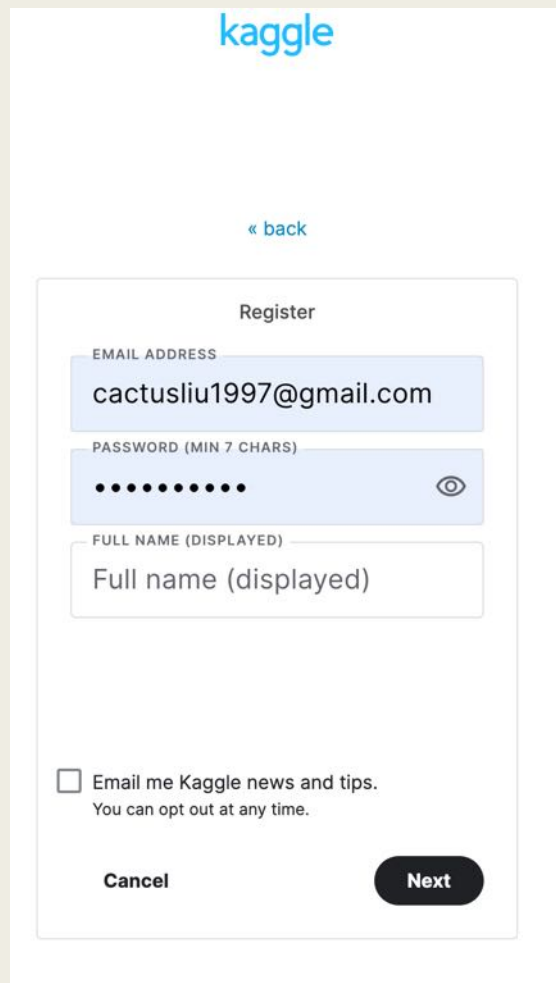


The image shows the Kaggle registration form in Chinese. At the top, the Kaggle logo is displayed in blue. Below it is a blue link labeled « back ». The main form is titled "Register" and contains the following fields:

- EMAIL ADDRESS:** A text input field containing "cactusliu1997@gmail.com".
- PASSWORD (MIN 7 CHARS):** A password input field with ten black dots and an eye icon to the right.
- FULL NAME (DISPLAYED):** A text input field containing "Full name (displayed)".

Below the form fields, there is a checkbox labeled "进行人机身份验证" (Perform human identity verification) next to the reCAPTCHA logo and the text "reCAPTCHA 隐私权 - 使用条款".

At the bottom, there is another checkbox labeled "Email me Kaggle news and tips. You can opt out at any time." and two buttons: "Cancel" and "Next".



The image shows the Kaggle registration form in English. At the top, the Kaggle logo is displayed in blue. Below it is a blue link labeled « back ». The main form is titled "Register" and contains the following fields:

- EMAIL ADDRESS:** A text input field containing "cactusliu1997@gmail.com".
- PASSWORD (MIN 7 CHARS):** A password input field with ten black dots and an eye icon to the right.
- FULL NAME (DISPLAYED):** A text input field containing "Full name (displayed)".

Below the form fields, there is a checkbox labeled "Email me Kaggle news and tips. You can opt out at any time." and two buttons: "Cancel" and "Next".

# 回归分析：引入

- 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？
- 网飞公司：
  - 当今世界流媒体平台中的佼佼者：利用大数据，从DVD光盘邮寄租借服务商，一跃成为在全球范围内拥有2.04亿订阅用户、市值最高的传媒公司



Q：请大家想一想，如果你是网飞的数据科学家，你会想要收集、了解用户的哪些使用数据，来了解他们的偏好、改进网飞的算法，吸引更多用户持续付费？

# 回归分析：引入

- 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？
- 网飞公司：
  - 网飞收集哪些数据？
    - 用户观看过的内容、搜索记录、评分、地理位置....
    - 用户每次观看节目的日期、时间、设备....
    - 用户习惯的观看方式：一天看一集还是一次性疯狂看整季....
    - 用户观看时何时暂停？暂停了之后是否继续看？....
  - 网飞怎么利用这些收集的数据辅助营销决策？

# 回归分析：引入

- 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？
- 网飞公司：
  - 网飞收集哪些数据？
    - 用户观看过的内容、搜索记录、评分、地理位置....
    - 用户每次观看节目的日期、时间、设备....
    - 用户习惯的观看方式：一天看一集还是一次性疯狂看整季....
    - 用户观看时何时暂停？暂停了之后是否继续看？....
  - 网飞怎么利用这些收集的数据辅助营销决策？
    - 更精准的推荐：利用数据分析来制定推荐算法，根据用户的需求和偏好，给他们推送感兴趣内容：Netflix的推荐系统为订阅者贡献了超过80%的流媒体视频内容，产生了高达10亿美元的收入

# 回归分析：引入

- 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？
- 网飞公司：
  - 网飞收集哪些数据？
    - 用户观看过的内容、搜索记录、评分、地理位置....
    - 用户每次观看节目的日期、时间、设备....
    - 用户习惯的观看方式：一天看一集还是一次性疯狂看整季....
    - 用户观看时何时暂停？暂停了之后是否继续看？....
  - 网飞怎么利用这些收集的数据辅助营销决策？
    - 更精准的推荐：利用数据分析来制定推荐算法，根据用户的需求和偏好，给他们推送感兴趣内容：Netflix的推荐系统为订阅者贡献了超过80%的流媒体视频内容，产生了高达10亿美元的收入
    - 更精准的生产！



# 回归分析：引入

## ■ 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？

### ■ 网飞公司：

#### - 《纸牌屋》

#### ■ BBC版本的改编：1990.12

《纸牌屋》（英语：House of Cards）是1990年的政治惊悚片电视连续剧，一共有四集。故事时空设定为英国首相玛格丽特·撒切尔夫人的请辞后。本剧于11月18日至12月9日期间，由英国广播公司播出，旋即广受好评。

本剧乃由安德鲁·戴维斯改编保守党幕僚长迈克尔·多布斯的同名小说，内维尔·泰勒也在1996年为英国广播公司国际频道对多布斯的小说进行戏剧化。本剧有两部电视连续集－《戏王者》和《最后一击》。本剧的开场和结束的主题曲音乐是“弗朗西斯·厄克特的进行曲”<sup>[1]</sup>。



#### - 网飞通过分析喜欢《纸牌屋》BBC原版观众及更多用户的观影习惯，发现：

- **主演和导演很重要：**喜欢BBC版的观众会经常看大卫·芬奇拍的电影、是奥斯卡影帝凯文·史派西的忠实影迷
- **剧情很重要：**喜欢反派人物，对政治、权利、阴谋有着极强的好奇心和探求欲。尤其是对于公众领导者，人们除了希望他们品德高尚、行为端正之外，还会相信他们或多或少涉足灰色地带
- **观看方式更重要：**人们并不喜欢每天固定时间收看电视剧，反而喜欢将周播剧“攒起来”，一次性看完；而且每天固定播放容易造成用户的流失

# 回归分析：引入

■ 流媒体视频播放平台网飞(Netflix)公司如何打造成功剧集？

■ 网飞公司：

- 《纸牌屋》

■ BBC版本的改编：1990.12

- 网飞版本的《纸牌屋》

《纸牌屋》（英语：*House of Cards*）是1990年的政治惊悚片电视连续剧，一共有四集。故事时空设定为英国首相玛格丽特·撒切尔夫人的请辞后。本剧于11月18日至12月9日期间，由英国广播公司播出，旋即广受好评。

本剧乃由安德鲁·戴维斯改编保守党幕僚长迈克尔·多布斯的同名小说，内维尔·泰勒也在1996年为英国广播公司国际频道对多布斯的小说进行戏剧化。本剧有两部电视连续集—《戏王者》和《最后一击》。本剧的开场和结束的主题曲音乐是“弗朗西斯·厄克特的进行曲”<sup>[1]</sup>。

纸牌屋 <i>House of Cards</i>	
	
原作	《纸牌屋》 迈克尔·多布斯作品
编剧	安德鲁·戴维斯 迈克尔·多布斯
导演	保罗·斯德

# 回归分析：引入

《纸牌屋》（英语：*House of Cards*）是美国一部政治权谋题材的网络连续剧，是奈飞自制剧。本剧由鲍尔·威利蒙创作并改编自安德鲁·戴维斯的BBC同名电视剧。两部剧集都是基于迈克尔·多布斯同名小说

创作的。第一季全部13集于2013年2月1日在流媒体网站奈飞首播。<sup>[1][2]</sup>第二季的13集于2014年2月14日全部放出。<sup>[1][3][4]</sup>2014年2月4日，即第二季播出十天之前，奈飞宣布续订本剧第三季，<sup>[5][6]</sup>于2015年2月27日播出。<sup>[7]</sup>本剧的第四季于2016年3月4日全部放出。<sup>[8]</sup>2016年1月，奈飞宣布续订第五季，于2017年5月30日播出，同时宣布鲍尔·威利蒙于第四季后退出本剧。<sup>[9]</sup>2017年10月30日，由于凯文·史派西被指控性侵犯，奈飞宣布第六季共13集会安排在2018年播出，而这将会是本剧的最后一季。<sup>[10]</sup>2017年11月3日，宣布史派西已经从本剧中被解雇。<sup>[11]</sup>第六季于2018年11月2日全部放出。

本剧背景设置在现今的华盛顿哥伦比亚特区，讲述了众议院南卡罗来纳州第五国会选区民主党籍议员、多数党党鞭弗兰克·安德伍德（Frank Underwood，凯文·史派西饰演）在晋升为国务卿的希望破灭后，在妻子克莱尔（Claire Underwood，罗宾·怀特饰演）的帮助下，开始运用复杂的权术获得最高权力的故事。剧集主要涉及的主题是无情的实用主义<sup>[12]</sup>、操纵和权力。<sup>[13]</sup>本剧一大特色是男主经常打破第四面墙，与观众对话，直接对镜头说出内心的真实想法。<sup>[14]</sup>

本剧的第一季获得了9项黄金时段艾美奖提名，成为了首部获得艾美奖提名的网络电视剧集。<sup>[15]</sup>本剧获得的9项提名则包括“最佳剧情类剧集”、“最佳剧情类剧集男主角”（凯文·史派西）、“最佳剧情类剧集女主角”（罗宾·怀特）以及“最佳剧情类剧集导演”（大卫·芬奇）等。本剧还获得了4项金球奖提名，其中罗宾·怀特凭借剧中表演最终获得“最佳剧情类剧集女主角”的奖项，也使本剧成为首部获得主要影视表演类奖项的网络剧集。而本剧的第二季获得了13项黄金时段艾美奖的提名和3个金球奖提名，<sup>[16]</sup>其中凯文·史派西获得了金球奖“最佳剧情类剧集男主角”的奖项。

纸牌屋	
House of Cards	
	
	《纸牌屋》片头画面
类型	剧情 政治 政治惊悚
原作	《纸牌屋》（小说） 作者：迈克尔·多布斯 《纸牌屋》（英国电视剧） 编剧：安德鲁·戴维斯
开创	鲍尔·威利蒙
主演	凯文·史派西 罗宾·怀特 迈克尔·凯利
国家 / 地区	 美国
语言	英语
季数	6
集数	73（每集列表）
每集长度	43–59分钟
作曲	杰夫·比尔
制作	
执行制片	大卫·芬奇 凯文·史派西 艾瑞克·罗斯 乔什华·多恩 达纳·布努内蒂

从内容、制作团队、宣传发行方式，《纸牌屋》的成功处处离不开数据，以及网飞公司对于影响电视剧集成功因素的分析与求证

回归分析：即是帮助我们从事现实世界的观测数据入手，为某一特定变量（被解释变量或因变量）构建其对于另一个或多个变量（解释变量或自变量）依存关系的统计方法

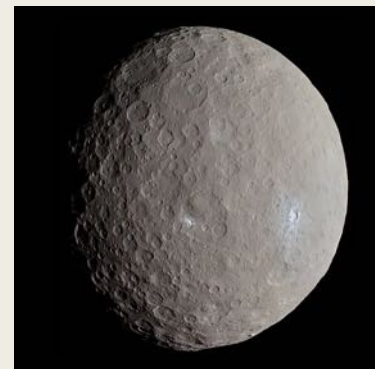
# 回归分析

约翰·卡尔·弗里德里希·高斯  
Carl Friedrich Gauß

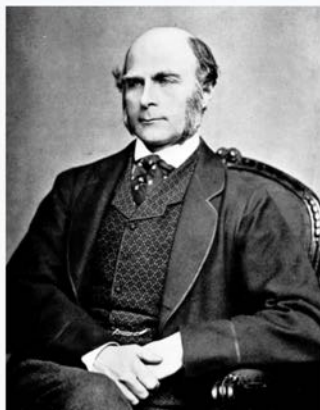


起初是由有着“数学王子”之称的高斯提出：

- 18岁的高斯发现了最小二乘法，并在此基础上创立的测量平差理论的帮助下，测算天体的运行轨迹，找到了小行星谷神星的运行轨迹。



弗朗西斯·高尔顿

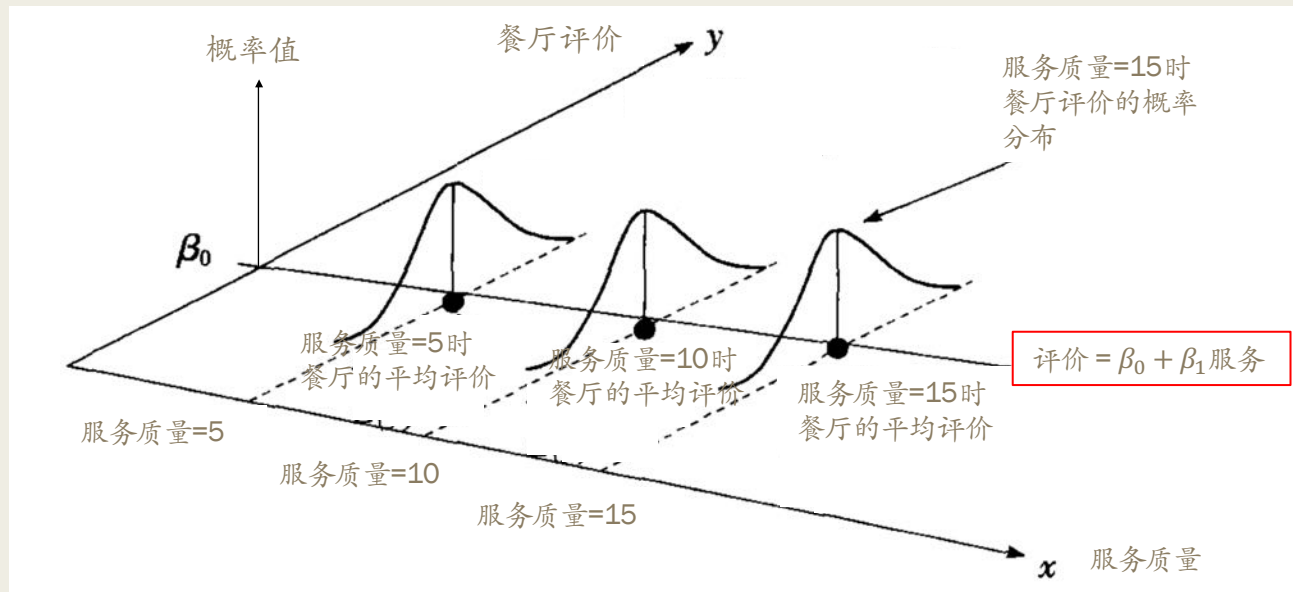


- 高尔顿在研究种子的生长规律时，提出了“regression”这一用语，不过其含义为“虽然父母的身高会遗传给孩子，但后代的身高会逐渐回归到中间”，与当前统计学中回归的含义并不相同。

# 回归分析

## ■ 回归分析：

- 研究如何用适当的数学模型去近似地表达或估计解释变量（服务质量）与被解释变量（餐厅评价）之间的平均变化关系



- 其目的是要：利用过去数据，揭示本质规律，并预测未来！
  - 根据已知的或固定的解释变量数值，去估计、预测所研究的被解释变量的总体平均值，判断某一项营销决策的价值

# 学术研究中的回归分析

Google Scholar

"regression" (source:"Marketing Science")



Articles

About 243 results (0.06 sec)

Any time  
Since 2022  
Since 2021  
Since 2018  
Custom range...

Sort by relevance  
Sort by date

Any type  
Review articles

include patents  
 include citations

**Frontiers: The persuasive effect of Fox News: Noncompliance with social distancing during the COVID-19 pandemic**

[PDF] informs.org

[A Simonov, S Sacher, JP Dubé...](#) - **Marketing Science**, 2022 - pubsonline.informs.org

... (panel (a)) and the corresponding reduced-form **regression** that replaces viewership with the channel position (panel (b)). The reduced-form **regression** tests whether we find a sig...

☆ Save Cite Cited by 4 Related articles All 4 versions

**Frontiers: The impact of ad-blockers on online consumer behavior**

[PDF] informs.org

[V Todri](#) - **Marketing Science**, 2022 - pubsonline.informs.org

... Difference-in-differences panel data **regression** results for consumer spending on brands consumers have not experienced in the past. The demographic fixed effects (FEs) include fixed ...

☆ Save Cite Cited by 5 Related articles

**Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer**

[PDF] informs.org

[P Anand, C Lee](#) - **Marketing Science**, 2022 - pubsonline.informs.org

... the following multiple **regression** framework with continuous independent variables of prices P and dependent variables of sales S, and we propose the following log-log **regression** in a ...

☆ Save Cite

回归分析是量化营销研究中最重要统计方法!

# 学术研究中的回归分析

Article

AM>  
AMERICAN MARKETING  
ASSOCIATION

## Do Spoilers Really Spoil? Using Topic Modeling to Measure the Effect of Spoiler Reviews on Box Office Revenue

Journal of Marketing  
1-19

© American Marketing Association 2020




Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0022242920937703

journals.sagepub.com/home/jmx



Jun Hyun (Joseph) Ryoo , Xin (Shane) Wang , and Shijie Lu 

### Empirical Analysis

#### Model of Box Office Revenue

Let  $i$  denote movies and  $t$  denote the days after release. The dependent variable is  $\ln(\text{DAILYREV})_{it}$ , which represents the log-transformed daily box office revenue for movie  $i$  on day  $t$ . To examine the relationship between spoiler reviews and box office revenue, we considered the following model specification:

$$\begin{aligned} \ln(\text{DAILYREV})_{it} = & \beta_1 \ln(\text{DAILYREV})_{i,t-1} \\ & + \beta_2 \ln(\text{INTENSITY})_{i,t-1} + \beta_3 \text{PROP}_{i,t-1} \\ & + \beta_4 \ln(\text{CUMRATING})_{i,t-1} + \beta_5 \ln(\text{CUMVOL})_{i,t-1} \\ & + \beta_6 \ln(\text{ADVERT})_{i,t-1} + \beta_7 \ln(\text{THEATERS})_{it} \\ & + \beta_8 t + \beta_9 \text{HOLIDAY}_{it} \\ & + \sum_{d=1}^6 \gamma_j I\{\text{DAYOFWEEK}_{it} = d\} + \omega_i + \epsilon_{it} \end{aligned}$$

(6)

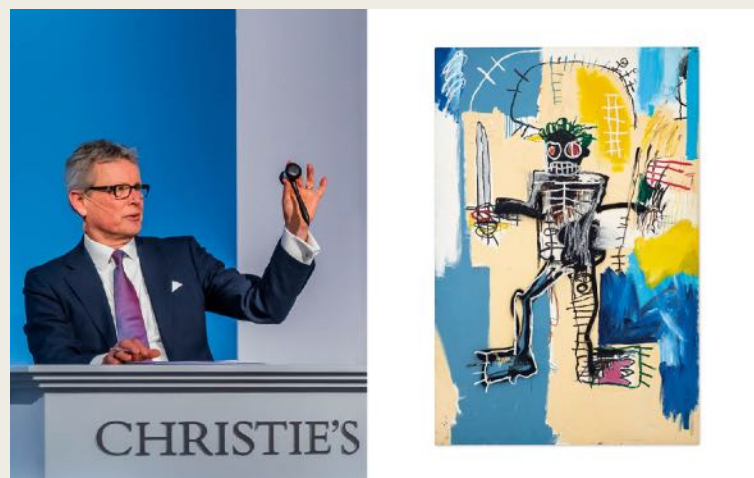
使用回归计量模型：量化电影评论的剧透强度，与其第二日票房之间的统计关系

结果发现：

- 前日电影评论中的剧透信息越多、次日电影票房越高
- 平均而言，剧透强度增加1个单位，次日票房增加0.18个单位

营销期刊 (JM) 2020, 评论中的剧透信息可以消减消费者的风险感知, 进而增加电影票房

# 学术研究中的回归分析



一项由中国人民大学、蚂蚁金服和Tilburg大学开展的研究，利用来自全球共计608个拍卖行、超过181万的拍卖数据，发现“画作转手的信息”可以有效建立起投资者的信任，可以将拍卖成功率提升4%、将落槌价提高54%，并使画作投资品的年化收益率提升5%~16%

“艺术品拍卖中的信任” (In Art We Trust)




# 学术研究中的回归分析

Article


**Media Coverage of Climate Change and Sustainable Product Consumption: Evidence from Hybrid Vehicle Market**


新闻中对于气候变化的报道，会影响人们的可持续性产品选择：  
以混合动力汽车市场为例

Yubo Chen , Mrinal Ghosh, Yong Liu, and Liang Zhao

**Abstract**  
As sustainable consumption becomes increasingly important, firms must better understand the drivers behind the consumption of these products. This article examines the effects of mass media in the context of the U.S. hybrid vehicle market. Drawing on monthly sales data, the authors provide evidence that the general coverage of climate change or global warming by major media outlets exerts an overall positive impact on the sales of hybrid vehicles. This impact mainly comes from the media reports that assert that climate change is occurring. In contrast, media coverage that either denies climate change or holds a neutral stance on the issue has little impact. The authors provide preliminary evidence that a social norm advocating for environmentally friendly consumption plays an important role in how media coverage affects consumer purchase. They provide implications for theory and practice and call for future research that examines the causal impact of media in general on consumer decisions, especially in domains that are crucial for the society.

**Keywords**  
climate change, global warming, hybrid vehicle, media, social marketing, sustainability

  
AMERICAN MARKETING ASSOCIATION

Journal of Marketing Research  
1-17  
© American Marketing Association 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/0022243719865898  
journals.sagepub.com/home/mrj  


研究发现，平均而言，前一个月主流大众媒体中与气候变化有关的新闻报道数量每增加1篇，下个月企业内混合动力汽车销量显著增加6辆

营销研究期刊 (JMR) 2019，大众媒体如何影响人们的消费行为？

# 学术研究中的回归分析

informs

<http://pubsonline.informs.org/journal/mnsc>

MANAGEMENT SCIENCE

Vol. 67, No. 12, December 2021, pp. 7291–7307  
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

通过低成本来刺激消费：来自COVID-19疫情期间的大规模政策实验

## Stimulating Consumption at Low Budget: Evidence from a Large-Scale Policy Experiment Amid the COVID-19 Pandemic

Qiao Liu,<sup>a</sup> Qiaowei Shen,<sup>a</sup> Zhenghua Li,<sup>b</sup> Shu Chen<sup>b</sup>

<sup>a</sup> Guanghua School of Management, Peking University, 100871 Beijing, China; <sup>b</sup> Research Institute, Ant Group, 310013 Hangzhou, China

Contact: qiao\_liu@gsm.pku.edu.cn, <https://orcid.org/0000-0003-0007-7190> (QL); qshen@gsm.pku.edu.cn, <https://orcid.org/0000-0003-3269-4003> (QS); sunny.lzh@antgroup.com (ZL); emily.cs@antgroup.com (SC)

Received: November 4, 2020

Revised: February 11, 2021

Accepted: March 12, 2021

Published Online in Articles in Advance:  
October 7, 2021

<https://doi.org/10.1287/mnsc.2021.4119>

Copyright: © 2021 INFORMS

数字消费券可以显著刺激消费，  
1元消费券带来3.4~5.8元的消费

**Abstract.** We use a novel panel with detailed transaction records of more than one million de-identified individuals to study the effect of a large-scale Chinese government-issued digital coupon program on consumer spending. Exploiting a difference-in-differences approach, we find that **the digital coupon is highly effective in stimulating consumption**. An effective government subsidy of RMB 1 can drive excess spending of RMB 3.4 to RMB 5.8, and the effect persists across multiple coupon issuance waves. **In explaining the results, we find that a behavioral model with mental accounting and loss aversion can match the empirical evidence from the field.** Our analysis, by illustrating the importance of embedding behavioral factors into the design and implementation of public policy, informs the current debate about cost-effective policy tools to recover the economy.



平均而言，1 RMB的政府补贴消费券，可以带动消费者3.4~5.8 RMB的额外花费

管理科学 (MASC) 2021, 数字消费券是否真的可以带动消费? 与支付宝平台的合作

# 简单回归分析

- 以画廊拍卖为例，假设我们有N次拍卖记录的数据(样本量为N)：
  - 针对每一次交易记录  $i$ ，我们知道 $Y_i$  和 $X_i$ ，其中：
    - $Y$ ：是我们想要解释、想要预测的变量，因此也被称为“被解释变量”(Dependent Variable)、“结果变量”(Outcome Variable)
      - $Y_i$  = 当前画作的成交价，是变量 $Y$ 的一个观测值
    - $X$ ：是我们认为会对被解释变量产生影响的变量，因此也被称为“自变量”(Independent Variable)、“解释变量”(Explanatory Variable)
      - $X_i$  = 画作交易历史的详尽程度，是变量 $X$ 的一个观测值
- 单变量回归模型： $Y = \beta_0 + \beta_1 X$
- 单变量回归方程： $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ，其中 $\epsilon_i$ 是样本(交易记录 $i$ )的随机误差项
  - 如果一个艺术品的交易详尽程度是 $k$ ，那么平均而言，该作品的成交价为 $\beta_0 + \beta_1 X_i$
  - 详尽程度每增加 1 个单位，作品成交价增加  $\beta_1$  个单位

# 简单回归分析

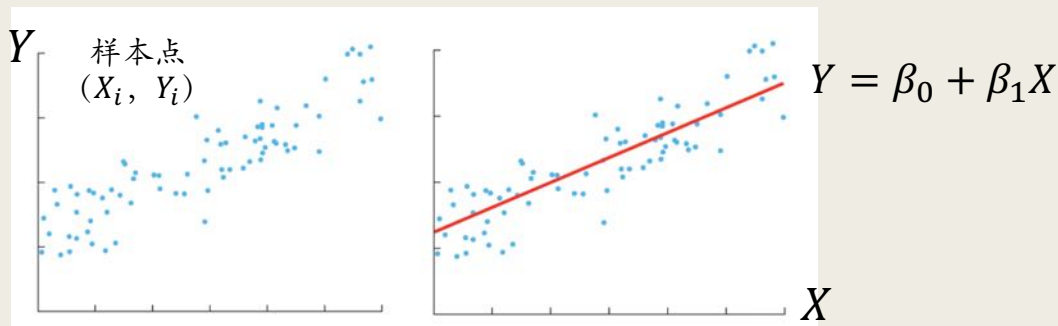
- 以画廊拍卖为例，假设我们有N次拍卖记录的数据(样本量为N)：
- 假设： $Price_i = 100 + 10 \times Details_i + \epsilon_i$ 
  - 如果我们已知一幅作品的背景信息详尽程度是20，那么我们可以预测，平均意义上而言，它的成交价格应为 $100 + 10 \times 20 = 300$
  - 如果随着某一史料的发现、一副作品的背景信息增加了1条，那么平均意义上而言，它的成交价格会上涨10倍。
- 假设： $Price_i = 100 - 30 \times Scandal_i + \epsilon_i$ 
  - $Scandal_i$ ：画作  $i$  创作者的丑闻
  - 平均意义上而言，丑闻数量增加1则，作品的成交价格会下降30倍！

# 简单回归分析

■ 单变量回归方程： $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ：

- 怎么计算  $\beta_0$  和  $\beta_1$  ？

- 单变量线性回归的本质：找到一条**最优的**线性函数，来拟合自变量与因变量之间相依关系。此时， $Y_i$ 的取值分成线性、可被解释的部分（ $\beta_0 + \beta_1 X_i$ ）加上当前模型设定中无法被解释的部分（误差项 $\epsilon_i$ ）



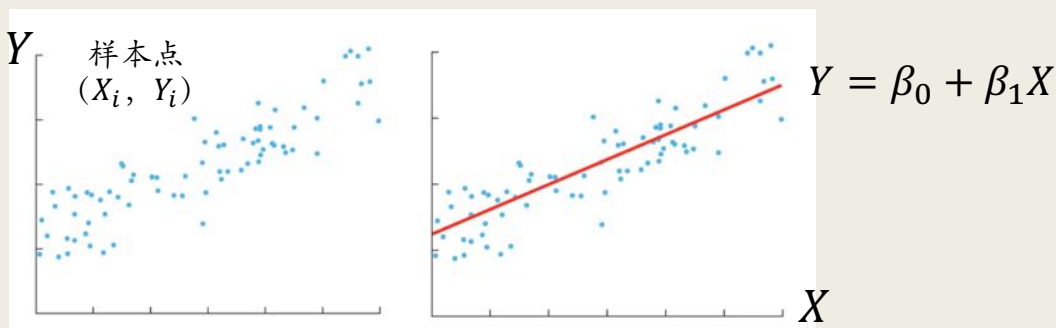
- 小调查：大家觉得，应该如何理解或计算“最优的线性拟合”？

# 简单回归分析

■ 单变量回归方程： $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ：

- 怎么计算  $\beta_0$  和  $\beta_1$  ？

- 单变量线性回归的本质：找到一条**最优**的线性函数，来拟合自变量与因变量之间相依关系。此时， $Y_i$ 的取值分成线性、可被解释的部分（ $\beta_0 + \beta_1 X_i$ ）加上当前模型设定中无法被解释的部分（误差项 $\epsilon_i$ ）



- 小调查：大家觉得，应该如何理解或计算“最优的线性拟合”？
  - **这个线性拟合足够好** =  $Y_i$ 中能被解释的部分足够多 =  $Y_i$ 中无法被解释的部分足够少 = 误差项  $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$ ，尽可能接近于0
  - $Q = \sum_{i=1}^N \epsilon_i^2 = \sum_{i=1}^N (Y_i - \beta_0 - \beta_1 X_i)^2$ ，找到 $\beta_0^*$ 和 $\beta_1^*$ ，使Q最小（最小二乘法）

# 简单回归分析

■ 单变量回归方程:  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$  :

- 怎么计算  $\beta_0$  和  $\beta_1$  ?

$$\beta_1^* = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

$$\beta_0^* = \bar{Y} - \beta_1^* \bar{X}$$

- 请大家结合回归分析的本质（使用线性的由解释变量主导的部分  $\beta_0 + \beta_1 X_i$ ，来理解  $Y_i$  的取值、解释  $Y_i$  的变差），思考：我们应该如何理解这两个式子？

- <https://www.youtube.com/watch?v=3g-e2aiRfbU>

# 简单回归分析

- 使用Excel进行回归分析（以众筹数据为例）：

The screenshot shows the Microsoft Excel interface with the 'Kickstarter-Project' workbook open. The '数据' (Data) tab is selected in the ribbon. The '分析工具' (Data Analysis) button in the ribbon is highlighted with a red box. Below the ribbon, a yellow warning message states: '可能的数据丢失 如果将此工作簿以逗号分隔(.csv)格式保存, 则某些功能可能会丢失。若要保留这些功能, 请以 Excel 文件格式保存。' (Possible data loss: If you save this workbook in comma-separated values (.csv) format, some features may be lost. To retain these features, save in Excel file format.)

The data table below shows columns for URL, Outcome, Target, FundingRaised, Backers, Comments, Location, Subtype, Duration, PhotosNumber, NumberOfProducts, Price, Gender, Created, and Backed. A dialog box titled '数据分析' (Data Analysis) is open, listing various analysis tools. The '回归' (Regression) option is selected and highlighted in blue.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	URL	Outcome	Target	FundingRaised	Backers	Comments	Location	Subtype	Duration	PhotosNumber	NumberOfProducts	Price	Gender	Created	Backed
2	https://www	1	68000	99943	238	12	NY	Hardware							0
3	https://www	1	5000	99720	759	364	CA	Hardware							0
4	https://www	0	45000	9933	114	7	CA	3DPrinting							0
5	https://www	0	60000	9911	17	1	CA	Software							0
6	https://www	0	100000	99065.5	739	8	NY	Technolog							0
7	https://www	0	110000	99	3	0	TX	Apps							0
8	https://www	0	5000	99	2	0	CA	Technolog							0
9	https://www	1	5000	9889	109	43	NY	Gadgets							2
10	https://www	0	5000	988	13	1	NY	Gadgets							3
11	https://www	0	3200	987	10	6	CA	Technolog							4
12	https://www	1	25000	98665.6	2162	190	CA	Hardware							2
13	https://www	1	5000	9865	227	9	TX	Web							5
14	https://www	1	500	986	19	0	CA	Hardware							15
15	https://www	0	91000	986	4	0	TX	Gadgets							0
16	https://www	0	30000	9852	50	6	NY	Hardware	30	8	11		40	U	2
17	https://www	0	150000	985	4	1	NY	Gadgets	40	30	6		929	U	0
18	https://www	0	17500	9830	469	37	CA	Hardware	30	5	9		24	M	0
19	https://www	0	6000	981	55	0	CA	Technology	40	16	5		15	M	3
20	https://www	1	50000	98033	693	302	CA	CameraEquip	30	37	7		99	U	3

数据分析

分析工具

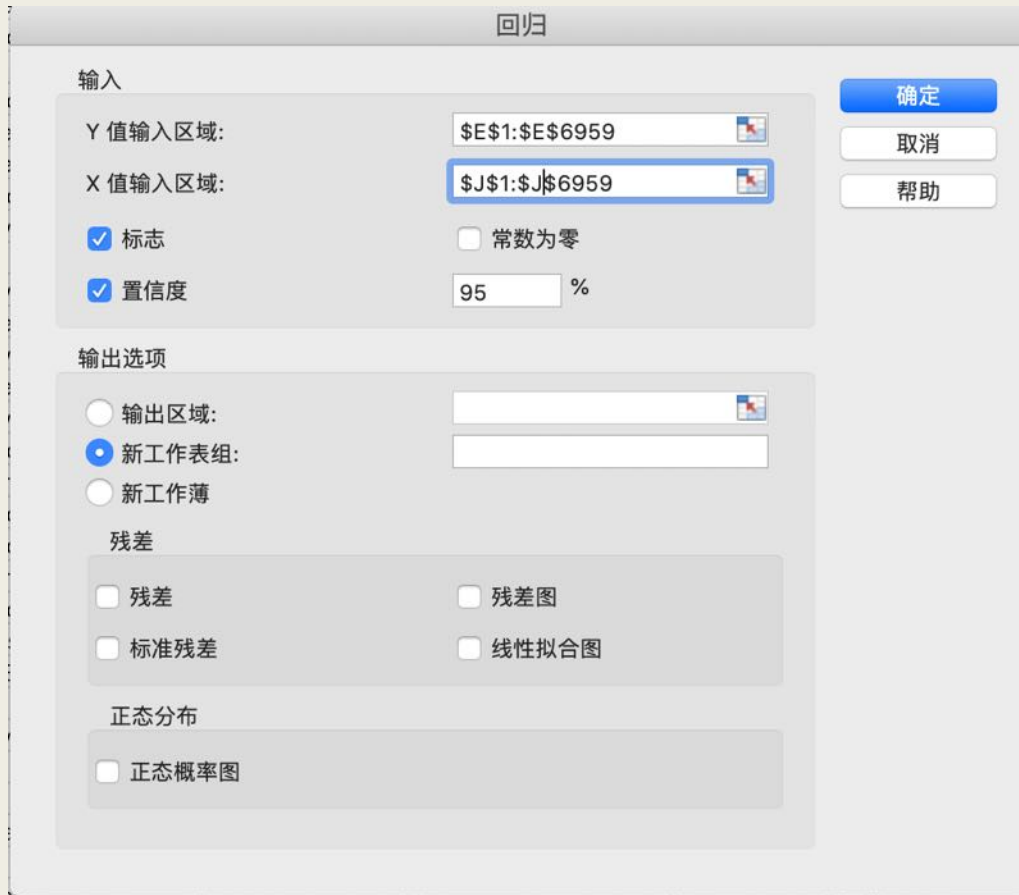
- 回归
- t-检验: 平均值的成对二样本分析
- t-检验: 双样本等方差假设
- t-检验: 双样本异方差假设
- z-检验: 双样本平均差检验

确定 取消 帮助



# 简单回归分析

- 使用Excel进行回归分析（以众筹数据为例）：



The image shows the '回归' (Regression) dialog box in Microsoft Excel. The dialog is titled '回归' and is divided into several sections:

- 输入 (Input):**
  - Y 值输入区域:
  - X 值输入区域:
  - 标志
  - 常数为零
  - 置信度:  %
- 输出选项 (Output Options):**
  - 输出区域:
  - 新工作表组:
  - 新工作簿
- 残差 (Residuals):**
  - 残差
  - 残差图
  - 标准残差
  - 线性拟合图
- 正态分布 (Normal Distribution):**
  - 正态概率图

On the right side of the dialog, there are three buttons: '确定' (OK), '取消' (Cancel), and '帮助' (Help).

# 简单回归分析

## ■ 使用Excel进行回归分析：

SUMMARY OUTPUT						
回归统计						
Multiple R	0.2292126	相关系数				
R Square	0.0525384	判定系数				
Adjusted R Square	0.0524022	调整后的判定系数				
标准误差	1860.8573					
观测值	6958					
方差分析						
	自由度	平方和	均方	F检验统计量	F检验统计量的显著性	
	df	SS	MS	F	Significance F	
回归分析	1	1.336E+09	1.336E+09	385.72264	1.261E-83	
残差	6956	2.409E+10	3462789.8			
总计	6957	2.542E+10				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	17.988604	27.151183	0.6625348	0.5076505	-35.236	71.213207
PhotosNum	33.438729	1.7025983	19.639823	1.261E-83	30.101117	36.776341

回归方程的截距

回归方程的斜率

检验回归系数的  
T统计量和P值

截距和斜率的置信区间

结果由三个部分构成：

1. 回归统计：展示了回归分析的常用统计量
2. 方差分析：是对回归分析的方差分析表。其主要作用是对回归方程的线性关系进行显著性检验
3. 回归参数估计的结果： $Backers = 17.99 + 33.44 \times PhotosNumber$

# 简单回归分析

- 使用Excel进行回归分析：

$$\textit{Backers} = 17.99 + 33.44 \times \textit{PhotosNumber}$$

- 含义：在Kickstarter上，项目图片每增加1张，项目的支持者数量平均增加约33个人，即项目是否有图片对于项目众筹成功而言是非常重要的积极影响因素！

# 简单回归分析

- 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

- 拟合优度：

SUMMARY OUTPUT					
回归统计					
Multiple R	0.2292126				
R Square	0.0525384				
Adjusted R Square	0.0524022				
标准误差	1860.8573				
观测值	6958				
方差分析					
	df	SS	MS	F	Significance F
回归分析	1	1.336E+09	1.336E+09	385.72264	1.26076E-83
残差	6956	2.409E+10	3462789.8		
总计	6957	2.542E+10			

- 判定系数 (R Square) 是对于估计回归方程拟合优度的度量。

$$R \text{ Square} = \frac{SSR \text{ (回归平方和)}}{SST \text{ (总平方和)}} = \frac{1.336E9}{2.409E10} = 0.052$$

- 含义：在项目支持人数 (Y) 的总变差中，有5.2%可以由项目使用的照片数量 (X) 与项目支持人数 (Y) 之间的线性关系来解释，即，在项目支持人数 (Y) 的取值变动中，有5.2%是由项目使用的照片数量 (X) 。可见二者之间有较强的线性关系
- R Square太低了，是为什么？

# 简单回归分析

- 使用Excel进行回归分析：

$$\textit{Backers} = 17.99 + 33.44 \times \textit{PhotosNumber}$$

- 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：
  - 请大家思考，为什么需要检验回归方程的质量？

# 简单回归分析

- 使用Excel进行回归分析：

$$\textit{Backers} = 17.99 + 33.44 \times \textit{PhotosNumber}$$

- 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：
  - 请大家思考，为什么需要检验回归方程的质量？
  - 因为这个回归结果是来自于随机抽取的样本数据的，这一回归方程是否反映了变量x与y之间的统计关系、或者变量x与y之间的统计关系是否显著，还有待于统计显著性检验的判定

# 简单回归分析

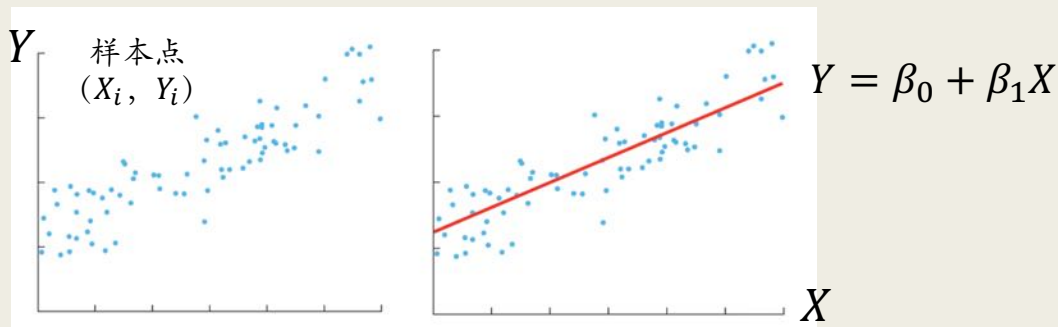
- 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

- 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：

- 检验1: 回归系数的检验（检验自变量对于因变量的影响是否显著）

- 请大家结合回归分析的本质（使用线性的由解释变量主导的部分  $\beta_0 + \beta_1 X_i$ ，来理解 $Y_i$ 的取值、解释 $Y_i$ 的变差），想想什么时候意味着自变量X对于因变量Y的影响不显著？



# 简单回归分析

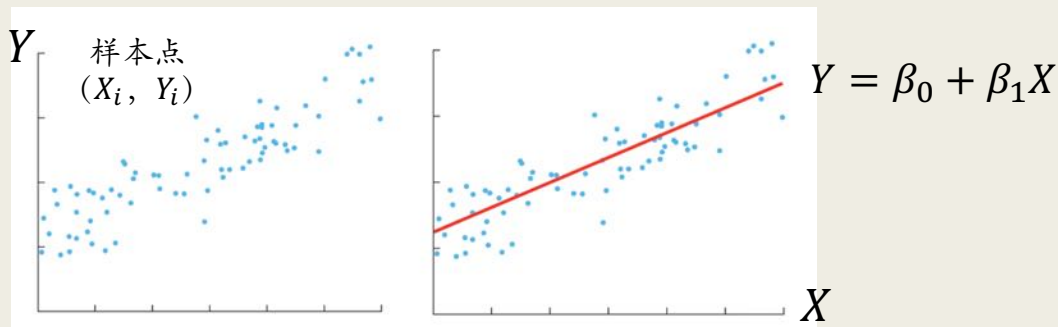
## ■ 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

## ■ 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：

### - 检验1: 回归系数的检验（检验自变量对于因变量的影响是否显著）

- 请大家结合回归分析的本质（使用线性的由解释变量主导的部分  $\beta_0 + \beta_1 X_i$ ，来理解  $Y_i$  的取值、解释  $Y_i$  的变差），想想什么时候意味着自变量X对于因变量Y的影响不显著？



- 回答：回归系数  $\beta_1 = 0$  的时候！此时回归线是一条水平线，因变量Y的取值不依赖于自变量X的取值，即两个变量之间没有线性关系



# 简单回归分析

## ■ 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

## ■ 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：

- 检验1: 回归系数的检验（检验自变量对于因变量的影响是否显著）

### ■ 检验的步骤：

- 待检验的假设： $H_0: \beta_1 = 0$ ；备择假设： $H_1: \beta_1 \neq 0$

- 计算检验统计量： $t = \frac{\beta_1^*}{se(\beta_1)} = \frac{\beta_1 \text{的估计值}}{\beta_1 \text{的标准误差}}$ （由数理统计理论支持）

- 做出决策：确定显著性水平， $\alpha = 0.05$ ，与p值比较大小：

■  $p \text{值} < \alpha$ ：有充足理由拒绝 $H_0$ ，说明X与Y之间存在显著线性关系

■  $p \text{值} > \alpha$ ：没有充足理由拒绝 $H_0$ ，没有证据表明X与Y之间存在显著线性关系

### ■ Excel数据分析中的结果：

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	17.988604	27.151183	0.6625348	0.5076505	-35.2359992	71.213207	-35.236	71.213207
PhotosNumber	33.438729	1.7025983	19.639823	1.261E-83	30.10111693	36.776341	30.101117	36.776341

# 简单回归分析

## ■ 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

## ■ 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：

- 检验2: 线性关系的检验（检验自变量和因变量之间的线性关系是否显著、二者之间的关系能否利用线性模型 $\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$ ）来解释

### ■ 检验的步骤：

- 待检验的假设： $H_0: \beta_1 = 0$ ；备择假设： $H_1: \beta_1 \neq 0$
- 计算检验统计量： $F = \frac{SSR/1}{SSE/(n-2)}$ （由数理统计理论支持）
- 做出决策：确定显著性水平， $\alpha = 0.05$ ，与p值比较大小：
  - $p\text{值} < \alpha$ ：有充足理由拒绝 $H_0$ ，说明X与Y之间存在显著线性关系
  - $p\text{值} > \alpha$ ：没有充足理由拒绝 $H_0$ ，没有证据表明X与Y之间存在显著线性关系

### ■ Excel数据分析中的结果：

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	1.336E+09	1.336E+09	385.72264	1.26076E-83
残差	6956	2.409E+10	3462789.8		
总计	6957	2.542E+10			

# 简单回归分析

- 使用Excel进行回归分析:

$$Backers = 17.99 + 33.44 \times PhotosNumber$$

- 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：
  - 检验2: 线性关系的检验（检验自变量和因变量之间的线性关系是否显著、二者之间的关系能否利用线性模型 $Backers = 17.99 + 33.44 \times PhotosNumber$ ）来解释
    - 请大家思考，这里的检验2（线性关系的检验）与检验1（回归系数的检验）是不是一回事？

# 简单回归分析

## ■ 使用Excel进行回归分析：

$$\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$$

## ■ 回归分析的显著性检验（上述估计方程是否能真实反映变量x和y之间的关系）：

- 检验2: 线性关系的检验（检验自变量和因变量之间的线性关系是否显著、二者之间的关系能否利用线性模型 $\text{Backers} = 17.99 + 33.44 \times \text{PhotosNumber}$ ）来解释
  - 请大家思考，这里的检验2（线性关系的检验）与检验1（回归系数的检验）是不是一回事？
  - 回答：
    - 在单变量回归分析中，检验1等价于检验2。X与Y之间线性关系的显著，等价于回归系数 $\beta_1$ 显著不为0，X的取值变化会影响Y的取值
    - 但是在多变量回归分析中，检验1和2的意义是不同的。检验2是检验回归方程整体上的显著性，而检验1是检验各个回归系数的显著性

# 简单回归分析

## ■ 随堂小练习：

- 刚刚的例子中，我们已经利用Excel探索了Kickstarter上众筹项目的成功因素，发现项目支持者数量与其发布的照片数量之间存在显著地正向关系。
- 请大家打开Excel，尝试自己做以下回归分析，并在聊天框中写出你得到的回归方程和拟合优度，并简单解释你得到的回归结果：
  - 视频内容的伤心程度（Upset）与项目支持者数量（Backers）之间存在何种统计关系？

# 简单回归分析

## ■ 随堂小练习：

- 刚刚的例子中，我们已经利用Excel探索了Kickstarter上众筹项目的成功因素，发现项目支持者数量与其发布的照片数量之间存在显著地正向关系。
- 请大家打开Excel，尝试自己做以下回归分析，并在聊天框中写出你得到的回归方程和拟合优度，并简单解释你得到的回归结果：
  - 视频内容的伤心程度（Upset）与项目支持者数量（Backers）之间存在何种统计关系？
$$\text{Backers} = 216.20 + 59.87 \times \text{Upset}$$

SUMMARY OUTPUT								
回归统计								
Multiple R	0.0698445							
R Square	0.0048783							
Adjusted R	0.0047352							
标准误差	1907.0864							
观测值	6958							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	1	124019034	124019034	34.099467	5.472E-09			
残差	6956	2.53E+10	3636978.6					
总计	6957	2.542E+10						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	216.1976	29.165331	7.4128285	1.383E-13	159.02465	273.37054	159.02465	273.37054
Upset	59.866276	10.251996	5.8394749	5.472E-09	39.769235	79.963316	39.769235	79.963316

# 简单回归分析

## ■ 随堂小练习：

- 刚刚的例子中，我们已经利用Excel探索了Kickstarter上众筹项目的成功因素，发现项目支持者数量与其发布的照片数量之间存在显著地正向关系。
- 请大家打开Excel，尝试自己做以下回归分析，并在聊天框中写出你得到的回归方程和拟合优度，并简单解释你得到的回归结果：
  - 视频长度（VideoLength）与项目筹集金额（FundingRaised）之间存在何种统计关系？

# 简单回归分析

## ■ 随堂小练习：

- 刚刚的例子中，我们已经利用Excel探索了Kickstarter上众筹项目的成功因素，发现项目支持者数量与其发布的照片数量之间存在显著地正向关系。
- 请大家打开Excel，尝试自己做以下回归分析，并在聊天框中写出你得到的回归方程和拟合优度，并简单解释你得到的回归结果：
  - 视频长度（VideoLength）与项目筹集金额（FundingRaised）之间存在何种统计关系？  $FundingRaised = 17669.71 + 168.38 \times VideoLength$

SUMMARY OUTPUT								
回归统计								
Multiple R	0.1176077							
R Square	0.0138316							
Adjusted R Squar	0.0136898							
标准误差	184922.29							
观测值	6958							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	1	3.336E+12	3.336E+12	97.561776	7.37717E-23			
残差	6956	2.379E+14	3.42E+10					
总计	6957	2.412E+14						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	17669.707	3119.6988	5.6639146	1.539E-08	11554.14605	23785.269	11554.146	23785.269
VideoLength	168.38359	17.047469	9.8773365	7.377E-23	134.9653481	201.80183	134.96535	201.80183



# 简单回归分析

- 思考：如果我们得到了Y与X之间的回归方程，并且检验结果全部显著，此时我们可以说X导致了Y吗？为什么？

# 简单回归分析

- 思考：如果我们得到了Y与X之间的回归方程，并且检验结果全部显著，此时我们可以说X导致了Y吗？为什么？

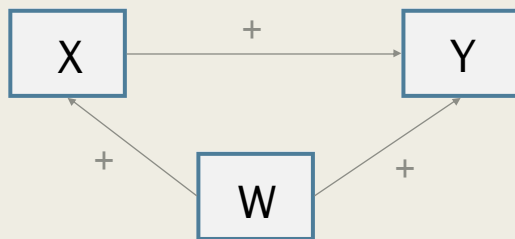
- 例子1：Y = 财富水平；X = 教育水平， $\beta_1 > 0$

- 虽然可以理解为，教育水平越高，这个人的财富水平越高，但也有可能是因为这个人的财富水平越高，他能接触和负担的教育资源越好



- 例子2：Y = 火灾造成的经济损失；X = 消防车的派出数量， $\beta_1 > 0$

- 两个变量之间满足很强的线性回归关系，但并不是说消防救援行动导致了火灾经济损失，而是X与Y都受到“火情严重程度”的影响



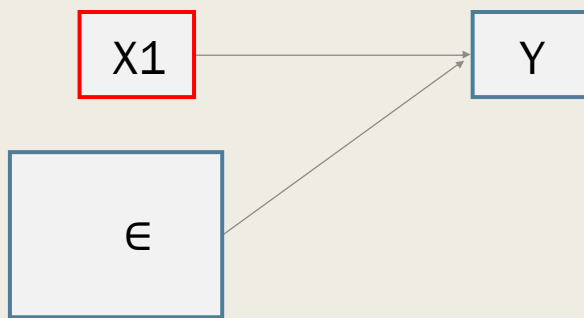
# 简单回归分析



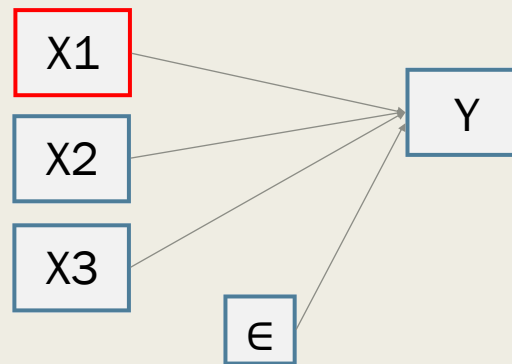
[https://www.bilibili.com/video/BV1mr4y1C7gs?spm\\_id\\_from=333.337.search-card.all.click&vd\\_source=19aecf57e19c27dfc37d8587a32cafdd](https://www.bilibili.com/video/BV1mr4y1C7gs?spm_id_from=333.337.search-card.all.click&vd_source=19aecf57e19c27dfc37d8587a32cafdd)

# 多元回归分析

- 单变量回归模型:  $Y = \beta_0 + \beta_1 X$
- 多变量回归模型:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$
- 为什么要进行多变量回归分析?
  - 在实际中影响因变量的因素往往有多个, 在单变量回归模型中, 我们实际上是把这些众多的因素都放入了随机误差项 $\epsilon_i$  (如: 火情的严重程度), 因此无法在这些变量都不变的条件下, 研究某一个解释变量 ( $X_1$ ) 对于被解释变量 ( $Y$ ) 的影响
  - 因此, 在我们只能使用“观测数据”展开实证研究的时候, 使用多变量回归分析, 可以帮助我们尽可能清楚地在控制其他变量不变的时候, 研究 $X_1$ 与 $Y$ 之间的关系



$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

# 多元回归分析

■ 单变量回归模型： $Y = \beta_0 + \beta_1 X$

■ 多变量回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

■ 以网飞的用户观影习惯为例，

- $Y$ ：用户对这部网剧的喜爱程度
- $X_1$ ：用户对导演和主演的喜爱程度
- $X_2$ ：剧情能在多大程度上满足人们的猎奇心理
- $X_3$ ：用户能在多大程度上自由支配刷剧节奏

# 多元回归分析

- 单变量回归模型： $Y = \beta_0 + \beta_1 X$

- 多变量回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- 以餐厅评价为例，

- $Y$ ：消费者对这家餐厅的喜爱程度
- $X_1$ ：餐厅的价格
- $X_2$ ：餐厅的服务
- $X_3$ ：餐厅的位置

- 如果不使用回归分析，我们可能认为这三个要素同等重要，

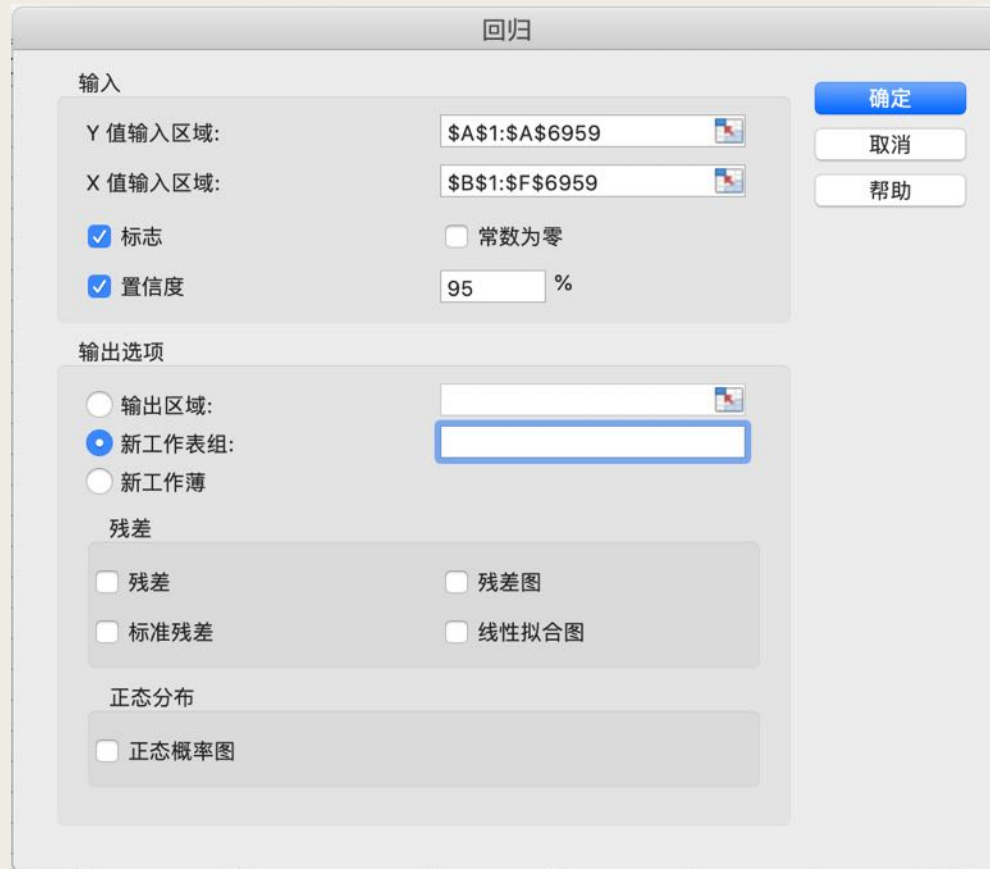
$$\text{Ratings} = 0.3 \times \text{Price} + 0.3 \times \text{Service} + 0.3 \times \text{Location}$$

- 通过回归我们可以发现，相比于其他要素而言，价格是更为重要的！

$$\text{Ratings} = 0.6 \times \text{Price} + 0.15 \times \text{Service} + 0.25 \times \text{Location}$$

# 多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：



The image shows the '回归' (Regression) dialog box in Excel. It is divided into several sections:

- 输入 (Input):**
  - Y 值输入区域: \$A\$1:\$A\$6959
  - X 值输入区域: \$B\$1:\$F\$6959
  - 标志
  - 常数为零
  - 置信度: 95 %
- 输出选项 (Output Options):**
  - 输出区域:
  - 新工作表组:
  - 新工作簿
- 残差 (Residuals):**
  - 残差
  - 残差图
  - 标准残差
  - 线性拟合图
- 正态分布 (Normal Distribution):**
  - 正态概率图

Buttons on the right: 确定 (OK), 取消 (Cancel), 帮助 (Help).

# 多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：

SUMMARY OUTPUT								
回归统计								
Multiple R	0.6381639							
R Square	0.4072532							
Adjusted R Square	0.4068268							
标准误差	1472.2847							
观测值	6958							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	5	10353532744	2070706549	955.2894167	0			
残差	6952	15069309547	2167622.2					
总计	6957	25422842292						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-0.21629	27.9362484	-0.0077423	0.993822841	-54.97987	54.547285	-54.97987	54.547285
PhotosNumber	12.232107	1.411839147	8.66395239	5.60212E-18	9.4644714	14.999743	9.4644714	14.999743
Price	-0.036487	0.024234931	-1.5055458	0.132229238	-0.083995	0.0110211	-0.083995	0.0110211
VideoLength	0.2241586	0.139029783	1.61230599	0.106940762	-0.048382	0.4966994	-0.048382	0.4966994
FbNumber	0.0254427	0.028328607	0.89812714	0.369148873	-0.03009	0.0809754	-0.03009	0.0809754
Comments	1.9150575	0.029765039	64.3391576	0	1.856709	1.9734061	1.856709	1.9734061

*Backers*

$$= -0.22 + 12.23 \times PhotosNumber - 0.04 \times Price + 0.22 \times VideoLength + 0.03 \times FbNumber + 1.92 \times Comments$$



# 多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：

*Backers*

$$= -0.22 + 12.23 \times PhotosNumber - 0.04 \times Price + 0.22 \times VideoLength + 0.03 \times FbNumber + 1.92 \times Comments$$

SUMMARY OUTPUT					
回归统计					
Multiple R	0.6381639				
R Square	0.4072532				
Adjusted R Square	0.4068268				
标准误差	1472.2847				
观测值	6958				
方差分析					
	df	SS	MS	F	Significance F
回归分析	5	10353532744	2070706549	955.2894167	0
残差	6952	15069309547	2167622.2		
总计	6957	25422842292			

- 拟合优度：

- 判定系数 (R Square) 是对于估计回归方程拟合优度的度量。

$$R \text{ Square} = \frac{SSR \text{ (回归平方和)}}{SST \text{ (总平方和)}} = \frac{10353532744}{25422842292} = 0.407$$

- 含义：在项目支持人数的总变差中，有40.7%可以由此处5个变量组成的线性关系来解释，即，在项目支持人数的取值变动中，有40.7%是由这5个变量的取值决定的。可见二者之间有较强的线性关系

# 多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：

*Backers*

$$= -0.22 + 12.23 \times \text{PhotosNumber} - 0.04 \times \text{Price} + 0.22 \times \text{VideoLength} + 0.03 \times \text{FbNumber} + 1.92 \times \text{Comments}$$

- 对回归结果进行检验：

- 检验1：回归系数的检验

- 检验的步骤：

- 待检验的假设： $H_0: \beta_1 = 0$ ；备择假设： $H_1: \beta_1 \neq 0$
- 计算检验统计量： $t = \frac{\beta_1^*}{se(\beta_1)} = \frac{\beta_1 \text{的估计值}}{\beta_1 \text{的标准误差}}$ （由数理统计理论支持）
- 做出决策：确定显著性水平， $\alpha = 0.05$ ，与比较p值比较大小：
  - **p值 <  $\alpha$** ：有充足理由拒绝 $H_0$ ，说明X与Y之间存在显著线性关系
  - **p值 >  $\alpha$** ：没有充足理由拒绝 $H_0$ ，没有证据表明X与Y之间存在显著线性关系

- Excel数据分析中的结果：

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.21629	27.9362484	-0.0077423	0.993822841	-54.97987	54.547285
PhotosNumber	12.232107	1.411839147	8.66395239	<b>5.60212E-18</b>	9.4644714	14.999743
Price	-0.036487	0.024234931	-1.5055458	0.132229238	-0.083995	0.0110211
VideoLength	0.2241586	0.139029783	1.61230599	0.106940762	-0.048382	0.4966994
FbNumber	0.0254427	0.028328607	0.89812714	0.369148873	-0.03009	0.0809754
Comments	1.9150575	0.029765039	64.3391576	<b>0</b>	1.856709	1.9734061

在5个变量中，仅有“图片数量”、“评论数量”的影响是显著的，其余3个变量都没有通过显著性检验、对Y的影响不大

# 多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：

*Backers*

$$= -0.22 + 12.23 \times PhotosNumber - 0.04 \times Price + 0.22 \times VideoLength + 0.03 \times FbNumber + 1.92 \times Comments$$

- 对回归结果进行检验：

- 检验2：线性关系的检验

- 检验的步骤：

- 待检验的假设： $H_0: \beta_1 = 0$ ；备择假设： $H_1: \beta_1 \neq 0$

- 计算检验统计量： $F = \frac{SSR/1}{SSE/(n-2)}$ （由数理统计理论支持）

- 做出决策：确定显著性水平， $\alpha = 0.05$ ，与比较p值比较大小：

- **p值 <  $\alpha$** ：有充足理由拒绝 $H_0$ ，说明X与Y之间存在显著线性关系

- **p值 >  $\alpha$** ：没有充足理由拒绝 $H_0$ ，没有证据表明X与Y之间存在显著线性关系

- Excel数据分析中的结果：

方差分析	df	SS	MS	F	Significance F
回归分析	5	10353532744	2070706549	955.2894167	0
残差	6952	15069309547	2167622.2		
总计	6957	25422842292			

解释变量的组合与Y之间的线性关系显著

# 多元回归分析

- 思考：既然拟合优度（R Square）反映的是5个变量的线性组合对于Y的解释力度，而通过对变量回归系数展开逐一检验我们发现，有3个变量对Y的影响并不显著，只有2个变量是有效变量。那么大家觉得，如果此刻我们丢掉这3个变量、仅使用2个有效变量，回归输出结果中的哪些部分会发生显著的变化？哪些不会？理由是什么？

SUMMARY OUTPUT						
回归统计						
Multiple R	0.6381639					
R Square	0.4072532					
Adjusted R Square	0.4068268					
标准误差	1472.2847					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	5	10353532744	2070706549	955.2894167	0	
残差	6952	15069309547	2167622.202			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.21629	27.9362484	-0.00774228	0.993822841	-54.97986537	54.547285
PhotosNumber	12.232107	1.411839147	8.663952385	5.60212E-18	9.46447141	14.999743
Price	-0.036487	0.024234931	-1.50554582	0.132229238	-0.083994662	0.0110211
VideoLength	0.2241586	0.139029783	1.612305992	0.106940762	-0.048382265	0.4966994
FbNumber	0.0254427	0.028328607	0.898127143	0.369148873	-0.030090027	0.0809754
Comments	1.9150575	0.029765039	64.33915763	0	1.85670897	1.9734061

# 多元回归分析

SUMMARY OUTPUT						
回归统计						
Multiple R	0.6381639					
R Square	0.4072532					
Adjusted R Square	0.4068268					
标准误差	1472.2847					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	5	10353532744	2070706549	955.2894167	0	
残差	6952	15069309547	2167622.202			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.21629	27.9362484	-0.00774228	0.993822841	-54.97986537	54.547285
PhotosNumber	12.232107	1.411839147	8.663952385	5.60212E-18	9.46447141	14.999743
Price	-0.036487	0.024234931	-1.50554582	0.132229238	-0.083994662	0.0110211
VideoLength	0.2241586	0.139029783	1.612305992	0.106940762	-0.048382265	0.4966994
FbNumber	0.0254427	0.028328607	0.898127143	0.369148873	-0.030090027	0.0809754
Comments	1.9150575	0.029765039	64.33915763	0	1.85670897	1.9734061

SUMMARY OUTPUT						
回归统计						
Multiple R	0.63776537					
R Square	0.40674467					
Adjusted R Square	0.40657407					
标准误差	1472.59835					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	2	10340605504	5170302752	2384.225639	0	
残差	6955	15082236788	2168545.908			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	23.4742664	21.48638819	1.092518026	0.274643313	-18.645611	65.5941434
PhotosNumber	12.7152339	1.385207408	9.179299671	5.62263E-20	9.99980471	15.4306631
Comments	1.91647187	0.029740387	64.44004546	0	1.85817164	1.9747721

# 多元回归分析

SUMMARY OUTPUT						
回归统计						
Multiple R	0.6381639					
R Square	0.4072532					
Adjusted R Square	0.4068268					
标准误差	1472.2847					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	5	10353532744	2070706549	955.2894167	0	
残差	6952	15069309547	2167622.202			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.21629	27.9362484	-0.00774228	0.993822841	-54.97986537	54.547285
PhotosNumber	12.232107	1.411839147	8.663952385	5.60212E-18	9.46447141	14.999743
Price	-0.036487	0.024234931	-1.50554582	0.132229238	-0.083994662	0.0110211
VideoLength	0.2241586	0.139029783	1.612305992	0.106940762	-0.048382265	0.4966994
FbNumber	0.0254427	0.028328607	0.898127143	0.369148873	-0.030090027	0.0809754
Comments	1.9150575	0.029765039	64.33915763	0	1.85670897	1.9734061

SUMMARY OUTPUT						
回归统计						
Multiple R	0.63776537					
R Square	0.40674467					
Adjusted R Square	0.40657407					
标准误差	1472.59835					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	2	10340605504	5170302752	2384.225639	0	
残差	6955	15082236788	2168545.908			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	23.4742664	21.48638819	1.092518026	0.274643313	-18.645611	65.5941434
PhotosNumber	12.7152339	1.385207408	9.179299671	5.62263E-20	9.99980471	15.4306631
Comments	1.91647187	0.029740387	64.44004546	0	1.85817164	1.9747721

再进一步，请大家思考：

如果我此时删掉Comments这个有效变量，大家觉得回归输出结果中的哪些部分会发生显著变化？哪些不会？理由是什么？

# 多元回归分析

SUMMARY OUTPUT						
回归统计						
Multiple R	0.6381639					
R Square	0.4072532					
Adjusted R Square	0.4068268					
标准误差	1472.2847					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	5	10353532744	2070706549	955.2894167	0	
残差	6952	15069309547	2167622.202			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.21629	27.9362484	-0.00774228	0.993822841	-54.97986537	54.547285
PhotosNumber	12.232107	1.411839147	8.663952385	5.60212E-18	9.46447141	14.999743
Price	-0.036487	0.024234931	-1.50554582	0.132229238	-0.083994662	0.0110211
VideoLength	0.2241586	0.139029783	1.612305992	0.106940762	-0.048382265	0.4966994
FbNumber	0.0254427	0.028328607	0.898127143	0.369148873	-0.030090027	0.0809754
Comments	1.9150575	0.029765039	64.33915763	0	1.85670897	1.9734061

SUMMARY OUTPUT						
回归统计						
Multiple R	0.63776537					
R Square	0.40674467					
Adjusted R Square	0.40657407					
标准误差	1472.59835					
观测值	6958					
方差分析						
	df	SS	MS	F	Significance F	
回归分析	2	10340605504	5170302752	2384.225639	0	
残差	6955	15082236788	2168545.908			
总计	6957	25422842292				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	23.4742664	21.48638819	1.092518026	0.274643313	-18.645611	65.5941434
PhotosNumber	12.7152339	1.385207408	9.179299671	5.62263E-20	9.99980471	15.4306631
Comments	1.91647187	0.029740387	64.44004546	0	1.85817164	1.9747721

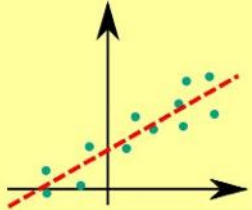
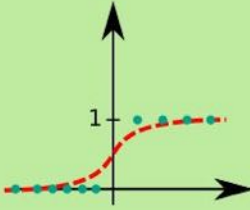
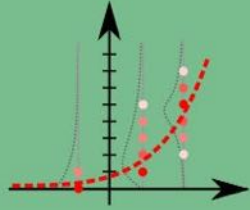
SUMMARY OUTPUT									
回归统计									
Multiple R	0.2292126								
R Square	0.0525384								
Adjusted R Square	0.0524022								
标准误差	1860.8573								
观测值	6958								
方差分析									
	df	SS	MS	F	Significance F				
回归分析	1	1335676426	1335676426	385.7226404	1.26076E-83				
残差	6956	24087165866	3462789.8						
总计	6957	25422842292							
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%	
Intercept	17.988604	27.15118343	0.66253479	0.507650495	-35.2359992	71.213207	-35.236	71.213207	
PhotosNumber	33.438729	1.702598305	19.6398228	1.26076E-83	30.10111693	36.776341	30.101117	36.776341	

# 与回归有关的其他讨论

- 回归分析到这里就终止了吗？



# 与回归有关的其他讨论

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
<ul style="list-style-type: none"> <li>① Econometric modelling</li> <li>② Marketing Mix Model</li> <li>③ Customer Lifetime Value</li> </ul>	<ul style="list-style-type: none"> <li>① Customer Choice Model</li> <li>② Click-through Rate</li> <li>③ Conversion Rate</li> <li>④ Credit Scoring</li> </ul>	<ul style="list-style-type: none"> <li>① Number of orders in lifetime</li> <li>② Number of visits per user</li> </ul>
		
<p>Continuous ⇒ Continuous</p>	<p>Continuous ⇒ True/False</p>	<p>Continuous ⇒ 0,1,2,...</p>
$y = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y = \frac{1}{1 + e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y \sim \text{Poisson}(\lambda)$ $\ln \lambda = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$
<p>lm(y ~ x1 + x2, data)</p>	<p>glm(y ~ x1 + x2, data, family=binomial())</p>	<p>glm(y ~ x1 + x2, data, family=poisson())</p>
<p>1 unit increase in x increases y by <math>\alpha</math></p>	<p>1 unit increase in x increases log odds by <math>\alpha</math></p>	<p>1 unit increase in x multiplies y by <math>e^\alpha</math></p>

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing.



自变量：连续  
因变量：连续

自变量：连续  
因变量：是或否  
(如：众筹项目是否成功？  
消费者最终是否决定购买？)

自变量：连续  
因变量：计数变量  
(如：消费者重复购买次数？  
消费者共享单车使用次数？  
本月混合动力汽车售出多少台？)

# 与回归有关的其他讨论

- 回归分析到这里就终止了吗？
- 有哪些渠道可以让我自学回归分析（计量经济学）？
  - *Seeing Theory* (A visual introduction to probability and statistics):  
<https://seeing-theory.brown.edu/index.html#firstPage>
  - UCLA 统计学教材: <https://stats.oarc.ucla.edu/>
  - Minitab 的线上教程: <https://support.minitab.com/zh-cn/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/before-you-start/example/>
  - ....

# 课后小作业

## ■ 请各位同学：

- 利用今天讲的回归模型以及Excel实操，分析变量之间的统计关系
- 结合研究问题的具体情境，对回归模型的结果做出解读

DDL：8月19日晚12点，提交回归分析的结果、以及结合研究情景对于回归结果的解读（[joeyliu1997@163.com](mailto:joeyliu1997@163.com)）

请各位同学发作业时，在邮件主题上注名，谢谢大家！

也欢迎大家就任何课程问题与我们沟通～