中国商业统计学会
Commerce Statistical Society of China

Credamo 见数®

市场调研与数据分析

“正大杯”第十二届全国市场调查与分析大赛公益培训

2.3 数据建模与分析（三）

数据建模与分析（三）

主要内容

1. 逻辑回归
2. 中介效应
3. 调节效应
4. 聚类分析

1 逻辑回归

分类型因变量

客户流失：Y = 流失与否

征信：Y = 是否逾期

购买决策：Y = 是否购买

逻辑回归

- **逻辑回归**（logistic regression）又称logistic回归分析，是一种广义的线性回归分析模型。通过逻辑回归分析能得到**自变量的权重**，从而了解哪些自变量是因变量的影响因素，同时也可**预测**在确定条件下因变量所处状态的可能性。

- 逻辑回归的因变量可以是二分类（即“是”或“否”，“有”或“无”等），也可以是多分类，但是二分类的更为常用，也更加容易解释。实际中最为常用的就是二分类的逻辑回归。

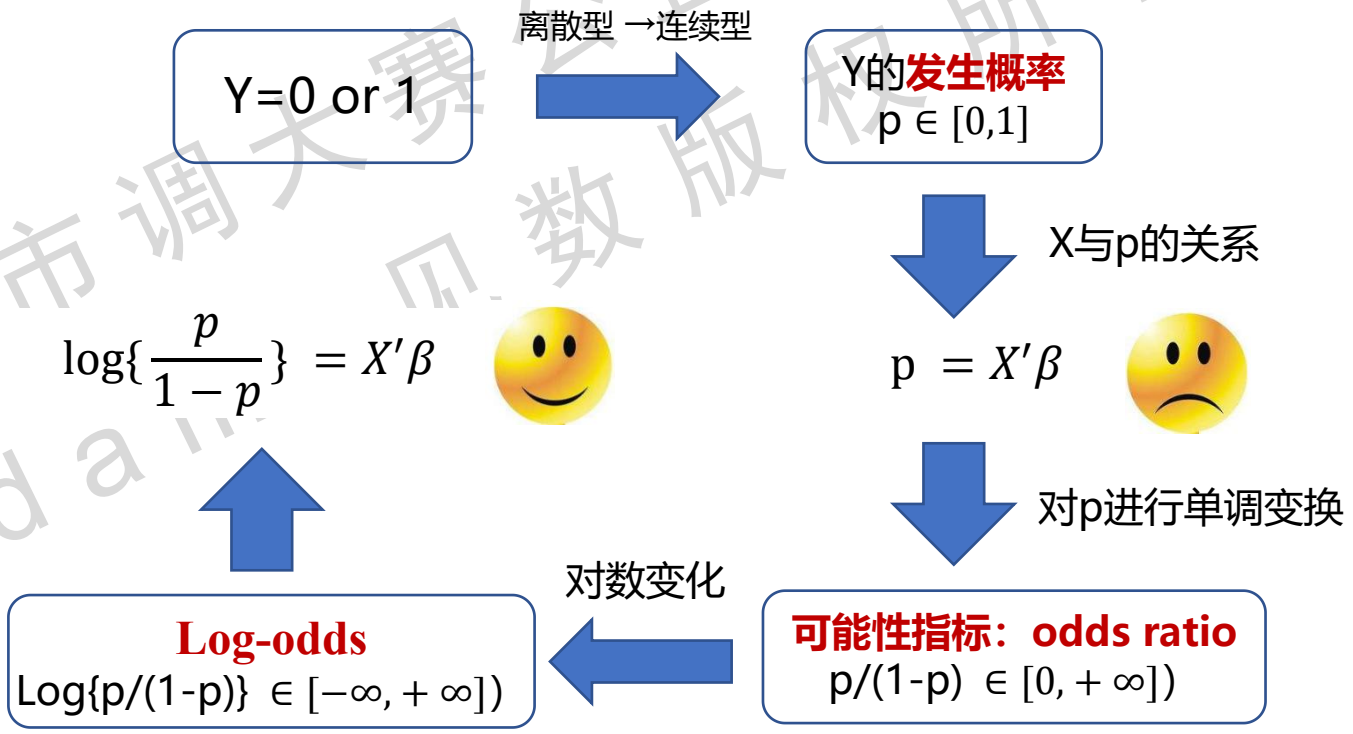
- 逻辑回归的自变量既可以是连续变量，也可以是分类变量。

- **应用**：常用于数据挖掘，疾病自动诊断，顾客判定，经济预测等领域。

逻辑回归

线性回归模型：$Y=\beta_0+\beta_1X_1+\ldots+\beta_pX_p+\varepsilon=X'\beta+\varepsilon$

逻辑回归模型：



逻辑回归

1. 对Y=1的概率进行建模，即：p=P(Y=1)=E(Y)
2. 对p进行logit变换，即：$Z = \log\{\frac{p}{1-p}\}$
3. 对Z建立线性回归模型，即$Z = X'\beta$

等价形式：$$P(Y=1) = p = \frac{\exp(Z)}{1+\exp(Z)} = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$$

逻辑回归

案例演示

目的：预测在手机广告中，广告的**趣味性、信息性、信任度和说服力**对**推荐意愿**的影响。

方法：利用Credamo的建模工具，在逻辑回归中因变量选中**推荐意愿**，自变量选中**趣味性、信息性、信任度和说服力**，系统即可自动生成逻辑回归模型。

逻辑回归

案例演示—结果分析（一）

- 首先可以关注模型的拟合情况，Deviance为偏差，Null_Deviance为无效偏差，无效偏差和剩余偏差之间的差异越大模型拟合越好。

- 其次，为了避免模型过度拟合，可以关注模型的**AIC值与BIC值**两个指标。建构回归模型时往往需要多次的回归分析，选取拥有更低的AIC值与BIC值的模型，有助于防止模型复杂度过高，缺乏普适性。

数据描述

下载数据

样本数	Deviance	Null_Deviance	Log-Likelihood	AIC值	BIC值
200	118.0352	213.2655	-59.0176	128.0352	-915.1367

逻辑回归

案例演示—结果分析（二）

参数摘要

下载数据

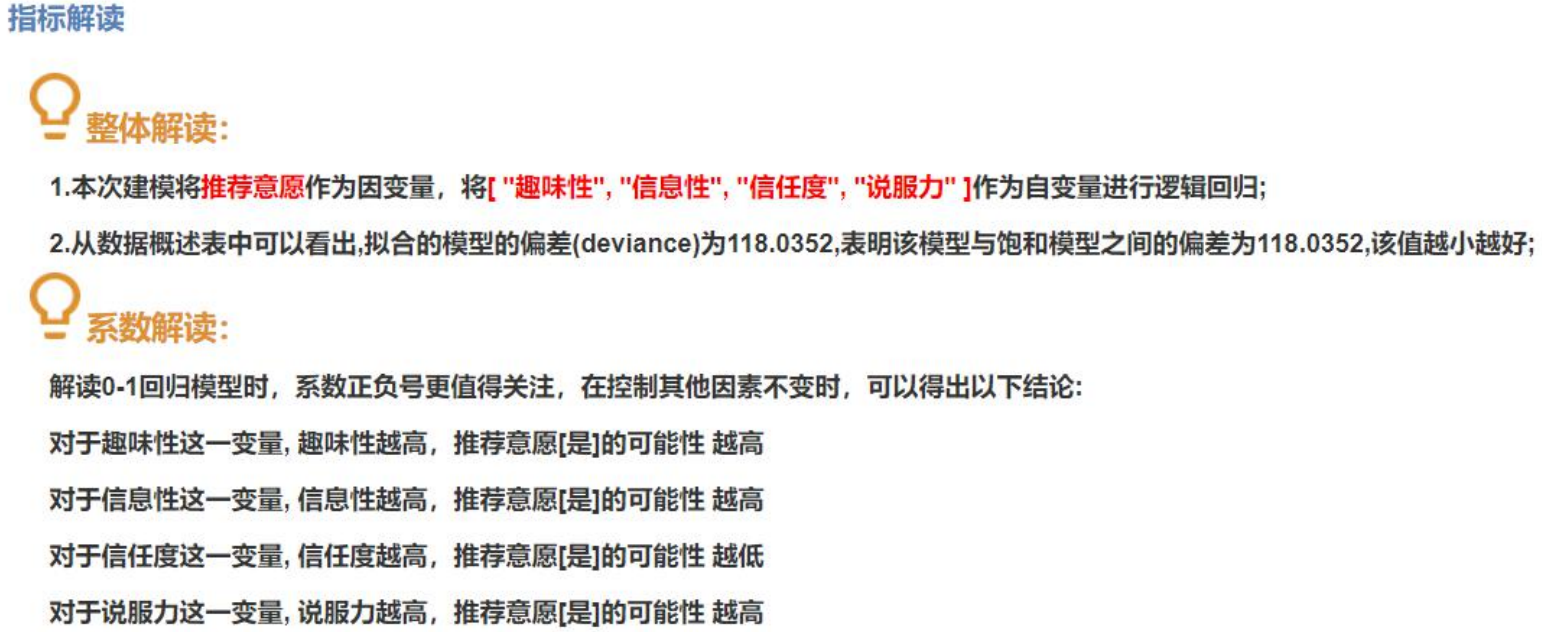
参数名称	系数	标准误	t值	P值	[0.025	0.975]
截距项	-8.8718	1.9210	-4.6183	0.0000	-12.6369	-5.1067
趣味性	0.6548	0.2053	3.1901	0.0014	0.2525	1.0572
信息性	0.3175	0.2558	1.2413	0.2145	-0.1838	0.8188
信任度	-0.3976	0.4715	-0.8431	0.3991	-1.3217	0.5266
说服力	1.5946	0.4651	3.4282	0.0006	0.6830	2.5063

- 对于模型中变量间的关系影响，可以关注参数**P值**与参数**系数**，P值说明变量关系的解释力度，系数正负解释了变量关系的正负影响，数值解释了影响的大小；P值小于0.05即可认为关系显著。

逻辑回归

案例演示—结果分析（三）

- 在Credamo见数上进行的建模分析可以**自动生成结果解读**；
- 由下图可知，在原假设是“推荐意愿为否”时，手机广告的**趣味性**、**信息性**和**说服力**正面影响消费者的**推荐意愿**，而广告的**信任度**负面影响消费者的推荐意愿。



2 中介效应

中介效应

- **中介效应（mesomeric effect）**并非一种分析方法，它是指自变量X对因变量Y的影响是通过变量M来实现的，也就是说M是X的函数，Y是M的函数，即X-M-Y。

- **中介变量（mediator）**是一个重要的统计概念，如果自变量X通过某一变量M对因变量Y产生一定影响，则称M为X和Y的中介变量。

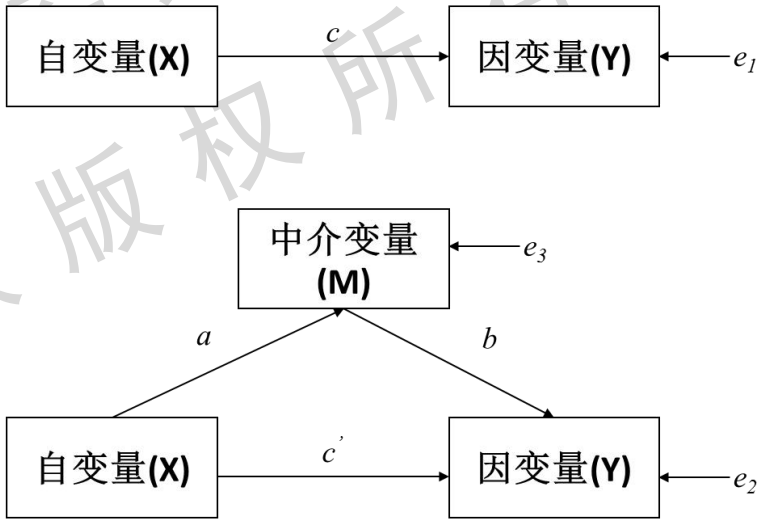
- **检验方法：**因果步骤法和系数乘积法

中介效应

因果步骤法

- 因果步骤法分为三步。第一，X对Y的回归，检验回归系数c的显著性；第二，X对M的回归，检验回归系数a的显著性；第三，X和M对Y的回归，检验回归系数b和c’的显著性。

- 如果系数c，a 和b都显著，就表示存在中介效应。此时如果系数c’不显著，就称这个中介效应是完全中介效应（full mediation）；如果回归系数c’显著，但c’<c，就称这个中介效应是部分中介效应（partial mediation）。中介效应的效果量（effect size）常用ab/c或ab/c’来衡量。



中介效应

案例演示

目的：预测在手机广告中，广告的**说服力**为中介变量时**信息性**与**购买意愿**的关系。

方法：在中介分析中因变量选中**购买意愿**，自变量选中**信息性**，中介变量选中**说服力**，系统即可自动生成中介效应分析模型。

中介效应

案例演示—结果分析（一）

- 对于模型中变量间的关系影响，可以关注参数**P值**与参数**系数**，P值说明变量关系的解释力度，系数说明变量关系的正负影响；P值小于0.05即可认为关系显著。

- 在模型1中，自变量信息性的P值明显小于0.05，且参数系数为0.5776，可认为信息性正向显著影响广告的说服力。

自变量对中介变量回归结果

下载数据

参数名称	系数	标准误	t值	P值	[0.025	0.975]	LLCI	ULCI
截距项	2.2430	0.2748	8.1630	0.0000	1.7011	2.7848	1.3490	3.1370
信息性	0.5776	0.0506	11.4257	0.0000	0.4779	0.6773	0.4232	0.7320

中介效应

案例演示—结果分析（二）

自变量、中介变量对因变量回归结果

下载数据

参数名称	系数	标准误	t值	P值	[0.025	0.975]	LLCI	ULCI
截距项	0.8779	0.5209	1.6855	0.0935	-0.1493	1.9051	-0.2762	2.0320
信息性	0.0480	0.0356	1.3484	0.1791	-0.0222	0.1182	-0.0138	0.1098
说服力	0.5603	0.0388	14.4258	0.0000	0.4837	0.6369	0.4755	0.6452

- 在模型2中，中介变量说服力的P值明显小于0.05，且参数系数为0.5603，可认为说服力正向显著影响购买意愿。信息性的P值为0.1791不显著，因此可认为说服力起到正向的完全中介效应。

中介效应

案例演示—结果分析（三）

间接效应Bootstrap检验结果

下载数据

参数名称	间接效应	Bootstrap标准误	LLCI	ULCI
信息性	0.4855	0.0732	0.3451	0.6334

- 使用Bootstrap检验方法时，主要关注的指标为LLCI（置信区间最低值）和ULCI（置信区间最高值），当LLCI与ULCI之间不包含0时，表明中介效应显著。

中介效应

案例演示—结果分析（四）

指标解读

逐步法

按照温忠麟 (2014) 提出的中介效应检验流程进行中介效应检验，在95%置信区间下，信息性对说服力的回归中，回归系数显著（p=0.0 < 0.05），且在信息性、说服力对购买意愿的回归中，说服力的回归系数显著（p=0.0 < 0.05）,表明说服力在信息性对购买意愿的影响中发挥了显著的间接效应。在信息性、说服力对购买意愿的回归中，信息性的回归系数不显著（p=0.4065 >= 0.05），表明控制了说服力之后，信息性对购买意愿的直接效应不显著。因此，说服力在信息性对购买意愿的影响中发挥了完全中介效应。

Bootstrap法

按照Zhao等（2010）提出的中介效应分析程序，参照Preacher和Hayes（2004）和Hayes（2013）提出的Bootstrap方法进行中介效应检验，样本量选择5000，在95%置信区间下，中介检验的结果不包含0（LLCI=0.4013，ULCI=1.0701）,表明在信息性对购买意愿的影响中，说服力的中介效应显著，且中介效应大小为0.7037。此外，控制了中介变量说服力之后，自变量信息性对因变量购买意愿的影响不显著，区间（LLCI=-0.2645，ULCI=0.5074）包含0。因此说服力在信息性对购买意愿的影响中发挥了中介作用，且是唯一的中介变量。

- 在自动生成的指标解读中，不论是逐步法还是Bootstrap法，都可以得到结论说服力在信息性对购买意愿的影响中发挥了正向的中介作用。

3 调节效应

调节效应

- **调节效应（moderating effect）**是指自变量X影响因变量Y时，受到调节变量M的干扰，即M会干扰X对Y的影响。

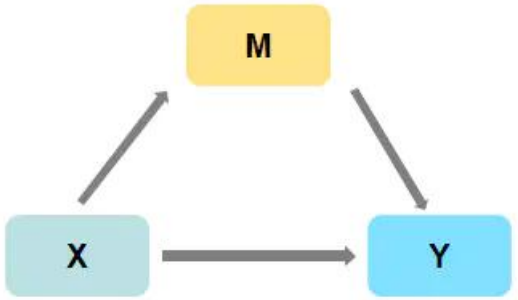
- **调节变量（moderator variable）**是指干扰自变量X对因变量Y影响的变量。调节变量可以是定性的(如性别、种族、学校类型等)，也可以是定量的(如年龄、受教育年限、刺激次数等)，它影响因变量和自变量之间关系的方向(正或负)和强弱。

调节效应

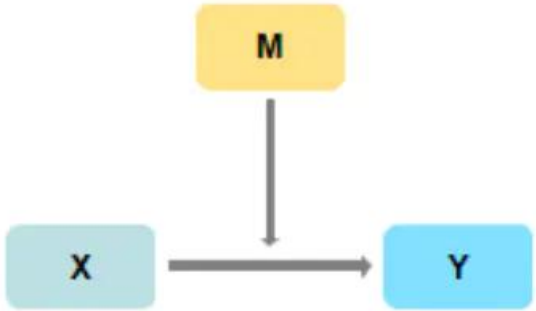
与中介效应的区别

- 中介作用是研究X对Y的影响时，是否会先通过中介变量M，再去影响Y；即是否有X->M->Y这样的关系，如果存在此种关系，则说明具有中介效应。比如工作满意度（X）会影响到创新氛围（M），再影响最终工作绩效（Y），此时创新氛围就成为了这一因果链当中的中介变量。如图①

- 调节作用是研究X对Y的影响时，是否会受到调节变量Z的干扰；比如开车速度（X）会对车祸可能性（Y）产生影响，这种影响关系受到是否喝酒（Z）的干扰，即喝酒时的影响幅度，与不喝酒时的影响幅度 是否有着明显的不一样。如图②



① 中介效应链路图



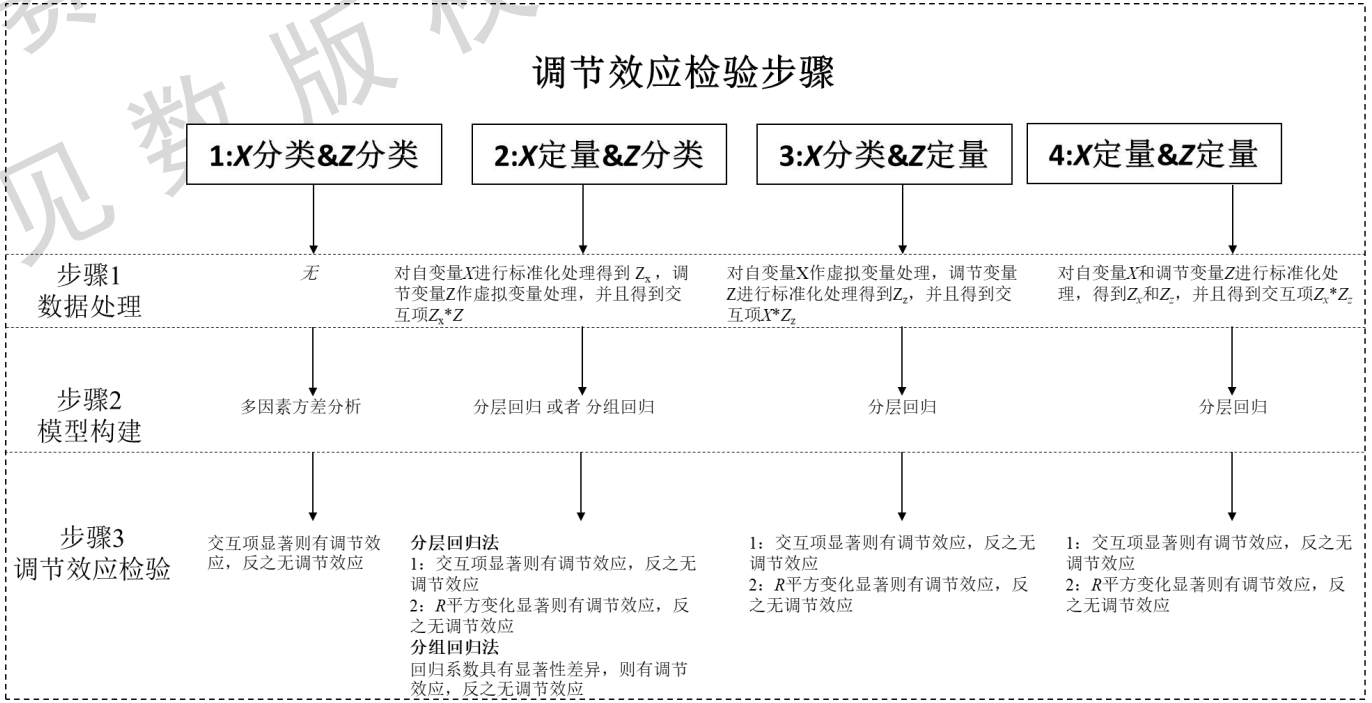
② 调节效应链路图

调节效应

调节效应分析步骤

- ① 识别X和M的数据类别，选择合适的研究方法。
- ② 调节效应检验
- ③ 进行分析

调节变量(Z)	自变量(X)	
	分类	定量
分类	多因素方差分析(ANOVA)	分层回归分析 或 分组回归
定量	分层回归分析	分层回归分析



调节效应

案例演示

目的：预测在手机广告中，购买者**性别**作为调节变量时**说服力**与**购买意愿**的关系。

方法：在回归分析中因变量选中**购买意愿**，自变量选中**说服力**和**性别**，系统即可自动生成回归分析模型。

调节效应

案例演示

方法：而后，在生成的模型估计中添加交互项**说服力**和**性别**。确认添加后即可生成含有交互项说服力×性别的回归模型。

调节效应

案例演示—结果分析（一）

数据描述

下载数据

样本数	R方	调整后的R方	F值	AIC值	BIC值
200	0.6637	0.6586	128.9415	477.1025	490.2958

- 首先关注模型的解释效力，重点关注**R方**指标。R方为模型的拟合优度，R方数值越接近1（一般认为0.5以上即可），说明模型的拟合程度与解释力越强。
- 其次，为了避免模型过度拟合，可以关注模型的**AIC值与BIC值**两个指标。建构回归模型时往往需要多次的回归分析，选取拥有更低的AIC值与BIC值的模型，有助于防止模型复杂度过高，缺乏普适性。

调节效应

案例演示—结果分析（二）

参数摘要

下载数据

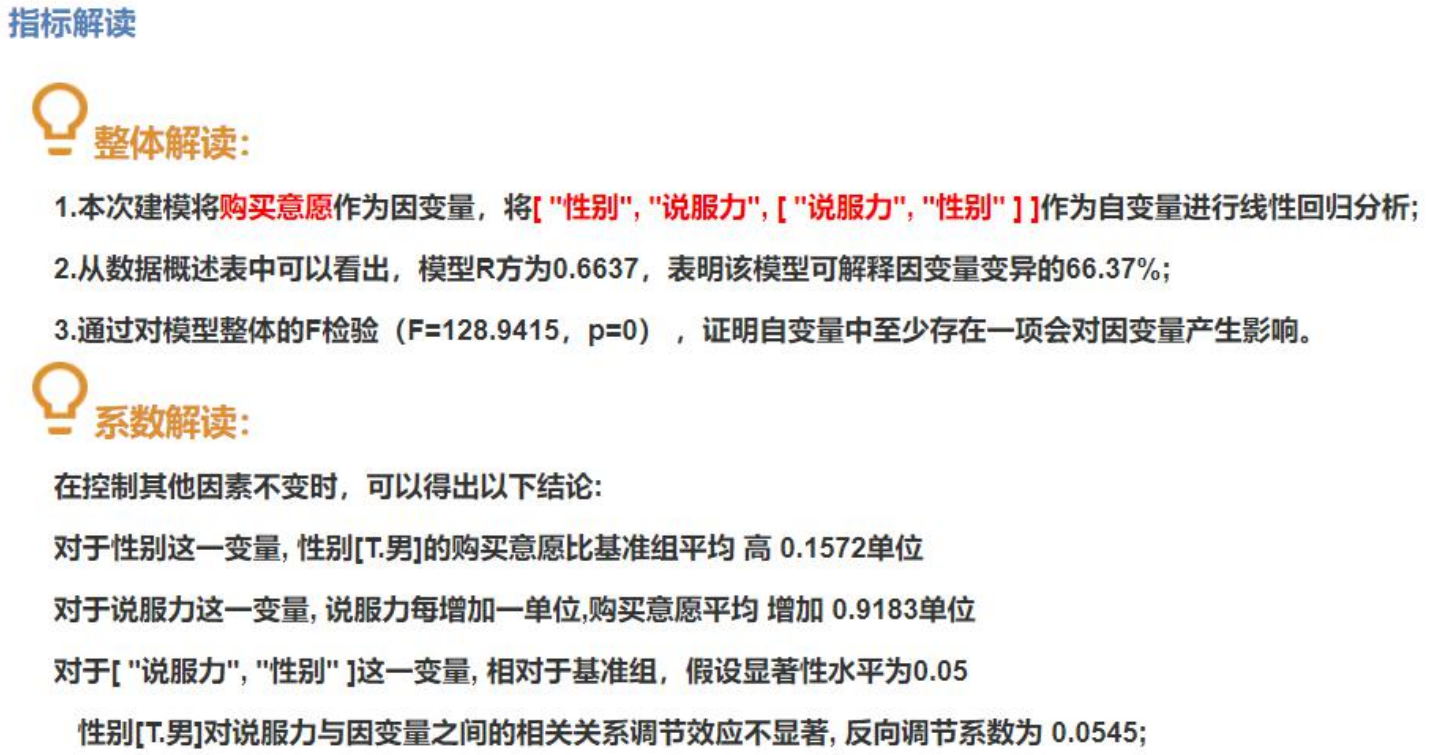
参数名称	系数	标准误	t值	P值	[0.025	0.975]
截距项	0.4709	0.3543	1.3291	0.1854	-0.2278	1.1695
性别[T.男]	0.1572	0.4926	0.3192	0.7499	-0.8142	1.1286
说服力	0.9183	0.0654	14.0442	0.0000	0.7894	1.0473
说服力:性别[T.男]	-0.0545	0.0908	-0.6006	0.5488	-0.2335	0.1245

- 在如上结果中，仅有变量**说服力**的P值小于0.05，而性别及性别与说服力的交互项的P值大于0.05，也可根据LLCI与ULCI的取值判断变量之间的关系是否显著。

调节效应

案例演示—结果分析（三）

- 由credamo自动生成的指标解读中可知，该模型有一定的解释力度，同时也说明了性别对娱乐性与购买意愿之间的相关关系调节效应**不显著**。



谢谢大家

关注右边公众号

及时了解大赛资讯和进程

随时学习大赛公益培训

中国商业统计学会

官方网址： http://www.china-cssc.org

公众号：

Credamo见数

官方网址： www.credamo.com

公众号：