课程报告

|  |  |
| --- | --- |
| **课程名称:** | 商务数据分析 |
| **课程学期:** | 2018年秋季 |
| **院系：** | 计算机科学与技术学院 |
| **班级：** | 1636101 |
| **姓名：** | 陈翔 |
| **学号：** | 1161800218 |

**回归模型数据分析案例1: 上市公司财务报表分析与预测**

**1. 研究目的**

通过对上市公司公开财务报表信息的统计分析，预测该企业来年的盈利状况。

**2. 背景介绍**

上市公司要求的财务信息披露，是高度发达的市场经济下所有权与经营权分离的必然结果，在两权分离后，由于企业的外部利益集团并不直接参与企业的经营管理活动，因此只有通过解读公司的财务信息来了解企业的财务状况，而想要更深的了解那些蕴含在数字背后的信息，进而预测企业下一年的盈利状况，就要对财务报表数字进行分析。

**3. 指标设计**

随机抽取某年度500 家上市企业的相关数据。对所有样本，解释变量来自当年，而因变量来自下一年。因变量是该企业的净资产收益率，而解释变量有：资产周转率，利润率=主营业务利润/主营业务收入，债务资本比率，成长速度=主营业务收入增长率，市倍率，收入质量=应收账款/主营业务收入，存货率=存货/资产总计，对数变换后的资产总计，以及当年净资产收益率。

资产周转率（ X1）

资产周转率是衡量企业资产管理效率的重要财务比率，在财务分析指标体系中具有重要地位。总资产周转率是考察企业资产运营效率的一项很重要指标， 体现企业经营期间全部资产从投入到产出的流转速度， 反映企业全部资产的管理质量和利用效率。通过该指标的对比分析，能反映企业本年度以及以前年度总资产的运营效率和变化， 发现企业和同类企业在资产利用上的差距， 促进了企业挖掘潜力、积极的创收、提高了产品市场占有率、也提高资产利用效率、一般情况下，这个数值越高，表明了企业总资产周转速度越快。销售能力就越强，资产利用效率就越高。

利润率（ X2）

利润率是反映企业一定时期利润水平的相对指标。利润率=主营业务利润/主营业务收入

债务资本比率（ X3 ）

债务资本比率反映了企业长期偿债能力。比率越小说明企业长期还债能力越强，比率越大则说明企业债务负担重对债券人不利，而且企业倒闭的风险大。

成长速度（ X4）

成长速度=主业务收入增长率主营业务收入增长率可以用来衡量公司的产品生命周期， 判断公司发展所处的阶段。一般的说， 如果主营业务收入增长率超过10%，说明公司产品处于成长期，将继续保持较好的增长势头，尚未面临产品更新的风险，属于成长型公司。如果主营业务收入增长率在5%～10%之间，说明公司产品已进入稳定期，不久将进入衰退期，需要着手开发新产品。如果该比率低于5%，说明公司产品已进入衰退期，保持市场份额已经很困难， 主营业务利润开始滑坡，如果没有已开发好的新产品，将步入衰落。

市倍率(X5)

即市净率，可用于投资分析，一般来说市净率较低的股票，投资价值较高，相反，则投资价值较低；但在判断投资价值时还要考虑当时的市场环境以及公司经营情况、盈利能力等因素.

收入质量（ X6）

收入质量侧重于观察企业收入的成长性和波动性。成长性越高，收入质量越好说明企业通过主营业务创造现金流量的能力越强。收入质量=应收账款/主营业务收入

存货率（ X7）

存货率=存货/资产总计

资产规模（ X8）

企业可以控制的资产总额

净资产收益率（ X9）

即公司税后利润除以净资产得到的百分比率， 该指标反映股东权益的收益水平， 用衡量公司运用自有资本的效率。指标值越高，说明投资带来的收益越高。该指标体现了自有资本获得净收益的能力，是企业报告期末总负债与所有者权益合计之比。它反映了企业长期偿债能力，同时也反映了企业的资本结构和企业利用外借资金的程度。另外，企业资本分为权益资本和债务资本，因此债务资本比重=债务资本/总资本=负债/总资产。

**4. 描述分析**

简单描述分析

N NU SD MIN MED MAX

y 500 0.0591940 0.48890668 -1.324 0.0450 1.6010

x1 500 0.5374740 0.36762055 0.000 0.4925 1.8830

x2 500 0.2127000 0.19786442 -0.345 0.2090 0.7500

x3 500 0.5263304 0.39808338 0.002 0.4552 2.0478

x4 500 1.7630300 1.30903920 0.000 1.4570 7.7850

x5 500 3.6324360 9.31886368 -22.952 3.8635 27.4220

x6 500 0.1490440 0.11450220 0.000 0.1256 0.6834

x7 500 0.1413180 0.09832033 0.000 0.1300 0.6090

x8 500 21.1220600 0.90792336 18.462 21.1585 23.6850

x9 500 0.3711660 0.49491645 -1.164 0.3950 1.9860

样本量是500 个，所有变量都没有缺失。

根据第二行统计分析可以看到，资产周转率的均值为0.5374740，中位数为0.4925.资产

周转速度一般。

第三行的结果告诉我们， 利润率为21.3％，中位数也仅为20.9％。这说明该时期企业利

润水平指标较低，即企业盈利较少。

第四行的结果表明，债务资本比率均值0.526 中位数0.45.偿还债务能力中等，债权人

处于不利状态，企业有倒闭的风险。

第五行结果表明，均值为17.6％，超过了10％， 可知公司产品处于成长期，将继续保

持较好的增长势头， 尚未面临产品更新的风险， 属于成长型公司，主营业务收入增长率

高。

第六行结果表明，均值为3.63 ，该市倍率较低，投资价值较高。但由于该公司为成长

型公司，盈利能力较差，所以投资价值有待考量。

第七行结果表明， 该公司虽为成长型， 但成长性不高， 企业通过主营业务创造现金流量

的能力较差。

第八行的结果14.13 ％表示存货不足。

第九行结果还原后分别为：均值： 1.4899 ，标准差： 2.4769 ，最小值： 1.042 ，中位数：

1.5453 ，最大值： 1.9331 ，可知资产规模中等。

第十行结果表明， 当年净资产收益率不是太高， 由第十一行可得下一年的净资产收益率

有了明显下降，大约下降了31 ％。说明企业成长极度不稳定，有破产风险。

**5. 模型分析**

在描述分析的基础上，我们通过普通线性回归对各个因素同相对利润变化之间的关系做了模型分析

Call:

lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9,

data = a)

Coefficients:

(Intercept) x1 x2 x3 x4 x5

-0.789521 0.030850 0.104957 -0.079449 0.002338 -0.003310

x6 x7 x8 x9

-0.005953 -0.165517 0.031483 0.589544

方差分析如下

Analysis of Variance Table

Response: y

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

x1 1 0.188 0.188 1.3422 0.247215

x2 1 1.103 1.103 7.8729 0.005218 \*\*

x3 1 5.348 5.348 38.1677 1.366e-09 \*\*\*

x4 1 0.004 0.004 0.0307 0.861045

x5 1 4.295 4.295 30.6514 5.042e-08 \*\*\*

x6 1 0.108 0.108 0.7694 0.380842

x7 1 0.173 0.173 1.2357 0.266852

x8 1 1.277 1.277 9.1136 0.002669 \*\*

x9 1 38.126 38.126 272.1184 < 2.2e-16 \*\*\*

Residuals 490 68.654 0.140

---

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

参数估计如下

Call:

lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9,

data = a)

Residuals:

Min 1Q Median 3Q Max

-1.15866 -0.23263 -0.00475 0.23598 1.33988

Coefficients:

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

(Intercept) -0.789521 0.407504 -1.937 0.0533 .

x1 0.030850 0.047506 0.649 0.5164

x2 0.104957 0.087241 1.203 0.2295

x3 -0.079449 0.044838 -1.772 0.0770 .

x4 0.002338 0.012839 0.182 0.8556

x5 -0.003310 0.001976 -1.675 0.0945 .

x6 -0.005953 0.147331 -0.040 0.9678

x7 -0.165517 0.173899 -0.952 0.3417

x8 0.031483 0.019233 1.637 0.1023

x9 0.589544 0.035739 16.496 <2e-16 \*\*\*

---

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

Residual standard error: 0.3743 on 490 degrees of freedom

Multiple R-squared: 0.4244, Adjusted R-squared: 0.4138

F-statistic: 40.14 on 9 and 490 DF, p-value: < 2.2e-16

从分析结果中可以看出，债务资本比率，市倍率，当年净资产收益的因素显著（ p 值<0.2 ）具体理解如下：

（1）债务资本比率的系数估计为-0.079. 这说明，在给定其他特征不变的情况下。债务资本比率高的企业，相比债务资本比率低的企业，来年可以有更高的盈利

（2）市倍率的参数估计为-0.003. 这说明在给定其他特征不变的情况下。市倍率高的企业相

比市倍率低的企业，来年可以带来更高的盈利。

（3）当年净资产收益率的参数估计为0.031. 这说明在给定其他特征不变的情况下。当年净

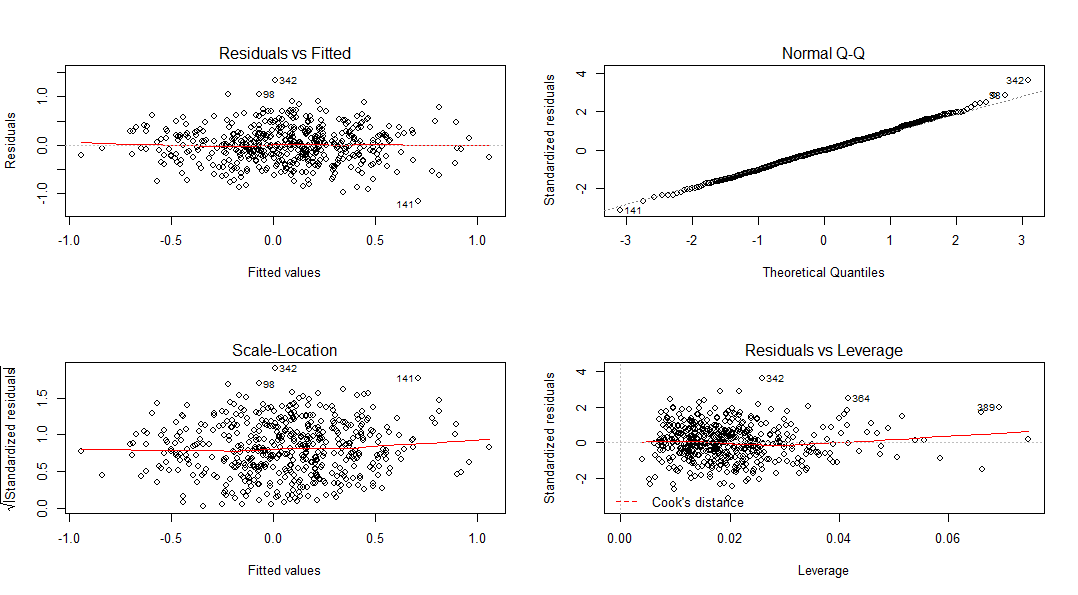
资产收益率高的企业相比当年净资产收益率低的企业，来年有更高的盈利。

（4）由于资产周转率，利润率成长速度，收入质量，存货率，下一年净资产收益率的p 值

不显著，因此没有证据证明这些因素的价值相关。当然， 不能排除这种可能性：可能因为样

本量不够大，发现不了或是其他原因。所以，目前不对以上因素下任何结论。

回归诊断: (1)残差图; (2) Q-Q图; (3)标准化残差方根散点图; (4) Cook距离图



**6. 总结讨论**

本研究通过对上市公司公开财务报表信息的统计分析。通过资产周转率，利润率，债务资本比率，成长速度，市倍率，收入质量，存货率，当年净资产收益率，预测该企业来年的盈利状况。本研究对数据做了描述分析以及回归分析， 其中回归分析的判决系数良好。研究发现，债务资本比率， 市倍率，当年净资产收益率起到正相关作用， 而缺乏足够证据刻画资产周转率，利润率成长速度，收入质量，存货率，下一年净资产收益率所起到的作用。

附录：源代码

a<-read.csv('E1.csv')

names(a)=c('y','x1','x2','x3','x4','x5','x6','x7','x8','x9')

summary(a)

N=sapply(a,length)

NU=sapply(a,mean)

SD=sapply(a,sd)

MIN=sapply(a,min)

MED=sapply(a,median)

MAX=sapply(a,max)

result=cbind(N,NU,SD,MIN,MED,MAX)

result

lm1=lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9,data=a)

lm1

anova(lm1)

summary(lm1)

par(mfrow=c(2,2))

plot(lm1)

**回归模型数据分析案例2: 非学位职业培训机构的学员数据分析**

**1. 研究目的**

通过对某非学位职业培训机构的178 名学员进行数据分析，了解什么样的学员有可能获得更好的学习效果（成绩）。以此为依据，来指导未来的招生计划。

**2. 背景介绍**

非学位职业培训主要是以在职人员为主要对象、以管理科学和管理技术为主要内容、以面授、影音和书籍为载体、以赢利为目的的商业行为。可以粗略的分为：个体户、演化型、中介型、营销型、产品型和综合型。

每一个类型都包括以市、省或全国为目标地区的机构。培训机构、讲师和客户是非学位培训行业的三种核心力量， 三者的力量对比关系决定着培训业的运营模式。由于三者的分布非常广泛，每一群体都包括实力、规模等差异巨大的众多个体，所以培训业的运营模式复杂多变，产品名目繁多、培训机构形式各异、讲师良莠不齐、竞争态势犬牙交错。

研究非学位培训产业有三个方面：

一、判断需求是非学位培训业的首要任务—购买动机

二、品牌是客户选择非学位培训的核心指标—购买行为

三、管理科学是非学位培训的核心内容—持续购买

**3. 指标设计**

我们对某年度随机抽取的178 名已结业的学院数据进行分析， 其中因变量为学员的课程平均成绩，解释性变量有五个，分别是：

性别（ X1）

该指标刻画了学员的性别区别，分为男女两个取值，希望通过分析判断出性别是否是影响学员平均成绩的因素。

出生年代（ X2）

参加非学位职业培训的人员年龄差异比较大， 这里我们根据原始数据中学员的出生年月日将学员分为不同的出生年代， 不同的出生年代也就意味着人员年龄的差异，但在划分这一变量的时候，由于每个年代意味着最多的10 年的年龄差别，这一年龄跨度比较大，有可能会影响到分析结果。出生年代的取值有三个：五十年代、六十年代和七十年代。

企业性质（ X3）

该指标刻画了学员所在企业的性质是民企、国企还是外企。不同的企业在招人的时候对人员的素质会有不同的要求， 而且不同性质的企业有着不同的企业文化和工作风格，这些差异和学员的学习能力——课程成绩是否会有较大的关系，我们需要通过数据分析来确定。

最高学历（ X4）

学员的最高学历的取值有四个，分别为大专、本科、硕士和硕士及以上。最高学历直接衡量了学员在参加非学位职业培训前的受教育程度， 是和学员的学习能力关联性较大的一个因素， 一般来说学员之前的最高学历越高受教育程度就越大，相应的学位知识的学习能力也就越强。但学校知识的学习能力越强是否就意味着职业培训中的平均成绩越高？毕竟学位学习和非学位学习的知识结构和学习风格会存在着一些较为明显的差异， 但经过初步判断这一解释变量将是一个比较重要的影响变量。

最高学历毕业年代（ X5）

最后一个指标是学员最高学历的毕业年代， 这一指标也是从原始数据中的学员的毕业年月日转化而来的， 在划分年代时也存在着和出生年代划分时同样的取值跨度较大的问题。这一指标的取值有三个： 八十年代、九十年代和21世纪（一零年代）。

**4. 描述分析**

在正式的模型分析之前，我们首先对因变量以及自变量作必要的描述分析。

核心发现以及结论如下。

x1 N mean SD min med max

*<fct>* *<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 男 133 4.37 0.0946 3.91 4.39 4.49

2 女 45 4.40 0.0749 4.17 4.41 4.51

关于性别，首先可以看到的是样本量。从中可以看到，男的样本是133个。女的样本是45 个。男的样本是明显多。以中位数计，男的成绩为exp(4.39)=11.8233分,女的成绩为exp(4.41)=11.98638分。相对差异不大。

x3 N mean SD min med max

*<fct>* *<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 国企 95 4.38 0.0906 3.91 4.39 4.49

2 民企 43 4.38 0.0963 4.09 4.39 4.49

3 外企 40 4.37 0.0858 4.09 4.38 4.51

关于企业性质，首先可以看到样本分布不良。分别为95（国企）、43（民企）和40（外企）。相较而言，国企的比例稍高于民企和外企。以中位数计，国企的成绩为exp(4.38)=11.90484分, 民企的成绩为exp(4.38)=11.90484分, 外企的成绩为exp(4.37)=11.87766 分。国企和民企的成绩完全相同，而外企的成绩稍微低。

x4 N mean SD min med max

*<fct>* *<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 本科 148 4.39 0.0858 3.91 4.41 4.51

2 大专 25 4.31 0.0985 4.09 4.32 4.48

3 硕士或以上 5 4.38 0.0387 4.33 4.38 4.42

关于最高学历，我们可以发现样本分布非常不均匀。分别为148（本科）、25（大专）、2（硕士）、3（硕士或以上）。本科成绩的均价为exp(4.39)=11.93202分， 大专成绩的均价为exp(4.31)=11.71458， 硕士成绩的均价为exp(4.42)=12.01356，硕士或以上成绩的均价为exp(4.36)=11.85048。可见，硕士成绩的均值比其他稍微高。

x2 N mean SD min med max

*<fct>* *<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 50年代 10 4.27 0.124 4.09 4.31 4.39

2 60年代 85 4.38 0.0909 3.91 4.39 4.51

3 70年代 83 4.38 0.0781 4.09 4.39 4.49

x5 N mean SD min med max

*<fct>* *<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 00年代 26 4.37 0.0744 4.19 4.37 4.49

2 80年代 47 4.35 0.113 3.91 4.38 4.49

3 90年代 105 4.39 0.0813 4.09 4.41 4.51

关于出生年代和毕业年代，首先需要说明的是，对于出生年代和最高学历毕业年代，我们都是在原始数据——即学员的具体出生年月日和最高学历毕业年月日的基础上经过初步的数据处理， 按照每十年一个年代的常规分类方法把具体的年月日转化为了有一定时间跨度的年代， 这种分类有利于我们对较为散乱的年月日进行分类，方便统计分析，找出定性数据的共性。出生年代的具体分类方法是：1950-1959年划分为五十年代；1960-1969划分为六十年代；1970-1979划分为七十年代。最高学历毕业年代划分方法： 1980-1989 为八十年代，1990-1999 为九十年代； 2000-2009 为21 世纪（一零年代）。

关于出生日期，也就是年龄范围，样本集中在60年代和70年代，分别为85和83。50年代仅有10名学员，且他们的分数，以中位数计，明显低于另外两个年龄范围的学员。这可能说明年龄超过50岁的人们很少愿意参加职业培训，且由于衰老等原因，他们的学习成绩偏低。

关于毕业时间，我们可以发现有105名学员毕业于90年代，毕业于80年代和00年代的学员只有47名和26名。这表明人们在毕业后10至20年的时间内，或许出于职业发展得需要，他们更愿意参加职业培训。但是毕业于不同年代的学员的学习成绩，从最大值和中位数上看，差距都不明显。

**5. 模型分析**

在描述分析的基础上，通过方差分析对各个因素同学习成绩之间的关系作了模型分析。

由于生理、心理、家庭等方面的原因，不同年龄的男女之间的差别是不一样的，比如对于生活状态而言，40岁的男性和女性之间的差别明显大于20岁的男性和女生之间的差别，所以对性别同出生日期（年代）的交互作用特别感兴趣。因此，在分析中加入了该交互作用。为理解分析简便，没有再考虑其他的交互作用。

Analysis of Variance Table

Response: y

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

x1 1 0.02513 0.025129 3.4616 0.0645792 .

x2 2 0.11640 0.058201 8.0174 0.0004744 \*\*\*

x3 2 0.01459 0.007293 1.0046 0.3683940

x4 2 0.06211 0.031054 4.2779 0.0154323 \*

x5 2 0.01073 0.005365 0.7390 0.4791477

x1:x2 2 0.01898 0.009488 1.3071 0.2733843

Residuals 166 1.20504 0.007259

---

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

Call:

lm(formula = y ~ x1 \* x2 + x3 + x4 + x5, data = a)

Residuals:

Min 1Q Median 3Q Max

-0.46125 -0.03376 0.01458 0.05246 0.16369

Coefficients:

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

(Intercept) 4.304567 0.039942 107.770 < 2e-16 \*\*\*

x1女 0.071462 0.072671 0.983 0.32686

x260年代 0.084621 0.033938 2.493 0.01363 \*

x270年代 0.084898 0.036967 2.297 0.02289 \*

x3民企 -0.007357 0.016376 -0.449 0.65385

x3外企 -0.016135 0.016626 -0.970 0.33323

x4大专 -0.057896 0.020094 -2.881 0.00448 \*\*

x4硕士或以上 0.001034 0.040039 0.026 0.97942

x580年代 -0.015912 0.025476 -0.625 0.53309

x590年代 0.002670 0.020043 0.133 0.89419

x1女:x260年代 -0.023456 0.076431 -0.307 0.75931

x1女:x270年代 -0.069372 0.075344 -0.921 0.35852

---

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

Residual standard error: 0.0852 on 166 degrees of freedom

Multiple R-squared: 0.1706, Adjusted R-squared: 0.1157

F-statistic: 3.105 on 11 and 166 DF, p-value: 0.0007968

方差分析表明，模型整体的F-检验高度显著（p-值：0.0005），这说明学员的学习成绩确实同目前考虑的五个因素（即：性别、出生日期、企业性质、最高学历、毕业时间）中的至少1个因素有关。为了进一步甄别到底哪个因素重要，对各个因素作第2型方差分析。从中可以看到，性别和出生日期的交互作用不显著，此外企业性质和毕业时间也不显著（假设10%的显著性水平）。因此，在后面的分析中，将这三个因素提出。重新拟合的结果表明，剩下的三个因素（即：性别、出生日期和最高学历）在10%的水平下都比较显著。模型的拟合优度比较一般，判决系数只有15%，说明学员成绩的影响因素的确纷繁复杂，难以捉摸。

Analysis of Variance Table

Response: y

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

x1 1 0.02513 0.025129 3.4802 0.0638109 .

x2 2 0.11640 0.058201 8.0605 0.0004507 \*\*\*

x4 2 0.06952 0.034760 4.8140 0.0092408 \*\*

Residuals 172 1.24193 0.007220

---

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

Call:

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

Residuals:

Min 1Q Median 3Q Max

-0.47124 -0.03526 0.01538 0.05714 0.15565

Coefficients:

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

(Intercept) 4.29804 0.02994 143.546 < 2e-16 \*\*\*

x1女 0.02414 0.01471 1.641 0.10266

x260年代 0.08522 0.03010 2.831 0.00519 \*\*

x270年代 0.08103 0.03035 2.669 0.00833 \*\*

x4大专 -0.05924 0.01952 -3.035 0.00278 \*\*

x4硕士或以上 0.01206 0.03901 0.309 0.75757

---

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

Residual standard error: 0.08497 on 172 degrees of freedom

Multiple R-squared: 0.1453, Adjusted R-squared: 0.1204

F-statistic: 5.846 on 5 and 172 DF, p-value: 5.152e-05

**6. 总结讨论**

本研究使用某非学位职业培训机构的178位学员的数据，对影响学员的学习效果的多个因素作了方差分析。通过性别、出生日期、最高学历、毕业时间刻画了各个学员的自身特征，同时还兼顾了学员所在企业的所有制性质。本研究发现，性别、出生日期（年代）和最高学历都对学员的学习效果有着显著的影响，但是判决系数不是很高，说明决定学员成绩的因素错综复杂，并不是仅仅由这几个因素决定。而最重要的因素比如智商、学习态度等又很难客观评定和具体化，所以造成拟合优度不是很高。

在未来的后续研究中，我们可以考虑两方面的改进。第一，采用更加科学的评定学习效果的指标。比如，进行入学测试和结业测试，从而计算每个学员这两个成绩的差值，来度量学员的进步水平，从而来评价学习效果。更精确的做法可以在入学时和结业时都进行多次测试计算平均成绩，来尽量避免学员超常发挥或者发挥失常的可能性，得到对学员成绩更客观的评价。第二，采集更多的可以作为解释变量的因素，比如每个学员可用于学习的时间的多少，或者可以做一些简单的智力测验来给学员的智力分等级。另外应该根据具体职业培训的门类制定解释变量，比如如果是IT培训，那么对于学员的成绩，理科生和文科生的区别很可能也是一个显著的解释变量。

附录：源代码

a<-read.csv('E2.csv')

names(a)=c("y","x1","x2","x3","x4","x5")

a$y=log(a$y)

library('dplyr')

group\_by(a,x1) %>% summarise(N=n(),mean=mean(y),SD=sd(y),min=min(y),med=median(y),max=max(y))

group\_by(a,x2) %>% summarise(N=n(),mean=mean(y),SD=sd(y),min=min(y),med=median(y),max=max(y))

group\_by(a,x3) %>% summarise(N=n(),mean=mean(y),SD=sd(y),min=min(y),med=median(y),max=max(y))

group\_by(a,x4) %>% summarise(N=n(),mean=mean(y),SD=sd(y),min=min(y),med=median(y),max=max(y))

group\_by(a,x5) %>% summarise(N=n(),mean=mean(y),SD=sd(y),min=min(y),med=median(y),max=max(y))

lm<-lm(y~x1\*x2+x3+x4+x5,data=a)

anova(lm)

summary(lm)

lm1<-lm(y~x1+x2+x4,data=a)

anova(lm1)

summary(lm1)

**回归模型数据分析案例3: 移动通信客户流失规律分析**

**1. 研究目的**

通过对某移动通信公司客户的流失数据分析，了解客户流失规律，建立流失预警系统，为客户关系管理服务。

**2. 背景介绍**

某年度随机抽取的1000 个移动通信客户。因变量是他们来年的流失行为（ 0=未流失， 1=流失）。为了能够预测客户的未来行为，我们采集了下面这些来自当年的指标：客户等级（区分VIP 客户等级）： 1,2,3,4；主叫次数（ %）：7 日内日均主叫次数/90 日内日均主叫次数；被叫次数（ %）：7 日内日均被叫次数/90 日内日均通话时长；费用（ %）：7 日内日均通话费用/90 日内日均通话费用。

移动通信行业的现有企业中，一般情况下客户月流失率在3%左右，如果静态计算，则所有客户会在 2 － 3 年内全部流失。在降低客户流失率方面，哪怕仅仅降低 1 ％就意味着你至少可以有百万元的收入增长！客户是一个公司最宝贵的财富，因此保持客户并增长客户就是头等重要的事情，同是又是很困难的一项任务。

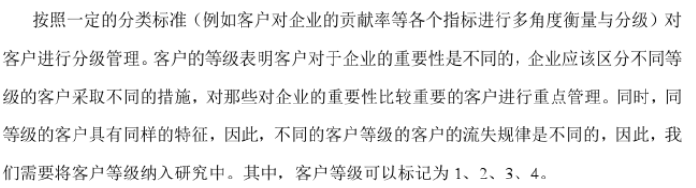
在2011 年底，中国的人口已达1347350000，同时手机量达到1006923000，也就是说中国的手机普及率差不多达到了75%，现有的用户数几乎已经接近人口总数。在一个如此成熟和饱和的市场中，开拓新用户的难度可想而知。根据美国市场营销学会顾客满意手册的统计数据表明，吸引一个新顾客所耗费的成本大概相当于保持一个现有客户的5 倍，而且从传统意义上来讲，移动通信行业保留旧客户利润率为开发一位新客户之16 倍，尤其对于剩余客户市场日渐稀疏的移动通信市场来说，减少客户流失就意味着用更少的成本减少利润的流失，这点已经为运营商所广为接受。

由此可见客户保持的重要性，也就是说保留旧客户比开发、吸收新客户更重要。在成熟期的产品市场中，要开拓新客户很不容易。客户的忠诚度应该是一个企业能够生存发展的最大资产之一，拥有忠诚度的客户，会因客户有学习的效果，而使企业可以花费较少的成本来服务客户，降低了公司在服务成本上的支出，而且忠诚的客户也会宣传正面的口碑效应以作为他人的参考，进而替企业创造新的交易。

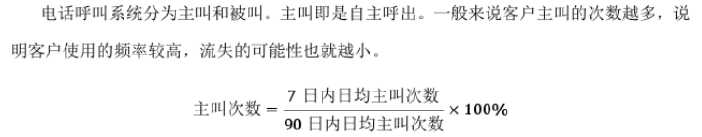
因此本文试图通过逻辑回归模型来对某移动通信公司客户的流失数据分析，了解客户流失规律，建立流失预警系统，为客户关系管理服务。

**3. 指标设计**

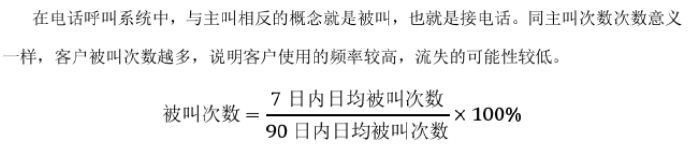
客户等级（X1）



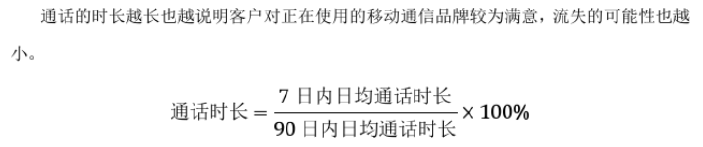
主叫次数（X2）



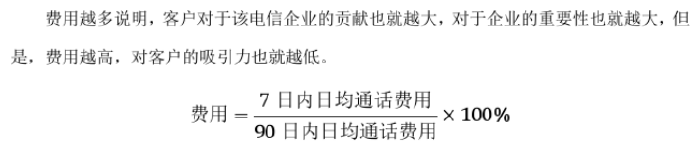
被叫次数（X3）



通话时长（X4）



费用（X5）



**4. 描述分析**

整体分析

x1 x2 x3

Min. :1.000 Min. :0.0000 Min. :0.0000

1st Qu.:1.000 1st Qu.:0.3960 1st Qu.:0.3683

Median :1.000 Median :0.8466 Median :0.8751

Mean :1.494 Mean :0.8597 Mean :0.8207

3rd Qu.:2.000 3rd Qu.:1.1983 3rd Qu.:1.1668

Max. :4.000 Max. :5.3277 Max. :4.2692

x4 x5 y

Min. :0.001113 Min. : 0.0000 Min. :0.0

1st Qu.:0.348605 1st Qu.: 0.3382 1st Qu.:0.0

Median :0.863261 Median : 0.7982 Median :0.0

Mean :0.862159 Mean : 0.8842 Mean :0.4

3rd Qu.:1.219975 3rd Qu.: 1.1565 3rd Qu.:1.0

Max. :5.814740 Max. :13.3367 Max. :1.0

N NU SD MIN MED MAX

x1 1000 1.4940000 0.5969257 1.00000000 1.0000000 4.000000

x2 1000 0.8597230 0.6187290 0.00000000 0.8465707 5.327703

x3 1000 0.8207132 0.5698695 0.00000000 0.8751006 4.269231

x4 1000 0.8621592 0.6286330 0.00111328 0.8632615 5.814740

x5 1000 0.8841646 0.9123904 0.00000000 0.7981747 13.336689

y 1000 0.4000000 0.4901431 0.00000000 0.0000000 1.000000

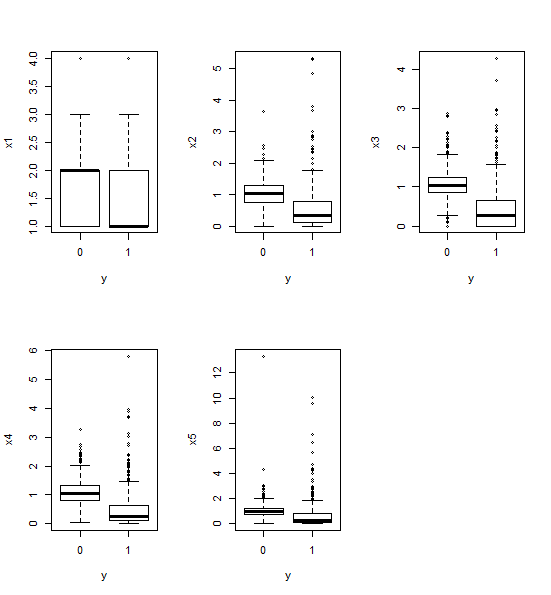
从中可见，1000个样本中，流失的用户涨了40%，客户流失的情况比较严重。

费用的标准差很大，说明不同用户之间的消费水平相差很大，因此，移动通信运营商应该根据这一特点推出不同的套餐服务，以满足不同用户的消费需求。

通话时间以及主叫次数和被叫次数之间的最小值都是非常小，标准差以非常大。

用户的等级平均数为1.49，说明客户的等级总体水平并不高，并且客户的等级标准差为平均数的1/3，说明客户之间的差别不是很大。

将各个解释变量按照流失状态做盒装图对比。



可以看出，流失用户的最近主叫次数，被叫次数，通话时长，费用数据未流失的都比流失的几种。合理的推测，主叫次数、被叫次数、通话时长以及费用对判断客户流失具有重要作用。

**5. 模型分析**

Call:

glm(formula = y ~ x1 + x2 + x3 + x4 + x5, family = binomial(link = logit),

data = a)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.0989 -0.7296 -0.4592 0.6907 4.2962

Coefficients:

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

(Intercept) 2.5426 0.2658 9.565 < 2e-16 \*\*\*

x1 -0.5068 0.1346 -3.766 0.000166 \*\*\*

x2 -0.7039 0.1952 -3.606 0.000311 \*\*\*

x3 -2.1962 0.2428 -9.045 < 2e-16 \*\*\*

x4 -0.5348 0.3075 -1.739 0.082045 .

x5 0.5333 0.1899 2.809 0.004975 \*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1346.02 on 999 degrees of freedom

Residual deviance: 991.11 on 994 degrees of freedom

AIC: 1003.1

Number of Fisher Scoring iterations: 5

进行逻辑回归，从中可以看出，客户等级，主叫次数，被叫次数，通话时长以及费用这几个解释变量都是非常显著的。尤其是被叫次数。说明被叫次数对于预测客户流失状态具有重要的作用，被叫次数越大，客户流失的可能性就越低。其次，客户等级、主叫次数、通话时间越长，客户流失的可能性就越低。费用越高，客户流失的可能性就越高。

以概率阈值\alpha =0.5 预测是否ST并计算混淆矩阵

y1

0 1

0 531 69

1 109 291

Overall accuracy = 0.822

Confusion matrix

Predicted (cv)

Actual [,1] [,2]

[1,] 0.885 0.115

[2,] 0.272 0.728

可以看出，这是一个很好的阈值。

**6. 总结讨论**

本研究分析了移动通信客户相关特征数据，建立了对移动通信客户未来流失情况具有一定预测能力的逻辑回归模型。分析表明客户的被叫次数、主叫次数、通话时长、费用、客户等级等因素是影响客户流失状况的重要因素，并且客户的被叫次数是最重要的因素。

附录：源代码

a<-read.csv('E3.csv')

names(a)=c('x1','x2','x3','x4','x5','y')

par(mfrow=c(2,3))

boxplot(x1~y,xlab="y",ylab="x1",data=a)

boxplot(x2~y,xlab="y",ylab="x2",data=a)

boxplot(x3~y,xlab="y",ylab="x3",data=a)

boxplot(x4~y,xlab="y",ylab="x4",data=a)

boxplot(x5~y,xlab="y",ylab="x5",data=a)

glm=glm(y~x1+x2+x3+x4+x5,family=binomial(link=logit),data=a)

summary(glm)

pred=predict(glm,a)

prob=exp(pred)/(1+exp(pred))

library('DAAG')

y1=1\*(prob>0.5)

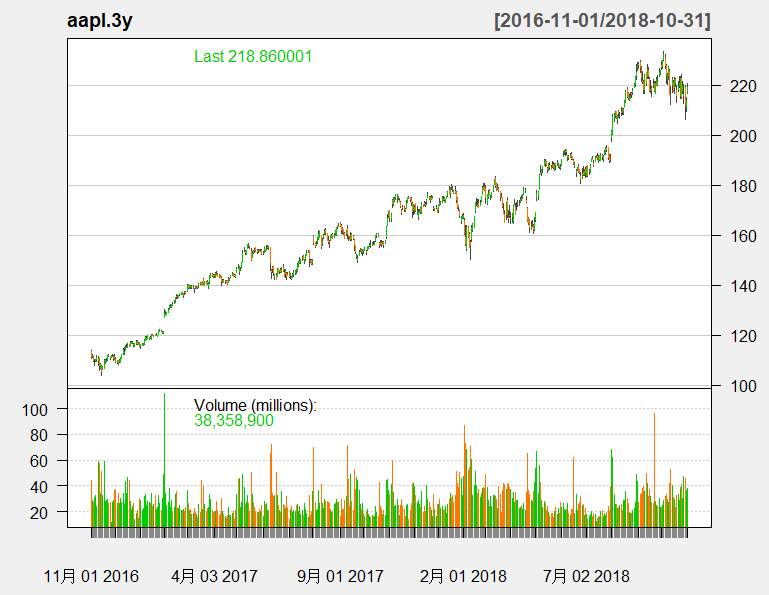
table(a$y,y1)

confusion(a$y,y1)

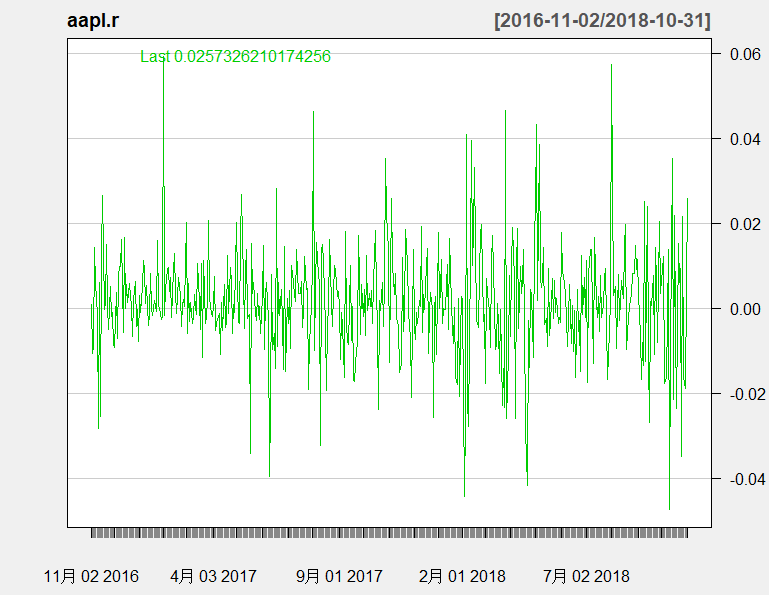
**金融数据分析和可视化: 基于公司收益和股票交易数据**

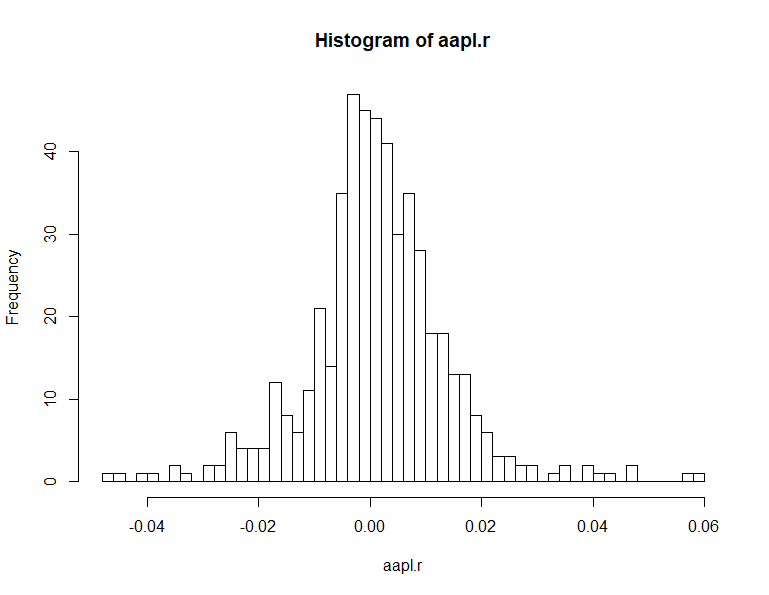
**1. 股票交易数据分析和可视化**

**日收盘价及成交量时序图**

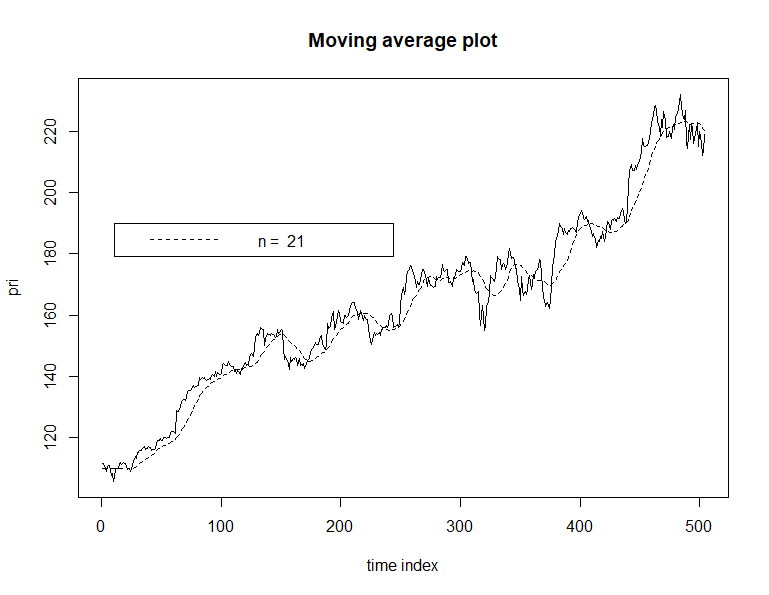


**根据计算出的日对数收益率，画出的时序图以及直方图**

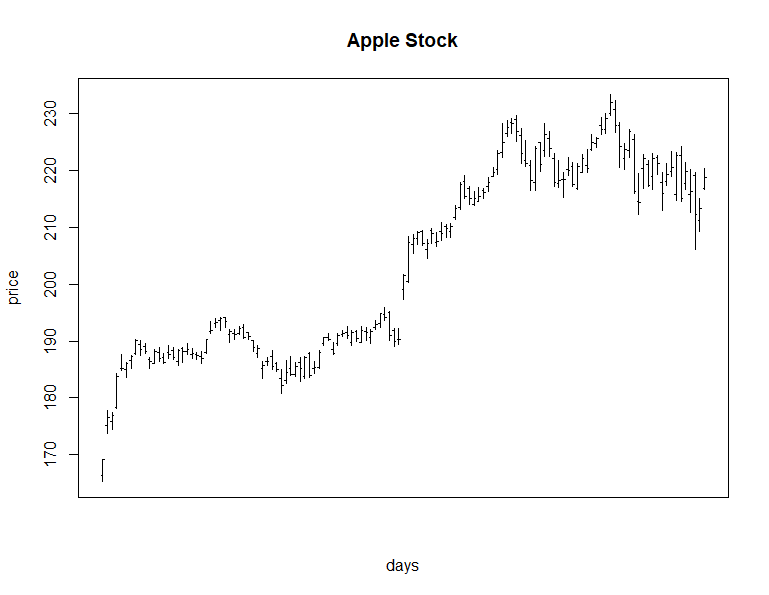




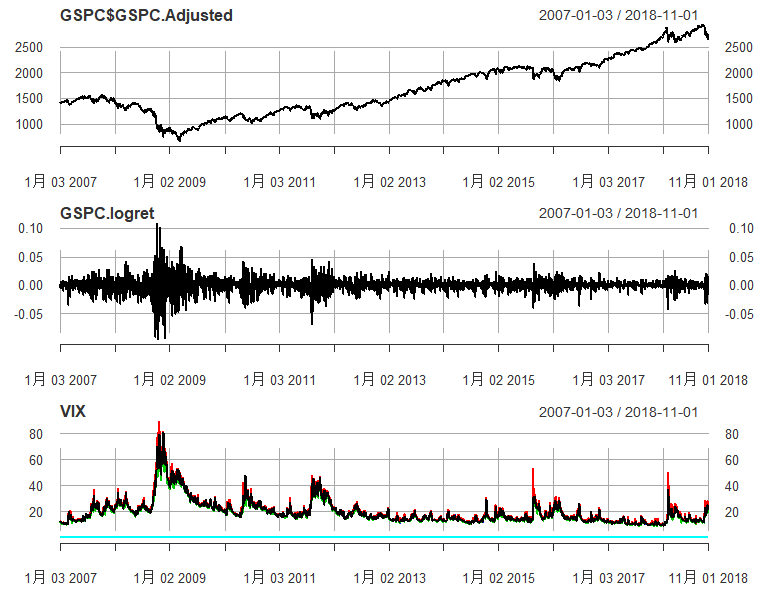
**股票收盘价及移动平均曲线图**



**绘制过去半年的AAPL股票数据条形图**



**比较2007年以来S&P500指数(代码^GSPC)日对数收益率时序图和波动率指数(代码^VIX)时序图**



**计算过去5年Google(股票代码GOOG)股票的Sharpe Ratio**

定义计算

0.02892251

内置函数计算

GOOG.Adjusted

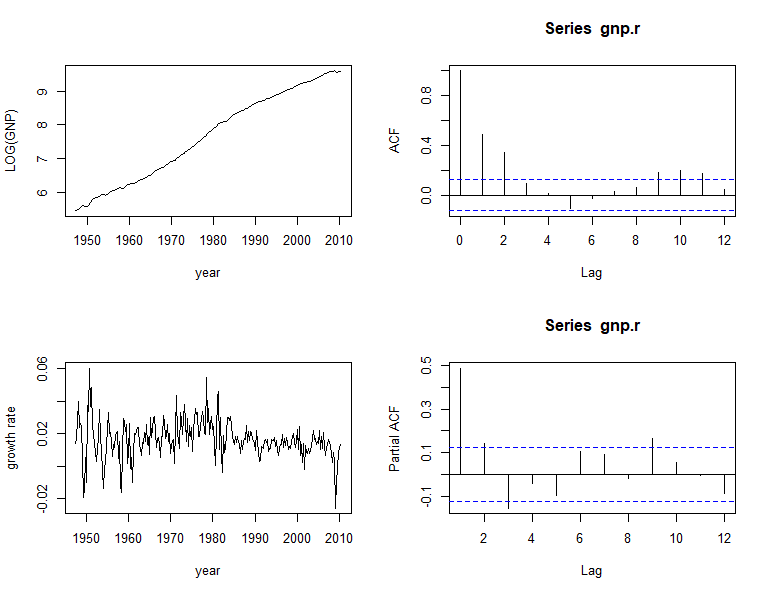
StdDev Sharpe (Rf=0%, p=95%): 0.02892251

内置函数计算年化Sharpe Ratio

GOOG.Adjusted

Annualized StdDev Sharpe (Rf=3%, p=95%): 0.3596502

**2. 基于某公司收益数据的自回归分析**



Call:

arima(x = gnp.r, order = c(3, 0, 0))

Coefficients:

ar1 ar2 ar3 intercept

0.4386 0.2063 -0.1559 0.0163

s.e. 0.0620 0.0666 0.0626 0.0012

sigma^2 estimated as 9.549e-05: log likelihood = 808.56, aic = -1607.12