03

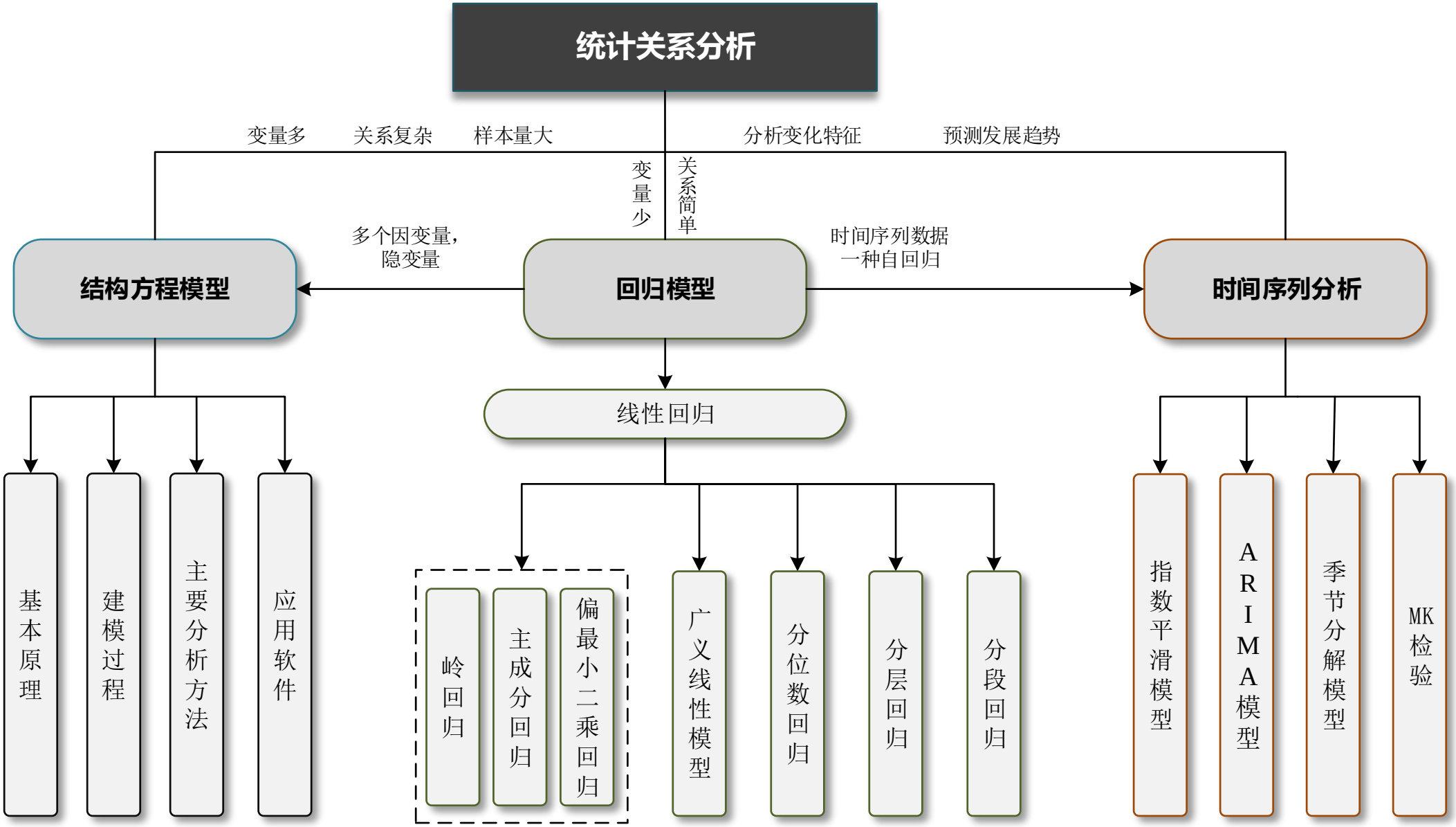
统计关系分析

统计关系分析模型

对统计关系进行定量分析

回归分析模型	结构方程模型	时间序列分析
根据已得的试验结果以及以往的经验来建立统计模型	基于变量的协方差矩阵来分析变量之间的关系	根据时间特征分析变化趋势

本章结构


统计关系分析
变量多
关系复杂
样本量大
分析变化特征
预测发展趋势
变量少
关系简单
多个因变量，
隐变量
时间序列数据
一种自回归
结构方程模型
回归模型
时间序列分析
基本原理
建模过程
主要分析方法
应用软件
线性回归
岭回归
主成分回归
偏最小二乘回归
广义线性模型
分位数回归
分层回归
分段回归
指数平滑模型
ARIMA模型
季节分解模型
MK检验

线性回归

线性回归模型

回归分析：通过自变量的给定值来估计或预测因变量的均值。 $Y = f(x) + \varepsilon$

线性回归模型的一般形式：

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

对于一个实际问题，如果获取了n个空间单元上的观测数据，则线性回归方程可以表示为

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon \\ \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon \end{cases}$$

矩阵表达形式为 $y = X\beta + \varepsilon$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

X是一个$n \times (p+1)$阶矩阵，称为回归设计矩阵。

线性回归要点

要点总结

01 自变量与因变量之间必须有线性关系；

02 多元回归存在多重共线性、自相关性和异方差性；

03 线性回归对异常值非常敏感。它会严重影响回归线，最终影响预测值；

04 多重共线性会增加系数估计值的方差，使得估计值对于模型的轻微变化异常敏感，结果就是系数估计值不稳定；

05 在存在多个自变量的情况下，我们可以使用向前选择法，向后剔除法和逐步筛选法来选择最重要的自变量。

.

线性回归

回归方程的参数估计

	最小二乘估计	最大似然估计
原理	寻找参数 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，…， 使离差平方和达到极小。	线性回归模型的似然函数为 $L=(2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta))$ 最大似然估计就是取似然函数L达到最大的$\hat{\beta}$和σ^2
回归参数β的估计值	$\hat{\beta}=(X'X)^{-1}X'y$	$\hat{\beta}=(X'X)^{-1}X'y$

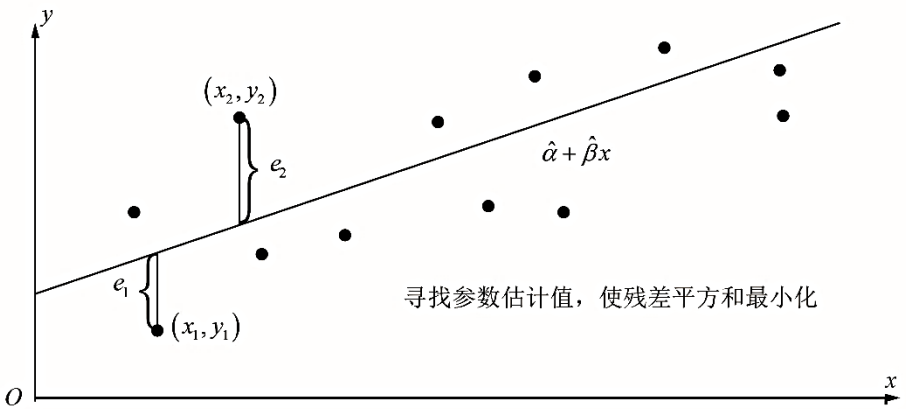
线性回归

回归方程的参数估计

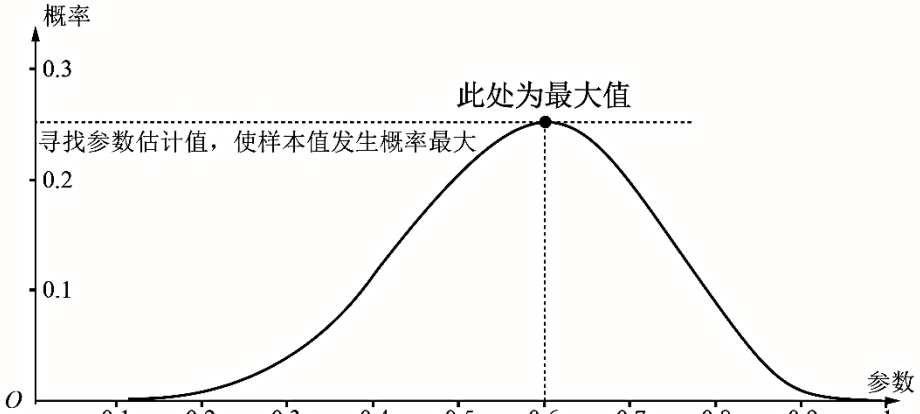
回归方程的因变量y的拟合值为 $\hat{y}=X\hat{\beta}=X(X'X)^{-1}X'y$

其回归残差的计算公式为 $e=\hat{y}-y$

最小二乘估计和最大似然估计两种方法总结如下图



最小二乘估计



最大似然估计

回归方程的显著性检验

显著性检验	**F检验**	***t*检验**
原理	看自变量对随机变量*y*是否具有明显的影响	对每个自变量进行显著性检验，从回归方程中剔除那些次要的、可有可无的变量
统计量	$F = \frac{SSR / p}{SSE / (n-p-1)}$	$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$
临界值	$F_a(p, n-p-1)$	$t_{a/2}$
解释	当$F > F_a(p, n-p-1)$时，回归方程是显著；反之，当$F < F_a(p, n-p-1)$时，则认为回归方程不显著。	当$\lvert t_j \rvert \geq t_{a/2}$时，认为自变量$x_i$对$y$的线性效果显著；当$\lvert t_j \rvert < t_{a/2}$时，认为自变量$x_i$对$y$的线性效果不显著。

线性回归方程的拟合优度

样本决定系数

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2越接近1，表明回归拟合的效果越好

注意：

- 一般认为R^2等于0.7左右才算通过了拟合优度检验。
- R^2与回归方程中自变量的数目以及样本量n有关

样本复相关系数

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

表示回归方程对原有数据拟合程度的好坏

线性回归

应用实例

以中国民航客运量回归模型为例，探究线性模型的建立和检验。为研究我国民航客运量的变化趋势及成因，以民航客运量（万人）作为y，国民收入（亿元）、消费额（亿元）、铁路客运量（万人）、民航航线历程（万人）、来华旅游入境人数（万人）为影响民航客运量的主要因素，依次作为自变量x_1，x_2，x_3，x_4和x_5。

Step 1：提出自变量与因变量，收集数据

年份	y	x_1	x_2	x_3	x_4	x_5
1978	231	3010	1888	81491	14.89	180.92
1979	298	3350	2195	86389	16	420.39
1980	343	3688	2531	92204	19.53	570.25
1981	401	3941	2799	95300	21.82	776.71
1982	445	4258	3054	99922	22.27	792.43
1983	391	4736	3358	106044	22.91	947.7
1984	554	5652	3905	110353	26.02	1285.22
1985	744	7020	4879	112110	27.72	1783.3
1986	997	7859	5552	108579	32.43	2281.95
1987	1310	9313	6386	112429	38.91	2690.23
1988	1442	11738	8038	122645	37.38	3169.48
1989	1283	13176	9005	113807	47.19	2450.14
1990	1660	14384	9663	95712	50.68	2746.2
1991	2178	16557	10969	95081	55.91	3335.65
1992	2886	20223	12985	99693	83.66	3311.5
1993	3383	24882	15949	105458	96.08	4152.7

Step 2：做相关分析，设定理论模型

计算增广相关阵，可以发现，y与x_1，x_2，x_4和x_5的相关系数均在0.9以上，说明所选自变量与y高度相关，用y与自变量做多元线性回归是适合的。

	y	x1	x2	x3	x4	x5
y	1.000	0.989	0.985	0.227	0.987	0.924
x1	0.989	1.000	0.999	0.258	0.984	0.930
x2	0.985	0.999	1.000	0.289	0.978	0.942
x3	0.227	0.258	0.289	1.000	0.213	0.504
x4	0.987	0.984	0.978	0.213	1.000	0.882
x5	0.924	0.930	0.942	0.504	0.882	1.000

相关系数

	y	x1	x2	x3	x4	x5
y		0.0000	0.0000	0.3981	0.0000	0.0000
x1	0.0000		0.0000	0.3350	0.0000	0.0000
x2	0.0000	0.0000		0.2777	0.0000	0.0000
x3	0.3981	0.3350	0.2777		0.4285	0.0464
x4	0.0000	0.0000	0.0000	0.4285		0.0000
x5	0.0000	0.0000	0.0000	0.0464	0.0000	

P值

线性回归

应用实例

Step 3：使用R软件计算多元回归模型，输出计算成果

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 450.909240 178.077719   2.532 0.029764 *
x1            0.353898   0.085230   4.152 0.001973 **
x2           -0.561476   0.125384  -4.478 0.001183 **
x3           -0.007254   0.002067  -3.510 0.005633 **
x4           21.577860   4.030051   5.354 0.000322 ***
x5            0.435188   0.051560   8.440 7.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.49 on 10 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9973
F-statistic:  1128 on 5 and 10 DF,  p-value: 2.03e-13
```

Step 4：进行回归诊断

①回归方程的构建：

$$\hat{y} = 450.9 + 0.354x_1 - 0.561x_2 - 0.007x_3 + 21.578x_4 + 0.435x_5$$

②回归方程的拟合优度：复相关系数R=0.999，决定系数R^2=0.998，由决定系数可知，回归方程高度显著；

③回归方程的显著性检验：方差分析中，F=1128，P=$2.03e^{-13}$，表明回归方程高度显著，说明x_1，x_2，x_3，x_4和x_5整体上对y有高度显著的线性影响；

④回归系数的显著性检验：可以发现自变量x_1，x_2，x_3，x_4和x_5均对y有显著影响。尽管x_3的铁路客运量的P值最大，但仍然在1%的显著性水平上对y高度显著，这也充分说明多元线性回归中，不能仅凭借相关系数的大小来决定变量的取舍。

多重共线性

多重共线性检验

多重共线性（Multi-collinearity）:在多元线性回归中，自变量之间存在较强的相关关系，使得模型失真或者难以估计准确。

方差膨胀因子

一个变量与其他变量间越大，则表明多重共线性越强。

特征根判定法

当矩阵至少有一个特征根近似为0时，X的列向量必定存在多重共线性。

直观判定

当增加或剔除一个自变量，或者改变一个观测值，回归系数的估计值发生较大变化，则存在严重的多重共线性。

多重共线性

消除多重共线性的方法

在涉及到自变量较多时	→	剔除一些不重要的解释变量
当回归方程中的全部自变量都通过显著性检验后仍然存在严重的多重共线性	→	重复剔除方差膨胀因子最大者所对应的自变量，直到回归方程中不再存在严重的多重共线性
当样本数据较少时	→	增大样本数据

以有偏估计为代价提高模型稳定性

- 岭回归法
- 主成分回归法
- 偏最小二乘法

多重共线性

岭回归

模型思想

对于设计矩阵*X*有| *X'X* |≈0时，若给其加上一个正整数矩阵*kI*，则新矩阵*X'X* + *kI*则会比*X'X*接近奇异的程度小的多。

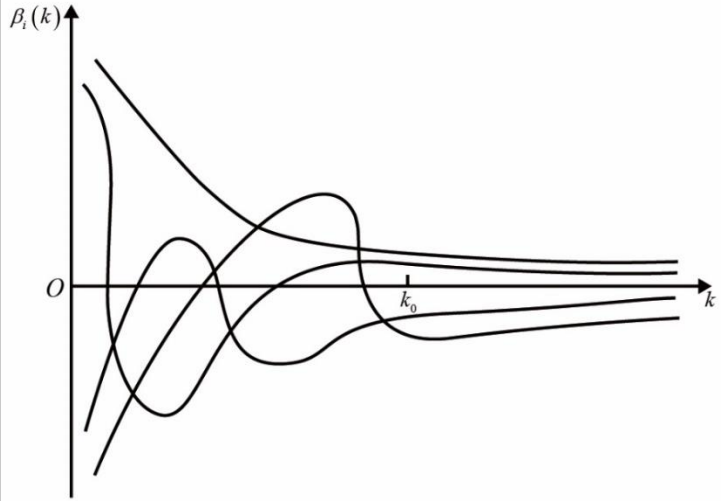
$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y$$

特性

有偏性是岭回归估计的一个重要特性。

适用情况

实际应用中只有当最小二乘回归结果不满意时才考虑使用岭回归。



岭迹图

岭参数k的选择

可以根据岭迹曲线的变化来确定适当的*k*值并进行自变量的选择。

选择*k*值的原则

- 各回归系数的岭估计基本稳定；
- 用最小二乘估计时符号不合理的回归系数，使用岭估计后其符号变得合理；
- 回归系数没有不合乎实际意义的绝对值
- 残差平方和增加不太多。

岭回归选择变量

岭回归的一个重要应用是选择变量，选择变量通常的原则是：

01 在岭回归的计算中，假定设计矩阵X已经中心化和标准化，这样可以直接比较标准化岭回归系数的大小，可以剔除掉标准化岭回归系数比较稳定且绝对值很小的变量。

02 当k值较小时，标准化岭回归系数的绝对值并不会很小但是会不稳定，因此随着k的增加会迅速趋于0，这样的岭回归系数并不稳定，也可以予以剔除。

03 删除标准化岭回归系数很不稳定的自变量，若有若干个岭回归系数不稳定，究竟剔除哪个或者哪几个，并无一般原则可循，需要根据剔除某个变量后重新进行岭回归分析的效果来确定。

主成分回归

基本思想

将线性相关的一类变量转化成为线性无关的一类新的综合变量，利用这些新的综合变量来反映原来多个变量的信息。

Step 1	Step 2	Step 3	Step 4
用主成分分析法计算出主成分表达式和主成分得分变量。	将因变量对主成分得分变量回归。	将主成分的表达式代回到回归模型中，得到标准化自变量与因变量的回归模型。	将标准化自变量转为原始自变量，得到主成分回归模型。

多重共线性

偏最小二乘回归

偏最小二乘：寻找x_1，x_2，…，x_p的线性函数时，需要考虑与y的相关性，选择与y相关性较强又能方便算得的线性函数。

提出目的

在解释变量里寻找某些线性组合，能更好地反映变量的变异信息。

适用情况

当两组变量的个数很多，且都存在多重相关性，而样本量又较少时。

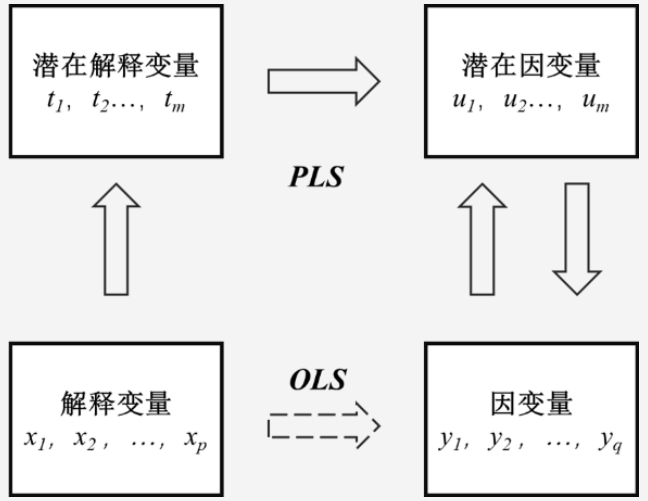
偏最小二乘回归

↑ 间接建立

因变量与解释变量之间的关系

↓ 直接建立

普通线性回归



偏最小二乘与最小二乘法的关系

偏最小二乘回归计算流程

Step 1

分别在自变量X与自变量Y中提取出成分t_1和u_1。提取成分有两个要求：

- t_1和u_1应尽可能大地携带他们各自数据表中的变异信息；
- t_1和u_1的相关程度能够达到最大。

相比于主成分分析方法具有的优势

Step 2

分别实施X对t_1的回归以及Y对u_1的回归。重复进行成分提取，直到能达到一个较满意的精度为止。

Step 3

若最终对X共提取了 m个成分，实施y_k对这m个成分的回归，然后再表达成y_k关于原变量x的回归方程。

多重共线性

应用实例

本例依旧沿用上节的民航数据集，探讨多重共线性及其模型优化。

Step 1：多重共线性的度量

①使用方差膨胀因子和条件数分别对该例的多重共线性进行度量。

Index	x_1	x_2	x_3	x_4	x_5
VIF	1963.33686	1740.507552	3.171186	55.488301	25.192748

- x_1是国民收入，x_2是消费额，查看上一节，二者的简单相关系数高达0.999。
- 一般情况下，回归方程的多重共线性的存在就是由方差膨胀因子超过10的数个变量引起的，说明这些自变量间有一定的多重共线性。

②对上节建立的多元回归方程使用条件数进行多重共线性的度量。

Condition Index	Variance Decomposition Proportions				
	x_1	x_2	x_3	x_4	x_5
1	0	0	0.003	0.001	0.002
2.069	0	0	0.3	0.001	0.001
7.827	0	0	0.324	0.096	0.354
18.889	0.009	0.014	0.14	0.591	0.582
121.221	0.991	0.986	0.234	0.311	0.06

- 最大的条件数为121.221
- 最大条件数主要是由x_1，x_2贡献的

多重共线性

应用实例

Step 2：剔除自变量的方法

①x_1的VIF值最大，首先去除x_1，重新建立回归方程。相应地，再次计算VIF值。

Index	x_2	x_3	x_4	x_5
VIF	77.545528	2.319334	33.811545	24.468681

②最大的VIF值依然远大于10，回归方程依然存在多重共线性，需要继续进行变量剔除，建立新的回归方程，并计算VIF值。

Index	x_3	x_4	x_5
VIF	1.983698	6.649848	8.513852

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 591.875903 257.729834   2.296  0.04045 *
x3           -0.010366   0.002635  -3.934  0.00199 **
x4           26.435810   2.249140  11.754 6.09e-08 ***
x5            0.317384   0.048321   6.568 2.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 79.79 on 12 degrees of freedom
Multiple R-squared:  0.9945,    Adjusted R-squared:  0.9931
F-statistic: 720.8 on 3 and 12 DF,  p-value: 8.263e-14
```

③此时的回归模型不存在明显的多重共线性，可以作为最终的回归模型。最终的回归方程为

$$\hat{y} = 591.876\text{-}0.010x_3 + 26.436x_4 + 0.317x_5$$

缺点：剔除了多个变量，损失了部分原始信息

多重共线性

应用实例

岭回归

①设定岭参数k使其在0~1之间以0.05的步长变化，得到一份岭迹图

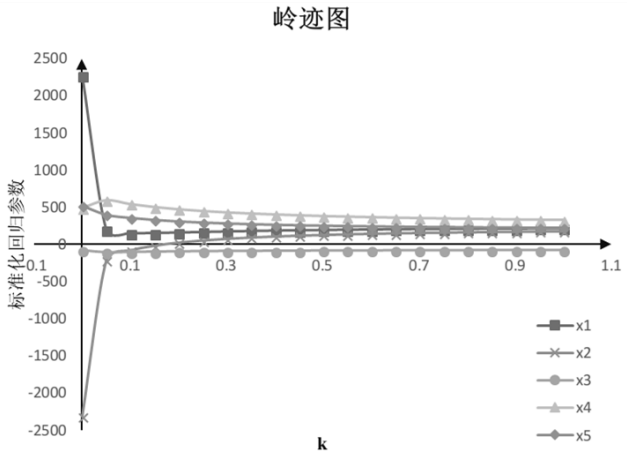


表 3-6　不同岭参数下的岭回归估计参数及拟合优度

岭参数	调整 R^2	截距	x_1	x_2	x_3	x_4	x_5
k=0	0.997	445.737	0.351	−0.555	−0.007	21.401	0.429
k=0.05	0.947	380.600	0.032	0.041	−0.007	15.053	0.192
k=0.1	0.912	323.843	0.035	0.048	−0.006	13.172	0.175
k=0.15	0.879	283.024	0.035	0.051	−0.006	12.321	0.169
k=0.95	0.495	70.065	0.030	0.046	−0.001	8.882	0.139
k=1	0.477	68.709	0.030	0.046	−0.001	8.768	0.138

- 在k=0.05前后，各变量的回归参数在岭迹图上趋于平稳，意味着最优的岭参数k落在这个区间里。
- 结合回归参数表，发现各项系数逐渐趋于平稳，相应地，方程的拟合优度在逐渐下降，这主要是由于岭回归有偏估计的特性。

②x_1与x_2的系数发生了剧烈变化，结合方差膨胀因子表，发现k=0时x_1的方差膨胀因子比x_2大，应考虑首先剔除变量x_1。

表 3-7　不同岭参数下的各个自变量的方差膨胀因子

岭参数	x_1	x_2	x_3	x_4	x_5
k=0	2021.476	1777.495	3.179	57.942	25.011
k=0.05	0.741	0.936	1.210	2.892	3.434
k=0.1	0.283	0.332	0.949	1.247	1.612
k=0.15	0.176	0.194	0.811	0.731	0.959
k=0.95	0.050	0.048	0.244	0.077	0.082
k=1	0.049	0.047	0.232	0.073	0.077

③近似地选定k=0.05，建立相应的回归方程。

$$\hat{y} = 457.338+0.063x_3-0.008x_3+18.295x_4+0.231x_5$$

优点：多保留一些自变量

多重共线性

应用实例

主成分回归

①对原始数据集进行主成分分析，得到方差贡献分析结果。在将原数据集分解为2个主成分时，可以解释98.47%的原始信息。

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.9981 0.9654 0.25381 0.10544 0.01627
Proportion of Variance 0.7984 0.1864 0.01288 0.00222 0.00005
Cumulative Proportion  0.7984 0.9848 0.99772 0.99995 1.00000
```

②以使用前两个主成分代表原始的5 个变量的信息，则对y与PC1和PC2构建线性关系。

$$\hat{y} = 1159.12 - 453.94PC1 + 186.42PC2$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1159.12      27.99   41.41 3.42e-15 ***
C1            453.94      14.01   32.41 8.07e-14 ***
C2            186.42      28.99    6.43 2.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112 on 13 degrees of freedom
Multiple R-squared:  0.9882,    Adjusted R-squared:  0.9864
F-statistic: 545.7 on 2 and 13 DF,  p-value: 2.886e-13
```

③还原回归方程

$$\hat{y} = 416.76 + 0.03976x_1 + 0.06082x_2 - 0.00765x_3 + 11.37x_4 + 0.1628x_5$$

优点：保留了所有的变量，回归方程显示的自变量与因变量间的关系符合实际意义

偏最小二乘回归

①对原始数据集进行偏最小二乘回归，采用留一验证法进行验证，得到交叉验证分析表和方差贡献表。

```
VALIDATION: RMSEP
Cross-validated using 16 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           1.033   0.2109   0.1268   0.1415   0.1112  0.07731
adjCV        1.033   0.2091   0.1261   0.1384   0.1098  0.07612

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps
X    79.70    98.47    99.00    99.99   100.00
Y    96.64    98.85    99.39    99.52    99.82
```

②可以发现，前两个主成分可以解释原始自变量98.84%的信息。利用R 软件的“plsr”包，设置主成分数为2，重新构建得到如下的结果

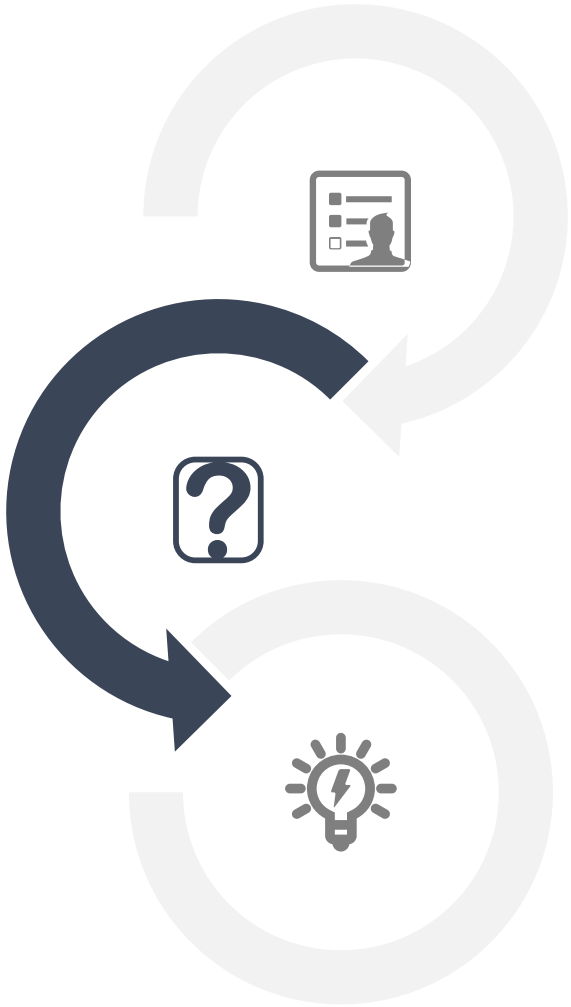
$$\hat{y} = 421.324 + 0.039x_1 + 0.059x_2 - 0.008x_3 + 11.708x_4 + 0.164x_5$$

广义线性模型

普通多元回归模型的问题

问题

- 残差不再服从零均值的正态分布
- 因变量的取值区间受限制
- ……



要求

因变量应是连续型数值变量。

解决方式

Nelder and Wedderburn提出了广义线性模型，放宽了限制：

- 因变量可以是连续或非连续的类型。
- 自变量的线性预测值仅是因变量的函数估计。

广义线性模型

指数分布族

指数分布族的概率分布形式如下

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\mathrm{a}(\phi)} + c(y, \phi)\right)$$

θ —— 自然参数

μ —— 均值

ϕ —— 散布参数

用来表示因变量的分布

每个观测值都有一个自然参数，而所有观测值的散布参数均为Φ，这表明，每个观测值都来自同一个分布，其均值因为自然参数的变化而不断变化，则称y服从**指数族分布**。

指数族分布
- 泊松分布
- 二项分布
- 指数分布
- 伽马分布
- 几何分布
- 负二项分布

广义线性模型

广义线性模型基本思想

01.系统成分

$$\eta = X\beta$$

系统成分是解释变量的线性组合

02.随机成分

$$\varepsilon = Y - \eta$$

随机成分指的是因变量y或随机误差ε的分布服从比正态分布更一般的概率分布

03.连接函数

$$g(u_i) = \sum_{j=1}^{p} X_{ij}\beta_{ij} \qquad i = 1, 2, \ldots, n$$

连接函数g是将自变量第i组观测值的线性组合与因变量的第i个观测值$_i$的数学期望联系起来的函数

广义线性模型

广义线性模型一般形式

广义线性模型的一般形式： $g(\mu)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_m x_m+\varepsilon$

广义线性模型的联系函数

模型	分布	联系函数	$\eta=g(\mu)$
一般线性模型	正态分布	恒等函数	$\eta=\mu$
逻辑回归模型	二项分布	Logit函数	$\eta=\ln(\mu/(1-\mu))$
泊松回归模型	泊松分布	对数函数	$\eta=\ln(\mu)$

广义线性模型参数估计和假设检验

参数估计

由于不再符合最小二乘的前提条件，广义线性模型采用极大似然估计法进行参数估计，其似然方程为

$$\frac{\partial l}{\partial \beta_i} = 0 = \sum_i \frac{\partial}{\partial \theta_i} \frac{(y_i \theta_i - b(\theta_i))}{a(\phi)} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i})$$

$$\Rightarrow \frac{\partial l}{\partial \beta_i} = \sum_i \frac{(y_i - \mu_i)}{a(\phi)} \frac{1}{\mathrm{b}''(\theta_i)} \frac{1}{\mathrm{g}'(\mu_i)} x_{ij}$$

假设检验

- Wald检验
- 约束检验
- 拟似然比检验

广义线性模型

01.概念

目标概率取值在0~1之间，但是回归方程的因变量取值却落在实数集当中

↓ Logit变换

取值区间变成整个实数集

→ **逻辑回归**

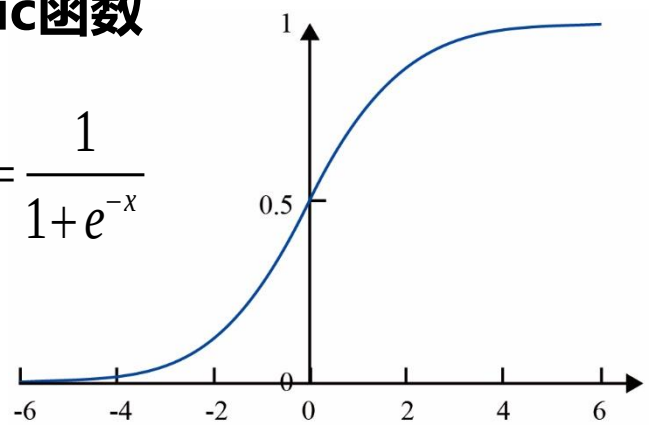
逻辑回归模型思想

02.适用情形

因变量是二分类变量，此时因变量服从二项分布，回归函数区间为[0，1]。

03.Logistic函数

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



04.回归模型

回归函数$E(y_i)$表示在自变量为x_i的条件下y_i的平均值，而y_i是0-1型随机变量，因而$E(y_i)$是在自变量为x_i的条件下y_i等于1的比例。可以用y_i等于1的比例P代替y_i本身作为因变量。

$$P=\frac{e^{g(\mu)}}{1+e^{g(\mu)}}$$

05.与普通线性模型的区别

- 因变量和自变量之间不存在线性关系。
- 可以是离散变量、连续变量甚至是哑变量。

广义线性模型

二元Logistic回归

p_i：在第i个观测中事件发生的概率
$1-p_i$：第i个观测中事件不发生的概率

事件的发生比： $p_i/(1-p_i)$，事件发生与不发生的概率之比，简记为Odds。对Odds做对数变换，就能够得到Logistic回归模型的线性模式：

$$\ln(\frac{p_i}{1-p_i}) = \alpha + \sum_{i=1}^{m} \beta_i x_i$$

其中x_i为影响y的m个自变量

Logistic要点

- Logistic回归广泛用于分类问题；
- Logistic回归不要求自变量和因变量存在线性关系。它可以处理多种类型的关系，因为它对预测的相对风险指数使用了一个非线性的 log转换；
- 为了避免过拟合和欠拟合，我们应该包括所有重要的变量。有一个很好的方法来确保这种情况，就是使用逐步筛选方法来估计Logistic回归；
- Logistic回归需要较大的样本量，因为在样本数量较少的情况下，极大似然估计的效果比普通的最小二乘法差；
- 自变量之间应该互不相关，即不存在多重共线性。然而，在分析和建模中，我们可以选择包含分类变量相互作用的影响；
- 如果因变量的值是定序变量，则称它为序Logistic回归；
- 如果因变量是多类的话，则称它为多元Logistic回归。

广义线性模型

应用实例

本例使用iris数据集中的部分数据进行建模。

通过对原始数据进行预处理，提取种类（Species）中的两种花Versicolor和Virginica作为因变量，花瓣的长度（Petel.Length）和宽度（Petal.Width）作为解释变量建立模型。

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -45.272     13.610  -3.327 0.000879 ***
Petal.Length    5.755      2.306   2.496 0.012565 *
Petal.Width    10.447      3.755   2.782 0.005405 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  20.564  on 97  degrees of freedom
AIC: 26.564
```

结果表明，花瓣的长度和宽度能够有效区分这两种花。构建的模型的形式为

$$g(\mu) = -45.272 + 5.755x_1 + 10.447x_2$$

即

$$P = \frac{e^{-45.272+5.755x_1+10.447x_2}}{1+e^{-45.272+5.755x_1+10.447x_2}}$$

广义线性模型

泊松回归

01.Poisson回归

基于事件计数变量而建立的回归模型，能对离散型随机变量进行回归建模，计数变量是指事件发生的次数。

02.应用场景

常用于单位时间或单位空间内某稀有事件发生数的影响因素分析。

03.回归模型

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

04.特点

- 因变量的对数与自变量呈线性关系
- 各观测量之间相互独立
- 各自变量水平上的因变量的方差与均值相等

广义线性模型

应用实例

本例以Breslow癫痫数据集中的部分数据作为数据源。数据记录了在对癫痫症患者实施药物治疗的最初八周内，抗癫痫药物对癫痫发病数的影响，提取的数据包括sumY（八周内的癫痫发病数），治疗条件（Trt），年龄（Age）和前八周内的基础癫痫发病数（Base）。

对这些数据使用泊松回归进行拟合

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9488259  0.1356191  14.370  < 2e-16 ***
Base          0.0226517  0.0005093  44.476  < 2e-16 ***
Age           0.0227401  0.0040240   5.651 1.59e-08 ***
Trtprogabide -0.1527009  0.0478051  -3.194   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2122.73  on 58  degrees of freedom
Residual deviance:  559.44  on 55  degrees of freedom
AIC: 850.71
```

泊松回归

```
(Intercept)        Base         Age Trtprogabide
  7.0204403   1.0229102   1.0230007   0.8583864
```

回归参数指数化

注：Trt这一字段的记录是一个二分类变量，为placebo（对照组）和progabide（实验组）。进一步证实，泊松回归对于自变量的分布没有要求。

- 治疗条件（Trt），年龄（Age）和前八周内的基础癫痫发病数（Base）三个自变量是显著影响治疗效果（用药八周内的癫痫发病数）的。
- 年龄+1岁⟶癫痫发病数的对数均值+0.0227

↓ 将参数指数化

无药物　　年龄+1岁 ⟶ 期望的癫痫发病数×1.023

有药物　　年龄+1岁 ⟶ 期望的癫痫发病数× 0.86

在其他条件不变的情况下，实验组相对于对照的发病数降低了14%，药物对于降低癫痫的发病次数有显著效果。

分位数回归

分位数回归

01.定义

利用解释变量的多个分位数（例如四分位、十分位、百分位等）来得到因变量的条件分布相对应的分位数方程。

02.特点

- 能够捕捉分布的尾部特征，更加全面的刻画分布的特征。
- 得到设定个数的分位数函数，能挖掘更加丰富的信息。.
- 对误差项并不要求很强的假设条件，能够更加全面的描述因变量条件分布的全貌.
- 解释变量对不同水平因变量的影响不同，且估计结果对离群值则表现的更加稳健。

分位数回归

分位数回归的模型思想

总体分位数

假设条件分布y|x的累积分布函数严格单调递增，总体条件τ分位数依赖于x记做$y_\tau(x)$,如果随机误差满足同方差的假设，或者其异方差形式为乘积，则 $y_\tau(x)$是x的线性函数。

$$y_\tau(x) = x[\beta + \alpha F_\varepsilon^{-1}(\tau)]$$

样本分位数

对于随机变量y，其总体τ分位数未知，可使用样本τ分位数来估计$y_{(\tau)}$。第τ分位数的回归方程表达式是

$$\hat{y}_\tau = \mathrm{X}\hat{\beta}_\tau$$

$$\hat{\beta}_\tau = \arg\min_{\beta \subset R^k} \left\{ \sum_{i=1}^{n} \tau(Y_i - x_i'\beta_\tau) + \sum_{i=1}^{n} (1-\tau)(Y_i - x_i'\beta_\tau) \right\}$$

分位数回归估计的检验

与普通线性回归类似的检验

可以对单个分位数回归方程中的参数有效性进行检验

分位数回归估计特殊要求的检验

斜率相等检验，即检验对于不同的分位点估计得到的结构参数是否相等。

分位数回归

应用实例

本例采用的Engel数据集源自19世纪德国统计学家恩格尔的关于食品支出和个人消费支出关系的研究。数据集包含收入（Income）和食品支出（Foodexp）两个维度共235个观测数据。

OLS回归

对数据集中的两个变量做普通最小二乘回归发现二者呈现正相关的关系。且回归方程的R^2很高，说明回归方程的拟合优度很好，食品支出和收入二者整体上的线性关系很强。

```
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:  0.8296
F-statistic:  1141 on 1 and 233 DF,  p-value: < 2.2e-16
```

中位数回归

首先对该数据集进行中位数回归，并对参数进行显著性检验。

```
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 81.48225 25.18563    3.23527  0.00139
   income       0.56018  0.03242   17.27747  0.00000
```

- 食品支出与收入的中位数回归，回归参数为0.56，截距为81.48，且均显著。说明在中位数水平上收入越高食品支出越多，符合全体样本的线性关系的趋势。
- 线性回归方程的系数略小于中位数回归的系数，说明数据局部分布的可能与整体分布所得到的关系不一致，即传统的OLS回归可能掩盖了数据的部分分布特征。

分位数回归

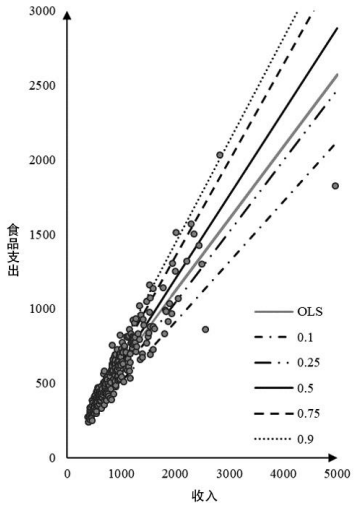
应用实例

分位数回归

Step 1：对原始样本数据进行在0.05，0.10，0.25，0.5，0.75，0.9和0.95共七个分位上进行分位数回归，得到的系数如下。

```
Coefficients:
             tau= 0.05   tau= 0.10  tau= 0.25  tau= 0.50  tau= 0.75
(Intercept) 124.8800408 110.1415742 95.4835396 81.4822474 62.3965855
income        0.3433611   0.4017658  0.4741032  0.5601806  0.6440141
            tau= 0.90  tau= 0.95
(Intercept) 67.3508721 64.1039632
   income      0.6862995  0.7090685
```

Step 2：结合普通线性回归的相关结果，将回归曲线绘制出来。



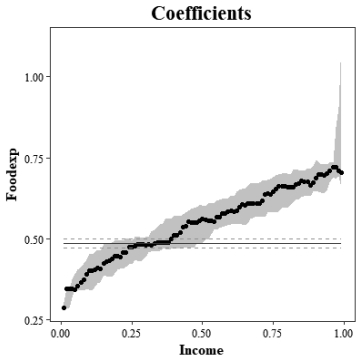
- 在不同分位下的分位数回归曲线不一致，即收入不同的家庭在食品支出上较为不同。

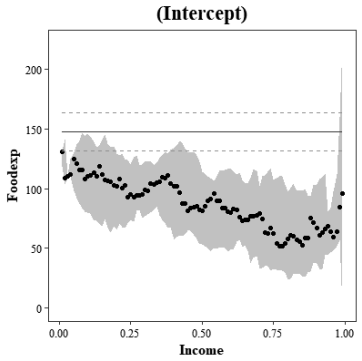
Step 3：使用偏差进行斜率相等性检验。

```
Quantile Regression Analysis of Deviance Table

Model: foodexp ~ income
Joint Test of Equality of Slopes: tau in {  0.05 0.1 0.25 0.5 0.75 0.9
0.95  }

  Df Resid Df F value    Pr(>F)
1  6     1639  11.998 3.175e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

不同分位点下收入对食品支出的影响机制是不相同的。

Step 4：绘制出在1~99个分位下分位数回归的系数和相应的截距。





不同的分位上，不同家庭的收入和支出间的线性关系的特征不同。

分层回归

分层回归简介

01.概念

分层回归：一种用于多层嵌套结构数据的线性统计方法

02.名称

Harvey Goldstein → 多层分析(Multilevel Analysis)

Stephen W.Raudenbush等人 → 分层线性模型结构(Hierarchical Linear Modeling)

03.影响

系统地解决了困扰社会科学半个多世纪的生态谬误（Ecological Fallacy）问题。

04.优点

- 能对多层数据进行综合分析
- 减少传统最小二元回归方法分析的统计误差
- 避免由人为选择分析单位可能出现的错误
- 各层样本均可作为分析单位，并且还可以研究不同层次之间的交互作用，从而拓宽了各专业的研究范围，深化了各专业的研究思路。

分层回归

分层回归的基本统计原理

- OLS估计
- 加权最小二乘估计。

公式：

$Y_i = \beta_0 + \beta_1 X_i + \gamma_i$

$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \mu_{0j} + \mu_{1j} X_{ij} + \gamma_{ij}$

HLM将残差项进行了分解，更符合实际情况。

样本要求

01 参数估计方法：收缩估计

02 样本大小与统计判断与假设检验有关

03 普通最小二乘回归

04 样本个数同样本量的比例

05 HLM的基本形式

适用数据形式：镶嵌型数据

分层回归

HLM的基本模型形式

01.零模型

方程 { 个体差异造成的部分
组间差异造成的部分

第一层方程： $Y_{ij} = \beta_{0j} + \gamma_{ij}$

第二层方程： $\beta_{0j} = \gamma_{00} + \mu_{0j}$

又叫作方差成分分析

跨级相关系数：$\rho = \tau_{00} /(\tau_{00}+\sigma_2)$

确定Y的总体变异中有多大程度是由于第二层或者组间差异造成的

02.完整模型

完整模型 { 第一层的预测变量
第二层的预测变量

第一层方程： $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \gamma_{ij}$

第二层方程： $\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \mu_{0j}$
$\beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + \mu_{1j}$

构建分析模型 { 随机效应回归模型
协方差模型

03.协方差分析模型

定义：在零模型和完整模型之间，可以通过向各层方程中增加不同的变量，设定不同的随机成分和固定成分来构建各种分析模型。

第一层： $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - X_{均}) + \gamma_{ij}$

第二层： $\beta_{0j} = \gamma_{00} + \mu_{0j}$
$\beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + \mu_{1j}$

重要前提：协方差对因变量的回归系数的组间一致性

分层回归

HLM的基本模型形式

04.随机效应回归模型

第一层： $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \gamma_{ij}$

β_{0j}和β_{1j}是随机的

第二层： $\beta_{0j} = \gamma_{00} + \mu_{0j}$

$\beta_{1j} = \gamma_{10} + \mu_{1j}$

没有预测变量

目的：寻找第一层的截距、斜率在第二层单位上的变异

05.发展模型

当不同时间的观察结果（第一层）嵌套于被观察的个体（第二层）。

- 零模型
- 完整模型
- 协方差分析模型
- 随机效应回归模型

×

发展模型:把多次的观察结果作为时间的某种数学函数来建构模型

06.三层及以上多层模型

三层模型是前面讨论的二层模型的直接扩展。以选择使用零模型和完整模型之间的任何一种模型。

第一层： $Y_{ij} = \beta_{0jk} + \gamma_{ijk}$

第二层： $\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$

第二层： $\gamma_{00k} = \pi_{000} + e_{0jk}$

在三层的零模型中，主要关注的是三个层之间的方差分解

分层回归

模型应用范围

多层模型可以广泛应用于组织和管理研究。

多层模型的第二个应用体现在对个体进行追踪、多测观测的发展研究中。

多层模型的第三种应用可以视为前面两种应用的综合，在教育研究中广为适用。

多层模型还可以用来做文献综述，即对众多的研究成果进行定量综合。

多层模型的另一种应用是利用多层的数据来回答单层数据的问题，这种方法充分利用多层模型中较为高级的统计估计方法，来改善单层回归的估计和分析。

分层回归

分层回归评述

优点

避免了没有考虑数据混合时的分析谬误。一元回归（或散点图）揭示出的规律可能是假的，因为没有控制其它变量的影响。

数据具有簇(面板或表格)结构

经典统计 ⟶ 观测变量都是独立且恒等分布的

混合效应模型

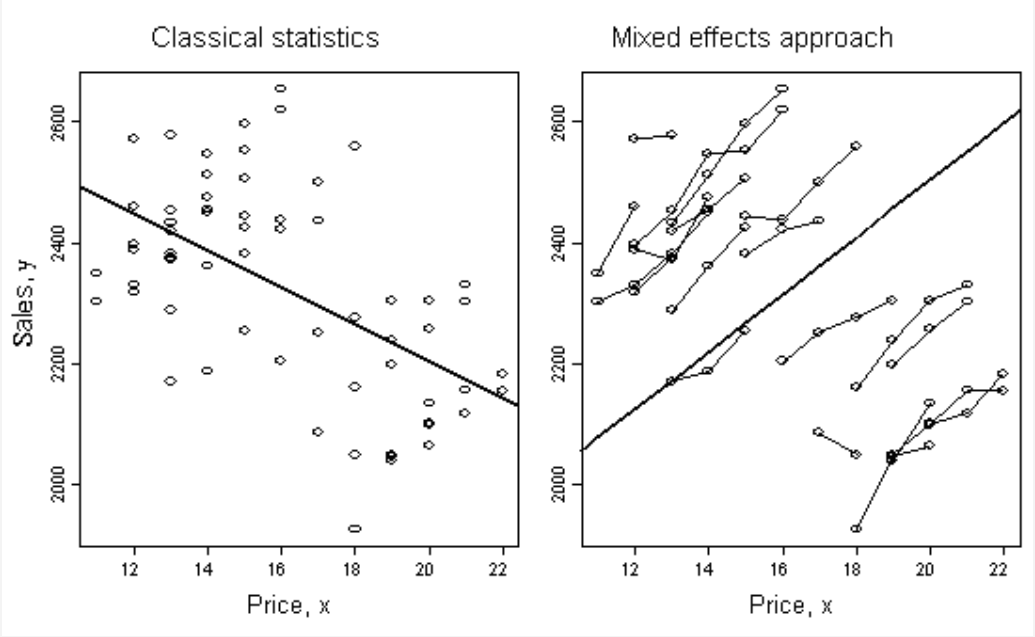
与经典统计中意义相同

随机的，用后验概率估算

变异

簇内部—总体平均水平　　不同簇间—簇特异性

销售额悖论例子



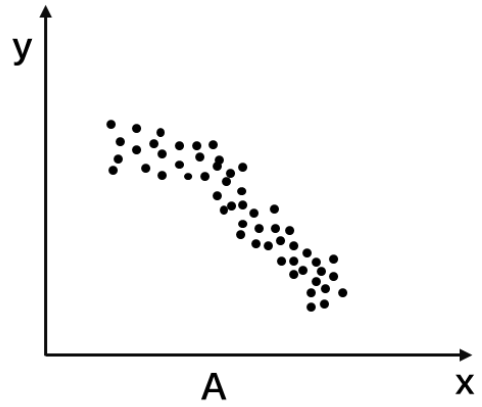
OLS模型　　产品价格↑ ⟶ 销售额↓

分层回归模型　　产品价格↑ ⟶ 销售量↑

分段回归

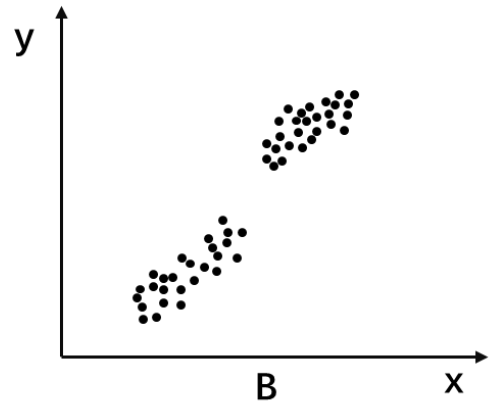
概述

分段函数：是指在函数定义域的不同部分不是用一个解析式表示，而是用几个不同的解析式来表达的函数，有时可能要用无穷多个解析式。



(1)

以A为分界点



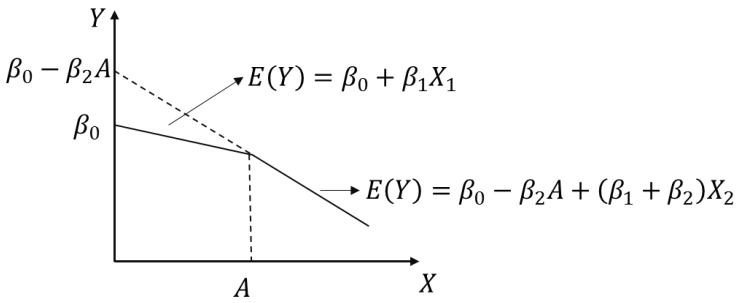
(2)

以B为分界点

分段回归

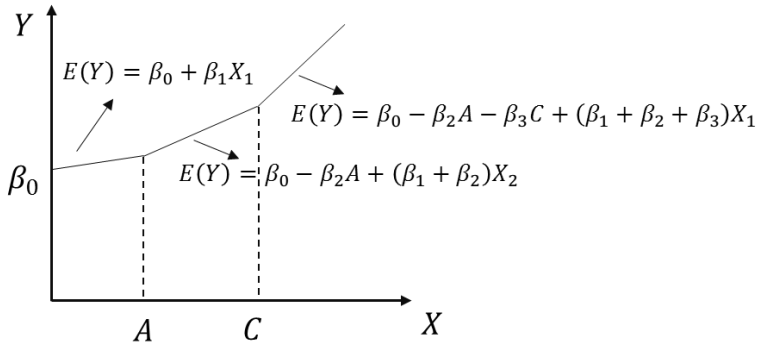
分段回归方法

有一个转折点的分段线性回归



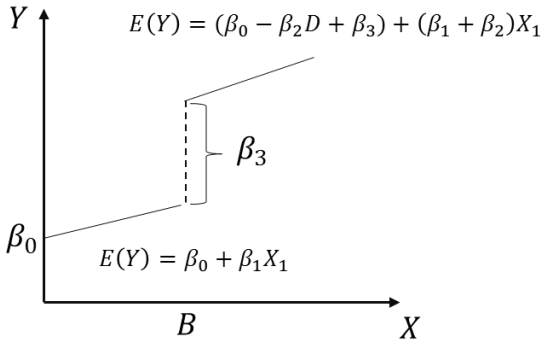
回归曲线在A处为转折点

两个以上转折点的分段回归



回归曲线在A和C为转折点

回归函数不连续的情况



分段直线在B处有一个跳跃点

结构方程模型

结构方程模型基本原理

01.定义

结构方程模型是一种建立、估计和检验因果关系模型的方法，是基于变量的协方差矩阵来分析变量之间关系的一种综合性统计方法，因此又称为协方差结构分析。

02.要求

- 样本数据
  - 符合多变量正态性
- 数据
  - 正态分布数据
- 测量指标
  - 线性关系

03.实质

- 统计方法
  - 因素分析
  - 路径分析
- 检验模型
  - 可观测的显性变量
  - 无法直接观测的潜在变量
  - 干扰或误差变量
- 获得自变量对依变量影响
  - 直接效果
  - 间接效果
  - 总效果

结构方程模型

结构方程模型组成部分

01.测量变量

可以直接测量的变量

- 内生测量变量
- 外生测量变量

02.潜在变量

无法直接观测并测量的变量

- 内生潜在变量
- 外生潜在变量

03.外生变量

只起解释变量作用的变量

- 不受其他变量的影响
- 不产生测量误差

04.内生变量

受其它变量包括外生变量和内生变量影响的变量

- 影响其它变量
- 产生测量误差

05.测量方程

使用测量变量来建构潜在变量的模型

$$x = \Lambda_x \xi + \delta$$
$$y = \Lambda_y \eta + \varepsilon$$

06.结构方程

建立潜在变量与潜在变量之间的关系

$$\eta = B_\eta + \Gamma_\xi + \zeta$$

结构方程模型

结构方程模型建模过程

1

理论
理论是假设模型成立的主要解释依据

2

模型设定
设定一个关于一组变量之间的相互关系（常常是因果关系）的模型

3

模型识别
基本原则：模型的自由参数不能多余观测数据的方差和协方差总数。

4

模型估计
推导出一个方差协方差矩阵∑，使∑与S之间的差异最小化。

5

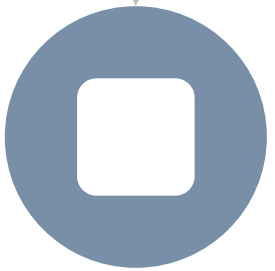
模型适配度评价
在已有的证据与理论范围内，考察所提出的模型拟合样本数据的程度。

6

模型修正
目的：改进初始模型的适合程度

7

分析解释
是对模型的统计结果进行分析、解释。



结构方程模型

结构方程模型分析方法

- 完全根据原始数据，通过迭代算法，找到各个变量之间完全的线性关系，而且估计出全部隐变量的值。
- 假定变量之间相关，图上及公式中无须注明。

偏最小二乘法
PLS

协方差方法
ML-LISREL

- 在对数据的多元正态性的假定之下，加上图模型所确定的协方差关系，得到似然函数，然后以最大似然法来得到各种估计。
- 可以仅用均值、协方差矩阵及样本量来得到所有结果。

PLS结构方程模型与ML-LSREL完全模型一样，可以表示为：

$$\eta = B\eta + \Gamma\xi + \zeta$$

$$x = \Lambda x\xi + \sigma \qquad y = \Lambda y\eta + \varepsilon$$

结构方程模型

PLS方法

01.定义

是将主成分分析与多元回归结合起来的迭代估计，是一种因果建模的方法。

02.基本思路

1. 对不同隐变量的测量变量子集抽取主成分估计潜变量得分
2. 使用普通最小二乘估计载荷系数和路径系数

03.PLS路径估计的具体步骤

Step 1：用迭代方法估计权重和潜变量得分。

①对潜变量进行标准化变化后，计算它的外部估计

②对潜变量进行标准化变化后，计算它的内部估计

③计算内部权重

④计算外部权重

Step 2：估计路径系数和载荷系数。

Step 3：估计位置参数

ML-LSREL方法

基本思路

从变量之间的协方差结构入手，通过拟合模拟模拟估计协方差与样本协方差来估计模型参数。

方法

- 极大似然法（Maximum Likelihood，ML）
- 非加权最小二乘（Unweighted Least Squares，ULS）
- 广义最小二乘（Generalized Least Squares，GLS）
- 其他方法

↓

模型估计协方差与样本协方差的拟合函数 —迭代方法→ 使拟合函数最优的参数估计

PLS方法与ML-LSREL方法的区别

01 PLS方法不用对数据做任何分布假定，而ML-LSREL方法必须假定数据服从多元正态分布。

02 PLS方法假定所有隐变量都是相关的 (即使在图模型中它们之间没有箭头) 就假定它们之间的相关严格为零，并体现在后续的计算之中。

03 PLS方法用全部数据建模，而ML-LISREL方法由于假定了分布，只要有各变量的均值，协方差矩阵和样本量就可以计算。

04 如果假定了数据变量的正态性，则ML-LISREL方法可以输出多达四十多种不同的统计量，检验的p值等，而PIS方法无法做这些检验，PLS有一些其他指标。

05 PLS适用于关注隐变量得分的情况，比如满意度指数。则ML-LISREL方法无法直接得到隐变量得分，因此各国在计算满意度指数时都用PLS方法。

06 PLS适用于小样本情形。

07 PLS由于收敛速度快，因此适用于较大、较复杂的结构方程模型，计算效率比ML-LISREL更高。

08 ML-LISREL方法是理论导向的，强调从探索到确认性分析的转换，PLS主要是在高度复杂但又没有什么理论信息时做因果预测分析。

结构方程模型

通径分析

01.目的

检验一个假想的因果模型的准确和可靠程度

测量变量间因果关系的强弱

回答
- 两变量x_j与x_i间是否存在相关关系？
- 两者间是否有因果关系？
- 若x_j影响x_i，是直接影响的、间接影响或两者都有？
- 直接影响与间接影响两者大小如何？

02.通径分析的基本模型

- SEM的单变量形式
- 只有观测变量而无潜在变量

$$\begin{cases} P_{1y}+r_{12}P_{2y}+r_{13}P_{3y}+\cdots\cdots+r_{1k}P_{ky}=r_{1y} \\ r_{12}P_{1y}+P_{2y}+r_{23}P_{3y}+\cdots\cdots+r_{2k}P_{ky}=r_{2y} \\ r_{31}P_{1y}+r_{32}P_{2y}+P_{3y}+\cdots\cdots+r_{3k}P_{ky}=r_{3y} \\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ r_{1k}P_{1y}+r_{k2}P_{2y}+r_{k3}P_{3y}+\cdots\cdots+P_{ky}=r_{ky} \end{cases}$$

03.剩余效应

实际工作中不可能把所有因变量的所有影响因素都包括在内，应进一步计算未研究的自变量和误差对因变量Y的通径效应系数P_{ry}。

04.通径系数可进行如下分析

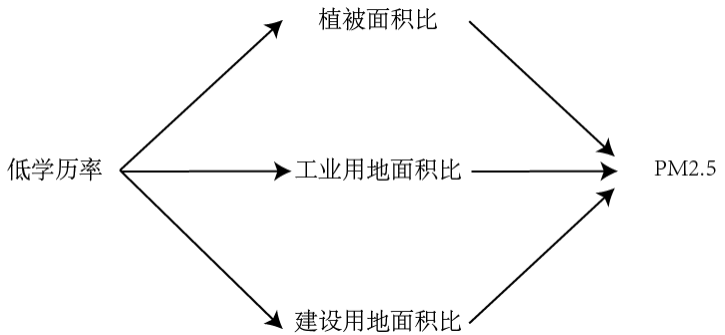
- 按绝对值的大小排列通径系数，用以说明每一通径效应对因变量Y的作用所占位置的相对重要性。
- 如果P_{iy}接近于r_{iy}说明r_{iy}反应了X_i和Y的真实关系，通过改变X_i的数量来改变Y是有效的。
- 如果r_{iy}大于0。但P_{iy}小于0。则说明间接效应是相关的主要原因，直接通过X_i改变Y是无效的，必须通过X_j方可成效。

结构方程模型

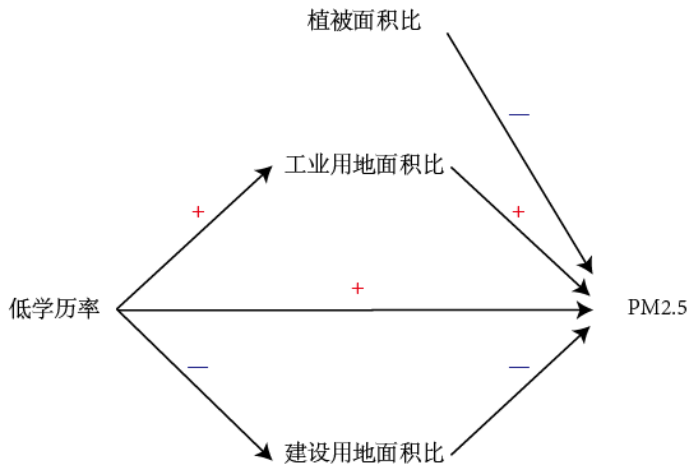
应用实例

借助结构方程模型分析空气污染暴露（本例中采用$PM_{2.5}$污染表征）、土地利用格局（本例采用植被覆盖率、工业用地覆盖率和建设用地覆盖率表征）和邻里社会经济地位（本例采用低学历率表征）三者之间的潜在关联机制。

Step 1：通过空气污染空间格局可以发现，研究区域内$PM_{2.5}$浓度并不是均匀分布的，不同社会经济地位的居民小区暴露$PM_{2.5}$浓度是不同的，这初步揭示了环境不平等现象的存在。
相关领域的文献指出，弱势群体（如低学历、低收入人群）聚集区的周边工业用地比重较高，绿化率较低，从而可能造成空气污染暴露。基于此，本例初步构建的三者关联假设模型如下图。



Step 2：在假设模型的基础上，准备好各小区评价指标数据，构建AMOS结构方程模型，并通过输出结果中提供的修正指标来修正假设关系。



时间序列分析

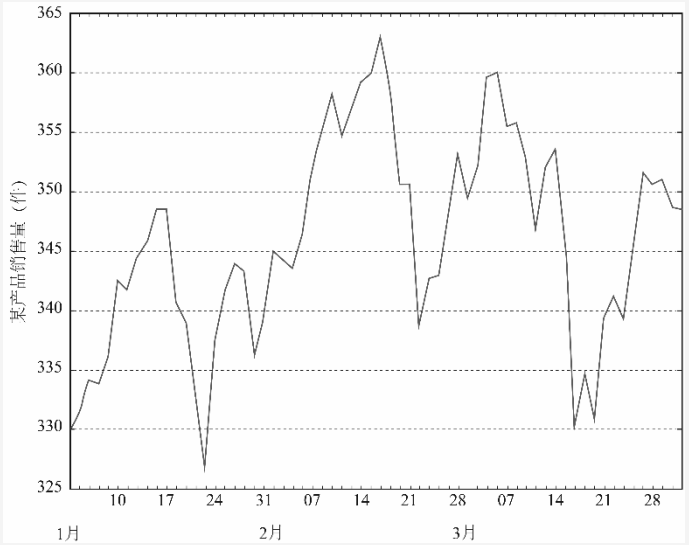
时间序列分析基本原理

01.定义

时间序列：同一现象在不同时间上的相继观察值排列而成的序列。

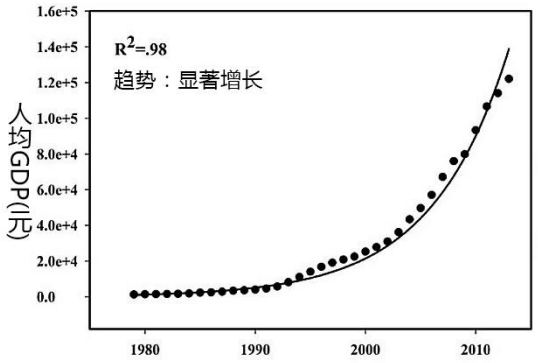
02.分类

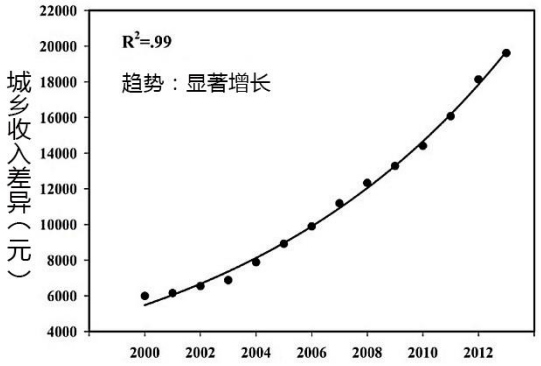
①平稳序列：各观察值基本上在某个固定的水平上波动，虽然在不同的时间段波动的程度不同，但并不存在某种规律，其波动可以看成是随机的。基本上不存在趋势。



03.趋势

时间序列在长时期内呈现出来的某种持续上升或持续下降的变动，也称长期趋势。





②非平稳序列：包含　　、　　或　　的序列，它可能只含有其中一种成分，也可能是几种成分的组合。

时间序列分析

时间序列分析的模型及检验

01.分类

按分析目的不同	时域分析	频域分析
序列	历史值的函数	不同频率的正弦或余弦波叠加的结果
分析目标	事物随时间发展变迁的趋势	频率特征

02.发展

早期
- 移动平均法
- 指数平滑法

1970后 ARIMA（求和自回归滑动平均模型）

03.分析内容

- 预测
- 缺失值填补
- ……

04.模型

- 指数平滑模型
- ARIMA模型
- 季节分解模型

指数平滑模型

性质 非线性

目标 使预测值和观测值之间的均方误差（MSE）达到最小

预测方程 $X_{t+1} = T_t$

把第t期的指数平滑值作为第t+1期的预测值

优点

- 继承了加权平均法重视近期数据的思想
- 克服了权重不易确定的局限性

01.简单加权平均

$$x_{t+1} = \frac{w_1 x_t + w_2 x_{t-1} + \ldots + w_t x_1}{w_1 + w_2 + \ldots + w_t}$$

缺点

- 主观性较大
- 计算复杂

02.自动加权平均

思路 自当前期向前，让各期权重按指数规律下降，为各期观测赋权重，使权重之和等于1；考虑t充分大的情形，把滞后1期的估计值单独提出。

$$T_t = \alpha x_t + (1-\alpha) T_{t-1}$$

a 平滑指数

T_t 时间序列xt第t期的指数平滑值

ARIMA模型

01.优势

以对含有季节成分的时间序列数据进行分析

02.主要参数

自回归阶数(p)
差分阶数(d)
移动平均阶数(q)
→ ARIMA(p,d,q)

03.差分

分类
- 一般性差分
- 季节性差分

算子
- 一阶差分 $\nabla y_t = (1-B)y_t = y_t - y_{t-1}$
- d阶差分 $\nabla^d y_t = \nabla(\nabla^{d-1} y_t) = (1-B)^d y_t$
- 季节差分 $\nabla_T y_t = y_t - y_{t-T}$

数据
- 非季节性数据 ⟶ 一阶差分
- 周期为12的季节性数据
  - 相加属性⟶差分算子$\nabla\nabla_{12}$
  - 相乘属性⟶差分算子∇_{12}^2
- 以季度为周期的数据⟶ 差分算子∇

ARIMA模型分类

	模型的一般形式	偏自相关函数	自相关函数
自回归模型	$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t$	截尾性	拖尾性
移动平均模型	$x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$	拖尾性	截尾性
自回归移动平均模型	$x_t = \theta_1 x_{t-1} + \theta_2 x_{t-2} + ... + \theta_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$	拖尾性	拖尾性

AR(p)模型、MA(q)模型都是ARMA(p,q)模型的特例，有 $\mathrm{AR}(p) = \mathrm{ARMA}(p,0), \mathrm{MA}(q) = \mathrm{ARMA}(0,q)$

时间序列分析

季节分解模型

01.定义

通过某些手段把时间序列中的4种变动趋势分解出来，并分别对其加以分析，再将分析结果综合起来组成个对原始时间序列的总模型。

02.时间序列的4种成分

- 长期趋势：表示序列取值随时间逐渐增加、减少或不变的长期发展趋势。
- 季节趋势：表示由于受到季节因素或某些习俗的影响，而出现的有规则的变化规律。
- 循环趋势：表示序列取值沿着趋势线有如钟摆般循环变动的规律。
- 不规则趋势：表示把时间序列中的长期趋势、季节趋势和循环趋势都去除后余下的部分。

季节分解模型的种类

	加法模型	**乘法模型**
各成分之间有无影响	无	有
模型	$Y=T+C+S+R$	$Y=T\times C\times S\times R$
季节因素、周期因素和不规则因素的取值	正值或负值	大于或小于1
反映	各因素对时间序列的影响方式和程度	各因素在长期趋势的基础上对原始序列的相对影响方式和程度

时间序列分析

MK检验

特点　非参数检验方法（无分布检验）

- 不需要样本遵从一定的分布
- 不受少数异常值的干扰
- 更适用于类型变量和顺序变量
- 计算比较简便。

计算

Step 1　趋势检验的统计量　$S=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{sgn}\left(x_j-x_i\right)$

当$n \geq 8$时，S 大致地服从正态分布，其均值为0，方差为：

$$\operatorname{Var}(S)=\frac{n(n-1)(2n+5)-\sum_{i=1}^{n} t_i(i-1)(2i+5)}{18}$$

Step 2　标准化统计量

$$Z_c=\begin{cases}\dfrac{S-1}{\sqrt{\operatorname{Var}(S)}}, & S>0\\ 0, & S=0\\ \dfrac{S+1}{\sqrt{\operatorname{Var}(S)}}, & S<0\end{cases}$$

标准正态分布

Step 3　衡量趋势大小

$$\beta=\operatorname{Median}\left(\frac{x_i-x_j}{i-j}\right)$$

- 正：上升
- 负：下降

Step 4　趋势检验

零假设H_0：β=0，当 $\left|Z_c\right|>Z_{(1-\alpha)/2}$ 时，拒绝零假设。

$Z_{(1-\alpha)/2}$　标准正态方差

α　显著性检验水平

Thank you