61 29

male_height

185, 173, 175, 182, 173, 181, 184, 179, 181, 187, 169, 178,
183, 168, 181, 175, 175, 186, 186, 182, 178, 177, 172, 168,
173.5, 184, 183, 175, 168, 174, 181, 170, 166, 178, 177, 181,
163, 172, 160, 173, 185, 172, 183, 180, 175, 178, 169, 175,
165, 169, 170, 183, 184, 174, 170, 173, 170, 182, 178, 170, 179

male_weight

65, 62, 80.3, 74.3, 55.7, 60, 59, 79, 62, 80, 56, 60.5,
73, 46, 65, 91, 64, 88, 63, 64, 65, 75.6, 64, 65.5,
58.8, 59, 71, 75, 60, 61, 75, 58, 56, 94.5, 87, 71,
47.5, 59, 57, 65, 67, 60, 85, 65, 73, 70, 55, 75,
55, 65, 65, 72, 99, 75, 53, 70, 58, 63, 92, 48, 69

male_armspan

188, 182, 183, 189, 185, 170, 179, 165, 171, 196, 168, 177,
179, 167, 175, 171, 170, 183, 180, 182, 180, 177, 176, 170,
167, 179, 186, 168, 160, 171.5, 181, 172, 146, 177, 168, 183,
162, 170.5, 166.7, 173, 176, 167, 188, 178, 178, 175, 169, 171,
164, 175, 161, 174, 188, 171, 166.6, 171, 169, 160, 175, 164, 169

male_leglength

102, 93, 107, 114, 107, 101, 98, 99, 98, 119, 97, 101,
102, 99, 103, 97, 82, 101, 95, 97, 91, 95.5, 95, 82,
83, 98, 98, 105, 93, 96.5, 104.5, 85, 77, 105, 94, 102,
77, 95.8, 83.5, 89, 106, 98, 102, 93, 98, 90, 86, 98,
99, 99, 92, 100, 108, 102, 83, 101, 102, 80, 85, 92, 98

male_footlength

25.1, 23.5, 26.5, 26, 26, 25, 26.5, 26, 26, 29, 26, 26.5,
27, 24.5, 25.5, 22, 24, 28, 26.5, 26.5, 26.5, 27, 25, 25.5,
24, 26.5, 26, 27, 22, 23.4, 25.5, 25.4, 22, 26.5, 22, 26,
23.8, 23, 22.8, 25.5, 25.8, 24.5, 26.5, 24, 26.5, 29, 23, 26,
22, 23, 22, 27.5, 26.5, 26, 20, 25.5, 24.5, 24, 25.5, 23, 25

female_height

159, 172, 163, 165, 168, 165, 163, 165, 160, 158, 168, 162,
161, 172, 168, 168, 174, 161, 162, 166, 162, 162, 170, 168,
159, 168, 164, 155, 166

female_weight

47.9, 54, 60, 53, 52, 53, 58.5, 55, 50, 46.5, 58, 46,
51, 63, 66.5, 52, 56, 44, 57.5, 54.5, 53, 48, 60, 59,
54.5, 68, 55, 42, 51

female_armspan

158.8, 173, 163, 164, 169, 166, 156, 165, 161, 152, 166, 148,
152, 174, 162, 166, 179.5, 160, 164, 159, 160, 155, 169, 168,
136.6, 149, 158, 155, 165

female_leglength

100, 105, 102, 97, 97.5, 94, 88, 94, 89, 92, 95, 90,
91, 104, 99, 97, 105, 90, 88, 94, 94, 97, 104, 82,
83, 100, 90, 85, 95

female_footlength

22.8, 23.5, 24, 22.6, 24.8, 21.5, 23.5, 23.5, 23, 22.5, 23.5, 21.4,
22.6, 24.5, 24.5, 23.5, 24.5, 23, 24.3, 24, 24, 23, 24.5, 24,
24.7, 23.5, 23.5, 25, 23

顺序统计量、均值、五数（最小值、第三 4 分位数、中位数、第一 4 分位数、最大值）、方差、标准差是数据的主要的统计量。

在 R 中，函数 sort( )可以将单组数据按从小到大排列，函数 summary( )可以计算出单组数据的均值和五数，函数 var( )和 sd( )分别用来求出单组数据的方差和标准差。

对这十组数据执行这些操作。输入命令和输出结果见附录 1.1。由此可得十组数据的顺序统计量（见下列数据，仅列出男生身高，其他略）和其他主要统计量（见下表）。男生身高的均值为 176.1，最小值为 160，第三 4 分位数为 172，中位数为 175，第一 4 分位数为 181，最大值为 187，方差为 40.48197，标准差为 6.362544。

男生身高的顺序统计量：

[1] 160.0 163.0 165.0 166.0 168.0 168.0 168.0 169.0 169.0 169.0 170.0 170.0
[13] 170.0 170.0 170.0 172.0 172.0 172.0 173.0 173.0 173.0 173.0 173.5 174.0
[25] 174.0 175.0 175.0 175.0 175.0 175.0 175.0 177.0 177.0 178.0 178.0 178.0
[37] 178.0 178.0 179.0 179.0 180.0 181.0 181.0 181.0 181.0 181.0 182.0 182.0
[49] 182.0 183.0 183.0 183.0 183.0 184.0 184.0 184.0 185.0 185.0 186.0 186.0
[61] 187.0

	最小值	第一 4 分位数	中位数	均值	第三 4 分位数	最大值	方差	标准差
男生身高	160	172	175	176.1	181	187	40.482	6.363
男生体重	46	59	65	67.32	74.3	99	134.762	11.609
男生臂展	146	168	173	173.8	179	196	75.378	8.682
男生腿长	77	92	98	96.29	102	119	73.364	8.565
男生脚长	20	24	25.5	25.1	26.5	29	3.425	1.851
女生身高	155	162	165	164.6	168	174	20.672	4.547
女生体重	42	51	54	54.1	58	68	38.122	6.174
女生臂展	136.6	156	162	161.2	166	179.5	76.737	8.760
女生腿长	82	90	94	94.53	99	105	40.713	6.381
女生脚长	21.4	23	23.5	23.56	24.3	25	0.839	0.916

以男生身高为例。因男生身高最大值为 187，最小值为 160，根据课本公式 $m \approx 1.87(n-1)^{0.4}$（其中 n=61），应将样本等距分为 9 组，每组组距为 3。

R 中使用函数 hist( )来画直方图。输入命令见附录 1.2。输出结果如下

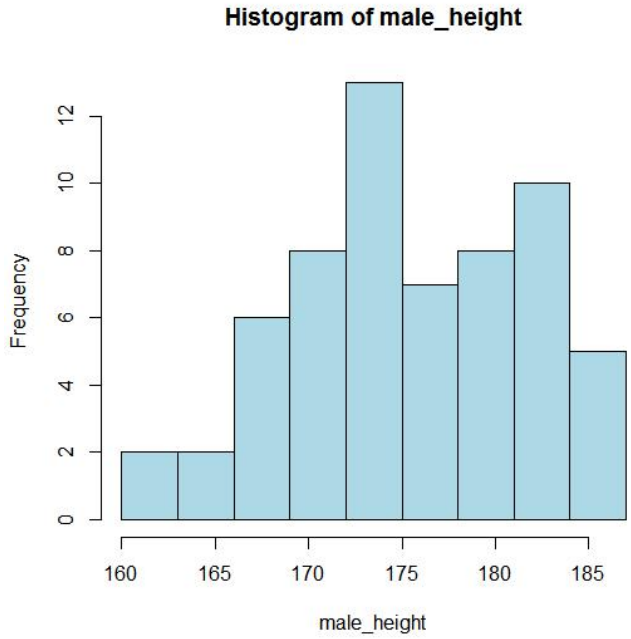
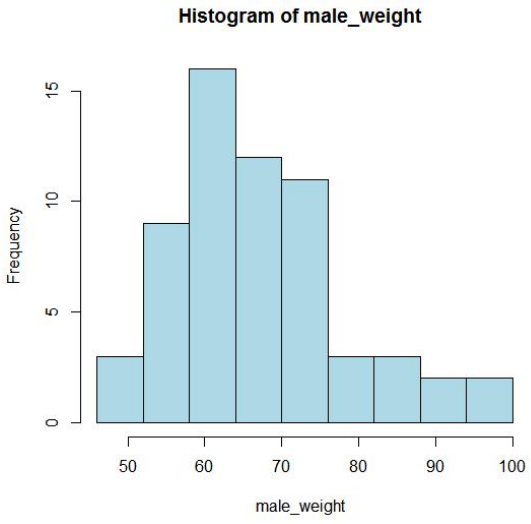


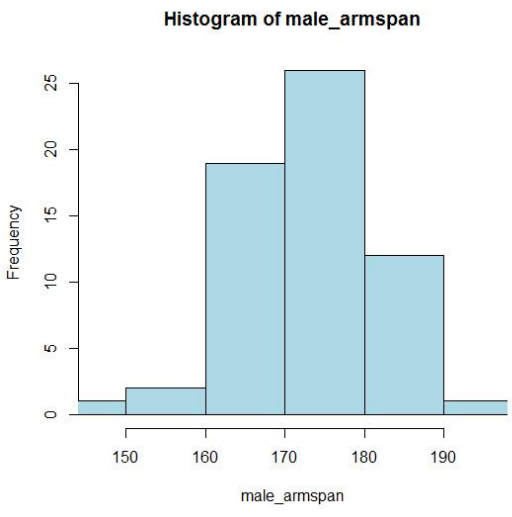
图 1.1

可见数据分布呈现中间多，两边少的特点，近似正态分布；但与正态分布的差异主要在：属于[163, 166]区间的数据较少，属于(175, 181]区间的数据较少，属于(181, 184]区间的数据较多。因此又不同于正态分布（具体分析见第三章 假设检验）。

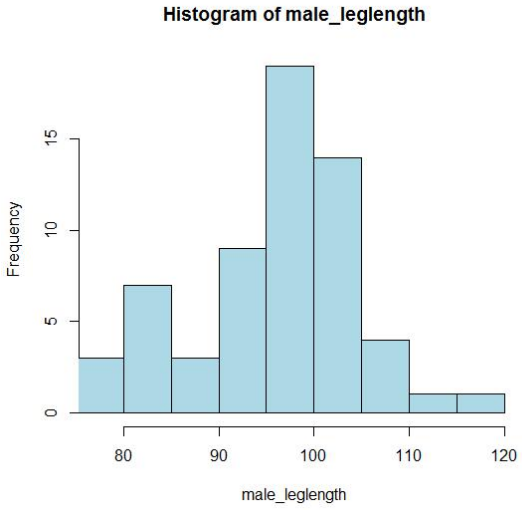
类似求得其他九组数据的直方图：



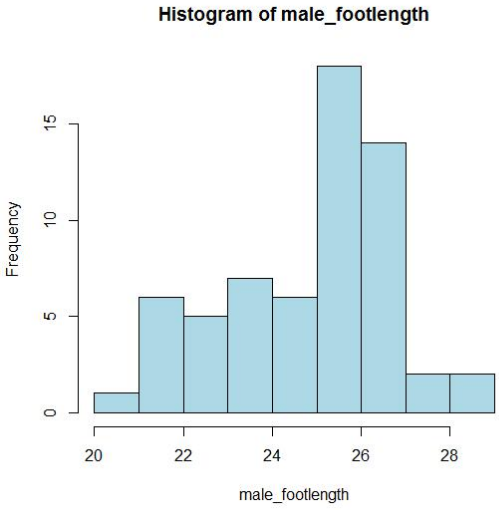
男生体重



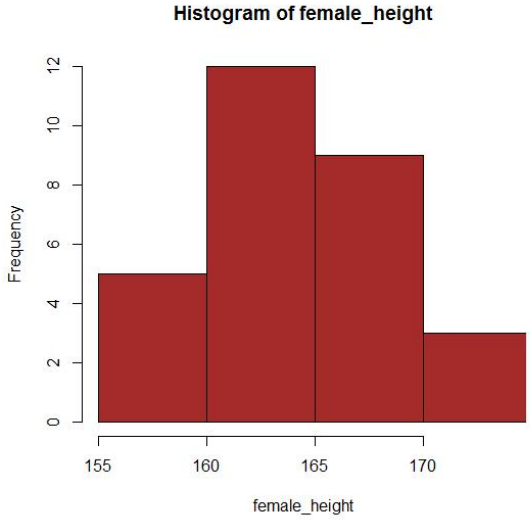
男生臂展



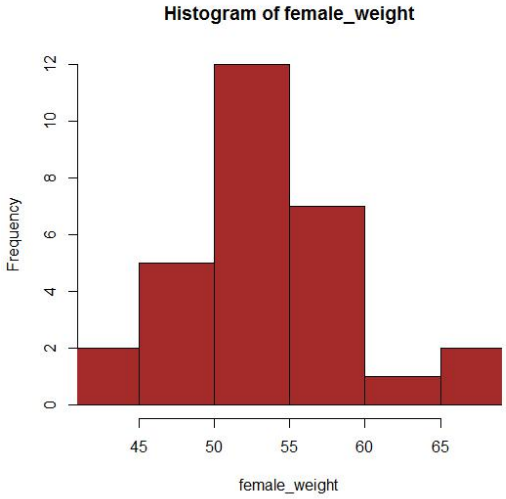
男生腿长



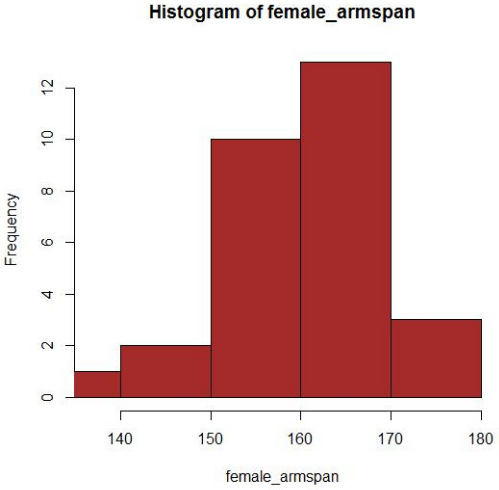
男生脚长



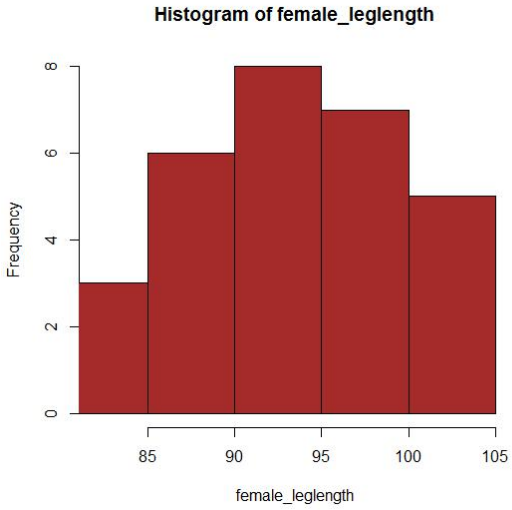
女生身高



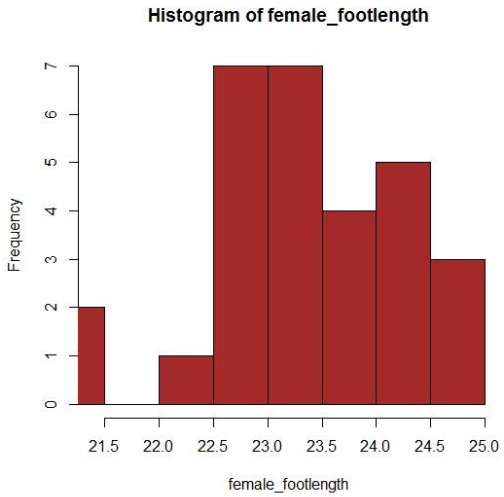
女生体重



女生臂展



女生腿长



女生脚长

可见数据分布都呈现中间多，两边少的特点，近似正态分布；但男生腿长、男生脚长、女生体重、女生脚长四组数据与正态分布有一定的差异。

以男生身高为例，在 R 中画经验分布函数。输入命令见附录 1.3。男生身高的经验分布函数如图 1.2。

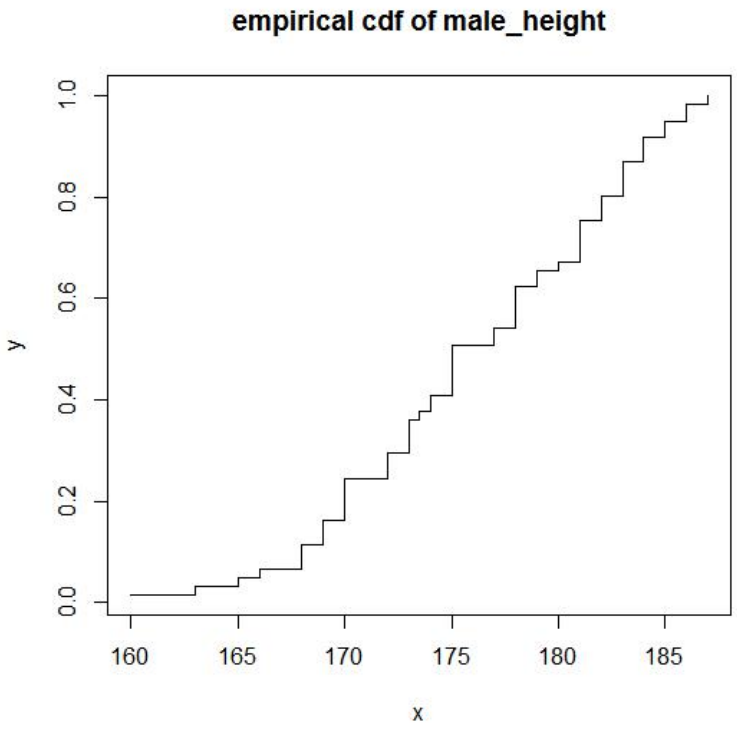


图 1.2

在第三章 假设检验中，我们将说明每组数据都近似服从正态分布。因此，我们可以用各类参数估计方法来估计其中未知的参数μ与σ^2。

正态分布的均值μ与方差σ^2的矩估计量分别为样本均值$\bar{\xi}$和样本方差S^2。故这十组数据所服从的正态分布的均值和方差的矩估计量如下表所示。

	均值	方差
男生身高	176.1	40.482
男生体重	67.32	134.762
男生臂展	173.8	75.378
男生腿长	96.29	73.364
男生脚长	25.1	3.425
女生身高	164.6	20.672
女生体重	54.1	38.122
女生臂展	161.2	76.737
女生腿长	94.53	40.713
女生脚长	23.56	0.839

正态分布的均值μ与方差σ^2的极大似然估计量也为样本的均值$\bar{\xi}$和样本方差S^2。故极大似然估计量和前面的矩估计量是一样的。

方差未知时，均值的置信度为1-α的置信区间为

$$\left(\bar{\xi}-\frac{\widetilde{S}t_{1-\alpha/2}(n-1)}{\sqrt{n}},\ \bar{\xi}+\frac{\widetilde{S}t_{1-\alpha/2}(n-1)}{\sqrt{n}}\right)$$

我们可以直接利用 R 语言的 t.test( )来求置信区间。输入命令与输出结果见附录 2.1。得到 95%置信度的均值区间估计为

	均值置信区间
男生身高	[174,44, 477.70]
男生体重	[64.35, 70.30]
男生臂展	[171.60, 176.05]
男生腿长	[94.10, 98.49]
男生脚长	[24.63, 25.58]
女生身高	[162.89, 166.35]
女生体重	[51.75, 56.45]
女生臂展	[157.84, 164.50]
女生腿长	[92.11, 96.96]
女生脚长	[23.19, 23.89]

均值未知时，方差的置信度为 1-α的置信区间为

$$\left(\frac{(n-1)\widetilde{S}^2}{\chi^2_{1-\alpha/2}(n-1)},\ \frac{(n-1)\widetilde{S}^2}{\chi^2_{\alpha/2}(n-1)}\right)$$

由于 R 中没有现成的程序，我们需要自己编写程序。输入命令与输出结果见附录 2.2。得到 95%置信度的方差区间估计为

	方差置信区间
男生身高	[29.16, 60.00]
男生体重	[97.07, 199.73]
男生臂展	[54.30, 111.72]
男生腿长	[52.84, 108,74]
男生脚长	[2.47, 5.08]
女生身高	[13.02, 37.81]
女生体重	[24.01, 69.73]
女生臂展	[48.33, 140.36]
女生腿长	[25.64, 74.47]
女生脚长	[0.53, 1.53]

本章中，由于内容较多，所以只对男生身高进行检验。其他九组数据的检验原理是一样的。

不论是男生（n=61）还是女生（n=29），样本个数都太小，偶然性太大，不能看做某些分布的近似分布。因此采用这些检验法效果并不好，很容易犯第一类错误和第二类错误。所以我们需要更换检验方法。

在 R 中可以直接调用函数 qqnorm( )来绘制 QQ 图。以男生身高为例，输入命令见附录 3.1。输出如下图片

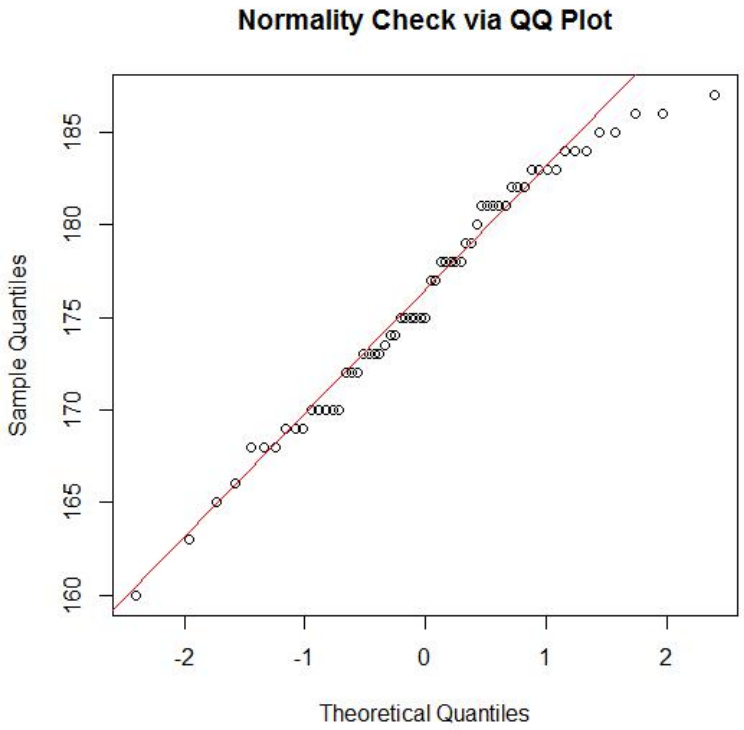


图 3.1

图 3.1 表明数据与正态性略有差异：区间(170, 175]与(180, 187]的数据都与正态性有偏差，特别是区间(184, 187]的数据（即图中右上方）与正态性有较大偏差。这与观察直方图所得直观感受相符合。

但除了区间(184, 187]外，其他数据都与正态性偏差不大，对此我们可以认为数据近似服从正态分布。出现偏差的原因可以解释为样本容量太小（n=61）。

在 R 中可以根据直方图拟合出数据近似的密度函数曲线，由此可与正态分布密度函数曲线作比较。输入命令见附录 3.2。输出结果为

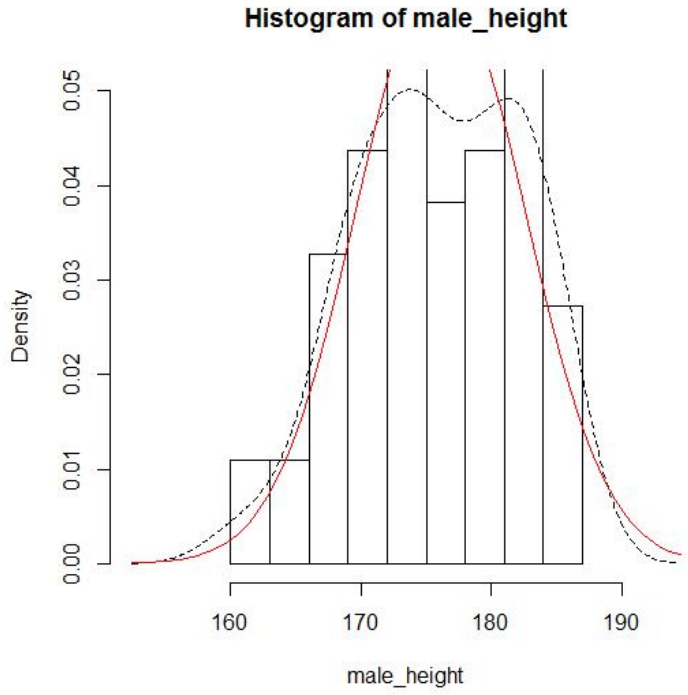


图 3.2

图 3.2 也表明数据与正态性略有差异。属于[163, 166]区间的数据较少，属于(175, 181]区间的数据较少，属于(184, 187]区间的数据较多。

我们可以把经验分布函数与正态分布函数作比较。输入命令见附录 3.3，输出结果如下

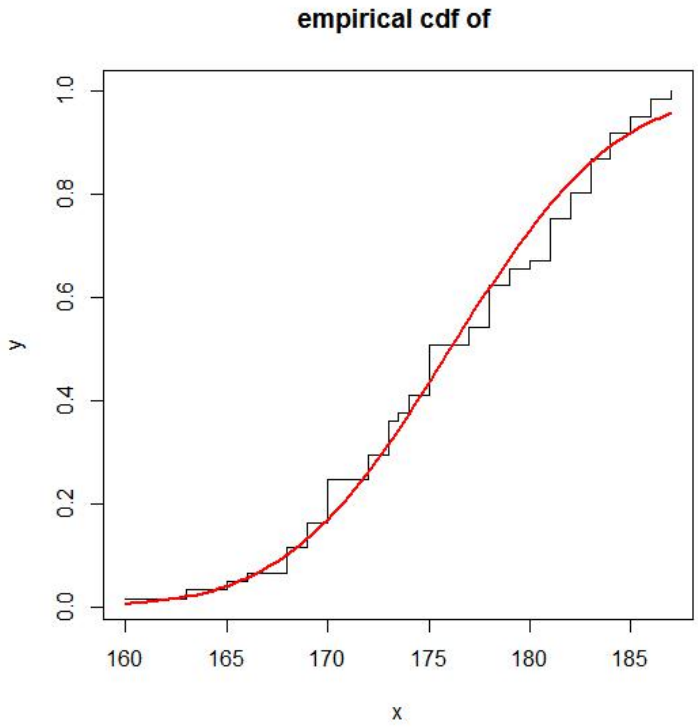


图 3.3

结论与前面类似。

假设男生身高所服从的正态分布方差为σ^2（未知），对此正态分布的均值μ进行假设检验。查阅国家统计局数据得 20-24 岁年龄区间的全国男子平均身高为 171.9，故想考察样本是否服从均值为 171.9 的正态分布。提出假设：

$$H_0 : \mu = \bar{\xi} = 176.1,\ H_1 : \mu \neq \bar{\xi} = 176.1$$

则 H_0 的置信度为 1-α的拒绝域为

$$\left\{ \left| \frac{\bar{\xi} - 171.9}{S/\sqrt{n-1}} \right| > t_{1-\alpha/2}(n-1) \right\}$$

我们可以直接利用 R 语言的 t.test( )来求在 H_0 成立时 $\left| \frac{\bar{\xi} - 171.9}{S/\sqrt{n-1}} \right| > t_{1-\alpha/2}(n-1)$ 的概率。输入命令与输出结果见附录 3.4。结果为：

$$\left| \frac{\bar{\xi} - 171.9}{S/\sqrt{n-1}} \right| > t_{1-\alpha/2}(n-1)$$

的概率为 3.36e-06。所以否定 H_0，认为样本不服从均值为 171.9 的正态分布。

先以身高为例。我们想知道男女身高所属分布的均值是否相同？记 a_1、a_2 分别为男生和女生身高所服从的分布的均值，则我们需要检验假设：

$$H_0 : a_1 = a_2, H_1 : a_1 \neq a_2$$

由方差分析知识，我们知道 H_0 的拒绝域为

$$\left\{ \frac{(n-r)Q_A}{(r-1)Q_e} > F_{1-\alpha}(r-1, n-r) \right\}$$

其中 Q_A 为组间偏差平方和，Q_e 为组内偏差平方和，n 为男生女生身高样本总数（n=90），r 为分组数（r=2）。

R 中函数 aov( )提供了方差分析的计算与检验。输入命令见附录 4.1，得到输出结果为

```
             Df Sum Sq Mean Sq F value     Pr(>F)
A             1   2578  2578.3   75.44 1.82e-13 ***
Residuals    88   3008    34.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

上述结果中, Df 表示自由度；sum Sq 表示平方和；Mean Sq 表示均方和；F value 表示 F 检验统计量的值, 即 F 比；Pr(>F)表示检验的 p 值；A 就是因素“性别”；Residuals 为残差（误差）。将其画成单因素方差分析表。

方差来源	平方和	自由度	样本方差	F 值	p 值
因素	2578	1	2578.3	75.44	1.82e-13
误差	3008	88	34.2		
总和	5586	89			

可以看出，$F=75.44>F_{0.05}(2-1,90-2)$，或者 $p=1.82e-13<<0.05$，说明有理由拒绝原假设，即认为男生女生身高均值有显著差异。

再通过函数 plot( )绘图可直观描述男女身高的差异。输入命令见附录 4.2。得到图 4.1。从图形上也可以看出，男生女生的身高有显著差异。

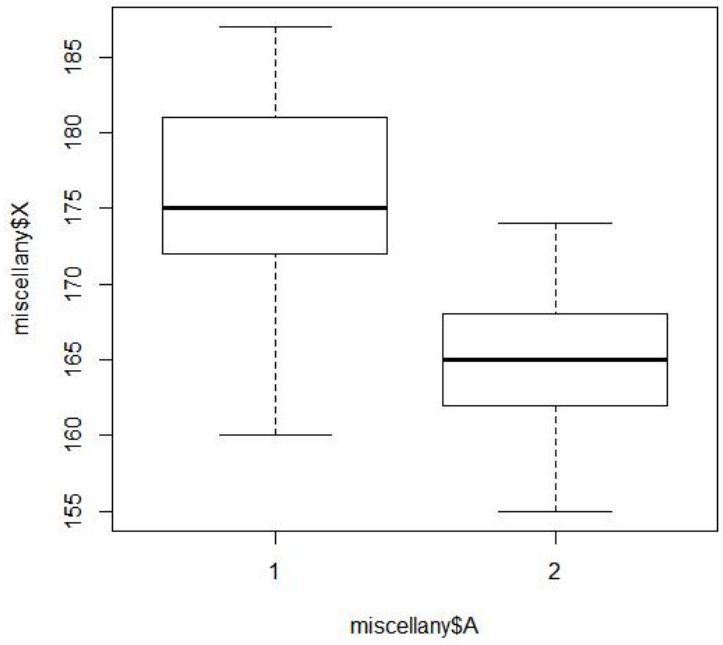


图 4.1

类似，可以求得男女生体重、臂展的均值也都有显著差异。

在进行相关分析和回归分析之前，可先通过不同变量之间的散点图直观地了解它们之间的关系和相关程度。若图中数据点分布在一条直线（曲线）附近，表明可用直线（曲线）近似地描述变量间的关系。若有多个变量，常制作多幅两两变量间的散点图来考察变量间的关系。

先以男生体重和身高为例，画出散点图。R 中使用函数 plot( )可以方便地画出两个样本的散点图，从而直观地了解对应随机变量之间的相关关系和相关程度。输入命令见附录 5.1。输出结果如下

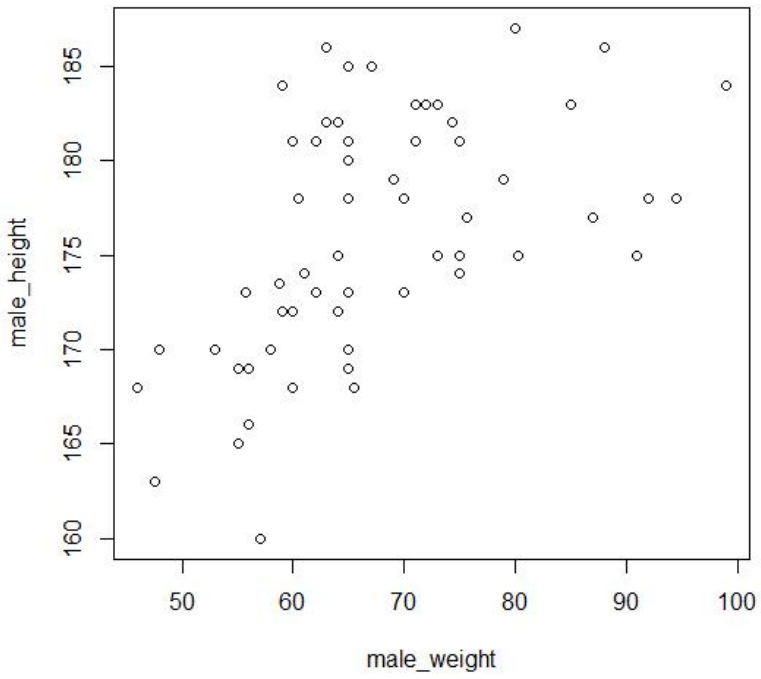


图 5.1

从图上可以直观看出，数据点分布相对较为分散，但观察所有点的分布趋势，又可能存在某种递增的趋向，所以可推测男生体重和身高之间有某种正相关关系。

类似，画出男生身高、体重、臂展、腿长、脚长两两变量间的散点图。从图 5.2 中可以直观看出，任意两组数据点分布都可能存在某种递增的趋向。可推测男生身高、体重、臂展、腿长、脚长两两之间都有正相关关系。女生也类似。

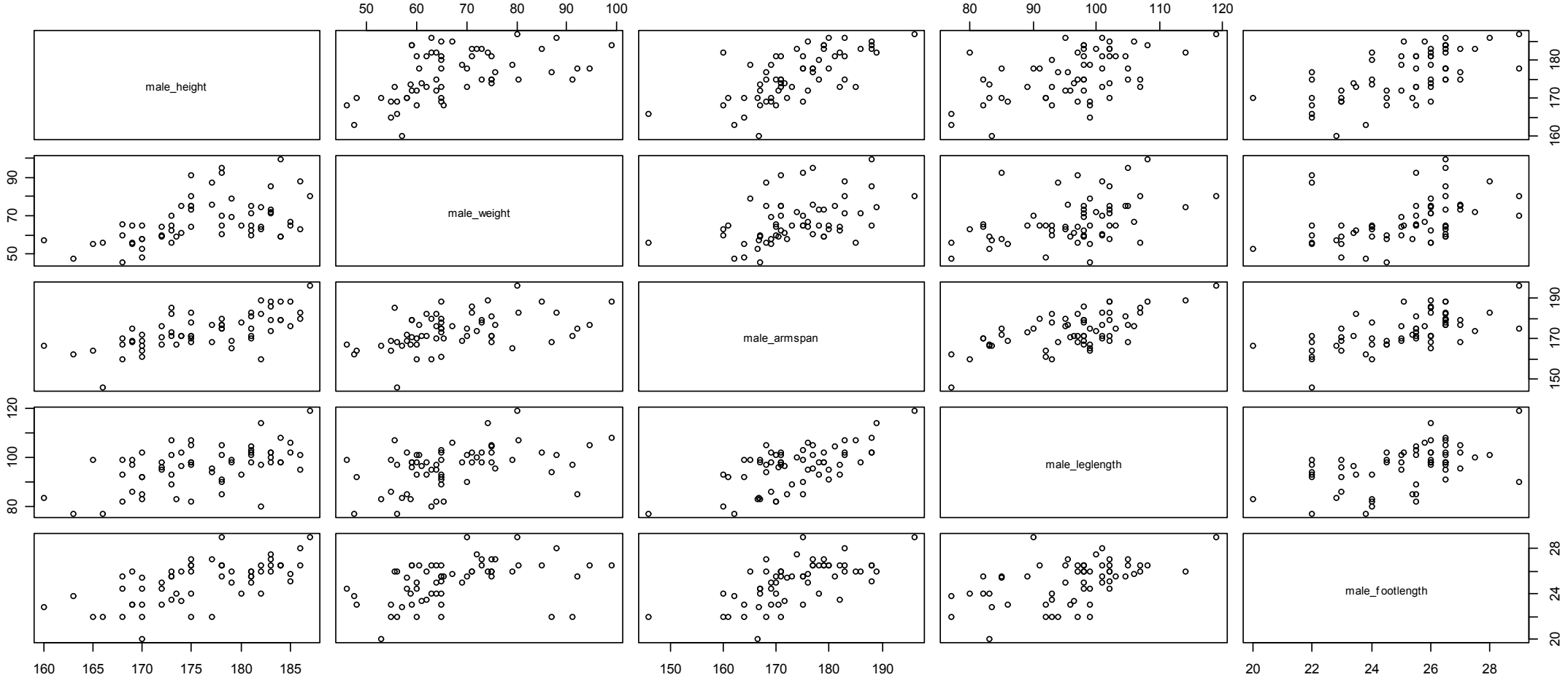

male_height
male_weight
male_armspan
male_leglength
male_footlength

图 5.2

散点图是一种最为有效最为简单的相关性分析工具。若通过散点图可以基本明确它们之间存在直线关系，则可通过线性回归进一步确定它们之间的函数关系。它们之间的相关程度可以用 Pearson 相关系数来刻划。Pearson 相关系数反映了变量间的线性相关程度的大小。

在 R 软件中，cor.test( )提供了 Pearson 相关性检验。以男生身高、体重为例。输入命令和输出结果见附录 5.2。

运行结果为

```
        Pearson's product-moment correlation

data:  male_weight and male_height
t = 4.5635, df = 59, p-value = 2.601e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2971703 0.6756816
sample estimates:
      cor
0.5107684
```

因为 p 值＝2.601e-05<0.05，故拒绝原假设，从而认为身高与体重相关。

类似可得，男生的身高、体重、臂展、腿长、脚长两两变量间，女生的身高、体重、臂展、腿长、脚长两两变量间都相关。

相关分析只能得出两个变量之间是否相关，但却不能回答在两个变量之间存在相关关系时，它们之间是如何联系的，即无法找出刻画它们之间因果关系的函数关系。回归分析就可以解决这一问题。

在 R 中，由函数 lm( )可以非常方便地求出回归方程，函数 confint( )可求出参数的置信区间。与回归分析有关的函数还有 summary( )，用于提取模型计算结果。

本节中仅以男生的体重（因变量）和身高（自变量）的关系为例。

函数 lm( )表示使用线性回归模型 $y=\beta_0+\beta_1 x$。函数 summary( )为提取模型计算结果，包括所求 β_0，β_1 和用于显著性检验的 p 统计量。以男生的体重和身高为例。输入命令与输出结果见附录 5.3。所求得结果如下：

回归系数的估计：回归系数的估计为 $\hat{\beta}_0$=159.22710，$\hat{\beta}_1$=0.27994；相应的标准差为 Sd($\hat{\beta}_0$)=4.18987，Sd($\hat{\beta}_1$)=0.06134。

回归系数的显著性检验：两个回归系数的 p 值均很小（分别为 2e-16 和 2.6e-05），故是非常显著的。

回归方程的显著性检验：F 分布的 p 值为 2.601e-05，故是非常显著的。

可以看出通过显著性检验，由此得到回归方程：male_weight = 157.22710 + 0.27994 male_height

R 中函数 abline( )可以在图中画出回归方程的直线。输入命令见附录 5.4，得到结果为

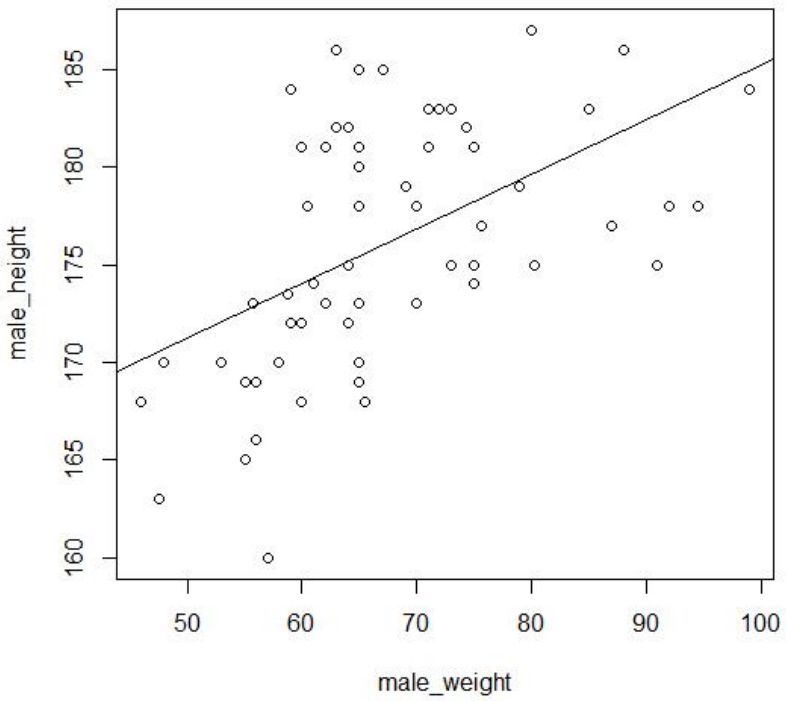


图 5.3

执行如下命令

```
op<-par(mfrow=c(2, 2))
plot(lm.reg)
par(op)
```

运行结果见图 5.4。上面的命令 plot(lm.reg)实际上使用了四次 plot(x, y)，产生四个图形，它们分别为：

1) Residual vs fitted 为拟合值 $\hat{y}$ 对残差的图形。可以看出，数据点都基本均匀地分布在直线 y=0 的两侧，无明显趋势；

2) Normal QQ-plot 图中数据点分布趋于一条直线，说明残差服从正态分布；

3) Scale－Location 图显示了标准化残差(standardized residuals)的平方根的分布情况。最高点为残差最大值点；

4) Cook 距离(Cook's distance)图显示了对回归的影响点。

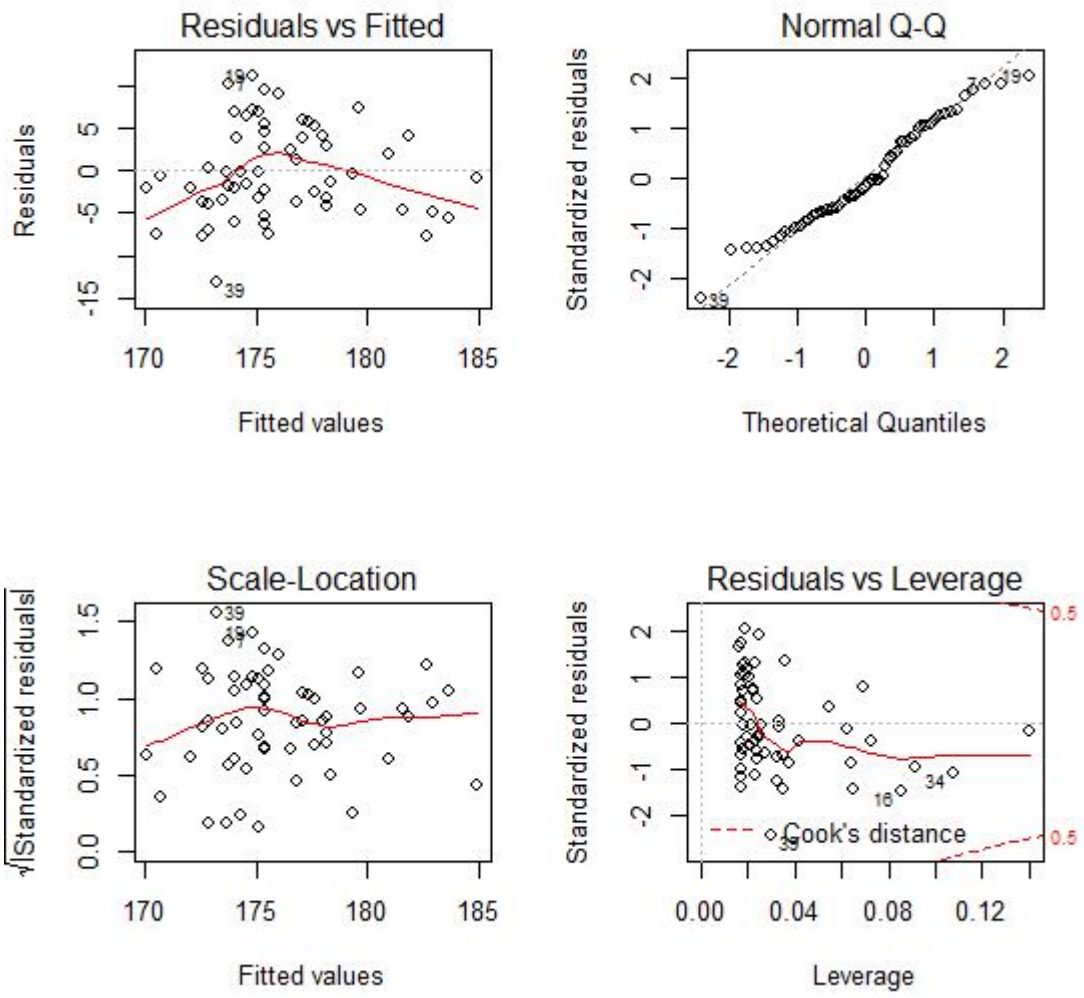


图 5.4

函数 confint( )用来求回归系数的区间估计。输入命令与输出结果见附录 5.5，可得β_0，β_1 的置信度为 95%的区间分别为[148.84, 165,61]和[0.157, 0.403]。

在多元线性回归分析中，我们可以检验自变量：身高、臂展、腿长、脚长与因变量：体重的关系。本节中仅以男生为例。

在 R 中执行如下命令

```
male_weight<-data.frame(male_height, male_weight, male_armspan,
 male_leglength, male_footlength)
```

```
lm.reg<-lm(male_weight~male_height+male_armspan
+male_leglength+male_footlength, data=male_weight)
summary(lm.reg)
```

输出结果为

```
Call:
lm(formula = male_weight ~ male_height + male_armspan +
male_leglength + male_footlength, data = male_weight)

Residuals:
    Min      1Q  Median      3Q     Max
-16.015  -5.906  -2.442   4.850  25.674

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -89.0880    37.5850  -2.370   0.0212 *
male_height       0.6076     0.2988   2.034   0.0468 *
male_armspan      0.1134     0.2380   0.477   0.6354
male_leglength    0.2162     0.2041   1.059   0.2941
male_footlength   0.3544     0.9862   0.359   0.7207
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 10.07 on 56 degrees of freedom
Multiple R-squared:  0.2981,    Adjusted R-squared:  0.2479
F-statistic: 5.945 on 4 and 56 DF,  p-value: 0.0004641
```

可见回归方程的系数的显著性不高：只有身高通过了检验。这说明如果选择全部变量构造方程，效果并不好。这就涉及到变量选择的问题，以建立“最优”的回归方程。

R 软件提供了获得“最优”回归方程的方法——“逐步回归法”。其计算函数为 step( )，它是以 Akaike 信息统计量为准则（简称 AIC 准则），通过选择最小的 AIC 信息统计量，来达到删除或增加变量的目的。

在本例中输入如下指令

```
lm.step<-step(lm.reg)
```

输出结果为

```
Start:  AIC=286.52
male_weight ~ male_height + male_armspan + male_leglength +
male_footlength

                  Df Sum of Sq    RSS    AIC
- male_footlength  1     13.09 5688.6 284.66
- male_armspan     1     23.03 5698.5 284.76
- male_leglength   1    113.69 5789.2 285.73
<none>                        5675.5 286.51
- male_height      1    419.09 6094.6 288.86

Step:  AIC=284.66
male_weight ~ male_height + male_armspan + male_leglength

                  Df Sum of Sq    RSS    AIC
- male_armspan     1     36.42 5725.0 283.05
- male_leglength   1    125.10 5813.7 283.98
<none>                        5688.6 284.66
- male_height      1    539.64 6228.2 288.18

Step:  AIC=283.05
male_weight ~ male_height + male_leglength

                  Df Sum of Sq    RSS    AIC
<none>                        5725.0 283.05
- male_leglength   1    251.29 5976.3 283.67
- male_height      1    930.56 6655.5 290.23
```

用全部变量作回归方程时，AIC 统计量的值为 286.52，如果去掉变量 footlength（脚长），AIC 统计量的值为 284.66；如果去掉变量 armspan（臂展），AIC 统计量的值为 284.76；依次类推,。由于去掉 footlength 使 AIC 统计量达到最小，因此 R 软件会自动去掉变量 footlength 进入下一轮计算。

在下一轮中，去掉 armspan 使得 AIC 统计量达到最小，故去掉 armspan。

在下一轮中，无论去掉哪一个变量，AIC 统计量的值均会升高，因此 R 软件自动终止计算，得到“最优”回归方程。

再用函数 summary( )提取相关回归信息。

```
> summary(lm.step)
```

提取结果为：

```
Call:
lm(formula = male_weight ~ male_height + male_leglength, data =
male_weight)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-16.189    -6.139    -1.703     5.024    26.454

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -88.4312       35.8991    -2.463    0.01675 *
male_height          0.7304         0.2379     3.070    0.00325 **
male_leglength     0.2819         0.1767     1.596    0.11602
---
Signif. codes:    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.935 on 58 degrees of freedom
Multiple R-squared:    0.292,         Adjusted R-squared:    0.2675
F-statistic: 11.96 on 2 and 58 DF,    p-value: 4.484e-05
```

可见回归系数的显著性水平有很大提高，所有的检验均是显著的或较为显著的，由此得到“最优”的回归方程：

male_weight = -88.4312 + 0.7304 male_height + 0.2819 male_leglength

下面研究对回归模型产生较大影响的异常值问题，其主要内容有：残差分析、异常点识别、影响分析、共线性诊断等。

计算在 5.3 节多元回归分析中的残差和标准残差（但不必输出），并画出相应的残差散点图（图 5.5）和标准残差散点图（图 5.6）。输入命令见附录 5.6。

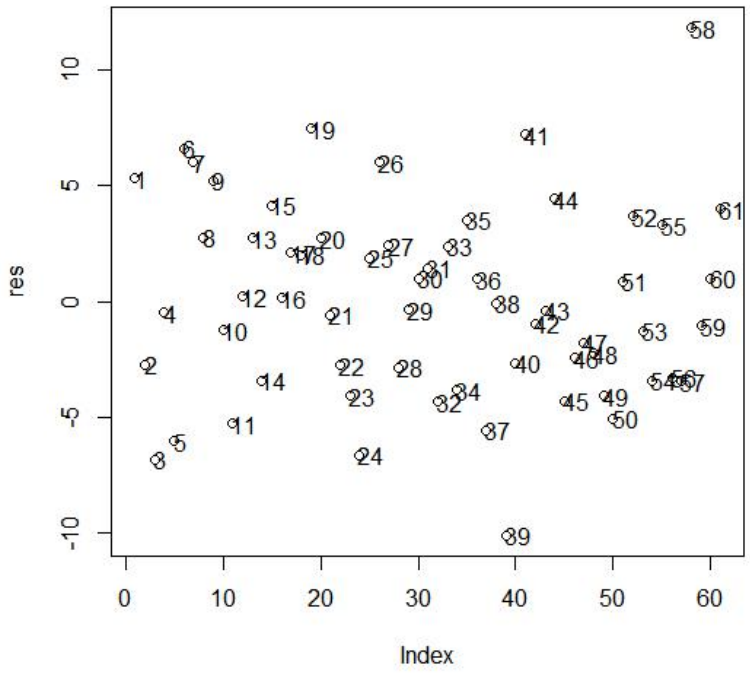


图 5.5

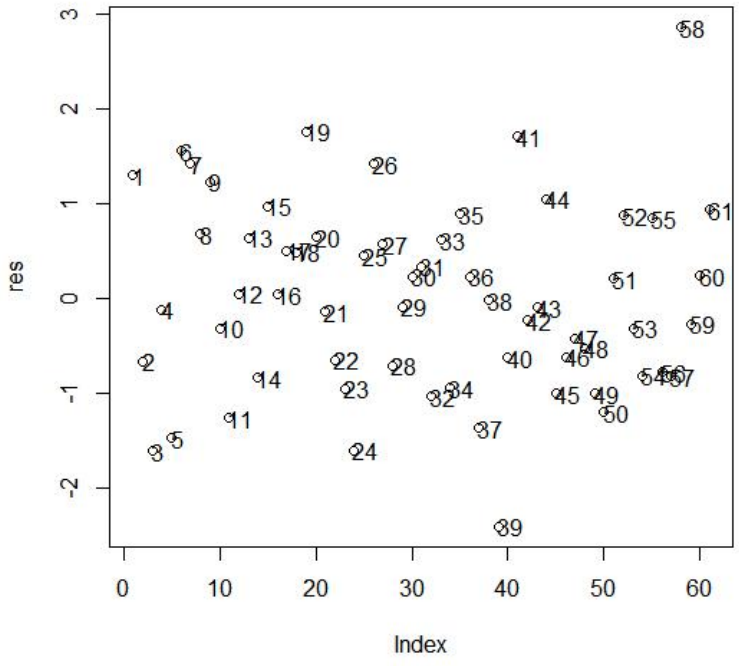


图 5.6

如果多元线性回归模型的假定成立，从理论上可证明 r1；r2；……；rn 相互独立且近似服从 N(0,1)，故关于观测值等的残差图中散点应随机的分布在-2 到+2 的带子里，并称之为正常残差图，否则称为异常残差图。从图 5.6 来看，本例符合正常残差图（仅 39 号和 58 号数据不在-2 到+2 的带子里）。

如果拟合后的模型能够很好地描述这组数据，那么残差对预测值的散点图应该像一些随机散布的点。可是，若某个观测不能和其它数据一起用这个模型表示，那么那个观测的残差通常很大。这里“很大”指的是残差的绝对值。因为一个“很大”的残差可能是正的也可能是负的。如果只有占很小百分比的观测出现大的残差，那么这些观测可能是异常点（outliers）它们不能用来与其余数据一起拟合模型。因此对数据中有残差“很大”的观测点，必须仔细地检查。

一般把标准化残差的绝对值≥2 的观测点认为是可疑点；而标准化残差的绝对值≥3 的观测点认为是异常点。由例的计算结果并结合图形可以看出，第 58 和 39 个点的残差比较大，被认定为异常点。

这里再做一个简单处理，去掉第 58 和 39 观测样本点，并重复上述回归分析及残差分析的过程，得到新的标准化残差图 5.7。与图 5.6 相比，现在残差点的分布已有了很大的改进，它们完全上落在[-2，2]的带状区域内。

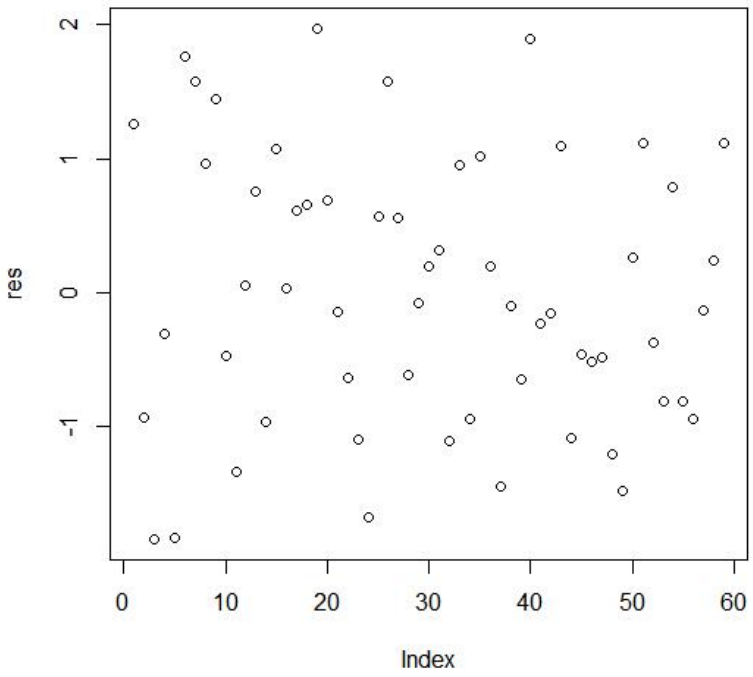


图 5.7

从分析观测点对回归结果的影响入手，找出对回归结果影响很大的观测点的分析方法称为影响分析。在 R 软件中，函数 influence.measures( )可以做回归诊断中影响分析的概括。输入命令与输出结果见附录 5.7，可以看出，第 10、16、33、45、52、54、57 个观测点为强影响点，（附录的结果中已用“*”号标出）。

R

将数据输入 R

```
male_height<-c(185, 173, 175, 182, 173, 181, 184, 179, 181, 187, 169, 178,
  183, 168, 181, 175, 175, 186, 186, 182, 178, 177, 172, 168,
  173.5, 184, 183, 175, 168, 174, 181, 170, 166, 178, 177, 181,
  163, 172, 160, 173, 185, 172, 183, 180, 175, 178, 169, 175,
  165, 169, 170, 183, 184, 174, 170, 173, 170, 182, 178, 170,
  179)

male_weight<-c(65, 62, 80.3, 74.3, 55.7, 60, 59, 79, 62, 80, 56, 60.5,
  73, 46, 65, 91, 64, 88, 63, 64, 65, 75.6, 64, 65.5,
  58.8, 59, 71, 75, 60, 61, 75, 58, 56, 94.5, 87, 71,
  47.5, 59, 57, 65, 67, 60, 85, 65, 73, 70, 55, 75,
  55, 65, 65, 72, 99, 75, 53, 70, 58, 63, 92, 48,
  69)

male_armspan<-c(188, 182, 183, 189, 185, 170, 179, 165, 171, 196, 168, 177,
  179, 167, 175, 171, 170, 183, 180, 182, 180, 177, 176, 170,
  167, 179, 186, 168, 160, 171.5, 181, 172, 146, 177, 168, 183,
  162, 170.5, 166.7, 173, 176, 167, 188, 178, 178, 175, 169, 171,
  164, 175, 161, 174, 188, 171, 166.6, 171, 169, 160, 175, 164,
  169)

male_leglength<-c(102, 93, 107, 114, 107, 101, 98, 99, 98, 119, 97, 101,
  102, 99, 103, 97, 82, 101, 95, 97, 91, 95.5, 95, 82,
  83, 98, 98, 105, 93, 96.5, 104.5, 85, 77, 105, 94, 102,
  77, 95.8, 83.5, 89, 106, 98, 102, 93, 98, 90, 86, 98,
  99, 99, 92, 100, 108, 102, 83, 101, 102, 80, 85, 92,
  98)

male_footlength<-c(25.1, 23.5, 26.5, 26, 26, 25, 26.5, 26, 26, 29, 26, 26.5,
  27, 24.5, 25.5, 22, 24, 28, 26.5, 26.5, 26.5, 27, 25, 25.5,
  24, 26.5, 26, 27, 22, 23.4, 25.5, 25.4, 22, 26.5, 22, 26,
  23.8, 23, 22.8, 25.5, 25.8, 24.5, 26.5, 24, 26.5, 29, 23, 26,
  22, 23, 22, 27.5, 26.5, 26, 20, 25.5, 24.5, 24, 25.5, 23,
  25)

female_height<-c(159, 172, 163, 165, 168, 165, 163, 165, 160, 158, 168, 162,
  161, 172, 168, 168, 174, 161, 162, 166, 162, 162, 170, 168,
```

```
  159, 168, 164, 155, 166)

female_weight<-c(47.9, 54, 60, 53, 52, 53, 58.5, 55, 50, 46.5, 58, 46,
  51, 63, 66.5, 52, 56, 44, 57.5, 54.5, 53, 48, 60, 59,
  54.5, 68, 55, 42, 51)

female_armspan<-c(158.8, 173, 163, 164, 169, 166, 156, 165, 161, 152, 166,
148,
  152, 174, 162, 166, 179.5, 160, 164, 159, 160, 155, 169, 168,
  136.6, 149, 158, 155, 165)

female_leglength<-c(100, 105, 102, 97, 97.5, 94, 88, 94, 89, 92, 95, 90,
  91, 104, 99, 97, 105, 90, 88, 94, 94, 97, 104, 82,
  83, 100, 90, 85, 95)

female_footlength<-c(22.8, 23.5, 24, 22.6, 24.8, 21.5, 23.5, 23.5, 23, 22.5, 23.5,
21.4,
  22.6, 24.5, 24.5, 23.5, 24.5, 23, 24.3, 24, 24, 23, 24.5, 24,
  24.7, 23.5, 23.5, 25, 23)
```

1.1 总体描述（仅列出男生身高的输入命令和输出结果，其他类似）

```
> sort(male_height)

 [1] 160.0 163.0 165.0 166.0 168.0 168.0 168.0 169.0 169.0 169.0 170.0 170.0
[13] 170.0 170.0 170.0 172.0 172.0 172.0 173.0 173.0 173.0 173.0 173.5 174.0
[25] 174.0 175.0 175.0 175.0 175.0 175.0 175.0 177.0 177.0 178.0 178.0 178.0
[37] 178.0 178.0 179.0 179.0 180.0 181.0 181.0 181.0 181.0 181.0 182.0 182.0
[49] 182.0 183.0 183.0 183.0 183.0 184.0 184.0 184.0 185.0 185.0 186.0 186.0
[61] 187

> summary(male_height)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  160.0   172.0   175.0   176.1   181.0   187.0

> var(male_height)

[1] 40.48197

> sd(male_height)

[1] 6.362544
```

1.2 绘制男生身高直方图

```
> hist(male_height, breaks=160+(0:9)*3,
+ xlim=c(min(male_height),max(male_height)), col='lightblue')
```

1.3 绘制男生身高经验分布函数

```
> x <- sort(male_height)
> n <- length(x)
> y <- (1:n)/n
> m <- mean(male_height)
> s <- sd(male_height)
> plot(x,y, type='s', main="empirical cdf of male_height")
```

2.1 方差未知时均值区间估计（仅列出男生身高的输入命令和输出结果，其他类似）

```
>    t.test(male_height)

        One Sample t-test

data:    male_height
t = 216.14, df = 60, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  174.4442 177.7033
sample estimates:
mean of x
  176.0738
```

2.2 均值未知时方差区间估计（仅列出男生身高的输入命令和输出结果，其他类似）

```
> interval_var1<-function(x,mu=Inf,alpha=0.05){
+ n<-length(x)
+ if (mu<Inf){
+ S2 <- sum((x-mu)^2)/n; df <- n
+ }
+ else{
+ S2 <- var(x); df <- n-1
+ }
+ a<-df*S2/qchisq(1-alpha/2,df)
+ b<-df*S2/qchisq(alpha/2,df)
+ data.frame(var=S2, df=df, a=a, b=b)
```

```
+ }
> interval_var1(male_height)

          var df        a        b
1 40.48197 60 29.15949 60.00032
```

3.1 绘制 QQ 图

```
> qqnorm(male_height,
+ main="Normality Check via QQ Plot")
> qqline(male_height, col='red')
```

3.2 与正态分布密度函数比较

```
dens <- density(male_height)
xlim <- range(dens$x); ylim<-range(dens$y)
hist(male_height,breaks=160+(0:9)*3,
xlim=xlim,ylim=ylim,
probability=T)

lines(dens,col=par('fg'),lty=2)
m <- mean(male_height)

s <- sd(male_height)
curve( dnorm(x, m, s), col='red', add=T)
hist(male_height,breaks=160+(0:9)*3,
xlim=xlim,ylim=ylim,
probability=T)

lines(dens,col=par('fg'),lty=2)
m <- mean(male_height)
s <- sd(male_height)
curve( dnorm(x, m, s), col='red', add=T)
```

3.3 经验分布函数与正态分布函数比较

```
> x <- sort(male_height)
> n <- length(x)
> y <- (1:n)/n
> m <- mean(male_height)
> s <- sd(male_height)
> plot(x,y, type='s', main="empirical cdf of ")
> curve(pnorm(x,m,s),col='red', lwd=2, add=T)
```

3.4 方差未知时检验均值

```
> t.test(male_height, mu=171.9)

        One Sample t-test

data:  male_height
t = 5.1235, df = 60, p-value = 3.36e-06
alternative hypothesis: true mean is not equal to 171.9
95 percent confidence interval:
 174.4442 177.7033
sample estimates:
mean of x
 176.0738
```

4.1 平方和的检验与分解

```
X<-c(185, 173, 175, 182, 173, 181, 184, 179, 181, 187, 169, 178,
  183, 168, 181, 175, 175, 186, 186, 182, 178, 177, 172, 168,
  173.5, 184, 183, 175, 168, 174, 181, 170, 166, 178, 177, 181,
  163, 172, 160, 173, 185, 172, 183, 180, 175, 178, 169, 175,
  165, 169, 170, 183, 184, 174, 170, 173, 170, 182, 178, 170,
  179,
  159, 172, 163, 165, 168, 165, 163, 165, 160, 158, 168, 162,
  161, 172, 168, 168, 174, 161, 162, 166, 162, 162, 170, 168,
  159, 168, 164, 155, 166)
A<-factor(c(rep(1,61),  rep(2,29)) )
miscellany<-data.frame(X, A)
aov.mis<-aov(X~A, data=miscellany)
summary(aov.mis)

            Df Sum Sq Mean Sq F value     Pr(>F)
A            1   2578  2578.3   75.44 1.82e-13 ***
Residuals   88   3008    34.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4.2 绘图

```
> plot(miscellany$X~miscellany$A)
```

5.1 绘制散点图

```
> level <- data.frame(male_weight, male_height)
> plot(level)
```

5.2 男生身高、体重的相关分析

```
> attach(level)
> cor.test(male_weight, male_height)

        Pearson's product-moment correlation

data:  male_weight and male_height
t = 4.5635, df = 59, p-value = 2.601e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2971703 0.6756816
sample estimates:
      cor
0.5107684
```

5.3 男生体重与身高的β0，β1 的估计与显著性检验

```
lm.reg<-lm(male_height~1+male_weight)
summary(lm.reg)

Call:
lm(formula = male_height ~ 1 + male_weight)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1839  -3.8232  -0.9416   4.1378  11.1364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.22710    4.18987  37.526  < 2e-16 ***
male_weight   0.27994    0.06134   4.563  2.6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.516 on 59 degrees of freedom
Multiple R-squared:  0.2609,    Adjusted R-squared:  0.2484
F-statistic: 20.83 on 1 and 59 DF,  p-value: 2.601e-05
```

5.4 画出回归方程的直线

```
abline(lm.reg)
```

5.5β0，β1 的区间估计

```
confint(lm.reg, level=0.95)
```

```
                    2.5 %       97.5 %
(Intercept) 148.8431982 165.6110086
male_weight   0.1571935   0.4026948
```

5.6 计算残差和标准残差，并画出相应的残差散点图

```
res<-residuals(lm.reg)
plot(res)
for(i in 1:61)
{text(i, res[i], labels=i, adj=(.05))}

res<-rstandard(lm.reg)
plot(res)
for(i in 1:61)
{text(i, res[i], labels=i, adj=(.05))}
```

5.7 影响分析

```
height<-data.frame(male_height, male_weight, male_armspan,
  male_leglength, male_footlength)
lm.reg<-lm(male_height~male_weight+male_armspan
+male_leglength+male_footlength, data=blood)
lm.step<-step(lm.reg)
summary(lm.step)
influence.measures(lm.reg)
```

最后一句命令的结果如下：

```
Influence measures of
        lm(formula = male_height ~ male_weight + male_armspan +
male_leglength +     male_footlength, data = blood) :

     dfb.1_    dfb.ml_w    dfb.ml_r    dfb.ml_l dfb.ml_f     dffit cov.r     cook.d
1  -0.28362 -0.139347    3.66e-01 -0.021482 -0.20362    0.4327 1.057 3.70e-02
2   0.11743    0.066427  -2.85e-01    0.118043    0.20755  -0.3341  1.145
2.24e-02
3   0.20558 -0.147290 -4.98e-02 -0.164003    0.02763 -0.4360 0.837 3.63e-02
4   0.06737    0.013978 -4.38e-02 -0.064053    0.04300 -0.1131 1.233 2.60e-03
```

5 0.37499 0.479750 -2.90e-01 -0.249670 0.07470 -0.6695 0.898 8.57e-02
6 0.11077 -0.172490 -1.99e-01 0.269900 0.04887 0.4065 0.860 3.18e-02
7 -0.16136 -0.246137 1.03e-01 -0.032603 0.14620 0.3700 0.915 2.66e-02
8 0.17612 0.168512 -2.94e-01 0.100977 0.14567 0.3499 1.138 2.45e-02
9 0.05689 -0.110240 -1.55e-01 0.084180 0.17191 0.2981 0.940 1.74e-02
10 0.16678 0.027030 -5.79e-02 -0.093319 -0.02921 -0.2123 1.292 9.15e-03
11 -0.09989 0.190432 2.10e-01 -0.108040 -0.22194 -0.3725 0.998 2.73e-02
12 -0.00320 -0.007001 -6.56e-04 0.003086 0.00509 0.0112 1.150 2.55e-05
13 -0.05210 0.007088 -1.86e-02 0.020868 0.07620 0.1435 1.079 4.15e-03
14 -0.08084 0.255161 1.10e-01 -0.174761 -0.04085 -0.3383 1.128 2.29e-02
15 -0.01068 -0.068379 -4.90e-02 0.135452 0.01042 0.2039 1.021 8.29e-03
16 0.00426 0.011232 9.32e-05 0.000845 -0.00974 0.0147 1.396 4.41e-05
17 0.03880 0.023748 5.88e-02 -0.157960 -0.01082 0.1864 1.159 7.03e-03
18 -0.07240 0.120310 7.53e-04 -0.052674 0.08849 0.2053 1.162 8.52e-03
19 -0.22424 -0.191368 2.14e-01 -0.209856 0.16907 0.4638 0.796 4.07e-02
20 -0.09554 -0.066326 9.03e-02 -0.056601 0.03809 0.1581 1.106 5.05e-03
21 0.01547 0.007569 -1.94e-02 0.027368 -0.01299 -0.0384 1.179 3.00e-04
22 0.02338 -0.047197 1.12e-02 0.061629 -0.08626 -0.1426 1.112 4.11e-03
23 0.03404 0.049610 -7.93e-02 0.050016 0.02368 -0.1720 1.005 5.90e-03
24 -0.09378 -0.067342 -5.00e-02 0.487912 -0.21756 -0.5736 0.934 6.35e-02
25 0.05688 -0.012570 1.57e-02 -0.108935 0.01088 0.1512 1.142 4.63e-03
26 -0.16136 -0.246137 1.03e-01 -0.032603 0.14620 0.3700 0.915 2.66e-02
27 -0.10215 -0.009644 1.23e-01 -0.060711 -0.02706 0.1484 1.143 4.46e-03
28 -0.07392 -0.044129 2.00e-01 -0.125498 -0.12811 -0.2390 1.221 1.16e-02
29 -0.02018 -0.000978 1.21e-02 -0.009819 0.01130 -0.0259 1.207 1.36e-04
30 0.00938 -0.009923 5.52e-03 0.011230 -0.02470 0.0392 1.140 3.13e-04
31 -0.02609 0.012102 1.78e-02 0.024634 -0.02304 0.0639 1.132 8.29e-04
32 -0.00838 0.088035 -7.74e-02 0.241951 -0.11084 -0.3208 1.060 2.05e-02
33 0.46713 0.076246 -3.30e-01 -0.085470 0.06538 0.5016 1.286 5.04e-02
34 -0.01793 -0.290469 1.24e-01 -0.069965 -0.02578 -0.3543 1.154 2.52e-02
35 0.19116 0.364587 -3.30e-02 0.000948 -0.29334 0.4762 1.213

4.53e-02

36	-0.02479	-0.003745	2.16e-02	0.001305	-0.00579	0.0372	1.136	2.82e-04
37	-0.23303	0.195306	-7.32e-04	0.375407	-0.13705	-0.5834	1.041	6.66e-02
38	-0.00635	0.006666	-3.34e-03	-0.006348	0.01520	-0.0230	1.155	1.08e-04
39	-0.00635	0.003536	-2.26e-02	0.099351	-0.04673	-0.1366	1.104	3.77e-03
40	-0.04156	-0.109398	-1.30e-01	0.335766	0.03032	0.4370	0.818	3.63e-02
41	-0.02513	0.015047	2.98e-02	-0.027477	-0.00422	-0.0495	1.146	5.00e-04
42	0.02885	-0.021714	-2.78e-02	0.012535	0.00912	-0.0471	1.198	4.51e-04
43	-0.06015	-0.022175	1.94e-01	-0.106445	-0.14600	0.2611	1.039	1.36e-02
44	0.05715	-0.037151	-9.98e-03	0.044821	-0.08242	-0.1863	1.011	6.92e-03
45	0.01907	-0.001495	4.14e-02	0.103539	-0.18373	-0.2105	1.307	8.99e-03
46	-0.03604	0.040215	-5.00e-02	0.064098	0.05127	-0.1352	1.146	3.70e-03
47	-0.03284	-0.045425	6.51e-02	-0.014020	-0.04547	-0.1019	1.123	2.11e-03
48	-0.21378	0.115182	1.20e-01	-0.273961	0.23181	-0.4381	1.084	3.81e-02
49	-0.01066	0.025452	-1.43e-01	-0.112766	0.32750	-0.3950	0.956	3.05e-02
50	0.06216	0.020382	-3.27e-02	0.019786	-0.03988	0.0797	1.194	1.29e-03
51	0.00329	0.019009	-1.58e-01	0.032803	0.23398	0.2933	1.045	1.71e-02
52	0.05763	-0.118390	-3.56e-02	-0.007227	0.03996	-0.1587	1.282	5.12e-03
53	-0.05916	-0.066010	1.41e-01	-0.094379	-0.06784	-0.1970	1.092	7.81e-03
54	0.10763	-0.038090	1.69e-01	-0.102181	-0.29855	0.3819	1.279	2.94e-02
55	-0.05182	-0.020186	1.06e-01	-0.090208	-0.03803	-0.1613	1.073	5.24e-03
56	-0.07414	0.109356	1.12e-01	-0.182591	0.00952	-0.2482	1.082	1.23e-02
57	-0.00511	-0.051442	-8.18e-03	0.046188	-0.00239	-0.0669	1.378	9.13e-04
58	0.03350	-0.041119	-1.38e-02	0.016618	-0.01312	0.0676	1.182	9.32e-04
59	0.10581	0.041077	-1.41e-01	0.083860	0.03356	0.2077	1.012	8.59e-03

	hat	inf
1	0.1048	
2	0.1151	
3	0.0510	
4	0.1175	
5	0.1132	
6	0.0489	
7	0.0509	
8	0.1159	
9	0.0402	
10	0.1677	*
11	0.0708	
12	0.0457	
13	0.0350	
14	0.1086	
15	0.0346	

16 0.2133 *
17 0.0851
18 0.0914
19 0.0498
20 0.0502
21 0.0705
22 0.0490
23 0.0239
24 0.1010
25 0.0669
26 0.0509
27 0.0669
28 0.1320
29 0.0908
30 0.0399
31 0.0391
32 0.0768
33 0.2161 *
34 0.1243
35 0.1786
36 0.0366
37 0.1363
38 0.0500
39 0.0431
40 0.0482
41 0.0461
42 0.0853
43 0.0541
44 0.0283
45 0.1764 *
46 0.0651
47 0.0434
48 0.1159
49 0.0654
50 0.0865
51 0.0647
52 0.1548 *
53 0.0552
54 0.1902 *
55 0.0380
56 0.0655
57 0.2048 *
58 0.0759
59 0.0336