**回归分析**

目录

[一．一元线性回归分析 1](#_Toc432247862)

[二．多元线性回归分析 3](#_Toc432247863)

[三．回归变量的选择方法 5](#_Toc432247864)

[四．模型的进一步分析 7](#_Toc432247865)

[五．logistic模型 11](#_Toc432247866)

# 1 一元线性回归分析

步骤：

A.建立回归模型；

B.求解回归模型中的参数；

C.对回归模型进行检验。

R中，与线性模型有关的函数有：lm()、summary()、anova()和predict()。我们由例子入手，逐步学习这些函数。

例1：

财政收入与税收有密切的依存关系。d4.3给出我们1978年改革开放以来到2008年共31年的税收（x，百亿元）和财政收入（y，百亿元）数据，试分析税收与财政收入之间的依存关系。

（1）读入数据

|  |
| --- |
| > yx=read.table("clipboard",header=T) ##在d4.3中选取B1：C32区域，然后拷贝 |

（2）拟合模型

|  |
| --- |
| > fm=lm(y~x,data=yx)  > fm |

Call:

lm(formula = yx$y ~ yx$x)

Coefficients:

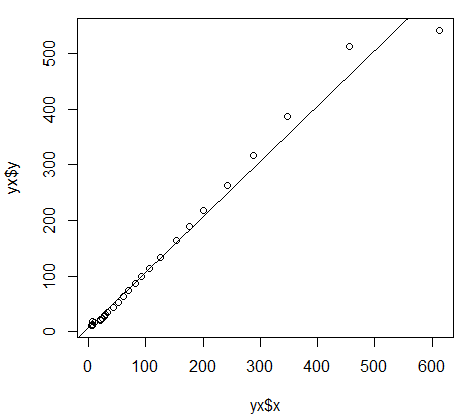
(Intercept) yx$x

6.7336 0.9982

**于是得到回归方程：**

（3）作回归直线

|  |
| --- |
| > plot(yx$x, yx$y)  > abline(fm) |



（4）回归方程的假设检验

1）模型的方差分析

|  |
| --- |
| > anova(fm) |

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 712077 712077 27428 < 2.2e-16 \*\*\*

Residuals 29 753 26

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

由于p<0.05，于是在0.05水平处拒绝原假设，即本例回归系数有统计学意义，x与y间存在直线回归关系。

2）回归系数的显著性检验

|  |
| --- |
| > summary(fm) |

Call:

lm(formula = yx$y ~ yx$x)

Residuals: ##残差的最小值，0.25分位点，中位数点，0.75分位点和最大值

Min 1Q Median 3Q Max

-76.763 -5.627 -1.264 3.003 51.066

Coefficients:

## Estimate是参数估计值，Std. Error表示参数的标准差，t value为t值，Pr(>|t|)为p值

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.73358 4.38003 1.537 0.135 ##常数项

yx$x 0.99824 0.02436 40.983 <2e-16 \*\*\* ##一次项

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.46 on 29 degrees of freedom ##残差的标准差

Multiple R-squared: 0.983, Adjusted R-squared: 0.9824 ##R方与调整R方

F-statistic: 1680 on 1 and 29 DF, p-value: < 2.2e-16 ##F值和p值

由于p<0.05，于是在0.05水平处拒绝原假设，即本例回归系数有统计学意义，x与y间存在回归关系。

注：本例中，当df=1时，t值得平方等于F值（dfe即为t的自由度n-2）。所以说当自变量只有一个时，方差分析与t检验的结果是等价的。但在下面的多元分析中，方差分析与t检验的结果并不等价

（5）预测

当经过检验，回归方程是有意义时，可以用它作预测。

|  |
| --- |
| > new<-data.frame(x=700) ##输入新的点x=700，这里即时是一个点，也要采用数据框形式  > lm.pred<-predict(fm,new,interval="prediction",level=0.95)##给出预测值，interval="prediction"指给出预测区间，level=0.95表示相应概率为0.95  > lm.pred ##fit为预测值，lwr是95%下限，upr是95%上限 |

fit lwr upr

1 705.5033 655.461 755.5456

# 2 多元线性回归分析

例2：

考察财政收入和国内生产总值x1，税收x2，进出口贸易总额x3，经济活动人口x4之间的数量关系，建立多元线性回归方程。

（1）读入数据

##在d4.4中选取B1：F32区域，然后拷贝

|  |
| --- |
| > yX=read.table("clipboard",header=T) |

（2）拟合模型

|  |
| --- |
| > (fm=lm(y~x1+x2+x3+x4,data=yX)) |

Call:

lm(formula = y ~ x1 + x2 + x3 + x4)

Coefficients:

(Intercept) yX$x1 yX$x2 yX$x3 yX$x4

23.5321088 -0.0033866 1.1641150 0.0002919 -0.0437416

于是得到多元线性回归方程：



（3）方程的标准化系数

由于自变量与因变量都是有单位的，从数值上来看，它们样本取值的极差会有很大的差异，均数与标准差也各不相同，所以不能由偏回归系数的大小直接说明对因变量线性影响的大小。对于这个问题常用变量标准化与计算标准化偏回归系数的方法来处理。标准化后常数项为0，且各变量的标准差相同，可用偏回归的系数的值来反映各自变量在其他自变量固定时对因变量线性影响的大小，相互之间可进行比较。

常用的统计软件都能给出标准化偏回归系数，但R语言中并不包含计算标准回归系数的函数。因此需要自编。

|  |
| --- |
| > library(mvstats)  > coef.sd(fm) |

$coef.sd

x1 x2 x3 x4

-0.0174513678 1.0423522972 0.0009628564 -0.0371053994

由标准化偏回归系数可见，税收对财政收入的线性影响最大。

（4）回归方程的假设检验

|  |
| --- |
| > summary(fm) |

Call:

lm(formula = y ~ x1 + x2 + x3 + x4, data = yX)

Residuals:

Min 1Q Median 3Q Max

-5.0229 -2.1354 0.3297 1.2639 6.9690

##系数的t值和p值，系数的显著性检验

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 23.5321088 4.5990714 5.117 2.47e-05 \*\*\*

x1 -0.0033866 0.0080749 -0.419 0.678

x2 1.1641150 0.0404889 28.751 < 2e-16 \*\*\*

x3 0.0002919 0.0085527 0.034 0.973

x4 -0.0437416 0.0092638 -4.722 7.00e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.79 on 26 degrees of freedom

Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997

F-statistic: 2.289e+04 on 4 and 26 DF, p-value: < 2.2e-16 ##方程的F值和p值，方程的显著性检验

模型的p<0.0001，故本例回归模型是有意义的。

偏回归系数b2，b4的p值都小于0.01，可认为解释变量税收x2和经济活动人口x4显著；b1，b3的p值大于0.50，不能否定b1=0，b3=0的假设。可认为国内生产总值x1和进出口贸易总额x3对财政收入y没有显著的影响。我们可以看到，国内生产总值、经济活动人口所对应的偏回归系数都为负，这与经济现实是不相符的。出现这种结果的可能原因是这些解释变量之间存在高度的共线性。

（5）预测

|  |
| --- |
| > new<-data.frame(x1=30,x2=40,x3=50,x4=100)  > lm.pred<-predict(fm,new,interval="prediction",level=0.95)  > lm.pred |

fit lwr upr

1 65.63555 56.8971 74.37399

# 3 回归变量的选择方法

1.全局择优法

对每组子集， RSS越小、R2越大、校正R2越大、AIC BIC越小，模型越好。

|  |
| --- |
| > library(leaps) ##安装包leaps  > varsel=regsubsets(y~ x1+ x2+ x3+ x4,data=yX)  > result=summary(varsel)  >data.frame(result$outmat,RSS=result$rss,R2=result$rsq,adjR2=result$adjr2,Cp=result$cp,BIC=result$bic) |

x1 x2 x3 x4 RSS R2 adjR2 Cp BIC

1 ( 1 ) \* 752.8849 0.9989438 0.9989074 69.745044 -205.5777

2 ( 1 ) \* \* 203.8835 0.9997140 0.9996936 1.198844 -242.6410

3 ( 1 ) \* \* \* 202.3451 0.9997161 0.9996846 3.001165 -239.4418

4 ( 1 ) \* \* \* \* 202.3360 0.9997162 0.9996725 5.000000 -236.0092

2.逐步回归法

向前引入法、向后剔除法、逐步筛选法。

|  |
| --- |
| > fm=lm(yX$y~ yX$x1+ yX$ x2+ yX$x3+ yX$ x4)  > fm.step=step(fm,direction="forward") #forward为向前引入法 |

Start: AIC=68.15

y ~ x1 + x2 + x3 + x4

|  |
| --- |
| > fm.step=step(fm,direction="backward") #backward为向后剔除法 |

Start: AIC=68.15

y ~ x1 + x2 + x3 + x4

Df Sum of Sq RSS AIC

- x3 1 0.0 202.3 66.156

- x1 1 1.4 203.7 66.363

<none> 202.3 68.154

- x4 1 173.5 375.8 85.351

- x2 1 6433.1 6635.4 174.352

Step: AIC=66.16

y ~ x1 + x2 + x4

Df Sum of Sq RSS AIC

- x1 1 1.5 203.9 64.390

<none> 202.3 66.156

- x4 1 197.3 399.6 85.253

- x2 1 7382.2 7584.5 176.496

Step: AIC=64.39

y ~ x2 + x4

Df Sum of Sq RSS AIC

<none> 204 64.39

-x4 1 549 753 102.89

-x2 1 367655 367859 294.82

|  |
| --- |
| > fm.step=step(fm,direction="both") #both为逐步筛选法 |

Start: AIC=68.15

y ~ x1 + x2 + x3 + x4

Df Sum of Sq RSS AIC

- x3 1 0.0 202.3 66.156

- x1 1 1.4 203.7 66.363

<none> 202.3 68.154

- x4 1 173.5 375.8 85.351

- x2 1 6433.1 6635.4 174.352

Step: AIC=66.16

y ~ x1 + x2 + x4

Df Sum of Sq RSS AIC

-x1 1 1.5 203.9 64.390

<none> 202.3 66.156

+ x3 1 0.0 202.3 68.154

- x4 1 197.3 399.6 85.253

- x2 1 7382.2 7584.5 176.496

Step: AIC=64.39

y ~ x2 + x4

Df Sum of Sq RSS AIC

<none> 204 64.390

+ x1 1 2 202 66.156

+ x3 1 0 204 66.363

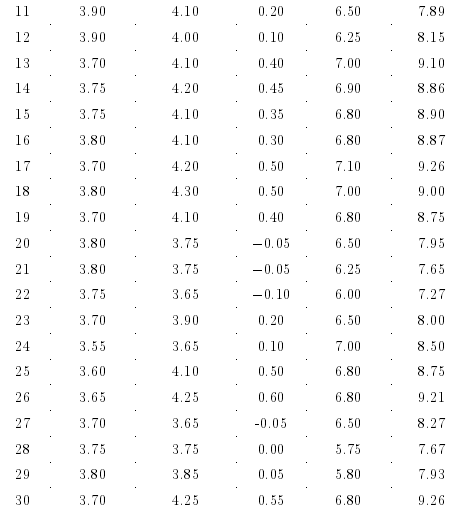
- x4 1 549 753 102.888

- x2 1 367655 367859 294.825

# 四．模型的进一步分析

例3：某大型牙膏制造企业想找出公司生产的牙膏销售量与销售价格、广告投入等之间的关系，从而预测在不同价格和广告费用下的销售量。为此，收集一下数据。试根据这些数据建立一个数学模型，分析牙膏销售量与其他因素的关系。





对于大多数顾客来说，在购买同类产品的牙膏时，更多地会关心不同品牌之间的价格差，而不是它们的价格本身。因此，在研究各个因素对销售量的影响时，用价格差代替公司销售价格各其他厂家平均价格更为合适。

1.模型1

|  |
| --- |
| >toothpaste<-data.frame(  X1=c(-0.05, 0.25,0.60,0, 0.25,0.20, 0.15,0.05,-0.15, 0.15,  0.20, 0.10,0.40,0.45,0.35,0.30, 0.50,0.50, 0.40,-0.05,  -0.05,-0.10,0.20,0.10,0.50,0.60,-0.05,0, 0.05, 0.55),  X2=c( 5.50,6.75,7.25,5.50,7.00,6.50,6.75,5.25,5.25,6.00,  6.50,6.25,7.00,6.90,6.80,6.80,7.10,7.00,6.80,6.50,  6.25,6.00,6.50,7.00,6.80,6.80,6.50,5.75,5.80,6.80),  Y =c( 7.38,8.51,9.52,7.50,9.33,8.28,8.75,7.87,7.10,8.00,  7.89,8.15,9.10,8.86,8.90,8.87,9.26,9.00,8.75,7.95,  7.65,7.27,8.00,8.50,8.75,9.21,8.27,7.67,7.93,9.26)  )  > lm.sol<-lm(Y~X1+X2,data=toothpaste) ##建立y=b0+b1\*x1+b2\*x2的线性模型  > summary(lm.sol)##模型检验 |

Call:

lm(formula = Y ~ X1 + X2, data = toothpaste)

Residuals:

Min 1Q Median 3Q Max

-0.49779 -0.12031 -0.00867 0.11084 0.58106

##系数的显著性检验

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.4075 0.7223 6.102 1.62e-06 \*\*\*

X1 1.5883 0.2994 5.304 1.35e-05 \*\*\*

X2 0.5635 0.1191 4.733 6.25e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

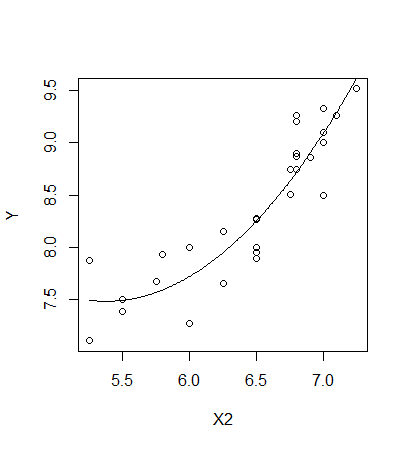
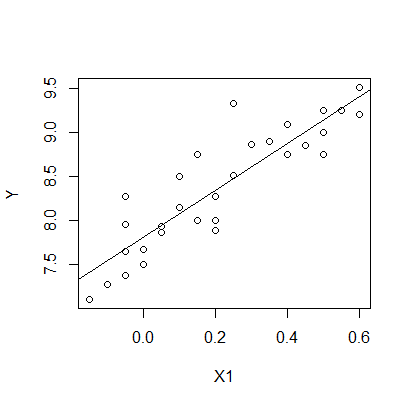
Residual standard error: 0.2383 on 27 degrees of freedom

Multiple R-squared: 0.886, Adjusted R-squared: 0.8776

F-statistic: 105 on 2 and 27 DF, p-value: 1.845e-13 ##方程的显著性检验

计算结果通过了回归系数检验和回归方程检验，由此得到销售量与价格差与广告费之间的关系为：y=4.4075+1.5883x1+0.5635x2

为了进一步分析回归模型，我们画出y与x1和y与x2散点图。从散点图上可以看出，y与x1用直线拟合较好，y与x2则用二次曲线拟合较好。



绘制上两图的命令如下：

|  |
| --- |
| > attach(toothpaste)  > plot(Y~X1); abline(lm(Y~X1)) |

##绘制y与x1的散点图和回归直线

##绘制y与x2的散点图和回归曲线

|  |
| --- |
| > lm2.sol<-lm(Y~X2+I(X2^2))  > x<-seq(min(X2), max(X2), len=200)  > y<-predict(lm2.sol, data.frame(X2=x))  > plot(Y~X2); lines(x,y) |

其中I(X2^2)表示模型中x2的平方项。

我们作相应的回归分析。

2.模型2

|  |
| --- |
| > lm.new<-update(lm.sol, .~.+I(X2^2)) ##将二次项加入回归方程中Y ~ X1 + X2 + I(X2^2)  > summary(lm.new) |

Call:

lm(formula = Y ~ X1 + X2 + I(X2^2), data = toothpaste)

Residuals:

Min 1Q Median 3Q Max

-0.40330 -0.14509 -0.03035 0.15488 0.46602

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.3244 5.6415 3.071 0.00495 \*\*

X1 1.3070 0.3036 4.305 0.00021 \*\*\*

X2 -3.6956 1.8503 -1.997 0.05635 .

I(X2^2) 0.3486 0.1512 2.306 0.02934 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2213 on 26 degrees of freedom

Multiple R-squared: 0.9054, Adjusted R-squared: 0.8945

F-statistic: 82.94 on 3 and 26 DF, p-value: 1.944e-13

此时，模型残差的标准差有所下降，R2有所上升。说明模型修正合理。模型检验的p值1.944e-13<0.05，模型显著。系数显著性检验中，x2的系数的p值>0.05。进一步修改模型，去掉x2项：

3.模型3

|  |
| --- |
| > lm2.new<-update(lm.new, .~.-X2)###-x2表示去掉x2这一项  > summary(lm2.new) |

Call:

lm(formula = Y ~ X1 + I(X2^2), data = toothpaste)

Residuals:

Min 1Q Median 3Q Max

-0.4859 -0.1141 -0.0046 0.1053 0.5592

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.07667 0.35531 17.102 5.17e-16 \*\*\*

X1 1.52498 0.29859 5.107 2.28e-05 \*\*\*

I(X2^2) 0.04720 0.00952 4.958 3.41e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2332 on 27 degrees of freedom

Multiple R-squared: 0.8909, Adjusted R-squared: 0.8828

F-statistic: 110.2 on 2 and 27 DF, p-value: 1.028e-13

此模型虽然通过了模型的显著性检验和系数的显著性检验，但与上一模型对比来看，残差的标准差上升，R2下降，这又是此模型的不足之处。

做进一步修正，考虑x1与x2的交互作用：

4.模型4

|  |
| --- |
| > lm3.new<-update(lm.new, .~.+X1\*X2) ###添加交互项X1\*X2  > summary(lm3.new) |

Call:

lm(formula = Y ~ X1 + X2 + I(X2^2) + X1:X2, data = toothpaste)

Residuals:

Min 1Q Median 3Q Max

-0.43725 -0.11754 0.00489 0.12263 0.38410

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29.1133 7.4832 3.890 0.000656 \*\*\*

X1 11.1342 4.4459 2.504 0.019153 \*

X2 -7.6080 2.4691 -3.081 0.004963 \*\*

I(X2^2) 0.6712 0.2027 3.312 0.002824 \*\*

X1:X2 -1.4777 0.6672 -2.215 0.036105 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2063 on 25 degrees of freedom

Multiple R-squared: 0.9209, Adjusted R-squared: 0.9083

F-statistic: 72.78 on 4 and 25 DF, p-value: 2.107e-13

模型通过了模型的显著性检验和系数的显著性检验，并且残差的标准差减少，R2增加。因此，最终模型为：



# 五．logistic模型

R中logistic模型的公式为：

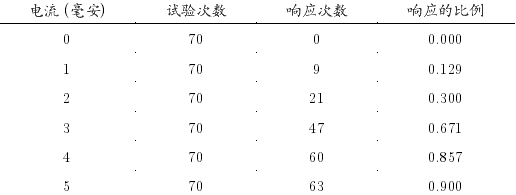
fm<-glm(formula,family=binomial(link=logit),data=data.frame)

式中的link=logit可以不写，因为logit是二项分布连接函数是缺省状态。

在用函数glm()作logistic回归模型时，对于公式formula有两种输入方法：一种是输入成功和失败的次数；另一种是线性模型通常数据的输入方式。

例5：

为研究高压电线对牲畜的影响，R.Norell研究小的电流对农场动物的影响。实验中选择了7头牛，6种电击强度（0、1、2、3、4、5mA）。每头牛被电击30下，每种强度5下，按随机次序进行。然后重复整个实验，每头牛总共被电击60下。对每次电击，响应变量——嘴巴运动或出现、或不出现。下图为数据：



这里，响应变量是分类的，它只有两个值：出现及未出现，这种情况下可用logistic回归：

，其中x是电流强度。概率p取0到1之间的值。

|  |
| --- |
| >norell<-data.frame(  x=0:5, n=rep(70,6), success=c(0,9,21,47,60,63)) ###输入数据  > norell$Ymat<-cbind(norell$success, norell$n-norell$success)##按成功和失败的次数的方式  >glm.sol<-glm(Ymat~x, family=binomial, data=norell)##logistic模型  > summary(glm.sol) |

Call:

glm(formula = Ymat ~ x, family = binomial, data = norell)

Deviance Residuals:

1 2 3 4 5 6

-2.2507 0.3892 -0.1466 1.1080 0.3234 -1.6679

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.3010 0.3238 -10.20 <2e-16 \*\*\*

x 1.2459 0.1119 11.13 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 250.4866 on 5 degrees of freedom

Residual deviance: 9.3526 on 4 degrees of freedom

AIC: 34.093

Number of Fisher Scoring iterations: 4

即：，，并且回归方程通过了检验，因此回归模型为：

----------------------ln(P/(1-P))=?

与线性回归模型相同，在得到回归模型后可以作预测，例如，当电流强度为3.5mA时，有响应的牛的概率为多少？

|  |
| --- |
| > pre<-predict(glm.sol, data.frame(x=3.5))  > p<-exp(pre)/(1+exp(pre));p |

1

0.742642

即74.26%。

可以作控制，如有50%的牛有响应，其电流强度为多少？

当p=0.5，即，所以

|  |
| --- |
| > X<- - glm.sol$coefficients[1]/glm.sol$coefficients[2];X |

(Intercept)

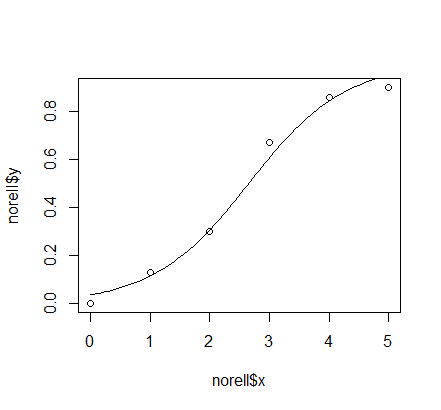
2.649439

即2.65mA的电流强度，可以使50%的牛有响应。

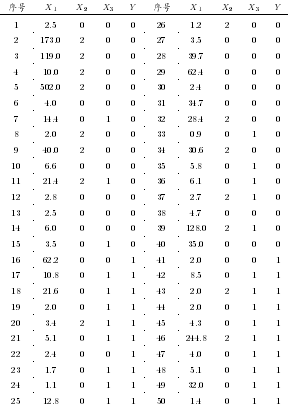
最后画出相应的比例与logistic回归曲线。

命令：

|  |
| --- |
| > d<-seq(0,5, len=100) ##d给出曲线横坐标的点  > pre<-predict(glm.sol, data.frame(x=d)) ##pre计算预测值  > p<-exp(pre)/(1+exp(pre)) ##p是相应的预测概率  > norell$y<-norell$success/norell$n  > plot(norell$x,norell$y) ##散点图  > lines(d,p) ##预测曲线 |



例6：50位急性淋巴细胞性白血病病人，在入院治疗时取得了细胞数X1（千个/mm3）；淋巴结浸润等级X2（分为0、1、2、3级）；出院后有无巩固治疗X3（1表示有，0表示无）。通过随访取得病人的生存时间，并以变量Y=0表示生存时间在1年以内，Y=1表示生存时间在1年或1年以上。关于X1、X2、X3和Y的观测数据如下。试用logistic回归模型分析病人生存时间长短的概率与X1、X2、X3的关系。



命令：

|  |
| --- |
| >life<-data.frame(  X1=c(2.5, 173, 119, 10, 502, 4, 14.4, 2, 40, 6.6,  21.4, 2.8, 2.5, 6, 3.5, 62.2, 10.8, 21.6, 2, 3.4,  5.1, 2.4, 1.7, 1.1, 12.8, 1.2, 3.5, 39.7, 62.4, 2.4,  34.7, 28.4, 0.9, 30.6, 5.8, 6.1, 2.7, 4.7, 128, 35,  2, 8.5, 2, 2, 4.3, 244.8, 4, 5.1, 32, 1.4), X2=rep(c(0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,  0, 2, 0, 2, 0, 2, 0),  c(1, 4, 2, 2, 1, 1, 8, 1, 5, 1, 5, 1, 1, 1, 2, 1,  1, 1, 3, 1, 2, 1, 4)),  X3=rep(c(0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1),  c(6, 1, 3, 1, 3, 1, 1, 5, 1, 3, 7, 1, 1, 3, 1, 1, 2, 9)),  Y=rep(c(0, 1, 0, 1), c(15, 10, 15, 10))  )  > glm.sol<-glm(Y~X1+X2+X3, family=binomial, data=life)  > summary(glm.sol) |

Call:

glm(formula = Y ~ X1 + X2 + X3, family = binomial, data = life)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6960 -0.5842 -0.2828 0.7436 1.9292

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.696538 0.658635 -2.576 0.010000 \*\*

X1 0.002326 0.005683 0.409 0.682308

X2 -0.792177 0.487262 -1.626 0.103998

X3 2.830373 0.793406 3.567 0.000361 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 67.301 on 49 degrees of freedom

Residual deviance: 46.567 on 46 degrees of freedom

AIC: 54.567

Number of Fisher Scoring iterations: 5

即回归模型为

用上述回归模型作观测，若一个病人的前两项的指标观测值为X1=5，X2=2，若无巩固治疗（X3=0），则1年以上的存活概率为：

|  |
| --- |
| > pre<-predict(glm.sol, data.frame(X1=5,X2=2,X3=0))  > p<-exp(pre)/(1+exp(pre));p |

1

0.03664087

为3.66%。

若进行了巩固治疗（X3=1），则1年以上的存活概率为：

|  |
| --- |
| > pre<-predict(glm.sol, data.frame(X1=5,X2=2,X3=1))  > p<-exp(pre)/(1+exp(pre));p |

1

0.3920057

为39.20%。

> 0.3920057/0.03664087

[1] 10.69859

比没有巩固治疗提高了10.69859倍。

实际上，用上述模型作预测还存在一些问题，这是因为在得到logistic回归模型时，X1的系数没有通过检验，其p值为0.6823，可以类似于线性模型，用step()进行变量筛选。

|  |
| --- |
| >glm.new<- step(glm.sol) |

Start: AIC=54.57

Y ~ X1 + X2 + X3

Df Deviance AIC

- X1 1 46.718 52.718

<none> 46.567 54.567

- X2 1 49.502 55.502

- X3 1 63.475 69.475

Step: AIC=52.72

Y ~ X2 + X3

Df Deviance AIC

<none> 46.718 52.718

- X2 1 49.690 53.690

- X3 1 63.504 67.504

Call: glm(formula = Y ~ X2 + X3, family = binomial, data = life)

Coefficients:

(Intercept) X2 X3

-1.642 -0.707 2.784

Degrees of Freedom: 49 Total (i.e. Null); 47 Residual

Null Deviance: 67.3

Residual Deviance: 46.72 AIC: 52.72

再用summary()函数显示模型的细节：

|  |
| --- |
| > summary(glm.new) |

Call:

glm(formula = Y ~ X2 + X3, family = binomial, data = life)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6849 -0.5949 -0.3033 0.7442 1.9073

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.6419 0.6381 -2.573 0.010082 \*

X2 -0.7070 0.4282 -1.651 0.098750 .

X3 2.7844 0.7797 3.571 0.000355 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 67.301 on 49 degrees of freedom

Residual deviance: 46.718 on 47 degrees of freedom

AIC: 52.718

Number of Fisher Scoring iterations: 5

从计算结果来看，所有系数均通过了检验（α=0.1），此时回归模型为



再作预测分析

|  |
| --- |
| > pre<-predict(glm.new, data.frame(X2=2,X3=0))  > p<-exp(pre)/(1+exp(pre));p |

1

0.04496518

|  |
| --- |
| > pre<-predict(glm.new, data.frame(X2=2,X3=1))  > p<-exp(pre)/(1+exp(pre));p |

1

0.4325522

|  |
| --- |
| > 0.4325522/0.04496518 |

[1] 9.619715

因此巩固治疗比没有巩固治疗提高了9.619倍。