

回归分析

线性回归模型

回归分析是对客观事物数量依存关系的分析，是统计中的一个常用的方法，被广泛的应用于社会经济现象变量之间的影响因素和关联的研究。根据自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。本部分主要介绍线性回归的原理、估计方法、参数估计量的性质以及在实际中的应用和R语言的实现。

一元线性回归

案例1

为了研究某社区家庭月消费支出与家庭月可支配收入之间的关系，随机抽取并调查了12户家庭的相关数据，见表。通过调查所得的样本数据能否发现家庭消费支出与家庭可支配收入之间的数量关系，以及如果知道了家庭的月可支配收入，能否预测家庭的月消费支出水平呢？

ID	1	2	3	4	5	6	7	8	9	10
Income	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
consume	594	638	1122	1155	1408	1595	1969	2078	2585	2530

案例2

医学上认为一个人的最大心率和年龄是有很大关系的，一般有这样的经验公式 $MaxRate=220-Age$ 来决定的。现在收集了15个来自不同年龄层的人接受了最大心率测试的数据，如表所示。

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39
MaxRate	202	186	187	180	156	169	174	172	153	199	193	174	198	183

概述

一元线性回归是回归分析模型中最简单的一种形式，也是学习回归分析的基础，只有掌握好一元线性回归，才能更好地理解多元线性回归和非线性回归等。

回归分析(regression analysis)是研究一个变量关于另一个（些）变量的具体依赖关系的计算方法和理论。
通常前一个变量被称为被解释变量（Explained Variable）或因变量（Dependent Variable）或响应变量（Response），后一个（些）变量被称为解释变量（Explanatory Variable）或自变量（Independent Variable）或者协变量（Covariate）。因变量往往又更加形象地称之为输出变量（Output variable），自变量称为输入变量（Input variable）。

回归分析关心的是根据解释变量的已知或给定值，考察被解释变量的总体均值

在一个假想的社区有100户家庭组成，要研究该社区每月家庭消费支出与每月家庭可支配收入的关系。现在我们想要研究家庭月收入增加，其平均月消费支出是如何变化的？

表:某社区家庭月可支配收入和消费支出

	每月家庭可支配收入X（元）										
每月家庭 消费支出 Y（元）	800	1100	1400	1700	2000	2300	2600	2900	3200	3500	
	561	638	869	1023	1254	1408	1650	1969	2090	2299	
	594	748	913	1100	1309	1452	1738	1991	2134	2321	
	627	814	924	1144	1364	1551	1749	2046	2178	2530	
	638	847	979	1155	1397	1595	1804	2068	2266	2629	
		935	1012	1210	1408	1650	1848	2101	2354	2860	
		968	1045	1243	1474	1672	1881	2189	2486	2871	
			1078	1254	1496	1683	1925	2233	2552		
			1122	1298	1496	1716	1969	2244	2585		
			1155	1331	1562	1749	2013	2299	2640		
			1188	1364	1573	1771	2035	2310			
			1210	1408	1606	1804	2101				
				1430	1650	1870	2112				
				1485	1716	1947	2200				
						2002					
总计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510	

从表可以看出：

1. 可支配收入相同的家庭，其消费支出不一定相同，即收入和消费支出的关系不是完全确定的；
2. 将居民消费支出看成是其可支配收入的线性函数时，总体回归函数为

$$E(Y|X_i) = \beta_0 + \beta_1X_i$$

其中， β_0, β_1 是未知参数，称为回归系数（regression coefficients）。

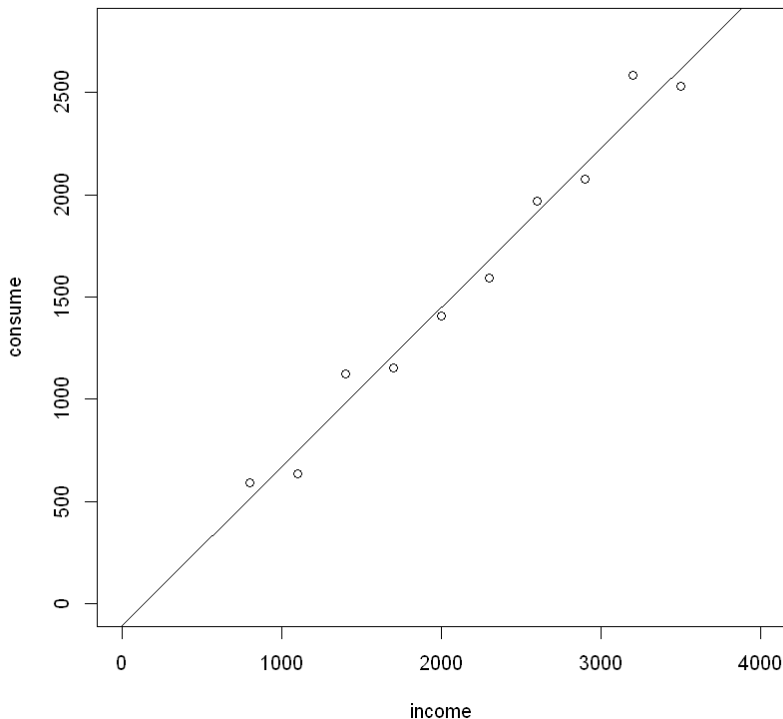
$\mu_i = Y_i - E(Y|X_i)$ 这是一个不可观测的随机变量，称为随机干扰项（stochastic disturbance）或随机误差项（stochastic error）

个别家庭的消费支出为： $Y_i = E(Y|X_i) + \mu_i = \beta_0 + \beta_1X_i + \mu_i$

案例1中样本的散点图。

In [32]:

```
income<-c(800, 1100, 1400, 1700, 2000, 2300, 2600, 2900, 3200, 3500)
consume<-c(594, 638, 1122, 1155, 1408, 1595, 1969, 2078, 2585, 2530)
plot(income, consume, xlim=c(0, 4000), ylim=c(0, 2800))
abline(lm(consume~income))
```



这些散点近似于一条直线，自然的想法是能否画一条直线尽可能好地拟合这些散点，这条直线称为样本回归线（sample regression lines）。

$\hat{Y}_i = f(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 称为样本回归函数（sample regression function, SRF）。

样本回归函数也有如下的随机形式： $\hat{Y}_i = \hat{Y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$ 其中， e_i 称为残差（residual）代表了其他影响 \hat{Y}_i 的随机因素的集合，可以看成是 $\hat{\mu}_i$ 的估计量 $\hat{\mu}_i$

参数估计

最小二乘法估计 (OLS)

$$\min: Q = \sum_{i=1}^n (Y_i - \bar{Y}) = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$


高斯—马尔可夫定理(Gauss-Markov theorem): 在给定经典线性回归的假定下, 最小二乘估计量是具有最小方差的线性无偏估计量 (best linear unbiased estimator, BLUE)

R里OLS的估计可用lm()函数

最大似然估计(MLE)

极大似然估计(Maximum Likelihood Estimation, MLE)的基本原理是, 当从模型总体随机 抽取n组样本观测值后, 最合理的参数估计量应该使得从模型中抽取该n组样本观测值的概率最大。

$$L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2) = P(Y_1, \dots, Y_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2} \Rightarrow$$

数似然函数:

$$l \sim \ln(L) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2} \sum Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

R里求MLE可以使用maxLik包里的maxLik () 函数

参数估计量的概率分布及随机干扰项方差的估计

R里求OLS的方差估计量 $\hat{\sigma}^2$ ，需要用summary()函数先将lm()的结果保存在slm对象里，然后提取sigma成分，即为 $\hat{\sigma}^2$ 。若要计算参数 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准差，先提取出coef矩阵，然后再提取矩阵的第二列即为参数 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准差。

In [33]:

```
lm1<-lm(consume~income)
lm1
```

Call:

```
lm(formula = consume ~ income)
```

Coefficients:

(Intercept)	income
-103.172	0.777

检验

拟合优度的检验

拟合优度检验是对回归拟合值与观测值之间拟合程度的一种检验。度量拟合优度的指标主要是判定系数（可决系数） R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 的取值范围为[0, 1], R^2 越接近1, 说明实际观测点离样本线越近, 拟合优度越高。

R里求拟合优度只要在`summary()`里提取`r.squared`成分即可。

变量显著性检验

已知参数分布为 $\hat{\beta}_1 \sim N(\beta_1, \frac{\hat{\sigma}^2}{\sum x_i^2})$, $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$, 构造检验统计量:

$$t \sim \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / n \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

R里是`summary()`函数会自动提供线性回归的t检验, 可以通过`slm$coef`提取得到回归系数估计值、标准差、t值和相应的p-value。

In [34]:

```
slm<-summary(lm1)
slm
```

Call:

```
lm(formula = consume ~ income)
```

Residuals:

Min	1Q	Median	3Q	Max
-113.54	-82.81	-52.80	69.66	201.74

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-103.17172	98.40598	-1.048	0.325
income	0.77701	0.04249	18.289	8.22e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115.8 on 8 degrees of freedom
Multiple R-squared: 0.9766, Adjusted R-squared: 0.9737
F-statistic: 334.5 on 1 and 8 DF, p-value: 8.217e-08

预测

点预测

对于拟合得到的一元线性回归模型 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, 在给定样本以外的解释变量的观测值 X_0 , 可以得到被解释变量的预测值 \hat{y}_0 , 可以此作为其条件均值 $E(Y|X = X_0)$ 或个值 \hat{y}_0 的一个近似估计, 称之为点预测。

In [35]:

```
$coef[,1][1]+ slm$coef[,1][2]*4000slm
```

Error in parse(text = x, srcfile = src): <text>:1:1: 意外的'\$'
1: \$
^

Traceback:

In []:

```
coef(lm1)[1]+coef(lm1)[2]*4000
```

也可以利用predict()函数，但需要基于回归结果lm1基础上，另外 X_0 的格式需要是数据框格式。

In [36]:

```
predict(lm1, newdata=data.frame(income=4000))
```

1: 3004.86868686869

如果求样本内给定X下的 可以在lm1上调用fitted()函数，就会返回样本内的拟合值

In [37]:

```
fitted(lm1)
```

```
1
518.436363636363
2
751.539393939394
3
984.642424242424
4
1217.74545454545
5
1450.84848484848
6
1683.95151515152
7
1917.05454545455
8
2150.15757575758
9
2383.26060606061
10
2616.36363636364
```

In [38]:

```
resid(lm1)
```

```
1
75.5636363636365
2
-113.539393939394
3
137.357575757576
4
-62.7454545454545
5
-42.8484848484849
6
-88.9515151515152
7
51.9454545454545
8
-72.1575757575757
9
201.739393939394
10
-86.3636363636364
```

区间预测

由于

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0, \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum x_i^2}\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2\right)$$

于是，在 $(1 - \alpha)$ 置信度下，总体均值 $E(Y_0|X_0)$ 的置信区间是：

$$\hat{Y}_0 - t_{1-\frac{\alpha}{2}} \times S\hat{Y}_0 < E(Y_0|X_0) < \hat{Y}_0 + t_{1-\frac{\alpha}{2}} \times S\hat{Y}_0$$

这也称为 $E(Y_0|X_0)$ 的区间预测

在R里求均值预测区间可以用predict()函数，但要在interval参数设为confidence，如果求个值的预测区间也可以用predict()函数，但要在interval参数设为prediction，另外可以从参数level设置不同的置信水平。

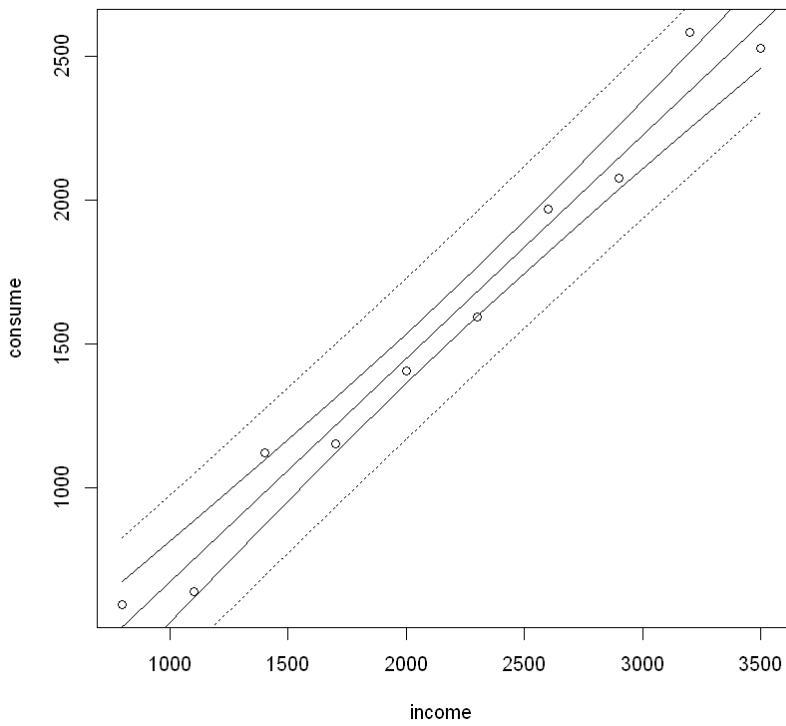
In [39]:

```
predict(lm1, newdata=data.frame(income=4000), interval="prediction")
```

	fit	lwr	upr
1	3004.869	2671.336	3338.401

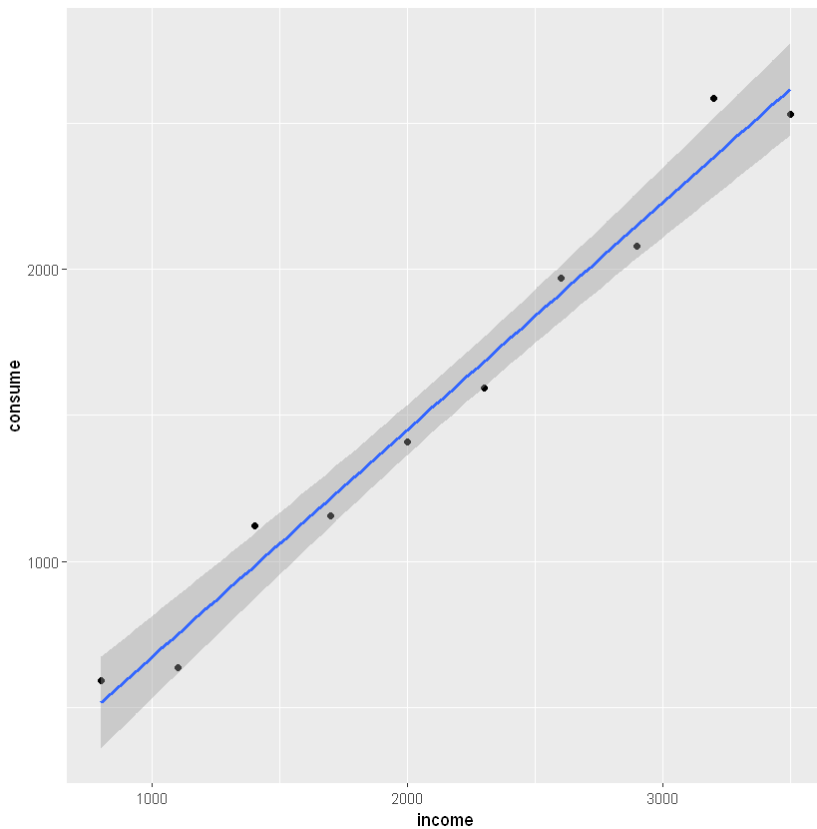
In [40]:

```
opar<-par(no.readonly=TRUE)
par(mfrow=c(1,1))
sx=sort(income) # 把自变量先从小到大排序
conf = predict(lm1, data.frame(income=sx), interval="confidence")
#均值的预测区间
pred = predict(lm1, data.frame(income=sx), interval="prediction")
#个值的预测区间
plot(income, consume);
abline(lm1) #添加回归线
lines(sx, conf[, 2]); lines(sx, conf[, 3]) # 均值置信区间
lines(sx, pred[, 2], lty=3); lines(sx, pred[, 3], lty=3)
par(opar)
```



In [41]:

```
library(ggplot2)
ggplot(mapping=aes(x=income, y=consume)) +
  geom_point() +
  geom_smooth(method=lm)
```



一元线性回归模型案例

医疗事业发展情况，医疗技术的提高、医疗保障水平的提高和覆盖范围的扩大，对居民潜在医疗服务需求产生较大的影响。各地在制定、实施规划的过程中，要按照医疗资源利用情况，对于医疗资源利用率低的医疗机构要适当缩小规模，或与其他医疗机构进行重组和调整；扩大医疗机构规模。要以提高医疗服务工作效率和医疗系统整体功能为主要手段满足增长的医疗服务需求。

为了给制定医疗机构的规划提供依据，分析比较医疗机构与人口数量的关系，建立卫生医疗机构数与人口数的回归模型。以人口数量为自变量，以医疗机构数为因变量，则一元线性回归模型为：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

其中 Y_i 表示卫生医疗机构数， X_i 表示人口数。变量采用四川省2000年各地的截面数据。这里 β_2 为人口每增加一万人医疗机构增加的数量， u_i 为随机误差项，即除了人口数以外，影响医疗机构数量的其他次要的、随机的因素。

四川省2000 年各地区医疗机构数与人口数

place	x	y
成都	1013.3	6304
自贡	315	911
攀枝花	103	934
泸州	463.7	1297
德阳	379.3	1085
绵阳	518.4	1616
广元	302.6	1021
遂宁	371	1375
眉山	339.9	827
宜宾	508.5	1530
广安	438.6	1589
达州	620.1	2403
雅安	149.8	866
巴中	346.7	1223
资阳	488.4	1361
阿坝	82.9	536

In [42]:

```
data=read.table("./data/medicine.txt",head=T)
attach(data)
plot(x, y, xlab="人口数", ylab="医疗机构数")
abline(lm(y~x))
cor(x, y)
```



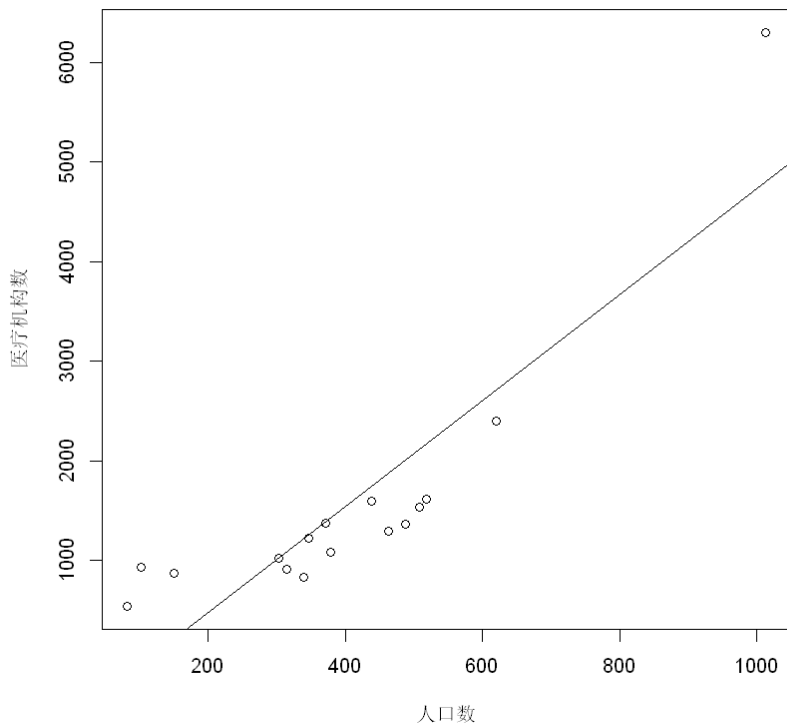
The following object is masked from data (pos = 4):

y

The following object is masked from data (pos = 5):

y

0.882638170249524



读入数据后，为了对解释变量和被解释变量之间的进行初步了解，采用描述统计方法对变量之间的关系进行探索。变量之间的pearson相关系数为0.8826382，说明变量之间有较强的相关关系。通过的散点图，并添加线性趋势直线可以看出，变量之间大都在一条直线附近波动，说明两变量之间存在线性关系。

为了进一步分析人口数的增加所需的医疗机构数增加数量，建立一元线性回归模型，并通过OLS估计方法对参数进行求解。并用使用summary()函数获取参数表及检验结果。Summary()函数的结果中，提供了残差的描述性统计量，参数值、参数的标准误差、t值和t检验p值，残差标准误，可决系数 R_2 和修正后的可决系数 \bar{R}_2 ，F值和F检验的p值。其得到的模型为：

$$Y_i = -587.2682 + 5.3211X_i + u_i$$

In [43]:

```
lm=lm(y~x)
lm.summary=summary(lm)
lm.summary
```



Call:
lm(formula = y ~ x)

Residuals:
 Min 1Q Median 3Q Max
-650.6 -434.6 -167.7 162.6 1499.4

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -587.2682 345.6418 -1.699 0.111
x 5.3211 0.7574 7.026 6e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Residual standard error: 651.3 on 14 degrees of freedom
Multiple R-squared: 0.7791, Adjusted R-squared: 0.7633
F-statistic: 49.36 on 1 and 14 DF, p-value: 5.998e-06

t检验的结果说明人口数量对医疗机构数量的影响是显著的，而截距项的存在却不是显著，这也与事实相符。因此我们应该剔除截距项后重新对回归模型进行求解，采用的模型为： $Y_i = \beta_1 + \beta_2 X_i + u_i$

在lm(y~x)命令中的x前加0即可得到不带截距项的的回归，从summary()函数得到的结果可以看出可决系数和修正后的可决系数都有所提高，F检验的p-value为0，方程整理显著，t检验的结果p-value也为0，说明人口数量对医疗结构数的影响显著。方程的结果为 $Y_i = 4.1860X_i + u_i$

In [44]:

```
lm2=lm(y~0+x)
lm2.summary=summary(lm2)
lm2.summary
```

Call:
lm(formula = y ~ 0 + x)

Residuals:
 Min 1Q Median 3Q Max
-683.44 -564.47 -246.33 -86.26 2062.32

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
x 4.1860 0.3785 11.06 1.31e-08 ***

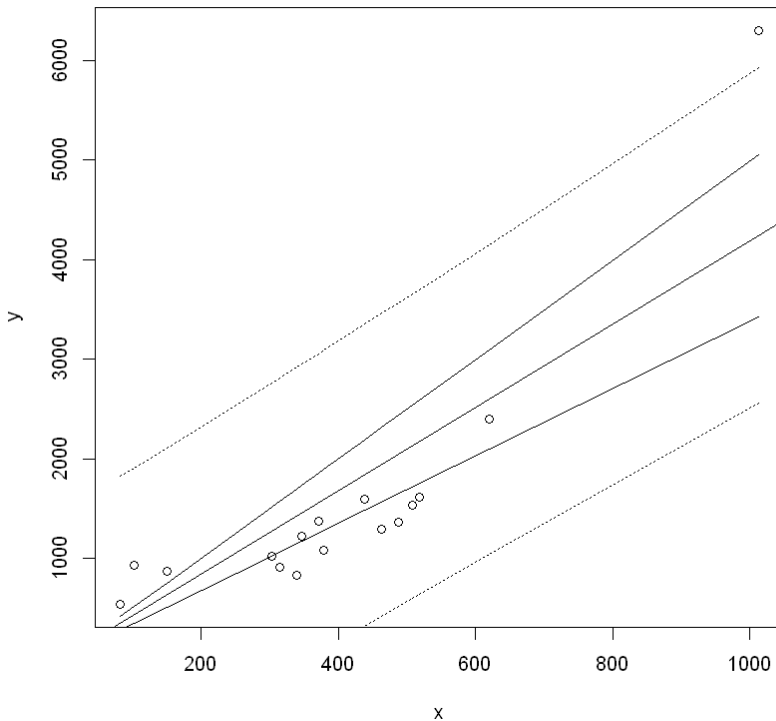
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 691 on 15 degrees of freedom
Multiple R-squared: 0.8907, Adjusted R-squared: 0.8
835
F-statistic: 122.3 on 1 and 15 DF, p-value: 1.309e-08

为了更好的规划医疗结构，人们常常需要根据人口数量预测医疗机构数。预测可以通过predict函数来实现，并且可将观测值，回归直线，均值预测区间，个值预测区间画在同意张图上。

In [45]:

```
opar<-par(no.readonly=TRUE)
par(mfrow=c(1,1))
sx=sort(x) #把自变量先从小到大排序
conf = predict(lm2,data.frame(x=sx),interval="confidence")
#求均值的预测区间
pred = predict(lm2,data.frame(x=sx),interval="prediction")
#求个值的预测区间
plot(x,y); #画散点图
abline(lm2) # 添加回归线
lines(sx, conf[, 2]); lines(sx, conf[, 3]) #用实线表示预测均值的95%置信带
lines(sx, pred[, 2], lty=3); lines(sx, pred[, 3], lty=3)
#用虚线表示预测均值的95%置信带
par(opar)
```



多元线性回归

为了研究影响中国税收收入增长的主要原因，分析中央和地方税收收入的增长规律，预测中国税收未来的增长趋势，需要建立计量经济模型。影响中国税收收入增长的因素很多，但据分析主要的因素可能有：（1）从宏观经济看，经济整体增长是税收增长的基本源泉。（2）公共财政的需求，税收收入是财政收入的主体，社会经济的发展和社会保障的完善等都对公共财政提出要求，因此对预算支出所表现的公共财政的需求对当年的税收收入可能会有一定的影响。（3）物价水平。我国的税制结构以流转税为主，以现行价格计算的GDP等指标和经营者的收入水平都与物价水平有关。（4）税收政策因素。选择包括中央和地方税收的“国家财政收入”中的“各项税收”（简称“税收收入”）作为被解释变量，以反映国家税收的增长；选择“国内生产总值（GDP）”作为经济整体增长水平的代表；选择中央和地方“财政支出”作为公共财政需求的代表；选择“商品零售物价指数”作为物价水平的代表。由于财税体制的改革难以量化，而且1985年以后财税体制改革对税收增长影响不是很大，可暂不考虑税制改革对税收增长的影响。所以解释变量设定为可观测的“国内生产总值”、“财政支出”、“商品零售物价指数”等变量。

年份	tax	GDP	expand	CPI
1978	519.28	3645.22	1122.09	100.7
1979	537.82	4062.58	1281.79	102
1980	571.7	4545.62	1228.83	106
1981	629.89	4889.46	1138.41	102.4
1982	700.02	5330.45	1229.98	101.9
1983	775.59	5985.55	1409.52	101.5
1984	947.35	7243.75	1701.02	102.8
1985	2040.79	9040.74	2004.25	108.8
1986	2090.73	10274.38	2204.91	106
1987	2140.36	12050.62	2262.18	107.3
1988	2390.47	15036.82	2491.21	118.5
1989	2727.4	17000.92	2823.78	117.8
1990	2821.86	18718.32	3083.59	102.1
1991	2990.17	21826.2	3386.62	102.9
1992	3296.91	26937.28	3742.2	105.4
1993	4255.3	35260.02	4642.3	113.2
1994	5126.88	48108.46	5792.62	121.7
1995	6038.04	59810.53	6823.72	114.8
1996	6909.82	70142.49	7937.55	106.1
1997	8234.04	78060.85	9233.56	100.8
1998	9262.8	83024.33	10798.18	97.4
1999	10682.58	88479.16	13187.67	97
2000	12581.51	98000.48	15886.5	98.5
2001	15301.38	108068.2	18902.58	99.2
2002	17636.45	119095.68	22053.15	98.7
2003	20017.31	134976.97	24649.95	99.9
2004	24165.68	159453.6	28486.89	102.8
2005	28778.54	183617.37	33930.28	100.8

年份	tax	GDP	expand	CPI
2006	34804.35	215904.41	40422.73	101
2007	45621.97	266422	49781.35	103.8
2008	54223.79	316030.34	62592.66	105.9
2009	59521.59	340319.95	76299.93	98.8
2010	73210.79	399759.54	89874.16	103.1
2011	89738.39	468562.38	109247.79	104.9
2012	100614.28	516282.06	125052.07	102

例中自变量个数不止一个，该如何建模分析？这就需要利用**多元回归分析** 方法。可以建立模型: $Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \varepsilon_i$

多元线性回归模型及假定

模型形式

线性模型的一般形式是：

$$Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + \cdots + \beta_kX_{ki} + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

矩阵形式记为 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 样本回归模型为: $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$

总体回归方程为 $E(\mathbf{Y} | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$,样本回归方程为 $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

假设

经典线性回归模型必须满足的假定条件：

- （1）零均值假定。假定随机干扰项 e 的期望向量或均值向量为零，即

$$E(\boldsymbol{\varepsilon}) = E \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} E\varepsilon_1 \\ E\varepsilon_2 \\ \vdots \\ E\varepsilon_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}$$

(2) 同方差和无序列相关假定。假定随机干扰项 ε 不存在序列相关且方差相同，即

$$\text{Var}(\boldsymbol{\varepsilon}) = E \left[(\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon})' \right] = E \left(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' \right) = \sigma^2 \mathbf{I}_n$$

(3) 随机干扰项 ε 与解释变量相互独立。即 $E(\mathbf{X}'\boldsymbol{\varepsilon}) = \mathbf{0}$

(4) 无多重共线性。假定数据矩阵 \mathbf{X} 列满秩，即 $\text{Rank}(\mathbf{X}) = k$

参数估计

普通最小二乘估计 (OLS)

R里OLS的估计可用lm()函数

极大似然估计(MLE)

$$\text{对数似然函数 } l = \ln(L) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

R里求MLE可以使用maxLik包里的maxLik()函数

In [46]:

```
#install.packages("maxLik")
```

In [47]:

```
library(maxLik)
dat=read.csv("./data/tax.csv", header=T)
lm3=lm(tax~GDP+expand+CPI, data=dat)
lm3
```

Call:

```
lm(formula = tax ~ GDP + expand + CPI, data = dat)
```

Coefficients:

(Intercept)	GDP	expand	CPI
-6.616e+03	4.729e-02	6.140e-01	5.821e+01

In [48]:

```
attach(dat)
loglik=function(para) {
  #para[5]=sigma
  N=length(dat$tax)
  e=dat$tax-para[1]-para[2]*dat$GDP-para[3]*dat$expand-para[4]*dat$CP
  I
  ll=-N*log(sqrt(2*pi)*para[5])-(1/(2*para[5]^2))*sum(e^2)
  #ll=-0.5*N*log(2*pi)-0.5*N*log(para[5]^2)-0.5*sum(e^2/para[5]^2)
  return(ll)
}

mle3=maxLik(loglik, start=c(0.1, 1, 1, 1, 1, 1), iterlim=10000)
mle3
coef(mle3)
summary(mle3)
detach()
```


[illegible]

Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -18288640615 (5 free parameter(s))
Estimate(s): 0.1003639 0.05877031 0.7948104 0.9995549 0.2983986

0.10036393810406 0.0587703099586589 0.794810418277517
0.999554904684784 0.298398604492347

Warning message in sqrt(diag(vc)):
"产生了NaNs"
Warning message in sqrt(diag(vc)):
"产生了NaNs"

Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -18288640615
5 free parameters
Estimates:

	Estimate	Std. error	t value	Pr(> t)
[1,]	1.004e-01	3.621e-04	277.2	<2e-16 ***
[2,]	5.877e-02	3.024e-06	19435.1	<2e-16 ***
[3,]	7.948e-01	1.280e-05	62079.7	<2e-16 ***
[4,]	9.996e-01	8.540e-04	1170.4	<2e-16 ***
[5,]	2.984e-01	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In [49]:

```
log(sqrt(2*pi))*0.1
```

0.0918938533204673

模型检验

(一) 拟合优度检验

拟合优度检验是对回归拟合值与观测值之间拟合程度的一种检验。度量拟合优度的指标主要是判定系数（可决系数） R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \text{ 样本可决系数}$$

$$\begin{aligned} R^{-2} &= 1 - \left(\frac{n-1}{n-k} \right) \frac{ESS}{TSS} \\ &= 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2) \\ &= R^2 - \left(\frac{k-1}{n-k} \right) (1 - R^2) \end{aligned}$$

$$R^{-2} \leq R^2$$

对于多元线性回归，建议不要直接使用可决系数，应该使用修正后的可决系数

R里求拟合优度只要在上文的slm对象里提取r.squared和adj.r.squared成分即可

In [50]:

```
slm3<-summary(lm3)
slm3$r.squared
slm3$adj.r.squared
```

0.998850835828579

0.998739626392635

(二) 方程整体显著性检验

方程的整体显著性检验，旨在对模型中被解释变量与解释变量之间的线性关系在总体上是否显著成立作出推断。一般使用 F 检验。

检验统计量：
$$F = \frac{ESS/k}{RSS/(n-k-1)} \sim F(k, n-k-1)$$

R里做回归的整体显著性检验，可以用summary()函数返回

In [51]:

```
summary(lm3)
```

Call:

```
lm(formula = tax ~ GDP + expand + CPI, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2551.29	-511.58	12.51	458.76	3033.79

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.616e+03	2.951e+03	-2.242	0.0323 *
GDP	4.729e-02	8.995e-03	5.257	1.03e-05 ***
expand	6.140e-01	3.880e-02	15.825	< 2e-16 ***
CPI	5.821e+01	2.771e+01	2.101	0.0439 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 946 on 31 degrees of freedom
Multiple R-squared: 0.9989, Adjusted R-squared: 0.9987
F-statistic: 8982 on 3 and 31 DF, p-value: < 2.2e-16



(三) 单个变量显著性检验

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad i = 1, 2, \cdots, n$$

检验统计量:

$$t = \frac{\hat{\beta}_j - \beta_j}{S\hat{\beta}_j} \sim t(n - k - 1)$$

In [52]:

```
summary(lm3)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.616014e+03	2.951208e+03	-2.241798	3.227172e-02
GDP	4.728764e-02	8.995209e-03	5.256980	1.027878e-05
expand	6.139577e-01	3.879721e-02	15.824790	2.125661e-16
CPI	5.820884e+01	2.770822e+01	2.100779	4.388931e-02

预测

(一) 单值预测

针对线性回归模型 $Y = X\beta + \varepsilon$ ，对给定的解释变量矩阵， $X_0 = \left(1, X_{20}, X_{30}, \cdots, X_{k0}\right)_{1 \times k}$ ，假设在预测期或预测范围内，有关系式 $Y_0 = X_0\beta + \varepsilon_0$ 。如果代入到样本回归模型 $\hat{Y} = X\hat{\beta}$ 中可得， $\hat{Y}_0 = X_0\hat{\beta}$ 。与一元线性回归模型类似， \hat{Y}_0 是 Y_0 的点估计值，也是 $E(Y_0)$ 的点估计值。

在R里求 Y_0 ，可以将给定的 X_0 代入到回归模型中。

也可以利用predict()函数，但需要基于回归结果lm1基础上，另外 X_0 的格式需要是数据框格式。

In [53]:

```
coef(lm3)
```

(Intercept)

-6616.01371155807

GDP

0.0472876393628013

expand

0.613957719973698

CPI

58.2088426107878

In [54]:

```
coef(lm3)[1]+coef(lm3)[2]*520000+coef(lm3)[3]*130000+coef(lm3)[4]*103
```

(Intercept): 103783.57314259

In [55]:

```
predict(lm3, newdata=data.frame(GDP=520000, expand=130000, CPI=103))
```

1: 103783.57314259

(二) 区间预测

Y_0 的 $1 - \alpha$ 的置信区间为 $\hat{Y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$

$E(Y_0)$ 的 $1 - \alpha$ 的置信区间为 $\hat{Y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$

在R里求均值预测区间可以用predict()函数，但要在interval参数设为confidence，如果求个值的预测区间也可以用predict()函数，但要在interval参数设为prediction。

In [56]:

```
predict(lm3, newdata=data.frame(GDP=520000, expand=130000, CPI=103),
        interval="confidence")
```

	fit	lwr	upr
1	103783.6	102275.7	105291.4

In [57]:

```
predict(lm3, newdata=data.frame(GDP=520000, expand=130000, CPI=103),
        interval="prediction")
```

	fit	lwr	upr
1	103783.6	101335	106232.2

In [58]:

```
predict(lm3, interval="confidence")
```



	fit	lwr	upr
1	106.9064	-484.77392	698.5867
2	300.3629	-258.89063	859.6165
3	523.5249	24.05868	1022.9911
4	274.7184	-259.58723	809.0240
5	322.6875	-222.31529	867.6902
6	440.6120	-114.98205	996.2061
7	754.7495	230.95263	1278.5464
8	1375.1484	867.61035	1882.6864
9	1393.6963	915.48918	1871.9034
10	1588.5233	1120.16408	2056.8826
11	2522.2875	1681.36040	3363.2145
12	2778.6029	1972.63943	3584.5663
13	2105.4482	1628.84515	2582.0512
14	2485.0272	2042.77020	2927.2841
15	3090.5513	2699.97303	3481.1295
16	4490.7663	3905.61724	5075.9154
17	6299.3617	5268.77720	7329.9462
18	7084.1358	6348.20206	7820.0695
19	7750.1374	7232.39305	8267.8817
20	8611.7664	8050.16943	9173.3634
21	9609.1781	8982.28492	10236.0713
22	11310.8864	10715.28402	11906.4888
23	13505.4080	12986.15300	14024.6629
24	15873.9785	15403.90869	16344.0482
25	18300.6543	17830.81376	18770.4949
26	20715.8190	20259.57746	21172.0606
27	24397.7857	23895.64388	24899.9275
28	28766.0269	28229.80924	29302.2446
29	34290.5364	33683.86618	34897.2067

	fit	lwr	upr
30	42588.1757	41794.24381	43382.1077
31	52921.8783	52116.13951	53727.6171
32	62072.8781	61400.52711	62745.2290
33	73467.9373	72707.39700	74228.4776
34	88720.8268	87719.72943	89721.9241
35	101064.8464	99739.49306	102390.1997

In [59]:

```
predict(lm3, interval="prediction")
```



```
Warning message in predict.lm(lm3, interval = "prediction"):  
"predictions on current data refer to _future_ responses  
"
```

	fit	lwr	upr
1	106.9064	-1911.0860	2124.899
2	300.3629	-1708.3611	2309.087
3	523.5249	-1469.3811	2516.431
4	274.7184	-1727.2032	2276.640
5	322.6875	-1682.1157	2327.491
6	440.6120	-1567.0962	2448.320
7	754.7495	-1244.3930	2753.892
8	1375.1484	-619.7959	3370.093
9	1393.6963	-593.9882	3381.381
10	1588.5233	-396.8149	3573.862
11	2522.2875	417.6816	4626.893
12	2778.6029	687.7215	4869.484
13	2105.4482	118.1490	4092.747
14	2485.0272	505.6841	4464.370
15	3090.5513	1122.1105	5058.992
16	4490.7663	2474.6792	6506.853
17	6299.3617	4112.0547	8486.669
18	7084.1358	5019.2372	9149.034
19	7750.1374	5752.5721	9747.703
20	8611.7664	6602.3887	10621.144
21	9609.1781	7580.5819	11637.774
22	11310.8864	9291.7406	13330.032
23	13505.4080	11507.4507	15503.365
24	15873.9785	13888.2360	17859.721
25	18300.6543	16314.9661	20286.343
26	20715.8190	18733.3045	22698.334
27	24397.7857	22404.2075	26391.364
28	28766.0269	26763.5941	30768.460
29	34290.5364	32268.0982	36312.975

	fit	lwr	upr
30	42588.1757	40501.9026	44674.449
31	52921.8783	50831.0835	55012.673
32	62072.8781	60029.7766	64115.979
33	73467.9373	71394.1414	75541.733
34	88720.8268	86547.2575	90894.396
35	101064.8464	98724.1693	103405.523

多元线性回归综合案例

近年来，中国旅游业一直保持高速发展，旅游业作为国民经济新的增长点，在社会经济发展中的作用日益显现。改革开放20 多年来，特别是进入90 年代后，中国的国内旅游收入年均增长14.4%，远高于同期GDP 9.76%的增长率。为了规划中国未来旅游产业的发展，需要定量地分析影响中国旅游市场发展的主要因素。经分析，影响国内旅游市场收入的主要因素 Y_t ，除了国内旅游人数和旅游支出以外，还可能与相关基础设施有关。为此，考虑的影响因素主要有国内旅游人数 X_1 ，城镇居民人均旅游支出 X_2 ，农村居民人均旅游支出 X_3 ，并以公路里程 X_4 和铁路里程 X_5 作为相关基础设施的代表。为此设定了如下对数形式的计量经济模型：

$$Y_t = \beta_0 + \beta_1X_{1t} + \beta_2X_{2t} + \beta_3X_{3t} + \beta_4X_{4t} + \beta_5X_{5t} + \varepsilon_t$$

年份 (亿元)	国内旅游收入 (亿元)	国内旅游人数 (万人次)	城镇居民人均 旅游支出 (元)	农村居民人均 旅游支出 (元)	公路里程 (万公里)	铁路里程 (万公里)
1994	1023.5	52400	414.7	54.9	111.78	5.9
1995	1375.7	62900	464	61.5	115.7	5.97
1996	1638.4	63900	534.1	70.5	118.58	6.49
1997	2112.7	64400	599.8	145.7	122.64	6.6
1998	2391.2	69450	607	197	127.85	6.64
1999	2831.9	71900	614.8	249.5	135.17	6.74
2000	3175.5	74400	678.6	226.6	140.27	6.87
2001	3522.4	78400	708.3	212.7	169.8	7.01
2002	3878.4	87800	739.7	209.1	176.52	7.19
2003	3442.3	87000	684.9	200	180.98	7.3

读入数据后，可先采用描述性统计的方法对解释变量与被解释变量之间的关系进行探索。各个自变量与被解释变量之间的pearson相关系数接近于1，说明各个自变量与因变量之间都有较强的相关关系。通过绘制各个自变量与因变量的散点图，并添加线性趋势直线，如图所示。可以看出，各个变量与因变量之间大都在一条直线附近波动，说明两变量之间存在线性关系。

In [60]:

```
data=read.table("./data/travel.txt", head=1)
attach(data)
par(mfrow=c(2, 3))
plot(x1, y, xlab="国内旅游人数", ylab="国内旅游收入");abline(lm(y~x1))
plot(x2, y, xlab="城镇居民人均旅游支出", ylab="国内旅游收入");abline(lm(y~x2))
plot(x3, y, xlab="农村居民人均旅游支出", ylab="国内旅游收入");abline(lm(y~x3))
plot(x4, y, xlab="公路里程", ylab="国内旅游收入");abline(lm(y~x4))
plot(x5, y, xlab="铁路里程", ylab="国内旅游收入");abline(lm(y~x5))
c(cor(x1, y), cor(x2, y), cor(x3, y), cor(x4, y), cor(x5, y))
pairs(data[, 3:7])
```



The following object is masked from data (pos = 3):

y

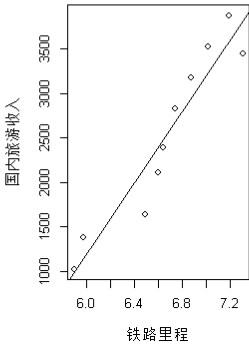
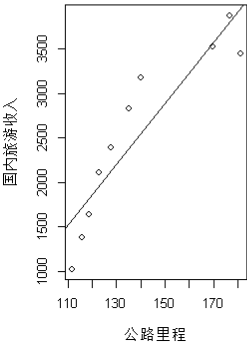
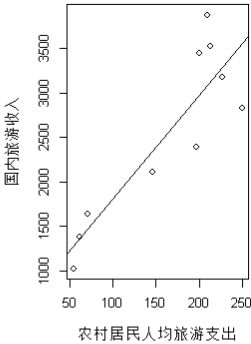
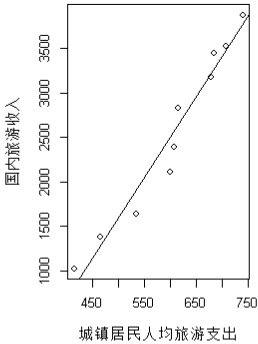
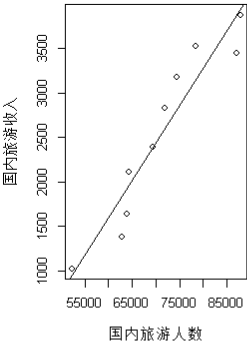
The following objects are masked from data (pos = 5):

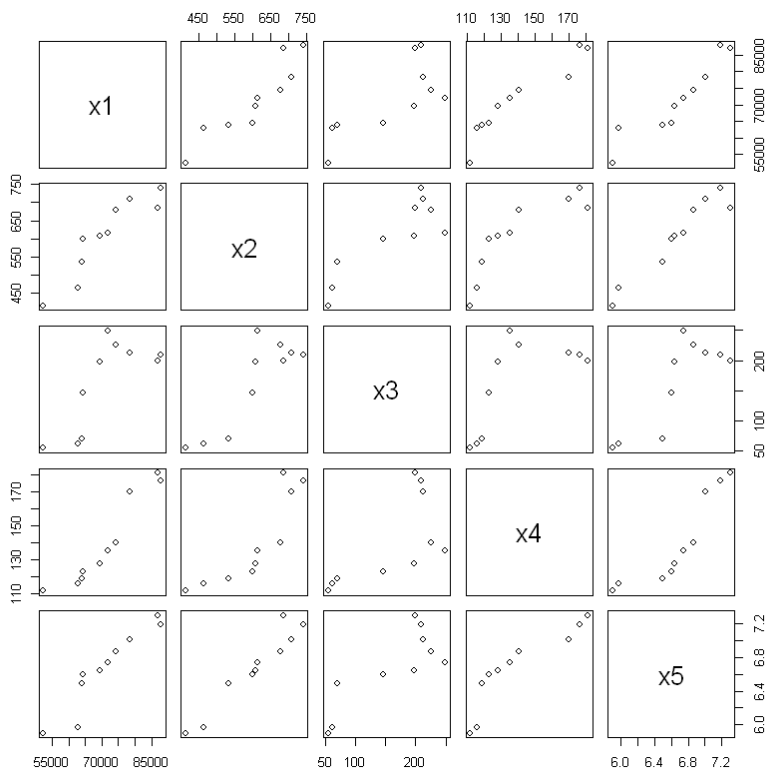
x1, x2, x3, x4, x5, y, year

The following objects are masked from data (pos = 6):

x1, x2, x3, x4, x5, y, year

0.95064557762311 0.977673291221279 0.878329860960568
0.916213790198967 0.951509291022565





为了进一步分析各解释变量对国内旅游收入的影响，建立多元线性回归模型，并通过OLS估计方法对参数进行求解。并用使用summary()函数获取参数表及检验结果。

In [61]:

```
lm=lm(y~x1+x2+x3+x4+x5)
lm.summary=summary(lm)
lm.summary
```

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

Residuals:

	1	2	3	4	5	6	7
8	9	10					
	47.98	-38.43	55.94	-70.41	-110.14	45.56	91.39
	-3						
	4.75	53.34	-40.48				

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-274.37728	1316.68972	-0.208	0.8451
x1	0.01309	0.01269	1.031	0.3607
x2	5.43819	1.38040	3.940	0.0170 *
x3	3.27177	0.94421	3.465	0.0257 *
x4	12.98624	4.17793	3.108	0.0359 *
x5	-563.10774	321.28299	-1.753	0.1545

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.1 on 4 degrees of freedom
Multiple R-squared: 0.9954, Adjusted R-squared: 0.9897
F-statistic: 173.4 on 5 and 4 DF, p-value: 9.19e-05



$$Y_t = -274.37728 + 0.0131X_{1t} + 5.4382X_{2t} + 3.2718X_{3t} + 12.9862X_{4t} - 563.10774X_{5t} + \varepsilon_t$$

得到回归模型后，要对模型进行检验

In [62]:

```
lm.summary$r.squared  
lm.summary$adj.r.squared
```

0.995406326047704

0.989664233607334

接下来，方程的整体显著性F检验显示F检验统计量为173.4，两个自由度为5和4，对应的p-value为0，说明方程整体是显著的。

对单个变量进行显著性检验。t检验的结果在summary()函数所得的参数表中。国内旅游人数，公路里程对国内旅游收入的影响不显著。城镇居民人均旅游支出，农村居民人居旅游支出，公路里程对国内旅游收入有显著影响。

t检验的结果说明了国内旅游人数，公路里程对国内旅游收入的影响不显著，这似乎与事实相互违背。但仔细观察可以发现，即国内旅游人数与城镇居民人均旅游支出和农村居民人居旅游支出所包含的信息具有重叠，且人均旅游支出与国内旅游收入更具有直接关系，通过计算相关系数可以看出国内旅游人数与城镇居民人均旅游支出和农村居民人居旅游支出成线性相关关系。同样的，公路旅程与铁路旅程也具有这样的关系。这实际是线性回归违背经典假设的一种情况—多重共线性，但通过t检验可以剔除信息重叠的变量。

In [63]:

```
c(cor(x1, x2), cor(x1, x3), cor(x4, x5))
```



0.9188508283606 0.751959906177393 0.897708265128704

由于截距项不显著可能是变量过多的信息重复所导致。因此剔除变量 X_1 ，变量 X_5 ，但因保留截距项，重新对回归模型进行求解，采用的模型为：

$$Y_t = \beta_0 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t$$

In [64]:

```
lm2=lm(y~x2+x3+x4)
lm2.summary=summary(lm2)
lm2.summary
```

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-115.74	-83.44	0.28	72.99	121.55

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2441.161	296.039	-8.246	0.000172	***
x2	4.216	1.069	3.945	0.007581	**
x3	3.222	1.050	3.068	0.022008	*
x4	13.629	2.904	4.693	0.003351	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.6 on 6 degrees of freedom

Multiple R-squared: 0.9914, Adjusted R-squared: 0.9872

F-statistic: 231.8 on 3 and 6 DF, p-value: 1.365e-06

可以看出方程的F值为231.8，p-value为0，方程整体检验。截距项和各个变量的t检验p-value均小于0.05，即城镇居民人均旅游支出，农村居民人均旅游支出，公路里程对国内旅游收入有显著影响。

$$Y_t = -2441.161 + 4.216X_{2t} + 3.222X_{3t} + 13.629X_{4t} + \varepsilon_t$$

在决策中我们常常希望通过城镇居民人均旅游支出 X_2 和农村居民人均旅游支出 X_3 ，公路里程 X_4 预测国内旅游收入 Y_t 。通过predict()函数，我们不仅可以求出均值和个值的预测值，还可以求出其预测区间。

In [65]:

```
conf = predict(lm2, data.frame(x2, x3, x4), interval="confidence") #求均值的预测区间
conf
```

	fit	lwr	upr
1	1007.512	801.3273	1213.696
2	1290.046	1134.2485	1445.843
3	1653.829	1472.0751	1835.582
4	2228.438	2071.1553	2385.721
5	2495.087	2361.0663	2629.107
6	2796.889	2580.9591	3012.819
7	3061.588	2911.4596	3211.716
8	3544.481	3408.8915	3680.071
9	3756.848	3586.4904	3927.206
10	3557.284	3359.5542	3755.013

In [66]:

```
pred = predict(lm2, data.frame(x2, x3, x4), interval="prediction") #求个值的预测区间
pred
```

	fit	lwr	upr
1	1007.512	665.3736	1349.649
2	1290.046	975.6905	1604.401
3	1653.829	1325.8336	1981.824
4	2228.438	1913.3440	2543.532
5	2495.087	2190.9357	2799.238
6	2796.889	2448.7910	3144.987
7	3061.588	2750.0032	3373.172
8	3544.481	3239.6353	3849.327
9	3756.848	3435.0281	4078.668
10	3557.284	3220.1734	3894.394

线性回归模型的扩展

上面提到的线性回归都是在经典假设下进行的。实际中经典假设不一定都能完全符合。如果不符合经典假设情况下该怎么处理？接下来主要涉及多重共线性、异方差和序列相关的检验和克服方法。

多重共线性

主要涉及什么是多重共线性？为何会出现多重共线性？如果出现多重共线性会有什么样的后果？如何检验多重共线性？如果存在多重共线性，该如何克服？

问题的提出

影响财政收入的因素众多复杂，但是通过研究经济理论对财政收入的解释以及对实践的考察，我们选取影响财政收入的因素为工业总产值（industry）、农业总产值（agriculture）、建筑业总产值(construction)、社会商品零售总产值(consumption)、人口总数(pop)和受灾面积(disaster)。将这六个变量作为解释变量，财政收入作为被解释变量，利用1998~2012年数据建立中国国家财政收入计量经济模型，资料如下表。

年份	工业	农业	建筑	消费	人口	受灾	财政收入
1998	34018.4	14241.9	10062	39229.3	124761	50145	9876
1999	35861.5	14106.2	11152.9	41920.4	125786	49980	11444.1
2000	40033.6	13873.6	12497.6	45854.6	126743	54688	13395.2
2001	43580.6	14462.8	15361.6	49435.9	127627	52215	16386
2002	47431.3	14931.5	18527.2	53056.6	128453	46946	18903.6
2003	54945.5	14870.1	23083.9	57649.8	129227	54506	21715.3
2004	65210	18138.4	29021.5	65218.5	129988	37106	26396.5
2005	77230.8	19613.4	34552.1	72958.7	130756	38818	31649.3
2006	91310.9	21522.3	41557.2	82575.5	131448	41091	38760.2
2007	110534.9	24658.2	51043.7	96332.5	132129	48992	51321.8
2008	130260.2	28044.2	62036.8	111670.4	132802	39990	61330.4
2009	135240	30777.5	76807.7	123584.6	133450	47214	68518.3
2010	160722.2	36941.1	96031.1	140758.7	134091	37426	83101.5
2011	188470.2	41988.6	116463.3	168956.6	134735	32471	103874.4
2012	199671	46940.5	137217.9	190423.8	135404	24960	117253.5

In [67]:

```
dat<-read.csv(file="./data/11-1.csv",header=T)
lm3=lm(revenue~industry+agriculture+construction+consumption+pop+disaster,
data=dat)
summary(lm3)
```

Call:

```
lm(formula = revenue ~ industry + agriculture + construction +
  consumption + pop + disaster, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1135.26	-418.22	22.56	374.25	1019.92

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.821e+04	7.081e+04	0.963	0.36359
industry	1.297e-01	1.032e-01	1.257	0.24426
agriculture	-7.065e-02	7.393e-01	-0.096	0.92622
construction	4.465e-02	2.038e-01	0.219	0.83206
consumption	6.011e-01	1.501e-01	4.005	0.00392 **
pop	-7.020e-01	5.007e-01	-1.402	0.19846
disaster	4.324e-02	5.569e-02	0.776	0.45986

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 811.3 on 8 degrees of freedom

Multiple R-squared: 0.9997, Adjusted R-squared: 0.9995

F-statistic: 4294 on 6 and 8 DF, p-value: 1.392e-13

从回归结果可以看出，调整后的 \bar{R}^2 是0.9995，说明拟合的非常好；F检验的p-value为1.392e-13，说明是显著的。但是除了消费变量（consumption）外，其余变量均不显著，这个结果很奇怪。这是为什么？难道该模型真的只有消费变量（consumption）对财政收入有影响？这明显不符合实际情况。实际上该模型存在着多重共线性，才导致这个奇怪的结果。

多重共线性定义及后果

“多重共线性”一词来由R.Frisch于1934年提出。它原指模型的解释变量间存在线性关系。

对于回归模型 $Y = X\beta + \varepsilon$,利用OLS估计的参数为 $\hat{\beta} = (X'X)^{-1}X'Y$, 其前提条件是 $X'X$ 是一个非退化矩阵, 即要求 $\text{rank}(X'X) = \text{rank}(X) = k < n$, 也就是说要求矩阵 X 列满秩, 否则无法求出参数的估计值 $\hat{\beta}$,这也是在多元线性回归模型的经典假设之一。

关于模型中解释变量之间的关系主要有三种:

(1) $r_{x_i x_j} = 0$ 解释变量间毫无线性关系, 变量间相互正交。这时多元回归的系数和每个参数 β_j 通过 Y 对 X_j 的一元回归估计结果一致。

(2) $|r_{x_i x_j}| = 1$ 解释变量间完全共线性, 即 $\text{rank}(X) < k$ 。此时模型参数将无法估计。

(3) $0 < r_{x_i x_j} < 1$ 解释变量间存在一定程度的线性关系。实际中碰到的主要是这种情形。当相关性较弱时, 可能影响不大, 但是随着解释变量间的共线性程度加强, 对参数估计值的准确性、稳定性带来影响。比如, 当 $|r_{x_i x_j}| \rightarrow 1$, $\hat{\beta}$ 仍具有无偏性, 即

$E(\hat{\beta}) = \beta$, 但是会丧失有效性, 因为此时 $|X'X| \rightarrow 0$, $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 会变得很大。以二元解释变量线性模型为例, 当 $r_{x_i x_j} = 0.8$ 时, $\text{Var}(\hat{\beta})$ 为 $r_{x_i x_j} = 0$ 时 $\hat{\beta}$ 方差的2.78倍。当 $r_{x_i x_j} = 0.95$ 时, $\text{Var}(\hat{\beta})$ 为 $r_{x_i x_j} = 0$ 时的10.26倍

多重共线性检验

多重共线性会造成一些严重的后果, 在建立线性回归模型的过程中, 有必要检验样本是否存在多重共线性。

检验多重共线性的常用方法主要有:

- 可决系数法

R^2 的值较大，且回归系数大多不显著。即： R^2 的值较大， F 值很高但每个回归参数估计值的方差都很大，即 t 很小，高度怀疑存在多重共线性。

- Klein判别法：

若有两个解释变量间的相关系数大于 R^2 ,高度怀疑存在多重共线性。

In [68]:

```
cor(dat[, -c(1, 8)]) #delete t and revenue
```

	industry	agriculture	construction	consumption	
industry	1.0000000	0.9899580	0.9888950	0.9948647	0
agriculture	0.9899580	1.0000000	0.9979530	0.9962729	0
construction	0.9888950	0.9979530	1.0000000	0.9980174	0
consumption	0.9948647	0.9962729	0.9980174	1.0000000	0
pop	0.9458216	0.9019155	0.9113149	0.9282159	1
disaster	-0.7907662	-0.8117106	-0.8081604	-0.8037320	-0

- 特征根法

根据矩阵行列式的性质，矩阵的行列式等于其特征根的连乘积。因而，当行列式 $|X'X| \approx 0$ 时，矩阵 $X'X$ 至少有一个特征根近似为零。反之可以证明，当矩阵 $X'X$ 至少有一个特征根近似为零时， X 的列向量间必存在复共线性，

In [69]:

```
x<- cbind(rep(1, length(dat[,1])), dat[, -c(1,8)])  
x<-as.matrix(x)  
eigen(t(x)%*%x)
```

eigen() decomposition

\$values

```
[1] 6.380013e+11 4.644425e+10 4.047931e+08 2.962096e+08 2.2  
90520e+07  
[6] 5.489892e+06 1.312736e-04
```

\$vectors

```
          [,1]          [,2]          [,3]          [,4]  
[1,] -4.612177e-06  5.536846e-06 -2.645779e-06  1.743727e-0  
6 -3.218785e-06  
[2,] -5.168233e-01 -4.110068e-01  6.750276e-01  2.609208e-0  
1 -1.856335e-01  
[3,] -1.249827e-01 -4.996705e-02 -1.270468e-01 -3.798663e-0  
2 -4.198784e-01  
[4,] -2.846459e-01 -3.988139e-01 -5.051496e-01 -3.449140e-0  
1 -4.901890e-01  
[5,] -4.813398e-01 -2.991077e-01 -2.666130e-01 -1.496394e-0  
1  7.334540e-01  
[6,] -6.063858e-01  6.713979e-01 -2.237648e-01  3.275633e-0  
1 -1.045087e-01  
[7,] -1.919843e-01  3.595603e-01  3.897019e-01 -8.257251e-0  
1 -8.961404e-03  
          [,6]          [,7]  
[1,] -8.472682e-06  1.000000e+00  
[2,]  7.593003e-02  1.268816e-06  
[3,] -8.876941e-01 -9.442331e-06  
[4,]  3.814409e-01  1.814270e-06  
[5,] -2.177554e-01 -4.925214e-07  
[6,]  1.150244e-01 -7.039225e-06  
[7,] -9.409348e-03 -5.139585e-07
```

- 条件指数法

特征根分析表明，当矩阵 $X'X$ 有一个特征根近似为零时，设计矩阵 X 的列向量间必存在复共线性。那么特征根近似为零的标准如何确定哪？这可以用条件数确定。记 $X'X$ 的最大特征根为 λ_m ，称

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, \quad i = 0, 1, 2, \dots, p$$

为特征根 λ_i 的条件数（Condition Index）。

一般认为：

- $0 < k < 10$ 时,设计矩阵 X 没有多重共线性;
- $10 \leq k < 100$ 时,认为 X 存在较强的多重共线性;
- 当 $k \geq 100$ 时,则认为 X 存在严重的多重共线性。

In [70]:

```
CI<-eigen(t(x)%%x)$values[1]/eigen(t(x)%%x)$values[7]
CI
```

4860087563798442

- 方差膨胀因子

对自变量做中心标准化，则 $X^*X^* = (r_{ij})$ 为自变量的相关阵。记

$C = (c_{jj}) = (X^*X^*)^{-1}$ 称其主对角线元素 $VIF_j = c_{jj}$ 为自变量 x_j 的方差扩大因子 (Variance Inflation Factor,简记为VIF)。

$$VIF_j = C_{jj} = \frac{1}{1-R_j^2}, j = 1, 2, \dots, p$$

经验表明,当 $VIF_j \geq 10$ 时,就说明自变量 x_j 与其余自变量之间有严重的多重共线性,且这种多重共线性可能会过度地影响最小二乘估计值。
还可用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性。当

$$VIF = \frac{1}{p} \sum_{j=1}^p VIF_j$$

远远大于1时就表示存在严重的多重共线性问题。

In [71]:

```
#install.packages("bstats") #R3.5不提供
#library(bstats) #VIF test
library(car)
vif(lm3)
```

Loading required package: carData

industry

729.200050293165

agriculture

1403.79631044332

construction

1462.0348013633

consumption

1108.21598023737

pop

59.4492423386617

disaster

4.93000349737597

多重共线性克服

(一) 逐步回归

逐步回归主要分为向前逐步回归（forward）、向后逐步回归(backward)和向后向前逐步回归（both）。逐步回归本身并不是一种新的回归或者参数的估计方法，所用到的参数估计方法都是原来的，是从众多的变量中选出最优模型的变量的一套方法。

向前逐步回归的思路是逐个引入变量。

后向逐步回归的思路是先引入全部自变量，然后逐个剔除不重要的变量，其剔除变量的思路和前向逐步回归的思路类似。

后向前向逐步回归先逐步剔除变量，但可以后面的步骤中重新引入原先被剔除的变量，其方向是双向的。而后向逐步回归的自变量一旦被提出后，在后面的步骤中就不会被重新引入，是单向的。

最优模型一般通过一些准则来确定，比如 F 值，可决系数 R^2 ， AIC 等。

In [72]:

```
step(lm3, direction="forward")
```

```
Start:  AIC=205.53
revenue ~ industry + agriculture + construction + consumption +
pop + disaster
```

```
Call:
lm(formula = revenue ~ industry + agriculture + construction +
consumption + pop + disaster, data = dat)
```

```
Coefficients:
(Intercept)      industry  agriculture  construction  consumption
6.821e+04      1.297e-01     -7.065e-02      4.465e-02
6.011e-01
pop      disaster
-7.020e-01  4.324e-02
```

In [73]:

```
step(lm3,direction="backward")
```

Start: AIC=205.53
revenue ~ industry + agriculture + construction + consumption +
pop + disaster

	Df	Sum of Sq	RSS	AIC
- agriculture	1	6011	5271671	203.55
- construction	1	31598	5297257	203.62
- disaster	1	396741	5662400	204.62
<none>			5265660	205.53
- industry	1	1039770	6305430	206.23
- pop	1	1294061	6559721	206.83
- consumption	1	10557615	15823275	220.03

Step: AIC=203.55
revenue ~ industry + construction + consumption + pop + disaster

	Df	Sum of Sq	RSS	AIC
- construction	1	37130	5308801	201.65
- disaster	1	742733	6014404	203.52
<none>			5271671	203.55
- industry	1	3630106	8901777	209.41
- pop	1	4733883	10005554	211.16
- consumption	1	12130608	17402279	219.46

Step: AIC=201.65
revenue ~ industry + consumption + pop + disaster

	Df	Sum of Sq	RSS	AIC
- disaster	1	707460	6016261	201.53
<none>			5308801	201.65
- industry	1	3889504	9198305	207.90
- pop	1	6439624	11748425	211.57
- consumption	1	103394549	108703350	244.94

Step: AIC=201.53
revenue ~ industry + consumption + pop

	Df	Sum of Sq	RSS	AIC
<none>			6016261	201.53
- industry	1	4820298	10836559	208.36
- pop	1	7281880	13298142	211.43
- consumption	1	108630708	114646969	243.74

Call:

```
lm(formula = revenue ~ industry + consumption + pop, data =  
dat)
```

Coefficients:

(Intercept)	industry	consumption	pop
71680.6334	0.1270	0.6209	-0.7217

In [74]:

```
step(lm3,direction="both")
```


Start: AIC=205.53
revenue ~ industry + agriculture + construction + consumption +
pop + disaster

	Df	Sum of Sq	RSS	AIC
- agriculture	1	6011	5271671	203.55
- construction	1	31598	5297257	203.62
- disaster	1	396741	5662400	204.62
<none>			5265660	205.53
- industry	1	1039770	6305430	206.23
- pop	1	1294061	6559721	206.83
- consumption	1	10557615	15823275	220.03

Step: AIC=203.55
revenue ~ industry + construction + consumption + pop + disaster

	Df	Sum of Sq	RSS	AIC
- construction	1	37130	5308801	201.65
- disaster	1	742733	6014404	203.52
<none>			5271671	203.55
+ agriculture	1	6011	5265660	205.53
- industry	1	3630106	8901777	209.41
- pop	1	4733883	10005554	211.16
- consumption	1	12130608	17402279	219.46

Step: AIC=201.65
revenue ~ industry + consumption + pop + disaster

	Df	Sum of Sq	RSS	AIC
- disaster	1	707460	6016261	201.53
<none>			5308801	201.65
+ construction	1	37130	5271671	203.55
+ agriculture	1	11544	5297257	203.62
- industry	1	3889504	9198305	207.90
- pop	1	6439624	11748425	211.57
- consumption	1	103394549	108703350	244.94

Step: AIC=201.53
revenue ~ industry + consumption + pop

	Df	Sum of Sq	RSS	AIC
<none>			6016261	201.53
+ disaster	1	707460	5308801	201.65
+ agriculture	1	124145	5892116	203.22

+ construction	1	1857	6014404	203.52
- industry	1	4820298	10836559	208.36
- pop	1	7281880	13298142	211.43
- consumption	1	108630708	114646969	243.74

```
Call:
lm(formula = revenue ~ industry + consumption + pop, data =
dat)
```

Coefficients:

(Intercept)	industry	consumption	pop
71680.6334	0.1270	0.6209	-0.7217

(二) 岭回归

当解释变量之间存在多重共线性时，即 $\left|\mathbf{X}'\mathbf{X}\right|\approx 0$ ，则 $\text{Var}(\hat{\boldsymbol{\beta}})=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ 就会增大，原因是 $\mathbf{X}'\mathbf{X}$ 接近奇异。如果将 $\mathbf{X}'\mathbf{X}$ 加上一个正常数对角阵 $\lambda\mathbf{I}$ ($\lambda>0$, \mathbf{I} 为单位矩阵) 即 $\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}$ ，使得 $\left|\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right|\approx 0$ 的可能性比 $\left|\mathbf{X}'\mathbf{X}\right|\approx 0$ 的可能性更小，那么 $\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}$ 接近奇异的程度就会比 $\mathbf{X}'\mathbf{X}$ 小的多，这就是岭回归的最初想法。可以证明 β 的岭回归估计为：

$$\tilde{\boldsymbol{\beta}}(\lambda)=\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

其中， $\tilde{\boldsymbol{\beta}}(\lambda)$ 称为 β 的岭回归估计量， λ 称为岭参数。我们知道此时 $\tilde{\boldsymbol{\beta}}(\lambda)$ 不是确定的向量，而随着 λ 取不同的值而变化。当 $\lambda=0$ 时， $\tilde{\boldsymbol{\beta}}(\lambda)=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ 就是普通最小二乘估计量。所以，从某种意义上讲，岭估计是最小二乘估计的推广，最小二乘估计是岭估计的特例。可以证明岭估计是一个有偏估计即：

$$E[\tilde{\boldsymbol{\beta}}(\lambda)]=E\left[\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}\right]=\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\neq\boldsymbol{\beta}$$

岭估计的关键是选取岭参数 λ ,岭迹法是选取岭参数 λ 的常用方法之一。若记 $\tilde{\beta}_i(\lambda)$ 为 $\tilde{\beta}(\lambda)$ 的第 i 个分量,它是 λ 的一元函数。当 λ 在 $[0, \infty)$ 上变化是, $\tilde{\beta}(\lambda)$ 的图形称为岭迹 (ridge trace) 。 $\tilde{\beta}(\lambda)$ 的每个分量 $\tilde{\beta}_i(\lambda)$ 的岭迹华仔同一个图上, 根据岭迹的变化趋势选择 λ 值, 使得各个回归系数的岭估计大体上稳定, 并且各个回归系数岭估计值的符号比较合理并符合实际。

R里MASS包的lm.ridge()函数可以用来做岭估计, 其用法与lm()用法类似。

当不指定 λ 值时候, λ 默认为0。

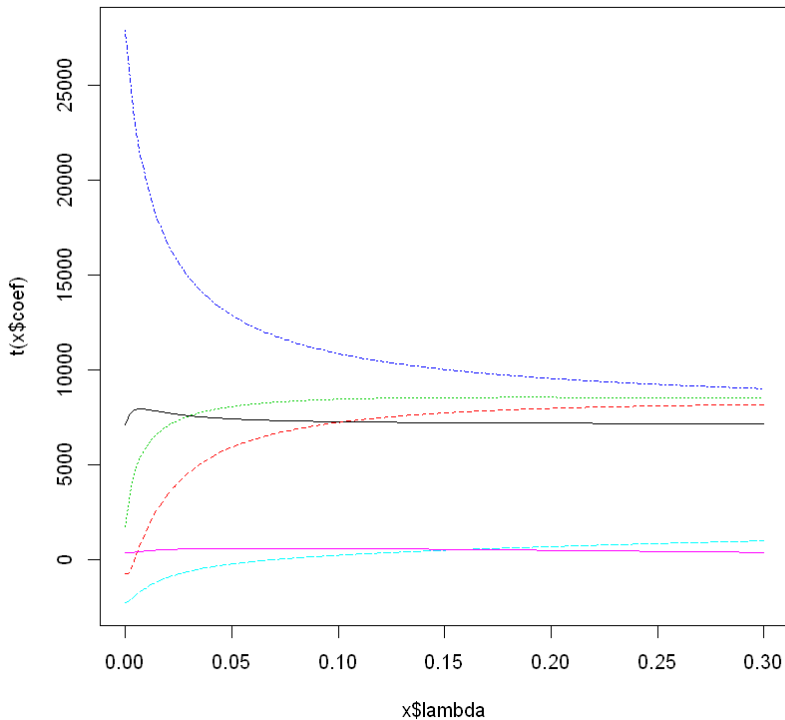
In [75]:

```
library(MASS)
lm.r<-lm.ridge(revenue~industry+agriculture+construction+consumption+pop+disaster, data=dat)
lm.r
```

	industry	agriculture	construction	c
onsumption	6.821406e+04	1.296967e-01	-7.065467e-02	4.465321e-02
	6.			
011086e-01				
pop		disaster		
-7.020226e-01		4.323570e-02		

In [76]:

```
plot(lm.ridge(revenue~industry+agriculture+construction
              +consumption+pop+disaster, data=dat, lambda=seq(0, 0.3, 0.001)))
```



In [77]:

```
select(lm.ridge(revenue~industry+agriculture+construction
                 +consumption+pop+disaster, data=dat, lambda=seq(0, 0.3, 0.001
)))
```

modified HKB estimator is 0.003136352
modified L-W estimator is 0.002329019
smallest value of GCV at 0.004

In [78]:

```
lm.ridge(revenue~industry+agriculture+construction+consumption
+pop+disaster,data=dat,lambda=0.004)
```

	industry	agriculture	construction	c
consumption				
	5.662106e+04	1.439691e-01	-3.539058e-03	1.146925e-01
	5.037839e-01			
	pop	disaster		
	-5.970767e-01	4.829628e-02		

异方差性

介绍什么是异方差？为何会出现异方差？如果出现异方差会有什么样的后果？如何检验异方差？如果存在异方差，该如何克服？

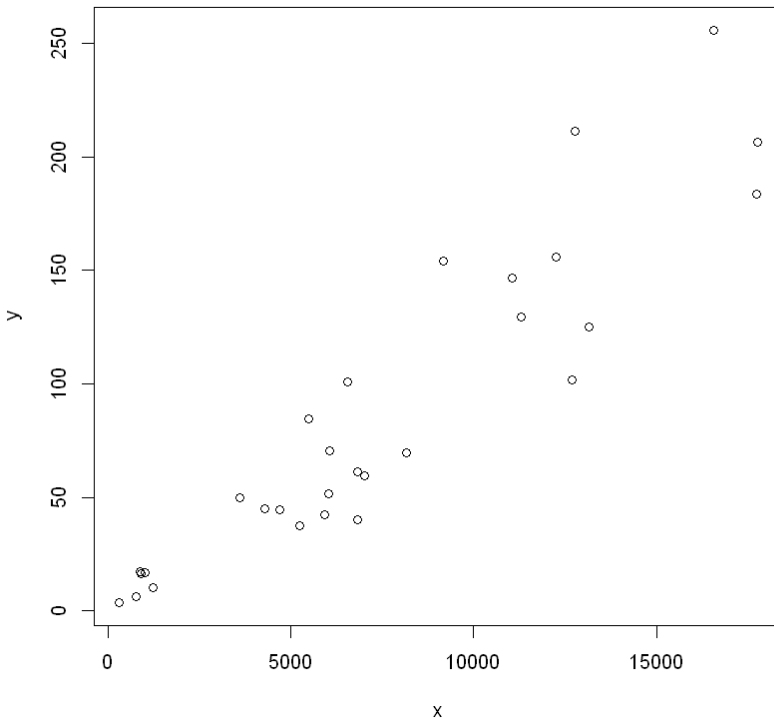
取某年中国29个省市自治区农作物种植业产值 Y （亿元）和农作物播种面积 X （万亩）数据，研究中国29个省市自治区农作物种植业产值和农作物播种面积的关系。

x	y
907.5	16.31
873.2	17.14
13159.2	125.24
5928.1	42.24
6834.4	40.28
5495.5	84.47
6055.2	70.7
12694.6	101.67
1018.5	16.83
12770.9	211.51
6542.7	101
12244.3	155.87
3601.5	49.72
8158.1	69.7
16564.5	255.92
17729.2	183.65
11061.5	146.79
11304.7	129.63
9166.2	154.28
6821.7	61.24
17779.6	206.5
4701.3	44.37
6036.1	51.79
316.5	3.53
7016.5	59.45
5252.5	37.29
761.7	6.33
1235.2	10.07

x	y
4275.1	44.78

In [79]:

```
agticul<-read.csv(file="./data/11-2.csv")
y=agticul[,2]
x=agticul[,1]
plot(x,y)
```



从散点图可以看出，农作物种植业产值与播种面积存在某种线性关系，说明可以用线性回归进行分析，但是我们发现一个问题，即农作物种植业产值的离散程度随着播种面积的增加而增大，在散点图上表现为“喇叭”型分布，这实际上是说明数据存在异方差，违反经典假设。

In [80]:

```
lm.a=lm(y~x)
summary(lm.a)
plot(resid(lm.a)~x)
```



```
Call:
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-48.924	-21.254	0.527	11.051	59.976

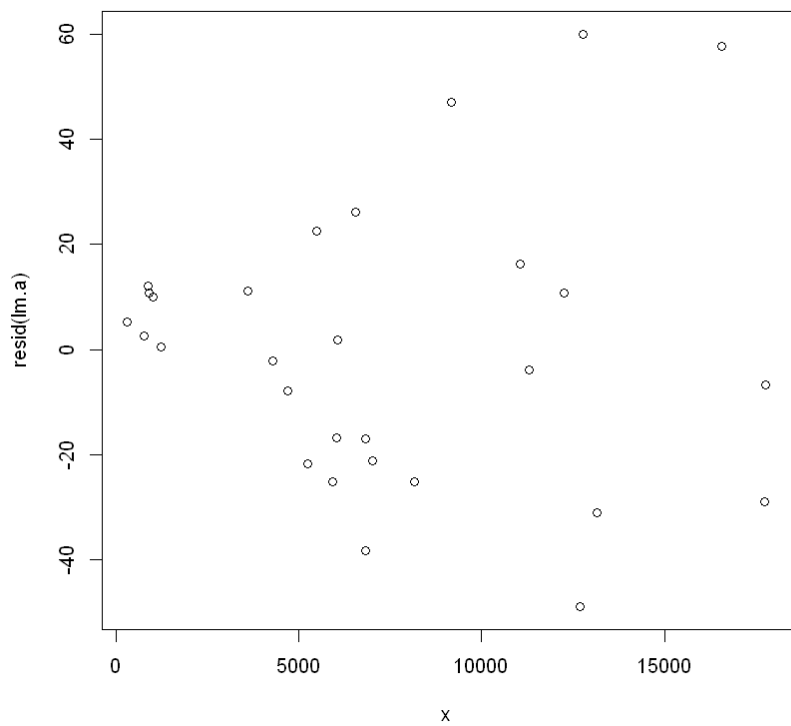
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.6610192	8.9241561	-0.634	0.531
x	0.0123088	0.0009888	12.449	1.07e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.06 on 27 degrees of freedom
Multiple R-squared: 0.8516, Adjusted R-squared: 0.8461
F-statistic: 155 on 1 and 27 DF, p-value: 1.066e-12





残差也是呈喇叭型分布，随着x的增加而增加。不符合经典假设。

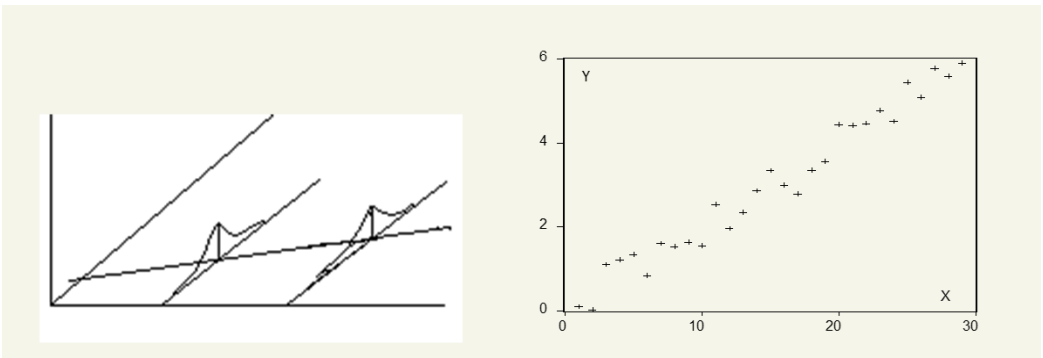
异方差性定义及后果

经典假设条件里， $\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$ ，即随机扰动项的协方差矩阵主对角线上的元素都是常数且相等，即每一随机扰动项的方差都是有限的相同值（同方差假定）；且非对角线上的元素为零（非自相关假定），但是如果当这个假定不成立时候，比如：

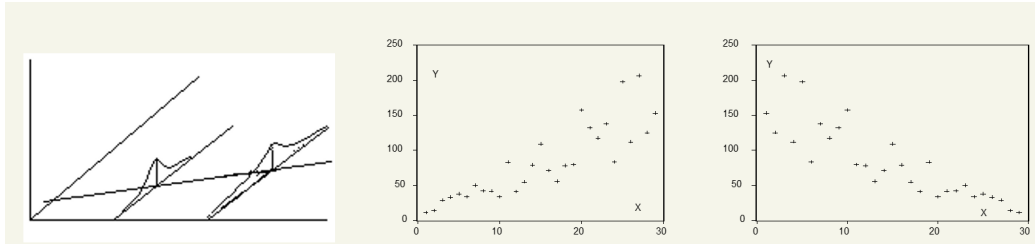
$$\text{Var}(\varepsilon) = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \sigma_{nn} \end{pmatrix} \neq \sigma^2 \mathbf{I}$$

ε 的协方差矩阵主对角线上的元素不相等时，称该随机扰动项存在异方差。

同方差的总体分布和散点图，随着解释变量的变化，相应的 ε_i 分布方差都是相同的。



递增型异方差的总体分布和散点图，随着解释变量的增加，相应 ε_i 的分布方差也是增加的，散点图呈喇叭状向外扩散



异方差的主要后果是回归参数估计量不再具有有效，因此回对模型的 F 检验和 t 检验带来问题。因此在计量经济分析中，有必要检验模型是否存在异方差。

异方差性检验

异方差的检验方法主要有散点图、残差图、Goldfeld-Quandt检验、Glejser检验和White检验。

(一) 散点图与残差图

定性分析主要利用散点图和残差图的形状来初步判断异方差的存在性。上例的散点图和残差图呈“喇叭”型分布，说明数据可能存在递增型异方差。但定性分析只能提供一个主观、初略的判断，还需进一步借助更加精确的检验方法。

(二) Goldfeld-Quandt检验

Goldfeld-Quandt检验是Goldfeld-Quandt于1965年提出的，所要检验的问题为

$H_0: \varepsilon_i$ 具有同方差, $H_1: \varepsilon_i$ 具有递增型异方差。

检验的具体步骤是：

第一，把原样本分成两个子样本。具体方法是把成对（组）的观测值按解释变量的大小顺序排列，略去 m 个处于中心位置的观测值。（通常 $n > 30$ 时，取 $m \approx n/4$ ，余下的 $n - m$ 个观测值分成容量相等 $(n - m)/2$ 的两个子样本。）

$$\{X_1, X_2, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_{n-1}, X_n\}$$

$$\tilde{n}_1 = (n-m)/2 \quad \tilde{n} = n/4 \quad \tilde{n} = (n-m)/2$$

第二，用两个子样本分别估计回归曲线，并计算残差平方和相对于 n_2 和 n_1 分别用 SSE_2 和 SSE_1 表示。

第三，构建 F 统计量

$$F = \frac{SSE_2 / (n_2 - k)}{SSE_1 / (n_1 - k)}, \quad (k \text{ 为模型中被估参数个数})$$

在 H_0 成立条件下， $F \sim F_{(n_2 - k, n_1 - k)}$

第四，判别规则如下：

若 $F \leq F_{\alpha}(n_2 - k, n_1 - k)$ ，接受 H_0 。（ u_t 具有同方差）

若 $F > F_{\alpha}(n_2 - k, n_1 - k)$ ，拒绝 H_0 （递增型异方差）

这里我们应该注意到，当模型含有多个解释变量时，应以每一个解释变量为基准检验异方差。此法的基本思路也适用于递减型异方差。另外，对于截面样本，计算 F 统计量之前，必须先把数据按解释变量的值从小到大排序。

对于示例数据，首先我们根据上面的检验步骤，自己编写一个`gq_test()`函数。检验的 F 统计量为11.197，对应 P values是0.00018，说明是显著的。即示例数据存在递增型异方差。

In [81]:

```
g_qtest=function(x,y){  #G-Q检验
  n=length(x)
  m=round(n/4)
  xs=sort(x)
  xs1=xs[1:ceiling((n-m)/2)]
  xs2=xs[floor(n-(n-m)/2+1):n]
  n1=length(xs1)
  n2=length(xs2)
  ii=1:length(xs1)
    for(i in 1:length(xs1)){
      ii[i]=which(x==xs1[i])
    }
  y1=y[ii]
  ii2=1:length(xs2)
  for(i in 1:length(xs2)){
    ii2[i]=which(x==xs2[i])
  }
  y2=y[ii2]
  lm1=lm(y1~xs1)
  lm2=lm(y2~xs2)
  res1=lm1$resid
  res2=lm2$resid
  FF=(sum(res2^2)/length(xs2))/(sum(res1^2)/length(xs1))
  p=1-pf(FF,n1,n2)
  v=c("Fvalues:",FF,"pvalues:",p)
  return(v)
}

g_qtest(x,y)
```

```
'Fvalues:' '11.1975072063171' 'pvalues:'
'0.000184993475651374'
```

另外R的lmtest包也提供G-Q方法的函数gqtest(), 其用法为:

```
gqtest(formula,point=0.5,fraction=0,alternative=c("greater","two.sided","less"),order.by
```

In [82]:

```
#install.packages("lmtest")
library(lmtest)
gqtest(lm.a, order.by=~x)
```

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

as.Date, as.Date.numeric

Goldfeld-Quandt test

data: lm.a

GQ = 8.5478, df1 = 13, df2 = 12, p-value = 0.000351

alternative hypothesis: variance increases from segment 1 to

2

(三) Glejser 检验

Glejser 检验基本思想是：检验 $\hat{\varepsilon}_i$ 是否与解释变量 X_i 存在函数关系。若存在函数关系，则说明存在异方差；若无函数关系，则说明不存在异方差。比如检验形式：

$$\left| \hat{\varepsilon}_i \right| = \alpha_0 + \alpha_1 X_i$$

$$\left| \hat{\varepsilon}_i \right| = \alpha_0 + \alpha_1 \frac{1}{X_i}$$

$$\left| \hat{\varepsilon}_i \right| = \alpha_0 + \alpha_1 X_i^2$$

Glejser 检验的特点是不仅能对异方差的存在进行判断，而且还能对异方差随某个解释变量变化的函数形式进行诊断。该方法即可检验递增型异方差，也可以检验递减型异方差。但该方法是属于穷举法，也存在很多缺陷。

In [83]:

```
library(lmtest)
re=resid(lm.a)
abre=abs(re)
summary(lm(abre~I(sqrt(x))))
```

Call:

```
lm(formula = abre ~ I(sqrt(x)))
```

Residuals:

Min	1Q	Median	3Q	Max
-28.507	-7.306	1.071	5.551	30.449

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.95857	6.94993	-0.282	0.7802
I(sqrt(x))	0.27861	0.08047	3.462	0.0018 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.14 on 27 degrees of freedom

Multiple R-squared: 0.3075, Adjusted R-squared: 0.2818

F-statistic: 11.99 on 1 and 27 DF, p-value: 0.0018

In [84]:

```
summary(lm(abre ~ -1 + I(sqrt(x)))) ##去掉不显著的截距项
```

Call:

```
lm(formula = abre ~ -1 + I(sqrt(x)))
```

Residuals:

Min	1Q	Median	3Q	Max
-27.6660	-8.5269	0.7122	4.4404	30.8635

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
I(sqrt(x))	0.2576	0.0299	8.616	2.32e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.91 on 28 degrees of freedom
Multiple R-squared: 0.7261, Adjusted R-squared: 0.7163
F-statistic: 74.23 on 1 and 28 DF, p-value: 2.317e-09



(四) White检验

White检验由H.White 1980年提出。White检验不需要对观测值排序，也不依赖于随机误差项服从正态分布，它是通过一个辅助回归式构造 χ^2 统计量进行异方差检验。

以二元回归模型 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$ 为例：检验的具体步骤如下：

第一，首先对上式进行OLS回归，求残差 $\hat{\varepsilon}_i$ 。并作如下辅助回归式：

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i1}^2 + \alpha_4 X_{i2}^2 + \alpha_5 X_{i1} X_{i2} + V_i$$

即用 $\hat{\varepsilon}_i^2$ 对原回归式中的各解释变量、解释变量的平方项，交叉项进行OLS回归。注意，上式中要保留常数项。求辅助回归式的可决系数 R^2

第二，怀特检验的零假设和备择假设是：

$H_0: \varepsilon_i$ 不存在异方差, $H_1: \varepsilon_i$ 存在异方差

第三，在不存在异方差假设条件下统计量 $nR^2 \sim \chi^2(5)$

其中 n 表示样本容量, R^2 是辅助回归式的OLS估计式的可决系数。自由度5表示辅助回归式中解释变量项数（不含常数项）。

第四，判别规则是：

若 $nR^2 \leq x^2\alpha(5)$ ，接受 $H_0(\varepsilon_i$ 具有同方差)

若 $nR^2 > x^2\alpha(5)$ ，拒绝 $H_0(\varepsilon_i$ 具有异方差)

怀特检验的特点是，不仅能够检验异方差的存在，同时，在多变量的情况下，还能够判断出是哪一个变量引起的异方差，通常适用于界面数据的情形该方法不需要异方差的先验信息，但要求观测值为大样本。

In [85]:

```
#install.packages("bstats")
#whilt.test(lm.a)
bptest(lm.a, ~x+I(x^2))
```

studentized Breusch-Pagan test

data: lm.a
BP = 8.0196, df = 2, p-value = 0.01814

In [86]:

```
## regression
lm2 <- lm(y ~ x)
## auxiliary regression
aux <- residuals(lm.a)^2
aux_lm <- lm(aux ~ x + I(x^2))
## test statistic
nrow(agticul) * summary(aux_lm)$r.squared
pchisq(nrow(agticul) * summary(aux_lm)$r.squared, df=2, lower.tail=F)
```

8.0195993407701

0.0181370282567751

异方差性克服

(一) 广义最小二乘法

设模型为 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

示例我们可以利用Glejser方法进行检验，发现残差绝对值与 X_i 存在 $|\hat{\varepsilon}_i| = 0.2576\sqrt{X_i}$ 。把每个变量都除以 $0.2576\sqrt{X_i}$ ，对变换后的数据做散点图，如图所示，已经不像变换前的散点图那样呈喇叭状。然后，我们对变换后的数据回归，做残差图，发现残差图也不呈喇叭型分布，说明基本消除了异方差。进一步，我们利用white 检验发现，其P-value为0.1354，进一步验证了，基本消除了示例的异方差。

In [87]:

```
ys<-y/(0.2576*sqrt(x))
xs<-x/(0.2576*sqrt(x))
plot(xs,ys)
lm.sa<-lm(ys~xs)
summary(lm.sa)
plot(xs,resid(lm.sa))
#white.test(lm.sa)
bptest(lm.sa,~xs+I(xs^2))
```

```
Call:
lm(formula = ys ~ xs)
```

Residuals:

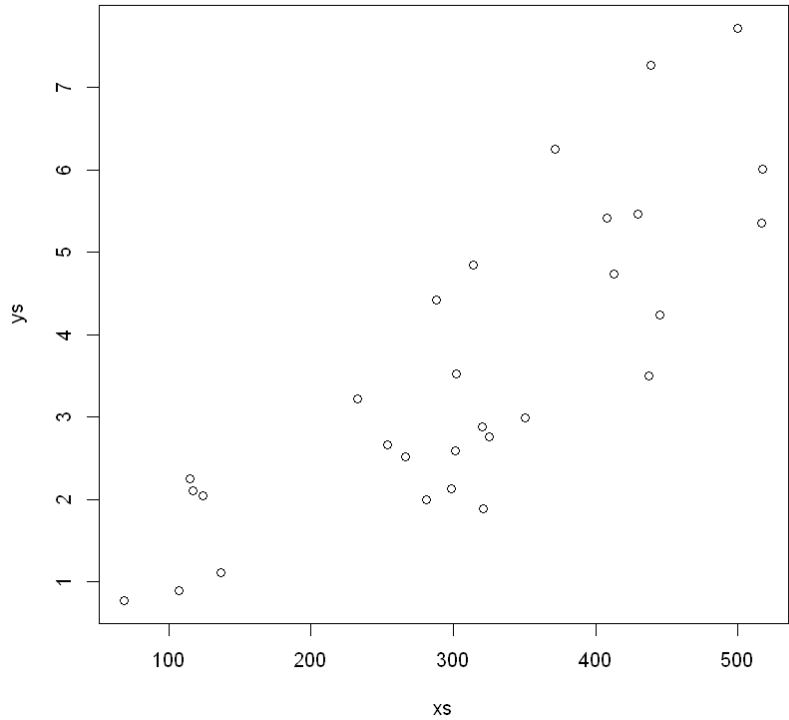
	Min	1Q	Median	3Q	Max
	-1.8039	-0.8791	-0.0481	0.6898	2.1778

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.099212	0.539252	-0.184	0.855
xs	0.011824	0.001608	7.351	6.59e-08 ***

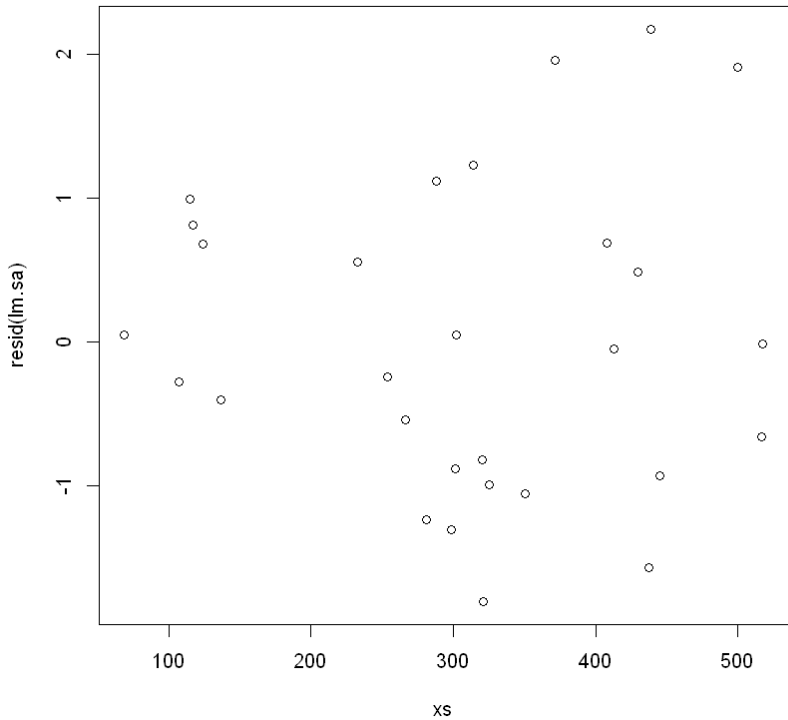
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.097 on 27 degrees of freedom
Multiple R-squared: 0.6668, Adjusted R-squared: 0.6545
F-statistic: 54.04 on 1 and 27 DF, p-value: 6.592e-08



studentized Breusch-Pagan test

data: lm.sa
BP = 3.9989, df = 2, p-value = 0.1354



(二) 取对数

在实际中，很多情况，通过对模型的变量取对数降低异方差性的影响。比如

$$\ln Y = \beta_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \cdots + \beta_k \ln X_k + \varepsilon^*$$

这是因为经过对数变换后的线性模型，其残差 ε^* 表示相对误差，而相对误差往往比绝对误差有较小的差异。

对示例的变量都取对数，然后画散点图，如图所示，其散点图就不像取对数前呈喇叭状。然后，我们对取对数后的数据回归，做残差图，发现残差图也不呈喇叭型分布，说明基本消除了异方差。进一步，我们利用white 检验发现，其P-value为0.275，进一步验证了，取对数后消除了示例的异方差。

In [88]:

```
lny<-log(y)
lnx<-log(x)
plot(lnx, lny)
lm.lna<-lm(lny~lnx)
summary(lm.lna)
plot(lnx, resid(lm.lna))

#white.test(lm.lna)
bptest(lm.lna, ~lnx+I(lnx^2))
```



```
Call:
lm(formula = lny ~ lnx)
```

Residuals:

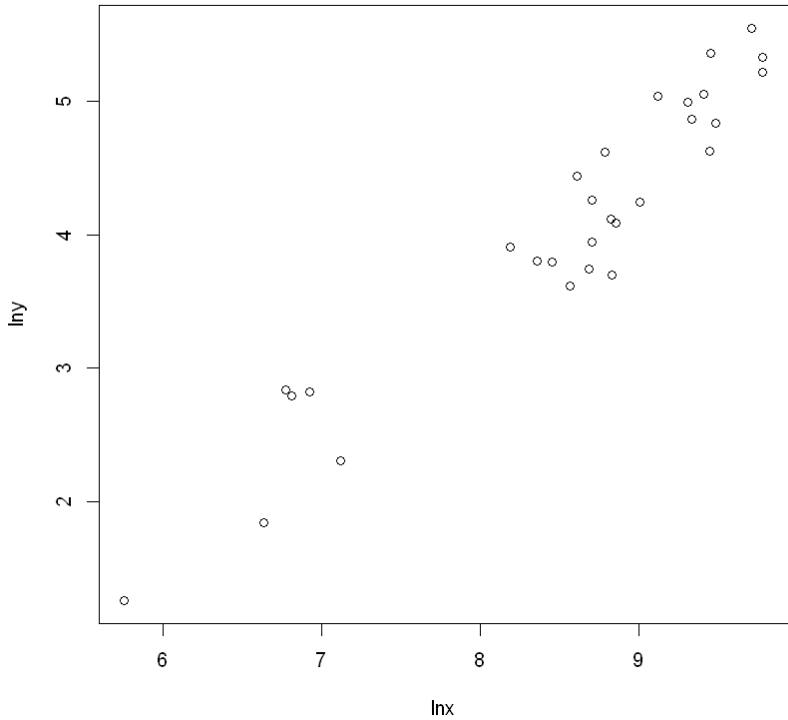
	Min	1Q	Median	3Q	Max
	-0.62287	-0.25197	-0.02322	0.32754	0.50315

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.18013	0.48969	-8.536	3.77e-09 ***
lnx	0.96253	0.05695	16.901	6.93e-16 ***

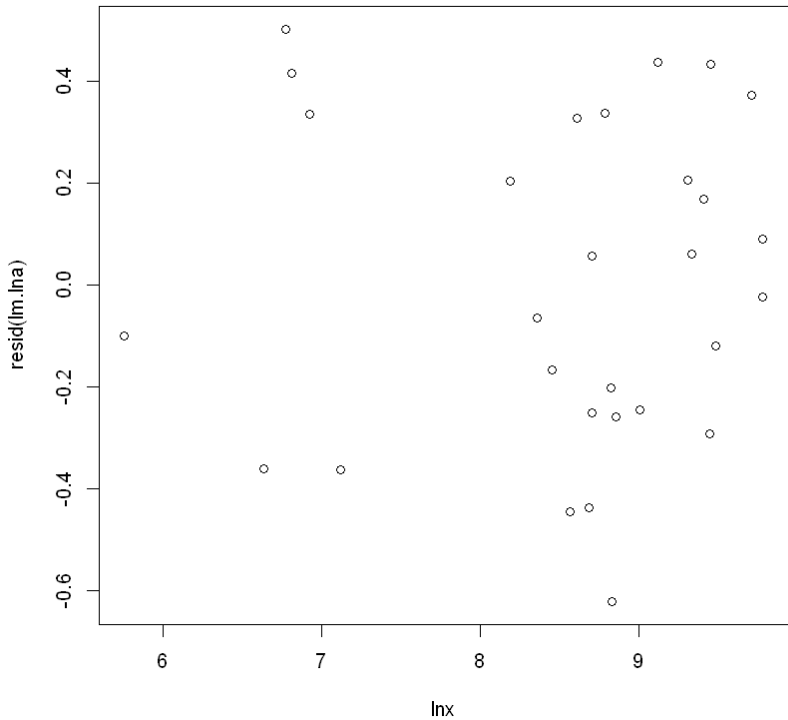
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3235 on 27 degrees of freedom
Multiple R-squared: 0.9136, Adjusted R-squared: 0.9104
F-statistic: 285.6 on 1 and 27 DF, p-value: 6.934e-16



studentized Breusch-Pagan test

data: lm.lna
BP = 2.5821, df = 2, p-value = 0.275



序列相关性

主要介绍什么是序列相关？为何会出现序列相关？如果出现序列相关会有什么样的后果？如何检验序列相关？如果存在序列相关，该如何克服？

2003年中国农村人口占59.47%，而消费总量却只占41.4%，农村居民的收入和消费是一个值得研究的问题。消费模型是研究居民消费行为的常用工具。通过中国农村居民消费模型的分析可判断**农村居民的边际消费倾向，这是宏观经济分析的重要参数。**同时，农村居民消费模型也能用于农村居民消费水平的预测。影响居民消费的因素很多，但由于受各种条件的限制，通常只引入居民收入一个变量做解释变量，即消费模型设定为

$$Y_t = \beta_1 + \beta_2 X_t + \mu_t$$

式中， Y_t 为农村居民人均消费支出， X_t 为农村人均居民纯收入， μ_t 为随机误差项。数据是从《中国统计年鉴》收集的中国农村居民1985-2003年的收入与消费数据。

t	income	expend	cpi
1985	397.6	317.42	100
1986	423.8	357	106.1
1987	462.6	398.3	112.7
1988	544.9	476.7	132.4
1989	601.5	535.4	157.9
1990	686.3	584.63	165.1
1991	708.6	619.8	168.9
1992	784	659.8	176.8
1993	921.6	769.7	201
1994	1221	1016.81	248
1995	1577.7	1310.36	291.4
1996	1923.1	1572.1	314.4
1997	2090.1	1617.15	322.3
1998	2162	1590.33	319.1
1999	2214.3	1577.42	314.3
2000	2253.4	1670	314
2001	2366.4	1741	316.5
2002	2475.6	1834	315.2
2003	2622.24	1943.3	320.2

为了消除价格变动因素对农村居民收入和消费支出的影响，不宜直接采用现价人均纯收入和现价人均消费支出的数据，而需要用经消费价格指数进行调整后的1985年可比价格计的人均纯收入和人均消费支出的数据作回归分析。使用普通最小二乘法估计消费模型得

In [1]:

```
dat=read.csv("./data/11-3.csv")
y=dat$expend
x=dat$income
p=dat$cpi
yp=y/p*100
xp=x/p*100
lm.in=lm(yp~xp)
summary(lm.in)
```

Call:

```
lm(formula = yp ~ xp)
```

Residuals:

Min	1Q	Median	3Q	Max
-27.808	-5.263	1.170	7.195	26.385

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	106.71298	12.20259	8.745	1.06e-07 ***
xp	0.59989	0.02136	28.090	1.09e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

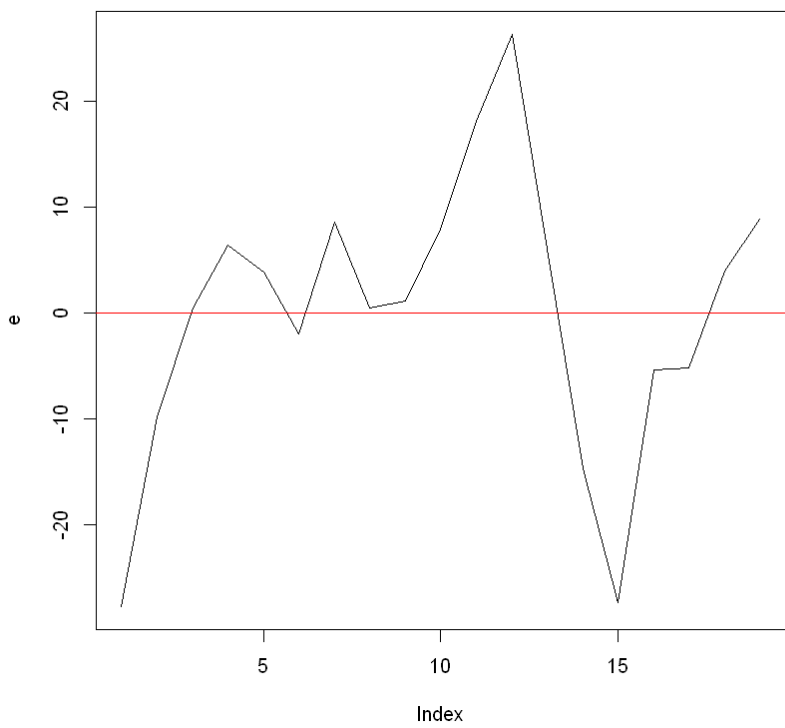
Residual standard error: 13.84 on 17 degrees of freedom

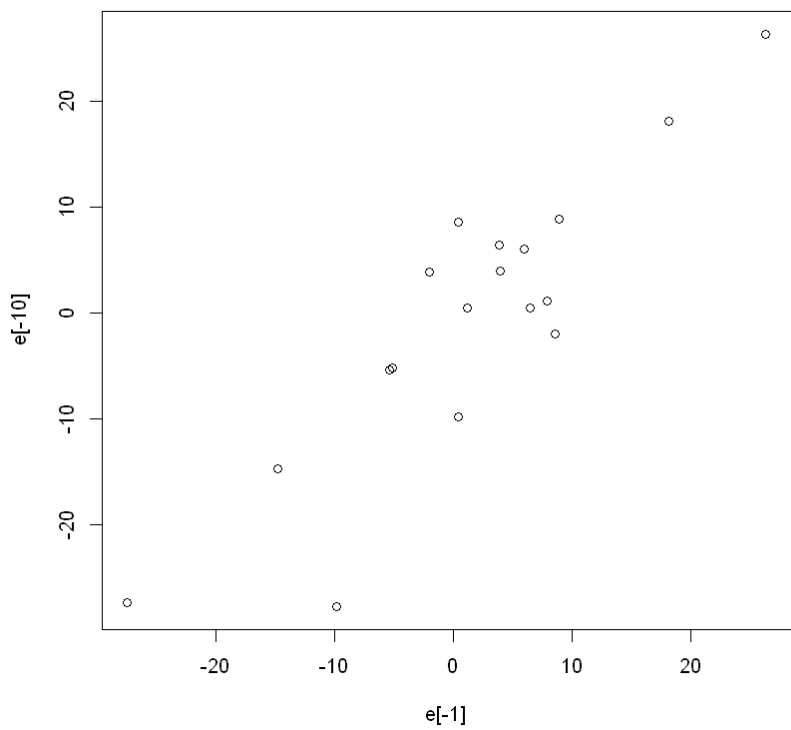
Multiple R-squared: 0.9789, Adjusted R-squared: 0.9777

F-statistic: 789.1 on 1 and 17 DF, p-value: 1.094e-15

In [90]:

```
e<-resid(lm.in)
plot(e, type="l")
abline(h=0, col="red")
plot(e[-1], e[-10])
```





该回归方程可决系数较高，回归系数均显著。从残差图可以看出残差的变动有系统模式，连续为正和连续为负。另外，从残差前后项之间的散点图可以看出，它们之间存在高度的正的线性关系。这些表明残差项存在一阶正自相关。如果数据间存在自相关，回归的结果是有问题的，需要我们做进一步的分析。

序列相关性定义及后果

针对线性模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad i = 1, 2, \dots, n$$

当 $\text{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0, \quad (i, j \in n, i \neq j)$ ，即误差项 ε_i 的取值在时间上是相互无关的称误差项 ε_i 非序列相关。如果

$$\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0, \quad (i \neq j)$$

则称误差项 ε_i 存在序列相关（自相关）。原指一随机变量在时间上与其滞后项之间的相关。这里主要是指回归模型中随机误差项 ε_i 与其滞后项的相关关系。

序列相关按形式可分为两类。

(1) 一阶自回归形式

当误差项 ε_i 只与其滞后一期值有关时，即

$$\varepsilon_i = f(\varepsilon_{i-1})$$

称 ε_i 具有一阶自回归形式。

(2) 高阶自回归形式

当误差项 ε_i 的本期值不仅与其前一期值有关，而且与其前若干期的值都有关系时，即

$$\varepsilon_i = f(\varepsilon_{i-1}, \varepsilon_{i-2}, \dots)$$

则称 ε_i 具有高阶自回归式。

序列相关的来源主要有：

- (1) 模型的数学形式不妥。若所用的数学模型与变量间的真实关系不一致，误差项常表现出自相关。
- (2) 变量的惯性。大多数时间序列都存在自相关。当期值往往受滞后值影响。突出特征就是惯性与低灵敏度。如国民生产总值，固定资产投资，国民消费等。
- (3) 回归模型中略去了带有自相关的重要解释变量。若丢掉了应该列入模型的带有自相关的重要解释变量，那么它的影响必然归并到误差项 ε_i 中，从而使误差项呈现自相关。

当误差项 ε_i 存在序列相关时，模型参数的最小二乘估计量仍具有无偏性，即 $E(\hat{\beta}) = \beta$ ，但丧失有效性，即 $\text{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 。

序列相关性检验

定性分析法

定性分析法就是依据残差序列图或残差散点图作出判断。由于残差 e_i 是对误差项 ε_i 的估计，所以尽管误差项 ε_i 观测不到，但可以通过 e_i 的变化判断 ε_i 是否存在序列相关。

若残差序列图或者残差散点图和下图类似，则说明存在正自相关；

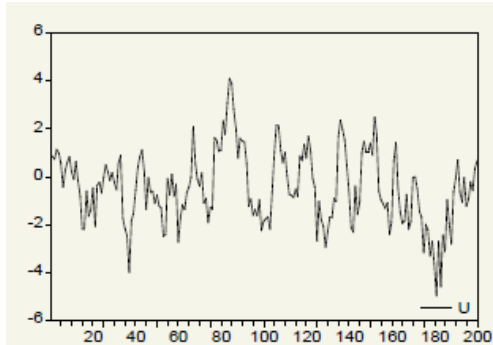


图 11-15. 正序列相关的序列图

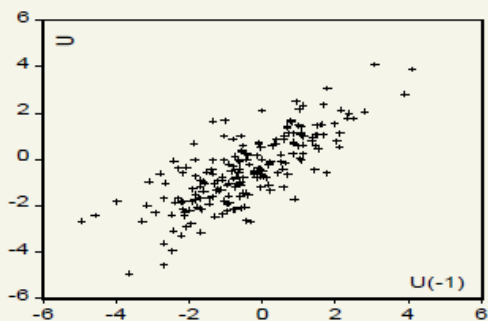


图 11-16. 正序列相关的散点图

若残差序列图或者残差散点图和下图类似，则说明存在负自相关。

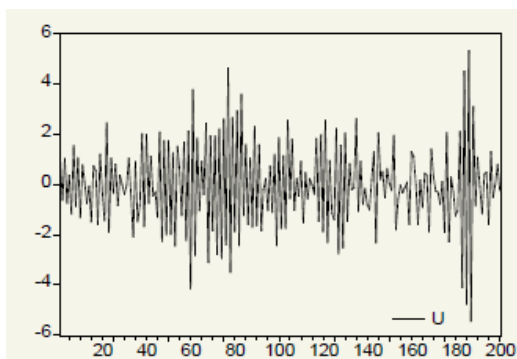


图 11-17. 负序列相关的序列图

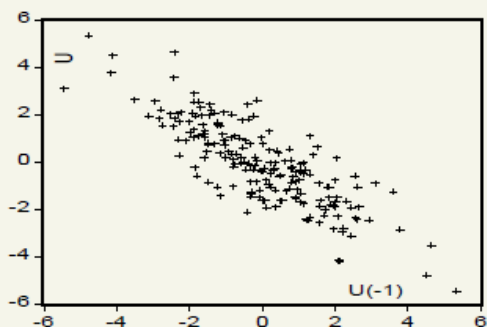


图 11-18. 负序列相关的散点图

示例的残差序列和散点图，说明存在序列自相关

DW (Durbin-Watson) 检验法

DW检验是J. Durbin, G. S. Watson于1950年提出的。它是利用残差 e_i 构成的统计量推断误差项 e_i 是否存在序列相关。使用DW检验，应首先满足如下三个条件：

- (1) 误差项 e_i 的自相关为一阶自回归形式。

(2) 因变量的滞后值 Y_{t-1} 不能在回归模型中作解释变量。

(3) 样本容量应充分大 ($n > 15$)

DW检验的基本思想如下。给出假设

$H_0: \rho = 0, (\varepsilon_i \text{不存在序列相关})$

$H_1: \rho \neq 0, (\varepsilon_i \text{存在一阶序列相关})$

用残差值 e_i 计算统计量DW。

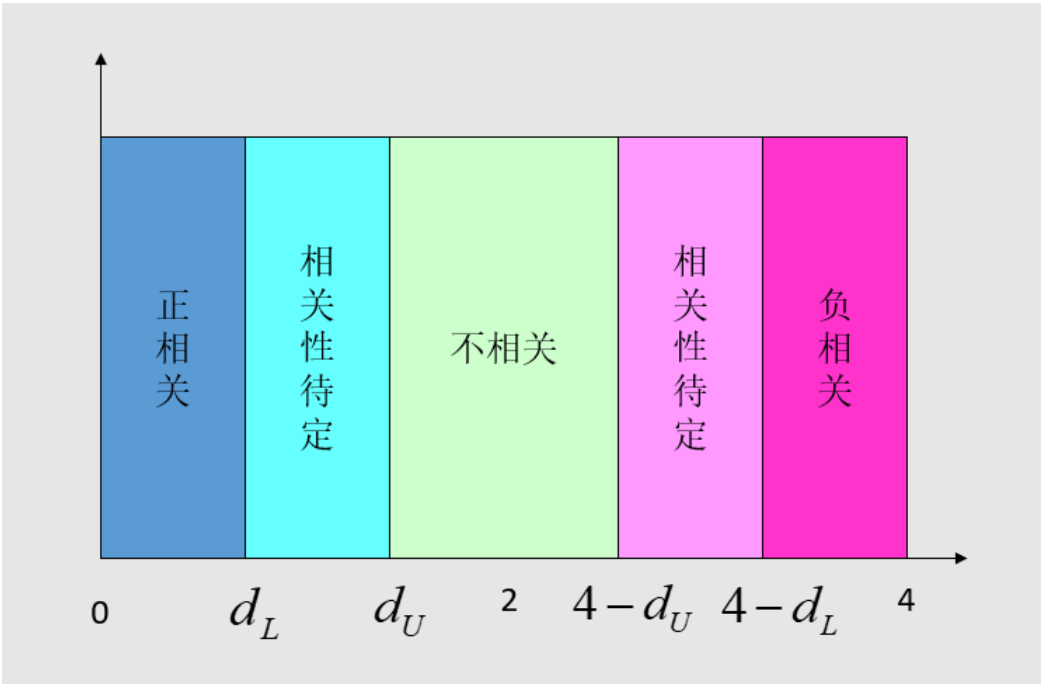
$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{\sum_{i=2}^n e_i^2 + \sum_{i=2}^n e_{i-1}^2 - 2 \sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} \approx \frac{2 \sum_{i=2}^n e_{i-1}^2 - 2 \sum_{i=2}^n e_i e_{i-1}}{\sum_{i=2}^n e_{i-1}^2}$$

因为 ρ 的取值范围是 $[-1, 1]$ ，所以DW统计量的取值范围是 $[0, 4]$ 。 ρ 与DW值的对应关系见表。

表 11-4 ρ 与 DW 值的对应关系及意义		
ρ	DW	ε_i 的表现
$\rho = 0$	$DW = 2$	ε_i 非序列相关
$\rho = 1$	$DW = 0$	ε_i 完全正序列相关
$\rho = -1$	$DW = 4$	ε_i 完全负序列相关
$0 < \rho < 1$	$0 < DW < 2$	ε_i 有某种程度的正序列相关
$-1 < \rho < 0$	$2 < DW < 4$	ε_i 有某种程度的负序列相关

实际中DW = 0, 2, 4 的情形是很少见的。当DW取值在 (0, 2) , (2, 4) 之间时, 怎样判别误差项 ε_i 是否存在序列相关呢? 推导统计量DW的精确抽样分布是困难的, 因为DW是依据残差 e_i 计算的, 而 e_i 的值又与 X_i 的形式有关。**DW检验与其它统计检验不同, 它没有唯一的临界值用来制定判别规则。**然而Durbin-Watson根据样本容量和被估参数个数, 在给定的显著性水平下, 给出了检验用的上、下两个临界值 d_U 和 d_L 。判别规则如下:

- (1) 若DW取值在 (0, d_L) 之间, 拒绝原假设 H_0 , 认为 ε_i 存在一阶正序列相关。
- (2) 若DW取值在 ($4 - d_L$, 4) 之间, 拒绝原假设 H_0 , 认为 ε_i 存在一阶负序列相关。
- (3) 若DW取值在 (d_U , $4 - d_U$) 之间, 接受原假设 H_0 , 认为 ε_i 非序列相关。
- (4) 若DW取值在 (d_L , d_U) 或 ($4 - d_U$, $4 - d_L$) 之间, 这种检验没有结论, 即不能判别 ε_i



当DW值落在“不确定”区域时，有两种处理方法。①加大样本容量或重新选取样本，重作DW检验。有时DW值会离开不确定区。②选用其它检验方法。

DW检验的判断一般根据DW检验临界值做出。R中lmtest的包提供了dwtest()函数，dwtest函数提供了对应的p值，我们可以根据p-value做出判断。

dwtest()函数的用法：

```
dwtest(formula, order.by = NULL, alternative = c("greater", "two.sided",  
"less"), iterations = 15, exact = NULL, tol= 1e-10, data = list())
```

In [91]:

```
library(lmtest)  
dw=dwtest(lm.in) ###DW检验  
dw
```

Durbin-Watson test

```
data: lm.in  
DW = 0.76951, p-value = 0.0003449  
alternative hypothesis: true autocorrelation is greater than  
0
```

示例中，DW检验统计量值为0.7695，对应的p-value是0.0003449，说明存在着正的序列相关性，这与定性检验的结果是一致的。

序列相关性克服

（一）序列相关的克服方法

如果模型的误差项存在序列相关，首先应分析产生序列相关的原因。如果序列相关是由于错误地设定模型的数学形式所致，那么就应当修改模型的数学形式。怎样查明序列相关是由于模型数学形式不妥造成的？一种方法是用残差 e_i 对解释变量的较高次幂进行回归，然后对新的残差作DW检验，如果此时序列相关消失，则说明模型的数学形式不妥。

如果序列相关是由于模型中省略了重要解释变量造成的，那么解决办法就是找出略去的解释变量，把它做为重要解释变量列入模型。怎样查明序列相关是由于略去重要解释变量引起的？一种方法是用残差 e_i 对那些可能影响因变量但又未列入模型的解释变量回归，并作显著性检验，从而确定该解释变量的重要性。如果是重要解释变量，应该列入模型。

只有当以上两种引起序列相关的原因都消除后，才能认为误差项 ε_i “真正”存在序列相关。在这种情况下，解决办法是变换原回归模型，使变换后的随机误差项消除序列相关，进而利用普通最小二乘法估计回归参数。

(二) 序列相关系数的估计

当序列相关系数未知时，通常使用科克伦-奥克特迭代法估计序列相关系数，其基本思想，是通过逐次迭代去寻求更为满意的序列相关系数的估计值，具体步骤如下：

第一，用普通最小二乘法对模型进行估计，然后对残差进行回归，即

$$e_i^{(1)} = \rho e_{i-1}^{(1)} + v_i$$

第二，用估计出来的 ρ 值进行广义差分，然后进行新的回归，即

$$Y_i^* = \beta_1(1 - \hat{\rho}) + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + \cdots + \beta_k X_{ki}^* + v_i$$

其中， $Y_i^* = Y_i - \hat{\rho}Y_{i-1}$ ； $X_{ji}^* = X_{ji} - \hat{\rho}X_{ji-1}$ ， $j = 1, 2, \cdots, k$ 。对变换后的方程进行估计，得到 $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ 。将这些修正过的参数代到原模型式，得到新的回归残差为

$$e_i^{(2)} = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \cdots - \hat{\beta}_k X_{ki}$$

第三，对新残差进行回归，即 $e_i^{(2)} = \rho e_{i-1}^{(2)} + v_i$ 得到新的 ρ 的估计值。这个迭代过程可以继续下去。

第四，将 ρ 的新估计值与前一个估计值比较，如果之间的差的绝对值小于容忍值，比如0.0001，就停止迭代。否则，继续迭代。

需要注意的是， ρ 的最后估计值不一定会使误差平方和最小，因为迭代法得到的可能是局部最小值，而不是全局最小值。

R中的`orcutt`包的`cochrane.orcutt()`函数可以求解序列相关系数 ρ 。示例的 ρ 通过Cochrane-Orcutt方法迭代后的值为0.4973338，该算法总共迭代了8次收敛，广义差分后的回归截距项为119.8985，回归系数为0.5834

In [2]:

```
#install.packages("orcutt")  
library(orcutt)  
cochrane.orcutt(lm.in)
```

Loading required package: lmtest

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

as.Date, as.Date.numeric

Cochrane-orcutt estimation for first order autocorrelation

Call:

```
lm(formula = yp ~ xp)
```

number of interaction: 8

rho 0.497334

Durbin-Watson statistic

(original): 0.76951 , p-value: 3.449e-04

(transformed): 1.39817 , p-value: 5.236e-02

coefficients:

(Intercept) xp

119.898510 0.583394

虚拟变量回归模型

哑变量 (Dummy Variable)，也叫虚拟变量，引入哑变量的目的是，将不能够定量处理的变量量化，如职业、性别对收入的影响，战争、自然灾害对GDP的影响，季节对某些产品（如冷饮）销售的影响等等。这种“量化”通常是通过引入“哑变量”来完成的。根据这些因素的属性类型，构造只取“0”或“1”的人工变量，通常称为哑变量 (dummy variables)，记为D。

举一个例子，假设变量“职业”的取值分别为：工人、农民、学生、企业职工、其他，5种选项，我们可以增加4个哑变量来代替“职业”这个变量，分别为D1 (1=工人/0=非工人)、D2(1=农民/0=非农民)、D3 (1=学生/0=非学生)、D4(1=企业职工/0=非企业职工)，最后一个选项“其他”的信息已经包含在这4个变量中了，所以不需要再增加一个D5 (1=其他/0=非其他)了。这个过程就是引入哑变量的过程，其实在结合分析 (conjoint analysis) 中，就是利用哑变量来分析各个属性的效用值的。

模型中引入虚拟变量的作用

- 1、**分离异常因素的影响**，例如分析我国GDP的时间序列，必须考虑“文革”因素对国民经济的破坏性影响，剔除不可比的“文革”因素。
- 2、检验不同属性类型对因变量的作用，例如工资模型中的文化程度、季节对销售额的影响。
- 3、提高模型的精度，相当于将不同属性的样本合并，扩大了样本容量（增加了误差自由度，从而降低了误差方差）

在模型中引入多个虚拟变量时，虚拟变量的个数应按下列原则确定：

- (1) 如果回归模型有截距项
有m种互斥的属性类型，在模型中引入 (m-1) 个虚拟变量。
- (2) 如果回归模型无截距项，有m个特征，设置m个虚拟变量

虚拟变量处理

利用caret包中的dummyVars函数进行虚拟变量处理

dummyVars函数:dummyVars建立一套完整的虚拟变量

先举一个简单的例子：

In [77]:

```
survey<-data.frame(service=c("very unhappy", "unhappy",  
                             "neutral", "happy", "very happy"))  
survey
```

service
very unhappy
unhappy
neutral
happy
very happy

可以直接增加一列rank，用数字代表不同情感

In [78]:

```
survey<-data.frame(service=c("very unhappy", "unhappy",  
                             "neutral", "happy", "very happy"), rank=c(1, 2, 3,  
4, 5))  
survey
```

service	rank
very unhappy	1
unhappy	2
neutral	3
happy	4
very happy	5

显然，对于单个变量进行如上处理并不困难，但是如果面对多个因子型变量都需要进行虚拟变量处理时，将会花费大量的时间。

下面用caret包中的dummyVars函数对因子变量进行哑变量处理。

In [88]:

```
#install.packages("caret")
library(caret)

customers<-data.frame(id=c(10, 20, 30, 40, 50), gender=c("male", "female",
                                                         "female", "male", "female"),
                      mood=c("happy", "sad", "happy",
                             "sad", "happy"), outcome=c(1, 1, 0, 0, 0))

customers
```

id	gender	mood	outcome
10	male	happy	1
20	female	sad	1
30	female	happy	0
40	male	sad	0
50	female	happy	0

利用dummyVars函数对customers数据进行哑变量处理

In [81]:

```
dmy<-dummyVars(~., data=customers)
```

对自身变量进行预测，并转换成data.frame格式

In [82]:

```
trsf<-data.frame(predict(dmy,newdata=customers))
trsf
```

id	gender.female	gender.male	mood.happy	mood.sad	outcome
10	0	1	1	0	1
20	1	0	0	1	1
30	1	0	1	0	0
40	0	1	0	1	0
50	1	0	1	0	0

从结果看，outcome并没有进行哑变量处理。

查看customers的数据类型

In [83]:

```
str(customers)
```

```
'data.frame':  5 obs. of  4 variables:
 $ id      : num  10 20 30 40 50
 $ gender  : Factor w/ 2 levels "female","male": 2 1 1 2 1
 $ mood    : Factor w/ 2 levels "happy","sad": 1 2 1 2 1
 $ outcome : num  1 1 0 0 0
```

可见，outcome的默认类型是numeric，现在这不是我们想要的。接下来将变量outcome转换成factor类型。

In [84]:

```
customers$outcome<-as.factor(customers$outcome)
str(customers)
```

```
'data.frame':  5 obs. of  4 variables:
 $ id      : num  10 20 30 40 50
 $ gender  : Factor w/ 2 levels "female","male": 2 1 1 2 1
 $ mood    : Factor w/ 2 levels "happy","sad": 1 2 1 2 1
 $ outcome : Factor w/ 2 levels "0","1": 2 2 1 1 1
```

customers中的变量outcome类型转换后，我们再次用dmy对该数据进行预测，并查看最终结果。

In [85]:

```
trsf<-data.frame(predict(dmy,newdata=customers))
trsf
```

id	gender.female	gender.male	mood.happy	mood.sad	outcome0
10	0	1	1	0	0
20	1	0	0	1	0
30	1	0	1	0	1
40	0	1	0	1	1
50	1	0	1	0	1

可见，outcome也已经进行了虚拟变量处理。

当然，也可以针对数据中的某一个变量进行虚拟变量（哑变量）处理。如我们需要对customers数据中的变量gender进行哑变量处理，可以执行以下操作：

In [86]:

```
dmy<-dummyVars(~gender,data=customers)
trfs<-data.frame(predict(dmy,newdata=customers))
trfs
```

gender.female	gender.male
0	1
1	0
1	0
0	1
1	0

对于两分类的因子变量，我们在进行虚拟变量处理后可能不需要出现代表相同意思的两列（例如：gender.female和gender.male）。这时候我们可以利用dummyVars函数中的fullRank参数，将此参数设置为TRUE。

In [87]:

```
dmy<-dummyVars(~., data=customers, fullRank=T)
trfs<-data.frame(predict(dmy, newdata=customers))
trfs
```

id	gender.male	mood.sad	outcome.1
10	1	0	1
20	0	1	1
30	0	0	0
40	1	1	0
50	0	0	0

虚拟变量线性回归

在线性回归分析中引入哑变量的目的是，可以考察定性因素对因变量的影响，引入哑变量有两种方式：加法方式与乘法方式。

所谓加法方式是指，哑变量作为单独的自变量，有独立的系数，从几何意义上来讲，就是只改变回归直线的截距（constant），不改变斜率（B）；

而乘法方式则正好相反，不改变截距，只改变斜率，因为哑变量在回归方程中不是作为一个独立的自变量，而是与其中某一个自变量相乘后作为一个自变量。

当然，也可以同时使用加法和乘法来引入哑变量，即同时改变截距和斜率。

In [97]:

```
dat=read.csv("./data/dummy.txt",header=T, sep=" ")
head(dat)
```

y	x1	x2	x3	x4	x5
29220	14010	98	115	15	f
29670	13260	98	26	8	m
136320	81240	96	199	19	m
111945	46260	96	120	19	m
24570	15510	95	46	12	m
36120	15810	93	8	16	f

将性别x5作为虚拟变量引入回归方程，建立当前年薪y关于受教育年限x4和性别虚拟变量x5的线性回归模型。

In [98]:

```
str(dat)
```

```
'data.frame': 36 obs. of 6 variables:
 $ y : int 29220 29670 136320 111945 24570 36120 41520 328
20 25620 32220 ...
 $ x1: int 14010 13260 81240 46260 15510 15810 20760 20010
16260 16260 ...
 $ x2: int 98 98 96 96 95 93 92 90 90 88 ...
 $ x3: int 115 26 199 120 46 8 168 205 191 252 ...
 $ x4: int 15 8 19 19 12 16 17 12 15 12 ...
 $ x5: Factor w/ 2 levels "f","m": 1 2 2 2 2 1 2 1 2 2 ...
```

1、加法模型

用relevel()函数,如

In [99]:

```
dat$x5=relevel(dat$x5, ref="m") #将性别x5作为虚拟变量, 以男性m作为参考水平,
女性f
fitlm=lm(y~x4+x5, data=dat) #建立y关于x4和x5的回归方程
summary(fitlm) #输出回归结果
```

Call:

```
lm(formula = y ~ x4 + x5, data = dat)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-29340	-13009	-3640	8341	63623

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11556.0	13143.5	-0.879	0.3856
x4	4434.4	915.5	4.844	2.92e-05 ***
x5f	-16840.5	6344.9	-2.654	0.0121 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18450 on 33 degrees of freedom

Multiple R-squared: 0.5247, Adjusted R-squared: 0.4958

F-statistic: 18.21 on 2 and 33 DF, p-value: 4.685e-06

得到回归方程： 男性： $y = -11556 + 4434.4x_4$ 女性: $y = -11556 + 4434.4x_4 - 16840.5$

2、乘法模型

In [100]:

```
dat$x5=relevel(dat$x5, ref="m") #将x5作为虚拟变量, 以男性m作为参考水平
fitlm2=lm(y~x4+x4:x5, data=dat)
summary(fitlm2)
```

Call:

```
lm(formula = y ~ x4 + x4:x5, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-30688	-11270	-362	8788	60865

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-15493.9	12076.9	-1.283	0.20845
x4	4786.8	866.2	5.526	3.9e-06 ***
x4:x5f	-1492.0	465.2	-3.207	0.00298 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17750 on 33 degrees of freedom

Multiple R-squared: 0.5602, Adjusted R-squared: 0.5336

F-statistic: 21.02 on 2 and 33 DF, p-value: 1.297e-06



得到回归方程： 男性： $y = -15493.9 + 4786.8x_4$ 女性： $y = -15493.9 + (4786.8 - 1492)x_4$

3、混合模型

In [101]:

```
fitlm3=lm(y~x4+x5+x4:x5, data=dat) #建立y关于x4和x5的回归方程
summary(fitlm3)
```

Call:

```
lm(formula = y ~ x4 + x5 + x4:x5, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-30950	-11624	-1141	7410	57183

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-28058	14352	-1.955	0.0594 .
x4	5642	1013	5.571	3.78e-06 ***
x5f	39195	25349	1.546	0.1319
x4:x5f	-4396	1932	-2.275	0.0298 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17390 on 32 degrees of freedom

Multiple R-squared: 0.5908, Adjusted R-squared: 0.5525

F-statistic: 15.4 on 3 and 32 DF, p-value: 2.236e-06



得到回归方程： 男性: $y = -28058 + 5642x_4$ 女性: $y = (-28058 + 39195) + (5642 - 4396)x_4$

非线性回归分析

非线性回归分析是线性回归分析的扩展，由于非线性回归的参数估计涉及非线性优化问题，计算比较困难，因此在计算机诞生前较少研究。20世纪七八十年代以来，随着计算机技术的发展，非线性回归的参数估计计算困难得到了克服，统计推断和预测分析技术也有很大发展。

主要介绍可线性化的非线性回归估计，不可线性化非线性回归的估计、应用以及R语言的实现。

我们知道线性回归模型的变量关系只是一种特例，现实中的变量关系大多是非线性的。关于非线性回归，不同的教材有不同的定义，有些教材定义只要因变量和自变量之间的关系是非线性就称为非线性回归，比如模型 $y_t = ae^{bx_t+u_t}$ 中 y_t 与 x_t 是非线性关系，称该模型为非线性回归；有些教材定义参数求解是非线性才称为非线性回归，比如模型 $y_t = ae^{bx_t+u_t}$ 中 y_t 与 x_t 虽然是非线性关系，但是通过变换后 $\ln y_t = \ln a + bx_t + u_t$ ，此时对参数 $\ln a$ 和 b 的求解是线性的，因此从参数求解的定义来看，该模型被归为线性回归，另外比如形如 $\mathbf{Y} = \mathbf{AK}^\alpha \mathbf{L}^\beta + \varepsilon$ 的非线性回归，就没法通过数学变换转换为线性回归，因此只能通过迭代算法等进行参数求解。

为了更全面地了解非线性回归，因此先涉及可线性化的非线性回归，然后再涉及不可先线性化的非线性回归。

可线性化的非线性回归

（一）Cobb-Douglas生产函数

经济中著名的Cobb-Douglas生产函数为 $Q = KL^\alpha C^{1-\alpha}$

其中， Q 表示产量； L 表示劳动力投入量； C 表示资本投入量； K 是常数； $0 < \alpha < 1$ 。这种生产函数是美国经济学家柯布和道格拉斯根据1899-1922年美国关于生产方面的数据研究得出的。更习惯的表达形式是

$$y_t = \beta_0 x_{t1}^{\beta_1} x_{t2}^{\beta_2} e^{u_t}$$

模型中的 y_t 与 x_t 是非线性关系，因此无法用OLS法直接估计，但可先作线性化处理。对式的两边同取对数，得：

$$\ln y_t = \ln \beta_0 + \beta_1 \ln x_{t1} + \beta_2 \ln x_{t2} + \mu_t$$

令 $y_t^* = \ln y_t, \beta_0^* = \ln \beta_0, x_{t1}^* = \ln x_{t1}, x_{t2}^* = \ln x_{t2}$,有 $y_t^* = \beta_0^* + \beta_1 x_{t1}^* + \beta_2 x_{t2}^* + \mu_t$.
上式为线性模型，可用OLS法估计后，再将参数代入到原模型。若回归参数 $\beta_1 + \beta_2 = 1$,称模型为规模报酬不变型; $\beta_1 + \beta_2 > 1$,称模型为规模报酬递增型;
 $\beta_1 + \beta_2 < 1$,称模型为规模报酬递减型。

示例：道格拉斯（Cobb-Douglas）生产函数

假设随机道格拉斯生产函数的模型为： $y_t = \beta_0 x_{t1}^{\beta_1} x_{t2}^{\beta_2} e^{u_t}, t = 1, \dots, T$, 其中其中 y_t 为产出, x_{t1} 为劳动投入, x_{t2} 为资本投入, μ_t 为随机干扰项。显然随机道格拉斯生产函数模型为非线性模型，需对方程两边取对数转化为线性模型

$\ln y_t = \ln \beta_0 + \beta_1 \ln x_{t1} + \beta_2 \ln x_{t2} + \mu_t$

收集了1958 - 1972年中国台湾地区农业部门的数据研究台湾农业部门的生产函数.

y	x2	x3
16607.7	275.5	17803.7
17511.3	274.4	18096.8
20171.2	269.7	18271.8
20932.9	267	19167.3
20406	267.8	19647.6
20831.6	275	20803.5
24806.3	283	22076.6
26465.8	300.7	23445.2
27403	307.5	24939
28628.7	303.7	26713.7
29904.5	304.7	29957.8
27508.2	298.6	31585.9
29035.5	295.5	33474.5
29281.5	299	34821.8
31535.8	288.1	41794.3

In [4]:

```
dat=read.csv(file="./data/douglas.csv")
lm1=lm(log(y)~log(x2)+log(x3), data=dat)
summary(lm1)
```

Call:

```
lm(formula = log(y) ~ log(x2) + log(x3), data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.15920	-0.02914	0.01179	0.04087	0.09640

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.3385	2.4495	-1.363	0.197939
log(x2)	1.4988	0.5398	2.777	0.016758 *
log(x3)	0.4899	0.1020	4.800	0.000433 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07481 on 12 degrees of freedom
Multiple R-squared: 0.889, Adjusted R-squared: 0.8705
F-statistic: 48.07 on 2 and 12 DF, p-value: 1.867e-06

从估计的结果可以看出1958 - 1972年台湾地区农业部门产出的劳动和资本弹性分别是1.4988和0.4899。即当保持资本投入不变，劳动投入增加1%，导致产出平均增加1.4988%；当保持劳动投入不变时，资本投入增加1%，导致产出平均增加0.4899%。两个弹性相加值为1.9887，就是规模报酬参数的取值，说明是规模报酬递增的。代回去可得 $\beta_0 = e^{-3.3385} = 0.0355$

可得柯布-道格拉斯（Cobb-Douglas）生产函数回归模型为

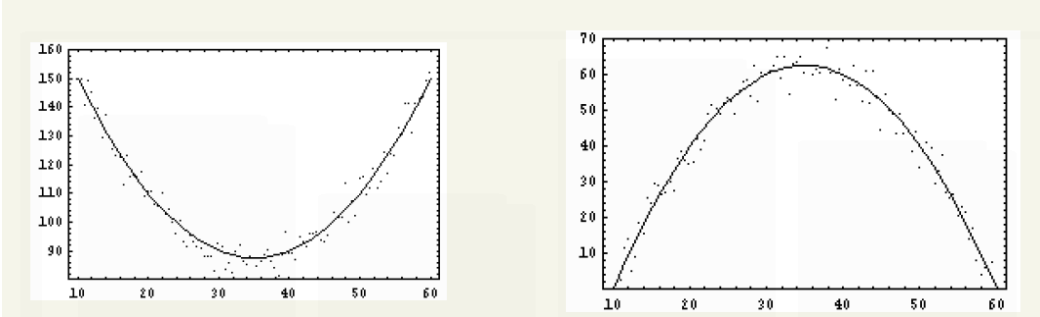
$$\hat{y}_i = 0.0355X_{2i}^{1.4988}X_{3i}^{0.4899}$$

(二) 多项式方程模型

二次项多项式方程的表达形式是

$$y_t = b_0 + b_1x_t + b_2x_t^2 + \mu_t$$

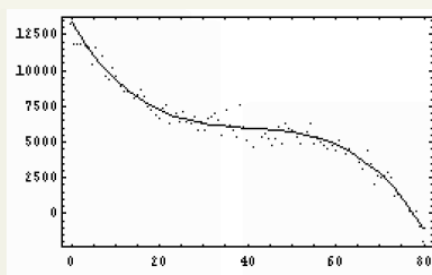
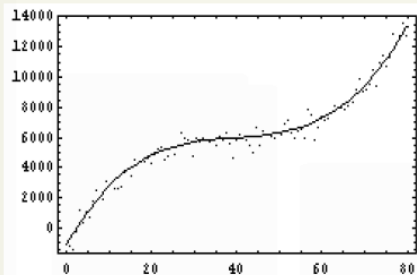
其中 $b_1 > 0, \quad b_2 > 0$ 和 $b_1 < 0, \quad b_2 < 0$ 情形的图形分别如图所示。令 $x_{2t} = x_t^2$ ，上式线性化为 $y_t = b_0 + b_1x_t + b_2x_{2t} + \mu_t$ 。经济学中的边际成本曲线、平均成本曲线与之相类似。



三次项多项式方程的表达形式是

$$y_t = b_0 + b_1x_t + b_2x_t^2 + b_3x_t^3 + \mu_t$$

其中 $b_1 > 0, \quad b_2 > 0, \quad b_3 > 0$ 和 $b_1 < 0, \quad b_2 < 0, \quad b_3 < 0$ 情形的图形分别如图所示。令 $x_{2t} = x_t^2, x_{3t} = x_t^3$ ，上式变为 $y_t = b_0 + b_1x_t + b_2x_{2t} + b_3x_{3t} + \mu_t$ ，这是一个三元线性回归模型。如经济学中的总成本曲线与图相似。



例:多项式回归

1609年，伽利略证明了一个物体在一个水平力的作用下，其下落轨道为一抛物线。为了验证这一事实，他做了一项实验并度量了两个变量：高度和距离，数据如下：

高度	距离
100	253
200	337
300	395
400	451
600	295
800	534
1000	574

通过数据描点，伽利略显然看到数据分布呈抛物线，且在数学上证明了它。在现代的眼光看来，如果确信为抛物线，我们可用二次回归模型得到那些系数。

In [5]:

```
x=c(100, 200, 300, 450, 600, 800, 1000)
y=c(253, 337, 395, 451, 495, 534, 574)
lm.1=lm(y~x) #一次模型y=a+bx
lm.2=lm(y~x+I(x^2)) #二次模型y=a+bx+cx2
lm.3=lm(y~x+I(x^2)+I(x^3)) #三次模型y=a+bx+cx2+dx3
```

In [6]:

```
summary(lm.1)$coef
summary(lm.2)$coef
summary(lm.3)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	269.4660734	24.18421016	11.142232	0.0001015488
x	0.3341268	0.04180624	7.992271	0.0004951455

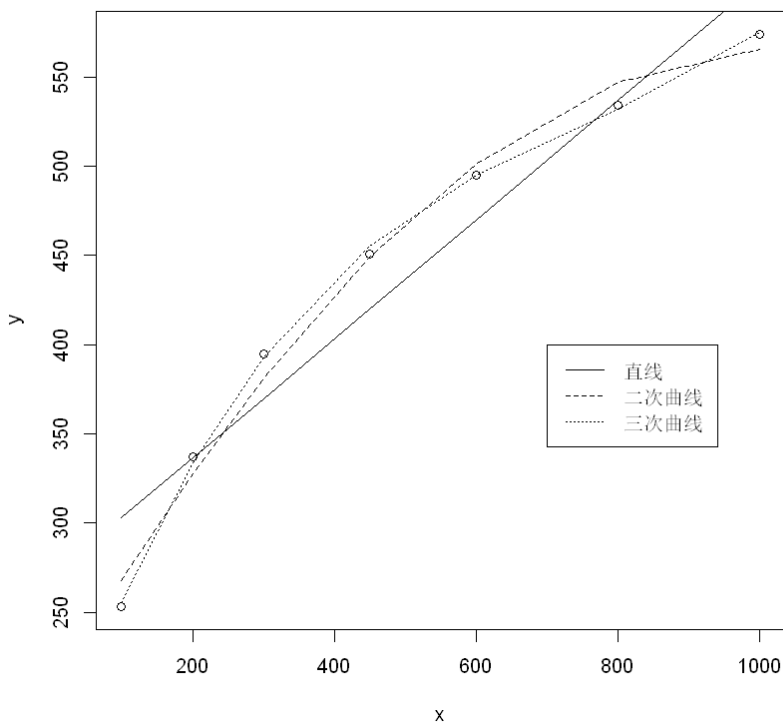
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.002120e+02	1.695062e+01	11.811481	0.0002940767
x	7.061816e-01	7.567631e-02	9.331607	0.0007341536
I(x^2)	-3.410076e-04	6.754293e-05	-5.048753	0.0072374199

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.554847e+02	8.182083e+00	19.003076	0.0003181899
x	1.118596e+00	6.453789e-02	17.332397	0.0004185186
I(x^2)	-1.254302e-03	1.360356e-04	-9.220394	0.0026985540
I(x^3)	5.550306e-07	8.183596e-08	6.782234	0.0065518838

注意必须使用结构I(x^2)，函数I()允许我们使用通常的幂符号。可尝试用二次和三次曲线去拟合数据，绘出图形

In [7]:

```
plot(x, y)
lines(x, fitted(lm.1), lty=1)
lines(x, fitted(lm.2), lty=2)
lines(x, fitted(lm.3), lty=3)
legend(700, 400, c("直线", "二次曲线", "三次曲线"), lty=1:3)
```



两条曲线似乎拟合得都不错，选择哪一条呢？可以比较它们的可决系数 (R^2)，哪个最大，哪个模型相对来说最好。

In [8]:

```
summary(lm.1)$r.squared
summary(lm.2)$r.squared
summary(lm.3)$r.squared
```

0.927406187106244

0.990153402234405

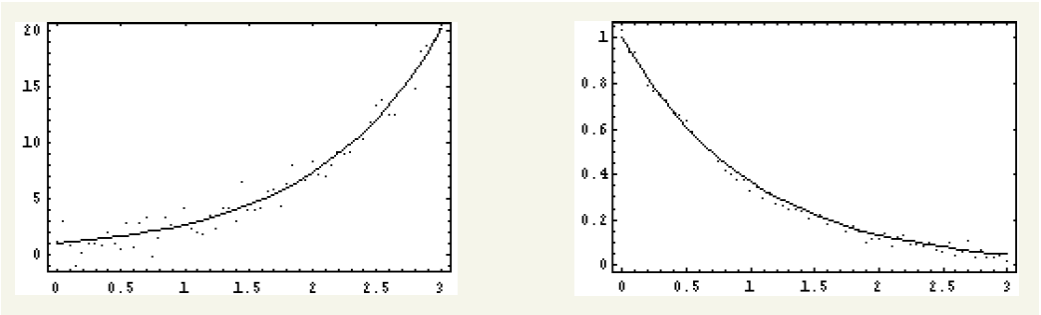
0.999397130995296

从中可以看出，相对来说，三次曲线的拟合效果最好。

(三) 指数函数模型

模型 $y_t = \beta_0 x_{t1}^{\beta_1} x_{t2}^{\beta_2} e^{u_t}, t = 1, \dots, T$ 的因变量和自变量之间的关系是一种指数函数关系，其中，其中 $b > 0$ 和 $b < 0$ 情形的图形分别如图所示。

$\ln y_t = \ln a + b x_t + u_t, \quad \wedge \quad y_t^* = \ln y_t, \quad a^* = \ln a$ 对上式等号两侧同取自然对数，得
 $y_t^* = a^* + b x_t + u_t$ ，对于参数 a^* 和 b 的求解可以通过线性回归的OLS得到，然后求
 $a = e^{a^*}$ 得到。



除此之外，还有半对数模型 $y_t = a + b \ln x_t + \mu_t$ ，双曲线模型 $1/y_t = a + b/x_t + \mu_t$ 等，这些模型的共同特点都是可以通过数学变换后转换为线性回归，然后我们可以利用线性回归方法求解参数，所以从参数求解上来看，本质上还是线性回归。

不可线性化的非线性回归

示例：中国粮食产量CES生产函数

我国是粮食大国，农业是我国的基础产业，粮食生产在我国农业生产中占有重要地位。影响我国粮食生产的主要因素有自然灾害、政策导向、生产投入和粮食流通等。国家政策导向对粮食生产的积极扶持作用，在时间序列样本区间内具有一致性；当我国粮食统一平价收购时，农民面临的是“无限的需求”。因此，假设影响我国粮食生产的主要因素是生产投入，即资本和劳动力。农业生产的特点决定了资本主要是土地和化肥；至于农业劳动力，我国一直是人工种植，但近年来呈现农业经济多种化经营的趋势，许多人从事副业生产。同时随着科技进步的影响，农业机械化水平有所提高，所以农业劳动力应包括农机动力和农业劳力。此外，气候等自然因素以及国家对农业和粮食生产的政策因素都粮食生产有影响。总的来说影响我国粮食产量的主要因素是：农业生产力、粮食播种面积和花费施用量。

年份	粮食产量	劳动力	播种面积	化肥
1975	28452	27561	121062	550000
1976	28631	27965	120743	597000
1977	28273	28124	120400	679000
1978	30477	28373	120587	884000
1979	33212	28692	119263	1086000
1980	32056	29181	117234	1269000
1981	32502	29836	114958	1335000
1982	35450	30917	113463	1513000
1983	38728	31209	114047	1660000
1984	40731	30927	112884	1740000
1985	37911	31187	108845	1776000
1986	39151	31311	110933	1931000
1987	40298	31720	111268	1999000
1988	39408	32308	110123	2141500
1989	40755	33284	112205	2357400
1990	44624	33336.4	113466	2590300
1991	43529	34186.3	112314	2805100
1992	44266	34037	110560	2930200
1993	45649	33258.2	110509	3151900
1994	44510	32690.3	109544	3317900
1995	46662	32334.5	110060	3593700
1996	50454	32260.4	110060	3827900
1997	49417	32434.9	112912	3980700
1998	51230	32626.4	113787	4083700
1999	50839	32911.8	113161	4124300
2000	46218	32732.5	108463	4146400
2001	45264	32451	106080	4253800

我们采用CES生产函数，取对数后模型如下：

$$\ln(Q_i/M_i) = \beta_0 + 1/\beta_3 \ln(\beta_1(L_i/M_i))^{\beta_3} + \beta_2(K_i/M_i)^{\beta_3} + \varepsilon_i$$

该模型是非线性的，无法转换为线性回归，因此需要用非线性参数估计方法求解。

In [13]:

```
fdat<-read.table(file="/data/fproduct.txt",header=T)
YM<-fdat$y/fdat$m#单位播种面积产量
LM<-fdat$l/fdat$m#单位播种面积农业劳动力
KM<-fdat$k/fdat$m#单位播种面积化肥使用量
summary(YM)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2348	0.2976	0.3622	0.3579	0.4185	0.4584

In [10]:

```
summary(LM)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2277	0.2660	0.2867	0.2780	0.2952	0.3079

In [11]:

```
summary(KM)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.543	12.474	19.446	21.330	31.470	40.100

（一）非线性模型的参数估计与迭代算法

参数估计是非线性回归分析的核心。非线性回归分析参数估计的方法也有多种，主要方法有非线性最小二乘估计(nonlinear least squares estimator, NLS)和非线性最大似然估计(Nonlinear maximum likelihood estimator, NMLE)。当非线性模型的随机误差项 $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 时，除了误差方差的估计以外，非线性回归的最大似然估计与最小二乘估计也是相同的。

实际上，非线性最小二乘估计和非线性极大似然估计都是非线性优化的问题，非线性优化有多种算法，比如有格点搜索法、二次爬坡法、高斯牛顿法和牛顿拉夫森法等。

(1) 高斯 - 牛顿法

高斯 - 牛顿法 (Gauss-Newton) 的基本思路是：非线性最小二乘估计的问题在于最小二乘函数 $S(\beta)$ 中的 f ，也就是回归模型

$$Y = f(X, \beta) + \varepsilon$$

的趋势部分不是参数向量 β 的线性函数，因此最优化问题

$$\min_{\hat{\beta}_1, \dots, \hat{\beta}_p} S(\beta) = [Y - f(X, \hat{\beta})]' [Y - f(X, \hat{\beta})]$$

的求解存在计算上的困难。当 f 是连续可微时，可以在某组参数初始值处作一阶泰勒级数展开，得到 f 的线性近似，把这个线性近似函数代入最小二乘函数得到参数的二次函数，克服参数估计计算的困难。但一阶泰勒级数展开得到的近似函数与原函数是有差异的，用上述级数展开近似的方法得到的参数估计也有偏差，偏差程度与泰勒级数展开的初始值与参数真实值的偏差相关。提高参数估计准确程度的途径是改进泰勒级数展开的初始值，方法是把已经得到的参数估计作为新的参数初始值，重新进行泰勒级数展开和参数估计。这种方法可以反复运用，直到得到比较理想的参数估计值。这种计算非线性回归参数估计的迭代算法称为“高斯 - 牛顿法”。

(2) 牛顿—拉夫森法

牛顿—拉夫森 (Newton - Raphson) 法的基本思想也是利用泰勒级数展开近似，通过迭代运算寻找最小二乘函数最优解的数值解法。牛顿—拉夫森法的迭代运算，相当于在前一个参数估计向量的基础上，按单位移动幅度（通常称为“步长”）搜索更好的参数估计值，因此牛顿—拉夫森法也是一种搜索法。牛顿—拉夫森法的优点是搜索方向和步长的确定比较科学，因此找到满足精度要求最优水平的搜索次数一般要小一些。牛顿—拉夫森方法的缺点是迭代运算中需要反复计算梯度向量，特别是海塞矩阵的逆矩阵，因此计算工作量很大。事实上，人们在实际应用中常常并不按照牛顿—拉夫森法进行搜索，而是根据一些简单法则确定搜索的步长，如“双向线性搜索法”就是其中常用的方法之一。

R里面优化的基本函数是optimize()和optim()。此外，R的stats包里提供专门做非线性最小二乘法的函数nls()。该函数的用法是nls(formula, data, start, control, algorithm, trace, subset, weights, na.action, model, lower, upper, ...) nls()函数的默认迭代算法是高斯-牛顿算法。Start设定初始值，参数trace=T时，会把参数迭代过程返回来。

In [14]:

```
n1.f<-nls(YM~A*(LM^a)*(KM^(1-a)),start=list(A=0.5,a=1),trace=T)
```

```
1.389927 : 0.5 1.0
0.1364105 : 0.0946038 0.4346868
0.1261383 : 0.1180032 0.4856855
0.1137523 : 0.1430868 0.5287686
0.1080431 : 0.1949993 0.6011286
0.08680564 : 0.2471957 0.6514697
0.06094021 : 0.3421454 0.7200617
0.01485043 : 0.4708669 0.7745319
0.006992903 : 0.4805984 0.7680351
0.006989582 : 0.4805334 0.7682233
0.006989582 : 0.4805400 0.7682266
```

In [15]:

```
summary(n1.f)
```

Formula: $YM \sim A * (LM^a) * (KM^{(1 - a)})$

Parameters:

	Estimate	Std. Error	t value	Pr(> t)
A	0.48054	0.04014	11.97	7.56e-12 ***
a	0.76823	0.01902	40.38	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01672 on 25 degrees of freedom

Number of iterations to convergence: 10

Achieved convergence tolerance: 5.875e-07

也可以使用极大似然法进行估计，maxLik包里提供了maxLik () 函数，但是需要先写出对数似然函数。其用法是

```
maxLik(logLik, grad = NULL, hess = NULL, start, method, constraints=NULL, ...)
```

In [16]:

```
#install.packages("maxLik")
library(maxLik)
loglik=function(para) {
  N=length(YM)
  e=YM-para[1]*LM^para[2]*KM^(1-para[2])
  ll=-0.5*N*(1+log(2*pi)-log(N))-0.5*N*log(sum(e^2))
  return(ll)
}
res=maxLik(loglik, start=c(0.5,1), method="NR")
summary(res)
```

Loading required package: miscTools

Please cite the 'maxLik' package as:

Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. Computational Statistics 26(3), 443-458. DOI 10.1007/s00180-010-0217-1.

If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum or 'tracker' at maxLik's R-Forge site:

<https://r-forge.r-project.org/projects/maxlik/>

Maximum Likelihood estimation

Newton-Raphson maximisation, 8 iterations

Return code 2: successive function values within tolerance limit

Log-Likelihood: 73.18747

2 free parameters

Estimates:

	Estimate	Std. error	t value	Pr(> t)
[1,]	0.48054	0.03895	12.34	<2e-16 ***
[2,]	0.76823	0.01846	41.62	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
NULL
```

(二) 初始值选取

在利用迭代算法进行非线性回归参数估计时，初始值的选择是一个值得重视的问题，如果我们想要得到较好的结果和提高工作效率，必须认真对待参数估计值的选择。但参数初始值的选择并没有一般法则。尽量接近参数真实值或最终估计值，最好是参数真实值的一致估计，是正确的初始值选择原则。但该原则的实用价值不大，因为参数真实值不可能知道，而一致估计量正是我们要求出的最小二乘估计量。在实践中，人们常常运用的是如下的经验方法：

- 1、利用参数的经济意义。
- 2、模型函数在特定点的性质。
- 3、降维法。

(三) 收敛性

理论上，非线性优化的迭代运算应该在梯度向量等于0，也就是满足最优化的一阶条件处终止。但实际上这通常做不到，因为函数和导数的计算都有累积的舍入误差。因此，迭代算法一般是以某种收敛标准作为终止迭代的信号，而不是真正满足一阶条件。

判断收敛和终止迭代并没有一致接受的标准。常用的标准主要有：

- (1) 目标函数（最小二乘函数）的改进已小于给定的小正数，即

$$|S(\boldsymbol{\beta}^{i+1}) - S(\boldsymbol{\beta}^i)| \leq \varepsilon, \varepsilon \text{ 即任意小正数};$$

- (2) 第二，参数值的变化小于给定的小正数。当模型只有一个参数时即
$$\|\boldsymbol{\beta}^{i+1} - \boldsymbol{\beta}^i\| \leq \varepsilon$$

- (3) 第三，梯度向量的模小于给定的小正数，即
$$\|\mathbf{g}(\boldsymbol{\beta}^{i+1})\| \leq \varepsilon$$

非线性回归评价和假设检验

(一) 可决系数

由于反映线性回归模型的可决系数 $R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$ 和调整的可决系数 $R^{-2} = 1 - \frac{n-1}{n-k} (1 - R^2)$ 。不涉及参数估计量的分布性质，也不需要做以这些分布性质为基础的假设检验，因此非线性导致的问题并不影响该统计量在评价回归方程拟合度方面的作用，仍然是评价非线性模型合理程度的基本指标。它们在非线性回归分析中的使用方法仍然是与在线性回归分析中相同的。

(二) 参数显著性的F检验

除了对高斯 - 牛顿法非线性回归可以利用最后一次线性近似函数线性回归的t检验以外，检验非线性模型参数的显著性还有多种其他方法，下面这个渐近F分布的统计量就是其中的一种方法，即

$$F(g, n - k) = \frac{[S(\beta_R) - S(\beta)] / g}{S(\beta) / (n - k)}$$

这个统计量分子、分母中的 β 是未对非线性模型参数施加约束时的参数估计， β_R 则是对模型的某些参数施加0假设约束后的参数估计， $S(\beta)$ 和 $S(\beta_R)$ 分别是对应两种参数估计的残差平方和， g 是0约束参数的数量。

很显然，如果施加0约束的参数本身对模型的影响没有显著性，那么上述F统计量的数值会很小，如果这些施加0约束的参数对模型的影响是明显的，那么该统计量的数值会较大，就会有显著性。因此，我们可以通过检验该统计量的显著性来判断模型参数的显著性。

虽然上述 F 统计量与线性回归模型的 F 统计量形式是相似的，但因为模型是非线性的，因此 $S(\boldsymbol{\beta})$ 和 $S(\boldsymbol{\beta}_R)$ 并不服从 χ^2 分布，该统计量并不严格服从 F 分布，只是近似服从 F 分布。在样本容量较大时，该统计量的分布与 F 分布很接近。我们可以利用 F 分布检验该统计量的显著性，但检验结果论的准确程度会受到一定影响，运用时应该加以注意。

(三) 似然比检验

似然比检验与 F 检验在本质上一致的另一种非线性模型参数显著性检验。似然比检验的统计量为

$$\lambda = -2 \left(\ln L(\boldsymbol{\beta}_R) - \ln L(\boldsymbol{\beta}) \right) = -2 \ln \frac{L(\boldsymbol{\beta}_R)}{L(\boldsymbol{\beta})}$$

式中 $\boldsymbol{\beta}$ 与 $\boldsymbol{\beta}_R$ 的含义与上述 F 检验统计量中同， $L(\boldsymbol{\beta})$ 与 $L(\boldsymbol{\beta}_R)$ 则分别是它们各自对应的非线性模型被自变量的似然函数值。似然函数即随机变量得到特定观测值序列的联立分布概率密度函数。

我们假设非线性模型的误差项服从均值为0的正态分布，那么上述统计量中的对数似然函数为

$$\ln L(\boldsymbol{\beta}) = -\frac{n}{2} \left[1 + \ln(2\pi) + \ln \left(\frac{\mathbf{e}'\mathbf{e}}{n} \right) \right]$$

式中， \mathbf{e} 是残差向量。如果参数估计采用的是最大似然估计，那么其中的 $\frac{\mathbf{e}'\mathbf{e}}{n}$ 实际上就是误差方差的估计。相应的有约束时，模型的对数似然函数为

$$\ln L(\boldsymbol{\beta}_R) = -\frac{n}{2} \left[1 + \ln(2\pi) + \ln \left(\frac{\mathbf{e}_R'\mathbf{e}_R}{n} \right) \right]$$

因此，统计量 λ 为

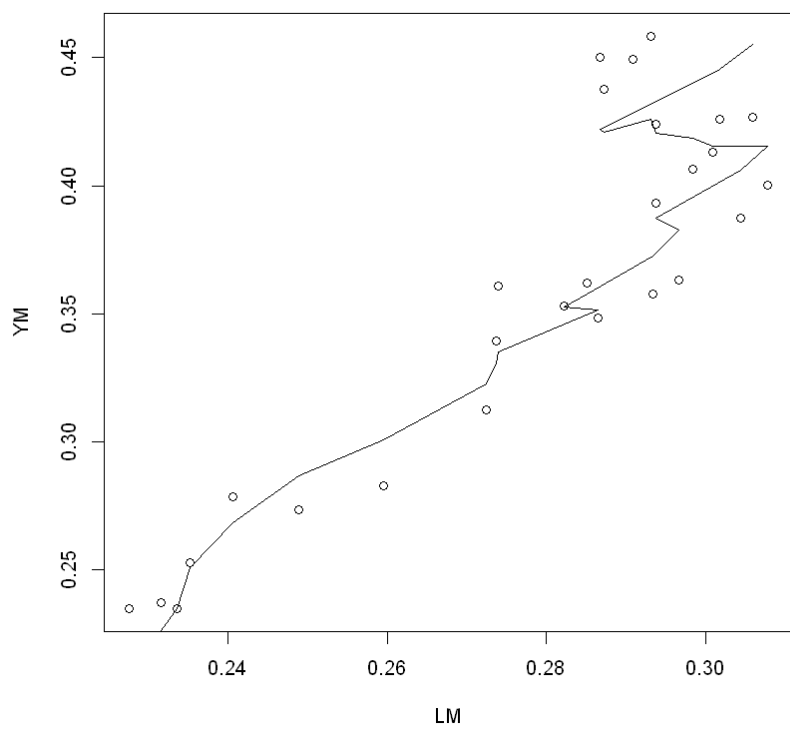
$$\lambda = n \left(\ln \left(\frac{\mathbf{e}' \mathbf{e}}{n} \right) - \ln \left(\frac{\mathbf{e}'_R \mathbf{e}_R}{n} \right) \right) = n \ln \left(\frac{\mathbf{e}' \mathbf{e}}{\mathbf{e}'_R \mathbf{e}_R} \right)$$

对于大样本来说，该统计量渐近服从自由度为约束数量 g 的 χ^2 分布，因此可以根据 χ^2 分布检验 λ 的显著性。当该统计量比给定显著水平的 χ^2 分布临界值大时，拒绝 H_0 假设，认为所检验的参数是显著的，否则认为检验的参数是不显著的。

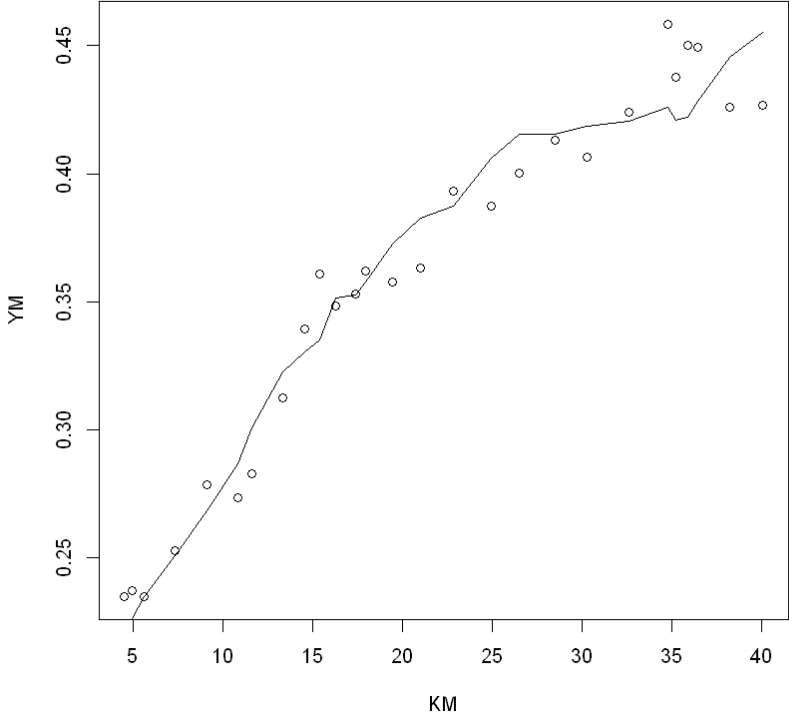
除了上述检验方法以外，非线性回归还有其他一些检验方法，如沃尔德检验和拉格朗日乘数检验，但多数方法在本质上都是相似的。

In [19]:

```
plot(YM~LM)
lines(LM, fitted(nl.f))
plot(YM~KM)
lines(KM, fitted(nl.f))
Rsq<-1-sum(resid(nl.f)^2)/(sum((YM-mean(YM))^2))
Rsq
```



0.947829258119358



In [20]:

```
adjRsqr<-1-(length(YM)-1)/(length(YM)-2)*(1-Rsqr)
adjRsqr
```

0.945742428444132

In [21]:

```
lm.f<-lm(YM~LM+KM)
anova(nl.f, lm.f) #F test
```

Res.Df	Res.Sum Sq	Df	Sum Sq	F value	Pr(>F)
25	0.006989582	NA	NA	NA	NA
24	0.006661378	1	0.0003282038	1.182472	0.2876539

In [22]:

```
Q=-2*(logLik(nl.f)-logLik(lm.f))
df.Q=df.residual(nl.f)-df.residual(lm.f)
1-pchisq(Q, df.Q)
```

'log Lik.' 0.2544787 (df=3)

二元选择模型

前面介绍了连续型的因变量建模分析，但实际中，并非所有的变量都是连续型的数据，有时因变量是离散型的数据，这时候我们需要用广义线性模型（generalizedlinearmodel,GLM）。

离散因变量（Discrete Dependent Variable）是指取值为0、1、2....等离散值的变量。在多数情况下，这些取值一般没有实际的意义，仅代表某一事件的发生，或者是用于描述某一事件发生的次数。根据取值的特点，离散因变量可以分为二元变量（binary variable）、多分变量和计数变量(count variable)。二元变量的取值一般为1和0，当取值为1时表示某件事情的发生，取值为0则表示不发生，比如信用卡客户发生违约的记为1，不违约的记为0。因变量为二元变量的模型称为二元选择模型(BinaryChoice Model)。

示例：为了考察一种新的经济学教学方法对学生成绩的影响，进行了调查，共得到了32个样本数据。数据见表。GRADE取1表示新近学习成绩提高，0表示其他；GPA是平均积分点；TUCE是以往经济学成绩；PSI取1表示受到新的经济学教学方法的指导，0表示其他。假如想要了解GPA，TUCE和PSI因素对学生成绩是否有影响？以及根据学生的GPA，TUCE和PSI预测学生成绩是否会提高？该如何建模分析？

表：新教学方法对成绩的影响数据

obs	GRADE	GPA	TUCE	PSI
1	0	2.66	20	0
2	0	2.89	22	0
3	0	3.28	24	0
4	0	2.92	12	0
5	1	4.00	21	0
6	0	2.86	17	0
7	0	2.76	17	0
8	0	2.87	21	0
9	0	3.03	25	0
10	1	3.92	29	0
11	0	2.63	20	0
12	0	3.32	23	0
13	0	3.57	23	0
14	1	3.26	25	0
15	0	3.53	26	0
16	0	2.74	19	0
17	0	2.75	25	0
18	0	2.83	19	0
19	0	3.12	23	1
20	1	3.16	25	1
21	0	2.06	22	1
22	1	3.62	28	1
23	0	2.89	14	1
24	0	3.51	26	1
25	1	3.54	24	1
26	1	2.83	27	1
27	1	3.39	17	1
28	0	2.67	24	1

obs	GRADE	GPA	TUCE	PSI
29	1	3.65	21	1
30	1	4.00	23	1
31	0	3.10	21	1
32	1	3.66	19	1

线性概率模型原理

假设二元选择模型：

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i \quad i = 1, 2, \dots, N$$

线性概率模型（LinearProbabilityModel，LPM）。实际上就是用普通的线性回归方法对二元取值的因变量直接建模。

其中， $Y_i = \begin{cases} 1 & \text{某一事件发生} \\ 0 & \text{某一事件不发生} \end{cases}$

是二元取值的因变量。

\mathbf{X}_i 是包含常数项的 k 元设计矩阵，假设在给定 \mathbf{X}_i 的时候，某一事件发生是概率为 p ,不发生的概率为 $1 - p$ ，即

$$\text{Prob}\left(Y_i = 1 \mid \mathbf{X}_i\right) = p, \text{Prob}\left(Y_i = 0 \mid \mathbf{X}_i\right) = 1 - p$$

对于线性概率模型，可以采用普通最小二乘法进行估计，但是会存在一些问题：

- 对式的拟合的结果是对某一事件发生的平均概率的预测，但是，拟合值并不能保证在0和1之间，完全有可能出现大于1和小于0的情形。

$$\hat{Y}_i = \text{Prob}\left(Y_i | X_i\right) = X_i' \hat{\beta}$$

- 由于Y是二元变量，因此扰动项也应该是二元变量，它应该服从二项分布，而不是我们通常假定的正态分布。

$$\varepsilon_i = \begin{cases} 1 - X_i' \beta & (Y_i = 1) \\ -X_i' \beta & (Y_i = 0) \end{cases}$$

- 在LPM中，扰动项是异方差的。

$$\text{Var}(\varepsilon_i) = (1 - X_i' \beta)^2 \cdot p + (-X_i' \beta)^2 \cdot (1 - p) = (1 - X_i' \beta) X_i' \beta \neq \text{常数}$$

- 由于因变量是二元选择的结果，因此按传统线性回归模型所计算的判定系数 R^2 不再有实际的意义。

$$Count_R^2 = \frac{\text{正确观测的个数}}{\text{总观测值个数}}$$

- 边际效应的分析：LPM的边际效应是一个常数，它与解释变量取值的大小无关。

$$\frac{\partial E(Y | X_i)}{\partial X_i} = \beta$$

由于LPM存在诸多问题，因此对于二元离散因变量一般不推荐使用LPM，而是需要其他更为科学的方法。

Probit模型原理

在LPM中， $\mathbf{X}_i' \hat{\beta}$ 不能保证概率的取值在0和1之间。为了保证估计的概率能在0-1，一个直接的想法就是在外套上分布函数。如果将 $F(\mathbf{X}_i' \beta)$ 用标准正态分布函数 $\Phi(\cdot)$ ，即

$$\text{Prob}(Y_i = 1 | \mathbf{X}_i) = \Phi(\mathbf{X}_i' \beta) = \int_{-\infty}^{\mathbf{X}_i' \beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

$\Phi(\mathbf{X}_i' \beta)$ 是正态分布的分布函数，其取值范围是[0,1]，这时的概率模型称为Probit模型。

二元选择模型也可以从潜变量回归模型去解释，考察以下模型

$$Y_i^* = \mathbf{X}_i' \beta + \varepsilon_i \quad i = 1, 2, \dots, T$$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

其中， Y_i^* 是潜变量或隐变量(Latent Variable)，上式称为潜变量反应函数 (Latent Response Function) 或指示函数(Index Function)。

假设：

$$A1: E(\varepsilon_i | \mathbf{X}_i) = 0$$

A2: ε_i 是*i. i. d.*的正态分布

$$A3: \text{rank}(\mathbf{X}_i) = k$$

在A1—A3的假定之下，考察模型中 Y_i 的概率特征：

$$\begin{aligned}\text{Prob}\left(Y_i = 1 \mid \mathbf{X}_i\right) &= \text{Prob}\left(Y_i^* > 0 \mid \mathbf{X}_i\right) \\ &= \text{Prob}\left(\mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i > 0 \mid \mathbf{X}_i\right) \\ &= \text{Prob}\left(\varepsilon_i > -\mathbf{X}_i'\boldsymbol{\beta} \mid \mathbf{X}_i\right) \\ &= \int_{-\mathbf{X}_i'\boldsymbol{\beta}}^{\infty} f\left(\varepsilon_i\right) d\varepsilon_i\end{aligned}$$

当 $f\left(\varepsilon_i\right)$ 为标准正态分布的概率密度函数

$\phi\left(\varepsilon_i\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_i^2}{2}\right)$ 时，上式可以写成：

$$\begin{aligned}\text{Prob}\left(Y_i = 1 \mid \mathbf{X}_i\right) &= 1 - \int_{-\infty}^{-\mathbf{X}_i'\boldsymbol{\beta}} \phi\left(\varepsilon_i\right) d\varepsilon_i \\ &= 1 - \Phi\left(-\mathbf{X}_i'\boldsymbol{\beta}\right) \\ &= \Phi\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\end{aligned}$$

这样，上式正是Probit模型。

Logit模型原理

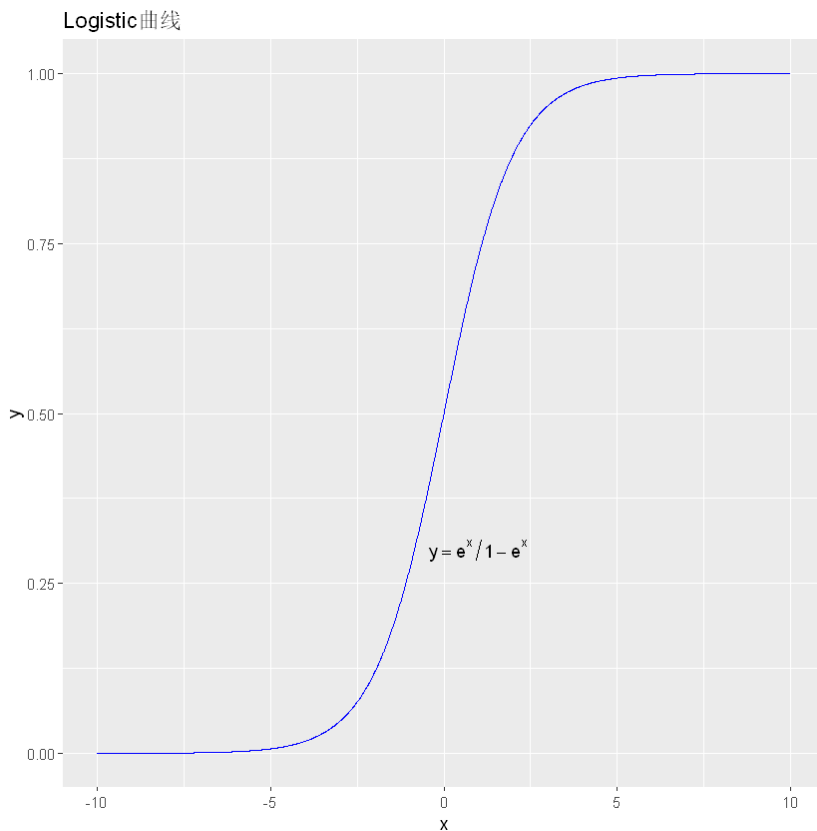
如果 $F\left(\mathbf{X}_i'\boldsymbol{\beta}\right)$ 取Logistic分布函数 $\Lambda(\cdot)$ ，则产生的概率模型为Logit模型：

$$\text{Prob}(Y_i = 1 | \mathbf{X}_i) = \Lambda(\mathbf{X}_i' \boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i' \boldsymbol{\beta})}$$

同样， $\Lambda(\cdot)$ 的取值也在0和1之间。

In [41]:

```
x <- seq(from = -10, to = 10, by = 0.01)
y = exp(x)/(1+exp(x))
library(ggplot2)
p <- ggplot(data = NULL, mapping = aes(x = x, y = y)) +
  geom_line(colour = 'blue') +
  annotate('text', x = 1, y = 0.3, label = 'y==e^x / 1-e^x', parse = TRUE) +
  ggtitle('Logistic曲线')
p
```



假设：

$$A1: E(\varepsilon_i | \mathbf{X}_i) = 0$$

A2: ε_i 是*i. i. d.* 的Logistic分布；

$$A3: \text{rank}(\mathbf{X}_i) = k$$

在A1—A3的假定之下，考察模型中中 Y_i 的概率特征：

$$\begin{aligned} \text{Prob}(Y_i = 1 | \mathbf{X}_i) &= \text{Prob}(Y_i^* > 0 | \mathbf{X}_i) \\ &= \text{Prob}(\mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i > 0 | \mathbf{X}_i) \\ &= \text{Prob}(\varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta} | \mathbf{X}_i) \\ &= 1 - \int_{-\infty}^{-\mathbf{X}_i' \boldsymbol{\beta}} f(\varepsilon_i) d\varepsilon_i \\ &= 1 - \frac{\exp(-\mathbf{X}_i' \boldsymbol{\beta})}{1 + \exp(-\mathbf{X}_i' \boldsymbol{\beta})} \\ &= \frac{\exp(\mathbf{X}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i' \boldsymbol{\beta})} = \Lambda(\mathbf{X}_i' \boldsymbol{\beta}) \end{aligned}$$

边际效应分析

对于Probit模型来说，其边际效应为：

$$\frac{\partial \text{Prob}(Y_i = 1 | \mathbf{X}_i)}{\partial \mathbf{X}_i} = \Phi'(\mathbf{X}_i' \boldsymbol{\beta}) \boldsymbol{\beta} = \phi(\mathbf{X}_i' \boldsymbol{\beta}) \boldsymbol{\beta}$$

对于Logit模型，其边际效应为：

$$\frac{\partial \text{Prob} \left(Y_i = 1 \mid \mathbf{X}_i \right)}{\partial \mathbf{X}_i} = \Lambda' \left(\mathbf{X}_i' \boldsymbol{\beta} \right) \boldsymbol{\beta} = \Lambda \left(\mathbf{X}_i' \boldsymbol{\beta} \right) \left(1 - \Lambda \left(\mathbf{X}_i' \boldsymbol{\beta} \right) \right) \boldsymbol{\beta}$$

其中： $\Lambda'(\cdot) = \Lambda(\cdot)(1 - \Lambda(\cdot))$

从两式中可以看到，Probit和Logit模型中解释变量对 Y_i 取值为1的概率的边际影响不是常数，它会随着解释变量取值的变化而变化。所以对于probit和logit模型，其回归系数的解释就没有线性回归那么直接了。其边际效应分析也没有那么线性回归那么直接。一种常用的方法，是计算其平均边际效应。即对于非虚拟的解释变量，一般是用其样本均值代入到两式中，估计出平均的边际影响。但是，对于虚拟解释变量而言，则需要先分别计算其取值为1和0时 $\text{Prob}(Y_i = 1 \mid \mathbf{X}_i)$ 的值，二者的差即为虚拟解释变量的边际影响。

最大似然估计

Probit和Logit模型的参数估计常用最大似然法。

对于Probit或Logit模型来说，

$$\text{Prob}(Y_i = 1 \mid \mathbf{X}_i) = F(\mathbf{X}_i' \boldsymbol{\beta})$$

$$\text{Prob}(Y_i = 0 \mid \mathbf{X}_i) = 1 - F(\mathbf{X}_i' \boldsymbol{\beta})$$

所以似然函数为： $L = \prod_{i=1}^N F(\mathbf{X}_i' \boldsymbol{\beta})^{Y_i} (1 - F(\mathbf{X}_i' \boldsymbol{\beta}))^{1 - Y_i}$

对数似然函数为：

$$\log L = \sum_{i=1}^N \left\{ Y_i \cdot \log F(\mathbf{X}_i' \boldsymbol{\beta}) + (1 - Y_i) \cdot \log [1 - F(\mathbf{X}_i' \boldsymbol{\beta})] \right\}$$

最大化logL的一阶条件为：

$$\begin{aligned}\frac{\partial \log L}{\partial \beta} &= \sum_{i=1}^N \left\{ Y_i \cdot \mathbf{X}_i \frac{f_i}{F_i} + (1 - Y_i) \cdot \mathbf{X}_i \frac{-f_i}{1 - F_i} \right\} \\ &= \sum_{i=1}^N \left\{ \mathbf{X}_i f_i \frac{Y_i - F_i}{F_i(1 - F_i)} \right\} = 0\end{aligned}$$

由于上式不存在封闭解，所以要用非线性求解的叠代法求解。常用的有Newton-Raphson法或二次爬坡法(Quadratic hill climbing)。

似然比检验和拟合优度

似然比检验类似于检验模型整体显著性的F检验，原假设为全部解释变量的系数都为0，检验的统计量LR为：

$$LR = 2 \left(\ln L - \ln L_0 \right)$$

其中，lnL为对概率模型进行MLE估计的对数似然函数值，lnL0为只有截距项的模型的对数似然函数值，往往也称为空模型，即模型中不包含任何解释变量。当原假设成立时，LR的渐近分布是自由度为 $k - 1$ （即除截距项外的解释变量的个数）的 χ^2 分布。

对于Probit和Logit模型，同样可以计算 $Count_R^2$ 以反映模型的拟合优度。此外，还可以计算类似于传统 R^2 的McFadden似然比指数(McFadden's Likelihood Ratio Index)来度量拟合优度。似然比指数的定义为：

$$McFaddenR^2 = 1 - \frac{\ln L}{\ln L_0}$$

$McFaddenR^2$ 总是介于0和1之间。当所有的斜率系数都为0时， $McFaddenR^2=0$ ，但是， $McFaddenR^2$ 不会恰好等于1。 $McFaddenR^2$ 越大，表明拟合得越好。

R中可以用glm()函数拟合广义线性模型，包含probit模型和logit模型。Glm()的形式与lm()类似，只是多了一些参数。函数的基本形式为：`glm(formula,family=family(link=function),data=)`其中，formula是模型表达式，与lm()的表达式一致。family参数设置模型连接函数对应分布族，比如gaussian分布，Poisson分布等

glm()的参数

分布族(family)	连接函数
binomial	(link= “logit”或“probit”或“cauchit”)
gaussian	(link= “identity”)
gamma	(link= “inverse”或“identity”或“log”)
inverse.gaussian	(link= “1/mu^2”)
poisson	(link= “log” 或“identity”或“sqrt”)
quasi	(link= “identity”, variance=”constant”)
quasibinomial	(link= “logit”)
quasipoisson	(link= “log”)

与分析线性回归模型时lm()连用的许多函数在glm()中也有对应的形式，

函数	功能
summary()	给出拟合模型的信息摘要
coefficients()/coef()	列出拟合模型的参数
confint()	给出模型参数的置信区间
residuals()	列出拟合模型的残差值
anova()	生成两个拟合模型的方差分析表
plot()	生成评价拟合模型的诊断图
predict()	用拟合的模型对原有数据进行拟合或者对新数据进行预测
aic()	计算拟合模型的AIC值

我们利用LPM去分析经济学教学新方法效果的例子。首先，我们求出数据的基本描述统计量，R程序如下。

In [23]:

```
grade=read.table(file="./data/grade.txt",heade=T)
```

In [24]:

```
summarys=function(x) {  
  list(mean=mean(x), max=max(x), min=min(x), sd=sd(x))  
} #自编一个求基本描述统计量简单函数  
  
summarys(subset(grade, PSI==0)$GRADE) #subset() 筛选PSI=0的数据  
summarys(subset(grade, PSI==1)$GRADE)  
summarys(grade$GRADE)  
summarys(subset(grade, PSI==0)$GPA)  
summarys(subset(grade, PSI==1)$GPA)  
summarys(grade$GPA)  
summarys(subset(grade, PSI==0)$TUCE)  
summarys(subset(grade, PSI==1)$TUCE)  
summarys(grade$TUCE)  
summarys(subset(grade, PSI==0)$PSI)  
summarys(subset(grade, PSI==1)$PSI)  
summarys(grade$PSI)
```

\$mean
0.166666666666667
\$max
1
\$min
0
\$sd
0.383482494423685

\$mean
0.571428571428571
\$max
1
\$min
0
\$sd
0.513552591013095

\$mean
0.34375
\$max
1
\$min
0
\$sd
0.482558704434814

\$mean
3.10111111111111
\$max
4
\$min
2.63
\$sd
0.422985035041762

\$mean
3.13785714285714
\$max
4
\$min
2.06
\$sd
0.53351055892411

\$mean
3.1171875
\$max
4
\$min
2.06
\$sd
0.466712831337948

\$mean
21.55555555555556
\$max
29
\$min
12
\$sd
4.00326663998911

\$mean
22.4285714285714
\$max
28
\$min
14
\$sd
3.85734635197838

\$mean
21.9375
\$max
29
\$min
12
\$sd
3.9015092199748

\$mean
0
\$max
0
\$min
0
\$sd
0

\$mean
1
\$max
1
\$min
1
\$sd
0

\$mean
0.4375
\$max
1
\$min
0
\$sd
0.504016128774185

数据的基本描述

变量	均值	最大值	最小值	标准差
GRADE				
PSI=0	0.166667	1	0	0.383482
PSI=1	0.571429	1	0	0.513553
全部	0.34375	1	0	0.482559
GPA				
PSI=0	3.101111	4	2.63	0.422985
PSI=1	3.137857	4	2.06	0.533511
全部	3.117188	4	2.06	0.466713
TUCE				
PSI=0	21.55556	29	12	4.003267
PSI=1	22.42857	28	14	3.857346
全部	21.9375	29	12	3.901509
PSI				
PSI=0	0	0	0	0
PSI=1	1	1	1	0
全部	0.4375	0	1	0.504016

首先设定以下线性概率模型：

$$GRADE = \beta_0 + \beta_1 \cdot GPA + \beta_2 \cdot TUCE + \beta_3 \cdot PSI + \varepsilon$$

其中，GRADE取1表示新近学习成绩提高，0表示其他；GPA是平均积分点；TUCE是以往经济学成绩；PSI取1表示受到新的经济学教学方法的指导，0表示其他。

用OLS估计这一线性概率模型的R程序和结果如下

In [25]:

```
lpm=lm(GRADE~GPA+TUCE+PSI, data=grade)
summary(lpm)
```

Call:

```
lm(formula = GRADE ~ GPA + TUCE + PSI, data = grade)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.78153	-0.27731	0.00531	0.21089	0.81145

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.49802	0.52389	-2.859	0.00793 **
GPA	0.46385	0.16196	2.864	0.00784 **
TUCE	0.01050	0.01948	0.539	0.59436
PSI	0.37855	0.13917	2.720	0.01109 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3881 on 28 degrees of freedom

Multiple R-squared: 0.4159, Adjusted R-squared: 0.3533

F-statistic: 6.646 on 3 and 28 DF, p-value: 0.001571

从分析的结果看，在5%的显著性水平上，PSI对GRADE的影响是显著的。也就是说，当GPA和TUCE都一样的情况下，接受过新的教学方法的学生与没有接受过新的教学方法的学生相比，学习成绩提高的概率要多0.3786。此外，GPA对成绩提高的边际影响是0.46，也就是说，在其他条件相同的情况下，GPA每增加1,学习成绩提高的概率是46%。

In [26]:

```
grade=read.table(file="./data/grade.txt",header=T)
grade.probit=glm(GRADE~GPA+TUCE+PSI,family=binomial
(link="probit"),data=grade)###probit模型
summary(grade.probit)
```

Call:

```
glm(formula = GRADE ~ GPA + TUCE + PSI, family = binomial(1
ink = "probit"),
    data = grade)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9392	-0.6508	-0.2229	0.5934	2.0451

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-7.45231	2.57152	-2.898	0.00376 **
GPA	1.62581	0.68973	2.357	0.01841 *
TUCE	0.05173	0.08119	0.637	0.52406
PSI	1.42633	0.58695	2.430	0.01510 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 41.183 on 31 degrees of freedom
Residual deviance: 25.638 on 28 degrees of freedom
AIC: 33.638

Number of Fisher Scoring iterations: 6

从分析结果来看，Probit的设定形式是： $\text{Prob}(GRADE_i = 1 | X_i) = \Phi(X_i' \beta)$ ，其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数。将系数的估计结果代入，得到估计的模型为：

$$\text{Prob}(GRADE_i = 1 | \mathbf{X}_i) = \Phi(-7.452320 + 1.625810 \cdot GPA + 0.051729 \cdot TUCE + 1.426332$$

似然比检验和拟合优度

In [27]:

```
#install.packages("lmtest")  
library(lmtest) #需首先安装lmtest包，并载入包  
lrtest(grade.probit) #LR检验
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
4	-12.81880	NA	NA	NA
1	-20.59173	-3	15.54585	0.001404896

上面的检验结果还给出了有关模型的似然比检验和拟合优度的信息。 L 为Model1 logLik的值为-12.81880， L_0 为Model2 logLik的值，为-20.59173， LR 值即为 $Chisq$ ，为15.54585，它对应的P值为0.001405,因此，它是显著的，表明模型整体是显著的。

$McFaddenR^2$

In [28]:

```
McFa.Rsquare=function(glm.object){ #glm()估计结果作为函数的输入变量  
deviance=glm.object$deviance  
null.deviance=glm.object$null.deviance  
McFa.Rsquare=1-(deviance/null.deviance)  
list(McFadden.Rsquare=McFa.Rsquare)  
}  
  
McFa.Rsquare(grade.probit)
```

\$McFadden.Rsquare = 0.377478033279115

Logit 模型的R程序和估计结果：

$$\text{Prob}\left(\text{GRADE}_i = 1 \mid \mathbf{X}_i \right) = \Lambda\left(-13.02135 + 2.826113 \cdot \text{GPA} + 0.095158 \cdot \text{TUCE} + 2.378688 \right)$$

In [29]:

```
grade.logit=glm(GRADE~GPA+TUCE+PSI,
                family=binomial(link="logit"), data=grade) #注意link设为logit
summary(grade.logit)
```

Call:

```
glm(formula = GRADE ~ GPA + TUCE + PSI, family = binomial(link = "logit"),
    data = grade)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9551	-0.6453	-0.2570	0.5888	2.0966

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-13.02135	4.93127	-2.641	0.00828 **
GPA	2.82611	1.26293	2.238	0.02524 *
TUCE	0.09516	0.14155	0.672	0.50143
PSI	2.37869	1.06456	2.234	0.02545 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 41.183 on 31 degrees of freedom
Residual deviance: 25.779 on 28 degrees of freedom
AIC: 33.779

Number of Fisher Scoring iterations: 5

似然比检验和拟合优度

In [30]:

```
library(lmtest)
lrtest(grade.logit) #####LR检验
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
4	-12.88963	NA	NA	NA
1	-20.59173	-3	15.40419	0.001501879

$McFaddenR^2$

In [31]:

```
McFa.Rsquare(grade.logit)
```

\$McFadden.Rsquare = 0.374038295372105

Probit和Logit模型中的回归系数与线性概率模型不同，并没有实际的经济意义。但可以依据边际效应分析的表达式计算解释变量GPA和TUCE对GRADE的平均边际影响。

In [32]:

```
coe=coef(grade, probit) #提取probit模型系数
probit=dnorm(coe[1]+coe[2]*mean(grade$GPA)+coe[3]*mean(grade$TUCE)+coe[4]*
mean(grade$PSI)) #求probit模型平均边际影响
(m. gpa=coe[2]*probit) #求GPA平均边际影响
(m. tuce=coe[3]*probit)
(m. PSI=coe[4]*probit)
coe.l=coef(grade, logit) #提取logit模型系数
logit=dlogis(coe.l[1]+coe.l[2]*mean(grade$GPA)+coe.l[3]*mean(grade$TUCE)+c
oe.l[4]*mean(grade$PSI)) #求logit模型平均边际影响
(m. gpa.l=coe.l[2]*logit)
(m. tuce.l=coe.l[3]*logit)
(m. PSI.l=coe.l[4]*logit)
```

GPA: 0.533348351717009

TUCE: 0.0169695429837901

PSI: 0.467908342757711

GPA: 0.53385882207147

TUCE: 0.0179754893942193

PSI: 0.449339276828118

Probit 和Logit 模型边际影响分析对比

model	Probit模型	Logit模型	
	$F'\left(\begin{smallmatrix} -' \\ X_i\hat{\beta} \end{smallmatrix}\right) = f\left(\begin{smallmatrix} -' \\ X_i\hat{\beta} \end{smallmatrix}\right)$	$\phi\left(\begin{smallmatrix} -' \\ X_i\hat{\beta} \end{smallmatrix}\right) = 0.3281$	$\Lambda'\left(\begin{smallmatrix} -' \\ X_i\hat{\beta} \end{smallmatrix}\right) = 0.1889$

变量	回归系数	平均边际影响	回归系数	平均边际影响
GPA	1.625810	0.5333	2.826113	0.5339
TUCE	0.051729	0.0170	0.095158	0.0180
PSI	1.426332	0.4644	2.378688	0.4493

解释变量GPA和TUCE对因变量GRADE的边际影响是通过将相应的回归系数乘以

$F'\left(X_i\hat{\beta}\right)$ 的值得到的。例如，对于Probit模型，GPA和GRADE的边际影响等于

$1.625810 \cdot 0.3281 = 0.5333$ 。但是这一算法不适用于像PSI这类离散的解释变量。对于Logit模型，PSI对GRADE的平均边际影响是PSI分别取值为1和0，GRADE取值为1时的概率差，即：

$$\Phi(-7.452320 + 1.625810 \cdot GPA + 0.051729 \cdot TUCE + 1.426332 \cdot 1) - \Phi(-7.452320 + 1.625810 \cdot GPA + 0.051729 \cdot TUCE) = 0.4644$$

表中的边际影响分析取的是解释变量的均值，但实际上解释变量对因变量的影响是非线性的。例如，在Probit模型当中，PSI对GRADE的影响是随着GPA和TUCE取值的不同而不同。假设TUCE取均值，则这一边际影响的函数为：

$$\Phi(-7.452320 + 1.625810 \cdot GPA + 0.051729 \cdot TUCE + 1.426332 \cdot 1) - \Phi(-7.452320 + 1.6$$

用各样本的GPA值代入这一边际影响函数，可以得到在不同的GPA水平下，PSI对GRADE的边际影响。

In [33]:

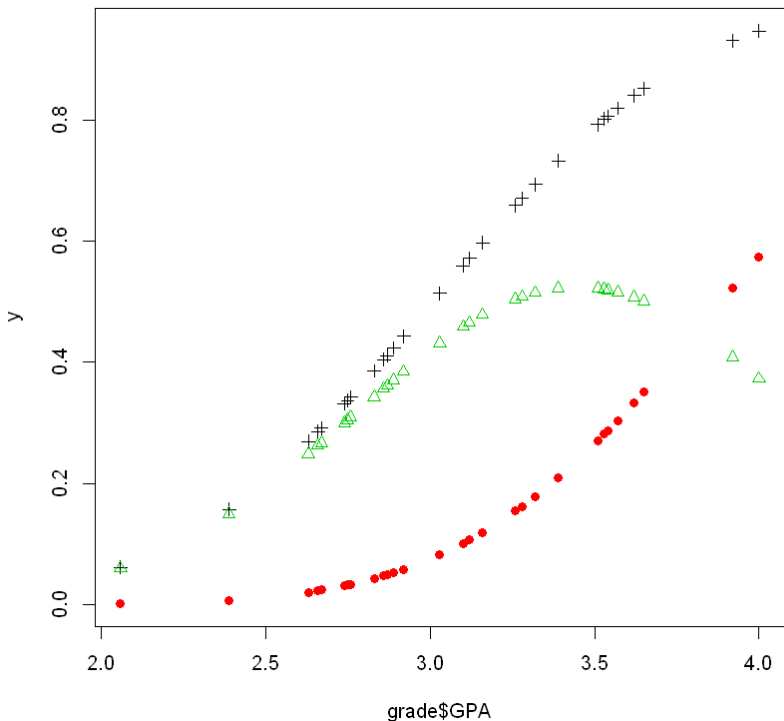
```
head(grade)
```

鐳繹bs	GRADE	GPA	TUCE	PSI
1	0	2.66	20	0
2	0	2.89	22	0
3	0	3.28	24	0
4	0	2.92	12	0
5	1	4.00	21	0
6	0	2.86	17	0

In [34]:

```
yz<-seq(0,1,0.1)
mean_TUCE=mean(grade$TUCE)
mean_TUCE
probit1=pnorm(coe[1]+coe[2]*grade$GPA+coe[3]*mean(grade$TUCE)+coe[4]*1) #求
probit模型平均边际影响
probit2=pnorm(coe[1]+coe[2]*grade$GPA+coe[3]*mean(grade$TUCE))
probit3=probit1-probit2
y=data.frame(probit1,probit2,probit3)
matplot(grade$GPA,y,pch=c(3,16,2))
```

21.9375



两曲线之间的差距并不是不变的，开始随着GPA的提高而增大，但当GPA高于一定水平后，这一差距又开始缩小。即绿色的曲线先上升又下降。

Logistic回归模型

在日常学习或工作中经常会使用线性回归模型对某一事物进行预测，例如预测房价、身高、GDP、学生成绩等，发现这些被预测的变量都属于连续型变量。然而有些情况下，被预测变量可能是二元变量，即成功或失败、流失或不流失、涨或跌等，对于这类问题，线性回归将束手无策。这个时候就需要另一种回归方法进行预测，即 Logistic回归。

在实际应用中，Logistic模型主要有三大用途：

- 1) 寻找危险因素，找到某些影响因变量的"坏因素"，一般可以通过优势比发现危险因素；
- 2) 用于预测，可以预测某种情况发生的概率或可能性大小；
- 3) 用于判别，判断某个新样本所属的类别。

Logistic模型实际上是一种回归模型，但这种模型又与普通的线性回归模型又有一定的区别：

- 1) Logistic回归模型的因变量为二分类变量；
- 2) 该模型的因变量和自变量之间不存在线性关系；
- 3) 一般线性回归模型中需要假设独立同分布、方差齐性等，而Logistic回归模型不需要；
- 4) Logistic回归没有关于自变量分布的假设条件，可以是连续变量、离散变量和虚拟变量；
- 5) 由于因变量和自变量之间不存在线性关系，所以参数(偏回归系数)使用最大似然估计法计算。

logistic回归模型概述

广义线性回归是探索“响应变量的期望”与“自变量”的关系，以实现非线性关系的某种拟合。这里面涉及到一个“连接函数”和一个“误差函数”，“响应变量的期望”经过连接函数作用后，与“自变量”存在线性关系。选取不同的“连接函数”与“误差函数”可以构造不同的广义回归模型。当误差函数取“二项分布”而连接函数取“logit函数”时，就是常见的“logistic回归模型”，在0-1响应的问题中得到了大量的应用。

Logistic回归主要通过构造一个重要的指标：发生比来判定因变量的类别。在这里我们引入概率的概念，把事件发生定义为 $Y=1$ ，事件未发生定义为 $Y=0$ ，那么事件发生的概率为 p ，事件未发生的概率为 $1-p$ ，把 p 看成 x 的线性函数；

回归中，最常用的估计是最小二乘估计，因为使得 p 在 $[0,1]$ 之间变换，最小二乘估计不太合适，希望有一种估计法能让 p 在趋近与0和1的时候变换缓慢一些（不敏感），这种变换是我们想要的，于是引入Logit变换,对 $p/(1-p)$ 也就是发生与不发生的比值取对数，也称对数差异比。经过变换后， p 对 x 就不是线性关系了。

logistic回归的公式可以表示为：

$$P = \frac{\exp \left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \right)}{1 + \exp \left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \right)}$$

其中 P 是响应变量取1的概率，在0-1变量的情形中，这个概率就等于响应变量的期望。

这个公式也可以写成：

$$\text{logit}(P) = \ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

可以看出，logistic回归是对0-1响应变量的期望做logit变换，然后与自变量做线性回归。参数估计采用极大似然估计，显著性检验采用似然比检验。

建立模型并根据AIC准则选择模型后，可以对未知数据集进行预测，从而实现分类。模型预测的结果是得到每一个样本的响应变量取1的概率，为了得到分类结果，需要设定一个阈值 p_0 ——当 p 大于 p_0 时，认为该样本的响应变量为1，否则为0。阈值大小对模型的预测效果有较大影响，需要进一步考虑。首先必须明确模型预测效果的评价指标。

对于0-1变量的二分类问题，分类的最终结果可以用表格表示为：

		预测值	
		0	1
实际值	0	a	b
	1	c	d

其中， d 是“实际为1而预测为1”的样本个数， c 是“实际为1而预测为0”的样本个数，其余依此类推。

显然地，主对角线所占的比重越大，则预测效果越佳，这也是一个基本的评价指标——总体准确率 $(a+d)/(a+b+c+d)$ 。

准确（分类）率=正确预测的正反例数/总数 $Accuracy=(a+d)/(a+b+c+d)$

误分类率=错误预测的正反例数/总数 $Error\ rate=(b+c)/(a+b+c+d)=1-Accuracy$

正例的覆盖率=正确预测到的正例数/实际正例总数

$Recall(True\ Positive\ Rate, \ or\ Sensitivity)=d/(c+d)$

正例的命中率=正确预测到的正例数/预测正例总数

$Precision(Positive\ Predicted\ Value,PV+)=d/(b+d)$

负例的命中率=正确预测到的负例个数/预测负例总数

$Negative\ predicted\ value(PV-)=a/(a+c)$

通常将上述矩阵称为“分类矩阵”。一般情况下，我们比较关注响应变量取1的情形，将其称为Positive（正例），而将响应变量取0的情形称为Negative（负例）。常见的例子包括生物实验的响应、营销推广的响应以及信用评分中的违约等等。针对不同的问题与目的，我们通常采用ROC曲线与lift曲线作为评价logistic回归模型的指标。

1) ROC曲线

设置了两个相应的指标：TPR与FPR。

TPR: True Positive Rate（正例覆盖率），将实际的1正确地预测为1的概率， $d/(c+d)$ 。

FPR: False Positive Rate，将实际的0错误地预测为1的概率， $b/(a+b)$ 。

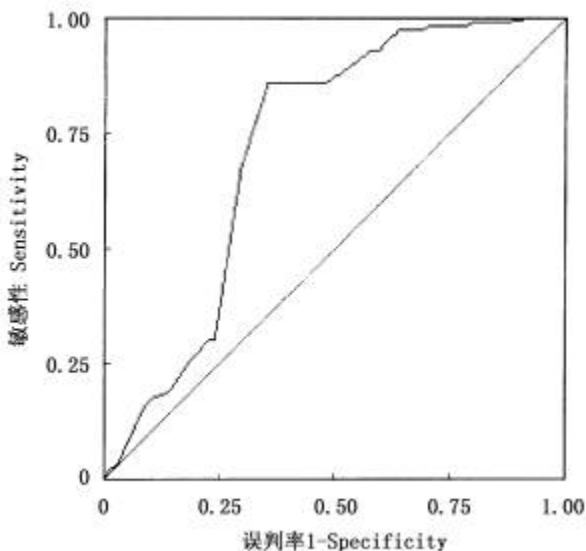
TPR也称为Sensitivity（即生物统计学中的敏感度），也可以称为“正例的覆盖率”——将实际为1的样本数找出来的概率。覆盖率是重要的指标，例如若分类的目标是找出潜在的劣质客户（响应变量取值为1），则覆盖率越大表示越多的劣质客户被找出。

类似地，1-FPR其实就是“负例的覆盖率”，也就是把负例正确地识别为负例的概率。

TPR与FPR相互影响，而我们希望能够使TPR尽量地大，而FPR尽量地小。影响TPR与FPR的重要因素就是上文提到的“阈值”。当阈值为0时，所有的样本都被预测为正例，因此TPR=1，而FPR=1。此时的FPR过大，无法实现分类的效果。随着阈值逐渐增大，被预测为正例的样本数逐渐减少，TPR和FPR各自减小，当阈值增大至1时，没有样本被预测为正例，此时TPR=0，FPR=0。

由上述变化过程可以看出，TPR与FPR存在同方向变化的关系（这种关系一般是非线性的），即，为了提升TPR（通过降低阈值），意味着FPR也将得到提升，两者之间存在类似相互制约的关系。我们希望能够牺牲较少FPR的基础上尽可能地提高TPR，由此画出了ROC曲线。

ROC曲线的全称为“接受者操作特性曲线”（receiver operating characteristic），其基本形式为：



当预测效果较好时，ROC曲线凸向左上角的顶点。平移图中对角线，与ROC曲线相切，可以得到TPR较大而FPR较小的点。模型效果越好，则ROC曲线越远离对角线，极端的情形是ROC曲线经过 $(0, 1)$ 点，即将正例全部预测为正例而将负例全部预测为负例。ROC曲线下的面积可以定量地评价模型的效果，记作AUC，AUC越大则模型效果越好。

当我们分类的目标是将正例识别出来时（例如识别有违约倾向的信用卡客户），我们关注TPR，此时ROC曲线是评价模型效果的准绳。

2) lift曲线

在营销推广活动中，我们的首要目标并不是尽可能多地找出那些潜在客户，而是提高客户的响应率。客户响应率是影响投入产出比的重要因素。此时，我们关注的不再是TPR（覆盖率），而是另一个指标：命中率。

回顾前面介绍的分类矩阵，正例的命中率是指预测为正例的样本中的真实正例的比例，即 $d/(b+d)$ ，一般记作PV。

在不使用模型的情况下，我们用先验概率估计正例的比例，即 $(c+d)/(a+b+c+d)$ ，可以记为k。

定义提升值 $lift = PV/k$ 。

lift揭示了logistic模型的效果。例如，若经验告诉我们10000个消费者中有1000个是我们的潜在客户，则我们向这10000个消费者发放传单的效率是10%（即客户的响应率是10%）， $k = (c+d)/(a+b+c+d) = 10\%$ 。通过对这10000个消费者进行研究，建立logistic回归模型进行分类，我们得到有可能比较积极的1000个消费者， $b+d=1000$ 。如果此时这1000个消费者中有300个是我们的潜在客户， $d=300$ ，则命中率PV为30%。此时，我们的提升值 $lift = 30\%/10\% = 3$ ，客户的响应率提升至原先的三倍，提高了投入产出比。

为了画lift图，需要定义一个新的概念depth深度，这是预测为正例的比例， $(b+d)/(a+b+c+d)$ 。

与ROC曲线中的TPR和FPR相同，lift和depth也都受到阈值的影响。

当阈值为0时，所有的样本都被预测为正例，因此 $depth=1$ ，而 $PV = d/(b+d) = (0+d)/(0+b+0+d) = k$ ，于是 $lift=1$ ，模型未起提升作用。随着阈值逐渐增大，被预测为正例的样本数逐渐减少，depth减小，而较少的预测正例样本中的真实正例比例逐渐增大。当阈值增大至1时，没有样本被预测为正例，此时 $depth=0$ ，而 $lift=0/0$ 。

由此可见，lift与depth存在相反方向变化的关系。在此基础上作出lift图：

In []:

```

```

与ROC曲线不同，lift曲线凸向 (0, 1) 点。我们希望在尽量大的depth下得到尽量大的lift（当然要大于1），也就是说这条曲线的右半部分应该尽量陡峭。

至此，我们对ROC曲线和lift曲线进行了描述。这两个指标都能够评价logistic回归模型的效果，只是分别适用于不同的问题：

如果是类似信用评分的问题，希望能够尽可能完全地识别出那些有违约风险的客户（不使一人漏网），我们需要考虑尽量增大TPR（覆盖率），同时减小FPR（减少误杀），因此选择ROC曲线及相应的AUC作为指标；

如果是做类似数据库精确营销的项目，希望能够通过对全体消费者的分类而得到具有较高响应率的客户群，从而提高投入产出比，我们需要考虑尽量提高lift（提升度），同时depth不能太小（如果只给一个消费者发放传单，虽然响应率较大，却无法得到足够多的响应），因此选择lift曲线作为指标。

相关R应用包

普通二分类 logistic 回归 用系统的 glm

因变量多分类 logistic 回归

有序分类因变量：用 MASS 包里的 polrb

无序分类因变量：用 nnet 包里的 multinom

条件logistic回归，用 survival 包里的 clogit

案例：本文用例来自于John Maindonald所著的《Data Analysis and Graphics Using R》一书，其中所用的数据集是anesthetic，数据集来自于一组医学数据，其中变量conc表示麻醉剂的用量，move则表示手术病人是否有所移动，而我们用nomove做为因变量，因为研究的重点在于conc的增加是否会使nomove的概率增加。

首先载入数据集并读取部分文件，为了观察两个变量之间关系，我们可以利cdplot函数来绘制条件密度图

In [104]:

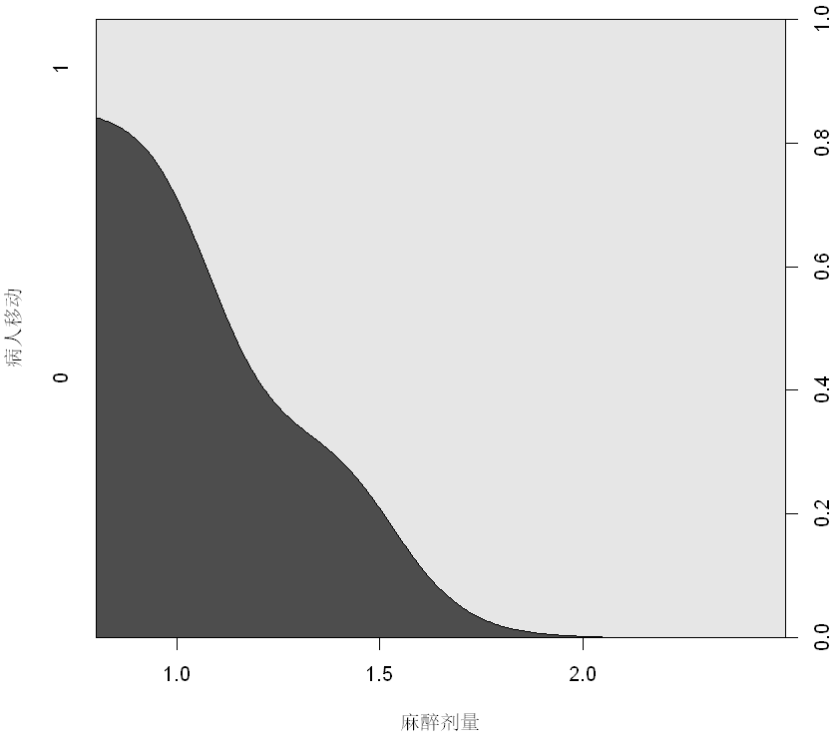
```
#install.packages("DAAG")
library(lattice)
library(DAAG)
head(anesthetic)
```

move	conc	logconc	nomove
0	1.0	0.0000000	1
1	1.2	0.1823216	0
0	1.4	0.3364722	1
1	1.4	0.3364722	0
1	1.2	0.1823216	0
0	2.5	0.9162907	1

In [103]:

```
cdplot(factor(nomove)~conc, data=anesthetic,  
        main='条件密度图', ylab='病人移动', xlab='麻醉剂量')
```

条件密度图



从图中可见，随着麻醉剂量加大，手术病人倾向于静止。下面利用logistic回归进行建模，得到intercept和conc的系数为-6.47和5.57，由此可见麻醉剂量超过1.16(6.47/5.57)时，病人静止概率超过50%。

In [105]:

```
anes1=glm(nomove~conc, family=binomial(link='logit'), data=anesthetic)

summary(anes1)
```

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

对模型做出预测结果

In [118]:

```
pre=predict(anes1, type='response')
```

将预测概率pre和实际结果放在一个数据框中

In [119]:

```
data=data.frame(prob=pre, obs=anesthetic$nomove)
```

将预测概率按照从低到高排序

In [120]:

```
data=data[order(data$prob),]  
  
n=nrow(data)  
  
tpr=fpr=rep(0, n)  
n  
tpr  
fpr
```

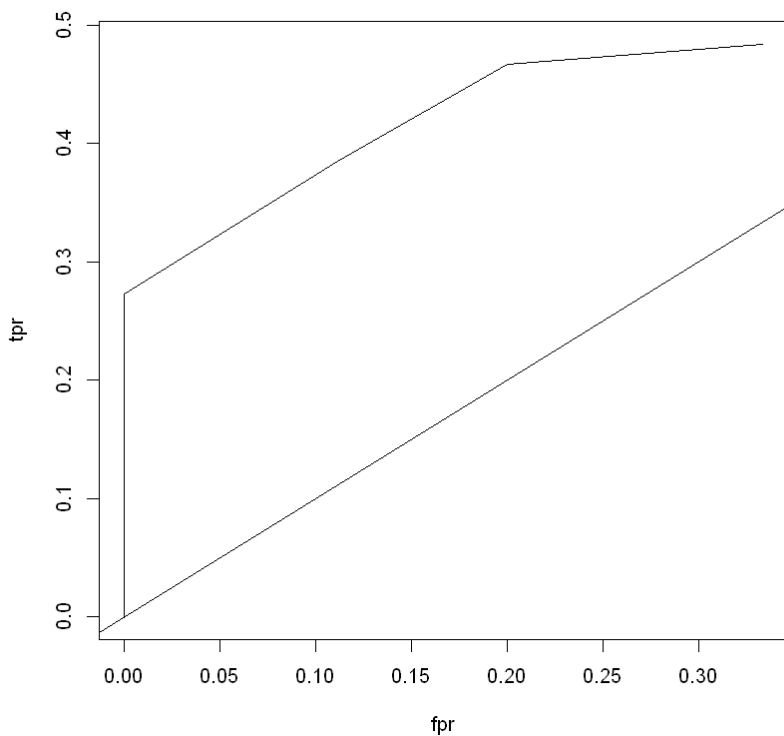
30

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0							
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0							

根据不同的临界值threshold来计算TPR和FPR，之后绘制成图

In [121]:

```
for (i in 1:n){  
  threshold=data$prob[i]  
  
  tp=sum(data$prob>threshold&data$obs==1)  
  
  fp=sum(data$prob>threshold&data$obs==0)  
  
  tn=sum(data$prob)  
  
  fn=sum(data$prob)  
  
  tpr[i]=tp/(tp+fn)  #真正率  
  
  fpr[i]=fp/(tn+fp)  #假正率  
  
}  
  
plot(fpr, tpr, type='l')  
  
abline(a=0, b=1)
```



R中也有专门绘制ROC曲线的包，如常见的ROCR包，它不仅可以用来画图，还能计算ROC曲线下面积AUC,以评价分类器的综合性能，该数值取0-1之间，越大越好。

In [124]:

```
install.packages("ROCR")  
library(ROCR)  
  
pred=prediction(pre, anesthetic$nomove)  
  
performance(pred, 'auc')@y.values  
  
perf=performance(pred, 'tpr', 'fpr')  
  
plot(perf)
```


Warning message:

"unable to access index for repository <http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5>:

无法打开URL' <http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES>' "

package 'ROCR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Zorg\AppData\Local\Temp\Rtmpvvh8YU\downloaded_packages

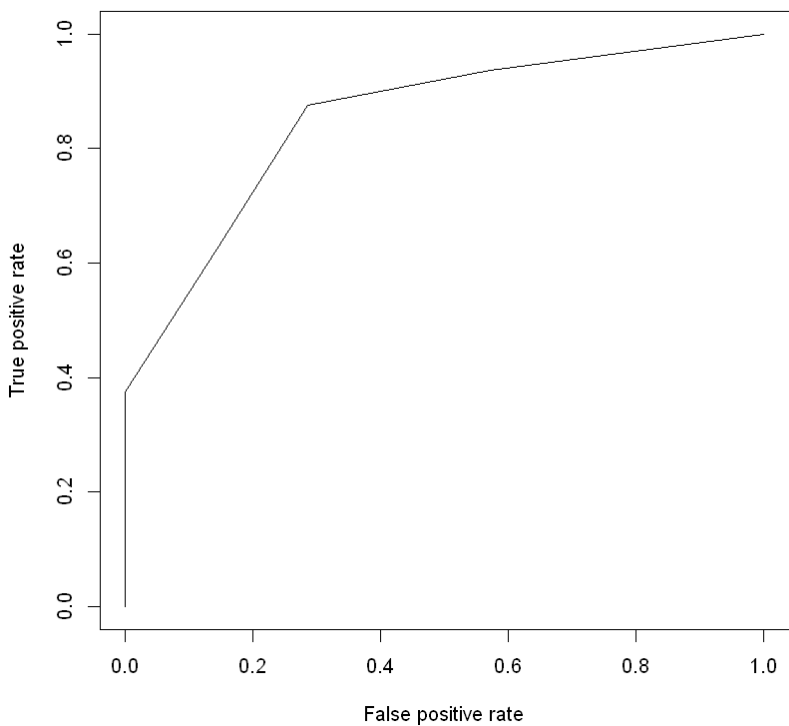
Loading required package: gplots

Attaching package: 'gplots'

The following object is masked from 'package:stats':

lowess

1. 0.852678571428571



还可以使用更加强大的pROC包，它可以方便的比较两个分类器，并且能自动标出最优临界点，图形看起来比较漂亮：

In [125]:

```
#install.packages("pROC")
```

```
library(pROC)
```

```
modelroc=roc(anesthetic$nomove, pre)
```

```
plot(modelroc, print.auc=TRUE, auc.polygon=TRUE,  
      grid=c(0.1, 0.2), grid.col=c("green", "red"),  
      max.auc.polygon=TRUE, auc.polygon.col="blue", print.thres=TRUE)
```

Warning message:

"unable to access index for repository <http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5>:

无法打开URL' <http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES>' "

package 'pROC' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

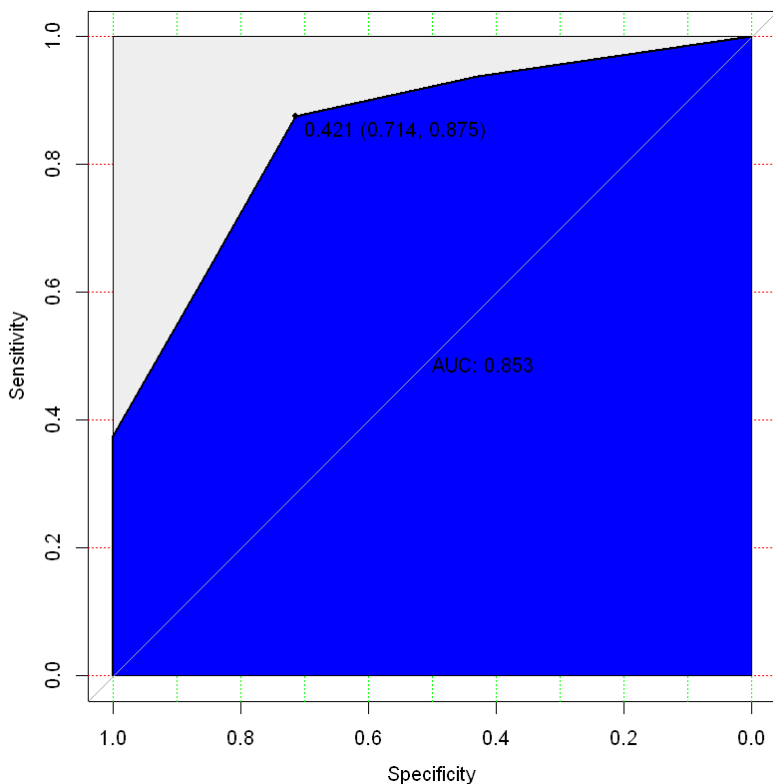
C:\Users\Zorg\AppData\Local\Temp\Rtmpwvh8YU\downloaded_packages

Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

The following objects are masked from 'package:stats':

cov, smooth, var



上面的方法是使用原始的0-1数据进行建模,即每一行数据均表示一个个体, 另一种是使用汇总数据进行建模, 先将原始数据按下面步骤进行汇总

In []:

```
anestot=aggregate(anesthetic[,c('move', 'nomove')],
                  by=list(conc=anesthetic$conc), FUN=sum)
```

In []:

```
anestot$conc=as.numeric(as.character(anestot$conc))

anestot$total=apply(anestot[,c('move', 'nomove')], 1, sum)

anestot$total
```

In []:

```
anestot$prop=anestot$nomove/anestot$total  
  
anestot$prop
```

对于汇总数据，有两种方法可以得到同样的结果，一种是将两种结果的向量合并做为因变量，如anes2模型。另一种是将比率做为因变量，总量做为权重进行建模，如anes3模型。这两种建模结果是一样的。

In []:

```
anes2=glm(cbind(nomove, move)~conc,  
          family=binomial(link='logit'), data=anestot)  
  
summary(anes2)
```

In []:

```
anes3=glm(prop~conc, family=binomial(link='logit'),  
          weights=total, data=anestot)
```

根据logistic模型，我们可以使用predict函数来预测结果，下面根据上述模型来绘图

In []:

```
x=seq(from=0, to=3, length.out=30)  
  
y=predict(anes1, data.frame(conc=x), type='response')  
  
plot(prop~conc, pch=16, col='red', data=anestot,  
      xlim=c(0.5, 3), main='Logistic回归曲线图', ylab='病人静止概率', xlab='麻醉剂量')  
  
lines(y~x, lty=2, col='blue')
```

In []:

多元选择模型

多分变量所取的离散值个数多于两个，如果各种结果之间没有自然顺序的话，称为无序变量。例如，当购买洗衣粉时，你可以在众多可供选择的洗衣粉品牌之间进行选择。如果各种结果之间有一个内在的自然的顺序，则为有序变量。例如，对债券等级的排序；老师给学生A、B、C、D和E五个等级的成绩。多分变量为因变量的模型称为多元选择模型(MultinomialModel)，其中又有有序选择模型(OrderedModel)、条件模型(ConditionalModel)、嵌套模型(NestedModel)等分类。本章主要介绍多元选择模型的条件模型、有序选择模型和嵌套模型。

有序选择模型

示例：一项研究为了了解影响本科生申请研究生的因素，向大三学生进行问卷调查，共收集了400份有效问卷。前10个观测值如下表。该数据共有四个变量，apply是有序变量，有三个水平，表示申请研究生的愿望程度，分别是unlikely, somewhatlikely, verylikely，并分别编码0,1,2；pared是0/1虚拟变量，表示父母是否至少有一个有研究生学位;public是0/1虚拟变量，1表示本科学校是公立的，0表示是私立的；gpa是学生的平均绩点。如何研究pared, public,gpa等因素对本科生申请研究生可能性的影响？

本科生申请研究生的部分数据

obs	apply	pared	public	gpa
1	Very likely	0	0	3.26
2	Somewhat likely	1	0	3.21
3	unlikely	1	1	3.94
4	Somewhat likely	0	0	2.81
5	Somewhat likely	0	0	2.53
6	unlikely	0	1	2.59
7	Somewhat likely	0	0	2.56
8	Somewhat likely	0	0	2.73
9	unlikely	0	0	3
10	Somewhat likely	1	0	3.5

(一) 有序选择模型

设定以下隐函数或指示函数模型：

$$Y_i^* = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i \quad i = 1, 2, \cdots, T$$

同样， Y_i^* 为无法观测到的连续型的隐变量（latent variable）。但是我们可以观测到的是

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ 1 & \text{if } 0 < Y_i^* \leq a_1 \\ 2 & \text{if } a_1 < Y_i^* \leq a_2 \\ \vdots & \\ M & \text{if } Y_i^* \geq a_{M-1} \end{cases}$$

为了表述简单，我们介绍M=2的情形为例，即 Y_i 有三种可能的选择：

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ 1 & \text{if } 0 < Y_i^* \leq a \\ 2 & \text{if } Y_i^* > a \end{cases}$$

式中 Y_i 的可能选择有M+1个。当M=1时，上式就等同于二项选择模型。

当 ε_i 是标准正态分布时，

$$\text{Prob}\left(Y_i = 0 \mid \mathbf{X}_i\right) = \int_{-\infty}^{-\mathbf{X}_i' \boldsymbol{\beta}} f\left(\varepsilon_i\right) d\varepsilon_i = \Phi\left(-\mathbf{X}_i' \boldsymbol{\beta}\right) = 1 - \Phi\left(\mathbf{X}_i' \boldsymbol{\beta}\right)$$

$$\begin{aligned} \text{Prob}\left(Y_i = 1 \mid \mathbf{X}_i\right) &= \int_{-\mathbf{X}_i' \boldsymbol{\beta}}^{a - \mathbf{X}_i' \boldsymbol{\beta}} f\left(\varepsilon_i\right) d\varepsilon_i = \Phi\left(a - \mathbf{X}_i' \boldsymbol{\beta}\right) - \Phi\left(-\mathbf{X}_i' \boldsymbol{\beta}\right) \\ &= \Phi\left(a - \mathbf{X}_i' \boldsymbol{\beta}\right) + \Phi\left(\mathbf{X}_i' \boldsymbol{\beta}\right) - 1 \end{aligned}$$

$$\begin{aligned} \text{Prob}\left(Y_i = 2 \mid \mathbf{X}_i\right) &= 1 - \text{Prob}\left(Y_i = 1 \mid \mathbf{X}_i\right) - \text{Prob}\left(Y_i = 0 \mid \mathbf{X}_i\right) \\ &= 1 - \Phi\left(a - \mathbf{X}_i' \boldsymbol{\beta}\right) \end{aligned}$$

这里称三个表达式为有序Probit模型（Ordered Probit Model）。

定义j=0,1,2为可能的结果， 设

$$Y_{i0} = \begin{cases} 1 & Y_i \text{ 等于 } 0, \text{ 发生的概率为 } F_{i0} \\ 0 & \text{其它, 发生的概率为 } 1 - F_{i0} \end{cases}$$

$$Y_{i1} = \begin{cases} 1 & Y_i \text{ 等于 } 1, \text{ 发生的概率为 } F_{i1} \\ 0 & \text{其它, 发生的概率为 } 1 - F_{i1} \end{cases}$$

$$Y_{i2} = \begin{cases} 1 & Y_i \text{ 等于 } 2, \text{ 发生的概率为 } F_{i2} \\ 0 & \text{其它, 发生的概率为 } 1 - F_{i2} \end{cases}$$

$$Y_{i0} = \begin{cases} 1 & Y_i \text{ 等于 } 0, \text{ 发生的概率为 } F_{i0} \\ 0 & \text{其它, 发生的概率为 } 1 - F_{i0} \end{cases}$$

其中, $F_{ij} = \text{Prob}(Y_i = j | X_i)$, $j = 0, 1, 2$, 则有序Probit和Logit模型的似然函数为:

$$\begin{aligned} L &= \prod_{i=1}^N \prod_{j=0}^2 F_{ij}^{Y_{ij}} \\ &= \prod_{i=1}^N F_{i0}^{Y_{i0}} F_{i1}^{Y_{i1}} F_{i2}^{Y_{i2}} \end{aligned}$$

对于有序probit和logit模型往往使用极大似然估计，即上式的对数达到最大时求得对应的参数。

有序Probit模型的边际效应为：

$$\frac{\partial \text{Prob} \left(Y_i=0 \mid \mathbf{X}_i \right)}{\partial \mathbf{X}_i} = -\phi \left(-\mathbf{X}_i' \boldsymbol{\beta} \right) \cdot \boldsymbol{\beta}$$

$$\frac{\partial \text{Prob} \left(Y_i=1 \mid \mathbf{X}_i \right)}{\partial \mathbf{X}_i} = \left[\phi \left(\mathbf{X}_i' \boldsymbol{\beta} \right) - \phi \left(a - \mathbf{X}_i' \boldsymbol{\beta} \right) \right] \cdot \boldsymbol{\beta}$$

$$\frac{\partial \text{Prob} \left(Y_i=2 \mid \mathbf{X}_i \right)}{\partial \mathbf{X}_i} = \phi \left(a - \mathbf{X}_i' \boldsymbol{\beta} \right) \cdot \boldsymbol{\beta}$$

同样，对有序Logit模型也可以进行类似的边际效应分析。

（二）、案例分析：本科生申请研究生的影响因素

以apply作为因变量，其他的三个变量作为自变量建立有序logit模型和有序probit模型。

In [45]:

```
#install.packages("foreign") #读入dta格式数据前需先安装foreign包
library(foreign)
dat <- read.dta("https://stats.idre.ucla.edu/stat/data/ologit.dta") #读入a
ta格式数据
head(dat)
```

apply	pared	public	gpa
very likely	0	0	3.26
somewhat likely	1	0	3.21
unlikely	1	1	3.94
somewhat likely	0	0	2.81
somewhat likely	0	0	2.53
unlikely	0	1	2.59

有序logit模型和有序probit模型可以用MASS包里的polr()函数；

In [37]:

```
library(MASS)
orm<-polr(~pared+public+gpa, data=dat, method="logistic") #logistic模型
summary(orm)
```

Re-fitting to get Hessian

Call:

```
polr(formula = apply ~ pared + public + gpa, data = dat, method = "logistic")
```

Coefficients:

	Value	Std. Error	t value
pared	1.04769	0.2658	3.9418
public	-0.05879	0.2979	-0.1974
gpa	0.61594	0.2606	2.3632

Intercepts:

	Value	Std. Error	t value
unlikely somewhat likely	2.2039	0.7795	2.8272
somewhat likely very likely	4.2994	0.8043	5.3453

Residual Deviance: 717.0249

AIC: 727.0249

In [42]:

```
orm_p<-polr(apply~pared+public+gpa, data=dat, method="probit") #probit模型  
summary(orm_p)
```

Re-fitting to get Hessian

Call:

```
polr(formula = apply ~ pared + public + gpa, data = dat, method = "probit")
```

Coefficients:

	Value	Std. Error	t value
pared	0.59811	0.1579	3.78881
public	0.01016	0.1728	0.05878
gpa	0.35815	0.1568	2.28479

Intercepts:

	Value	Std. Error	t value
unlikely somewhat likely	1.2968	0.4675	2.7738
somewhat likely very likely	2.5028	0.4766	5.2517

Residual Deviance: 717.4951

AIC: 727.4951

有序logit模型也可以用rms包中的lrm()函数拟合;

In [48]:

```
#install.packages("rms")
library(rms)
orm2<-lrm(apply~pared+public+gpa, data=dat)
orm2
```

Logistic Regression Model

```
lrm(formula = apply ~ pared + public + gpa, data = dat)
```

		Model Likelihood		Discriminatio	
n		Ratio Test		Indexes	
Rank Discrim.					
Indexes					
Obs	400	LR chi2	24.18	R2	0.07
0 C	0.605				
unlikely	220	d. f.	3	g	0.50
9 Dxy	0.210				
somewhat likely	140	Pr(> chi2)	<0.0001	gr	1.66
4 gamma	0.211				
very likely	40			gp	0.12
1 tau-a	0.119				
max deriv	3e-11			Brier	0.23
5					

	Coef	S.E.	Wald Z	Pr(> Z)
y>=somewhat likely	-2.2033	0.7795	-2.83	0.0047
y>=very likely	-4.2988	0.8043	-5.34	<0.0001
pared	1.0477	0.2658	3.94	<0.0001
public	-0.0587	0.2979	-0.20	0.8438
gpa	0.6157	0.2606	2.36	0.0182

在此例子中，对于有序logit模型，lrm()函数拟合的结果与polr()函数拟合的结果相同。并且lrm()函数还给出了似然比检验结果。

多元无序Logit模型

多元无序选择模型假设因变量是多于两类的分类变量，如品牌等。主要介绍多元无序Logit模型的原理和参数估计。

(一)、问题的提出

示例：收集了有关钓鱼模式选择的横截面数据，共有1182个观测值，6个变量，其中钓鱼模式(mode)有四种方式，即beach, pier, boat和charter。price.*表示每个人选择每种模式所付出的成本；catch.*表示每个人选择每种模式的抓到鱼的概率；income是月收入，数据形式见表。现在以钓鱼模式(mode)为因变量，四种选择项之间是没有顺序大小的，该如何研究其他因素对钓鱼模式选择的影响呢？

表:钓鱼模式选择的部分数据

ID	mode	price.beach	price.pier	price.boat	price.charter	catch.beach	catch.pier
1	charter	157.93	157.93	157.93	182.93	0.0678	0.0503
2	charter	15.114	15.114	10.534	34.534	0.1049	0.0451
3	boat	161.874	161.874	24.334	59.334	0.5333	0.4522
4	pier	15.134	15.134	55.93	84.93	0.0678	0.0789
5	boat	106.93	106.93	41.514	71.014	0.0678	0.0503
6	charter	192.474	192.474	28.934	63.934	0.5333	0.4522

(二)、多元无序logit模型

当 Y_i 的各种可能的选择是不考虑顺序的，且每个选择之间是相互独立时，可以用多元无序Logit模型。

考察随机效用模型

$$Y_{ij}^* = U_{ij} + \varepsilon_{ij}$$

$$Y_i = j \text{ 当且仅当 } Y_{ij}^* > Y_{il}^*, l \neq j$$

其中， Y_{ij}^* 可以理解为消费者*i*对商品*j*的效用，该效用在实际上往往是无法直接观测的。假设可供消费者选择的商品有*M*+ 1 个，即*j* = 0, 1, 2, ..., *M*。 U_{ij} 为效用的可观测部分，它可表示为解释变量*X*的线性组合； ε_{ij} 为效用的不可观测部分。

显然， $Y_{ij}^* = U_{ij} + \varepsilon_{ij}$ 是不可观测的。但是，如果消费者选择了第*j*种商品，即 $Y_i = j$, 则对该消费者来说，商品*j*的效用一定大于其他所有的商品，即 $Y_{ij}^* > Y_{il}^*, l \neq j$, 因此

$$\begin{aligned} F_{ij} &= Prob\left(Y_i = j | \mathbf{X}_i\right) = Prob\left(Y_{ij}^* > Y_{il}^*, \forall l \neq j | \mathbf{X}_i\right) \\ &= Prob\left(U_{ij} + \varepsilon_{ij} > U_{il} + \varepsilon_{il}, \forall l \neq j | \mathbf{X}_i\right) \\ &= Prob\left(U_{ij} - U_{il} + \varepsilon_{ij} > \varepsilon_{il}, \forall l \neq j | \mathbf{X}_i\right) \end{aligned}$$

假设式中 ε_{ij} 服从I型极端值分布（Type I Extreme-value Distribution），即

$$F\left(\varepsilon_{ij}\right)=\exp \left[-\exp \left(-\varepsilon_{ij}\right)\right], \text { 则可以推出, } Prob\left(Y_i=j \mid \mathbf{X}_i\right)=\frac{e^{U_{ij}}}{\sum_{l=0}^M e^{U_u}} j 0,1,2, \cdots M, \text { 称为多元无序Logit模型或者条件logit模型。}$$

多元无序Logit模型的估计一般也是采用最大似然法,R中估计多元无序logit模型使用mlogit包中的mlogit()函数

（三）、案例分析：关于钓鱼模式的选择

收集了有关钓鱼模式选择的横截面数据，共有1182个观测值，10个变量，其中钓鱼模式有四种方式，即beach，pier，boat和charter。

In [49]:

```
#install.packages("mlogit")
library(mlogit)
data("Fishing", package="mlogit")
dim(Fishing)
head(Fishing)
```

Attaching package: 'mlogit'

The following object is masked from 'package:rms':

lrtest

1182 10

mode	price.beach	price.pier	price.boat	price.charter	catch.beach
charter	157.930	157.930	157.930	182.930	0.0678
charter	15.114	15.114	10.534	34.534	0.1049
boat	161.874	161.874	24.334	59.334	0.5333
pier	15.134	15.134	55.930	84.930	0.0678
boat	106.930	106.930	41.514	71.014	0.0678
charter	192.474	192.474	28.934	63.934	0.5333

但在使用mlogit之前，需要把数据用mlogit.data（）转换成合适mlogit函数分析的格式。

In [50]:

```
Fish <-mlogit.data(Fishing, varying = c(2:9), shape = "wide", choice = "mode")
head(Fish) #转换后的数据前6行
```

	mode	income	alt	price	catch	chid
1.beach	FALSE	7083.332	beach	157.930	0.0678	1
1.boat	FALSE	7083.332	boat	157.930	0.2601	1
1.charter	TRUE	7083.332	charter	182.930	0.5391	1
1.pier	FALSE	7083.332	pier	157.930	0.0503	1
2.beach	FALSE	1250.000	beach	15.114	0.1049	2
2.boat	FALSE	1250.000	boat	10.534	0.1574	2

利用mlogit()以mode为因变量， price和catch为自变量建立多元选择无序logit模型

In [51]:

```
mlog_F=mlogit(mode ~ price + catch, data = Fish)
summary(mlog_F)
```

Call:

```
mlogit(formula = mode ~ price + catch, data = Fish, method
= "nr",
      print.level = 0)
```

Frequencies of alternatives:

```
beach    boat charter    pier
0.11337 0.35364 0.38240 0.15059
```

nr method

7 iterations, 0h:0m:0s

g' (-H) ^-lg = 6.22E-06

successive function values within tolerance limits

Coefficients :

	Estimate	Std. Error	z-value	Pr(> z)
boat:(intercept)	0.8713749	0.1140428	7.6408	2.154e-14 ***
charter:(intercept)	1.4988884	0.1329328	11.2755	< 2.2e-16 ***
pier:(intercept)	0.3070552	0.1145738	2.6800	0.0073627 **
price	-0.0247896	0.0017044	-14.5444	< 2.2e-16 ***
catch	0.3771689	0.1099707	3.4297	0.0006042 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1230.8

McFadden R^2: 0.17823

Likelihood ratio test : chisq = 533.88 (p.value = < 2.22e-16)

从上面的分析结果可以看出，该分析是以beach为基准，选择boat、charter和pier的截距项是不一样的，分别比beach高0.8713749、1.4988884和0.3070552。同时价格price对选择是负的影响，而捕获率catch对选择是有正的影响。McFadden R^2 为0.17823。LR检验的p值小于2.22e-16，说明该模型是统计显著的。

嵌套Logit模型

（一）、问题的提出：旅行交通方式选择

在条件Logit的设定中，假设随机效用模型中 $Y_i = j, j = 0, 1, 2, \dots, M$ 的各种可能的选择之间是相互独立的。但在实际中，可能出现这样的情况，即各种可能的选择之间可以分成若干的组，组与组之间是相互独立的，但组内的选择之间却是相互关联的。这时，则需要采用嵌套模型来估计各种选择的概率。

示例：Greene（2003）的教材Econometric Analysis（5thedition）研究了关于从悉尼到墨尔本的旅行交通方式选择的问题。市民出行时交通选择方式(mode)首先面临选择是公共交通还是私人交通方式，其中私人交通方式有air和car，公共交通方式有train和bus。该数据共有840个观测值。什么因素会影响市民出行时交通方式的选择呢？

表 旅行交通方式选择部分数据

ID	individual	mode	choice	wait	vcost	travel	gcost	income	size
1	1	air	no	69	59	100	70	35	1
2	1	train	no	34	31	372	71	35	1
3	1	bus	no	35	25	417	70	35	1
4	1	car	yes	0	10	180	30	35	1
5	2	air	no	64	58	68	68	30	2
6	2	train	no	44	31	354	84	30	2

比如，一个高中毕业生首先面临两种选择：不上大学和上大学。在上大学的选择中又存在着上公立学校和私立学样的选择。也就是说，他面临的是三种选择：不上大学、上公立大学或上私立大学。后两种选择与第一种选择之间是相互独立的，但是后两者之间却是相关联的。

又比如，在有关家庭医疗支出的调查中，问题如下：

- 1.倘若家里需要支付大笔医疗费用，您会选择缩减平时家庭消费支出吗？（若选“会”则继续第8题，否则跳至第9题）
A.会B.不会
- 2、您会选择缩减下列哪些支出以支付家庭大笔的医疗支出？
A.食品B.烟酒及用品C.居住D.交通通讯E.衣着F.家庭设备及维修支出G.娱乐教育支出

综合考虑1,2两个问题，实际上就是一个嵌套问题。对于这类问题，该如何建模分析？

（二）、嵌套logit模型原理

嵌套logit模型的原理，以M=2为例，即j=0，1，2三种选择时的情形。

考察随机效用模型

$$Y_{ij}^* = U_{ij} + \varepsilon_{ij}$$

$$Y_i = j \text{ 当且仅当 } Y_{ij}^* > Y_{il}^*, l \neq j$$

设j=0，1，2。假设 $\varepsilon_{i0}, \varepsilon_{i1}, \varepsilon_{i2}$ 可分为两组， ε_{i0} 一组， ε_{i1} 和 ε_{i2} 一组。两组之间相互独立，但 ε_{i1} 和 ε_{i2} 的相关系数为 $1 - \rho^2$ 。假设 ε_{i1} 和 ε_{i2} 的联合分布为以下II型极端值分布

$$F(\varepsilon_{i1}, \varepsilon_{i2}) = \exp \left\{ - \left[\exp \left(-\rho^{-1} \varepsilon_{i1} \right) + \exp \left(-\rho^{-1} \varepsilon_{i2} \right) \right]^\rho \right\}$$

ε_{i0} 仍为I型极端值分布：

$$F(\varepsilon_{i0}) = \exp\left[-\exp(-\varepsilon_{i0})\right]$$

此时的嵌套Logit模型为：

$$\begin{aligned} \text{Prob}(Y_i = 0 | \mathbf{X}_i) &= \text{Prob}(Y_{i0}^* > Y_{i1}^*, Y_{i0}^* > Y_{i2}^* | \mathbf{X}_i) \\ &= \frac{e^{U_{i0}}}{e^{U_{i0}} + \left(e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}\right)^\rho} \end{aligned}$$

$$\begin{aligned} \text{Prob}(Y_i = 1 | \mathbf{X}_i) &= \text{Prob}(Y_{i1}^* > Y_{i0}^*, Y_{i1}^* > Y_{i2}^* | \mathbf{X}_i) \\ &= \frac{e^{\rho^{-1}U_{i1}}}{e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}} \cdot \frac{\left(e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}\right)^\rho}{e^{U_{i0}} + \left(e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}\right)^\rho} \\ &= \text{Prob}(Y_i = 1 | Y_i \neq 0, \mathbf{X}_i) \cdot \text{Prob}(Y_i \neq 0 | \mathbf{X}_i) \end{aligned}$$

$$\begin{aligned} \text{Prob}(Y_i = 2 | \mathbf{X}_i) &= \text{Prob}(Y_{i2}^* > Y_{i0}^*, Y_{i2}^* > Y_{i1}^* | \mathbf{X}_i) \\ &= \frac{e^{\rho^{-1}U_{i2}}}{e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}} \cdot \frac{\left(e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}\right)^\rho}{e^{U_{i0}} + \left(e^{\rho^{-1}U_{i1}} + e^{\rho^{-1}U_{i2}}\right)^\rho} \\ &= \text{Prob}(Y_i = 2 | Y_i \neq 0, \mathbf{X}_i) \cdot \text{Prob}(Y_i \neq 0 | \mathbf{X}_i) \end{aligned}$$

(三)、案例分析：旅行交通方式选择

Greene (2003) 的教材Econometric Analysis (5th edition) 关于从悉尼到墨尔本的旅行交通方式选择的数据在R的AER包中有对应的数据TravelMode。wait是等待交通工具的时间；vcost是交通工具的成本；travel是旅行的时间；gcost是总的成本；income是家庭收入；size是人数规模。

In [53]:

```
#install.packages("AER")
library("AER")
data("TravelMode", package="AER")
data("TravelMode", package="AER")
dim(TravelMode)
```

```
Loading required package: car
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following objects are masked from 'package:rms':
```

```
Predict, vif
```

```
Loading required package: sandwich
```

840 9

在分析前，我们对数据进行一些转换，转换过程中使用with()函数进行，使用方法是with(data, expr, ...)，即对data进行表达式运算。转换过程和转换后的结果如下

In [54]:

```
TravelMode$avincome<-with(TravelMode, income*(mode=="air"))
TravelMode$time<-with(TravelMode, travel+wait)/60
TravelMode$timeair<-with(TravelMode, time*I(mode=="air"))
TravelMode$income<-with(TravelMode, income/10)
TravelMode$incomeother<-with(TravelMode, ifelse(mode%in%c('air','car'), income, 0))
head(TravelMode)
```

individual	mode	choice	wait	vcost	travel	gcost	income	size
1	air	no	69	59	100	70	3.5	1
1	train	no	34	31	372	71	3.5	1
1	bus	no	35	25	417	70	3.5	1
1	car	yes	0	10	180	30	3.5	1
2	air	no	64	58	68	68	3.0	2
2	train	no	44	31	354	84	3.0	2



然后再利用嵌套logit模型进行分析，在R里仍然使用mlogit()函数分析，但是在nested里需要进行设置。以choice为因变量，gcost,wait,incomeother为自变量建立嵌套logit模型，结果如下

In [55]:

```
nl<-mlogit(choice~gcost+wait+incomeother, TravelMode,  
shape='long', alt.var='mode',  
nests=list(public=c('train', 'bus'), other=c('car', 'air')))  
summary(nl)
```

```
Call:
mlogit(formula = choice ~ gcost + wait + incomeother, data
= TravelMode,
        nests = list(public = c("train", "bus"), other = c("ca
r",
                    "air")), shape = "long", alt.var = "mode")
```

Frequencies of alternatives:

	air	train	bus	car
	0.27619	0.30000	0.14286	0.28095

bfgs method
12 iterations, 0h:0m:0s
g' (-H)^-1g = 1.14E-07
gradient close to zero

Coefficients :

	Estimate	Std. Error	z-value	Pr(> z)
train:(intercept)	0.0056402	0.6363849	0.0089	0.9929285
bus:(intercept)	-0.7735521	0.8089799	-0.9562	0.3389677
car:(intercept)	-6.1536851	1.1783823	-5.2221	1.769e-07 *
**				
gcost	-0.0195488	0.0060331	-3.2402	0.0011943 *
*				
wait	-0.1064623	0.0205635	-5.1773	2.252e-07 *
**				
incomeother	0.4256608	0.1131906	3.7606	0.0001695 *
**				
iv:public	0.9694958	0.3047966	3.1808	0.0014687 *
*				
iv:other	1.7244022	0.5186970	3.3245	0.0008858 *
**				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Log-Likelihood: -188.43
McFadden R^2: 0.33594
Likelihood ratio test : chisq = 190.65 (p.value = < 2.22e-16)



从分析结果可以看出，以air作为基准，train,bus,car的截距项分别比air高0.0056402，-0.7735521，-6.1536851；gcost和wait对选择方式具有负的影响，而incomeother，iv.public和iv.other对选择方式具有正的影响。McFadden R^2 为0.33594。LR检验的p值小于2.22e-16，说明该模型是统计显著的。

计数模型和受限因变量模型

计数变量主要用于描述某一事件发生的次数，它仅取整数值。例如，每户家庭的子女数。因变量为计数变量的模型称为计数模型(CountModel)。

受限因变量 (limiteddependentvariable) 是指因变量的观测值是连续的，但是受到某种限制，其抽样并非完全随机的，得到的观测值并不完全反应因变量的真实情况。选择性样本 (selectivesample) 是受限因变量的主要形式，其样本观测值是在选择性限制的情况下抽取的。受限因变量常见的两类数据：截断 (truncation) 数据和审查 (Censoring) 数据。受限因变量模型主要包括截断模型(TruncatedModel)和审查模型(CensoredData)两类。这两类模型多应用在调查数据的分析当中。

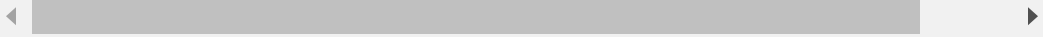
计数模型

(一)、问题的提出：轮船事故的计数数据模型

示例：表中给出了4个制造期，2个服务期的5种轮船的事故发生次数。表中的数据涉及4个制造期，2个服务期的五种类型的轮船的发生事故的次数数据。其中，TYPE表示轮船类型；TA，TB，TC，TD，TE是表示轮船类型的虚拟变量；T6064，T6569，T7074，T7579是制造期间虚拟变量；O6064，O7579是运营期间虚拟变量；Mon是服务量的测量；Acc是发生事故的次数。由于在75-79年制造的轮船不可能在60-64年期间运营，因此每一类型的轮船都有一个缺失数据。此外，由于原数据引用处未交待的原因，第五类型的轮船还有一个缺失数据，因此实际样本容量只有34。

obs	Type	TA	TB	TC	TD	TE	T6064	T6569	T7074	T7579	O6074	O7579	Miles
1	1	1	0	0	0	0	1	0	0	0	1	0	1
2	1	1	0	0	0	0	1	0	0	0	0	1	
3	1	1	0	0	0	0	0	1	0	0	1	0	10
4	1	1	0	0	0	0	0	1	0	0	0	1	10
5	1	1	0	0	0	0	0	0	1	0	1	0	15
6	1	1	0	0	0	0	0	0	1	0	0	1	33
7	1	1	0	0	0	0	0	0	0	1	1	0	↑
8	1	1	0	0	0	0	0	0	0	1	0	1	22
9	2	0	1	0	0	0	1	0	0	0	1	0	448
10	2	0	1	0	0	0	1	0	0	0	0	1	171
11	2	0	1	0	0	0	0	1	0	0	1	0	286
12	2	0	1	0	0	0	0	1	0	0	0	1	203
13	2	0	1	0	0	0	0	0	1	0	1	0	70
14	2	0	1	0	0	0	0	0	1	0	0	1	130
15	2	0	1	0	0	0	0	0	0	1	1	0	↑
16	2	0	1	0	0	0	0	0	0	1	0	1	71
17	3	0	0	1	0	0	1	0	0	0	1	0	11
18	3	0	0	1	0	0	1	0	0	0	0	1	5
19	3	0	0	1	0	0	0	1	0	0	1	0	7
20	3	0	0	1	0	0	0	1	0	0	0	1	6
21	3	0	0	1	0	0	0	0	1	0	1	0	7
22	3	0	0	1	0	0	0	0	1	0	0	1	19
23	3	0	0	1	0	0	0	0	0	1	1	0	↑
24	3	0	0	1	0	0	0	0	0	1	0	1	2
25	4	0	0	0	1	0	1	0	0	0	1	0	2
26	4	0	0	0	1	0	1	0	0	0	0	1	1
27	4	0	0	0	1	0	0	1	0	0	1	0	2
28	4	0	0	0	1	0	0	1	0	0	0	1	1

obs	Type	TA	TB	TC	TD	TE	T6064	T6569	T7074	T7579	O6074	O7579	Mi
29	4	0	0	0	1	0	0	0	1	0	1	0	3
30	4	0	0	0	1	0	0	0	1	0	0	1	12
31	4	0	0	0	1	0	0	0	0	1	1	0	1
32	4	0	0	0	1	0	0	0	0	1	0	1	20
33	5	0	0	0	0	1	0	0	0	1	0	1	
34	5	0	0	0	0	1	1	0	0	0	0	1	1
35	5	0	0	0	0	1	0	1	0	0	1	0	7
36	5	0	0	0	0	1	0	1	0	0	0	1	4
37	5	0	0	0	0	1	0	0	1	0	1	0	11
38	5	0	0	0	0	1	0	0	1	0	0	1	21
39	5	0	0	0	0	1	0	0	0	1	1	0	1
40	5	0	0	0	0	1	0	0	0	1	0	1	5



我们感兴趣的是轮船发生事故的次数以及其影响因素，轮船发生事故的次数是正整数，不同于普通回归的因变量是连续的实数，该如何分析？

计数数据模型的设定

当因变量为计数数据时，一般是以泊松分布来描述它的概率，此时，因变量Y的概率分布函数为：

$$\text{Prob}\left(Y = Y_i\right) = \frac{e^{-u_i} u_i^{Y_i}}{Y_i!}$$

其中， u_i 一般设定为： $u_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ 或 $\log u_i = \mathbf{X}_i'\boldsymbol{\beta}$

其中 \mathbf{x}_i 是包括常数项在内的k个解释变量。可以证明，

$$E\left(Y_i | \mathbf{X}_i\right)=\operatorname{Var}\left(Y_i | \mathbf{X}_i\right)=u_i=e^{\mathbf{X}_i \boldsymbol{\beta}}$$

对 u_i 的设定可以保证 Y_i 的预测值是非负的。

$$\text{可得}\frac{\partial \log E\left(Y_i | \mathbf{X}_i\right)}{\partial X_i}=\boldsymbol{\beta}$$

可以认为 $\boldsymbol{\beta}$ 是解释变量的变动对 Y 的变动率的平均影响。例如，假设第 j 个解释变量的系数是 β_j ，则它表明在其他变量不变的情况下， X_j 每增加一个单位，则 Y 发生的次数将平均增加 β_j 。如果第 j 个解释变量是虚拟变量的话，则它从0到1发生变化时，它 Y 发生的次数将平均增加 $100 \cdot \left[\exp\left(\beta_j\right)-1\right] \%$

在 u_i 的设定中加入随机扰动项，得到计数数据模型

$$Y_i=e^{\mathbf{X}_i^{\prime} \boldsymbol{\beta}+\varepsilon_i}$$

计数数据模型的估计

对模型进行参数的估计，虽然可以对对数方程直接进行OLS估计，但是会丢失那些因变量取值为0的数据，因为0是无法取对数的。另一种选择是直接采用非线性的叠代方法，但往往无法克服模型中的异方差特性，因此更多的是采用最大似然法进行估计。其对数似然函数为：

$$\ln L=\sum_{i=1}^N\left[-\exp \left(\mathbf{X}_i^{\prime} \boldsymbol{\beta}\right)+Y_i \cdot \mathbf{X}_i^{\prime} \boldsymbol{\beta}-\ln Y_i !\right]$$

将参数的估计结果代入 $\operatorname{Prob}\left(Y=Y_i\right)=\frac{e^{-u_i} u_i^{Y_i}}{Y_i !}$ ，可以得到在给定解释变量 X 的值的条件下 Y 的各种取值的概率的估计： $\operatorname{Prob}\left(Y=Y_i\right)=\frac{e^{-\mathbf{X}_i^{\prime} \hat{\boldsymbol{\beta}}}\left(\mathbf{X}_i^{\prime} \hat{\boldsymbol{\beta}}\right)^{Y_i}}{Y_i !}, Y_i=0,1,2, \cdots$

R里泊松回归的估计也用`glm()`,只是对应的`family`改为`poission()`即可。

对示例轮船事故进行数据分析。

In [56]:

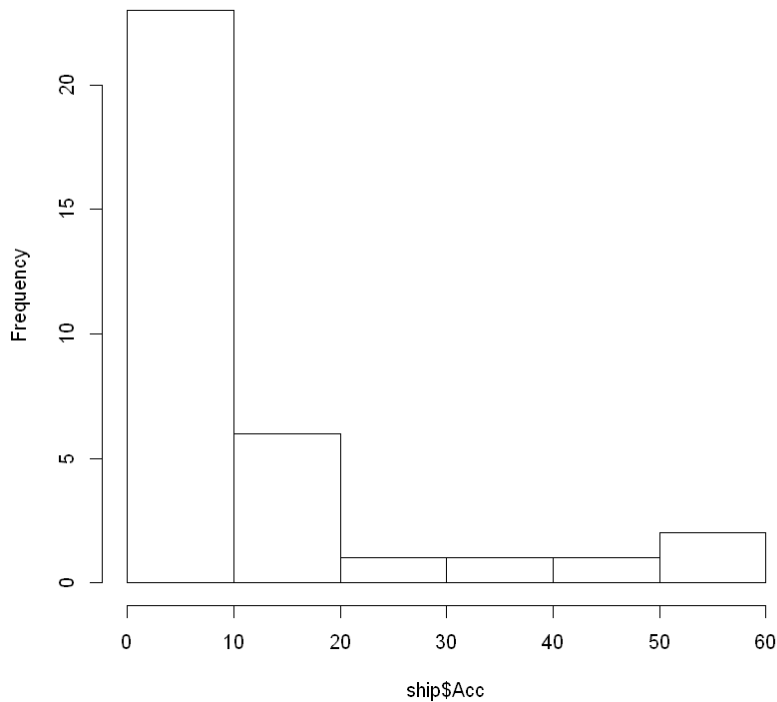
```
ship=read.table(file="./data/ship.txt",header=T) #读入数据
summary(ship[,-1])
hist(ship$Acc)
```


Type		TA		TB		TC	
TD							
Min. :0.0	:1	Min. :0.0		Min. :0.0		Min. :0.0	Min.
1st Qu. :0.0	:2	1st Qu. :0.0		1st Qu. :0.0		1st Qu. :0.0	1st
Median :0.0	:3	Median :0.0		Median :0.0		Median :0.0	Medi
Mean :0.2	:3	Mean :0.2		Mean :0.2		Mean :0.2	Mean
3rd Qu. :0.0	:4	3rd Qu. :0.0		3rd Qu. :0.0		3rd Qu. :0.0	3rd
Max. :1.0	:5	Max. :1.0		Max. :1.0		Max. :1.0	Max.

TE		T6064		T6569		T7074	
T7579							
Min.	:0.0	Min.	:0.000	Min.	:0.00	Min.	:0.00
Min.	:0.000						
1st Qu.	:0.0	1st Qu.	:0.000	1st Qu.	:0.00	1st Qu.	:0.00
1st Qu.	:0.000						
Median	:0.0	Median	:0.000	Median	:0.00	Median	:0.00
Median	:0.000						
Mean	:0.2	Mean	:0.225	Mean	:0.25	Mean	:0.25
Mean	:0.275						
3rd Qu.	:0.0	3rd Qu.	:0.000	3rd Qu.	:0.25	3rd Qu.	:0.25
3rd Qu.	:1.000						
Max.	:1.0	Max.	:1.000	Max.	:1.00	Max.	:1.00
Max.	:1.000						

06074		07579		Mon		Acc	
Min.	:0.000	Min.	:0.000	Min.	: 45	Min.	:
:0.00							
1st Qu.	:0.000	1st Qu.	:0.000	1st Qu.	: 371	1st Qu.	:
:1.00							
Median	:0.000	Median	:1.000	Median	: 1095	Median	:
:4.00							
Mean	:0.475	Mean	:0.525	Mean	: 4811	Mean	:1
:0.47							
3rd Qu.	:1.000	3rd Qu.	:1.000	3rd Qu.	: 2223	3rd Qu.	:1
:1.75							
Max.	:1.000	Max.	:1.000	Max.	:44882	Max.	:5
:8.00							
				NA's	:6	NA's	:6

Histogram of ship\$Acc



为了考察轮船的类型、制造期间、服务期间、服务量对事故发生次数的影响，估计以下泊松分布计数数据模型：

$$E(Acc|X) = \exp\left(\beta_1 + \beta_2 \cdot TB + \beta_3 \cdot TC + \beta_4 \cdot TD + \beta_5 \cdot TE + \beta_6 \cdot T6569 + \beta_7 \cdot T7074 + \beta_8 \right.$$

其中，TA，TB，TC，TD，TE是表示轮船类型的虚拟变量；T6064，T6569，T7074，T7579是制造期间虚拟变量；O6064，O7579是运营期间虚拟变量；Mon是服务量的测量；Acc是发生事故的次数。

In [57]:

```
fit<-glm(Acc~TB+TC+TC+TD+TE+T6569+T7074+T7579+07579+log(Mon),
         family=poisson(),data=ship)
summary(fit)
```

Call:

```
glm(formula = Acc ~ TB + TC + TC + TD + TE + T6569 + T7074
+
      T7579 + 07579 + log(Mon), family = poisson(), data = sh
ip)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6569	-0.8872	-0.4843	0.4789	2.7436

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.6185	0.8737	-6.431	1.27e-10 ***
TB	-0.3586	0.2697	-1.330	0.18360
TC	-0.7621	0.3383	-2.253	0.02426 *
TD	-0.1314	0.2970	-0.442	0.65820
TE	0.2697	0.2419	1.115	0.26492
T6569	0.6618	0.1539	4.301	1.70e-05 ***
T7074	0.7602	0.1781	4.268	1.97e-05 ***
T7579	0.3615	0.2473	1.462	0.14379
07579	0.3699	0.1182	3.129	0.00175 **
log(Mon)	0.9062	0.1017	8.906	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 614.539 on 33 degrees of freedom
Residual deviance: 38.132 on 24 degrees of freedom
(6 observations deleted due to missingness)
AIC: 156

Number of Fisher Scoring iterations: 5

在泊松回归中，因变量以条件均值的对数形式来建模。比如log(Mon)增加一个单位时，轮船事故发生次数的对数平均值将增加0.9062。截距项表示当所有自变量都为0时，轮船事故发生次数的对数均值。通常在因变量的初始尺度（轮船事故发生数而非发生数的对数）上解释回归系数比较容易。因此，经常需要对回归系数进行指数化。

首先利用coef()提取回归模型的系数

In [58]:

```
coef(fit)
```

```
(Intercept)
-5.61852108464911
TB
-0.358599482485301
TC
-0.762129479033722
TD
-0.131403212308766
TE
0.269666556710552
T6569
0.661819940451711
T7074
0.760204940204911
T7579
0.361462115929213
O7579
0.369862727188898
log(Mon)
0.906170175346311
```

然后把所有回归系数指数化，由于回归系数的小数位较多，如果我们只需要保留3位小数即可，我们用round()函数保留3位小数：

In [59]:

```
round(exp(coef(fit)), 3)
```

(Intercept)

0.004

TB

0.699

TC

0.467

TD

0.877

TE

1.31

T6569

1.938

T7074

2.139

T7579

1.435

O7579

1.448

log(Mon)

2.475

可以看出，当 $\log(\text{Mon})$ 增加一个单位时，轮船事故发生次数将增加2.475次。

受限因变量模型

（一）、截断模型问题的提出

截断数据即不能从全部个体，而只能从一部分个体中随机抽取因变量的样本观测值。比如，我们对农村居民在过去一年中的家庭医疗支出进行随机调查，在所调查的8000户家庭里，其中有2000户在一年内有发生医疗支出，而剩余的6000户没有医疗支出，如果仅以2000户的样本进行分析，显然没有医疗支出的6000户被截断了。

截断模型原理

设以下隐变量（latent variable）模型

$$Y_i^* = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i \quad i = 1, 2, \cdots, T$$

其中， Y_i^* 是隐变量，不可直接观测。对于截断数据 Y_i ,

$$Y_i = Y_i^* \text{ 当且仅当 } Y_i^* > 0$$

Y_i 是可以直接观测到的。因此，对 Y_i 来说，

$$\begin{aligned} Y_i &= E\left(Y_i | \mathbf{X}_i, Y_i > 0\right) + \varepsilon_i \\ &= E\left(Y_i^* | \mathbf{X}_i, Y_i^* > 0\right) + \varepsilon_i \end{aligned}$$

其中，

$$\begin{aligned} &E\left(Y_i^* | \mathbf{X}_i, Y_i^* > 0\right) \\ &= E\left(\mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i | \mathbf{X}_i, \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}\right) \\ &= \mathbf{X}_i' \boldsymbol{\beta} + E\left(\varepsilon_i | \mathbf{X}_i, \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}\right) \\ &= \mathbf{X}_i' \boldsymbol{\beta} + \int_{-\mathbf{X}_i' \boldsymbol{\beta}}^{\infty} \varepsilon_i f\left(\varepsilon_i | \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}\right) d\varepsilon_i \end{aligned}$$

其中 $f\left(\varepsilon_i | \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}\right)$ 是 ε_i 的条件概率密度函数。因为

$$\int_{-\mathbf{X}_i' \boldsymbol{\beta}}^{\infty} \varepsilon_i f\left(\varepsilon_i | \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}\right) d\varepsilon_i = \lambda\left(\mathbf{X}_i' \boldsymbol{\beta}\right) \neq \int_{-\infty}^{+\infty} \varepsilon_i f\left(\varepsilon_i\right) d\varepsilon_i = E\left(\varepsilon_i\right) = 0$$

$$\text{所以}E\Big(Y_i^* \mid \mathbf{X}_i, Y_i^* > 0\Big) = \mathbf{X}_i'\boldsymbol{\beta} + \lambda\Big(\mathbf{X}_i'\boldsymbol{\beta}\Big)$$

代入

$$\begin{aligned} Y_i &= \mathbf{X}_i'\boldsymbol{\beta} + \lambda\Big(\mathbf{X}_i'\boldsymbol{\beta}\Big) + \varepsilon_i \\ &= \mathbf{X}_i'\boldsymbol{\beta} + V_i \end{aligned}$$

$$\text{其中, } V_i = \lambda\Big(\mathbf{X}_i'\boldsymbol{\beta}\Big) + \varepsilon_i$$

如果直接将Y对X进行OLS回归，由于 $E\Big(V_i\Big) = E\Big[\lambda\Big(\mathbf{X}_i'\boldsymbol{\beta}\Big)\Big] \neq 0$ ，所以OLS估计量 $\hat{\beta}_{OLS}$ 不是一致估计量。

当 $\varepsilon_i > -\mathbf{X}_i'\boldsymbol{\beta}$ 时， ε_i 的条件概率密度函数

$$\begin{aligned} &f\Big(\varepsilon_i \mid \varepsilon_i > -\mathbf{X}_i'\boldsymbol{\beta}\Big) \\ &= \frac{f\Big(\varepsilon_i\Big)}{\text{Prob}\Big(\varepsilon_i > -\mathbf{X}_i'\boldsymbol{\beta}\Big)} \\ &= \frac{f\Big(\varepsilon_i\Big)}{\int_{-\infty}^{\infty} \mathbf{X}_i'\boldsymbol{\beta} f\Big(\varepsilon_i\Big) d\varepsilon_i} \end{aligned}$$

因此:

$$\begin{aligned} \lambda\Big(\mathbf{X}_i'\boldsymbol{\beta}\Big) &= \int_{-\mathbf{X}_i'\boldsymbol{\beta}}^{+\infty} \varepsilon_i f\Big(\varepsilon_i \mid \varepsilon_i > -\mathbf{X}_i'\boldsymbol{\beta}\Big) d\varepsilon_i \\ &= \frac{\int_{-\mathbf{X}_i'\boldsymbol{\beta}}^{+\infty} \varepsilon_i f\Big(\varepsilon_i\Big) d\varepsilon_i}{\int_{-\infty}^{\infty} \mathbf{X}_i'\boldsymbol{\beta} f\Big(\varepsilon_i\Big) d\varepsilon_i} \end{aligned}$$

当假设 ε_i 服从均值为零，方差为 σ^2 的正态分布时，

$$\lambda\left(\mathbf{X}_i'\boldsymbol{\beta}\right)=\sigma\cdot\frac{\phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)}$$

$$\frac{\phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)}=\lambda_i$$

λ_i 称为逆Mills比率(Inverse Mills Ratio)。截断数据Y的实际的方程

$$Y_i=\mathbf{X}_i'\boldsymbol{\beta}+\sigma\cdot\lambda_i+\varepsilon_i$$

审查模型问题的提出

例如，在一次关于住院人数的调查中，其中一个问题：

- 1、在过去一年中，您家里人住院的次数是有几次？
A. 0次B.1次C.2次D.3次E.4次F.5次及以上

该问题中，家里人住院的次数应该是一个计数数据，如果以该变量作为因变量可以使用poission回归，但是此处特殊的地方是当住院次数在5次以上的时候被归并（censor）在5次，我们观察不到6次或6次以上的值，这就是一种审查数据（censoreddata）。该如何分析这样的数据？

示例：Fair(1978)研究了婚外性行为的问题。该数据是从大约2000份回收的电子问卷中抽取的601初婚且未离婚的样本对象。有关初始数据和Fair对该问题研究的论文可以从互联网：<http://fairmodel.econ.yale.edu/rayfair/worksd.htm>上下载。其中该数据的前10个观测值如下：<http://fairmodel.econ.yale.edu/rayfair/worksd.htm>上下载。其中该数据的前10个观测值如下：)

ID	affairs	gender	age	Yearsmarried	children	religiousness	education	occupati
4	0	male	37	10	no	3	18	
5	0	female	27	4	no	4	14	
11	0	female	32	15	yes	1	12	
16	0	male	57	15	yes	5	18	
23	0	male	22	0.75	no	2	17	
29	0	female	32	1.5	no	2	17	
44	0	female	22	0.75	no	2	12	
45	0	male	57	15	yes	2	14	
47	0	female	32	15	yes	4	16	
49	0	male	22	1.5	no	4	14	

该如何建模分析哪些因素会影响婚外性行为的次数？

审查模型原理

对隐变量模型：

$$Y_i^* = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i \quad i = 1, 2, \dots, T$$

其中， Y_i^* 是隐变量，如果实际获取的是审查数据

$$Y_i = \begin{cases} Y_i^* & \text{当 } Y_i^* > 0 \\ 0 & \text{当 } Y_i^* \leq 0 \end{cases}$$

即只能获得 Y_i^* 大于0时的数据，当 Y_i^* 小于0时，只能得到观测值0。因此，对 Y_i 来说，

$$\begin{aligned} Y_i &= E(Y_i | \mathbf{X}_i, Y_i \geq 0) + \varepsilon_i \\ &= \text{Prob}(Y_i = 0) \cdot 0 + \text{Prob}(Y_i > 0) \cdot E(Y_i | \mathbf{X}_i, Y_i > 0) + \varepsilon_i \\ &= \text{Prob}(Y_i > 0) \cdot E(Y_i^* | \mathbf{X}_i, Y_i^* > 0) + \varepsilon_i \\ &= \text{Prob}(Y_i > 0) \cdot [\mathbf{X}_i' \boldsymbol{\beta} + E(\varepsilon_i | \mathbf{X}_i, \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta})] + \varepsilon_i \end{aligned}$$

由于公式可以简化

$$E(\varepsilon_i | \mathbf{X}_i, \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}) = \lambda(\mathbf{X}_i' \boldsymbol{\beta}) = \sigma \cdot \frac{\phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}$$

因此

$$Y_i = \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \cdot \mathbf{X}_i' \boldsymbol{\beta} + \sigma \cdot \phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) + \varepsilon_i$$

即审查数据的实际模型。

最大似然估计

一、截断模型的似然函数

对于截断数据来说，当 $Y_i > 0$ 时，由 $Y_i = Y_i^* = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i > 0$ 可以推出 $\varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}$

$$f(\varepsilon_i | \varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta}) = \frac{f(\varepsilon_i)}{\text{Prob}(\varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta})} = \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}$$

截断模型的似然函数为

$$L = \prod_{i=1}^N \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}$$

二、审查模型的似然函数

对于审查数据来说，可以把数据分成两组： $N_1 = \{i | Y_i = 0\}$ 和 $N_2 = \{i | Y_i > 0\}$

因此，

$$\text{Prob}(Y_i = 0) = \text{Prob}(Y_i^* \leq 0) = \text{Prob}(\varepsilon_i \leq -\mathbf{X}_i' \boldsymbol{\beta}) = \Phi\left(-\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) = 1 - \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)$$

$$f(Y_i | Y_i > 0) = \frac{f(Y_i)}{\text{Pr ob}(Y_i > 0)} = \frac{f(\varepsilon_i)}{\text{Pr ob}(Y_i > 0)} = \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)}$$

审查模型的似然函数为：

$$\begin{aligned} L &= \prod_{i \in N_1} \left[1 - \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \right] \prod_{i \in N_2} \left[\text{Prob}(Y_i > 0) \cdot f(Y_i | Y_i > 0) \right] \\ &= \prod_{i \in N_1} \left[1 - \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \right] \prod_{i \in N_2} \left[\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \right] \end{aligned}$$

三、Heckman二阶段估计

对于审查模型来说，还可以用Heckman两阶段法进行参数的估计。第一阶段实际上是借用Probit模型把参数 β/σ 先估计出来，第二步再用OLS法对审查数据中大于临界值的部分用截断模型进行估计。具体思路如下。对于审查数据，设定以下虚拟变量

$$d_i = \begin{cases} 1, & Y_i > 0 \\ 0, & Y_i = 0 \end{cases}$$

当 $Y_i^* = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i$ 中的扰动项 $\varepsilon_i \sim N(0, \sigma^2)$ 时，

$$\begin{aligned} \text{Prob}(d_i = 1 | \mathbf{X}_i) &= \text{Prob}(Y_i > 0 | \mathbf{X}_i) = \text{Prob}(Y_i^* > 0 | \mathbf{X}_i) \\ &= \text{Prob}(\varepsilon_i > -\mathbf{X}_i' \boldsymbol{\beta} | \mathbf{X}_i) = \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

$$\begin{aligned}\text{Prob}\left(d_i = 0 \mid \mathbf{X}_i\right) &= \text{Prob}\left(Y_i^* \leq 0 \mid \mathbf{X}_i\right) \\ &= 1 - \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)\end{aligned}$$

第一阶段, 令 $\alpha = \beta/\sigma$, 估计Probit模型, 其似然函数为:

$$L = \prod_{i=1}^N \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)^{d_i} \left(1 - \Phi\left(\frac{\mathbf{X}_i' \boldsymbol{\beta}}{\sigma}\right)\right)^{1-d_i}$$

通过对Probit模型的MLE估计, 可以得到 α 的一致估计量 $\hat{\alpha}$

第二阶段主要集中于 $Y_i > 0$ 的数据, 即主要估计截断模型, 但是, 其中的 λ_i 用

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{X}_i' \hat{\alpha}\right)}{\Phi\left(\mathbf{X}_i' \hat{\alpha}\right)}.$$

也就是说, 这一阶段是对模型 $Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \sigma \cdot \hat{\lambda}_i + \varepsilon_i$ 进行OLS估计, 从而得到 β 和 σ 的一致估计量。

示例分析: AER包里有Affairs数据集。建立Tobit模型可以用使用censReg包里的censReg(),使用方法和运行结果如下:

In [60]:

```
#install.packages("censReg")
data("Affairs", package="AER")
library(censReg)
estResult<-censReg(affairs~age+yearsmarried+religiousness+
occupation+rating, data=Affairs)
summary(estResult)
```

Please cite the 'censReg' package as:

Henningsen, Arne (2017). censReg: Censored Regression (Tobit) Models. R package version 0.5. <http://CRAN.R-Project.org/package=censReg>.

If you have questions, suggestions, or comments regarding the 'censReg' package, please use a forum or 'tracker' at the R-Forge site of the 'sampleSelection' project:

<https://r-forge.r-project.org/projects/sampleselection/>

Call:

```
censReg(formula = affairs ~ age + yearsmarried + religiousness +
         occupation + rating, data = Affairs)
```

Observations:

Total	Left-censored	Uncensored	Right-censored
601	451	150	0

Coefficients:

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	8.17420	2.74145	2.982	0.00287 **
age	-0.17933	0.07909	-2.267	0.02337 *
yearsmarried	0.55414	0.13452	4.119	3.80e-05 ***
religiousness	-1.68622	0.40375	-4.176	2.96e-05 ***
occupation	0.32605	0.25442	1.282	0.20001
rating	-2.28497	0.40783	-5.603	2.11e-08 ***
logSigma	2.10986	0.06710	31.444	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 7 iterations

Return code 1: gradient close to zero

Log-likelihood: -705.5762 on 7 Df

Tobit模型也可以用AER包里的tobit(), 使用方法和运行结果如下:

In [61]:

```
library(AER)
fm.tobit<-tobit(affairs~age+yearsmarried+religiousness+
                occupation+rating,data=Affairs)
summary(fm.tobit)
```

Call:
tobit(formula = affairs ~ age + yearsmarried + religiousness +
 occupation + rating, data = Affairs)

Observations:
Total Left-censored Uncensored Right-censored
601 451 150

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.17420 2.74145 2.982 0.00287 **
age -0.17933 0.07909 -2.267 0.02337 *
yearsmarried 0.55414 0.13452 4.119 3.80e-05 ***
religiousness -1.68622 0.40375 -4.176 2.96e-05 ***
occupation 0.32605 0.25442 1.282 0.20001
rating -2.28497 0.40783 -5.603 2.11e-08 ***
Log(scale) 2.10986 0.06710 31.444 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale: 8.247

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -705.6 on 7 Df
Wald-statistic: 67.71 on 5 Df, p-value: 3.0718e-13

分位数模型

传统的线性回归模型描述了因变量的条件均值分布与自变量X的关系，为了和分位数回归相区别，因此把传统的回归又称为均值回归（MeanRegression）。其中，OLS是估计回归系数的最基本方法。如果模型的随机误差项来自均值为零、方差相同的分布，那么回归系数的OLS估计为最佳线性无偏估计（BLUE）；如果随机误差项是正态分布，那么回归系数的OLS估计与MLE估计一致，均为最小方差无偏估计（MVUE）。此时它具有无偏性、有效性等优良性质。

但实际中，假设不能够满足时，为了弥补普通最小二乘法（OLS）在回归分析中的缺陷，1818年Laplace提出了中位数回归（MedianRegression），利用最小绝对偏差估计（Leastabsolutedeviance,LAD）。在此基础上，1978年Koenker和Bassett把中位数回归推广到了一般的分位数回归（QuantileRegression）上。分位数回归是估计一组回归变量X与被解释变量Y的分位数之间关系的建模方法。主要介绍基本的分位数回归及其应用。

问题的提出

示例：恩格尔定律

德国统计学家恩格尔(Engel)使用收集的235个比利时家庭的收入与食物支出数据得出其著名的恩格尔定律其著名的恩格尔定律:收入越高的家庭将其收入用于食物支出的比例越低。

No.	收入	消费
1	420.1577	255.8394
2	541.4117	310.9587
3	901.1575	485.68
4	639.0802	402.9974
5	750.8756	495.5608
6	945.7989	633.7978
7	829.3979	630.7566
8	979.1648	700.4409


```
In [ ]:
```

```
#install.packages("quantreg")
library(quantreg)
data(engel)
attach(engel)
hist(foodexp)
curve(density(x~foodexp), add=T)
plot(income, foodexp, xlab="Household Income", ylab="Food Expenditure",
      type="n", cex=.5)
points(income, foodexp, cex=.5, col="blue")
```

从家庭消费支出的核密度函数图可以看出消费支出不符合正态性假设，是一个右偏的分布。另外从收入消费散点图来看，消费支出和收入之间存在着异方差，即随着收入的增加，消费之间的差异在扩大。对于这样的数据，如果直接利用传统的均值回归方法会有问题，此处利用分位数回归是比较合适的

总体分位数和总体中位数

定义：对于一个连续随机变量 y ，其总体第 τ 分位数 $y_{(\tau)}$ 的定义是： y 小于等于 $y_{(\tau)}$ 的概率是 τ ，即

$$\tau = P(y \leq y_{(\tau)}) = F(y_{(\tau)})$$

其中 $P()$ 表示概率， $F(y_{(\tau)})$ 表示 y 的累积分布函数(cdf)。

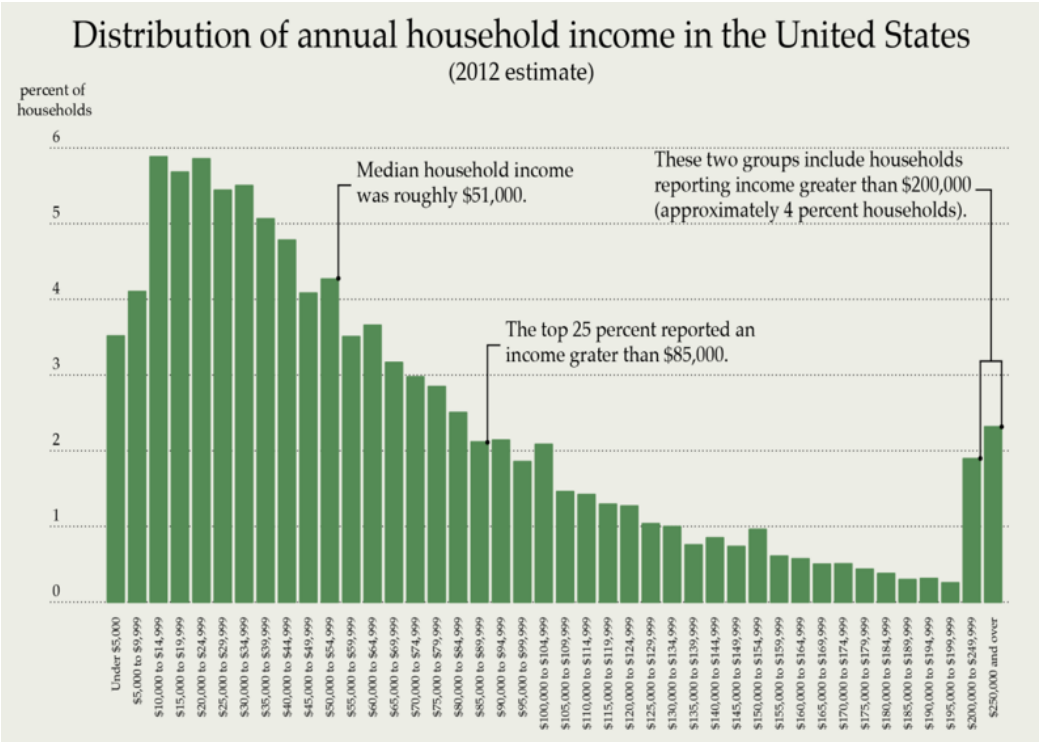
比如 $y_{(0.25)} = 3$ ，则意味着 $y \leq 3$ 的累积概率是0.25，即 $P(y \leq 3) = 0.25$ ，并且，

$$y_{(\tau)} = F^{-1}(y_{(\tau)}), \text{即} F(y_{(\tau)}) \text{的反函数是} y_{(\tau)}$$

当 $\tau = 0.5$ 时， $y_{(\tau)}$ 是 y 的中位数。当 $\tau = 0.75$ 时， $y_{(\tau)}$ 是 y 的3/4分位数。当 $\tau = 0.25$ 时， $y_{(\tau)}$ 是 y 的1/4分位数。若 y 服从标准正态分布，则

$$y_{(0.5)} = 0, \quad y_{(0.95)} = 1.645, \quad y_{(0.975)} = 1.96$$

另外，如果随机变量 分布是对称的，那么其均值与中位数是相同的。当其中位数小于均值时，分布是右偏的。反之，分布是左偏的。一般来讲，工资的分布是右偏的，所以如果单纯以平均工资来反映工资的话，这是很不恰当的，因此美国等一些国家除了公布平均工资外，还会同时公布工资的中位数和1/4、3/4分位数等。



经验分位数估计

对一个离散的随机变量 y ，取其容量为 T 的样本序列 (y_1, \dots, y_T) ，计算第 τ 分位数的方法如下：

首先将数据从小到大排序，标号为 $i, i = 1, 2, \dots, T$ ，然后利用下方所列的方法计算随机变量 y 的第 τ 分位数的排列序号的 i ；如果 i 为整数，则随机变量 y 的第 τ 分位数即为 y_i ，如果 i 不是整数，则随机变量 y 的第 τ 分位数为：

$$y_{(\tau)} = y_{[i]} + (i - [i]) \left(y_{[i]+1} - y_{[i]} \right)$$

其中 $[i]$ 表示不大于 i 的最大整数。给定一个具体的随机变量 y ，对于一个容量为 T 的样本，则 y 的第 τ 分位数的序号 i 的计算方法如下。在大样本情况下，各方法收敛到同一值。

连续样本的经验分位数利用连续样本经验分位数表中列的方法计算。

```
quantile(x, probs= seq(0, 1, 0.25), na.rm = FALSE, names = TRUE, type = 7, ...)
```

其中，type是1到9的取值，代表不同的经验分位数的算法，默认是第7种算法，具体的各种算法如下表。

(1) 离散样本分位数

对于types 1, 2 and 3, $Q[i](p)$ 是关于 p 的离散函数,当 $i= 1$ 和 2 时, $m = 0$, 当 $i= 3$ 时, $m = -1/2$.

表：离散样本经验分位数

Type1:经验分布函数的反函数. 假如 $g = 0$,则 $\gamma = 0$; g 取其他值时, γ 取1。
 Type2:与Type 1类似，但是在非连续处取均值，当 $g = 0$,则 $\gamma = 0.5$; g 取其他值时, γ 取1。
 Type3:SAS定义方法：最近的偶数顺序统计量。假如 $g = 0$,则 $\gamma = 0$ 并且 j 是偶数, $\gamma = 0$; 其他情况, γ 取1。

(2) 连续样本分位数

对于types 4-9, $Q[i](p)$ 是关于 p 的连续函数, 以及对应的 $\gamma = g$ 和 m 详见下表。样本（经验）分位数可以通过点 $(p[k], x[k])$ 之间进行线性插值得到，其中 $x[k]$ 是第 k 个顺序统计量。关于 $p[k]$ 的具体表达式详见下表。

表:连续样本经验分位数

Type4: $m = 0$, $p[k] = k/n$. 也就说经验分布函数的线性插值。

Type5: $m = 1/2$, $p[k] = (k - 0.5)/n$.这是一个分段线性回归函数。水文研究比较常用该方法。

Type6: $m = p$, $p[k] = k/(n + 1)$.因此, $p[k] = E[F(x[k])]$.MinitabandSPSS用这种方法

Type7: $m = 1 - p$, $p[k] = (k - 1)/(n - 1)$.此时, $p[k] = mode[F(x[k])]$.S语言使用此方法.

Type8: $m = (p + 1)/3$, $p[k] = (k - 1/3)/(n + 1/3)$.则 $p[k] = \sim median[F(x[k])]$.这个分位数估计方法近似与中位数无偏, 而不管x的分布。

Type9: $m = p/4 + 3/8$, $p[k] = (k - 3/8)/(n + 1/4)$.当x是正态分布是, 该估计结果对期望顺序统计量是近似无偏的。

我们利用quantile()函数求经验分位数

In [64]:

```
quantile(x, probs= c(0.1, 0.25, 0.5, 0.75), type=2)
```

10%
-8
25%
-5
50%
0
75%
5

In []:

分位数回归原理

离差绝对值 $\sum |y - \alpha|$ 在中位数时取到最小值。因此, 中位数回归估计量可以通过最小绝对离差法 (least absolute deviation, LAD) 估计

对于线性回归模型 $y_t = \mathbf{X}_t' \boldsymbol{\beta} + \mu_t$ ，通过求 $\sum \left| y_t - \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(0.5)} \right|$ 最小，得到 $\boldsymbol{\beta}$ 的中位数回归系数估计量 $\text{hat}\boldsymbol{\beta}_{(0.5)}$ ，从而得到 y_t 的中位数回归拟合值 $\left(\hat{y}_{(0.5)t} | \mathbf{X} \right) = \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(0.5)}$ 。

现在我们把中位数回归推广到分位数回归。对于回归模型，被解释变量 y 对以 X 为条件的第 τ 分位数用函数 $y_{(\tau)} | X$ 表示，其含义是：以 X 为条件的 y 小于等于 $y_{(\tau)} | X$ 的概率是 τ ，即 $p(y \leq y_{(\tau)} | X) = F(y_{(\tau)} | X) = \tau$ ，或者可以写成 $y_{(\tau)t} | X = F^{-1}(y_{(\tau)t} | X)$

其中 $F(y_{(\tau)} | X)$ 和 $F^{-1}(y_{(\tau)t} | X)$ 分别是 y 在给定 X 条件下的累积概率分布函数(cdf)和其反函数。则 $y_{(\tau)} | X$ 称作被解释变量 y_t 对 X 的条件分位数函数。而 $F'(y_{(\tau)t} | X) = f(y_{(\tau)t} | X)$ 则称作分位数概率密度函数。

Koenker和Bassett(1978)证明，若用 $\hat{y}_{(\tau)}$ 表示 y 的 τ 分位数回归估计量，则对于以检查函数 (check function) w_τ 为权数， y 对任意值 α 的加权离差绝对值和 $\sum w_\tau |y_t - \alpha|$ 只有在 $\alpha = \hat{y}_{(\tau)}$ 时取得最小值。其中

$$\sum w_\tau |y_t - \alpha| = - \sum_{i: y_i < \alpha}^T (1 - \tau)(y_t - \alpha) + \sum_{i: y_i \geq \alpha}^T \tau(y_t - \alpha)$$

其中 $\tau \in (0, 1)$ 。因此，分位数回归可以通过加权的最小绝对离差和法 (weighted least absolute deviation, WLAD) 进行估计。

根据公式，对于线性回归模型 $y_t = \mathbf{X}_t' \boldsymbol{\beta} + \mu_t$ ，求第 τ 分位数回归方程系数的估计量 $\hat{\boldsymbol{\beta}}_{(\tau)}$ 的方法就相当于求使得下式 (目标函数) 达到最小时的解，

$$\begin{aligned} Q &= - \sum_{\hat{u}_{(\tau)t} < 0}^T (1 - \tau) \hat{u}_{(\tau)t} + \sum_{\hat{u}_{(\tau)t} \geq 0}^T \tau \hat{u}_{(\tau)t} \\ &= - \sum_{t: y_t < \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(\tau)}}^T (1 - \tau) (y_t - \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(\tau)}) + \sum_{t: y_t \geq \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(\tau)}}^T \tau (y_t - \mathbf{X}_t' \hat{\boldsymbol{\beta}}_{(\tau)}) \end{aligned}$$

其中， $\hat{u}_{(\tau)t}$ 表示第 τ 分位数回归方程对应的残差。

第 τ 分位数的回归方程表达式是

$$\hat{y}_{(\tau)t} = \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(\tau)}$$

当 $\tau = 0.5$ 时, Q 变为

$$Q = - \sum_{t: y_t < \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)}} 0.5 \left(y_t - \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)} \right) + \sum_{t: y_t \geq \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)}} 0.5 \left(y_t - \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)} \right) = \sum_{t=1}^T 0.5 \left| y_t - \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)} \right|$$

$\hat{y}_{(0.5)t} = \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(0.5)}$ 称作中位数回归方程, $\hat{\boldsymbol{\beta}}_{(0.5)}$ 是中位数回归系数估计量。

一旦得到估计的分位数回归方程, 就可以计算分位数回归的残差 $\hat{u}_{(\tau)t}$

$$\hat{u}_{(\tau)t} = y_t - \hat{y}_{(\tau)t} = y_t - \mathbf{X}'_t \hat{\boldsymbol{\beta}}_{(\tau)}$$

对一个样本, 估计的分位数回归式越多, 对被解释变量条件分布的理解就越充分。以一元回归为例, 如果用LAD法估计的中位数回归直线与用OLS法估计的均值回归直线有显著差别, 则表明被解释变量的分布是非对称的。如果散点图上侧分位数 (较大分位数) 回归直线之间与下侧分位数 (较小分位数) 回归直线之间相比, 相互比较接近, 则说明被解释变量的分布是左偏倚的。反之是右偏倚的。对于不同分位数回归函数如果回归系数的差异很大, 说明在不同分位数上解释变量对被解释变量的影响是不同的。

正如普通最小二乘OLS回归估计量的计算是基于最小化残差平方和一样, 分位数回归估计量的计算也是基于一种非对称形式的绝对值残差最小化, 其中, 中位数回归运用的是最小绝对值离差估计(LAD, least absolute deviations estimator)。它和OLS主要区别在于回归系数的估计方法和其渐近分布的估计。在模型设定、回归系数检验、残差检验、预测等方面则基本相同。

分位数回归的优点是：

(1) 能够更加全面的描述被解释变量条件分布的全貌，而不是仅仅分析被解释变量的条件期望（均值），也可以分析解释变量如何影响被解释变量的中位数、分位数等。

(2) 不同分位数下的回归系数估计量常常不同，即解释变量对不同水平被解释变量的影响不同。

(3) 与最小二乘法相比，估计结果对离群值则表现的更加稳健，而且，分位数回归对误差项并不要求很强的假设条件，因此对于非正态分布而言，分位数回归系数估计量则更加稳健。

R做分位数回归需要安装包`quantreg`，分位数回归的函数是`rq()`，使用方法是`rq(formula, tau=.5, data, subset, weights, na.action, method="br", model = TRUE, contrasts, ...)`

其中`formula`是模型表达式，与`lm()`等函数类似；`tau`是设置分位数的参数，比如0.5则表示中位数回归，也可以设置一个分位数的向量，比如`(.05,.1,.25,.75,.9,.95)`等。

我们对示例的恩格尔数据进行中位数回归估计：

In [65]:

```
q_0.5=rq(foodexp~ income, tau=0.5, data=engel)
summary(q_0.5)
```

```
Call: rq(formula = foodexp ~ income, tau = 0.5, data = engel)
```

```
tau: [1] 0.5
```

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	81.48225	53.25915	114.01156
income	0.56018	0.48702	0.60199

则中位数回归估计结果：

$$\begin{aligned}\hat{food}_t &= \hat{\beta}_{0(0.5)} + \hat{\beta}_{1(0.5)} income \\ &= 81.48 + 0.56018 income\end{aligned}$$

而其此时均值回归方程为：

In [66]:

```
lmf<-lm(foodexp~income, data=engel)
summary(lmf)
```

```
Call:
lm(formula = foodexp ~ income, data = engel)

Residuals:
    Min       1Q   Median       3Q      Max
-725.70  -60.24   -4.32   53.41  515.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.47539   15.95708    9.242  <2e-16 ***
income        0.48518    0.01437   33.772  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 114.1 on 233 degrees of freedom
Multiple R-squared:  0.8304,    Adjusted R-squared:  0.8
296
F-statistic: 1141 on 1 and 233 DF,  p-value: < 2.2e-16
```

我们发现中位数回归和均值回归方程不管是截距还是回归系数差异挺大的，进一步说明解释变量 y_t 的分布是非对称的。

下面同时求(.05,.1,.25,.75,.9,.95)上分位数回归的系数。

In [69]:

```
taus<-c(.05,.1,.25,.75,.9,.95)
xx <-seq(min(income),max(income),100)
rqss=rq((foodexp)~(income),tau=taus)
summary(rqss)
```

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.05

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	124.88004	98.30212	130.51695
income	0.34336	0.34333	0.38975

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.1

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	110.14157	79.88753	146.18875
income	0.40177	0.34210	0.45079

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.25

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	95.48354	73.78608	120.09847
income	0.47410	0.42033	0.49433

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.75

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	62.39659	32.74488	107.31362
income	0.64401	0.58016	0.69041

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.9

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	67.35087	37.11802	103.17399
income	0.68630	0.64937	0.74223

Call: rq(formula = (foodexp) ~ (income), tau = taus)

tau: [1] 0.95

Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	64.10396	46.26495	83.57896
income	0.70907	0.67390	0.73444

下面在上面的散点图基础上添加均值回归拟合线和分位数回归拟合线。

In []:

```
f <-coef(rqss) #提取分位数回归系数
yy<-cbind(1,xx)%*%f #计算分位数回归拟合值
for(i in 1:length(taus)){
  lines(xx,yy[,i],col = "gray")
}
abline(lm(foodexp~ income),col="red",lty= 2)
abline(rq(foodexp~ income), col="blue")
legend(2500,500,c("mean (LSE) fit","median (LAE) fit"),
      col = c("red","blue"),lty= c(2,1))
```

从图可以看出，中位数回归线在均值回归线上方，也就说中位数回归线的回归系数更大些，另外，散点图上侧分位数（较大分位数）回归直线之间与下侧分位数（较小分位数）回归直线之间相比，相互比较疏远，则说明被解释变量 y_i 的分布是右偏倚的，这与上文的结论一致。

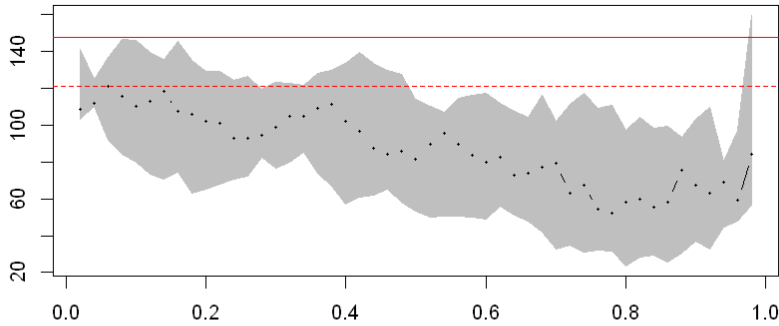
接下来，我将不同的分位数下的回归系数画一个趋势图，可以了解在不同的分位数下回归系数的变化情况。

In [73]:

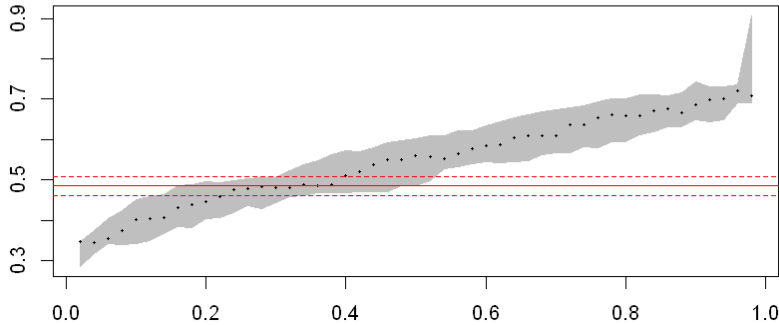
```
plot(summary(rq(foodexp~income, tau = 1:49/50,data=engel)))
```

Warning message in rq.fit.br(x, y, tau = tau, ci = TRUE, ...):
"Solution may be nonunique"
Warning message in rq.fit.br(x, y, tau = tau, ci = TRUE, ...):
"Solution may be nonunique"
Warning message in rq.fit.br(x, y, tau = tau, ci = TRUE, ...):
"Solution may be nonunique"
Warning message in rq.fit.br(x, y, tau = tau, ci = TRUE, ...):
"Solution may be nonunique"

(Intercept)



income



是恩格尔数据不同分位数下的截距和回归系数的走势图，可以发现，随着收入的增加，截距项是在逐渐下降，而回归系数是在不断上升的。