多元统计分析主要研究一个变量（因变量）和多个变量（自变量）之间是否有线性关系?如果有,那么如何由数据来估计这种关系的一种统计方法,是一元线性回归的扩展.在学习的时候,要弄清楚多元回归分析和一元回归分析在哪些地方是相同的,哪些是多元回归下才有的东西.下面我们主要介绍多元回归模型的定义,如何由数据对多元回归模型的参数进行估计,如何对参数进行检验,如何对检验因变量和自变量之间存在线性关系,如何选取和因变量存在显著线性关系的自变量,如何由估计的模型进行预测等.

1.多元回归模型

因变量 Y 与自变量$X_1, X_2, \cdots, X_p$ 之间的关系为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

其中ϵ 称为随机误差,且满足$E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 称为回归系数.

2.模型假设

- 若 $X_1, X_2, \cdots, X_p$为随机向量,则假设 $(X_1, X_2, \cdots, X_p)'$ 与 ϵ 相互独立且在$X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p$的条件下, $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$, 从而 $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
- 若$X_1, X_2, \cdots, X_p$为向量，则在 $X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p$条件下， $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$,从而 $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

假设因变量Y和自变量 $X_1, X_2, \cdots, X_p$样本数据如下：

$$\begin{array}{cc} Y & X_1, X_2, \cdots, X_p \\ y_1 & x_{11}, x_{12}, \cdots . x_{1p} \\ y_2 & x_{21}, x_{22}, \cdots . x_{2p} \\ \vdots & \vdots \\ y_n & x_{n1}, x_{n2}, \cdots . x_{np} \end{array}$$

则$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, i = 1, 2, \cdots, n$,其中随机误差 ϵ_i是不可观测的. 假设$\epsilon_1, \epsilon_2, \cdots, \epsilon_n \sim N(0, \sigma^2)$且相互独立.我们称该模型为多元回归模型或者Gauss-Markov模型。

令

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots & \cdots & \cdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

则$Y = X\beta + \varepsilon$.

根据模型假设我们有

- $E(\varepsilon) = 0;$
- $Var(\varepsilon) = \sigma^2 I;$
- $\varepsilon \sim N(0, \sigma^2 I).$

3.线性关系的诊断

如何确定 Y 与 $X_1, X_2, \cdots, X_p$ 之间是线性关系呢?

- 若 $p=1$ 时，可以用散点图和相关系数法给出初步的鉴别;
- 若 $p \geq 2$ 时,散点图法就失效了.但可以通过计算 Y 与 $X_1, X_2, \cdots, X_p$ 之间的相关系数来初步判定是否存在线性关系?如果 Y 和 $X_1, X_2, \cdots, X_p$ 之间存在显著地线性关系，则可以考虑用多元线性回归来对数据建模.

4.参数估计：最小二乘估计

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2$$

其中, $||.||$ 为n维向量的欧式长度.由数学分析或者线性代数的相关知识可得

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$\hat{\beta}$ 称为参数 β 的最小二乘估计.

5. Y 的拟合值,残差,随机误差方差的估计

- Y 的拟合值： $\hat{Y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n]^T = X\hat{\beta}$.
- 残差: $e_i = y_i - \hat{y}_i, i = 1, 2, \cdots, n$.
- 残差向量为

$$e = [e_1, e_2, \cdots, e_n]^T = Y - \hat{Y} = Y - X\hat{\beta}$$

- 随机误差 σ^2 方差的估计为： $\hat{\sigma}^2 = \frac{||e||^2}{n-p-1}$.

6.回归系数的显著性检验

对 $\forall, j = 1, 2, \cdots, p$,统计假设为

$$H_{j0} : \beta_j = 0 \quad vs \quad H_{j1} : \beta_j \neq 0$$

统计量为 $t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}$,其中 c_{jj} 为 $C = (X^T X)^{-1}$ 的对角线上第 j 个元素, $j = 1, 2, \cdots, p$.

若 H_{j0} 成立,则 $t \sim t(n-p-1)$.

7.回归方程显著性检验

我们通过 F 检验来说明 Y 和 $X_1, X_2, \cdots, X_p$ 之间是否存在线性关系. Y 和 $X_1, X_2, \cdots, X_p$ 存在线性关系可表述为 $\beta_1, \beta_2, \cdots, \beta_p$ 不全为0。因此 Y 和 $X_1, X_2, \cdots, X_p$ 之间是否存在线性关系可表述为以下统计假设

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \ \ vs \ \ H_1 : \beta_1, \beta_2, \cdots, \beta_p \text{不全为} 0$$

- 检验统计量为 $F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)}$,其中 $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $SSE = \sum_{i=1}^{n} e_i^2$;
- 在 H_0 成立的条件下, $F \sim F(p, n-p-1)$.

8.回归预测

回归预测分别为点预测和区间预测.点预测就是对因变量均值的预测.点预测很简单,就是把自变量的值直接带入估计的回归方程,便可得到因变量均值的预测估计.

- 令$x=(x_1,x_2,\cdots,x_p)^T$,那么$E(y)$的估计值为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$$

- **均值置信区间的估计**
  给定$x=(x_1,x_2,\cdots,x_p)^T$和显著性水平α,对$E(y)$的$(1-\alpha)100$%置信区间估计为

$$\hat{y}\pm t_{1-\alpha/2}(n-p-1)s_x$$

其中$s_x=\hat{\sigma}\sqrt{x^T(X^TX)^{-1}x}$.

- **点预测区间**
  给定$x=(x_1,x_2,\cdots,x_p)^T,y$是一个随机变量.给定显著性水平$\alpha,y$的$(1-\alpha)100$%预测区间为

$$\hat{y}\pm t_{1-\alpha/2}(n-p-1)s'_x$$

其中$s'_x=\hat{\sigma}\sqrt{1+x^T(X^TX)^{-1}x}$.

9.变量的选择

我们在对Y和$X_1,X_2,\cdots,X_p$的关系建模时,并不需要把所有自变量都选入模型.有些自变量对因变量线性影响比较小,因此需要对自变量进行选择,挑选那些对因变量影响显著的变量.我们纪要挑选那些对Y影响比较显著的自变量,尽可能地使得模型越简单越好,也要使得模型拟合数据的效果要好.因此需要有一个合理的标准,根据这个标准来评判所建的模型是否更优?

- AIC**信息准则**AIC信息准则即Akaike information criterion，是衡量统计模型拟合优良性的一种标准,由于它为日本统计学家赤池弘次创立和发展的,因此又称赤池信息量准则,它可以权衡所估计模型的复杂度和此模型拟合数据的优良性.

$$AIC=2p+n\ln(SSE/n)$$

AIC值越小,说明模型越简单并且拟合数据越好.

10.常见变量选择的方法

- **一切子回归法**
  对所有自变量的子集关于Y做回归建模,找到最小的AIC所对应的模型.假设自变量有个p,这样就需要建立2^p-1个模型,然后选出最小AIC所对应的模型.若p比较大时,该方法是不适合的.
- **前进法**
  从一个变量开始,逐步增加自变量,直至变量增加后AIC没有变小.
- **后退法**
  首先考虑所有变量，逐步减少变量，直至变量减少后AIC没有变小.
- **逐步回归法**
  把前进法和后退法结合起来的一种变量选择的方法.
  在实际应用,逐步回归法是经常使用的变量选择的方法.

11. 案例

根据下面数据回答下面问题

y	x1	x2	x3	x4
79220	14010	98	115	15
79670	13260	98	26	8
186320	81240	96	199	19
161945	46260	96	120	19
74570	15510	95	46	12
86120	15810	93	8	16
91520	20760	92	168	17
82820	20010	90	205	12
75620	16260	90	191	15
82220	16260	88	252	12
78020	14760	88	38	12
76370	14010	87	123	16
78020	14760	86	367	12
120570	43740	85	134	20
83270	16260	85	438	8
77570	16860	85	171	8
68420	11460	85	72	12
75320	14010	85	59	15
71120	11460	83	75	8
91520	22260	81	3	16
76220	12510	81	0	12
74420	12510	81	13	12
85220	17760	79	94	12
98570	22500	74	45	16
77420	12810	74	2	12
110720	35010	74	272	12
69020	11460	72	184	8
87920	19260	71	12	16
75770	13710	69	12	12
76520	20010	68	344	8
81620	17010	68	155	8
86570	14760	67	6	15
72170	14760	67	181	12
137570	46260	66	50	18
121320	23010	65	19	16
77570	17010	64	69	12

- （1）建立 y 关于 x_1, x_2, x_3, x_4 回归方程,并对回归方程和回归系数进行显著性检验；
- （2）采用逐步回归法建立 y 关于 x_1, x_2, x_3, x_4 线性回归方程,并对回归方程和回归系数进行显著性检验；
- （3）给定 $x_1 = 20000, x_2 = 85, x_3 = 290, x_4 = 20$,根据逐步回归建立的线性回归方程给出 y 的预测值以及 $E(y)$ 的95%的置信区间和 y 的95%的预测区间。

解:

- （1）
  R程序及结果

```
>data<-read.table("clipboard",header=T) #将数据读入到data中
>lma<-lm(y~x1+x2+x3+x4,data=data)
#建立y关于x1、x2、x3和x4的线性回归方程,数据为data
>summary(lma)   #模型汇总,给出模型回归系数的估计和显著性检验等
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = data)
Residuals:
 Min       1Q   Median       3Q      Max
-12924.2  -4588.1   -269.6   1756.2  25215.7
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48386.0620 11237.2882   4.306 0.000155 ***
x1              1.6831     0.1302    12.929   5.01e-14 ***
x2            -34.5520   130.2602  -0.265  0.792570
x3            -13.0004    13.7882  -0.943   0.353043
x4            808.3223   547.8017  1.476   0.150144

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7858 on 31 degrees of freedom
Multiple R-squared:  0.919,     Adjusted R-squared:  0.9086
F-statistic: 87.95 on 4 and 31 DF,  p-value: < 2.2e-16
```

结果分析：
回归方程为 $y = 48386.0620 + 1.6831x_1 - 34.5520x_2 - 13.0004x_3 + 808.3223x_4$.

回归方程的显著性检验：F值 $= 87.95, p$值 $< 2.2 \times 10^{-16} < 0.01$,因此 x_1, x_2, x_3, x_4 对 y 非常显著的线性影响.
回归系数 x_1, x_2, x_3, x_4 的 t 检验:

变量	x_1	x_2	x_3	x_4
p 值	5.01×10^{-14}	0.792570	0.353043	0.150144
t 值	12.929	-0.265	-0.943	1.476

若显著性水平为 $\alpha = 0.05$，那么从上面可知只有的 x_1 系数显著不为0.

- （2）
  R程序及结果

```
>lm.step<-step(lma,direction="both") #用“一切子集回归法”来进行逐步回归
Start:  AIC=650.41
y ~ x1 + x2 + x3 + x4
        Df   Sum of Sq       RSS      AIC
 - x2   1    4.3448e+06  1.9186e+09  648.49
- x3    1    5.4896e+07  1.9692e+09  649.43
<none>                   1.9143e+09  650.41
- x4    1    1.3445e+08  2.0487e+09  650.85
- x1    1    1.0323e+10  1.2237e+10  715.19

Step:  AIC=648.49
```

```
y ~ x1 + x3 + x4

       Df   Sum of Sq       RSS      AIC
- x3   1    6.2078e+07   1.9807e+09  647.64
<none>                   1.9186e+09  648.49
- x4   1    1.3011e+08   2.0487e+09  648.85
+ x2   1    4.3448e+06   1.9143e+09  650.41
- x1   1    1.0341e+10   1.2259e+10  713.26

Step:  AIC=647.64
y ~ x1 + x4
       Df   Sum of Sq       RSS      AIC
 <none>                   1.9807e+09  647.64
+ x3    1    6.2078e+07  1.9186e+09  648.49
+ x2    1    1.1527e+07  1.9692e+09  649.43
- x4    1    2.9640e+08  2.2771e+09  650.66
- x1    1    1.1654e+10  1.3635e+10  715.09
```

利用逐步回归得到最优回归模型，即y 关于 x_1,x_4 回归方程.

```
>summary(lm.step)
Call:
    m(formula = y ~ x1 + x4, data = data)
    Residuals:
Min     1Q Median     3Q    Max
-13632  -4759   -615   1761  25076
Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  42097.165   5265.218   7.995 3.18e-09 ***
x1              1.631      0.117  13.934 2.22e-15 ***
x4           1039.260    467.671   2.222   0.0332 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7747 on 33 degrees of freedom
Multiple R-squared:  0.9162,    Adjusted R-squared:  0.9111
F-statistic: 180.4 on 2 and 33 DF,  p-value: < 2.2e-16
```

结果分析：y 关于x_1,x_4 回归方程为$y=42097.165+1.631x_1+1039.260x_4$.

F检验：F值$=180.4$，p值$<2.2\times10^{-16}<0.01$,因此$x1,x4$ 对y 非常显著的线性影响。
回归系数t检验：

变量	x_1	x_4
t值	13.934	2.222
p值	2.22×10^{-15}	0.0332
若显著性水平为$\alpha=0.05$，那么从上面可值x_1,x_2的系数都显著不为0。		

- （3）
  R程序及结果

```
>preds<-data.frame(x1=20000,x4=20)  #给定解释变量x1和x4的值
```

```
>predict(lm.step,newdata=preds,interval="c",level=0.95)
#均值估计和均值的95%置信区间

     fit          lwr        upr
1 95493.09     88348.34    102637.8
>predict(lm.step,newdata=preds,interval="prediction",level=0.95)#预测值与预测区间

       fit       lwr        upr
1 95493.09    78187.28    112798.9
```

结果分析：均值估计值为 95493.09 ，均值95%的置信区间为 $[88348.34，102637.8]$ ，95%预测区间为 $[78187.28，112798.9]$ 。