第10章 含定性变量的回归模型

1. **自变量含定性变量的回归模型**
2. **自变量含定性变量的回归模型与应用**
3. 因变量是定性变量的回归模型
4. Logistic(逻辑斯蒂)回归模型

备注： **考试范围1-2；** 授课范围1-4

10.1 自变量含定性变量的回归模型

10.1.1 简单情况

首先讨论定性变量只取两类可能值的情况，例如研究粮食产量问题，y 为粮食产量，x 为施肥量，另外再考虑气候问题，分为正常年份和干旱年份两种情况，对这个问题的数量化方法是引入一个0-1型变量 D，令：

$D_i=1$	表示正常年份
$D_i=0$	表示干旱年份

10.1 自变量含定性变量的回归模型

粮食产量的回归模型为：

$$y_i=\beta_0+\beta_1x_i+\beta_2D_i+\varepsilon_i \qquad (10.1)$$

其中干旱年份的粮食平均产量为：

$$E(y_i|D_i=0)=\beta_0+\beta_1x_i$$

正常年份的粮食平均产量为：

$$E(y_i|D_i=1)=(\beta_0+\beta_2)+\beta_1x_i$$

10.1 自变量含定性变量的回归模型

例10-1 某经济学家想调查文化程度对家庭储蓄的影响，在一个中等收入的样本数据中，随机调查了13户高学历家庭与14户低学历的家庭，因变量 y 为上一年家庭储蓄增加额，自变量 x_1 为上一年家庭总收入，自变量 x_2 表示家庭学历，高学历家庭 x_2=1，低学历家庭x_2=0，调查数据见表10-1：

10.1 自变量含定性变量的回归模型

表 10-1

序　　号	y(元)	x_1(万元)	x_2	e_i	de_i
1	235	2.3	0	−588	455
2	346	3.2	1	−220	−2 372
3	365	2.8	0	−2 371	−1 047
4	468	3.5	1	−1 246	−3 229
5	658	2.6	0	−1 313	−101
6	867	3.2	1	301	−1 851
7	1 085	2.6	0	−886	326
8	1 236	3.4	1	−96	−2 135
9	1 238	2.2	0	797	1 784
10	1 345	2.8	1	2 309	−67
11	2 365	2.3	0	1 542	2 585
12	2 365	3.7	1	−115	−1 985
13	3 256	4.0	1	−371	−2 074
14	3 256	2.9	0	137	1 517
15	3 265	3.8	1	403	−1 412
16	3 265	4.6	1	−2 658	−4 023
17	3 567	4.2	1	−826	−2 416
18	3 658	3.7	1	1 178	−692
19	4 588	3.5	0	−827	891
20	6 436	4.8	1	−252	−1 505
21	9 047	5.0	1	1 593	453
22	7 985	4.2	0	−108	2 002
23	8 950	3.9	0	2 005	3 947
24	9 865	4.8	0	−524	1 924
25	9 866	4.6	0	243	2 578
26	10 235	4.8	0	−154	2 294
27	10 140	4.2	0	2 047	4 157

10.1 自变量含定性变量的回归模型

建立 y 对 x_1和x_2的线性回归，R 软件的计算代码如下，其运行结果见输出结果10.1，其中残差 e_i 列于表10-1中。

```
data10.1<-read.csv("D:/data10.1.csv",head=TRUE)
lm10.1<-lm(y~x1+x2,data=data10.1)
summary(lm10.1)
resid(lm10.1)
```

10.1 自变量含定性变量的回归模型

输出结果 10.1

```
Call:
lm(formula = y ~ x1 + x2, data = data10.1)

Residuals:
      Min        1Q    Median        3Q       Max
 -2658.1    -706.9    -114.5       600.1    2309.0

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  -7976.8     1093.4       -7.295    1.55e-07 ***
x1            3826.1      304.6       12.562    4.82e-12 ***
x2           -3700.3      513.4       -7.207    1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1289 on 24 degrees of freedom
Multiple R-squared:  0.8793,    Adjusted R-squared:  0.8692
F-statistic: 87.43 on 2 and 24 DF,  p-value: 9.555e-12
```

10.1 自变量含定性变量的回归模型

两个自变量 x_1 与 x_2 的系数都是显著的，判定系数 R^2=0.879，回归方程为：

$$\hat{y} = -7977 + 3826x_1 - 3700x_2$$

这个结果表明，中等收入的家庭每增加1万元收入，平均拿出3826元作为储蓄。高学历家庭每年的平均储蓄额少于低学历的家庭，平均少3700元。

如果不引入家庭学历定性变量 x_2，仅用 y 对家庭年收入 x_1 做一元线性回归，得判定系数 R^2=0.618，拟合效果差很多。

10.1 自变量含定性变量的回归模型

家庭年收入 x_1 是连续型变量，它对回归的贡献也是不可缺少的。如果不考虑家庭年收入这个自变量，13户高学历家庭的平均年储蓄增加额为3009.31元，14户低学历家庭的平均年储蓄增加额为5059.36元，这样会认为高学历家庭每年的储蓄增加额比低学历的家庭平均少5059.36-3009.31=2050.05元，而用二元回归法算出的数值是3700元，两者并不相等。

10.1 自变量含定性变量的回归模型

用二元回归法算出的高学历家庭每年的平均储蓄增加额比低学历的家庭平均少3700元，这是在假设两者的家庭年收入相等的基础上的储蓄增加额差值，或者说是消除了家庭年收入的影响后的差值，因而反映了两者储蓄增加额的真实差异。而直接由样本计算的差值2050.05元是包含有家庭年收入影响在内的差值，是虚假的差值。所调查的13户高学历家庭的平均年收入额为3.8385万元，14户低学历家庭的平均年收入额为3.4071万元，两者并不相等。

10.1 自变量含定性变量的回归模型

10.1.2 复杂情况

某些场合定性自变量可能**取多类值**，例如某商厦策划营销方案，需要考虑销售额的季节性影响，**季节因素**分为春、夏、秋、冬4种情况。为了用定性自变量反应春、夏、秋、冬四季，我们初步设想引入如下4个0-1自变量：

$$\begin{cases} x_1 = 1, & \text{春季} \\ x_1 = 0, & \text{其它} \end{cases} \qquad \begin{cases} x_2 = 1, & \text{夏季} \\ x_2 = 0, & \text{其它} \end{cases}$$

$$\begin{cases} x_3 = 1, & \text{秋季} \\ x_3 = 0, & \text{其它} \end{cases} \qquad \begin{cases} x_4 = 1, & \text{冬季} \\ x_4 = 0, & \text{其它} \end{cases}$$

10.1 自变量含定性变量的回归模型

可是这样做却产生了一个新的问题，即$x_1+x_2+x_3+x_4=1$，构成**完全多重共线性**。

解决这个问题的方法很简单，我们只需去掉一个0-1型变量，只保留3个0-1型自变量即可。例如去掉x_4，只保留x_1、x_2、x_3。

对一般情况，一个定性变量有 k 类可能的取值时，需要引入 k-1 个0-1型自变量。当 k=2 时，只需要引入一个0-1型自变量即可。

10.2 自变量含定性变量的回归模型与应用

10.2.1 分段回归

例10-2 表10-2给出某工厂生产批量 x_i 与单位成本 y_i (美元)的数据。试用分段回归建立回归模型。

表 10-2

序　　号	y	$x(=x_1)$	x_2
1	2.57	650	150
2	4.40	340	0
3	4.52	400	0
4	1.39	800	300
5	4.75	300	0
6	3.55	570	70
7	2.49	720	220
8	3.77	480	0

10.2 自变量含定性变量的回归模型与应用

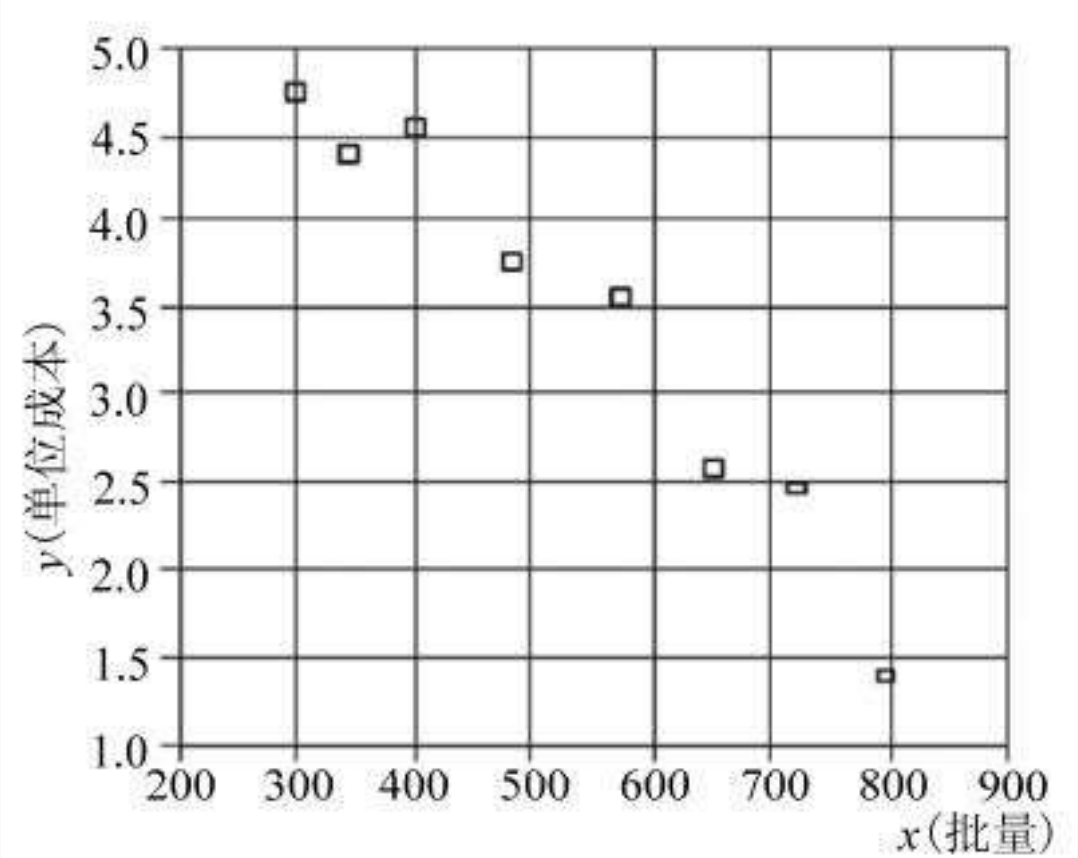


图 10-1 单位成本与批量的散点图

10.2 自变量含定性变量的回归模型与应用

由图10.1可看出数据在生产批量x_p=500时发生较大变化，即批量大于500时成本明显下降。我们考虑由两段构成的分段线性回归,这可以通过引入一个0-1型虚拟自变量实现。假定回归直线的斜率在x_p=500处改变，建立回归模型

$$y_i=\beta_0+\beta_1 x_i+\beta_2(x_i-500)D_i+\varepsilon_i \quad (10.2)$$

其中

$$\begin{cases} D_i=1, & \text{当 } x_i>500 \\ D_i=0, & \text{当 } x_i\le 500 \end{cases}$$

10.2 自变量含定性变量的回归模型与应用

引入两个新的自变量

$$x_{i1} = x_i, \quad x_{i2} = (x_i - 500)D_i$$

这样回归模型转化为标准形式的二元线性回归模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad (10.3)$$

（10.3）式可以分解为两个线性回归方程：

当$x_1 \leqslant 500$时，$E(y) = \beta_0 + \beta_1 x_1$ (10.4)

当$x_1 > 500$时，$E(y) = (\beta_0 - 500\beta_2) + (\beta_1 + \beta_2)x_1$ (10.5)

10.2 自变量含定性变量的回归模型与应用

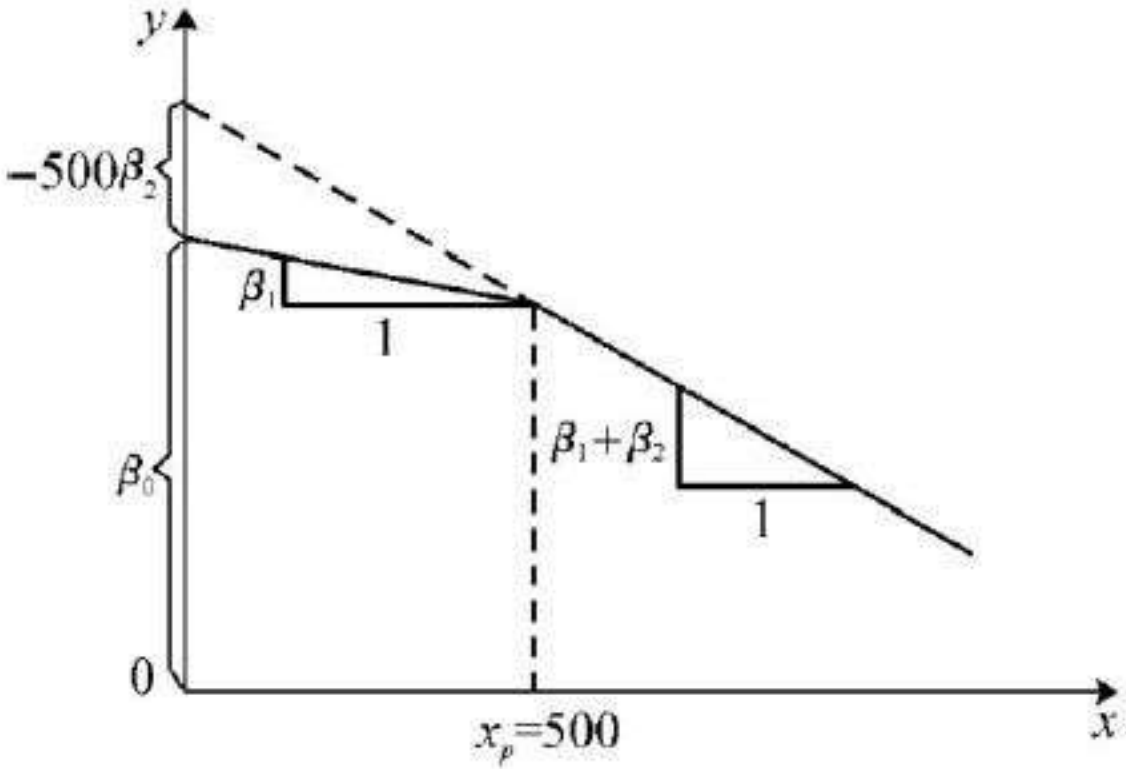


图 10-2

10.2 自变量含定性变量的回归模型与应用

用普通最小二乘法拟合模型(10.3)式得回归方程为：

$$=5.895-0.00395x_1-0.00389x_2 \qquad (10.6)$$

利用此模型可说明生产批量小于500时，每增加1个单位批量，单位成本降低0.00395美元；当生产批量大于500时，每增加1个单位批量，估计单位成本降低到
0.00395+0.00389=0.00784(美元)。

10.2 自变量含定性变量的回归模型与应用

以上只是根据散点图从直观上判断本例数据应该用折线回归拟合，这一点还需要做统计的显著性检验，这只需对（10.2）式的回归系数 β_2 做显著性检验。回归方程式 (10.6) 的相关计算代码及输出结果 10.2 如下所示。

```
data10.2<-read.csv("D:/data10.2.csv",head=TRUE)
#data10.2 中存储了表 10.2 中的数据
lm10.2<-lm(y~x+x2,data=data10.2)
summary(lm10.2)
anova(lm10.2)
```

输出结果 10.2

```
> summary(lm10.2)
Call:
lm(formula = y ~ x + x2, data = data10.2)

Residuals:
     1         2       3         4        5       6        7        8
-0.17160 -0.15117 0.20605 -0.17463 0.04068 0.18068 0.29765 -0.22765

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  5.895447    0.604213      9.757    0.000192 ***
x           -0.003954    0.001492     -2.650    0.045432 *
x2          -0.003893    0.002310     -1.685    0.152774
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2449 on 5 degrees of freedom
Multiple R-squared:  0.9693,    Adjusted R-squared:  0.9571
F-statistic: 79.06 on 2 and 5 DF,  p-value: 0.0001645

> anova(lm10.2)
Analysis of Variance Table
Response: y
           Df    SumSq    Mean Sq    F value      Pr(>F)
x          1    9.3159   9.3159    155.2779    5.902e-05 ***
x2         1    0.1704   0.1704      2.8397    0.1528
Residuals  5    0.3000   0.0600
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.2 自变量含定性变量的回归模型与应用

复决定系数 $R^2 = 0.969$，拟合效果很好。对 β_2 的显著性检验的 t 值= -1.685，显著性检验的概率 P 值 = 0.153，β_2没有通过显著性检验，不能认为 β_2 非零。这样，根据显著性检验，还不能认为本例数据适合拟合折线回归。

用 y 对 x 做一元线性回归，计算代码如下，其运行结果如输出结果10.3 所示。

```
lms10.2<-lm(y~x,data=data10.2)
summary(lms10.2)
anova(lms10.2)
```

输出结果 10.3

```
> lms10.2<-lm(y~x,data=data10.2)
Call:
lm(formula = y ~ x, data = data10.2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34983 -0.17335 -0.05465  0.24673  0.35694

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  6.7945511  0.3241223    20.96     7.68e-07 ***
x           -0.0063184  0.0005796   -10.90     3.53e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.28 on 6 degrees of freedom
Multiple R-squared:  0.9519,    Adjusted R-squared:  0.9439
F-statistic: 118.8 on 1 and 6 DF,    p-value: 3.534e-05

> anova(lms10.2)
Analysis of Variance Table
Response: y
           Df    Sum Sq   Mean Sq   F value    Pr(>F)
x          1     9.3159   9.3159    118.84     3.534e-05 ***
Residuals  6     0.4703   0.0784
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.2 自变量含定性变量的回归模型与应用

y 对 x 的一元线性回归的判定系数 $R^2=0.952$，回归方程为：

$$\hat{y}=6.795-0.006318x \qquad (10.7)$$

（10.7）式说明，批量每增加一件，成本平均下降0.006318美元，这个结论在自变量的样本范围300至800内都是适用的。

10.2 自变量含定性变量的回归模型与应用

10.2.2 回归系数相等的检验

例10.3 回到例10.1的问题，例10.1引入0-1型自变量的方法是假定储蓄增加额y对家庭收入的回归斜率β_1与家庭文化程度无关，家庭文化程度只影响回归常数项β_0，这个假设是否合理，还需要做统计检验。检验方法是引入如下含有交互效应的回归模型：

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i1}x_{i2}+\varepsilon_i \quad (10.8)$$

其中y为上一年家庭储蓄增加额，x_1为上一年家庭总收入，x_2表示家庭学历，

高学历家庭$x_2=1$，低学历家庭$x_2=0$。

10.2 自变量含定性变量的回归模型与应用

回归模型（10.8）式可以分解为对高学历和对低学历家庭的两个线性回归模型，分别为：

高学历家庭$x_2=1$,

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2+\beta_3 x_{i1}+\varepsilon_i$$
$$=(\beta_0+\beta_2)+(\beta_1+\beta_3)x_{i1}+\varepsilon_i \qquad (10.9)$$

低学历家庭$x_2=0$,

$$y_i=\beta_0+\beta_1 x_{i1}+\varepsilon_i \qquad (10.10)$$

10.2 自变量含定性变量的回归模型与应用

可见，高学历家庭的回归常数为$\beta_0+\beta_2$，回归系数为$\beta_1+\beta_3$；低学历家庭的回归常数为β_0，回归系数为β_1。要检验两个回归方程的回归系数是否相等，等价于对回归模型式(10.8)做参数的假设检验

$$H_0：\ \beta_3=0,$$

当拒绝H_0时，认为$\beta_3\neq0$，这时高学历与低学历家庭的储蓄回归模型实际上被拆分为两个不同的回归模型(10.9)和(10.10)式。

当不拒绝H_0时，认为$\beta_3=0$，这时高学历与低学历家庭的储蓄回归模型是如下形式的联合回归模型：

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i \qquad (10.11)$$

10.2 自变量含定性变量的回归模型与应用

(10.11)正是例10-1 所建立的回归模型。建立式(10.8)的 回归模型的计算代码及运行代码的输出结果10.4 如下所示。

```
lm10.3<-lm(y~x1+x2+I(x1*x2),data=data10.1)
summary(lm10.3)
```

输出结果 10.4

```
Call:
lm(formula = y ~ x1 + x2 + I(x1 * x2), data = data10.1)
Residuals:
    Min      1Q  Median      3Q     Max
-2234.2  -662.0  -281.5   728.8  2239.9

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -8763.9      1270.9   -6.896  4.96e-07 ***
x1            4057.2       359.3   11.292  7.36e-11 ***
x2            -776.9      2514.5   -0.309  0.760
I(x1 * x2)    -787.6       663.4   -1.187  0.247
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.2 自变量含定性变量的回归模型与应用

```
Residual standard error: 1278 on 23 degrees of freedom
Multiple R-squared:  0.8863,    Adjusted R-squared:  0.8714
F-statistic: 59.75 on 3 and 23 DF,    p-value: 5.187e-11
```

从输出结果10.4 中看到，对 β_3显著性检验的显著性概率$P=0.247$，应该不拒绝原假设H0：$\beta_3=0$，认为例10-1 采用的回归模型式(10.11)是正确的。

另外，输出结果10.4 中 x_2的回归系数 β_2的显著性概率为 0.760，也没有通过显著性检验，并且比 β_3的显著性更低，是否应该首先剔除 x_2 而保留 x_1x_2？回答是否定的，因为这样做与经济意义不符。

10.2 自变量含定性变量的回归模型与应用

对回归模型式(10.9)与式(10.10)，若$\beta_2=0$，表明两个回归方程的常数项相等；若$\beta_3=0$，表明两个回归方程的斜率相等。经济学家首先关心的是两个回归方程的斜率是否相等，其次才关心常数项是否相等。

通常认为，回归常数项是在自变量为零时y的平均值，但在本例中则**没有这种现实意义**。这是因为本例是对中等收入家庭的储蓄分析，收入为零的家庭的储蓄增加额超出了本模型所包含的范围。本例的回归常数项仅是与储蓄增加额的平均值有关的一个数值。

10.3 因变量是定性变量的回归模型

在许多社会经济问题中，所研究的因变量往往只有两个可能结果，这样的因变量也可用虚拟变量来表示，虚拟变量的取值可取0或1。

10.3.1 定性因变量的回归方程的意义

设因变量y是只取0，1两个值的定性变量，考虑简单线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (10.12)$$

在这种y只取0，1两个值的情况下，因变量均值 $E(y_i) = \beta_0 + \beta_1 x_i$ 有着特殊的意义。

10.3 因变量是定性变量的回归模型

由于y_i是0-1型贝努利随机变量，则得如下概率分布：

$$P(y_i=1)=\pi_i \qquad P(y_i=0)=1-\pi_i$$

根据离散型随机变量期望值的定义，可得

$$E(y_i)=1(\pi_i)+0(1-\pi_i)=\pi_i \qquad (10.13)$$

得到 $$E(y_i)=\pi_i=\beta_0+\beta_1 x_i$$

所以，作为由回归函数给定的因变量均值，$E(y_i)=\beta_0+\beta_1 x_i$是自变量水平为$x_i$时$y_i=1$的概率。

对因变量均值的这种解释既适用于这里的简单线性回归函数，也适用于复杂的多元回归函数。当因变量是0-1变量时，因变量均值总是代表给定自变量时$y=1$的概率。

10.3 因变量是定性变量的回归模型

10.3.2 定性因变量回归的特殊问题

1. 离散非正态误差项。

对一个取值为0和1的因变量，

误差项 $\varepsilon_i=y_i-(\beta_0+\beta_1 x_i)$ 只能取两个值：

当 $y_i=1$ 时，　　$\varepsilon_i=1-\beta_0-\beta_1 x_i=1-\pi i$

当 $y_i=0$ 时，　　$\varepsilon_i=-\beta_0-\beta_1 x_i=-\pi_i$

显然，误差项 ε_i 是两点型离散分布，当然正态误差回归模型的假定就不适用了。

10.3 因变量是定性变量的回归模型

2. 零均值异方差性。

当因变量是定性变量时，误差项ε_i仍然保持零均值，这时出现的另一个问题是误差项ε_i的方差不相等。0-1型随机变量ε_i的方差为

$$D(\varepsilon_i)=D(y_i)=\pi_i(1-\pi_i)$$

$$=(\beta_0+\beta_1 x_i)(1-\beta_0-\beta_1 x_i) \qquad (10.14)$$

ε_i的方差依赖于x_i，是异方差，不满足线性回归方程的基本假定。

10.3 因变量是定性变量的回归模型

3.回归方程的限制

当因变量为0、1虚拟变量时，回归方程代表概率分布，所以因变量均值受到如下限制：

$$0 \leqslant E(y_i) = \pi_i \leqslant 1$$

对一般的回归方程本身并不具有这种限制，线性回归方程$y_i=\beta_0+\beta_1 x_i$将会超出这个限制范围。

对于普通的线性回归所具有的上述三个问题，我们需要构造出能够满足以上限制的回归模型。

10.4 Logistic回归模型

10.4.1 分组数据的Logistic回归模型

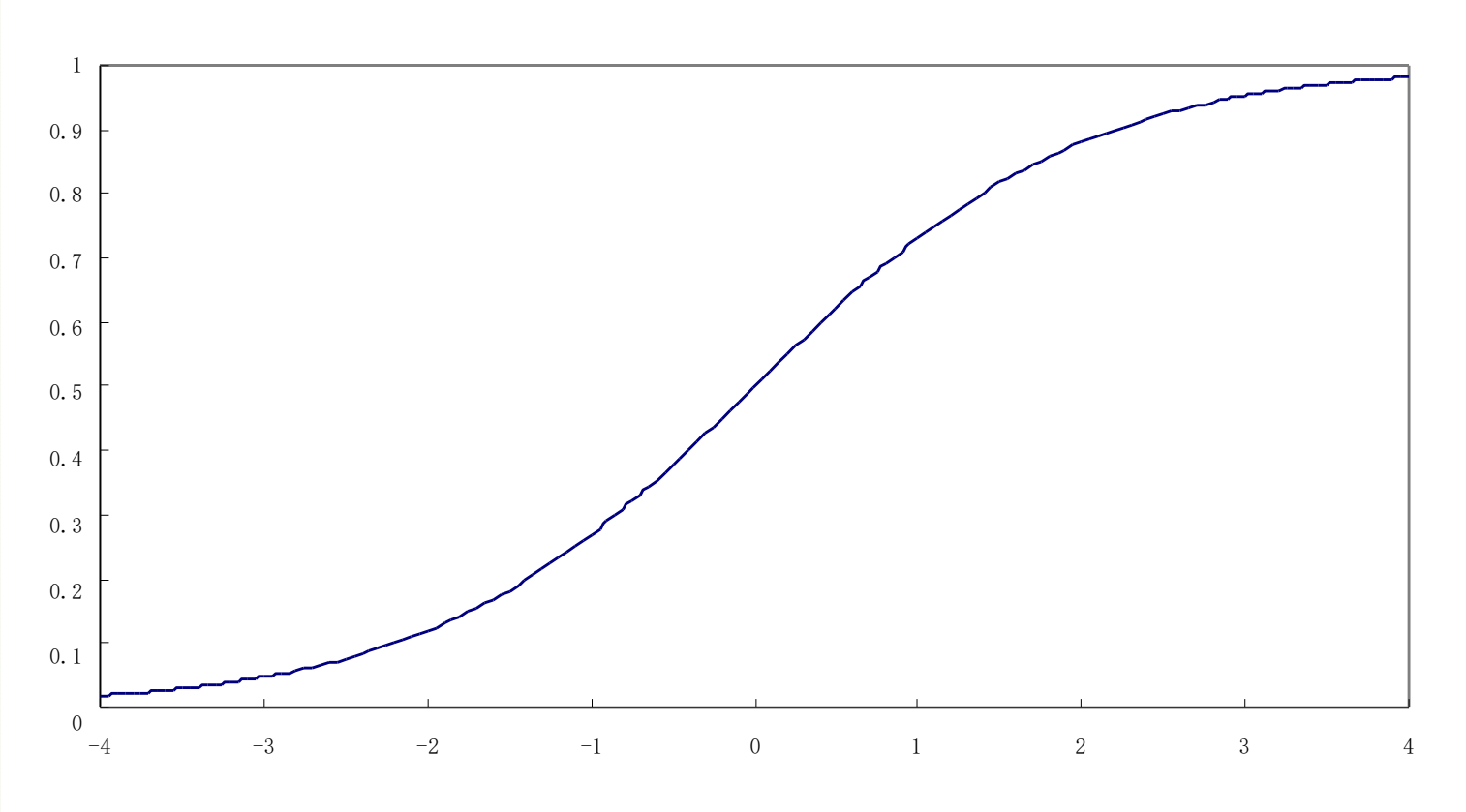
针对0-1型因变量产生的问题，我们对回归模型应该做两个方面的改进。

第一，回归函数应该改用限制在[0，1]区间内的连续曲线，而不能再沿用直线回归方程。限制在[0，1]区间内的连续曲线有很多，例如所有连续型随机变量的分布函数都符合要求，我们常用的是Logistic函数与正态分布函数。

Logistic函数的形式为

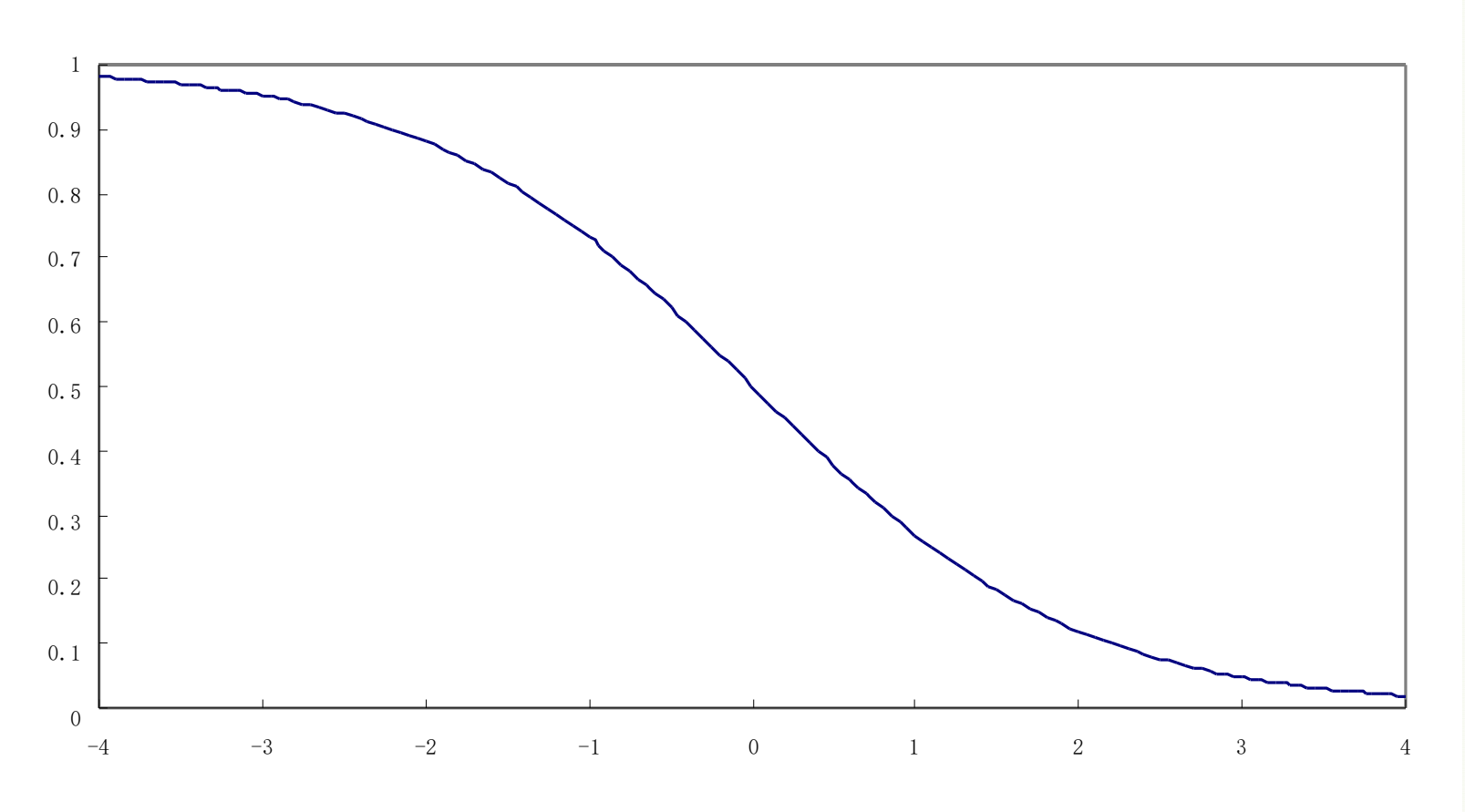
$$f(x)=\frac{e^{x}}{1+e^{x}}=\frac{1}{1+e^{-x}} \qquad (10.15)$$

10.4 Logistic回归模型



$$f(x)=\frac{1}{1+e^{-x}}$$

10.4 Logistic回归模型



$$f(x)=\frac{1}{1+e^{x}}$$

10.4 Logistic回归模型

第二，因变量y_i本身只取0、1两个离散值，不适于直接作为回归模型中的因变量。

由于回归函数$E(y_i)=\beta_0+\beta_1 x_i$ 表示在自变量为x_i 的条件下 y_i的平均值，而 y_i是0-1型随机变量，因而$E(y_i)=\beta_0+\beta_1 x_i$就是在自变量为$x_i$ 的条件下y_i 等于1的概率。 这提示我们可以用y_i 等于1的概率代替y_i 本身作为因变量。

下面通过一个例子来说明Logistic回归模型的应用。

10.4 Logistic回归模型

例10-4 在一次住房展销会上，与房地产商签定初步购房意向书的共有 n=313名顾客中，在随后的3个月的时间内，只有一部分顾客确实购买了房屋。购买了房屋的顾客记为1，没有购买房屋的顾客记为0。以顾客的年家庭收入（万元）为自变量x，对如下的数据，建立Logistic回归模型。

10.4 Logistic回归模型

表 10-3

序号	年家庭收入(万元) x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i = m_i/n_i$	逻辑变换 $p_i' = \ln\left(\frac{p_i}{1-p_i}\right)$	权重 $w_i = n_i p_i (1-p_i)$
1	1.5	25	8	0.320 000	−0.753 77	5.440
2	2.5	32	13	0.406 250	−0.379 49	7.719
3	3.5	58	26	0.448 276	−0.207 64	14.345
4	4.5	52	22	0.423 077	−0.310 15	12.692
5	5.5	43	20	0.465 116	−0.139 76	10.698
6	6.5	39	22	0.564 103	0.257 829	9.590
7	7.5	28	16	0.571 429	0.287 682	6.857
8	8.5	21	12	0.571 429	0.287 682	5.143
9	9.5	15	10	0.666 667	0.693 147	3.333

10.4 Logistic回归模型

Logistic回归方程为

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, 2, \cdots, c$$

其中c为分组数据的组数，本例c=9。做线性化变换，令

$$p_i' = \ln(\frac{p_i}{1-p_i})$$

上式的变换称为逻辑（Logit）变换，得

$$p_i' = \beta_0 + \beta_1 x_i + \varepsilon_i$$

10.4 Logistic回归模型

计算出经验回归方程为

$$\hat{p}' = -0.886 + 0.156x \qquad (10.19)$$

判定系数 r^2=0.9243，显著性检验P值≈0，高度显著。还原为（10.16）式的Logistic回归方程为

$$\hat{p}_i = \frac{\exp(-0.886 + 0.156x)}{1 + \exp(-0.886 + 0.156x)} \qquad (10.20)$$

利用（10.20）式可以对购房比例做预测，例如对x_0=8，

$$\hat{p}_i = \frac{\exp(-0.886 + 0.156 \times 8)}{1 + \exp(-0.886 + 0.156 \times 8)} = \frac{1.436}{1 + 1.436} = 0.590$$

10.4 Logistic回归模型

我们用Logistic回归模型成功地拟合了因变量为定性变量的回归模型，但是仍然存在一个不足之处，就是异方差性并没有解决，（10.18）式的回归模型不是等方差的，应该对（10.18）式用加权最小二乘估计。当n_i较大时，p_i'的近似方差为：

$$D(p_i') \approx \frac{1}{n_i \pi_i (1-\pi_i)} \qquad (10.21)$$

其中 $\pi_i = E(y_i)$，因而选取权数为：

$$w_i = n_i p_i (1-p_i)$$

10.4 Logistic回归模型

对例10.4 重新用加权最小二乘做估计，计算代码如下所示，其运行结果见输出结果10.5。

计算代码

```
data10.4<-read.csv("D:/data10.4.csv",head=TRUE)
#data10.4 中保存了表 10.3 中的数据，其中逻辑变换后的变量记为 p1
lm10.4<-lm(p1~x,weights=w,data10.4)     #使用加权最小二乘估计
summary(lm10.4)
```

10.4 Logistic回归模型

输出结果 10.5

```
Call:
lm(formula = p1 ~ x, data = data10.4, weights = w)

Weighted Residuals:
     Min       1Q       Median       3Q       Max
-0.47461  -0.30088    0.04359    0.26694  0.44923

Coefficients:
                  Estimate   Std. Error    t value   Pr(>|t|)
(Intercept)      -0.84887    0.11358      -7.474    0.000140 ***
x                 0.14932    0.02071       7.210    0.000176 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3862 on 7 degrees of freedom
Multiple R-squared:  0.8813,    Adjusted R-squared:  0.8644
F-statistic: 51.98 on 1 and 7 DF,  p-value: 0.0001759
```

10.4 Logistic回归模型

用加权最小二乘法得到的Logistic回归方程为

$$\hat{p}_i = \frac{\exp(-0.849 + 0.149x)}{1 + \exp(-0.849 + 0.149x)} \qquad (10.23)$$

对x_0=8时的购房比例做预测

$$\hat{p}_i = \frac{\exp(-0.849 + 0.149 \times 8)}{1 + \exp(-0.849 + 0.149 \times 8)} = \frac{1.409}{1 + 1.409} = 0.585$$

10.4 Logistic回归模型

10.4.2 未分组数据的Logistic回归模型

设y是0-1型变量，$x_1, x_2, \ldots, x_p$是与y相关的确定性变量，

n组观测数据为$(x_{i1}, x_{i2}, \ldots, x_{ip}; y_i)$，$i=1,2,\ldots,n$，

y_i与$x_{i1}, x_{i2}, \ldots, x_{ip}$的关系为：

$$E(y_i)=\pi_i=f(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip})$$

其中函数f（x）是值域在[0，1]区间内的单调增函数。对于Logistic回归

$$f(x)=\frac{e^x}{1+e^x}$$

10.4 Logistic回归模型

于是y_i是均值为$\pi_i=f(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip})$的0-1型分布，概率函数为：

$$P(y_i=1)=\pi_i \qquad P(y_i=0)=1-\pi_i$$

可以把y_i的概率函数合写为：

$$P(y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i},\ y_i=0,1; \quad i=1,2,\ldots,n \tag{10.24}$$

于是$y_1, y_2, \ldots, y_n$的似然函数为：

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{10.25}$$

10.4 Logistic回归模型

对数似然函数

$$\ln L=\sum_{i=1}^{n}[y_i\ln\pi_i+(1-y_i)\ln(1-\pi_i)]$$

$$=\sum_{i=1}^{n}[y_i\ln\frac{\pi_i}{(1-\pi_i)}+\ln(1-\pi_i)]$$

Logistic回归

$$\pi_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})}{1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})}$$

代入得

$$\ln L=\sum_{i=1}^{n}[y_i(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})$$

$$-\ln(1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}))] \quad (10.26)$$

极大似然估计就是选取$\beta_0,\beta_1,\beta_2,\ldots,\beta_p$的估计值使上式达极大。

10.4 Logistic回归模型

例10-5　临床医学中为了研究麻醉剂用量与患者是否保持静止的关系，对 30 名患者在手术前 15 分钟给予一定浓度的麻醉剂后的情况进行了记录。记录数据见表 10-4 中，其中麻醉剂浓度为自变量 x，患者是否保持静止为因变量 y，y 取1时表示患者静止，y 取0 时表示患者有移动，试建立 y 关于 x 的Logistic 回归模型。本例数据来自于 R 软件 DAAG 包中自带的 anesthetic 数据集。

10.4 Logistic回归模型

表 10-4

序号	麻醉剂浓度(x)	患者是否保持静止(y)	预测值 $\hat{p}$	序号	麻醉剂浓度(x)	患者是否保持静止(y)	预测值 $\hat{p}$
1	1.0	1	0.288 7	16	1.4	1	0.790 0
2	1.2	0	0.552 7	17	1.4	1	0.790 0
3	1.4	1	0.790 0	18	0.8	0	0.117 6
4	1.4	0	0.790 0	19	0.8	1	0.117 6
5	1.2	0	0.552 7	20	1.2	1	0.552 7
6	2.5	1	0.999 4	21	0.8	0	0.117 6
7	1.6	1	0.919 7	22	0.8	0	0.117 6
8	0.8	0	0.117 6	23	1.0	0	0.288 7
9	1.6	1	0.919 7	24	0.8	0	0.117 6
10	1.4	0	0.790 0	25	1.0	0	0.288 7
11	0.8	0	0.117 6	26	1.2	1	0.552 7
12	1.6	1	0.919 7	27	1.0	0	0.288 7
13	2.5	1	0.999 4	28	1.2	1	0.552 7
14	1.4	1	0.790 0	29	1.0	0	0.288 7
15	1.6	1	0.919 7	30	1.2	1	0.552 7

10.4 Logistic回归模型

在R中对0-1 型因变量做logistic 回归的函数为 glm()，该函数主要用来建立广义线性模型，当 glm()函数中的参数family=binomial(表明分布族为二项分布)，联系函数link="logit"时，建立的回归模型为Logistic 回归模型。对例10.5 中的数据建立Logistic回归模型的计算代码如下，运行代码后得到输出结果10.6。

```
install.packages("DAAG")
library(DAAG)
fm<-glm(nomove~conc,family=binomial(link="logit"),data=anesthetic)
#nomove 为表 10-4 中的 y，conc 为 x
summary(fm)
p=predict(fm,type="response")    #计算 y=1 的概率的预测值 p̂
```

10.4 Logistic回归模型

输出结果 10.6

```
Call:
glm(formula = nomove ~ conc, family = binomial(link = "logit"),
    data = anesthetic)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.76666  -0.74407  0.03413  0.68666  2.06900

Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -6.469       2.418   -2.675   0.00748 **
conc           5.567       2.044    2.724   0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 27.754  on 28  degrees of freedom
AIC: 31.754
Number of Fisher Scoring iterations: 5
```

10.4 Logistic回归模型

输出结果10.6 中的z value 的计算公式类似于线性回归中t value，即

$$Z = \frac{\hat{\beta}_j}{sd(\hat{\beta}_j)}$$

其中，$\hat{\beta}_j$是参数的估计值(Estimate)，$sd(\hat{\beta}_j)$是估计参数的标准差(Std. Error)。在假设 $\beta_j=0$成立时，Z 近似服从标准正态分布。

由该检验可知，回归系数是显著的，回归方程为

$$\hat{p} = \frac{\exp(-6.469 + 5.567x)}{1 + \exp(-6.469 + 5.567x)}$$

10.4 Logistic回归模型

10.4.3 Probit 回归模型

Probit 回归称为单位概率回归，与Logistic回归相似，也是拟合0-1型因变量回归的方法，其回归函数是

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad (10.28)$$

用样本比例 p_i 代替概率 π_i，表示为样本回归模型

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (10.29)$$

10.4 Logistic回归模型

例10-6　使用例 10.4 的购房数据，首先计算出$\Phi^{-1}(p_i)$的数值，见表10-5。以 $\Phi^{-1}(p_i)$ 为因变量，以年家庭收入 x 为自变量做普通最小二乘线性回归，得回归方程

$$\hat{\Phi}^{-1}(p_i) = -0.552 + 0.0970x$$

或等价地表示为

$$\hat{p}_i = \Phi(-0.552 + 0.0970x)$$

对 $x_0 = 8$，　$\hat{p}_0 = \Phi(-0.552 + 0.0970 \times 8) = \Phi(0.224) = 0.589$

与用Logistic 回归计算的预测值很接近。

10.4 Logistic回归模型

表 10-5

序号	年家庭收入 (万元)x	签订意向书 人数 n_i	实际购房 人数 m_i	实际购房比例 $p_i = m_i/n_i$	Probit 变换 $p'_i = \Phi^{-1}(p_i)$
1	1.5	25	8	0.320 000	−0.467 70
2	2.5	32	13	0.406 250	−0.237 20
3	3.5	58	26	0.448 276	−0.130 02
4	4.5	52	22	0.423 077	−0.194 03
5	5.5	43	20	0.465 116	−0.087 55
6	6.5	39	22	0.564 103	0.161 38
7	7.5	28	16	0.571 429	0.180 01
8	8.5	21	12	0.571 429	0.180 01
9	9.5	15	10	0.666 667	0.430 73

10.4 Logistic回归模型

使用 R 软件可以直接做 Probit 回归，做 Probit 回归的函数仍为 glm()，其中只需将联系函数设为 link="probit"，对于已整理的分组数据在使用 glm() 函数建立 Probit 模型时，需要以购房比例作为因变量，签订意向书人数作为权重，以下为相应的计算代码，运行后得到输出结果10.7。

```
data10.4<-read.csv("D:/data10.4.csv",head=TRUE)
glm10.6<-glm(p~x,weight=n,family=binomial(link="probit"),data=data10.4)
summary(glm10.6)
```

10.4 Logistic回归模型

输出结果 10.7

```
Call:
glm(formula = p ~ x, family = binomial(link = "probit"), data = data10.4,
    weights = n)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.47599  -0.30254  0.04287  0.27093  0.45008

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.53177    0.18144  -2.931  0.00338 **
x             0.09354    0.03307   2.829  0.00467 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9.1386  on 8  degrees of freedom
Residual deviance: 1.0441  on 7  degrees of freedom
AIC: 40.09
Number of Fisher Scoring iterations: 3
```

10.4 Logistic回归模型

由输出结果10.7 得回归方程

$$\hat{\Phi}^{-1}(p_i) = -0.53177 + 0.09354x$$

该结果与前面普通最小二乘的结果(10.30)很接近，在R 软件中也可以对该分组数据做Logistic 回归，具体代码如下：

```
glma10.6<-glm(p~x,weight=n,family=binomial(link="logit"),data=data10.4)
summary(glma10.6)
```

运行代码后，可得到回归方程为

$$\hat{p}' = -0.85178 + 0.14982x$$

这也与用最小二乘法所得到的Logistic 回归方程式(10.19)很接近。