《新媒体数据运营与分析》

Excel (10): 统计分析-回归分析（二）

教师：林志良
邮箱：linzhl@nfu.edu.cn
个人网站：www.zhilianglin.com

目录

- **Excel回归分析结果：**
  - 判定系数
  - 线性关系检验：F检验
  - 回归系数检验：t检验

目录

Excel回归分析结果

SUMMARY OUTPUT

回归统计	
Multiple R	0.44
R Square	0.20
Adjusted R Sq	0.13
标准误差	59.09
观测值	15

方差分析

	df	SS	MS	F	Significance F
回归分析	1	11100.44	11100.44	3.179037	0.10
残差	13	45392.9	3491.761		
总计	14	56493.33			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	558.28	90.44105	6.17283	0.00	362.8912439	753.6633	362.8912	753.6633
价格	-24.03	13.47956	-1.78299	0.10	-53.15467646	5.08696	-53.1547	5.08696

判定系数

回归模型诊断

回归系数检验

判定系数

- **R 方 (R Square)**：代表模型解释的方差比例。值越接近1，模型的解释力越强。如果 R 方为 0.8，这表示自变量解释了 80% 的因变量变化。【一元回归一般看此数值】
- **调整后的 R 方 (Adjusted R Square)**：与 R 方类似，但调整了自变量数量，适合比较多变量模型。该值也越高越好。【多元回归一般看此数值】

回归统计	
Multiple R	0.44
R Square	0.20
Adjusted R Sq	0.13
标准误差	59.09
观测值	15

R方

调整后R方

线性关系检验：F检验

- 方差分析表主要检验回归模型是否有意义。

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	11100.44	11100.44	3.179037	0.10
残差	13	45392.9	3491.761		
总计	14	56493.33			

回归模型检验的p值：

· $p < 0.05$ → 回归模型有意义

· $p > 0.05$ → 回归模型没有意义

线性关系检验：F检验

- **零假设与备择假设：**

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (所有回归系数都等于零)

H_1：不是所有回归系数都等于零。（回归模型成立（具有统计学显著意义）。/至少有一个自变量显著影响因变量。）

其中，k = 自变量个数

回归系数检验：t检验

- 对回归系数的估计及检验。

截距

斜率（回归系数）

截距是否显著检验的p值

(我们一般不管它)

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	558.28	90.44105	6.17283	0.00	362.8912439	753.6633	362.8912	753.6633
价格	-24.03	13.47956	-1.78299	0.10	-53.15467646	5.08696	-53.1547	5.08696

斜率（回归系数）是否显著检验的p值

· $p < 0.05$ → 结果显著

· $p > 0.05$ → 结果不显著

回归系数检验：t检验

- **零假设与备择假设：**

H_0: $\beta_1 = 0$ (存在线性关系)

H_1: $\beta_1 \neq 0$ (不存在线性关系)

多元回归方程

多元回归方程

用样本数据估计多元回归方程的回归系数

有k个自变量的多元回归方程

Y的预测值（估计值）

估计的截距

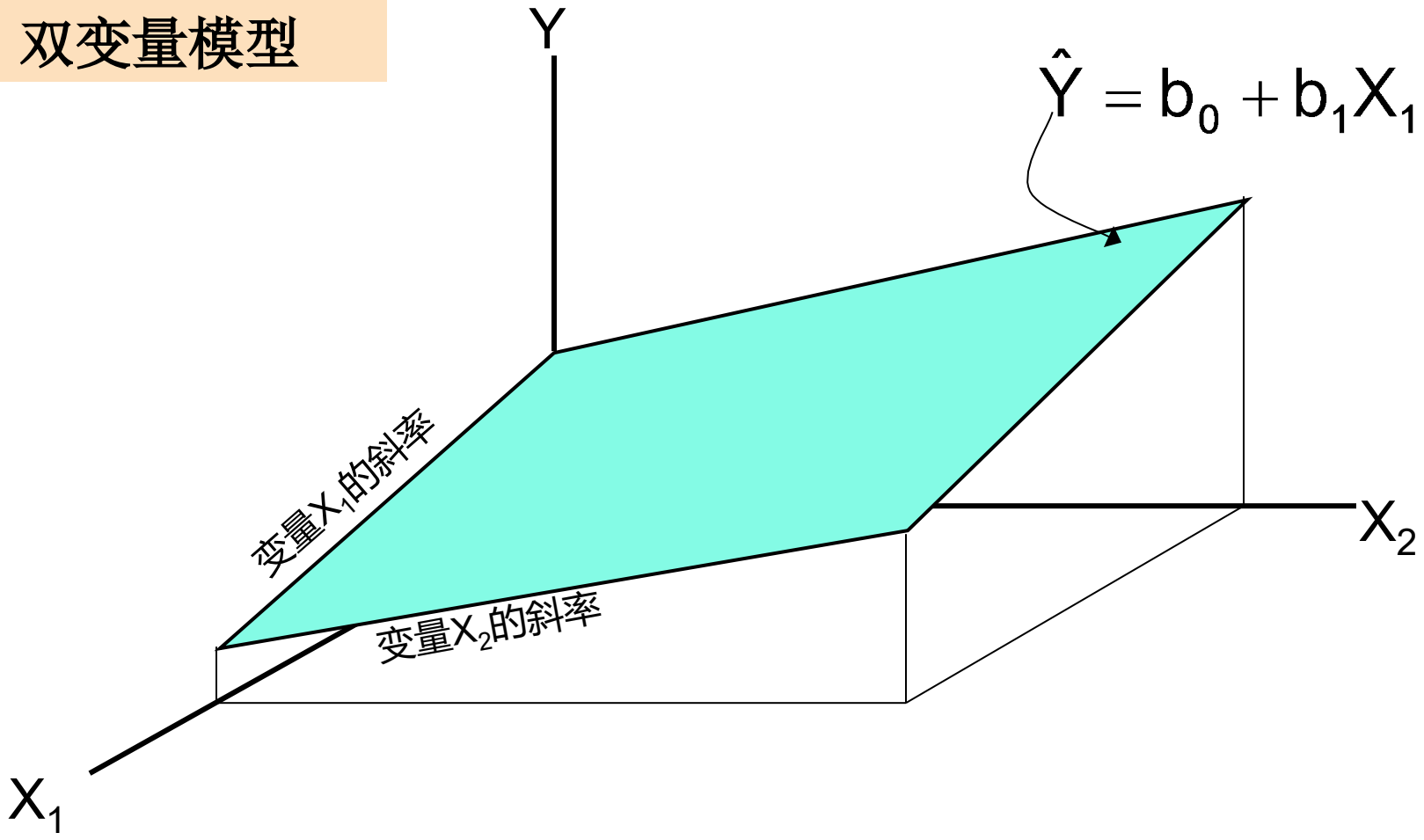
估计的回归系数

$$\hat{Y}_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

多元回归方程

多元回归方程

双变量模型



多元回归分析相对应的数据做可视化比较复杂，我们一般不做了

多元回归案例

周数	销量	价格 ($)	广告费 ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

馅饼销量

多元回归方程：

$$\widehat{\text{销量}} = b_0 + b_1\,(\text{价格}) + b_2\,(\text{广告费})$$

案例

判定系数：R^2 & 调整后的R^2

SUMMARY OUTPUT	
回归统计	
Multiple R	0.722134
R Square	0.521478
Adjusted R Square	0.441724
标准误差	47.46341
观测值	15

结论:

- 在考虑了样本量和自变量个数的前提下，馅饼价格、广告费可以解释销量44.17%的变异。

案例

线性关系检验：F检验

方差分析					
	df	SS	MS	F	Significance F
回归分析	2	29460.03	14730.01	6.538607	0.012006372
残差	12	27033.31	2252.776		
总计	14	56493.33			

结论:

回归模型成立（具有统计学显著意义）。/至少有一个自变量显著影响因变量。

案例

回归系数检验：t检验

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	306.5261933	114.2539	2.682851	0.019932	57.58834426	555.464	57.58834	555.464
价格($)	-24.97508952	10.83213	-2.30565	0.039788	-48.5762627	-1.37392	-48.5763	-1.37392
广告费($100s)	74.13095749	25.96732	2.854779	0.014494	17.55303206	130.7089	17.55303	130.7089

结论:

· **价格**：在控制其它变量的前提下，价格每提高1个单位销量平均下降24.98，结果显著

· **广告费**：在控制其它变量的前提下，广告费每提高1个单位销量平均增加74.13，结果显著

自变量为二分类型变量

- 当自变量为二分类型变量时，可以将自变量直接放进回归模型里面进行分析。

自变量为二分类型变量

二分类型变量回归系数解读

- 在回归结果中，**二分变量的系数表示相较于参考类别的差异效应**
- 假设变量“是否参加活动”被编码为 1（参加）和 0（不参加），并且回归系数结果如下：
  - **系数为 3.0**：这表示“参加活动”（1）与“不参加活动”（0）相比，因变量的平均值增加 3 个单位。
  - **负系数**：如果系数是负值（例如 -2.5），则表示“参加活动”与“不参加活动”相比，因变量的平均值减少 2.5 个单位。

自变量为二分类型变量

周数	销量	价格	广告费	假期
1	350	5.5	3.3	0
2	460	7.5	3.3	1
3	350	8	3	0
4	430	8	4.5	0
5	350	6.8	3	0
6	380	7.5	4	0
7	430	4.5	3	0
8	470	6.4	3.7	1
9	450	7	3.5	1
10	490	5	4	1
11	340	7.2	3.5	0
12	300	7.9	3.2	0
13	440	5.9	4	0
14	450	5	3.5	0
15	300	7	2.7	0

自变量为二分类型变量

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	317.24	74.33751	4.2675	0.00	153.6196	480.8511	153.6196	480.8511
价格	-22.60	7.066631	-3.1975	0.01	-38.1491	-7.04199	-38.1491	-7.04199
广告费	60.65	17.19178	3.528095	0.00	22.81538	98.4931	22.81538	98.4931
假期	76.70	18.39733	4.169041	0.00	36.20696	117.1915	36.20696	117.1915

在控制其它变量的前提下，假期（1）比非假期（0）平均高76.70，结果显著

自变量为分类型变量

- 当自变量为多分类型变量时，则需要将之处理成（n-1）个二分类型变量。

自变量为分类型变量

- **第一步**：将多分类型变量变成多个二分类型变量

F	G	H	I	J	K	L	M	N	O
领班			领班1	领班2	领班3		I列公式	J列公式	K列公式
							=IF(F2=1,1,0)	=IF(F2=2,1,0)	=IF(F2=3,1,0)
1			1	0	0				
1			1	0	0				
2			0	1	0				
2			0	1	0				
3			0	0	1				
3			0	0	1				
1			1	0	0				
1			1	0	0				
2			0	1	0				
2			0	1	0				
3			0	0	1				
3			0	0	1				
1			1	0	0				
1			1	0	0				
2			0	1	0				

自变量为分类型变量

- **第二步**：去掉其中一个二分类型变量（去掉的变量作为参照变量），将其它的二分类型变量放到数据集中。

周数	销量	价格	广告费	假期	领班1	领班2
1	350	5.5	3.3	0	1	0
2	460	7.5	3.3	1	1	0
3	350	8	3	0	0	1
4	430	8	4.5	0	0	1
5	350	6.8	3	0	0	0
6	380	7.5	4	0	0	0
7	430	4.5	3	0	1	0
8	470	6.4	3.7	1	1	0
9	450	7	3.5	1	0	1
10	490	5	4	1	0	1
11	340	7.2	3.5	0	0	0
12	300	7.9	3.2	0	0	0
13	440	5.9	4	0	1	0
14	450	5	3.5	0	1	0
15	300	7	2.7	0	0	1

自变量为分类型变量

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	236.70	82.14021	2.881696	0.02	50.88909	422.5172	50.88909	422.5172
价格	-13.89	8.243959	-1.68507	0.13	-32.5408	4.757443	-32.5408	4.757443
广告费	60.70	16.11043	3.767804	0.00	24.25662	97.14529	24.25662	97.14529
假期	67.09	18.43818	3.63861	0.01	25.37929	108.7994	25.37929	108.7994
领班1	44.41	23.58199	1.88317	0.09	-8.93728	97.75506	-8.93728	97.75506
领班2	22.82	20.96881	1.088357	0.30	-24.6132	70.25631	-24.6132	70.25631

· **领班1**：在控制其它变量的前提下，领班1的销量比领班3平均高44.41，结果不显著

· **领班2**：在控制其它变量的前提下，领班2的销量比领班3平均高22.82，结果不显著

参考资料

- Lizongzhang的个人空间-合集_·Excel 数据分析实战


谢谢！