论文课 LESSON5

科研论文的数据分析与写作

（刘佳妮, 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）

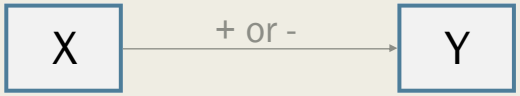
- 回归分析的基本原理与实操
  - 如何利用回归模型来刻画解释变量对于被解释变量的意义？
  - 回归模型背后的原理是什么？如何判断回归模型的好坏？
  - 如何实现每一步操作？如何解读回归结果？

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

- 学术论文的各个部分
  - 之前的课程：题目、摘要与关键词、引言、文献评述、模型、分析和讨论
  - 使用三线表报告回归分析结果
- 学术论文的投稿、修改和发表

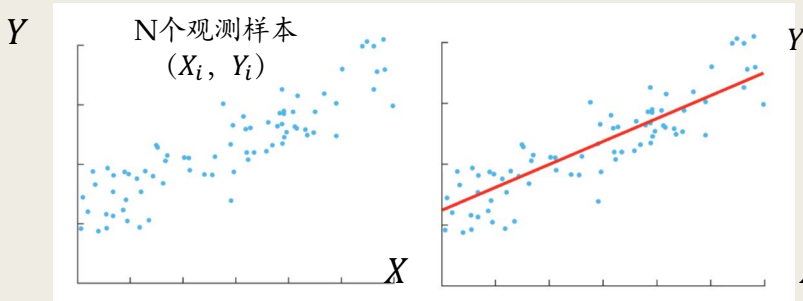
回归分析

- **回归分析：**
  - *是帮助我们从现实世界的观测数据入手，为某一特定变量（被解释变量Y）构建其对于另一个或多个变量（解释变量X）依存关系的统计方法*



  - 其目的是要：
    - 利用过去数据，揭示本质规律(X对Y有何影响？)，预测未来(给定X水平下，Y的取值)!
    - 可以帮助我们理解现象，判断某一项营销决策的价值

- **简单回归分析（单变量、单元、一元回归分析）：**
  - *假设我们有N个观测样本(X_i, Y_i)(样本量 = N)，关于每个样本我们知道：*
    - 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是样本i 被解释变量的取值
    - 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是样本i 被解释变量的取值



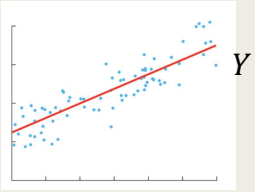
  - *回归分析帮助我们得到一条最优的线性函数 $Y = \beta_0 + \beta_1 X$，来解释X与Y之间的统计关系*
    - β_0：X=0时y的取值；纵轴上的截距项
    - β_1：X每变动一个单位，Y变动多少个单位
    - 回归分析的原理：将 Y 的取值变动情况分成两个部分，$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$：
      - $\beta_0 + \beta_1 X$：*Y中可以被X所解释的部分，及估计得到的回归方程*
      - ϵ_i：*随机的、无法被解释的部分*

      通过找到使误差平方和$\sum_{i=1}^{N} \epsilon_i^2$最小化的$\beta_0^*$和$\beta_1^*$，使 Y中可以被X所解释的部分足够大（$\beta_0 + \beta_1 X$），而随机的、无法被解释的部分足够小（$\epsilon_i$）；这一方法又被称作“最小二乘法”

- **简单回归分析（单变量、单元、一元回归分析）：**
  - *以众筹数据集为例，我们可以得到回归方程：*

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

    - 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！
  - *针对回归分析的结果，我们需要对其好坏展开评价与检验：*
    - 使用拟合优度来评价回归模型的好坏
      - $\text{R Square} = \frac{\text{SSR（回归平方和）}}{\text{SST（总平方和）}} = \frac{1.336E9}{2.409E10} = 0.052 = 5.2\%$
      - 含义：在Y的总变差中，有5.2%可以由X与Y之间的线性关系来解释，可见二者之间有较强的线性关系
    - 检验1：回归系数的显著性检验
      - H_0：$\beta_1 = 0$（此时意味着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）
      - ***p值< α：有充足理由拒绝H_0，说明X与Y之间存在显著线性关系***



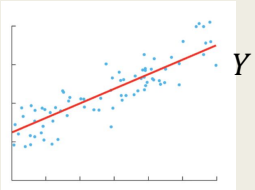
$Y = \beta_0 + \beta_1 X$

    - 检验2：回归方程的显著性检验
      - H_0：$\beta_1 = 0$（自变量和因变量之间的线性关系是否显著）
      - ***p值< α：有充足理由拒绝H_0，说明X与Y之间存在显著线性关系***
      - 在单变量回归分析中，检验1和检验2等价；但是在多变量回归分析中，二者并不等价

- **简单回归分析（单变量、单元、一元回归分析）：**
  - *以众筹数据集为例，我们可以得到回归方程：*

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

    - 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！
  - *针对回归分析的结果，我们需要对其好坏展开评价与检验：*
    - 使用拟合优度来评价回归模型的好坏
      - $\text{R Square} = \frac{\text{SSR（回归平方和）}}{\text{SST（总平方和）}} = \frac{1.336E9}{2.409E10} = 0.052 = 5.2\%$
      - 含义：在Y的总变差中，有5.2%可以由X与Y之间的线性关系来解释，可见二者之间有较强的线性关系
    - 检验1：回归系数的显著性检验
      - H_0：$\beta_1 = 0$（此时意味着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）
      - ***p值< α：有充足理由拒绝H_0，说明X与Y之间存在显著线性关系***



$Y = \beta_0 + \beta_1 X$

    - 检验2：回归方程的显著性检验
      - H_0：$\beta_1 = 0$（自变量和因变量之间的线性关系是否显著）
      - ***p值< α：有充足理由拒绝H_0，说明X与Y之间存在显著线性关系***
      - 在单变量回归分析中，检验1和检验2等价；但是在多变量回归分析中，二者并不等价

p值：当原假设成立的时候，我们观测到当前数据情况的概率。
如果p值很小、但我们又确实在现实中观测到了当前数据，
那么此时我们就有充足的理由拒绝掉原假设、认为它不对。

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
R Square	0.0525384
Adjusted R S	0.0524022
标准误差	1860.8573
观测值	6958

相关系数
判定系数
调整后的判定系数

方差分析	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			

自由度　平方和　均方　F检验统计量　F检验统计量的显著性

	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

- 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，$\beta_1 > 0$；教育水平可以提高薪酬水平吗？
  - 例子2：Y = 火灾造成的经济损失；X = 消防车的派出数量，$\beta_1 > 0$；消防车会加剧火灾的经济损失？

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

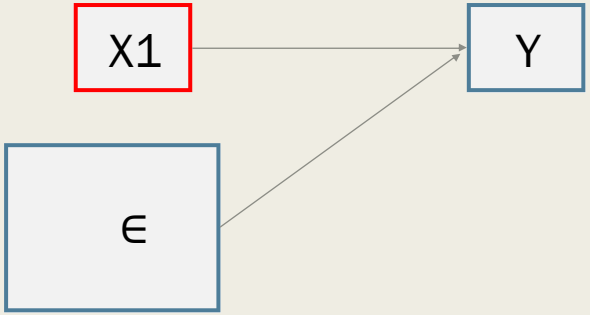
  - 例子2：Y = 火灾造成的经济损失；X = 消防车的派出数量，$\beta_1 > 0$
    - 两个变量之间满足很强的线性回归关系，但并不是说消防救援行动导致了火灾经济损失，而是X与Y都受到“火情严重程度”的影响

简单回归分析

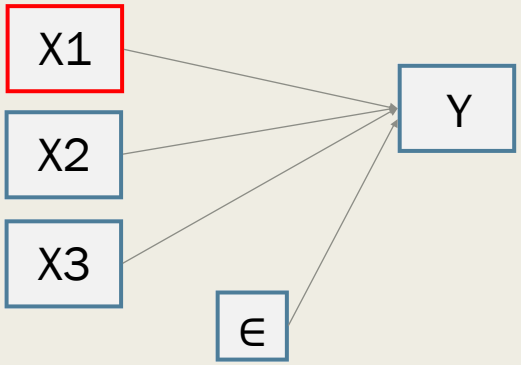
https://www.bilibili.com/video/BV1mr4y1C7gs?spm_id_from=333.337.search-card.all.click&vd_source=19aecf57e19c27dfc37d8587a32cafdd

多元回归分析

- 单变量回归模型：$Y = \beta_0 + \beta_1 X$
- 多元回归模型：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
- 为什么要进行多元回归分析？



简单回归分析：$Y = \beta_0 + \beta_1 X_1 + \in$



多元回归分析：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \in$

- 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$\in_i$（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 (X1) 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X1与Y之间的关系，此时X1为解释变量，X2和X3为控制变量

多元回归分析

- **单变量回归模型：** $Y = \beta_0 + \beta_1 X$
- **多元回归模型：** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
  - *假设我们有N个观测样本* $(X_{1i}, X_{2i}, X_{3i}, Y_i)$ *(样本量* $= N$ *)*

- **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，

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

- 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！

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

多元回归分析

- 使用Excel进行多元回归分析（以众筹数据为例）：

Backers	Comments	PhotosNumber	Price	FbNumber	VideoLength
238	12	52	179	0	250
759	364	22	110	8	145
114	7	5	75	0	196
17	1	0	100	0	599
739	8	5	375	0	220

导入每列数据时可以使用快捷键：shift + control + 下键

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

- 写出多元回归方程：

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

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。

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

  - 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）：**

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

- **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< α：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> α：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	5.60212E-18	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	0	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

A

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			

B

	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

C

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

A

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			

B

	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

C

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

A, B, C 均未发生很大变化

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

A

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			

B

	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

C

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

A, B, C 均未发生很大变化

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

线性回归方程的整体解释力度下降

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

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

线性回归方程的整体解释力度下降

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

多元回归分析

- 随堂小练习：请大家打开Excel，尝试自己做以下回归分析，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：

  - 利用项目筹集金额（FundingRaised）和捐赠者数量（Backers），计算人均捐款数额（AvgDonate），将计算结果保存在新的一列：
    - **6116个项目的捐赠者数量（Backers）> 0；样本量 = 6116**
  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

多元回归分析

- 随堂小练习：请大家打开Excel，尝试自己做以下回归分析，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：

  - 利用项目筹集金额（FundingRaised）和捐赠者数量（Backers），计算人均捐款数额（AvgDonate），将计算结果保存在新的一列：
    - **6116个项目的捐赠者数量（Backers）> 0；样本量 = 6116**
  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$AvgDonate = 47.17 + 0.001Comments + 3.66PhotoNumbers + 0.08Price - 0.005FbNumber + 0.14VideoLength$$

  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

多元回归分析

- 随堂小练习：请大家打开Excel，尝试自己做以下回归分析，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：

  - 利用项目筹集金额（FundingRaised）和捐赠者数量（Backers），计算人均捐款数额（AvgDonate），将计算结果保存在新的一列：
    - **6116个项目的捐赠者数量（Backers）> 0；样本量 = 6116**
  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$AvgDonate = 47.17 + 0.001Comments + 3.66PhotoNumbers + 0.08Price - 0.005FbNumber + 0.14VideoLength$$

  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$Backers = 4.31 + 1.91Comments + 12.22PhotoNumbers - 0.06Price + 0.03FbNumber + 0.24VideoLength$$

  - 并分析：FundingRaised 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

多元回归分析

- 随堂小练习：请大家打开Excel，尝试自己做以下回归分析，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：

  - 利用项目筹集金额（FundingRaised）和捐赠者数量（Backers），计算人均捐款数额（AvgDonate），将计算结果保存在新的一列：
    - **6116个项目的捐赠者数量（Backers）> 0；样本量 = 6116**
  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$AvgDonate = 47.17 + 0.001Comments + 3.66PhotoNumbers + 0.08Price - 0.005FbNumber + 0.14VideoLength$$

  - 并分析：AvgDonate 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$Backers = 4.31 + 1.91Comments + 12.22PhotoNumbers - 0.06Price + 0.03FbNumber + 0.24VideoLength$$

  - 并分析：FundingRaised 与 Comments、PhotosNumber、Price、FbNumber、VideoLength 之间存在何种统计关系？

$$FundingRaised = -10819 + 213.91Comments + 2234.67PhotoNumbers + 17.68\ Price + 0.61FbNumber + 35.51VideoLength$$

Q：有同学可以解释这个结果吗？

使用三线表展示回归结果

2. 被解释变量

Table 4. Estimation Results of the Model of Box Office Revenue.

1. 选择的模型：OLS，普通最小二乘回归模型

	OLS (1)	FE (2)	GMM with IVs for Lagged DV (3)	GMM with IVs for Lagged DV, WOM, and Marketing Mix (4)
Intercept	5.958***	-	-	-
	(.025)			
$\ln(\text{DAILYREV})_{i,\,t-1}$	-	.474***	.606***	.638***
		(.011)	(.013)	(.017)
$\ln(\text{INTENSITY})_{i,\,t-1}$	.180***	.045***	.077***	.060***
	(.010)	(.008)	(.014)	(.014)
$\text{PROP}_{i,\,t-1}$	.564***	−.016	.170**	.075
	(.041)	(.031)	(.055)	(.062)
$\ln(\text{CUMRATING})_{i,\,t-1}$	.457***	−.004	.202***	.171***
	(.012)	(.020)	(.029)	(.027)
$\ln(\text{CUMVOL})_{i,\,t-1}$	.140***	−.192***	.048***	.037***
	(.004)	(.014)	(.008)	(.010)
$\ln(\text{ADVERT})_{i,\,t-1}$	.123***	.018***	.051***	.097***
	(.002)	(.002)	(.003)	(.007)
$\ln(\text{THEATERS})_{it}$	.894***	.431***	.356***	.337***
	(.002)	(.010)	(.012)	(.020)
AGE(t)	−.033***	−.019***	−.012***	−.008***
	(3.33e−4)	(.001)	(.001)	(.001)
HOLIDAY_{it}	.759***	.572***	.558***	.533***
	(.023)	(.016)	(.018)	(.018)
DAYOFWEEK dummies	Yes	Yes	Yes	Yes
Movie fixed effects	No	Yes	Yes	Yes
Adjusted R-squared	.899	.952	-	-
Cluster-robust standard error	No	Yes	Yes	Yes
Number of observations			49,057	

4. 拟合优度

5. 样本量

*$p < .05$; **$p < .01$; ***$p < .001$.

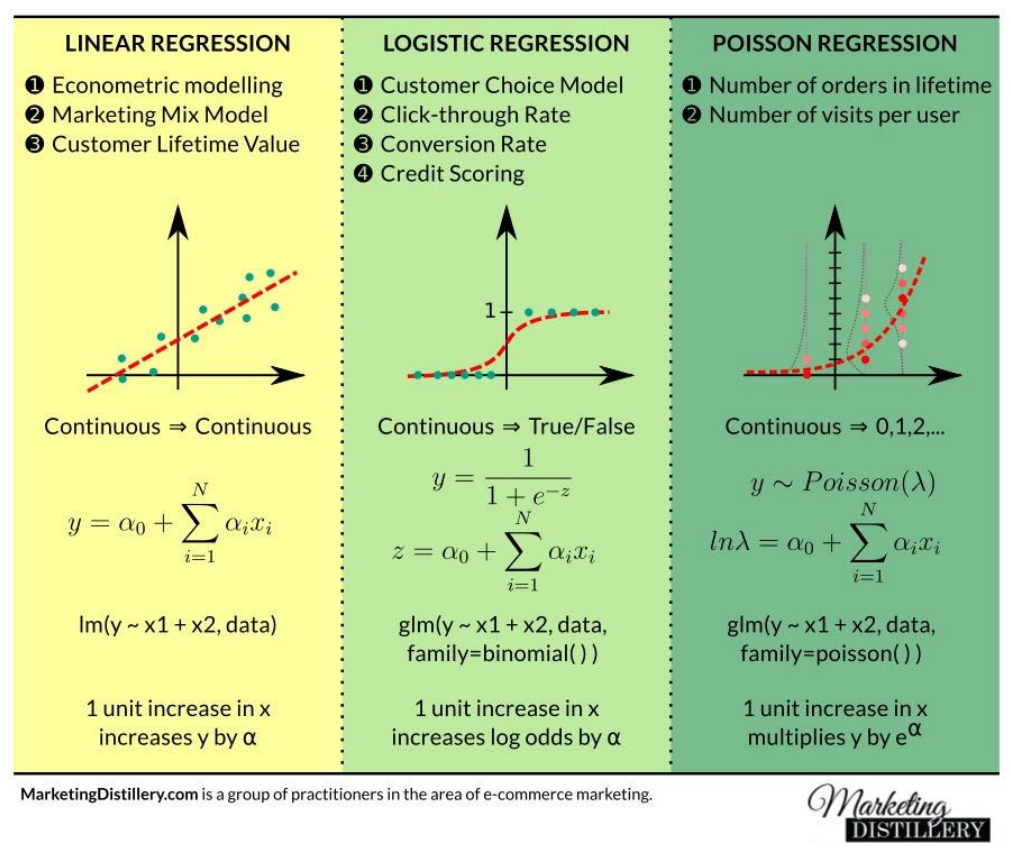
6. 表注：参数估计的显著性系数的符号含义

3. 解释变量与控制变量的估计结果：

- 估计值
- 估计值的标准误
- 参数的显著性检验结果

与回归有关的其他讨论

- 回归分析到这里就终止了吗？



....

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）？
  - *Bilibili & YouTube*
  - *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/*
  - ....

学术论文的写作

1. 研究问题的提出与聚焦：
    - *解释变量X;*
    - *被解释变量Y;*
    - *主效应X → Y*
    - *解释为何X与Y之间存在这种主效应？主效应在何时会被放大、何时会被抑制?*

2. 回顾现有文献中对于上述要素的讨论
    - *现有文献的结论（“评”）*
    - *现有文献的不足（“述”）*
    - *本研究的独特贡献（“点明价值”）*

学术论文的写作

3. 数据的来源、获取方式、合理性与价值（样本与总体、数据集的价值）

Setting and Data

We obtained a list of movies released in the United States between January 2013 and December 2017 from WildAboutMovies.com. From this list, we sampled 993 movies that have their daily box office revenue data available on BoxOfficeMojo.com.

We focused on the first eight weeks of daily box office revenue because 97% of total box office revenue is accrued within the first eight weeks of a movie's release (Liu 2006). We collected daily box office revenue and daily number of theaters in which a movie was playing, as well as other movie characteristics (e.g., Motion Picture Association of America rating, genre, and release type) from both BoxOfficeMojo.com and IMDb. We matched our movie sample with advertising spending data provided by Kantar Media.

We used IMDb to collect online WOM data for two reasons. First, IMDb is by far the most popular online movie review platform in the United States.[1] Second, IMDb requires users to label their reviews with spoiler warnings if a user believes that

学术论文的写作

4. 变量的定义、度量方式、与描述性统计特征

- *统计数据、统计变量与理论变量*
  - 我定义的变量可以研究我的问题吗？
  - 我定义的变量应该如何测量？需要哪些统计数据的支持？
    我目前收集到的数据可以支持我测量它们吗？

Table 1. Variable Definitions.

Variable Name	Description
DAILYREV	Box office revenue on day t for movie i.
INTENSITY	Spoiler intensity of spoiler reviews within the ten days prior to day t for movie i.
PROP	Moving average of proportion of spoiler reviews within the ten days prior to day t for movie i.
CUMRATING	Mean ratings of cumulative movie reviews on day t for movie i.
CUMVOL	Number of cumulative movie reviews on day t for movie i.
ADVERT	Average daily advertising expenditure on day t for movie i.
THEATERS	Number of theaters that screen movie i on day t.
AGE (t)	Number of days since the release of movie i in theaters.
HOLIDAY	Dummy variable for the ten federal holidays in the United States.
DAYOFWEEK	Indicator variables for each day of the week.

Table 2. Descriptive Statistics.

Variable	Mean	Standard Deviation	Minimum	Maximum
Daily level				
DAILYREV (in $)	1,039,985	3,265,580	5	119,119,282
INTENSITY	2.48	2.69	0	45.17
PROP	.18	.15	0	1
CUMRATING	6.27	1.48	0	10
CUMVOL	12.87	247.73	0	4,276
ADVERT (in $1,000)	126.3	621.7	0	6,807
THEATERS	1,240	1,309	1	4,535
Movie level				
MPAA ratings				
G & PG	.15	.36	0	1
PG-13	.40	.49	0	1
R	.40	.49	0	1
Unrated	.05	.21	0	1
Genres				
Action	.09	.28	0	1
Adventure/Sci-Fi	.10	.30	0	1
Comedy	.20	.40	0	1
Drama	.32	.47	0	1
Family	.10	.30	0	1
Foreign	.02	.14	0	1
Horror	.06	.24	0	1
Musical	.02	.12	0	1
Romance	.02	.14	0	1
Thriller	.08	.27	0	1
Release type				
Limited Release	.40	.49	0	1

学术论文的写作

4. 用表达式建立起回归模型

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:

解释变量X及其估计系数

被解释变量

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

随机误差项

学术论文的写作

5. 使用三线表记录回归模型的分析结果

2. 被解释变量

Table 4. Estimation Results of the Model of Box Office Revenue.

1. 选择的模型：OLS，普通最小二乘回归模型

	OLS (1)	FE (2)	GMM with IVs for Lagged DV (3)	GMM with IVs for Lagged DV, WOM, and Marketing Mix (4)
Intercept	5.958***	-	-	-
	(.025)			
$\ln(\text{DAILYREV})_{i,t-1}$	-	.474***	.606***	.638***
		(.011)	(.013)	(.017)
$\ln(\text{INTENSITY})_{i,t-1}$	.180***	.045***	.077***	.060***
	(.010)	(.008)	(.014)	(.014)
$\text{PROP}_{i,t-1}$	.564***	−.016	.170**	.075
	(.041)	(.031)	(.055)	(.062)
$\ln(\text{CUMRATING})_{i,t-1}$	.457***	−.004	.202***	.171***
	(.012)	(.020)	(.029)	(.027)
$\ln(\text{CUMVOL})_{i,t-1}$	.140***	−.192***	.048***	.037***
	(.004)	(.014)	(.008)	(.010)
$\ln(\text{ADVERT})_{i,t-1}$	.123***	.018***	.051***	.097***
	(.002)	(.002)	(.003)	(.007)
$\ln(\text{THEATERS})_{it}$	.894***	.431***	.356***	.337***
	(.002)	(.010)	(.012)	(.020)
AGE(t)	−.033***	−.019***	−.012***	−.008***
	(3.33e−4)	(.001)	(.001)	(.001)
HOLIDAY_{it}	.759***	.572***	.558***	.533***
	(.023)	(.016)	(.018)	(.018)
DAYOFWEEK dummies	Yes	Yes	Yes	Yes
Movie fixed effects	No	Yes	Yes	Yes
Adjusted R-squared	.899	.952	-	-
Cluster-robust standard error	No	Yes	Yes	Yes
Number of observations			49,057	

*$p < .05$; **$p < .01$; ***$p < .001$.

3. 解释变量与控制变量的估计结果：
- 估计值
- 估计值的标准误
- 参数的显著性检验结果

4. 拟合优度

5. 样本量

6. 表注：参数估计的显著性系数的符号含义

学术论文的写作

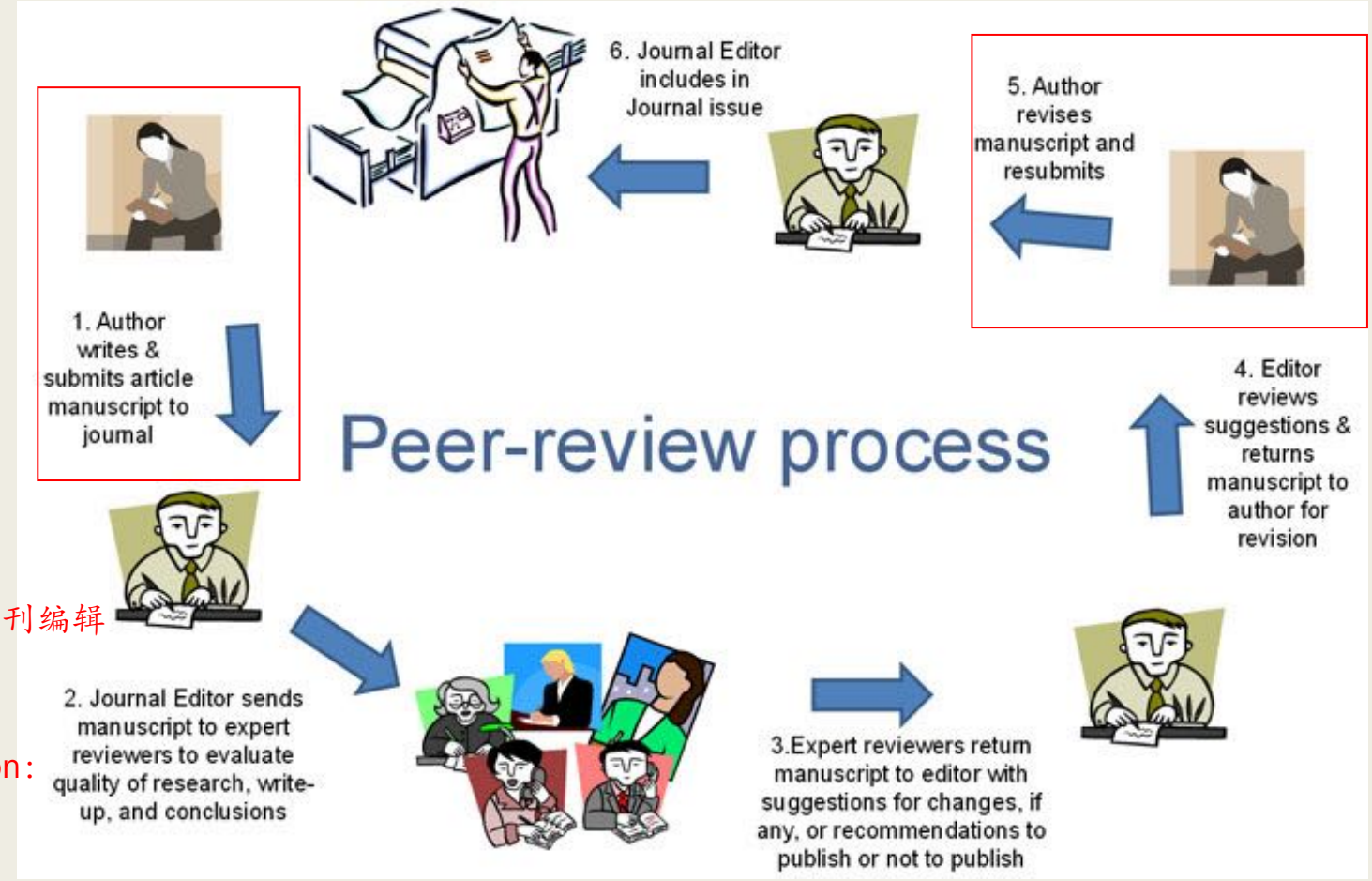
6. 对回归分析的结果展开解读
   - 系数的估计结果（估计值的大小与方向、p值）？
   - 系数估计结果的含义（数据表明，解释变量与被解释变量之间存在何种关系？）
   - 如何理解这一分析结果?

Empirical Findings

We begin with a standard ordinary least squares (OLS) regression of the model of box office revenue without the lagged dependent variable.[6] We report the results in Table 4, Column 1, which provides preliminary evidence that the association between spoiler intensity and box office revenue is positive and significant (.180, $p < .001$). We find that the association between spoiler volume and box office revenue is also positive and significant (.564, $p < .001$). Estimates for the control variables are of expected signs. For example, both CUMRATING and CUMVOL have positive associations with box office revenue. In addition, box office revenue is greater for movies that played in a larger number of theaters and spent more on advertising.

学术论文的投稿和修改

1. 期刊投稿流程
   - *同行评议*



期刊编辑

匿名审稿人

Desk Rejection：30～40%

学术论文的投稿和修改

1. 期刊投稿流程
    - *同行评议*
    - *投稿流程*
        - 仔细阅读期刊的投稿指南，包括投稿要求、评审流程、撰稿要求（题目、摘要、关键词、图标规范、参考文献、引用标志、字体与板式）
            - *营销科学学报：http://www.jms.org.cn:8081/jms/CN/column/column3.shtml*
            - *心理学报：https://journal.psych.ac.cn/xlxb/CN/column/column6.shtml*
            - *Journal of Marketing Research: https://journals.sagepub.com/author-instructions/MRJ*
        - 进入同行评议环节 (R&R, revision and resubmission; 2～4轮不等)，针对期刊编辑和审稿人提出的意见作出修改
            - *R&R分类：Major Revision 和 Minor Revision*
            - *从他人的R&R中学习经验：https://journal.psych.ac.cn/xlxb/CN/0439-755X/current.shtml*

学术论文的投稿和修改

1. 期刊投稿流程
2. 会议投稿流程（流程短、不需要反复修改、无出版物、可以参加学术会议获取修改建议）
    - *INFORMS Marketing Science Conference: https://www.informs.org/Meetings-Conferences/INFORMS-Conference-Calendar/2022-ISMS-Marketing-Science-Conference*
    - *JMS 中国市场营销学术年会暨博士生论坛: http://www.jmsmeeting.org.cn/index.php/subs?fid=23*
    - *CMIC 中国市场营销国际学术年会:*
    - *IFMSA 营销科学与应用国际会议*

课后小作业

论文辅导阶段		
课程主题	授课导师	是否有作业
科研论文写作的基本规范	副导师	1. 基于PPT或者学术期刊列表，挑选1-2篇顶刊论文（中英文皆可） 2. 详细阅读其各个部分，以及各部分内部的行文逻辑，并撰写论文阅读笔记 **DDL：8月5日晚12点，提交论文阅读笔记**
科研论文的选题与文献综述	副导师	1. 提出1-2个感兴趣的研究问题 2. 根据第一节课提供的学术期刊清单，找到相关的3-4篇顶刊论文 3. 在阅读文章内容后，重点学习文献综述部分；通过滚雪球的方式找到其他相关文章、并填写相关文献汇总表 **DDL：8月8日晚12点，提交选题及文献汇总表**（汇总表的格式可参考课程网站上的样例）
科研论文数据的采集和分析（一）	副导师	1. 选定1个研究问题，利用网络资源采集或者开放资源获取数据 2. 确定核心变量，对核心变量展开度量，并进行描述统计分析 **DDL: 8月17日晚12点，提交数据集的采集方式、描述、变量定义、及变量的描述统计分析结果**
科研论文数据的采集和分析（二）	副导师	1. 利用回归模型，分析变量之间的统计关系 2. 对回归模型的结果做出解读 **DDL: 8月19日晚12点，提交回归分析结果表以及对回归结果的解读**
完成一篇高质量的学术论文	副导师	整合之前的作业，完成一篇学术论文 **DDL: 8月21日晚12点**

请各位同学：撰写3～5页的分析报告，包含：研究问题的提出、文献评述、数据采集、变量定义及描述性分析结果、利用三线表汇报回归分析的结果、对于回归结果的解读

DDL：8月21日晚12点 (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～