

多元线性回归： 含有定性自变量

Multiple Linear Regression

王树佳 | 深圳大学经济学院

sjwang123@163.com

学习目标

当自变量中含有定性变量时，如何进行多元回归分析？

- 定性变量如何设置
- 如何解释系数
- 如果虚拟变量之间存在交互影响
- 如果定性变量有多个可能结果

内容

1. 自变量含有虚拟变量
 - A. 解释
 - B. 交互虚拟变量
 - C. 半对数模型的虚拟变量
2. 自变量含有多水平分类变量
 - A. 引入虚拟变量
 - B. 系数的解释

内容

1. 自变量含有虚拟变量
 - A. 解释
 - B. 交互虚拟变量
 - C. 半对数模型的虚拟变量
2. 自变量含有多水平分类变量
 - A. 引入虚拟变量
 - B. 系数的解释

虚拟变量

虚拟变量(Dummy Variable)：只取0和1两个值的变量

□ 也叫二值变量(Binary)，指示变量(Indicator)，0-1变量(Zero-One)

□ 政策评价：

A. 房价与二胎政策

$S=1$ ：二胎政策前； $S=0$ ：二胎政策后

B. 收入与大学类型

$X=1$ ：重点大学； $X=0$ ：非重点大学

C. 限牌与交通拥堵

$X=1$ ：限牌前； $X=0$ ：限牌后

例1：工资的影响因素

回归分析：wage 与 female

回归方程为

$$\text{wage} = 14.1 - 3.53 \text{ female}$$

自变量	系数	系数标准误	T	P
常量	14.1189	0.3025	46.68	0.000
female	-3.5252	0.4289	-8.22	0.000

S = 7.69995 R-Sq = 5.0% R-Sq (调整) = 4.9%

解释：Female系数=-3.53，表示女性的平均时薪比男性低3.53美元，该差异具有统计显著性。

- ✓ 上一章是如何研究Wage与性别之间关系的？
- ✓ 此结果是否足以证明存在性别歧视？

例1：工资的影响因素

回归分析: wage 与 female, nonwhite, union, education, exper
回归方程为

$$\text{wage} = -7.18 - 3.07 \text{ female} - 1.57 \text{ nonwhite} + 1.10 \text{ union} + 1.37 \text{ education} + 0.167 \text{ exper}$$

自变量	系数	系数标准误	T	P
常量	-7.183	1.016	-7.07	0.000
female	-3.0749	0.3646	-8.43	0.000
nonwhite	-1.5653	0.5092	-3.07	0.002
union	1.0960	0.5061	2.17	0.031
education	1.37030	0.06590	20.79	0.000
exper	0.16661	0.01605	10.38	0.000

S = 6.50814 R-Sq = 32.3% R-Sq (调整) = 32.1%

是否存在性别歧视？

Female系数=-3.07 : Female=1, Male=0

解释：表示在教育、经验等条件相同情况下，女性的时薪比男性平均少3.07美元，该差异具有高度统计显著性（T-检验统计量为-8.43，p-值为0）。因此，有理由认为存在性别歧视问题。

- ✓ 研究性别收入差距时，female是解释变量，其它变量称为**控制变量**。
- ✓ 男性(Male)称为**基准组或参考组**。
- ✓ **虚拟变量系数**=取值为1的组与基准组Wage之差

一般系数用截距解释

假设： $wage = \beta_0 + \beta_1 Female + \beta_2 Edu + \epsilon$

则

$$\text{基准组截距} = \beta_0$$

$$\text{Female组截距} = \beta_0 + \beta_1$$

$$\begin{aligned}\beta_1 &= E(wage|Femal = 1, edu) - E(wage|Femal = 0, edu) \\ &= (\beta_0 + \beta_1 + \beta_2 Edu) - (\beta_0 + \beta_2 Edu) \\ &= \text{截距之差}\end{aligned}$$

虚拟变量陷阱

能否建立如下模型？

$$wage = \beta_0 + \beta_1 Female + \beta_2 Male + \beta_3 Edu + \epsilon$$

其中：

Female=1，如果女性；Female=0，如果男性

Male=1，如果男性；Male=0，如果女性

□ 虚拟变量陷阱：

Male=1-Female，是线性关系

内容

1. 自变量含有虚拟变量

A. 解释

B. 交互虚拟变量

C. 半对数模型的虚拟变量

2. 自变量含有多水平分类变量

A. 引入虚拟变量

B. 系数的解释

交互虚拟变量

- 女性每小时收入显著低于男性
- 但是，白人女性和非白人女性的收入一样吗？
- 加入一个变量：女性，非白人。

- 交互虚拟变量(Interactive Dummy):
两个虚拟变量的乘积： $\text{Female} * \text{Nonwhite}$

例1：工资的影响因素

回归分析: wage 与 female, nonwhite, union, education, exper, femalenonw

回归方程为

$$\text{wage} = -7.09 - 3.24 \text{ female} - 2.16 \text{ nonwhite} + 1.12 \text{ union} + 1.37 \text{ education} + 0.166 \text{ exper} + 1.10 \text{ femalenonw}$$

自变量	系数	系数标准误	T	P
常量	-7.089	1.019	-6.95	0.000
female	-3.2401	0.3953	-8.20	0.000
nonwhite	-2.1585	0.7484	-2.88	0.004
union	1.1150	0.5064	2.20	0.028
education	1.37011	0.06590	20.79	0.000
exper	0.16586	0.01606	10.33	0.000
femalenonw	1.095	1.013	1.08	0.280

S = 6.50771 R-Sq = 32.4% R-Sq (调整) = 32.1%

解释

- 基准组为男性白人($Femal=0, nonw=0$) , 截距= -7.09
- 女性白人($F=1, Nw=0$) : 截距= $-7.09-3.24$
 - ✓ 平均工资比基准组低3.24美元
- 非白种男性 ($F=0, Nw=1$) : 截距= $-7.09-2.16$
 - ✓ 平均工资比基准低2.16美元
- 交互虚拟变量 $femalenonw$ 的系数约为1.10 , 但是它在统计上并不显著 , 因为 $p\text{-值}=0.28 > 0.05$
- 非白种女性($F=1, Nw=1$) : 截距= $-7.09-3.24-2.16+1.10$
 - ✓ 平均工资比基准低 : $-3.24-2.16+1.10=-4.30$ 美元
 - ✓ 换句话说 , 与基准组相比 , 非白人女性挣得的平均工资要比单纯是女性或单纯是非白人更低

内容

1. 自变量含有虚拟变量

A. 解释

B. 交互虚拟变量

C. 半对数模型的虚拟变量

2. 自变量含有多水平分类变量

A. 引入虚拟变量

B. 系数的解释

例1：半对数模型的虚拟变量

回归分析: lnwage 与 female, nonwhite, union, education, exper
回归方程为

$$\lnwage = 0.906 - 0.249 \text{ female} - 0.134 \text{ nonwhite} + 0.180 \text{ union} + 0.0999 \text{ education} \\ + 0.0128 \text{ exper}$$

自变量	系数	系数标准误	T	P
常量	0.90550	0.07417	12.21	0.000
female	-0.24915	0.02663	-9.36	0.000
nonwhite	-0.13354	0.03718	-3.59	0.000
union	0.18020	0.03695	4.88	0.000
education	0.099870	0.004812	20.75	0.000
exper	0.012760	0.001172	10.89	0.000

S = 0.475237 R-Sq = 34.6% R-Sq (调整) = 34.3%

不太准确的解释

$$\text{Inwage} = 0.906 - 0.249 \text{ female} - 0.134 \text{ nonwhite} + 0.180 \text{ union} + 0.0999 \text{ education} + 0.0128 \text{ exper}$$

1. 教育回报：教育年限每增加1年，在其他条件不变情况下，工资增长9.99%；
2. 工作经验回报：工作经验每增加1年，在其他条件不变情况下，工资增长1.28%；
3. 性别差异：女性工资比男性低24.9%；
4. 种族差异：非白人工资比白人低13.4%；
5. 工会回报：工会成员比非工会成员平均工资低18%。

虚拟变量系数更准确的解释

推导公式：

$$\ln(wage_F) - \ln(wage_M) = -0.249$$

$$\frac{wage_F - wage_M}{wage_M} = e^{-0.249} - 1 \approx -0.221$$

即女性工资比男性平均低22.1%。

一般情形：

如果因变量为 $\ln y$ ，虚拟变量系数为 β ，则虚拟变量1和0时，因变量的相对差异为百分之
 $100(e^\beta - 1)$

内容

1. 自变量含有虚拟变量
 - A. 解释
 - B. 交互虚拟变量
 - C. 半对数模型的虚拟变量
2. 自变量含有多水平分类变量
 - A. 引入虚拟变量
 - B. 系数的解释

例2：上市公司的价值评估

英国一家智囊机构希望建立一个“公司估值”模型来评估一个公司的价值。该机构收集了30家公司的相关信息。

变量名称

VALUE：公司价值（百万英镑）

SIZE：公司的规模（雇员人数，单位：千人）

PE：公司的市盈率

PROFIT：税前利润（百万英镑）

MN：企业类型：MN=0表示本国企业，MN=1多国企业

SECTOR：公司经营行业（1=农业，2=工业，3=休闲业，
4=金融业）

多水平定性变量如何引入虚拟变量？

以公司价值(VALUE)为因变量，以行业变量(SECTOR)为自变量进行线性回归，看看结果如何

回归方程为

$$\text{VALUE} = 116 - 20.2 \text{ SECTOR}$$

自变量	系数	系数标准误	T	P
常量	116.186	9.419	12.34	0.000
SECTOR	-20.208	3.518	-5.74	0.000

S = 21.5075 R-Sq = 54.1% R-Sq (调整) = 52.4%

方差分析

来源	自由度	SS	MS	F	P
回归	1	15259	15259	32.99	0.000
残差误差	28	12952	463		
合计	29	28211			

多水平定性变量如何引入虚拟变量？

线性回归方程为： $VALUE = 116 - 20.2 SECTOR$

方差分析的 F -检验和行业变量(SECTOR)的 t -检验的 p -值都是0.000，说明回归方程高度显著。

回归方程的可决系数为： $R-Sq = 54.1\%$ 。

回归系数的意义：

行业代码越大，公司价值就越低。当行业变量(SECTOR)的值每增加一个单位时，公司的价值减少20.2（百万英镑），即相邻两个行业的平均价值之差为（-20.2）（百万英镑）。

□ 以上回归方程合理吗？

为什么前面的方法是错误的？

描述性统计：VALUE

变量	SECTOR	平均值					
		平均值	标准误	标准差	中位数	偏度	峰度
VALUE	1	106.30	5.01	14.17	105.91	0.32	-1.41
	2	50.16	3.59	10.15	47.41	0.28	-2.27
	3	78.72	3.29	8.72	73.41	0.57	-1.99
	4	29.67	2.63	6.96	29.84	0.42	-1.48

- 四个行业的平均价值依次为：
106.3 , 50.16 , 78.72 , 29.67
- 相邻两个行业的平均市值之差有明显差异。
- 行业变量(SECTOR)的回归系数 (-20.2) 显然不合理。

如何引入虚拟变量？

因为行业变量(SECTOR)有四个水平，我们需要引入3个虚拟变量，分别为：

- S1 : $S1=1$ ，如果SECTOR=1，否则， $S1=0$ ；
- S2 : $S2=1$ ，如果SECTOR=2，否则， $S2=0$ ；
- S3 : $S3=1$ ，如果SECTOR=3，否则， $S3=0$ 。

基准组：SECTOR=4 意味着 $S1=0$ ， $S2=0$ ， $S3=0$

□ 如果引入4个，将导致**虚拟变量陷阱**。

内容

1. 自变量含有虚拟变量
 - A. 解释
 - B. 交互虚拟变量
 - C. 半对数模型的虚拟变量
2. 自变量含有多水平的分类变量
 - A. 引入虚拟变量
 - B. 系数的解释

虚拟变量系数解释

假设关于三个虚拟变量的回归方程为

$$VALUE = b_0 + b_1 S1 + b_2 S2 + b_3 S3$$

以上方程实际上包含了4个方程，即

$$\begin{aligned} E(VALUE | SECTOR=1) &= b_0 + b_1(1) + b_2(0) + b_3(0) \\ &= b_0 + b_1 \end{aligned}$$

$$\begin{aligned} E(VALUE | SECTOR=2) &= b_0 + b_1(0) + b_2(1) + b_3(0) \\ &= b_0 + b_2 \end{aligned}$$

$$\begin{aligned} E(VALUE | SECTOR=3) &= b_0 + b_1(0) + b_2(0) + b_3(1) \\ &= b_0 + b_3 \end{aligned}$$

$$\begin{aligned} E(VALUE | SECTOR=4) &= b_0 + b_1(0) + b_2(0) + b_3(0) \\ &= b_0 \text{ (基准组)} \end{aligned}$$

虚拟变量系数解释

- b_0 : 截距=基准组($S_1=S_2=S_3=0$)
- b_1 : S_1 与基准组之差
- b_2 : S_2 与基准组之差
- b_3 : S_3 与基准组之差

以公司市值(VALUE)为因变量，以三个虚拟变量S1、S2和S3为自变量进行线性回归

回归方程为

$$\text{VALUE} = 29.7 + 76.6 \text{ S1} + 20.5 \text{ S2} + 49.1 \text{ S3}$$

自变量	系数	系数标准误	T	P
常量	29.669	3.973	7.47	0.000
S1	76.626	5.441	14.08	0.000
S2	20.493	5.441	3.77	0.001
S3	49.053	5.619	8.73	0.000

S = 10.5123 R-Sq = 89.8% R-Sq (调整) = 88.6%

方差分析

来源	自由度	SS	MS	F	P
回归	3	25338.1	8446.0	76.43	0.000
残差误差	26	2873.2	110.5		
合计	29	28211.3			

模型结果

根据Minitab的输出结果，线性回归方程为：

$$\text{VALUE} = 29.7 + 76.6 S1 + 20.5 S2 + 49.1 S3$$

方差分析的F-检验和三个虚拟变量的t-检验的p-值都远小于0.05，说明回归方程各个系数均为高度显著。

回归方程的复可决系数为：R-Sq = 89.8%，说明方程拟合程度很好。

□ 比用行业变量(SECTOR)为一个自变量的回归方程的拟合情况 (R-Sq = 54.1%) 大为改善。

虚拟变量系数解释

- 基准组：金融业平均市值=29.7；
- 农业(SECTOR=1)：平均市值比金融业高76.6；
- 工业(SECTOR=2)：平均市值比金融业高20.5；
- 休闲业(SECTOR=3)：平均市值比金融业高49.1。

例2：上市公司价值评估：全部变量

回归分析：VALUE 与 SIZE, PE, PROFIT, MN, SECTOR_1, SECTOR_2, SECTOR_

回归方程为

$$\text{VALUE} = 38.0 + 0.55 \text{ SIZE} - 1.40 \text{ PE} + 1.11 \text{ PROFIT} - 2.11 \text{ MN} + 44.3 \text{ SECTOR}_1 + 8.78 \text{ SECTOR}_2 + 24.5 \text{ SECTOR}_3$$

自变量	系数	系数标准误	T	P
常量	37.96	13.69	2.77	0.011
SIZE	0.553	1.042	0.53	0.601
PE	-1.3988	0.3321	-4.21	0.000
PROFIT	1.1063	0.4092	2.70	0.013
MN	-2.113	7.036	-0.30	0.767
SECTOR_1	44.28	10.27	4.31	0.000
SECTOR_2	8.783	5.782	1.52	0.143
SECTOR_3	24.535	7.875	3.12	0.005

不显著，系数如何解释？

S = 7.93923 R-Sq = 95.1% R-Sq (调整) = 93.5%

例2：上市公司价值评估：显著变量

回归分析：VALUE 与 PE, PROFIT, SECTOR_1, SECTOR_2, SECTOR_3

回归方程为

$$\text{VALUE} = 38.6 - 1.44 \text{ PE} + 1.22 \text{ PROFIT} + 43.7 \text{ SECTOR}_1 + 7.05 \text{ SECTOR}_2 + 24.2 \text{ SECTOR}_3$$

自变量	系数	系数标准误	T	P
常量	38.606	9.383	4.11	0.000
PE	-1.4389	0.3110	-4.63	0.000
PROFIT	1.2155	0.3514	3.46	0.002
SECTOR_1	43.651	8.276	5.27	0.000
SECTOR_2	7.050	4.809	1.47	0.156
SECTOR_3	24.248	6.604	3.67	0.001

S = 7.66298 R-Sq = 95.0% R-Sq (调整) = 94.0%

Recap

1. 自变量含有虚拟变量
 - A. 解释
 - B. 交互虚拟变量
 - C. 半对数模型的虚拟变量
2. 自变量含有多水平分类变量
 - A. 引入虚拟变量
 - B. 系数的解释

讨论：什么是毕业生收入的决定性因素？

介绍：为了研究毕业生工作后收入情况及影响因素，英国某商学院在2004年针对1999年毕业的大学生进行了一项调查，以研究他们毕业5年后的情况变化。

抽样：把所有1999年毕业的商学院学生的名字和地址列表（包括全日制学生和在职学生）作为抽样框，然后按照随机数字表，随机抽取了400名毕业生进行问卷调查。

调查：首先设计好一份问卷，包括毕业情况、毕业后的第一份工作和现在的工作情况。然后写一封信，说明调研的目的以及请求被访者的合作。把信和问卷一并寄给被抽中的毕业生。如果6个星期后没有回复，再次寄出一封信和问卷，并电话询问。

数据编码

名称	说明 (DESCRIPTION)
ID-CODE	编号
GENDER	性别: 0=Female, 1=Male
MONTHS	毕业后几个月开始第一份工作
DEGREE	毕业等级: 1 - First Class Honours (一等荣誉学士学位) 2 - Upper Second Class Honours (二级甲等荣誉学士学位) 3 - Lower Second Class Honours (二级乙等荣誉学士学位) 4 - Third Class Honours (三等荣誉学士学位) 5 - Other.
TYPE	是否工读交替制课程 (sandwich course)? 0 = Yes, 1 = No.

数据编码

名称	说明 (DESCRIPTION)
SAL99	第一份工作的年薪
JOB99	第一份工作所从事的行业 (INDUSTRY) 行业 (INDUSTRY) 编号见后面
AREA99	第一份工作所在地区 (REGION) 地区 (REGION) 编号见后面
SAL04	2004年年薪
JOB04	2004年所从事的行业 (INDUSTRY)
AREA04	2004年工作所在地区 (REGION)
NO-JOBS	2004年前转换工作的次数

数据编码

REGION

1	Northern England.
2	Yorks/Humberside.
3	North West England.
4	East Anglia.
5	East Midlands.
6	West Midlands.
7	Greater London.
8	South East [Excl GL].
9	South West.

INDUSTRY

1	Manufacturing.
2	Public Utilities.
3	Retailing.
4	Financial Sector.
5	Public Administration.
6	Education.
7	Higher Degree.

问题

把Salary04作为因变量，把GENDER, MONTHS, TYPE, Salary99, NOJOBS, DEGREE, JOB04和 AREA04等其它影响因素作为自变量，建立多元线性回归模型。

1. 如何在Minitab中把Degree，JOB04等多水平定性变量转换为虚拟变量？
2. 建立、求解并解释多元线性回归模型；
3. 剔除不显著的自变量，把其它显著自变量留在方程，重新求解。结果如何？