学生学号	0121906280724	实验课成绩	

武汉理工大学

学 生 实 验 报 告 书

实验课程名称	自变量的选择
开 课 学 院	理学院
指导教师姓名	李丹
学 生 姓 名	张逸敏
学生专业班级	统计 2001

2022 -- 2023 学年 第 二 学期

实验教学管理基本规范

实验是培养学生动手能力、分析解决问题能力的重要环节；实验报告是反映实验教学水平与质量的重要依据。为加强实验过程管理，改革实验成绩考核方法，改善实验教学效果，提高学生质量，特制定实验教学管理基本规范。

1、 本规范适用于理工科类专业实验课程，文、经、管、计算机类实验课程可根据具体情况参照执行或暂不执行。

2、 每门实验课程一般会包括许多实验项目，除非常简单的验证演示性实验项目可以不写实验报告外，其他实验项目均应按本格式完成实验报告。

3、 实验报告应由实验预习、实验过程、结果分析三大部分组成。每部分均在实验成绩中占一定比例。各部分成绩的观测点、考核目标、所占比例可参考附表执行。各专业也可以根据具体情况，调整考核内容和评分标准。

4、 学生必须在完成实验预习内容的前提下进行实验。教师要在实验过程中抽查学生预习情况，在学生离开实验室前，检查学生实验操作和记录情况，并在实验报告第二部分教师签字栏签名，以确保实验记录的真实性。

5、 教师应及时评阅学生的实验报告并给出各实验项目成绩，完整保存实验报告。在完成所有实验项目后，教师应按学生姓名将批改好的各实验项目实验报告装订成册，构成该实验课程总报告，按班级交课程承担单位（实验中心或实验室）保管存档。

6、 实验课程成绩按其类型采取百分制或优、良、中、及格和不及格五级评定。

附表：实验考核参考内容及标准

	观测点	考核目标	成绩组成
实验预习	1. 预习报告 2. 提问 3. 对于设计型实验，着重考查设计方案的科学性、可行性和创新性	对实验目的和基本原理的认识程度，对实验方案的设计能力	20%
实验过程	1. 是否按时参加实验 2. 对实验过程的熟悉程度 3. 对基本操作的规范程度 4. 对突发事件的应急处理能力 5. 实验原始记录的完整程度 6. 同学之间的团结协作精神	着重考查学生的实验态度、基本操作技能；严谨的治学态度、团结协作精神	30%
结果分析	1. 所分析结果是否用原始记录数据 2. 计算结果是否正确 3. 实验结果分析是否合理 4. 对于综合实验，各项内容之间是否有分析、比较与判断等	考查学生对实验数据处理和现象分析的能力；对专业知识的综合应用能力；事实求实的精神	50%

实验课程名称：_实用回归分析_____

实验项目名称	自变量的选择			实验成绩	
实 验 者	张逸敏	专业班级	统计 2001	组　　别	
同 组 者	刘璇、马钟森、李耀祖、危景熙、焦鼎云			实验日期	2023 年 4 月 7 日

第一部分：实验数据与实验要求

1、　实验数据

选取数据集 peru.txt

```
> #导入数据
> peru <- read.delim("E:/R/Rwd/regression analysis/实验三/peru.txt")
> peru$prop<-peru$Years/peru$Age> attach(peru)
> x<-data.frame(Age,Years,prop,Weight,Height,Chin,Forearm,Pulse)
> y<-Systol
> head(x)
  Age Years      prop Weight Height Chin Forearm Pulse
1  21     1 0.04761905   71.0   1629  8.0     7.0    88
2  22     6 0.27272727   56.5   1569  3.3     5.0    64
3  24     5 0.20833333   56.0   1561  3.3     1.3    68
4  24     1 0.04166667   61.0   1619  3.7     3.0    52
5  25     1 0.04000000   65.0   1566  9.0    12.7    72
6  27    19 0.70370370   62.0   1639  3.0     3.3    72
> head(y)
[1] 170 120 125 148 140 106
```

对数据进行处理，依据数据说明选择

Age,Years,prop,Weight,Height,Chin,Forearm,Pulse,Systol 作为实验数据。

2、　实验要求：用 R 软件完成下列的计算分析：

（1）检测自变量是否存在多重共线性；

（2）对实验数据中的自变量分别用逐步回归法和子集最优法选择自变量，并对 R 软件中的每一步得到的结果加以解释，最后对比分析两种方

法的不同。

第二部分：实验过程记录

过程记录（包括操作的步骤或者代码，输出的结果或者图形）：

先检查整体的共线性，使用所有变量进行回归：

```
> #使用全部变量进行回归
> LM<-lm(y~.,x)
> #共线性
> library(car)
> vif(LM)
      Age     Years      prop    Weight    Height      Chin  Forearm     Pulse
3.168030 33.368288 23.446005  4.699351  1.886692  2.030225  2.574783  1.315549
 > sqrt(vif(LM))>2
   Age   Years    prop  Weight  Height    Chin Forearm   Pulse
 FALSE    TRUE    TRUE    TRUE   FALSE   FALSE   FALSE   FALSE
```

从结果可以看到，存在多重共线性。

一、前进法：

1.1 前进法原理：

前进法是决定备选自变量取舍的一种回归分析方法，其基本思路是从没有变量的模型开始，然后逐步将自变量添加到模型中。具体做法是：向前选择过程从模型中自变量个数为零开始，此时所有自变量都不在模型中。如果它们分别添加到模型中与因变量建立一元回归模型，我们选择小于 acrit（ α_{entry} ）值最低 p 值的一个自变量进入模型（即考虑回归模型中自变量的系数的显著性）。然后在一元回归模型的基础上，继续将剩余自变量分别添加到模型中建立二元回归模型，选择小于 acrit 值的最低 p 值的一个自变量进入模型。重复进行该过程，直到没有新的自变量可以添加。

1.2 编程实现：

在该实验中我们选取 acrit 值为 0.05，即 $\alpha_{\text{entry}} = 0.05$，每一步引入 p 值小于 0.05 的变量中 p 值最小的变量，直到没有变量满足添加的条件。

引进第一个变量：

```
> #前进法
> #第一次前进
> p1<-c()
> for(i in 1:8){
+   fit<-lm(y~x[,i])
+   pt<-summary(fit)$coefficients[2,4]
+
+   p1<-rbind(p1,pt)
+ }
> rownames(p1)<-c("p_Age","p_Year","p_prop","p_Weight","p_H
eight","p_Chin","p_Forearm","p_Pulse")
> p1
                [,1]
p_Age     0.9718297633
p_Year    0.5964186036
p_prop    0.0888139183
p_Weight  0.0006654447
p_Height  0.1801795967
p_Chin    0.3002709908
p_Forearm 0.0935587119
p_Pulse   0.4108898180
> which(p1==min(p1),arr.ind = TRUE)
         row col
p_Weight   4   1
> p1[which.min(p1)]
[1] 0.0006654447
```

由结果可以知道，$x_4(Weight)$ 的 p 值最小，且小于 0.05，因此将该变量引入模型中。

引进第二个变量：

```
> #第二次前进
> x4<-x[,4]
> x<-x[,-4]
> p2<-c()
```

```
> for(i in 1:7){
+   fit<-lm(y~x[,i]+x4)
+   pt<-summary(fit)$coefficients[2,4]
+
+   p2<-rbind(p2,pt)
+ }
> rownames(p2)<-c("p_Age","p_Year","p_prop","p_Height","p_C
hin","p_Forearm","p_Pulse")
> p2
                [,1]
p_Age     0.0831008499
p_Year    0.0043598571
p_prop    0.0006991283
p_Height  0.9024460380
p_Chin    0.2967489835
p_Forearm 0.9257371017
p_Pulse   0.8422073396
> which(p2==min(p2),arr.ind = TRUE)
       row col
p_prop   3   1
> p2[which.min(p2)]
[1] 0.0006991283
```

由结果可以知道$x_3(prop)$的 p 值最小，且小于 0.05，因此将该变量引入模型中。

引进第三个变量：

```
> #第三次前进
> x3<-x[,3]
> x<-x[,-3]
> p3<-c()
> for(i in 1:6){
+   fit<-lm(y~x[,i]+x3+x4)
+   pt<-summary(fit)$coefficients[2,4]
+
+   p3<-rbind(p3,pt)
+ }> rownames(p3)<-c("p_Age","p_Year","p_Height","p_Chin","
p_Forearm","p_Pulse")
> p3
              [,1]
p_Age     0.3120067
p_Year    0.2818559
p_Height  0.6250515
```

```
p_Chin    0.1534065
p_Forearm 0.4748916
p_Pulse   0.7927541
> which(p3==min(p3),arr.ind = TRUE)
      row col
p_Chin   4   1
> p3[which.min(p3)]
[1] 0.1534065
```

从结果可以看到，此时 p 值最小的变量为 $x_5(Chin)$，但其 p 值大于 0.05，因此停止引入变量，即用前进法得到模型为，以 $x_3(Weight), x_4(prop)$ 为自变量进行回归。

```
> #用 x3,x4 进行回归
> summary(lm(y~x3+x4))
Call:
lm(formula = y ~ x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-18.4330  -7.3070   0.8963   5.7275  23.9819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
x3          -26.7672     7.2178  -3.708 0.000699 ***
x4            1.2169     0.2337   5.207 7.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 9.777 on 36 degrees of freedom
Multiple R-squared:  0.4731,    Adjusted R-squared:  0.4438
F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06
```

从结果可以看到，前进法最终得到的方程为：

$$y = 60.8959 - 26.7672x_3 + 1.2169x_4$$

即：

$$Systol = 60.8959 - 26.7672\,prop + 1.2169Weight$$

且 t 检验以及 F 检验均显著，因此方程具有意义。再进行共线性检验：

```
> #共线性
> sqrt(vif(lm(y~x3+x4)))>2
   x3    x4
FALSE FALSE
```

可以看到此时不存在多重共线性。

二、后退法：

2.1 后退法原理：

后退法是所有变量选择过程中最简单的一种，其与前进法相反，基本思想是先用所有变量建立回归模型，然后再依次剔除变量。具体做法是：我们从包含所有自变量的模型开始，首先删除大于阈值 acrit($\alpha_{removal}$) 的最高 p 值的一个自变量，然后重新拟合模型，并删除剩余自变量中最不显著的自变量，删除标准依然是其 p 值大于 acrit。重复进行该过程，直至所有“不显著”的自变量被删除，从而完成变量选择过程，获得最佳模型。

2.2 编程实现：

依然选定 acrit 值为 0.05，即(α_{removal})=0.05，将不显著的变量进行剔除。

剔除第一个变量：

```
> #后退法
> x<-data.frame(Age,Years,prop,Weight,Height,Chin,Forearm,P
ulse)
> y<-Systol
> LM8<-lm(y~.,x)
> p<-summary(LM8)
> p
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-12.1891  -5.9730   0.1668   5.3803  14.6818

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  147.05587   48.21377   3.050 0.004750 **
Age          -1.10995    0.32011  -3.467 0.001610 **
Years         2.41595    0.79124   3.053 0.004711 **
prop       -113.54307   29.12747  -3.898 0.000505 ***
Weight        1.42678    0.42220   3.379 0.002031 **
Height       -0.03594    0.03604  -0.997 0.326657
Chin         -0.97163    0.72364  -1.343 0.189441
Forearm      -1.37112    0.96691  -1.418 0.166480
Pulse         0.11966    0.16695   0.717 0.479081
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.523 on 30 degrees of freedom
Multiple R-squared:  0.6664,    Adjusted R-squared:  0.5774
F-statistic:  7.49 on 8 and 30 DF,  p-value: 1.876e-05
> which.max(p$coefficients[,4])
Pulse
    9
> p$coefficients[which.max(p$coefficients[,4]),4]
[1] 0.4790808
```

从结果可以看到变量 $x_8(Pulse)$ 的 p 值最大，且大于 0.05，即该变量不显著，因此将该变量剔除。

剔除第二个变量：

```
> #第二次后退
> x<-x[,-8]
> LM7<-lm(y~.,x)
> p<-summary(LM7)
> p
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-12.7430  -5.8265   0.4344   5.5914  15.3665

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  153.68245   46.94640   3.274 0.002612 **
Age          -1.09745    0.31712  -3.461 0.001592 **
Years         2.35670    0.78072   3.019 0.005044 **
prop       -110.45540   28.58029  -3.865 0.000531 ***
Weight        1.46055    0.41626   3.509 0.001400 **
Height       -0.03713    0.03572  -1.040 0.306595
Chin         -1.03844    0.71196  -1.459 0.154743
Forearm      -1.14395    0.90629  -1.262 0.216277
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.456 on 31 degrees of freedom
Multiple R-squared:  0.6607,    Adjusted R-squared:  0.584
F-statistic: 8.622 on 7 and 31 DF,  p-value: 7.559e-06
> which.max(p$coefficients[,4])
Height
     6
> p$coefficients[which.max(p$coefficients[,4]),4]
[1] 0.306595
```

从结果可以看到变量 $x_5(Height)$ 的 p 值最大且大于 0.05，因此认为该变量不显著，将该变量剔除。

剔除第三个变量：

```
> #第三次后退
> x<-x[,-5]
> LM6<-lm(y~.,x)
> p<-summary(LM6)
> p
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-13.7352  -6.7084   0.6907   5.1713  15.4859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.3000    21.5299   5.123 1.39e-05 ***
Age          -1.0926     0.3175  -3.441 0.001631 **
Years         2.5523     0.7587   3.364 0.002005 **
```

```
prop        -116.6819    27.9808  -4.170 0.000217 ***
Weight         1.1681     0.3072   3.803 0.000608 ***
Chin          -0.8776     0.6958  -1.261 0.216343
Forearm       -0.8214     0.8526  -0.963 0.342561
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.466 on 32 degrees of freedom
Multiple R-squared:  0.6488,    Adjusted R-squared:  0.583
F-statistic: 9.854 on 6 and 32 DF,  p-value: 3.673e-06
> which.max(p$coefficients[,4])
Forearm
      7
> p$coefficients[which.max(p$coefficients[,4]),4]
[1] 0.3425611
```

从结果可以看到，变量$x_7(Forearm)$的 p 值最大且大于 0.05，因此认为该变量不显著，将该变量剔除。

剔除第四个变量：

```
> #第四次后退
> x<-x[,-6]
> LM5<-lm(y~.,x)
> p<-summary(LM5)
> p
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-14.520  -6.640  -1.093   4.893  16.366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  109.3590    21.4843   5.090 1.41e-05 ***
Age           -1.0120     0.3059  -3.308 0.002277 **
Years          2.4067     0.7426   3.241 0.002723 **
prop        -110.8112    27.2795  -4.062 0.000282 ***
Weight         1.0976     0.2980   3.683 0.000819 ***
Chin          -1.1918     0.6140  -1.941 0.060830 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

```
1

Residual standard error: 8.457 on 33 degrees of freedom
Multiple R-squared:  0.6386,    Adjusted R-squared:  0.5839
F-statistic: 11.66 on 5 and 33 DF,  p-value: 1.531e-06
> which.max(p$coefficients[,4])
Chin
   6
> p$coefficients[which.max(p$coefficients[,4]),4]
[1] 0.0608296
```

从结果可以看到，变量$x_6(Chin)$的 p 值最大，且大于 0.05，因此将该变量剔除。

剔除第五个变量：

```
> #第五次后退
> x<-x[,-5]
> LM4<-lm(y~.,x)
> p<-summary(LM4)
> p
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-16.890  -5.976   0.058   5.407  16.835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  116.8354    21.9797   5.316 6.69e-06 ***
Age           -0.9507     0.3164  -3.004 0.004971 **
Years          2.3393     0.7714   3.032 0.004621 **
prop        -108.0728    28.3302  -3.815 0.000549 ***
Weight         0.8324     0.2754   3.022 0.004742 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.795 on 34 degrees of freedom
Multiple R-squared:  0.5974,    Adjusted R-squared:   0.55
F-statistic: 12.61 on 4 and 34 DF,  p-value: 2.142e-06
> which.max(p$coefficients[,4])
Age
```

```
  2
> p$coefficients[which.max(p$coefficients[,4]),4]
[1] 0.004970545
```

从结果可以看到，变量$x_1(Age)$的 p 值最大，但小于 0.05，因此停止提出变量，最后得到的回归方程为：

$$y = 116.8354 - 0.9507x_1 + 2.3393x_2 - 108.0728x_3 + 0.8324x_4$$

即：

$$\text{Systol} = 116.8354 - 0.9507\,Age + 2.3393Years - 108.0728\,prop + 0.8324Weight$$

且所有变量均显著，再考虑多重共线性：

```
> sqrt(vif(LM4))>2
Age  Years   prop Weight
 FALSE   TRUE   TRUE  FALSE
```

可以看到此时存在多重共线性，这应该是可以预料到的，因为$prop = \frac{Years}{Age}$。

三、逐步回归法：

3.1 逐步回归法原理：

逐步回归是向后消除和向前选择的组合，在每个阶段可以添加或删除一个变量，使得变量有进有出。这解决了在流程早期添加或删除变量的变量在后来不能删除或添加的矛盾。

3.2 编程实现：

基于 AIC 准则：

首先考虑使用 R 语言自带的函数 step 进行逐步回归，该函数是以精确的 AIC 准则进行回归，考虑了模型的统计拟合度以及用来拟合的参数数目。AIC 值越小，模型越优，它说明模型用较少的参数获得了足够的拟

合度。step()函数不会评估所有可能模型的 AIC，但会使用搜索方法依次比较模型。

```
> #逐步回归法
> ##首先考虑基于标准方法(AIC 准则)
> x<-data.frame(Age,Years,prop,Weight,Height,Chin,Forearm,Pulse)
> y<-Systol
> LM<-lm(y~.,x)
> lstep<-step(LM,direction = "both")
Start:  AIC=174.9
y ~ Age + Years + prop + Weight + Height + Chin + Forearm + Pulse

         Df Sum of Sq    RSS    AIC
- Pulse    1     37.31 2216.4 173.56
- Height   1     72.23 2251.3 174.17
<none>                2179.1 174.90
- Chin     1    130.95 2310.1 175.18
- Forearm  1    146.06 2325.2 175.43
- Years    1    677.20 2856.3 183.46
- Weight   1    829.55 3008.7 185.48
- Age      1    873.30 3052.4 186.04
- prop     1   1103.76 3282.9 188.88

Step:  AIC=173.56
y ~ Age + Years + prop + Weight + Height + Chin + Forearm

         Df Sum of Sq    RSS    AIC
- Height   1     77.26 2293.7 172.90
- Forearm  1    113.91 2330.3 173.52
<none>                2216.4 173.56
- Chin     1    152.11 2368.5 174.15
+ Pulse    1     37.31 2179.1 174.90
- Years    1    651.50 2867.9 181.61
- Age      1    856.29 3072.7 184.30
- Weight   1    880.23 3096.7 184.61
- prop     1   1067.91 3284.3 186.90

Step:  AIC=172.9
y ~ Age + Years + prop + Weight + Chin + Forearm

         Df Sum of Sq    RSS    AIC
```

```
- Forearm  1    66.53 2360.2 172.01
- Chin     1   114.02 2407.7 172.79
<none>              2293.7 172.90
+ Height   1    77.26 2216.4 173.56
+ Pulse    1    42.35 2251.3 174.17
- Years    1   811.30 3105.0 182.71
- Age      1   848.93 3142.6 183.18
- Weight   1  1036.53 3330.2 185.44
- prop     1  1246.44 3540.1 187.83

Step:  AIC=172.02
y ~ Age + Years + prop + Weight + Chin

         Df Sum of Sq    RSS    AIC
<none>               2360.2 172.01
+ Forearm  1    66.53 2293.7 172.90
+ Height   1    29.88 2330.3 173.52
+ Pulse    1     9.84 2350.4 173.85
- Chin     1   269.48 2629.7 174.23
- Years    1   751.19 3111.4 180.79
- Age      1   782.65 3142.9 181.18
- Weight   1   970.26 3330.5 183.44
- prop     1  1180.14 3540.4 185.83
> summary(lstep)
Call:
lm(formula = y ~ Age + Years + prop + Weight + Chin, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-14.520  -6.640  -1.093   4.893  16.366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  109.3590    21.4843   5.090 1.41e-05 ***
Age           -1.0120     0.3059  -3.308 0.002277 **
Years          2.4067     0.7426   3.241 0.002723 **
prop        -110.8112    27.2795  -4.062 0.000282 ***
Weight         1.0976     0.2980   3.683 0.000819 ***
Chin          -1.1918     0.6140  -1.941 0.060830 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.457 on 33 degrees of freedom
```

```
Multiple R-squared:  0.6386,    Adjusted R-squared:  0.5839
F-statistic: 11.66 on 5 and 33 DF,  p-value: 1.531e-06
```

从结果可以看到依据 AIC 准则，依次删去了变量 Pulse，Height，Forearm，最后以 Age，Years，prop，Weight，Chin 建立了回归模型，而除了变量 Chin 意外其余所有变量均显著，可以看到这与前面的后退法所得到的结果类似，而后退法最后删除了不显著的变量 Chin，而根据 AIC 准则则保留了变量 Chin。但可以看到变量 Chin 在 0.1 的水平上是显著的。所以最后由基于 AIC 准则的逐步回归得到的回归模型为：

$$y = 109.3590 - 1.0120x_1 + 2.4067x_2 - 110.8112x_3 + 1.0976x_4 - 1.1918x_6$$

即：

$$\text{Systol} = 109.3590 - 1.0120\text{Age} + 2.4067\textit{Year} - 110.8112\,\textit{prop} + 1.0976\textit{Weight} - 1.1918\textit{Chin}$$

该回归模型的共线性：

```
> #共线性
> sqrt(vif(lstep))
>2
Age  Years   prop Weight   Chin
 FALSE   TRUE   TRUE  FALSE  FALSE
```

可以看到，此时模型存在多重共线性。

基于检验的方法：

由于基于 AIC 准则得到的回归模型 Chin 的 p 值大于 0.05，接下来考虑基于检验方法的逐步回归。

首先设置 $\alpha_{entry} = 0.05, \alpha_{removal} = 0.06$ 即若一个变量的 P 值小于 0.05 且是所有变量 P 值最小的我们才将其进行引入处理。但是在剔除变量时，如果变量的 P 值大于 0.06 且是最大值时我们就将其进行剔除处理。符合严进宽出准则。

引入第一个变量：

```
> ##因此考虑基于检验方法的逐步回归
> x<-data.frame(Age,Years,prop,Weight,Height,Chin,Forearm,P
ulse)> y<-Systol
> #引入第一个变量
> p1<-c()
> for(i in 1:8){
+   fit<-lm(y~x[,i])
+   pt<-summary(fit)$coefficients[2,4]
+
+   p1<-rbind(p1,pt)
+ }> rownames(p1)<-c("p_Age","p_Year","p_prop","p_Weight","
p_Height","p_Chin","p_Forearm","p_Pulse")
> p1
                [,1]
p_Age     0.9718297633
p_Year    0.5964186036
p_prop    0.0888139183
p_Weight  0.0006654447
p_Height  0.1801795967
p_Chin    0.3002709908
p_Forearm 0.0935587119
p_Pulse   0.4108898180
> which(p1==min(p1),arr.ind = TRUE)
         row col
p_Weight   4   1
> p1[which.min(p1)]
[1] 0.0006654447
```

由结果可得，变量$x_4(Weight)$的 p 值最小且小于 0.05，因此将该变量引入。

引入第二个变量：

```
> #引入第二个变量
> x4<-x[,4]
> x<-x[,-4]
> p2<-c()
> for(i in 1:7){
+   fit<-lm(y~x[,i]+x4)
+   pt<-summary(fit)$coefficients[2,4]
+
+   p2<-rbind(p2,pt)
+ }
> rownames(p2)<-c("p_Age","p_Year","p_prop","p_Height","p_C
```

```
hin","p_Forearm","p_Pulse")
> p2
                [,1]
p_Age     0.0831008499
p_Year    0.0043598571
p_prop    0.0006991283
p_Height  0.9024460380
p_Chin    0.2967489835
p_Forearm 0.9257371017
p_Pulse   0.8422073396
> which(p2==min(p2),arr.ind = TRUE)
      row col
p_prop   3   1
> p2[which.min(p2)]
[1] 0.0006991283
```

由结果可得，变量 $x_3(prop)$ 的 p 值最小且小于 0.05，因此将该变量引入。

检查引入变量是否符合条件：

```
> #检查引入变量
> summary(lm(Systol~prop+Weight))
Call:
lm(formula = Systol ~ prop + Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-18.4330  -7.3070   0.8963   5.7275  23.9819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
prop        -26.7672     7.2178  -3.708 0.000699 ***
Weight        1.2169     0.2337   5.207 7.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 9.777 on 36 degrees of freedom
Multiple R-squared:  0.4731,    Adjusted R-squared:  0.4438
F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06
```

可以看到引入的变量均显著，因此保留变量 $x_3(prop)$。

引入第三个变量：

```
> #引入第三个变量
> x3<-x[,3]
> x<-x[,-3]
> p3<-c()
> for(i in 1:6){
+   fit<-lm(y~x[,i]+x3+x4)
+   pt<-summary(fit)$coefficients[2,4]
+
+   p3<-rbind(p3,pt)
+ }> rownames(p3)<-c("p_Age","p_Year","p_Height","p_Chin","
p_Forearm","p_Pulse")
> p3
             [,1]
p_Age     0.3120067
p_Year    0.2818559
p_Height  0.6250515
p_Chin    0.1534065
p_Forearm 0.4748916
p_Pulse   0.7927541
> which(p3==min(p3),arr.ind = TRUE)
      row col
p_Chin   4   1
> p3[which.min(p3)]
[1] 0.1534065
```

可以看到此时变量 $x_6(Chin)$ 的 p 值最小为 0.1534065，但大于 0.05，因此停止变量引入。

可以看到最终得到的回归模型与前进法相同

```
> summary(lm(y~x3+x4))
Call:
lm(formula = y ~ x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-18.4330  -7.3070   0.8963   5.7275  23.9819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8959    14.2809   4.264 0.000138 ***
x3          -26.7672     7.2178  -3.708 0.000699 ***
x4            1.2169     0.2337   5.207 7.97e-06 ***
---
```

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

Residual standard error: 9.777 on 36 degrees of freedom
Multiple R-squared:  0.4731,    Adjusted R-squared:  0.4438
F-statistic: 16.16 on 2 and 36 DF,  p-value: 9.795e-06
```

因此得到的方程为：

$$y = 60.8959 - 26.7672x_3 + 1.2169x_4$$

即：

$$Systol = 60.8959 - 26.7672\,prop + 1.2169Weight$$

```
> #共线性
> sqrt(vif(lm(y~x3+x4)))>2
   x3    x4
FALSE FALSE
```

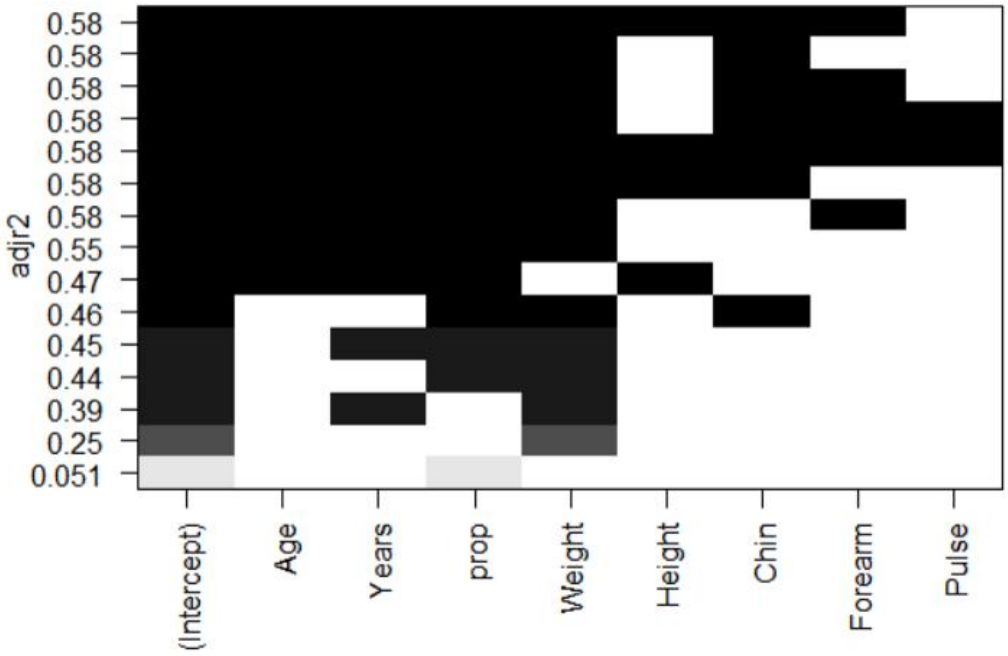
且此时不存在多重共线性。

四、全子集回归法：

虽然逐步回归可能会找到一个好的模型，但是不能保证模型就是最佳模型，因为不是每一个可能的模型都被评价了。为了克服这个限制，便有了全子集回归法。

顾名思义，全子集回归是指所有可能的模型都会被检验。在 R 语言中可以使用 leaps 包中的 regsubsets（）函数实现，可以通过 R 平方、调整 R 平方或 Mallows Cp 统计量等准则来选择最佳模型。

```
> #全子集回归
> library(leaps)
> x<-data.frame(Age,Years,prop,Weight,Height,Chin,Forearm,P
ulse)
> y<-Systol
> leaps<-regsubsets(y~.,x,nbest=2)
> plot(leaps,scale = "adjr2")
```



从结果可以看到，图中结果为选择对应的变量建立的回归模型对应的调整的 R 平方，由于选择 7 变量和 5 变量的模型所得到的调整的 R 平方差不多，因此我们趋向于选择更少的变量，因此以调整的 R 平方为标准得到的全子集回归模型为选择变量 Age,Years,prop,Weight,Chin 进行回归，得到回归模型：

```
> summary(lm(Systol~Age+Years+prop+Weight+Chin))
Call:
lm(formula = Systol ~ Age + Years + prop + Weight + Chin)

Residuals:
    Min      1Q  Median      3Q     Max
-14.520  -6.640  -1.093   4.893  16.366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  109.3590    21.4843   5.090 1.41e-05 ***
```

```
Age          -1.0120     0.3059  -3.308 0.002277 **
Years         2.4067     0.7426   3.241 0.002723 **
prop       -110.8112    27.2795  -4.062 0.000282 ***
Weight        1.0976     0.2980   3.683 0.000819 ***
Chin         -1.1918     0.6140  -1.941 0.060830 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 8.457 on 33 degrees of freedom
Multiple R-squared:  0.6386,    Adjusted R-squared:  0.5839
F-statistic: 11.66 on 5 and 33 DF,  p-value: 1.531e-06
```

此时与基于 AIC 准则得到的回归模型相同，为：

$$y = 109.3590 - 1.0120x_1 + 2.4067x_2 - 110.8112x_3 + 1.0976x_4 - 1.1918x_6$$

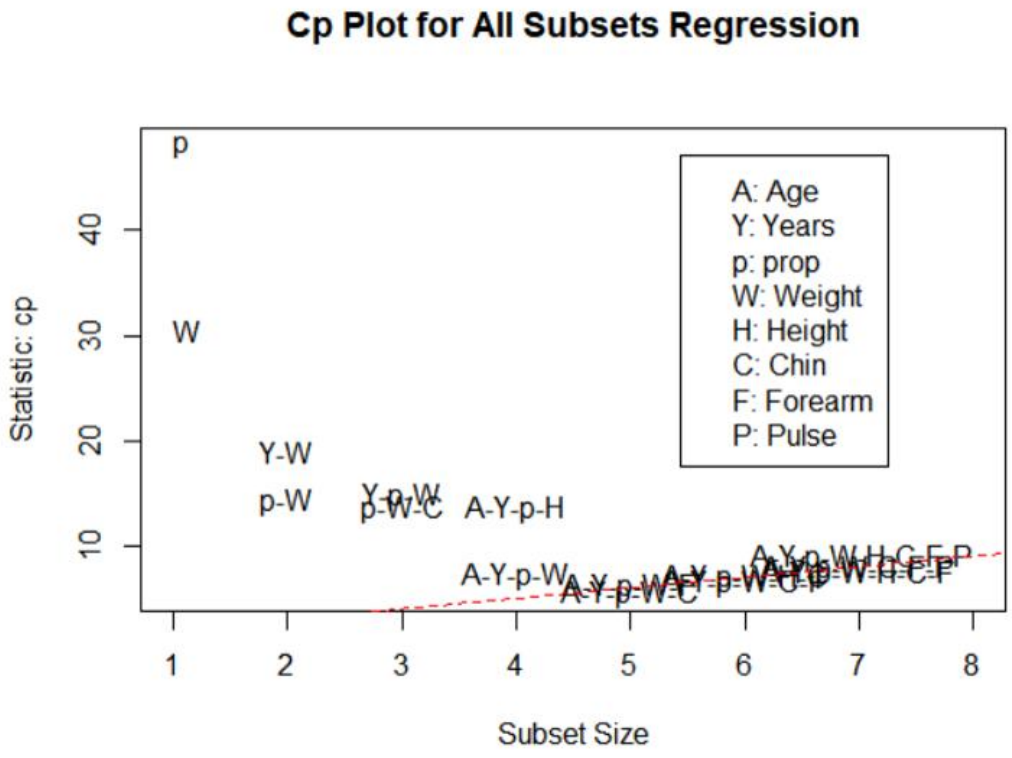
即：

$$\text{Systol} = 109.3590 - 1.0120\text{Age} + 2.4067\mathit{Year} - 110.8112\,\mathit{prop} + 1.0976\mathit{Weight} - 1.1918\mathit{Chin}$$

```
> sqrt(vif(lm(Systol~Age+Years+prop+Weight+Chin)))>2
   Age  Years   prop Weight   Chin
 FALSE   TRUE   TRUE  FALSE  FALSE
```

再以 Cp 准则进行全子集回归

```
> library(car)
> subsets(leaps,statistic = "cp",main="Cp Plot for All Subse
ts Regression")
> abline(1,1,lty=2,col="red")
```



在 图 中显示了对于同子集大小，基于 Mallows Cp 统计量的两个最佳模型。越好的模型离截距项和斜率项均为 1 的直线越近。因此，从上图可以看到，与使用调整的 R 平方为标准相同，此时变量增多时最优模型与红线差距不大，因此我们仍可以选择 5 变量模型作为我们的回归模型。

教师签字_____

第三部分　结果与讨论（可加页）

	回归模型	
前进法	$Systol = 60.8959 - 26.7672\,prop + 1.2169Weight$	
后退法	$\text{Systol} = 116.8354 - 0.9507\,Age + 2.3393Years - 108.0728\,prop + 0.8324Weight$	
逐步回归	AIC	$\text{Systol} = 109.3590 - 1.0120\text{Age} + 2.4067Year - 110.8112\,prop + 1.0976Weight - 1.1918Chin$
	p 值	$Systol = 60.8959 - 26.7672\,prop + 1.2169Weight$
全子集回归	$\text{Systol} = 109.3590 - 1.0120\text{Age} + 2.4067Year - 110.8112\,prop + 1.0976Weight - 1.1918Chin$	

1. 从各种回归方法得到的回归模型中可以看出，前进法与后退法得到的回归模型不同，而与逐步回归中基于检验方法得到的的回归模型相同，一方面这是由于前进法与后退法对变量只能进行引入与剔除，一旦将变量引入或者剔除，该变量就再也无法剔除或引入，而逐步回归虽然既能引入也能剔除，但每次也是针对一个变量，而逐步回归中选择的引入变量和剔除变量所对应的 p 值以及初始模型也对最终的模型有所影响，可以想到，如果初始选择的模型是所有变量进行的回归模型，最后得到的结果可能是与后退法相似的，同时如果对应的 p 值限制放大也会可能会得到不同的模型。另一方面，逐步回归也是前进法与后退法的结合，而用 p 值只考虑引入或剔除变量的显著性，如前进法得到的两变量模型和后退法得到的四变量模型得到的变量均显著，而由于每次只考虑一个变量，因此前进法再引入第三个变量对于严格的 p 值不显著，因此结束了

引入过程，得到了一个可能的最佳的模型，这也从侧面说明了逐步回归法可能会找到一个好的模型，但是不能保证模型就是最佳的模型，因为不是每一个可能的模型都被评价了（前进法和后退法得到的可能都是好的模型）

2. 而从全子集回归和 AIC 准则的得到的模型来看，后退法与两者选择的结果更接近，后退法剔除了不显著的变量 Chin，而全子集回归和 AIC 准则则是通过各自的标准得到的结果，全子集回归法相较于逐步回归来说，其计算量较大但却能得到一个最佳的模型，因为全子集回归法会将每一个可能的模型进行评估并给出最佳的回归模型。因此我们可能更倾向于选择四变量模型或五变量模型来建立我们的回归模型。

3. 在全子集回归模型中，尽管得到的结果是七变量模型更优，但由于其相较于五变量模型相差不大，因此我们倾向于选择变量较少的模型（这也蕴含了 AIC 准则，模型用较少的参数获得了足够的拟合度）。然而，在实际的情况中，尽管我们有多种变量选择的方法，但最终得到的模型有可能并不是我们想要的，因为在这些方法在满足相应条件时，可能会将我们所关注的因变量删去，或者是引入我们不需要或关注度很低的变量，因此，一般来说，变量自动选择应该被看做是对模型选择的一种辅助方法，而不是直接方法。拟合效果佳而没有意义的模型将毫无帮助。