多元回归模型

目录 CONTENTS

01

多元回归模型

4 多元回归模型

多元回归分析是研究因变量y 如何依赖两个或两个以上自变量的问题.

多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

多元回归方程

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

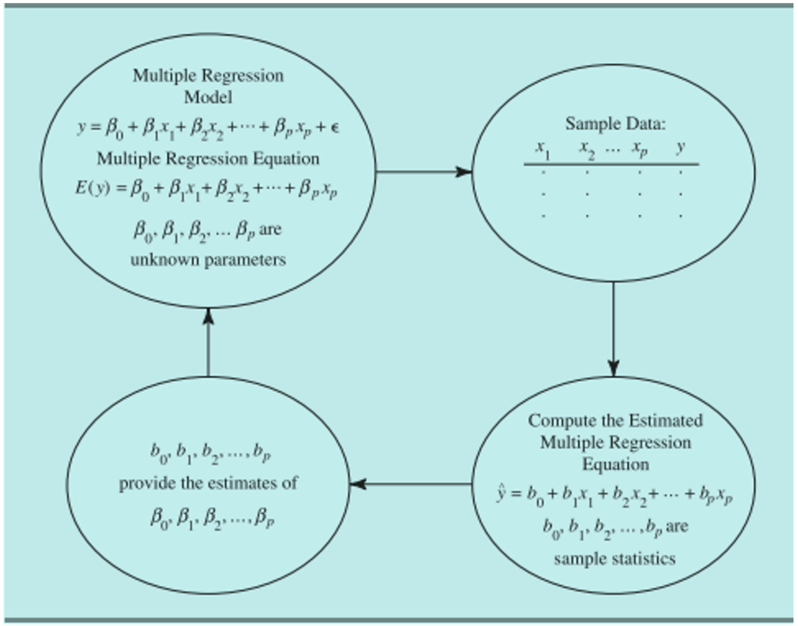
5

估计的多元回归方程

估计的多元回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

多元回归模型的估计步骤



6 最小二乘法

- 最小二乘法利用样本数据，通过使因变量的观测值y_i 与因变量的预测值$\hat{y}_i$之间的差的平方和达到最小的方法求得b_0，b_1，…， b_p的值。
- 但是由于计算过于复杂，超出我们目前知识掌握的范围。我们将重点介绍使用PYTHON求解的过程
- 后面高级的章节，我们会学习使用梯度下降以及矩阵计算的方式求解

最小二乘法准则

注：德国数学家高斯提出的

$$min\sum(y_i-\hat{y}_i)^2$$

式中，y_i为对于第i次观测因变量的观测值；$\hat{y}_i$为对于第i次观测因变量的预测值

7 案例分析：Butler运输公司

Butler运输公司面临的一个问题：

管理人员希望估计司机每天行驶的时间。管理人员认为司机的行驶时间可能与每天运送货物行驶的里程以及运送货物的次数有关系

Butler运输公司的数据

Driving Assignment	x_1 = Miles Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

8 案例分析：Butler运输公司

```
import pandas as pd
from statsmodels.formula.api import ols
df = pd.read_csv("C:/Users/feyman/Desktop/Butler.CSV")

df_model = ols("Time ~ Miles + Deliveries", data=df).fit()
print(df_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Time   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.876
Method:                 Least Squares   F-statistic:                     32.88
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           0.000276
Time:                        15:02:10   Log-Likelihood:                -6.8398
No. Observations:                  10   AIC:                             19.68
Df Residuals:                       7   BIC:                             20.59
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.8687      0.952     -0.913      0.392      -3.119       1.381
Miles          0.0611      0.010      6.182      0.000       0.038       0.085
Deliveries     0.9234      0.221      4.176      0.004       0.401       1.446
==============================================================================
Omnibus:                        0.039   Durbin-Watson:                   2.515
Prob(Omnibus):                  0.981   Jarque-Bera (JB):                0.151
Skew:                           0.074   Prob(JB):                        0.927
Kurtosis:                       2.418   Cond. No.                         435.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

$$\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2$$

02

多元判定系数

10 多元判定系数

多元判定系数

$$R^2 = \frac{SSR}{SST}$$

- 案列中的R^2=0.904，表示的是运输车辆行驶时间y中变异性的90.4%，能用运送货物的行驶里程和运送货物的次数作为自变量的估计的多元回归方程解释
- 对于仅有一个自变量，既每天运送货物的行驶里程的估计的多元回归方程，得出的R^2的值是66.41%。于是，当运送货物的次数作为第二个自变量进入模型后，运输车辆形式时间y的变异性中能被估计的多元回归方程解释的百分比由66.41%增加到90.38%
- 由于增加自变量将会影响到因变量中的变异性被估计的回归方程解释的百分比，为了避免高估这一影响，提出用自变量的数目去修正R^2的值。
- **修正多元判定系数：**

$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$，**n表示观测值的数目，p表示自变量的数目**

对于Butler公司的例子：R_a^2=0.8763

03

显 著 性 检 验

12 模型的假定

判定系数r^2即使较大，我们也需要对假定模型的合理性做出进一步的分析。确定是否合理的一个步骤：是要对变量之间关系的显著性进行检验。回归分析中的显著性检验是以对误差项ε的下列假定为依据进行的：

1. 误差项ε是一个平均值或期望值为0的随机变量，即$E(\varepsilon)=0$
2. 对所有的x值，ε的方差都是相同的。我们用σ^2表示ε的方差
3. ε的值是相互独立的
4. 对所有特定的x值，误差项ε是一个正态分布的随机变量

13 显 著 性 检 验

1. F检验用于确定在因变量和所有自变量之间是否存在一个显著的关系，我们把F检验称为总体的显著性检验。
2. 如果F检验已经表明了模型总体的显著性，那么t检验用来确定每一个单个的自变量是否为一个显著的自变量。对模型中的每一个单个的自变量，都要单独地进行t检验，我们把每一个这样的t检验都称为单个的显著性检验。

14

F 检 验

多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- F检验的假设与多元回归模型的参数有关：
  - $\boldsymbol{H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0}$
  - $\boldsymbol{H_a}$: 至少**有一个参数不等于零**
- 均方是一个平方和除以它所对应的自由度
  - 总的平方和有n-1个自由度
  - 回归平方和SSR有p个自由度
  - 误差平方和SSE有n-p-1个自由度
  - 因此，均方回归MSR是SSR/p，均方误差MSE是SSE/(n-p-1)

15 F 检验

- MSE给出了误差项ϵ的方差σ^2的一个无偏估计量。如果原假设H_0：$\beta_1=\beta_2=\cdots=\beta_p=0$成立，MSR也给出了$\sigma^2$的一个无偏估计量，并且MSR/MSE的值将接近于1。但是如果原假设H_0被拒绝，MSR将高估σ^2，这时MSR/MSE的值将变得比较大
- 如果H_0成立并且有关多元回归模型的假定都成立，那么MSR/MSE的抽样分布是一个分子的自由度为p，分母的自由度为n-p-1的F分布

总体显著性的F检验

F TEST FOR OVERALL SIGNIFICANCE

H_0: $\beta_1=\beta_2=\cdots=\beta_p=0$

H_a: One or more of the parameters is not equal to zero

TEST STATISTIC

$$F=\frac{\text{MSR}}{\text{MSE}} \tag{15.14}$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\le \alpha$
Critical value approach:	Reject H_0 if $F \ge F_\alpha$

where F_α is based on an F distribution with p degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator.

16

案 例 F 检 验

```python
import pandas as pd
from statsmodels.formula.api import ols
df = pd.read_csv("C:/Users/feyman/Desktop/Butler.CSV")

df_model = ols("Time ~ Miles + Deliveries", data=df).fit()
print(df_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Time   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.876
Method:                 Least Squares   F-statistic:                     32.88
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           0.000276
Time:                        15:02:10   Log-Likelihood:                -6.8398
No. Observations:                  10   AIC:                             19.68
Df Residuals:                       7   BIC:                             20.59
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.8687      0.952     -0.913      0.392      -3.119       1.381
Miles          0.0611      0.010      6.182      0.000       0.038       0.085
Deliveries     0.9234      0.221      4.176      0.004       0.401       1.446
==============================================================================
Omnibus:                        0.039   Durbin-Watson:                   2.515
Prob(Omnibus):                  0.981   Jarque-Bera (JB):                0.151
Skew:                           0.074   Prob(JB):                        0.927
Kurtosis:                       2.418   Cond. No.                         435.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

假设检验

- $H_0: \beta_1 = \beta_2 = 0$
- H_a: 至少有一个参数不等于零

检 验 量

- 统计量F=32.88
- p-value = 0.000276 小于 显著性检验水平0.01，所以我们应该拒绝原假设

结 论

在每天行驶的时间y和每天运送货物的行驶里程x_1、运送货物的次数x_2这两个自变量之间存在一个显著的关系。

17 t 检 验

t TEST FOR INDIVIDUAL SIGNIFICANCE

For any parameter β_i

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

REJECTION RULE

p-value approach: Reject H_0 if *p*-value $\leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a *t* distribution with $n - p - 1$ degrees of freedom.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.8687	0.952	-0.913	0.392	-3.119	1.381
Miles	0.0611	0.010	6.182	0.000	0.038	0.085
Deliveries	0.9234	0.221	4.176	0.004	0.401	1.446

t 检 验

如果F检验显示了多元回归关系在总体上是显著的,那么t检验就能帮助我们确定每一个单个参数的显著性问题

结 论

从计算结果的数据来看，我们发现β_1的p-value是0.000，β_2的p-value是0.004 均小于显著性水平0.01

因此我们拒绝原假设$H_0: \beta_1 = 0$ 和$H_0: \beta_2 = 0$

04

分 类 自 变 量

19 分类自变量

目前我们定义的自变量都是连续变量。有些情形下我们必须使用分类自变量，比如性别，付款方式等等。

案例

约翰逊过滤股份有限公司想要预估每次帮客户维修所需要的时间，管理人员认为维修时间依赖于两个因素：一个是从最近一次维修服务至今水过滤系统已经使用的时间，另一个是需要维修的故障类型。

约翰逊过滤公司的数据

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	electrical	2.9
2	6	mechanical	3.0
3	8	electrical	4.8
4	3	mechanical	1.8
5	2	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrical	4.5

20 分类自变量

我们定义下面的变量：

$$x_2 = \begin{cases} 0, & \text{机械} \\ 1, & \text{电子} \end{cases}$$

在回归分析中，$\boldsymbol{x_2}$**被称为虚拟变量或指标变量**，我们可以把多元回归模型写成如下形式：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

约翰逊过滤公司的数据

我们定义下面的变量：

$$x_2 = \begin{cases} 0, & \text{机械} \\ 1, & \text{电子} \end{cases}$$

在回归分析中，$\boldsymbol{x_2}$**被称为虚拟变量或指标变量**，我们可以把多元回归模型写成如下形式：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

21 PYTHON解析

```
import pandas as pd
from statsmodels.formula.api import ols
df = pd.read_csv("C:/Users/feyman/Desktop/Johnson.CSV")
df_model = ols("Time ~ Months  + C(Type)", data=df).fit()
print(df_model.summary())
```

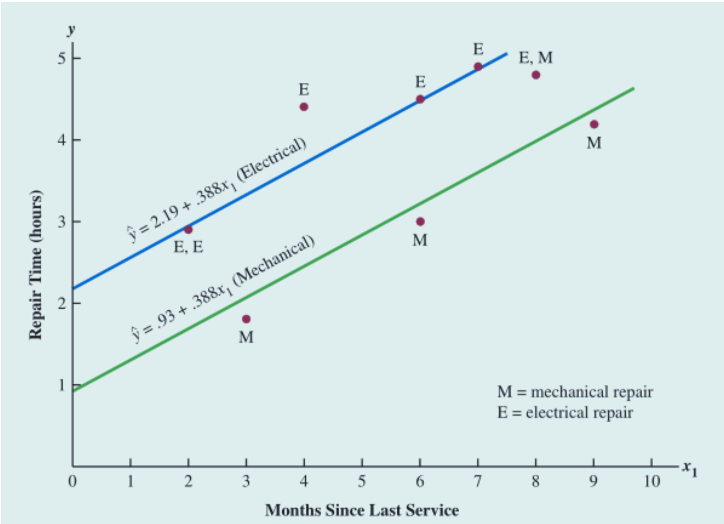
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Time   R-squared:                       0.859
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     21.36
Date:                Mon, 20 Aug 2018   Prob (F-statistic):            0.00105
Time:                        17:03:04   Log-Likelihood:                -4.6200
No. Observations:                  10   AIC:                             15.24
Df Residuals:                       7   BIC:                             16.15
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      0.9305      0.467      1.993      0.087      -0.174       2.035
C(Type)[T.1]   1.2627      0.314      4.020      0.005       0.520       2.005
Months         0.3876      0.063      6.195      0.000       0.240       0.536
==============================================================================
Omnibus:                        3.357   Durbin-Watson:                   1.136
Prob(Omnibus):                  0.187   Jarque-Bera (JB):                1.663
Skew:                           0.994   Prob(JB):                        0.435
Kurtosis:                       2.795   Cond. No.                         22.0
==============================================================================
```

- 在显著性水平0.05下，与F检验相关联的P-值是0.001，这就表明回归关系是显著的。
- t检验部分表明，两个自变量在统计上都是显著的
- 修正的R^2表明估计的回归方程很好地解释了维修时间的变异性

22 参数解释

- 多元回归方程：$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- 当x_2=0（机械）时，用E(y|机械)表述故障维修时间的平均值或者期望值：
  - E(y|机械)=$\beta_0 + \beta_1 x_1 + \beta_2 \times 0 = \beta_0 + \beta_1 x_1$ （公式一）
- 类似地，对于电子类型的故障（x_2=1）：
  - E(y|电子)=$\beta_0 + \beta_1 x_1 + \beta_2 \times 1 =$（$\beta_0 + \beta_2$）$+ \beta_1 x_1$（公式二）
- 比较上面的两个公式，我们发现无论是电子还是机械，平均维修时间都是x_1的线性函数。它们的斜率一样，但是截距不同。β_2的解释是表示电子类故障的平均维修时间与机械类故障的平均维修时间之间的差
  - ➢ β_2为负：电子类型的平均维修时间小于机械类
  - ➢ β_2为正：电子类型的平均维修时间大于机械类
  - ➢ β_2为0：电子类型的平均维修时间与机械类没有差别

23 参数解释



估计的多元回归方程：$\hat{y} = 0.93 + 0.3876x_1 + 1.263x_2$

- 机械类故障：$\hat{y} = 0.93 + 0.3876x_1$
- 电子类故障：$\hat{y} = 2.193 + 0.3876x_1$

我们得到了两个估计的回归方程。因为$b_2 = 1.263$，可得知：电子类型故障的维修时间要比机械类型故障的维修时间多用了1.263个小时

24

复杂分类

- 如果一个分类变量有超过2个值呢？ 该怎么处理
- 销售地区A、B、C三个值，因此我们需要定义3–1=2个虚拟变量
- $x_1 = \begin{cases} 1, & \text{如果销售地区为}B \\ 0, & \text{其它} \end{cases}$ $x_2 = \begin{cases} 1, & \text{如果销售地区是}C \\ 0, & \text{其它} \end{cases}$

销售地区	x_1	x_2
A	0	0
B	1	0
C	0	1

25 参数解释

- 销售数量的期望值E(y)关于虚拟变量的回归方程：
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- 回归方程的三种变化
  - $E(y|\text{销售地区}A) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0$
  - $E(y|\text{销售地区}B) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 = \beta_0 + \beta_1$
  - $E(y|\text{销售地区}C) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 = \beta_0 + \beta_2$
- 于是，β_0是地区A销售数量的期望值，β_1是地区B销售的期望值和地区A销售的期望值的差。β_2是C地区与A地区的期望值的差
- **重点：在多元回归分析中，如果一个分类变量有k个水平，那么需要定义k-1个虚拟变量**