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	线性回归模型的实现
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：_线性回归模型的实现___

实验项目名称	线性回归模型的实现			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、焦鼎云、危景熙			实验日期	2023.3.24

第一部分：实验要求

1、 在 data 文件夹中选择一个数据样本集;
2、 用 R 软件载入所需数据;
3、 拟合多元线性回归模型;
4、 多元线性回归模型的诊断;
5、 残差图分析;
6、 实验总结。
7、 评分参考标准

实验一		
评分内容	分值	得分
统计软件 R 的使用	10	
报告格式的规范程度	10	
原始数据的导入, 根据实际问题确定建模的因变量和自变量, 建立多元线性回归模型	10	
模型的检验, 包括各类决定系数, 有序平方和, 广义 F 检验, 失拟检验等。	20	
未通过检验的处理方法, 包括分析原因, 处理过程, 处理结果	5	
对输出结果中的各个量的解释	10	
利用残差图对模型的四个假设的直观检测	20	
加分项, 包括课堂之外新方法的使用, 对模型预测功能的探讨等	5	
实验总结, 包括实验收获、对未解问题的思考、对老师教学的建议	10	
总分	100	

第二部分：实验过程记录

过程记录：

一、实验原理

回归分析研究的主要对象是客观事物变量间的统计关系。它是建立在对客观事物进行大量实验和观察的基础上，用来寻找隐藏在看起来不确定的现象中的统计规律的统计方法。在回归分析中，解释变量称为自变量，被解释变量称为因变量，处于被解释的特殊地位。在回归分析中，因变量是随机变量，而自变量可以是随机变量，也可以是非随机变量。回归分析不仅可以揭示自变量对因变量的影响大小，还可以用回归方程进行预测和控制。

1.1 多元线性回归模型的一般形式

多元线性回归的方程如下：

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$$

其中 p 为自变量个数，β_0 为常数项，β_i 称为偏回归系数。当其它自变量固定时，β_i 反映了 x_i 每变动一个单位，y 的平均值变动的大小。

对一个实际问题，如果我们获得 n 组观测数据 $(x_{i1}, x_{i2}, \cdots, x_{ip};\ y_i)$，则线性回归模型可以表示为：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

写成矩阵形式，为：$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.

1.2 回归诊断

线性回归模型的成功建模，依赖于如下的假设：

(1) 线性模型假设：$y = X\beta + \varepsilon$

(2) 随机抽样假设：每个样本被抽到的概率相同且同分布

(3) 无完全共线性假设：X 满秩

(4) 严格外生性假设：$E(\varepsilon \mid X) = 0$

(5) 球形扰动项假设：$\mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n$

(6) 正态性假设：$\varepsilon \mid X \sim N(0, \sigma^2 I_n)$

回归诊断要研究的第一个问题就是考察我们的数据是否符合这些假设，如果假设不成立，探讨如何对数据进行修正以使它们（近似）满足这些假设。另一个重要问题是检测并处理对统计推断造成较大影响的数据点，也称为强影响点（influence case）。下面列举一些检验方法。

(1) 线性性检验。用残差-拟合值图像检验模型是否满足线性性假设，如果图像中显示残差对拟合值呈随机波动状，则说明模型拟合良好；若呈曲线状，则说明可

能需要在模型中增加一个二次项。

(2) 独立性检验。残差理论上是白噪声，不具有相关性。用 Durbin-Watson 检验：

H_0: 残差不存在自相关; H_1: 残差是相关的

$$\text{检验统计量}DW=\sum_{i=2}^{n}\frac{(\varepsilon_i-\varepsilon_{i-1})^2}{SSE}$$

如果$DW\approx 0$表示残差中存在正自相关; $DW\approx 4$表示残差中存在负自相关; $DW\approx 2$表示残差不存在自相关。

如果残差存在自相关性，则需要考虑给模型增加自回归项。

(3) 正态性检验。用学生化残差图进行检验。学生化残差公式如下：

$$SRE_i=\frac{e_i}{\sqrt{MS_{残}\cdot\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}}$$

用学生化残差图可以直观地判断误差项服从正态分布这一假定是否成立。若假定成立，学生化残差的分布也应服从标准正态分布。在学生化残差图中，大约有95.45%的学生化残差在$[-2,+2]$之间，这是因为$\Phi(2)-\Phi(-2)=0.9545$，$\Phi(\cdot)$为标准正态分布函数。

(4) 异方差检验。检验残差的异方差性，可用 Breusch-Pagan 检验，原假设是不存在异方差，备择假设是存在异方差。

异方差将导致回归系数的标准误估计错误，一种解决办法是使用稳健的标准误，另一种方法是在回归之前对数据y或x进行变换，实现方差稳定后再建模。原则上，当残差方差变化不太快时取开根号变换$\sqrt{y}$；当残差方差变化较快时取对数变换$\ln y$；当残差方差变化很快时取逆变换$1/y$；还有其他变换，如著名的 Box-Cox 变换或 Yeo-Johnson 变换，将非正态分布数据变换为正态分布。

(5) 共线性检验

多元线性回归建模，若自变量数据之间存在较强的线性相关性，即存在多重共线性。

多重共线性，会导致回归模型不稳定，这样得到的回归模型，是伪回归模型，就是并不反映自变量与因变量的真实影响关系。

共线性诊断可以用方差膨胀因子 VIF 来确定，若$VIF>10$，则判断有严重多重共线性。VIF 计算公式如下：

$$VIF=\frac{1}{1-R_j^2}$$

其中R_j是x_j和其余变量的决定系数。

多重共线性的解决办法：若两变量线性相关系数较大，则只用其中一个变量；采用逐步回归剔除冗余变量；主成分回归；岭回归。

1.3 异常观测值检验

一个全面的回归分析要覆盖对异常值的分析，包括离群点、高杠杆值点和强影响点。这些数据点需要更深入的研究，因为它们再一定程度上与其它观测点不同，

可能对结果产生较大的负面影响。

我们可以通过观察“残差与杠杆图”(Residuals vs Leverage)关注的单个观测点的信息，并从图形可以鉴别出离群点、高杠杆值点和强影响点。若一个观测点是离群点,表明拟合回归模型对其预测效果不佳(产生了巨大的或正或负的残差);若一个观测点有很高的杠杆值,表明它是一个异常的预测变量值的组合,因变量值不参与计算一个观测点的杠杆值;若一个观测点是强影响点(influential observation)，表明它对模型参数的估计产生的影响过大,非常不成比例。强影响点可以通过 Cook 距离即 Cook'sD 统计量来鉴别。

当判断出异常点后，可以通过删除异常观测点的方法提高模型拟合效果，删除离群点通常可以提高数据集对于正态假设的拟合度，而强硬想点会干扰结果，通常也会被删除。删除最大的离群点和强影响点后，模型需要重新拟合。若离群点或强影响点仍然存在，重复以上过程直至获得比较满意的拟合。

删除异常观测值点需要持有谨慎的态度。如果观测点是由于记录错误等原因造成，那么可以放心删去；但如果是其它情况，那异常点就有可能蕴含这一些没有注意到的规律，需要更深入地进行研究。

二、一元线性回归模型

由于多元回归模型一般存在多重共线性问题，本文首先建立一元回归模型，在第三部分“三、多元线性回归模型”中探讨多元线性回归，并比较两者区别。

2.1 数据导入与初步建模

导入 peru.txt 数据,并展示部分数据。

```
>peru <- read.table("E:/R/Rwd/regression analysis/data/peru.txt",header = T)>
>peru[1:10,]
```

	Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	Diastol
1	21	1	71.0	1629	8.0	7.0	12.7	88	170	76
2	22	6	56.5	1569	3.3	5.0	8.0	64	120	60
3	24	5	56.0	1561	3.3	1.3	4.3	68	125	75
4	24	1	61.0	1619	3.7	3.0	4.3	52	148	120
5	25	1	65.0	1566	9.0	12.7	20.7	72	140	78
6	27	19	62.0	1639	3.0	3.3	5.7	72	106	72
7	28	5	53.0	1494	7.3	4.7	8.0	64	120	76
8	28	25	53.0	1568	3.7	4.3	0.0	80	108	62
9	31	6	65.0	1540	10.3	9.0	10.0	76	124	70
10	32	13	57.0	1530	5.7	4.0	6.0	60	134	64

根据数据说明,删去舒张压（Diastol）数据并添加城市地区生活比例数据（proportion）.

```
> peru$Diastol<-NULL
> peru$proportion<-peru$Years/peru$Age
> peru[1:10,]
```

	Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	proportion
1	21	1	71.0	1629	8.0	7.0	12.7	88	170	0.04761905
2	22	6	56.5	1569	3.3	5.0	8.0	64	120	0.27272727
3	24	5	56.0	1561	3.3	1.3	4.3	68	125	0.20833333
4	24	1	61.0	1619	3.7	3.0	4.3	52	148	0.04166667

5	25	1	65.0	1566	9.0	12.7	20.7	72	140	0.04000000
6	27	19	62.0	1639	3.0	3.3	5.7	72	106	0.70370370
7	28	5	53.0	1494	7.3	4.7	8.0	64	120	0.17857143
8	28	25	53.0	1568	3.7	4.3	0.0	80	108	0.89285714
9	31	6	65.0	1540	10.3	9.0	10.0	76	124	0.19354839
10	32	13	57.0	1530	5.7	4.0	6.0	60	134	0.40625000

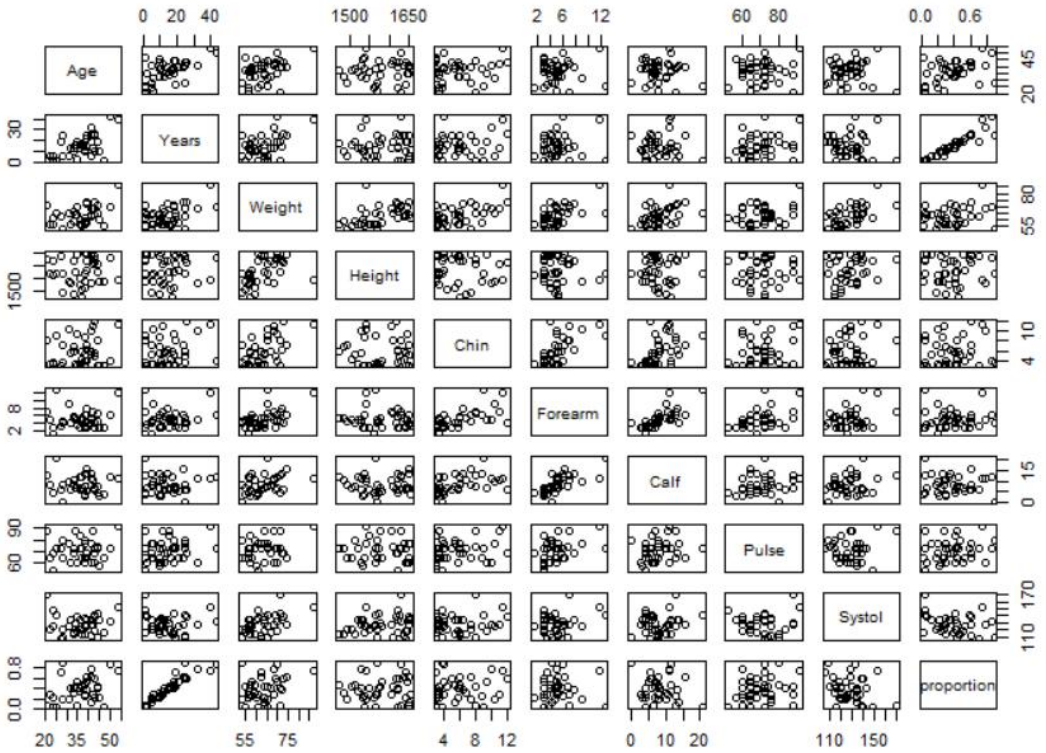
得到新的数据集,观察变量间的相关性,使用 cor()函数得到变量之间的相关系数矩阵。

```
> attach(peru)
> cor(peru)
```

	Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	proportion
Age	1.000000000	0.588212502	0.4316630	0.055777982	0.157908294	0.05520278	-0.005374411	0.090654502	0.005844807	0.36452333
Years	0.588212502	1.000000000	0.4811534	0.072594154	0.221697674	0.14302404	0.001099438	0.236904643	-0.087480460	0.93814540
Weight	0.431662982	0.481153366	1.0000000	0.450330307	0.561748764	0.54373244	0.391865474	0.311793359	0.521364290	0.29308303
Height	0.055777982	0.072594154	0.4503303	1.000000000	-0.007898078	-0.06893212	-0.002845856	0.007829993	0.219114553	0.05118739
Chin	0.157908294	0.221697674	0.5617488	-0.007898078	1.000000000	0.63788150	0.515999762	0.223100921	0.170192453	0.12009179
Forearm	0.055202779	0.143024038	0.5437324	-0.068932124	0.637881501	1.00000000	0.735525936	0.421907596	0.272280231	0.02801564
Calf	-0.005374411	0.001099438	0.3918655	-0.002845856	0.515999762	0.73552594	1.000000000	0.208715412	0.250789289	-0.11301572
Pulse	0.090654502	0.236904643	0.3117934	0.007829993	0.223100921	0.42190760	0.208715412	1.000000000	0.135477107	0.21419518
Systol	0.005844807	-0.087480460	0.5213643	0.219114553	0.170192453	0.27228023	0.250789289	0.135477107	1.000000000	-0.27614565
proportion	0.364523334	0.938145398	0.2930830	0.051187387	0.120091791	0.02801564	-0.113015720	0.214195184	-0.276145651	1.00000000

使用 plot()函数画两两变量之间的散点图。

```
> plot(peru)
```



通过自相关系数矩阵和散点图可以看到,因变量（收缩压 Systol）与体重（Weight）相关性比较大,相关系数为 0.52,而其他变量与因变量的相关系数均在 0.3 以下。

因此,先考虑通过体重（Weight)进行一元线性回归。

```
> LM1<-lm(Systol~Weight); LM1

Call:
lm(formula = Systol ~ Weight)

Coefficients:
(Intercept)      Weight
   66.5969       0.9629
```

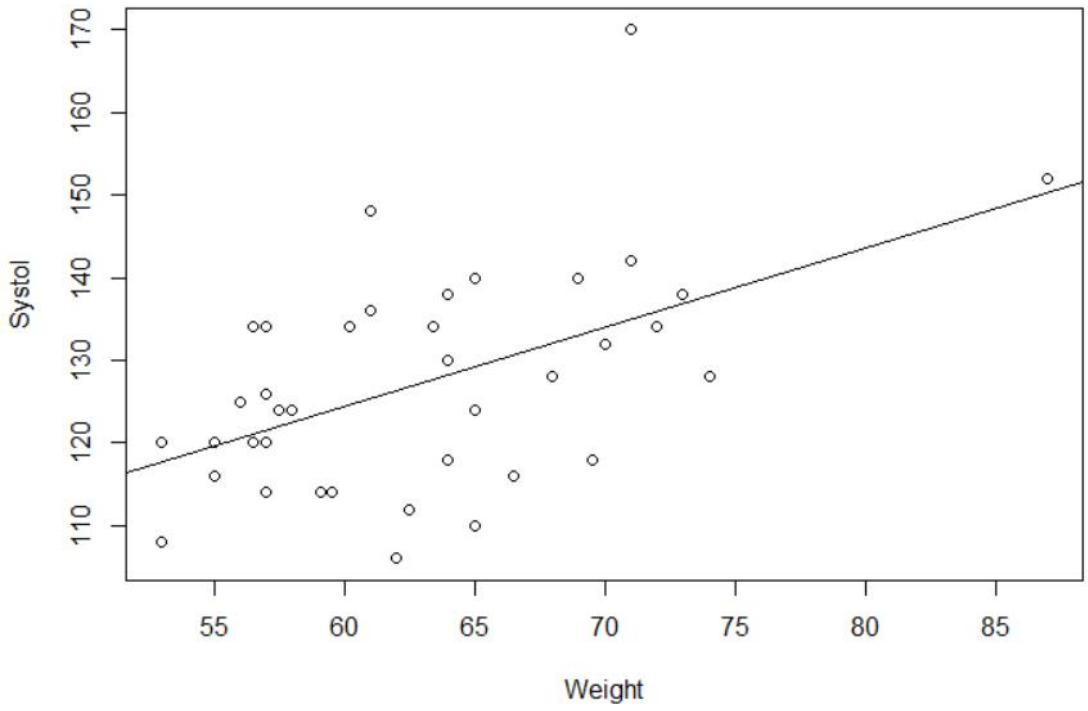
从输出结果可以看到，截距项的估计值$\hat{\beta}_0$为 66.5969，斜率项的估计值$\hat{\beta}_1$为 0.9629，即：

$$\widehat{Systol} = 66.5969 + 0.9629Weight$$

这说明体重每增加 1 公斤时,收缩压增加 0.9629（mmHg）。

下面将体重（Weight）和收缩压（Systol）的散点图和拟合直线图画出来。

```
> plot(Weight,Systol)
> abline(LM1)
```



2.2 回归诊断

2.2.1 相关系数检验

首先先对自变量和因变量的相关关系 r 进行检验。r 与其他统计指标一样,也有抽样误差,从同一总体内抽取若干大小相同的样本,各样本的相关系数总有波动,因此需判断不等于 0 的 r 值是来自总体相关系数$\rho = 0$还是$\rho \neq 0$的总体,由于来自$\rho = 0$的总体所有样本相关系数呈对称分布,故 r 的显著性可用 t 检验进行。

进行显著性检验步骤：

（1）建立检验假设,H_0: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = 0.05$

（2）计算相关系数 r 的 t 值。T 统计量的计算公式如下：

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

（3）计算 t 值和 p 值

```
> r<-cor(Weight,Systol)
> n<-length(Weight)
> tr<-r/sqrt((1-r^2)/(n-2))
> tr
[1] 3.716405
> p<-2*(1-pt(tr,n-2))
> p
[1] 0.0006654447
```

也可以用 cor.test()函数进行检验

```
> cor.test(Weight,Systol)

        Pearson's product-moment correlation

data:  Weight and Systol
t = 3.7164, df = 37, p-value = 0.0006654
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2463759 0.7186619
sample estimates:
      cor
0.5213643
```

以上结果显示,t 统计量的值为 3.7164, p 值为 0.0006654

所以在显著性水平$\alpha = 0.01$ 下, p 值小于 0.01,拒绝原假设,接受备择假设,有充足的证据来支持总体中收缩压（Systol）和体重（Weight）之间的线性关系,二者高度相关。

2.2.2 T 检验、F 检验和决定系数

summary()函数提供了最小值、最大值、四分位数和数值型变量的均值, 以及频数统计, 因此调用该函数获取相关描述性统计量。

```
> summary(LM1)

Call:
lm(formula = Systol ~ Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-20.294  -8.491   0.446   6.662  35.040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.5969    16.4639   4.045 0.000255 ***
Weight        0.9629     0.2591   3.716 0.000665 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.34 on 37 degrees of freedom
Multiple R-squared:  0.2718,    Adjusted R-squared:  0.2521
```

```
F-statistic: 13.81 on 1 and 37 DF,  p-value: 0.0006654
```

上述结果显示出了线性回归的系数常数项（Intercept）估计值$\hat{\beta}_0$为 66.5969,自变量（Weight）的斜率项的估计值$\hat{\beta}_1$为 0.9629,以及标准差, t 值和 p 值。

Multiple R-squared, 为相关系数R^2的检验,越接近 1 则越显著。

Adjusted R-squared, 为相关系数的修正系数, 解决多元回归自变量越多,相关系数R^2越大的问题。

F-statistic 表示 F 统计量, 自由度为(1,n-2)。

p-value 用于 F 检验判定, 匹配显著性标记。

从上表可以得到：

① T 检验法：

T 检验是检验模型某个自变量对于因变量的显著性（单参数检验）,通常用 P-value 判断显著性, 对于一元线性回归来说, 原假设和备择假设如下：

$$H_0：\ \beta_1=0；\ H_1:\beta_1\neq 0$$

检验统计量：$t=\dfrac{\hat{\beta}_1\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})}}{\sqrt{MSE}}\sim t(n-2)$

由上表结果可以得到 t=3.716,p=0.000665,远远小于 0.01,所以拒绝原假设,接受备择假设。

② F 检验法：

F 检验用于对所有的自变量在整体上看对于因变量的线性显著性（多参数检验）, 是对线性回归模型整体进行检验。

原假设和备择假设为：$H_0:\beta_0=\beta_1=0;\ H_1:\exists\,\beta_i\neq 0$

检验统计量：$F=\dfrac{\mathrm{SS}R/(p-1)}{\mathrm{SS}E/(n-p)}\sim F(p-1,n-p)$

由上表结果可以得到 F=13.81, p=0.000654。可以看到相关系数检验、T 检验、F 检验在一元线性回归时等价。

③ 决定系数：

决定系数为$R^2=SSR/SST$, 它用来判断回归方程的拟合程度,其取值是$[0,1]$, 越接近 1 说明拟合程度越好。

由上表可得决定系数$R^2=0.2718$,调整后的决定系数$Adj-R^2=0.2521$。

说明自变量体重（Weight）可以解释因变量收缩压（Systol）27.18%的波动,拟合效果较差。

2. 2. 3 失拟检验

失拟检验需要重复的观测值（称之为“复制”）,自变量的值至少重复一次,也就是说,若数据集中的每个 x 的值是独一无二的,那么失拟检验就不能进行。

H_0:模型合理，不存在失拟。H_1:模型不合理，存在失拟。

统计量：$F=\dfrac{MSLF}{MSPE}=\dfrac{SSLF/(c-2)}{SSPE/(n-c)}\sim F(c-2,n-c)$

```
> Full<-lm(Systol~as.factor(Weight))
> anova(LM1,Full)

Analysis of Variance Table
Model 1: Systol ~ Weight

Model 2: Systol ~ as.factor(Weight)
```

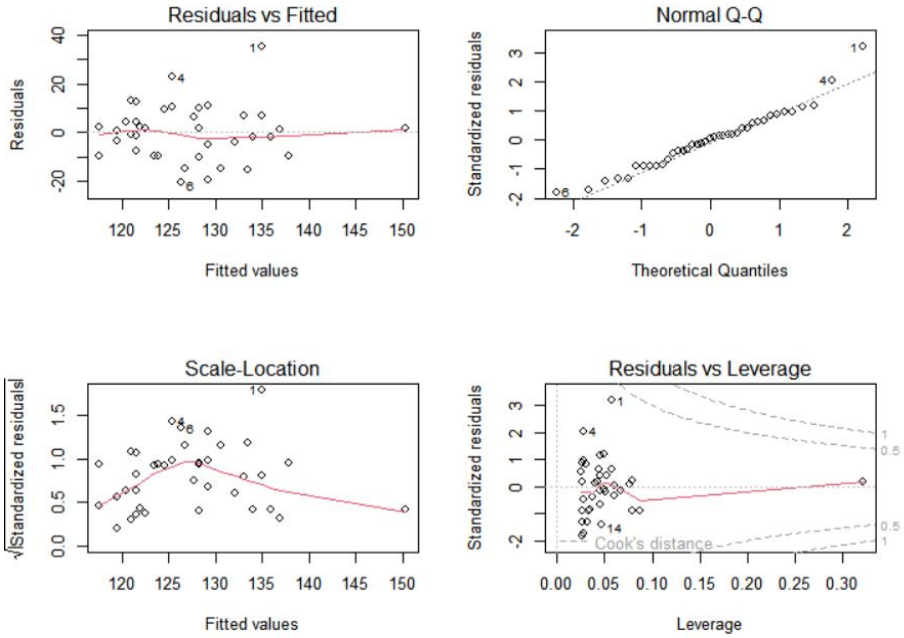
```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     37 4756.1
2     13 1514.3 24    3241.7 1.1595 0.4018
```

由上图可以得到,F 的值为 1.1595,p 值为 0.4018>0.05,因此无法拒绝原假设,即接受模型合理,不存在失拟。

2.2.4 模型适用性检验

R 基础安装中提供了大量检验回归分析中统计假设方法。最常见的方法就是对 lm()函数返回的对象使用 plot()函数,可以生成评价模型拟合情况的四幅图形。

```
> opar<-par(no.readonly = TRUE)
> par(mfrow=c(2,2))
> plot(LM1)
> par(opar)
```



① 从左上的残差与拟合值点图可以看到:

残差值在等于 0 的的直线上上下随机波动,说明自变量与因变量线性相关。

同时残差基本都落在以残差为 0 为中心的带形区域内,而且从左下的位置尺度图中,水平线周围的点基本上符合随机分布,说明误差项的方差满足同方差的假设。

红线呈现出一条平稳的曲线,且靠近 0,没有明显的形状,因此可以看到残差数据表现良好。

② 右上的正态 Q-Q 图中,残差基本都落在直线上,说明残差满足正态性假设。

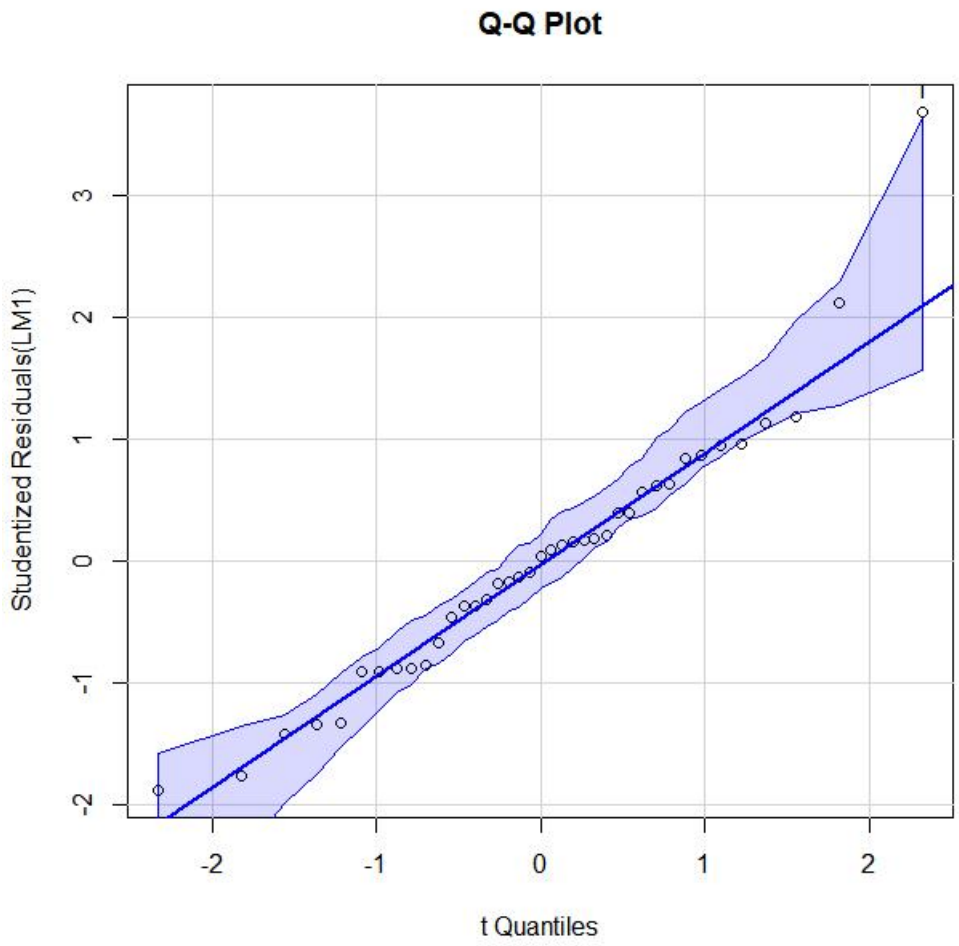
③ 右下的“残差与杠杆图”提供了单个观测点的信息,可以鉴别离群点、高杠杆值点,强影响点。

从上面可以看出,残差应该满足线性模型的四个基本假设。

除此以外,在 R 中 car 包提供了大量函数,也可以进行模型适用性检验,下面通过 car 包来进行检验:

① 正态性:

```
> library(car)
> qqPlot(LM1,labels=row.names(peru),id.method="identify",simulate=TRUE,main="Q-Q Plot")
```



从图中可以看到除了第 1 个观测值外,其他点都离直线很近,且落在置信区间中,这表明正态性假设符合得很好,而 1 为异常值点。

② 误差的独立性

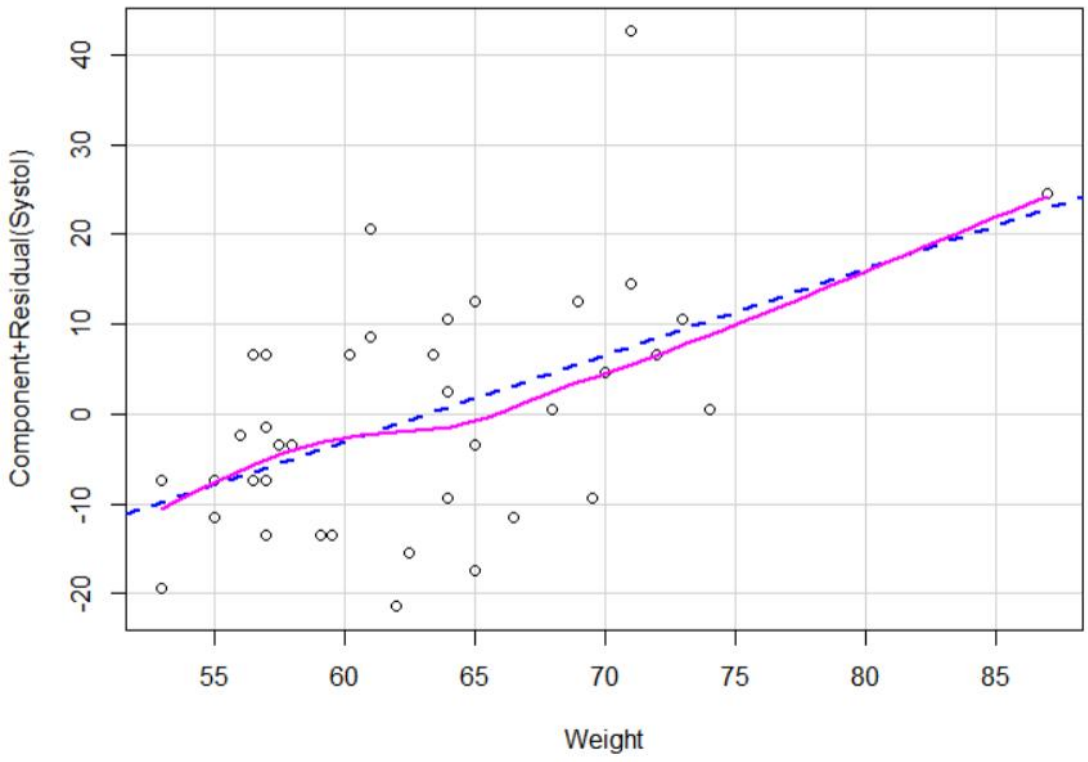
误差的独立性可以通过 Durbin-Watson 检验（D-W 检验）来检测误差的序列相关性。

```
> durbinWatsonTest(LM1)
 lag Autocorrelation D-W Statistic p-value
   1     -0.02276777       1.78682    0.42
 Alternative hypothesis: rho != 0
```

可以看到 p 值为 0.42,说明不拒绝原假设,即接受序列不相关,误差具有独立性。

③ 线性：

```
> crPlots(LM1)
```



从上图可以看出,线性假设是成立的。

④ 同方差性：

ncvTest()函数生成一个计分检验,零假设为误差方差不变,备择假设为误差方差随着拟合值水平的变化而变化。若检验显著，则说明存在异方差性（误差方差不恒定）。

```
> ncvTest(LM1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.175396, Df = 1, p = 0.2783
```

可以看到 p 值为 0.2783（>0.05）,说明满足方差不变假设。

综上,通过检验可以得到残差满足线性模型的四个基本假设,线性模型假设得到验证。

2. 2. 5. 异常值检验

一个全面的回归分析要覆盖对异常值的分析，包括离群点、高杠杆值点和强影响点。这些数 据点需要更深入的研究，因为它们在一定程度上与其他观测点不同，可能对结果产生较大的负面影响。

同时在模型适用性检验的过程中，我们也发现观测 1 和 4 可能是异常点，因此，接下来进行异常值检验。

① 离群点：

离群点是指那些模型预测效果不佳的观测点。它们通常有很大的、或正或负的

残差($Y_i - \hat{Y}_i$)。正的残差说明模型低估了响应值,负的残差则说明高估了响应值。car 包也提供了一种离群点的统计检验方法。outlierTest()函数可以求得最大标准化残差绝对值 Bonferroni 调整后的 p 值。

```
> outlierTest(LM1)
  rstudent unadjusted p-value Bonferroni p
1 3.685796        0.00074581     0.029086
```

可以看到 p 值为 0.029086（<0.05),较为显著,说明具有离群点,且离群点为第 1 个观测。

删去该离群点后在进行检验：

```
> outlierTest(lm(Systol[2:39]~Weight[2:39]))
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
3 2.598498         0.013611      0.51723
```

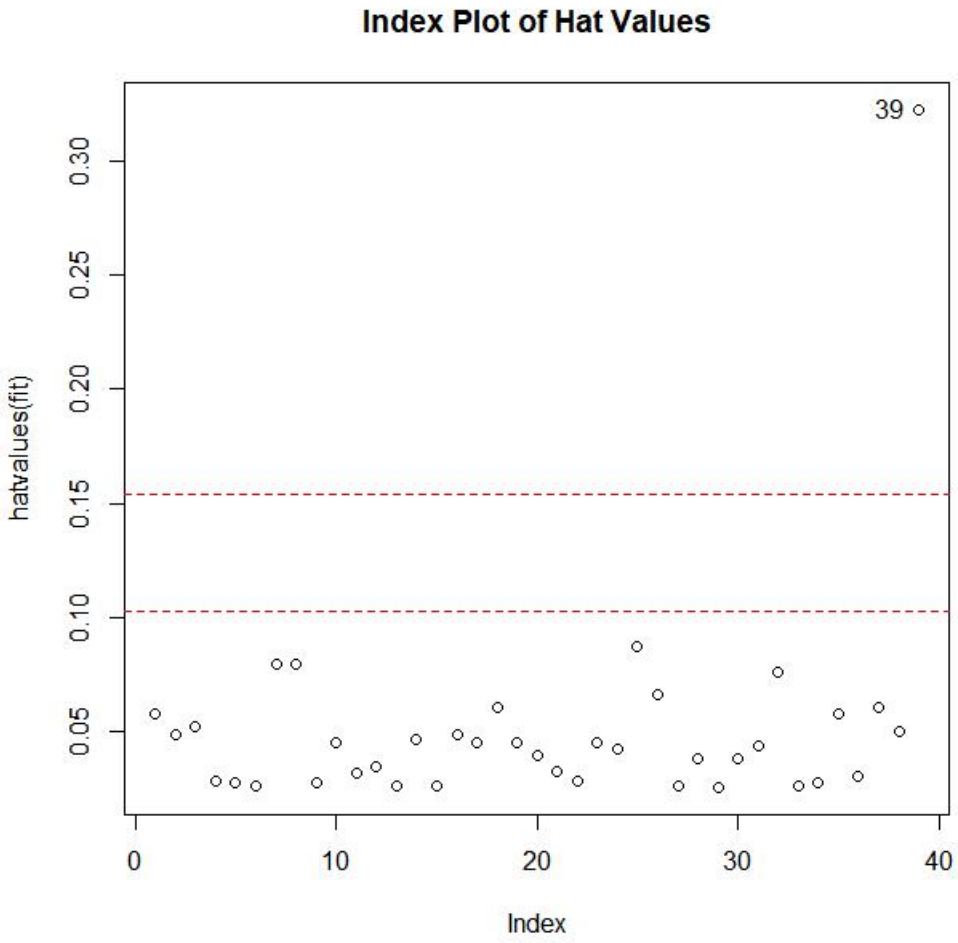
可以看到,此时数据集中不再存在离群点。

② 高杠杆值点

高杠杆值观测点,即与其他预测变量有关的离群点。换句话说,它们是由许多异常的预测变 量值组合起来的,与响应变量值没有关系。

高杠杆值的观测点可通过帽子统计量（hat statistic）判断。对于一个给定的数据集,帽子均值为 p/n,其中 p 是模型估计的参数数目（包含截距项）,n 是样本量。一般来说,若观测点的帽子值大于帽子均值的 2 或 3 倍,就可以认定为高杠杆值点。

```
> hat.plot <- function(fit) {
+   p <- length(coefficients(fit))
+   n <- length(fitted(fit))
+   plot(hatvalues(fit), main="Index Plot of Hat Values")
+   abline(h=c(2,3)*p/n, col="red", lty=2)
+   identify(1:n, hatvalues(fit), names(hatvalues(fit)))
+ }
> hat.plot(LM1)
```



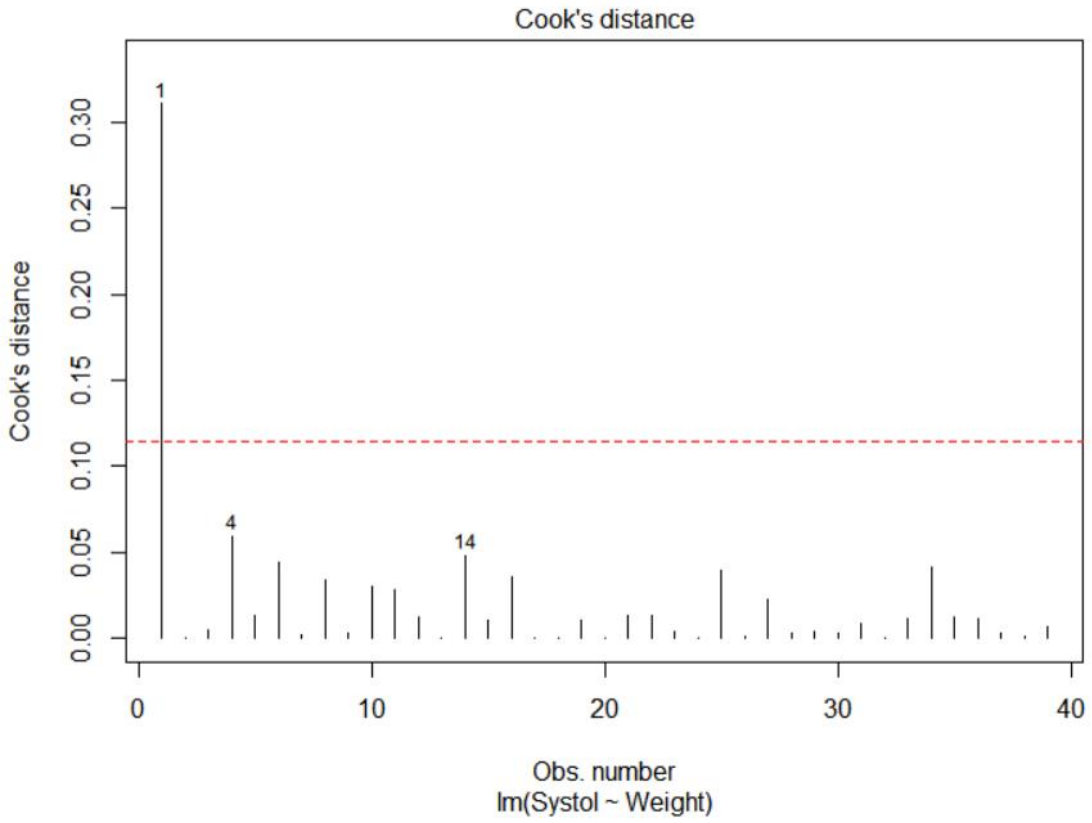
从上图可以看到观测 39 为高杠杆值点。

③ 强影响点

强影响点,即对模型参数估计值影响有些比例失衡的点。例如,若移除模型的一个观测点时 模型会发生巨大的改变,那么就需要检测一下数据中是否存在强影响点了。

有两种方法可以检测强影响点：Cook 距离,或称 D 统计量,以及变量添加图（added variable plot）。一般来说,Cook's D 值大于 $4/(n-k-1)$,则表明它是强影响点,其中 n 为样本量大小,k 是预测变量数目。

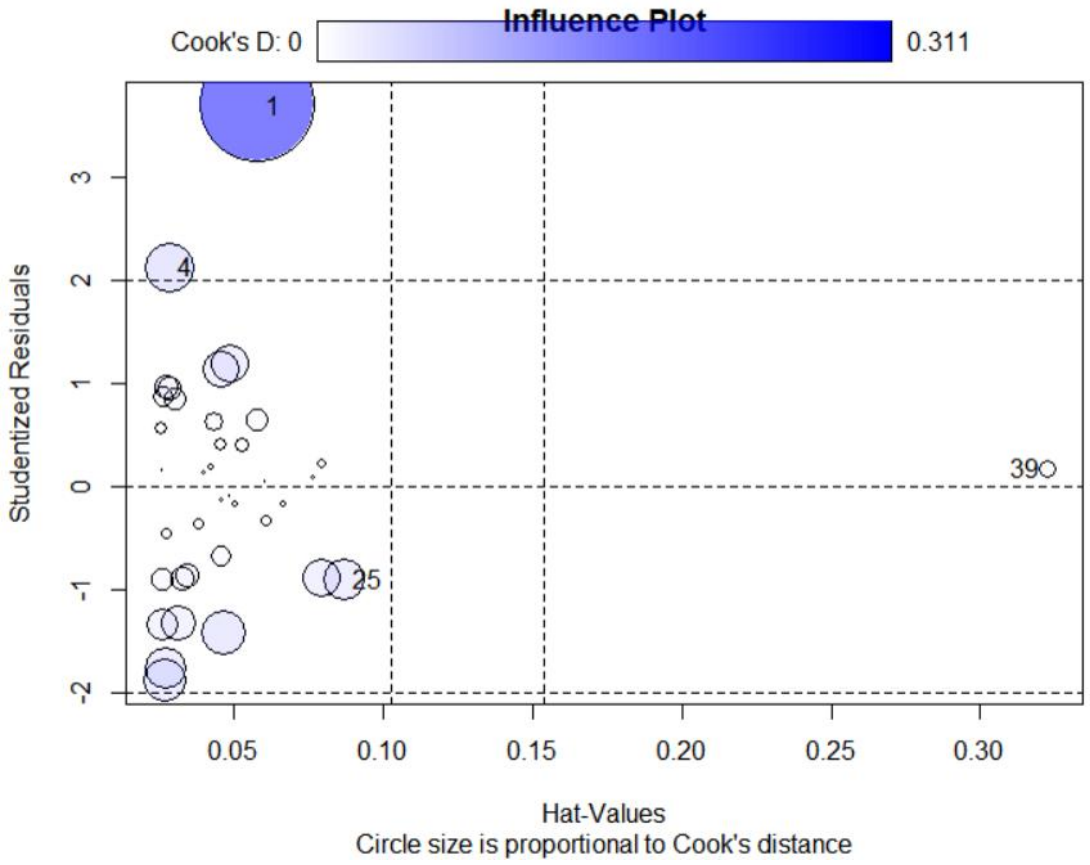
```
> cutoff <- 4/(nrow(peru)-length(LM1$coefficients)-2)
> plot(LM1, which=4, cook.levels=cutoff)
> abline(h=cutoff, lty=2, col="red")
```



可以看到观测 1 为强影响点。

最后,还可以将离群点、杠杆值和强影响点的信息整合到一幅图形中：

```
> influencePlot(LM1, id=(method="identify"), main="Influence Plot",
+             sub="Circle size is proportional to Cook's distance")
      StudRes       Hat     CookD
1   3.6857962 0.05774677 0.31062995
4   2.1218970 0.02807509 0.05940562
25 -0.9069408 0.08701409 0.03938595
39  0.1727920 0.32245675 0.00729610
```



可以看到观测 1 既是离群点又是强影响点，该点对拟合曲线的影响较大，同时观测 1 的数据为，体重为 71 公斤，血压收缩压为 170mmHg，远高于正常值，说明这个数据可能来自一名高血压患者。因此，可以将该点删除，可能可以使模型预测更加合理可信，获得比较满意的拟合结果。

因此，删去观测 1 再次进行回归分析，可以得到：

```
> Sl<-Systol[2:39]
> Wt<-Weight[2:39]
> newLM1<-lm(Sl~Wt)
> newLM1
Call:
lm(formula = Sl ~ Wt)

Coefficients:
(Intercept)          Wt
   75.2604       0.8106
> summary(newLM1)
Call:
lm(formula = Sl ~ Wt)

Residuals:
    Min      1Q  Median      3Q     Max
-19.5173  -7.4093   0.2668   7.0654  23.2933
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.2604    14.4148   5.221 7.63e-06 ***
Wt            0.8106     0.2276   3.562  0.00106 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.794 on 36 degrees of freedom
Multiple R-squared:  0.2606,  Adjusted R-squared:  0.24
F-statistic: 12.69 on 1 and 36 DF,  p-value: 0.001059
```

2.3 模型确定即模型解释

从上述分析及结果可以得到,最后的一元回归模型为

$$\widehat{Systol} = 75.2604 + 0.8106 \times Weight$$

模型表明，当体重每增加 1kg 时，收缩压平均而言上升 0.8106mmHg，但模型的决定系数 $R^2 = 0.2606$，说明只解释了因变量 26.06%的变动，解释能力较差。

2.4 模型预测

当自变量（体重）60 公斤时，因变量（收缩压 Systol）均值的置信区间和因变量新值的预测区间为：

```
> new<-data.frame(Wt=60)
> new
  Wt
1 60
> lm.conf<-predict(newLM1,new,interval="confidence",level = 0.95)> lm.conf
      fit     lwr     upr
1 123.8961 120.3976 127.3946
> lm.pred<-predict(newLM1,new,interval="prediction",level = 0.95)
> lm.pred
      fit     lwr     upr
1 123.8961 103.7277 144.0644
```

由结果可以看出,当体重为 60 公斤时,收缩压 Systol 平均值的置信区间为[120.3976,127.3946]，新值的预测区间为[103.7277,144.0644]，单位为 mmHg。

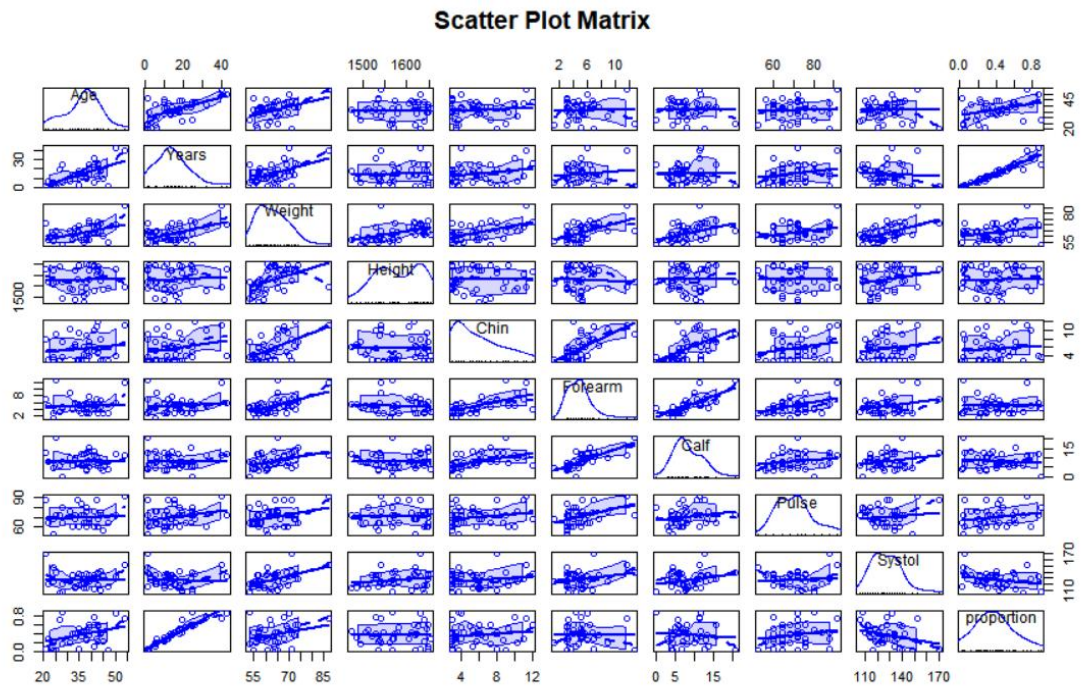
由上述检验及结果分析可以看到,只使用 Weight 作为自变量进行一元线性回归得到的模型拟合效果并不够好,因此,接下来考虑进行多元线性回归。

三、多元线性回归模型

3.1 变量选择与模型建立

仍从变量的相关关系出发，使用 `scatterplotMatrix()`可以生成散点图矩阵。

```
> scatterplotMatrix(peru,spread=FALSE,smooth.args=list(lty=2),main="Scatter Plot Matrix")
```

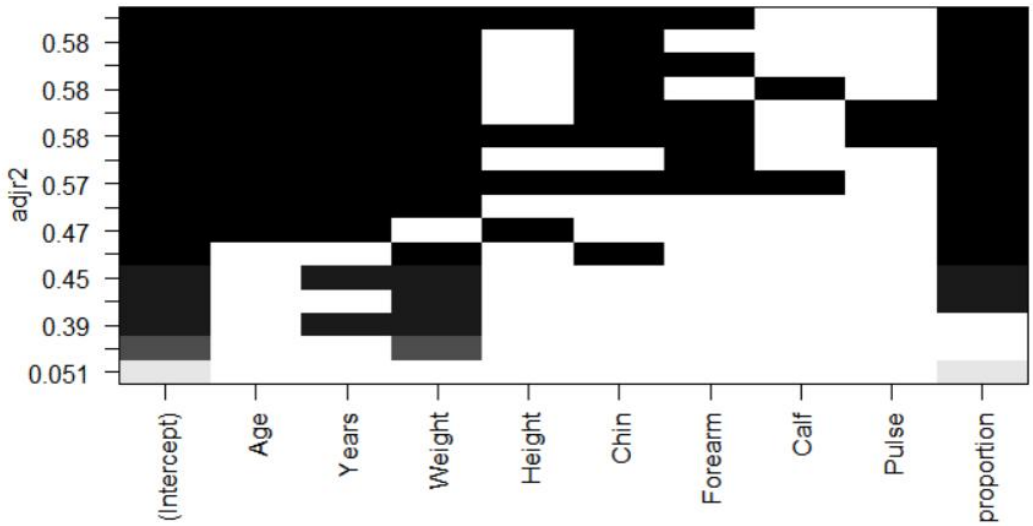


可以看到因变量（Systol)和 Weight、Forearm、Calf 等变量间的线性关系比较强。

因此,接下来我们进行选择“最佳”的回归模型。选择全子集回归法可能会帮我们找到比较适合的多元回归模型。

全子集回归法是指所有可能的模型都会得到检验。

```
> leaps<-regsubsets(Systol~Age+Years+Weight+Height+Chin+Forearm+Calf+Pulse+proportion,dat
a=peru,nbest=2)
> plot(leaps,scale = "adjr2")
```



由全子集回归图可以看到调整决定系数最高的两个模型分别是七预测变量模型和五变量预测模型。

七变量模型：Age,Years,Weight,Height,Chin,Forearm,proportion

五变量模型：Age,Years,Weight,Chin,proportion

两模型的调整决定系数相差不大，因此我们更倾向于选择变量较少的模型。

同时，通过向后逐步回归法也可以得到五变量模型比七变量模型更合适，代码如下：

```
> LM7<-lm(Systol~Age+Years+Weight+Height+Chin+Forearm+proportion);LM7

Call:
lm(formula = Systol ~ Age + Years + Weight + Height + Chin +
    Forearm + proportion)

Coefficients:
(Intercept)          Age        Years       Weight       Height         Chin
  153.68245     -1.09745      2.35670      1.46055     -0.03713     -1.03844
    Forearm   proportion
   -1.14395   -110.45540

> library(MASS)
> stepAIC(LM7,direction = "backward")
Start:  AIC=173.56
Systol ~ Age + Years + Weight + Height + Chin + Forearm + proportion

           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
```

```
- Years      1    651.50 2867.9 181.61
- Age        1    856.29 3072.7 184.30
- Weight     1    880.23 3096.7 184.61
- proportion  1   1067.91 3284.3 186.90

Step:  AIC=172.9
Systol ~ Age + Years + Weight + Chin + Forearm + proportion

           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
- Years      1    811.30 3105.0 182.71
- Age        1    848.93 3142.6 183.18
- Weight     1   1036.53 3330.2 185.44
- proportion  1   1246.44 3540.1 187.83

Step:  AIC=172.02
Systol ~ Age + Years + Weight + Chin + proportion

           Df Sum of Sq    RSS    AIC
<none>                  2360.2 172.01
- 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
- proportion  1   1180.14 3540.4 185.83

Call:
lm(formula = Systol ~ Age + Years + Weight + Chin + proportion)

Coefficients:
(Intercept)          Age        Years       Weight         Chin   proportion
    109.359       -1.012        2.407        1.098       -1.192     -110.811
```

向后逐步回归法采用 AIC 为评价指标，如果删除一个变量使得当前 AIC 减少，则删除，直到不能删除变量为止。从上面的运行结果可以看出，初始 AIC 为 173.56，剔除 Height 后变为 172.90，剔除 Forearm 变为 172.01，然后提出任何变量都会增大 AIC，方法结束，留下的变量为 Age、Years、Weight、Chin、proportion。

对五变量模型求解多元线性回归模型，结果如下

```
> LM5<-lm(Systol ~ Age + Years + Weight + Chin + proportion)
> LM5

Call:
lm(formula = Systol ~ Age + Years + Weight + Chin + proportion)
```

```
Coefficients:
(Intercept)         Age       Years      Weight        Chin   proportion
   109.359      -1.012       2.407       1.098      -1.192     -110.811
```

因此可以得到多元线性回归模型为：

$$\widehat{Systol} = 109.359 - 1.012Age + 2.407Years + 1.098Weight - 1.192Chin - 110.811proportion$$

3.2 回归诊断

3.2.1 决定系数和 F 检验

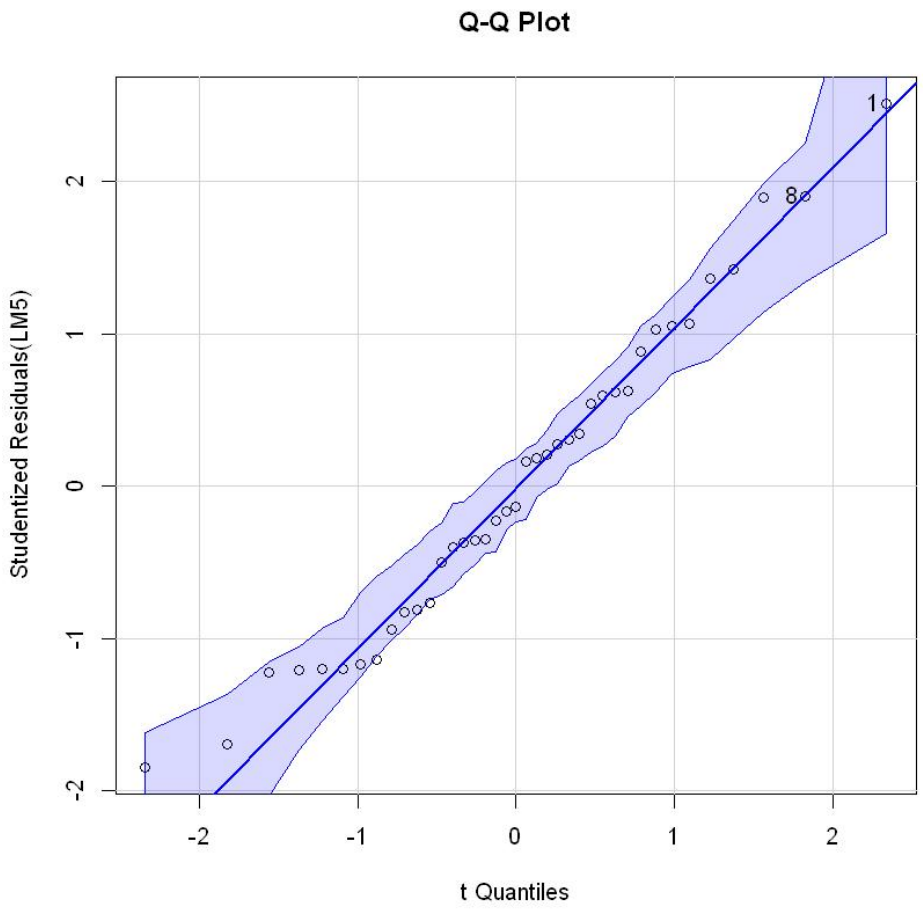
```
> summary(LM5)

Call:
lm(formula = Systol ~ Age + Years + Weight + Chin + proportion)

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 **
Weight        1.0976     0.2980   3.683 0.000819 ***
Chin         -1.1918     0.6140  -1.941 0.060830 .
proportion  -110.8112    27.2795  -4.062 0.000282 ***
---
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
```

由结果可以看出，除 Chin(下巴皮褶)外,其余变量均在 0.01 的水平上显著,而 Chin 的 p 值＜0.1,可以接受。此外,整个回归模型的 F 检验统计量为 11.66，p 值＜0.01，说明整个回归模型显著，而决定系数R^2为 0.6386,调整的决定系数$adj-R^2$为 0.5839,模型相较于一元回归模型解释能力显著增强。

3.2.2 模型适用性检验

① 正态性

```
> qqPlot(LM5,id=TRUE, simulate=TRUE,main="Q-Q Plot")
```



qqPlot()画出了在$n-p-1$个自由度的 t 分布下的学生化残差图形，图上的点大致趋于一条直线，直观上符合正态分布。

② 误差的独立性

利用 Durbin-Watson 检验判断误差是否独立：

```
> durbinWatsonTest(LM5)
 lag Autocorrelation D-W Statistic p-value
   1     -0.1732433     2.232518   0.576
 Alternative hypothesis: rho != 0
```

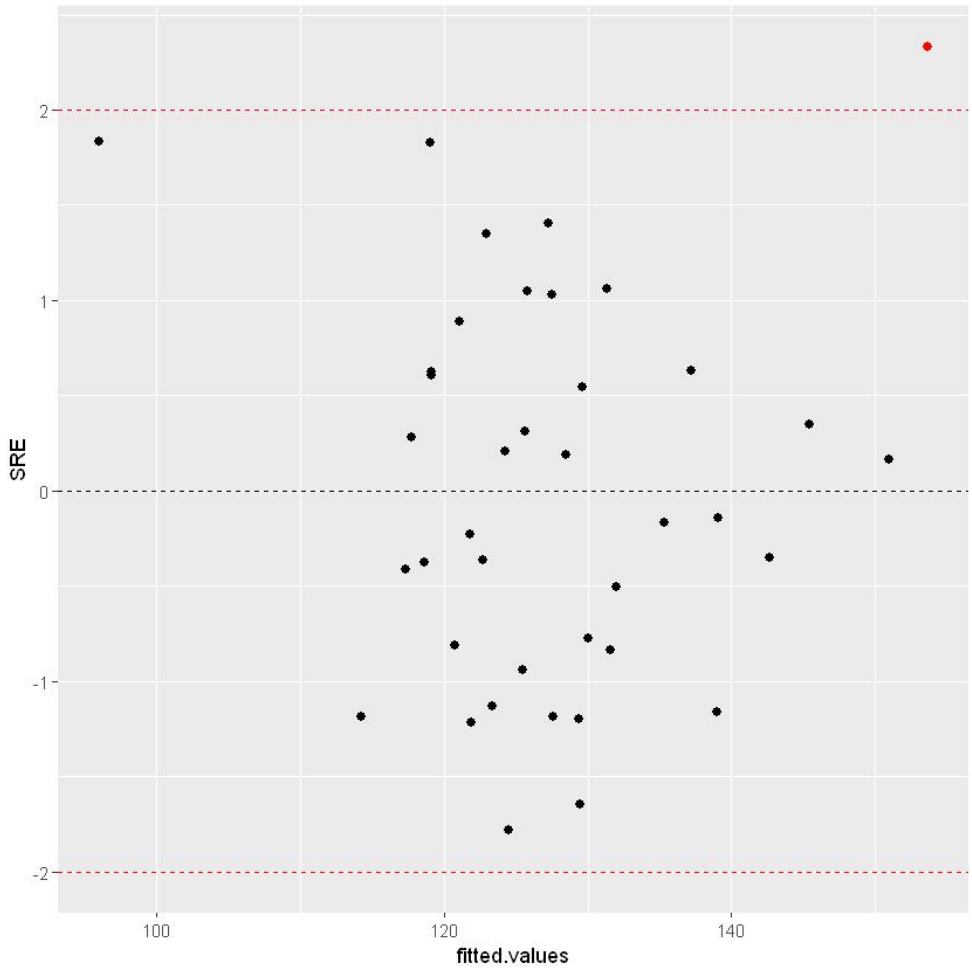
$p = 0.576 > 0.05$，结果不显著，接受原假设，即误差项之间独立。

③ 线性性

绘制学生化残差图如下，学生化残差图相比一般的残差图的优势在于不受单位的影响。

```
>library(dplyr)
>library(ggplot2)
>RSAdata = data.frame(fitted.values=predict(LM5), SRE=rstandard(LM5))
>RSAdata = mutate(RSAdata, new=if_else(abs(SRE) > 2,1,0))
>ggplot(data = RSAdata, aes(x = fitted.values,y = SRE, color = new))+
   geom_point(size = 2)+
```

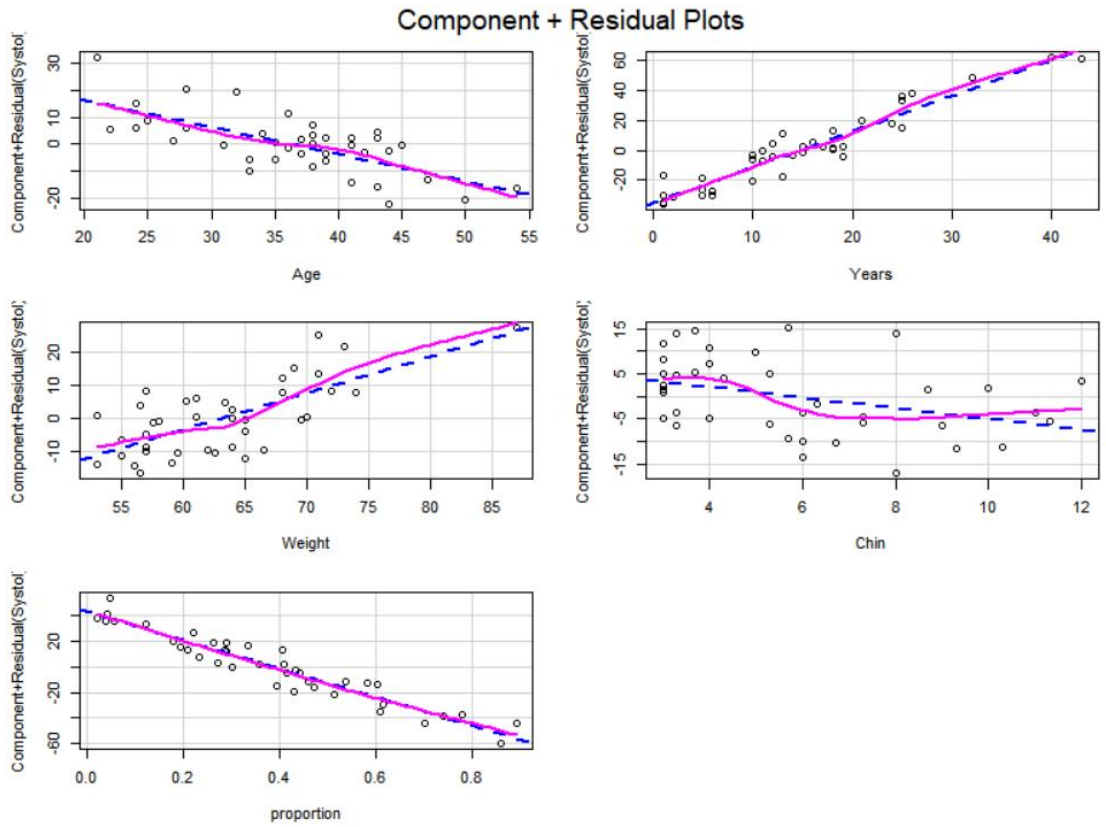
```
    geom_hline(yintercept = c(-2,0,2),linetype = 'dashed' , color = c('red', 'black','red'))+
    scale_color_gradient(low = 'black',high='red', guide = F)
```



一般认为超过±2的学生化残差为异常值，从上图看到，只有一个点为异常点。因此，可以认为残差在 0 附近随机波动，满足线性性假设。

也可以通过自变量的偏残差图来判断是否满足线性性，代码如下：

```
> crPlots(LM5)
```



可以看到图形大致呈线性，因此可以认为线性性假设成立。

④ 同方差性

```
> ncvTest(LM5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.07495813, Df = 1, p = 0.78425
```

利用 ncvTest()生成一个计分检验，$p = 0.78425 > 0.05$，结果不显著，接受原假设，认为误差方差恒定。

根据①②③④，可以认为线性模型的四个基本假设成立。

3.2.3 多重共线性检验

利用方差扩大因子法检验多重共线性。

```
> vif(LM5)
      Age     Years    Weight      Chin proportion
 2.938806 29.852840  2.377901  1.484336 20.886013
 > sqrt(vif(LM5))>2
      Age     Years    Weight      Chin proportion
    FALSE      TRUE     FALSE     FALSE      TRUE
```

可以发现,模型具有多重共线性（Years 和 proportion），这可能是显而易见的，因为$proportion = Years/Age$,但多重共线性对于预测结果的影响较小,而对自变量系数的估计值的解释有影响。当只关注预测值而不关注自变量变化时,我们可以接受多重共线性。

尝试删除 VIF 最大的 Years 变量，作四变量回归模型并检验多重共线性。

```
>LM4 = lm(Systol~Age + Weight + Chin + proportion); vif(LM4)
```

结果如下：

Age: 1.34428704283104　　Weight: 1.81402651092542

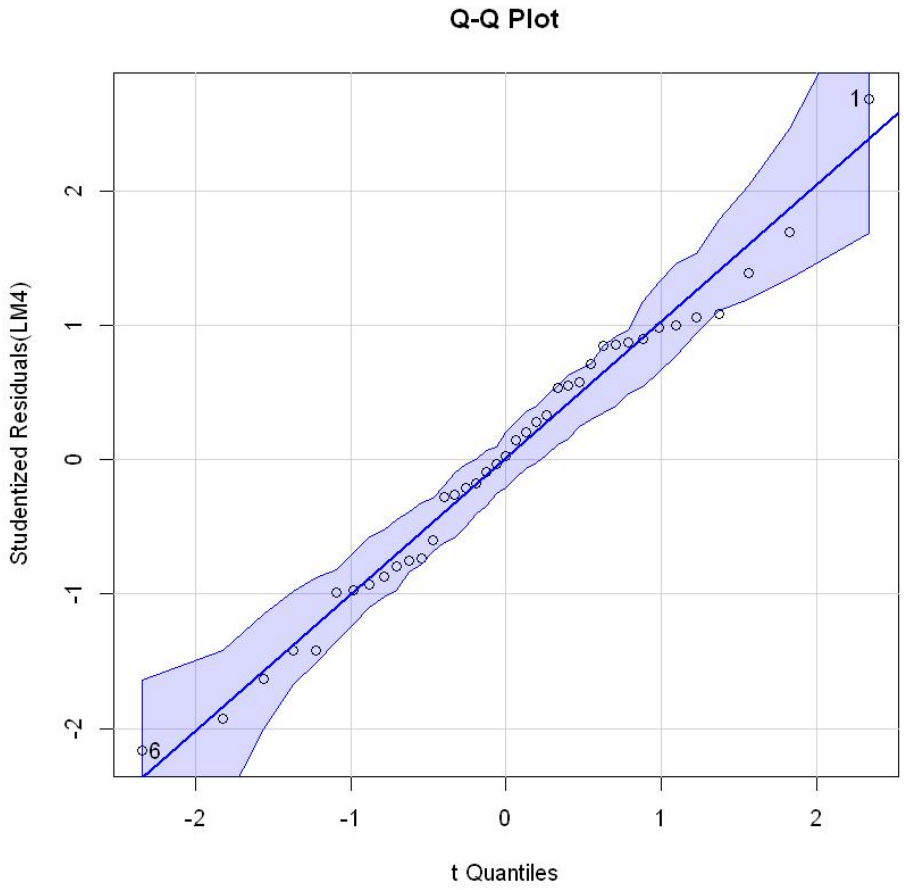
Chin: 1.48108171790358　　Proportion: 1.18497452054468

因此可以判断删除 Years 变量后模型不存在多重共线性。

下面检验该四变量模型的四个基本假设是否成立：

①正态性

```
>qqPlot(LM4,id=TRUE, simulate=TRUE,main="Q-Q Plot")
```



异常点为 1 和 6，其余点大致呈直线，可以认为满足正态性假设。

②独立性

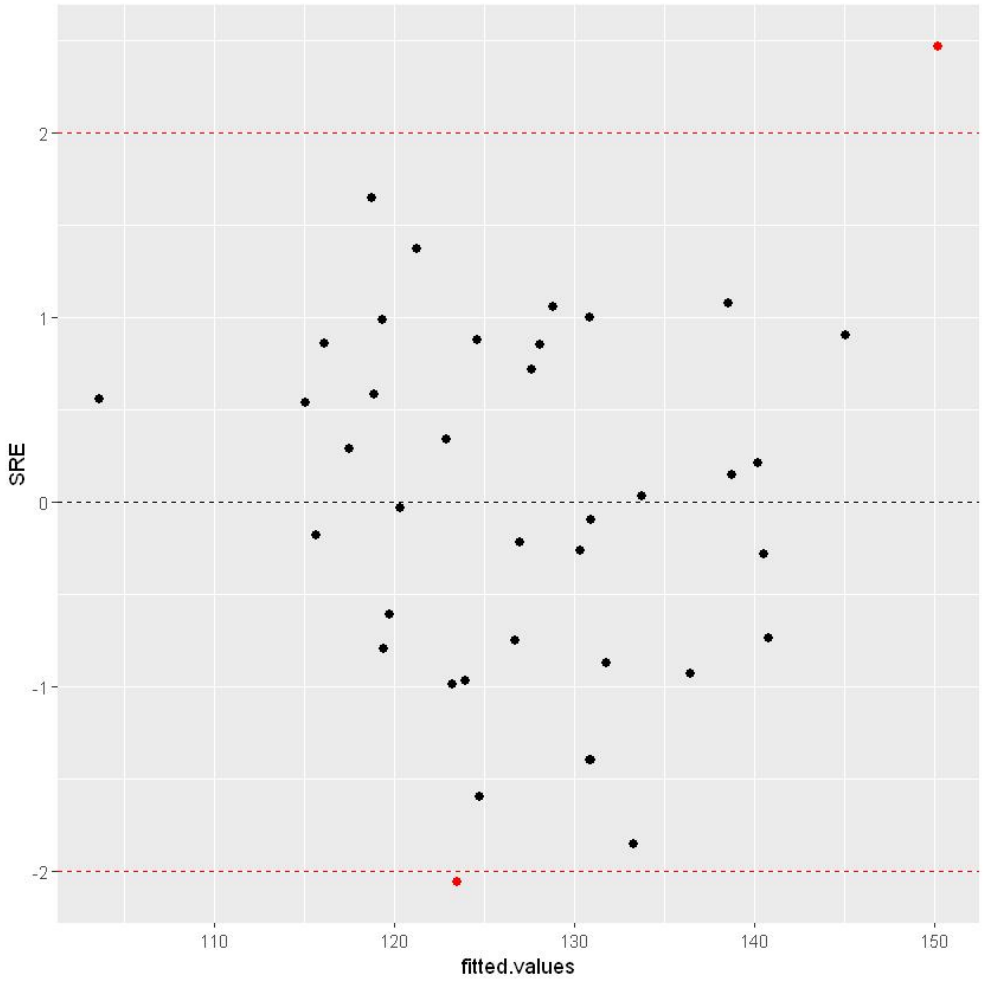
```
>durbinWatsonTest(LM4)
lag Autocorrelation D-W Statistic p-value
   1      -0.1375673      2.133176   0.842
 Alternative hypothesis: rho != 0
```

DW 检验不显著，接受原假设，误差项之间相互独立。

③线性性

```
>RSAdata1 = data.frame(fitted.values=predict(LM4), SRE=rstandard(LM4))
>RSAdata1 = mutate(RSAdata1, new=if_else(abs(SRE) > 2,1,0))
>ggplot(data = RSAdata1, aes(x = fitted.values,y = SRE, color = new))+
   geom_point(size = 2)+
```

```
    geom_hline(yintercept = c(-2,0,2),linetype = 'dashed' , color = c('red', 'black','red'))+
scale_color_gradient(low = 'black',high='red', guide = F)
```



根据学生化残差图，可以认为残差在 0 周围随机波动，线性性假设成立。

④同方差性

```
>ncvTest(LM4)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.158575, Df = 1, p = 0.28176
```

计数检验不显著，接受原假设，认为误差项同方差。

根据①②③④，四变量模型满足基本假设，模型求解结果为

$$\widehat{Systol} = 54.7373 - 0.2817 \times Age + 1.5679 \times Weight$$
$$-1.0986 \times Chin - 24.9480 \times proportion$$

决定系数 $R^2 = 0.5236$，$Adj - R^2 = 0.4676$，四变量模型的决定系数低于五变量模型，但是系数具有可解释性。因此，当需要判断影响收缩压的影响因素时，采用四变量模型；当需要预测收缩压时，采用五变量模型。

3.2.4 异常值检验

由于五变量模型调整决定系数更高，所以异常值检测按照五变量模型来进行。当进行变量系数解释时，才利用四变量模型。

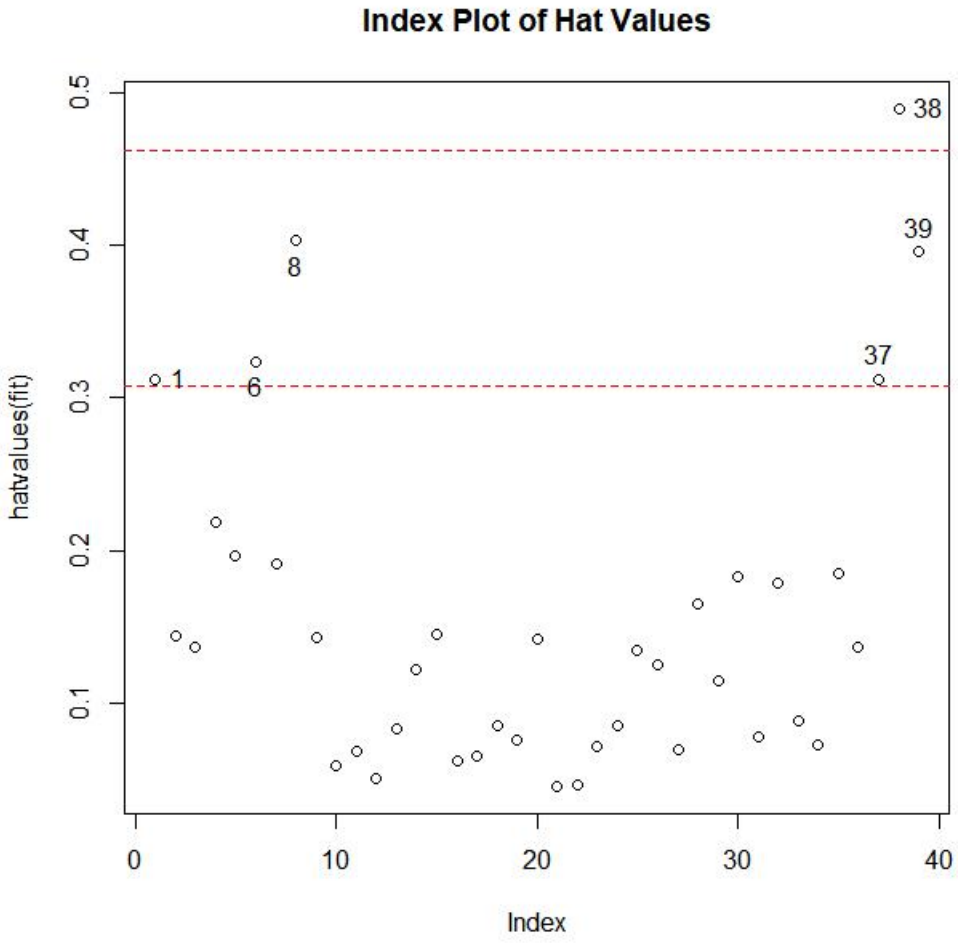
① 离群点

```
> outlierTest(LM5)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
1 2.513052         0.017201      0.67084
```

可以看到此时没有离群点存在。

② 高杠杆值点

```
> hat.plot(LM5)
[1]  1  6  8 37 38 39
```



可以看到 1 、6 、8、37、38、39 为高杠杆值点，它们由许多异常的自变量组合而成，例如点 1 的来到城镇居住时间为 1 年，低于其他人，而下巴褶皱又高于其他人。高杠杆值点可能是强影响点，也可能不是，这要看它们是否是离群点。

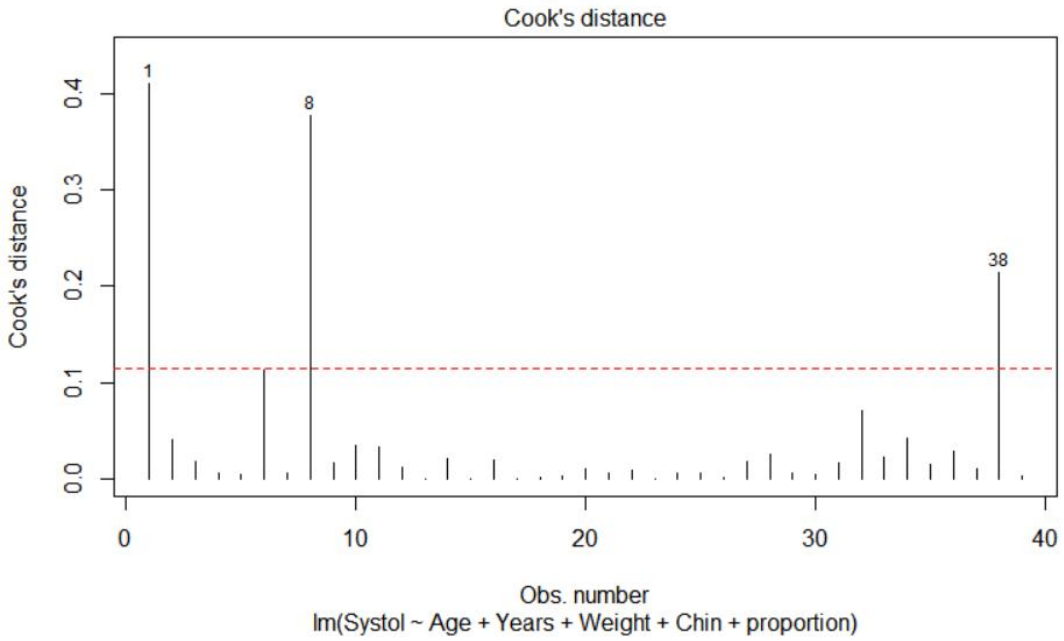
③ 强影响点

有两种方法可以检测强影响点，一种是 D 统计量法，另一种是变量添加图法。

首先进行 D 统计量检验：

```
> cutoff <- 4/(nrow(peru)-length(LM1$coefficients)-2)
> plot(LM5, which=4, cook.levels=cutoff)
```

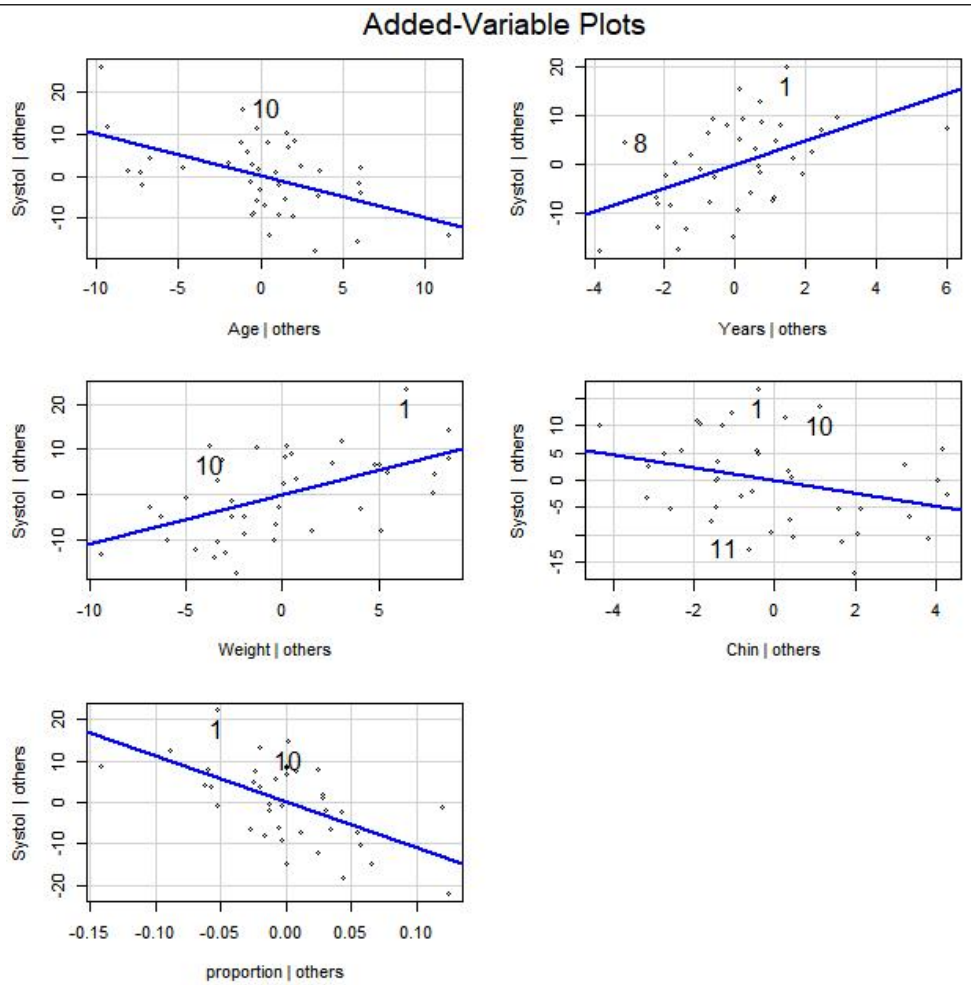
```
> abline(h=cutoff, lty=2, col="red")
```



可以看到观测 1,8,38 为强影响点，也就是说，移除这些点，模型参数估计值会发生显著变化。

接下来进行变量添加图检验：

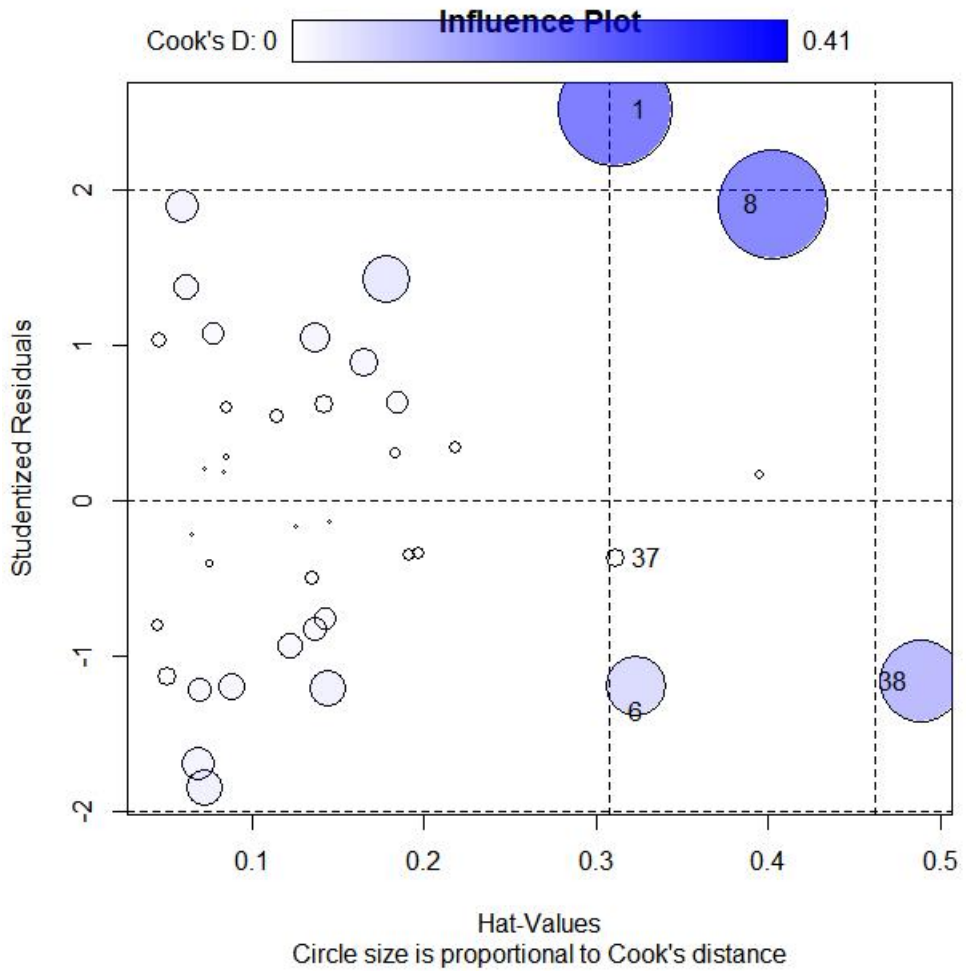
```
> avPlots(LM5,ask=FALSE,id.method="identify")
```



可以看到 1、8、10、11 为强影响点。

最后,将离群点、杠杆值和强影响点的信息整合到一幅图形中：

```
> influencePlot(LM5, id=list(method="identify"), main="Influence Plot",
+               sub="Circle size is proportional to Cook's distance")
```



尝试删去观测 1、8 和 38 再次进行回归：

```
> detach(peru)
> newperu<-peru[c(-1,-8,-38),]
> attach(newperu)
> nLM5<-lm(Systol ~ Age + Years + Weight + Chin + proportion)
> nLM5
Call:
lm(formula = Systol ~ Age + Years + Weight + Chin + proportion)

Coefficients:
(Intercept)          Age        Years       Weight         Chin   proportion
   122.5453      -0.9536       3.3932       0.8582      -1.4577    -146.4596
> summary(nLM5)
Call:
lm(formula = Systol ~ Age + Years + Weight + Chin + proportion)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9376  -4.3938   0.0379   4.4212  16.7469
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.5453    20.5538   5.962 1.55e-06 ***
Age           -0.9536     0.3150  -3.027 0.005035 **
Years          3.3932     0.8722   3.890 0.000516 ***
Weight         0.8582     0.2803   3.062 0.004611 **
Chin          -1.4577     0.5702  -2.557 0.015873 *
proportion  -146.4596    32.2829  -4.537 8.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.457 on 30 degrees of freedom
Multiple R-squared:  0.6119,  Adjusted R-squared:  0.5472
F-statistic: 9.459 on 5 and 30 DF,  p-value: 1.708e-05
```

可以看到决定系数R^2为 0.6119,且所有系数估计值均显著,因此可以尝试接受删去这些异常值点进行回归。

3.2.5 模型确定及模型解释

根据最后的分析,选择由更少的数据进行回归,则相较于原本的回归模型:

$$\widehat{Systol} = 109.359 - 1.012Age + 2.407Years + 1.098Weight - 1.192Chin - 110.811proportion$$

新模型变为:

$$\widehat{Systol} = 122.5453 - 0.9536Age + 3.3932Years + 0.8582Weight - 1.4577Chin - 146.4596proportion$$

模型决定系数为 0.6119,预测能力较强,由于模型存中 Years 与 proportion 存在多重共线性,该模型对斜率项估计值的解释意义不大。

利用 3.2.3 多重共线性检验中提出的四变量模型来进行系数解释。四变量模型如下:

$$\begin{aligned}\widehat{Systol} &= 54.7373 - 0.2817 \times Age + 1.5679 \times Weight \\ &-1.0986 \times Chin - 24.9480 \times proportion\end{aligned}$$

可以看到, 年龄每增加一岁, 收缩压平均而言下降 0.2817mmHg; 体重每增加 1kg, 收缩压平均而言增加 1.5679mmHg; 皮肤皱褶每增加 1 单位, 收缩压平均而言下降 1.0986mmHg; 在城镇生活时间占当前年龄的比例每增加 1%, 收缩压平均而言下降 0.24948mmHg。

3.2.6 模型预测

```
> new<-data.frame(Age=mean(Age),Years=mean(Years),Weight=mean(Weight),Chin=mean(Chin),pro
portion=mean(proportion))
> new
      Age    Years   Weight Chin proportion
1 36.83333 14.05556 63.03333  5.9   0.368515
> lm.conf<-predict(nLM5,new,interval="confidence",level = 0.95)
> lm.conf
      fit      lwr      upr
```

```
1 126.6389 124.1007 129.1771
> lm.pred<-predict(nLM5,new,interval="prediction",level = 0.95)
> lm.pred
      fit      lwr      upr
1 126.6389 111.1995 142.0783
```

可以看到，收缩压平均值的置信区间为$[124.1007, 129.1771]$，收缩压预测新值的预测区间为$[111.1995, 142.0783]$。

因变量均值在给定$x = x_h$的置信区间为：

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times (\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum (x_i - \overline{x})^2})}$$

因变量新值在给定$x = x_h$的预测区间为：

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times (1 \ + \ \frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum (x_i - \overline{x})^2})}$$

通过公式可以明显看出新值的预测区间要比均值的置信区间大一些。

教师签字_____

第三部分　实验总结

1. 本次实验，我们首先根据变量相关系数图选择了和因变量最密切相关的Weight变量进行一元线性回归，并且通过了线性性、独立性、正态性、同方差性检验，检测出的异常点数量也很少，但是模型决定系数不高，只有26％左右。

2. 一元线性回归结果较差，接下来准备建立多元线性回归模型。首先通过全子集选择法，得到了七变量模型和五变量模型(见3.1 变量选择与模型建立)的拟合效果都比较好。直观上来说在拟合效果差不多的情形下，选择变量较少的模型比较好，这样可以避免多元回归中严重多重共线性的问题。之后又将七变量模型记为全模型，进行向后逐步回归法，得到的结果恰好是全子集选择法中的五变量模型，给之前凭直观选择变量较少的模型提供了依据。

3. 本文在变量选择部分运用了全子集回归和逐步回归两种方法。大部分情况中，全子集回归要优于逐步回归，因为考虑了更多模型。但是，当有大量预测变量时，全子集回归会很慢。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对结果解释毫无帮助，主题背景知识的理解才是建立优良模型的基础。

4. 在针对五变量模型的回归诊断中，虽然该模型是逐步回归的结果，但仍然存在严重的多重共线性。对整个方程的F检验显示方程整体上显著，对单个变量的t检验，除Chin以外的变量均在0.01的水平上显著，变量Chin也在0.1的水平上显著，所以方程的整体预测能力并不受多重共线性的影响，只是对回归系数的解释变得不可靠。针对多重共线性问题，剔除方差膨胀因子最大的变量重新建立四变量模型，利用四变量模型进行回归系数解释，利用五变量模型进行预测。

5. 针对多重共线性，除了剔除方差膨胀因子最大的变量重新回归以外，还可以采用主成分回归、岭回归、Lasso回归等方法。采用上述方法后，回归系数具有了可解释性，但系数估计是有偏估计，而不是普通最小二乘的无偏估计。此外，对于主成分回归而言，主成分的现实意义难以解释；对于岭回归而言，系数k的取值难以确定。

6. 本文一开始删除了舒张压(Diastol)变量，只利用收缩压(Systol)作为因变量做连续型回归。可以尝试利用舒张压和收缩压计算出高血压分类(1级、2级、3级)，然后利用分类方法进行建模，比如逻辑回归、决策树、随机森林、支持向量机等方法。但由于本课程主要关注回归分析，对于分类算法涉及不多，因此在本实验中不展示分类算法的建模。