1 数据介绍

本文是基于一个关于子女与父母身高调查的分析报告，被调查者全部是大学生。原始数据共有 195 条记录，包含 6 个变量，其中问卷序号属于无关变量，5 个有效变量分别是：性别（sex）、本人身高（high）、本人体重（weight）、父亲身高（father）、母亲身高（mother）。主要研究目的有两个，一是基于调查数据分析大学男女生的体质差异以及后代与父母的身高差异；二是对子女身高的预测。

```
survey <- read.table("~/LearningR/Data/survey2014_student.csv",sep =
";",header = T,fileEncoding = "GB2312")
df <- survey
head(df)
```

```
   ##   no sex high weight father mother
## 1  1   男  165    100    163    155
## 2  2   男  180    155    158    162
## 3  3   女  170    115    175    158
## 4  4   男  181    152    178    176
## 5  5   女  153     72    178    160
## 6  6   女  160     90    176    156
```

2 数据清理

在进行分析工作之前，我们需要对数据进行清理。通过观察原始数据以及其描述统计量可以发现该数据中主要存在的几个问题：

（1）存在缺失值。由于数据中存在缺失值的观测记录只有 3 条，数量很少，所以我们直接删除包含缺失值的记录，不会对分析结果造成很大影响。

（2）“体重”变量最小值为 40，中位数为 104，存在单位不统一问题。需要注意的是，男女生在处理这一问题时需要区别对待。同样是体重 80，对男生来说，80 公斤的可能性大，而对于女生来说，却是 80 斤的可能性较大。观察数据后，我们将男生大于 80 和女生大于 70 的体重数除以 2，将体重单位统一为“kg”。

（3）数据中存在极大的异常值，说明存在不真实数据。BMI 值是用来判断肥胖程度的指标，如果用 BMI 过高或过低来决定数据是否真实，可能会将一些过胖或者过瘦的同学错判成虚假数据，另一方面，如果某个被调查者的身高和体重都填写了极大的异常值，其 BMI 值也有可能在正常范围内。因此，认为用 BMI 来判断数据真实性并不合理，这里我们认为成年人的正常身高范围为 130cm～200cm，正常体重范围为 30kg～100kg，删除每个变量下过大或过小的值。

```
summary(df[-1])
```

```
##  sex          high               weight             father
##  女:97   Min.   :1.110e+02   Min.   :4.000e+01   Min.   :1.110e+02
```

```
##  男:98   1st Qu.:1.620e+02   1st Qu.:7.600e+01   1st Qu.:1.680e+02
##          Median :1.690e+02   Median :1.040e+02   Median :1.720e+02
##          Mean   :5.815e+33   Mean   :1.203e+27   Mean   :2.303e+29
##          3rd Qu.:1.750e+02   3rd Qu.:1.250e+02   3rd Qu.:1.750e+02
##          Max.   :1.122e+36   Max.   :2.333e+29   Max.   :4.444e+31
##          NA's   :2           NA's   :1           NA's   :2
##      mother
##  Min.   :1.500e+01
##  1st Qu.:1.580e+02
##  Median :1.600e+02
##  Mean   :2.879e+30
##  3rd Qu.:1.650e+02
##  Max.   :5.556e+32
##  NA's   :2
```

```
#删除缺失值、无关变量及重复记录
df <- na.omit(df[-1])
df <- df[!duplicated(df),]
#统一体重单位
df[which(df$sex=="男" & df$weight>80),3] <- df[which(df$sex=="男" &
df$weight>80),3]/2
df[which(df$sex=="女" & df$weight>70),3] <- df[which(df$sex=="女" &
df$weight>70),3]/2
#删除异常值
df[c(2,4,5)][df[c(2,4,5)]>200|df[c(2,4,5)]<130] <- NA
df[3][df[3]>100|df[3]<30] <- NA
df <- na.omit(df)
```

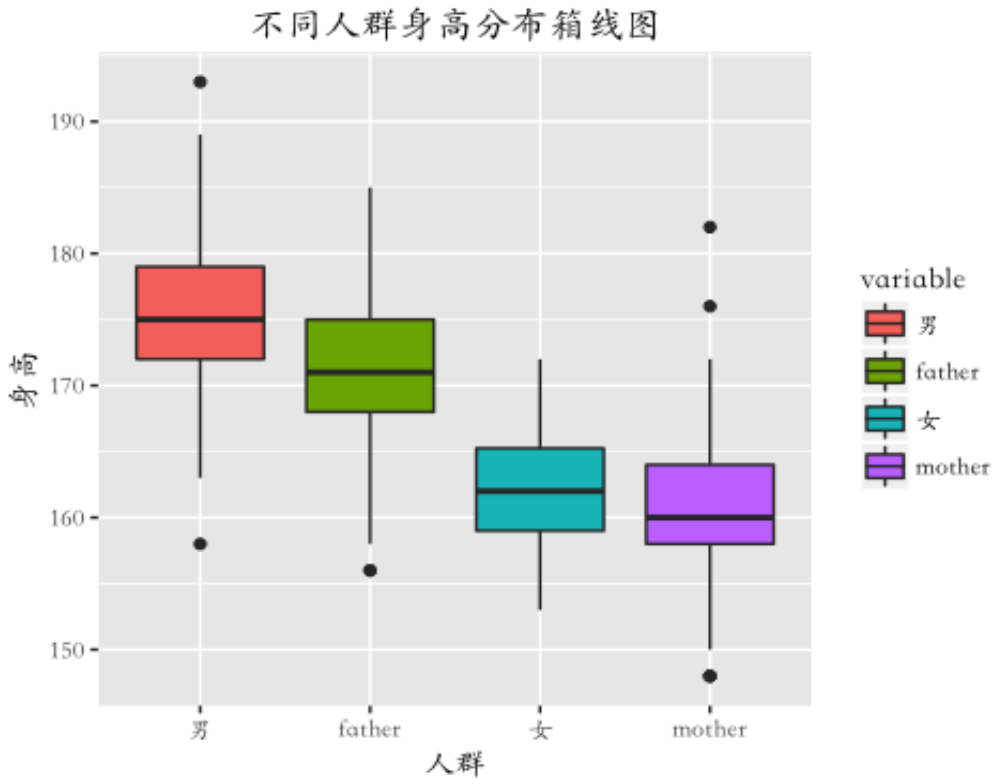
清理后的数据包含 177 条记录，5 个变量，通过描述统计量我们可以看出数据已经基本正常，可以使用。

3 描述性分析

为了对比不同人群的身高差异，我们做出男同学、女同学、父亲、母亲的身高分布箱线图如下：

```
library(ggplot2)
library(reshape2)
library(car)
library(gridExtra)
#男女生身高、体重差异
mdf <- melt(df,id=c("sex","weight"))
mdf$variable <- factor(mdf$variable,levels = c("男","high","father","女
","mother"))
mdf$variable[1:177] <- mdf$sex[1:177]
ggplot(mdf,aes(x=variable,y=value,fill=variable))+
  geom_boxplot()+
  labs(title="不同人群身高分布箱线图",x="人群",y="身高")+
```

```
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))
```



上图清晰的显示了男女身高的差异，男性的身高普遍高于女性身高，但男性身高的方差也比女性稍大。另一方面，虽然男生与父亲、女生与母亲的身高分布相似，但男（女）生身高的均值明显的超过了父（母）亲，说明身高受后天营养条件等的影响很大。

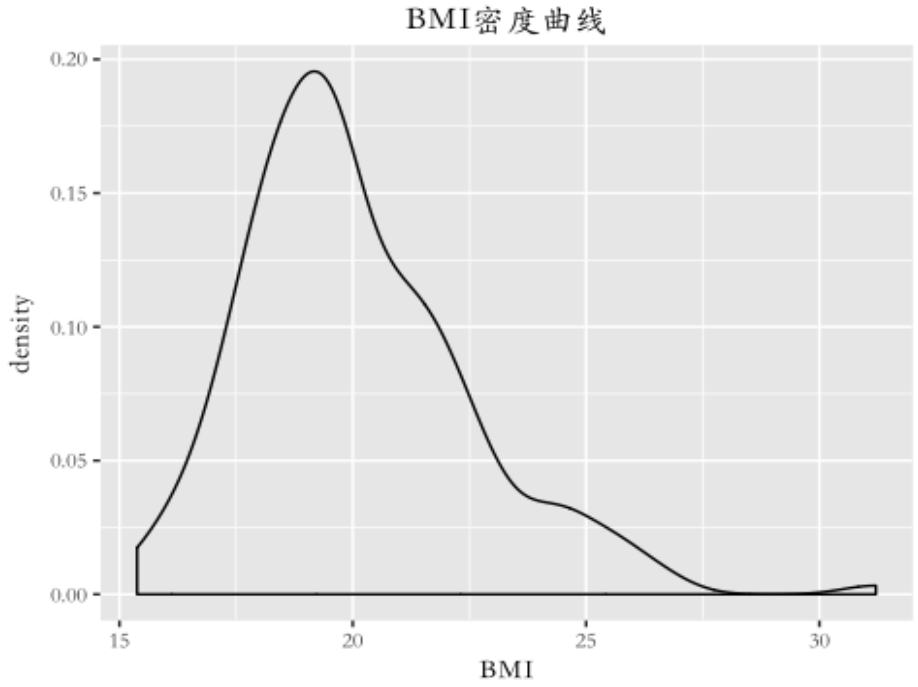
国际上常用 BMI 值，即体重（kg）与身高（米）平方的比值，作为衡量人体胖瘦程度以及是否健康的一个标准。成人的 BMI 标准为： 过轻：<18.5；正常：18.5-23.9；过重：24-27；肥胖：28-32；非常肥胖：>32。

我们可以根据调查结果计算出被调查者的 BMI 值，观察大学生群体的体质情况。

```
df$BMI <- df$weight/(df$high/100)^2
df$BMI1[df$BMI<18.5] <- "过轻"
df$BMI1[df$BMI>=18.5&df$BMI<24] <- "正常"
df$BMI1[df$BMI>24&df$BMI<=27] <- "过重"
df$BMI1[df$BMI>27&df$BMI<32] <- "肥胖"
df$BMI1[df$BMI>32] <- "非常肥胖"
summary(df$BMI)
```

```
    ##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.38   18.46   19.62   20.06   21.38   31.21
```

```
ggplot(df,aes(x=BMI))+geom_density()+
  labs(title="BMI 密度曲线")+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))
```



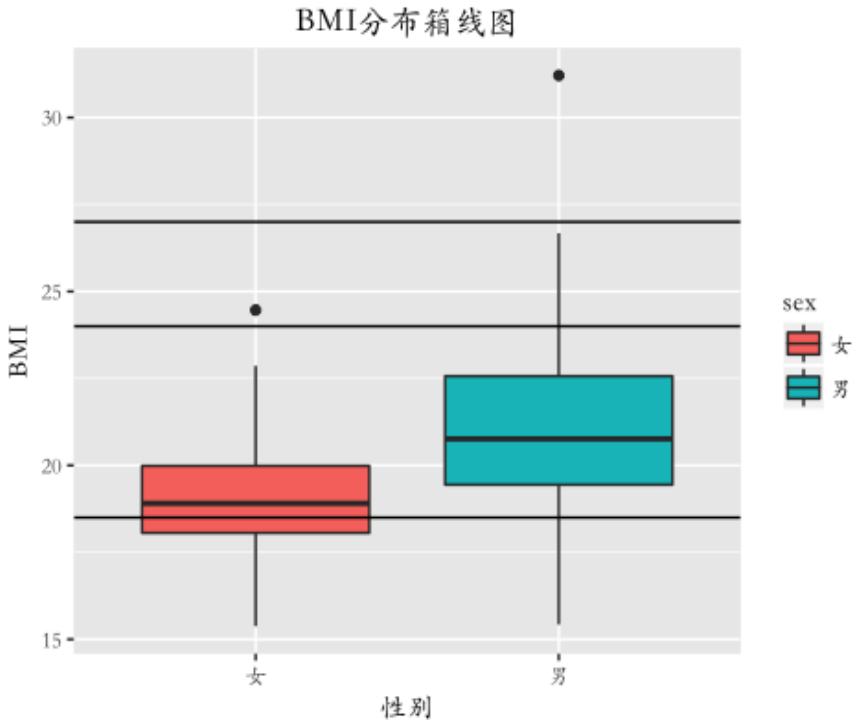

```
options(digits=3)
table(df$sex,df$BMI1)
##       正常 肥胖 过轻 过重
##   女   56    0   35    1
##   男   58    1   13   13

addmargins(prop.table(table(df$sex,df$BMI1)))
##           正常     肥胖     过轻     过重      Sum
##   女  0.31638 0.00000 0.19774 0.00565 0.51977
##   男  0.32768 0.00565 0.07345 0.07345 0.48023
##   Sum 0.64407 0.00565 0.27119 0.07910 1.00000
```

从 BMI 的密度曲线可以看出,大学生的 BMI 总体呈现右偏趋势，均值在 20 左右，大多数同学处在正常范围内，存在少数肥胖人群，具体分布比例可以通过列联表看出：在我们的调查人群中，65%左右的同学 BMI 正常;27%的同学过轻，其中女生占大多数；8%左右的同学过重，几乎全部集中在男生；肥胖人数较少，不存在非常肥胖人群。显然，男女生 BMI 特征不大相同，男生群体总体正常，个别过重；而超过 1/3 的女生体重过轻，需要引起注意。下图是男女生的 BMI 分布箱线图，更直观的显示了不同性别的特点。

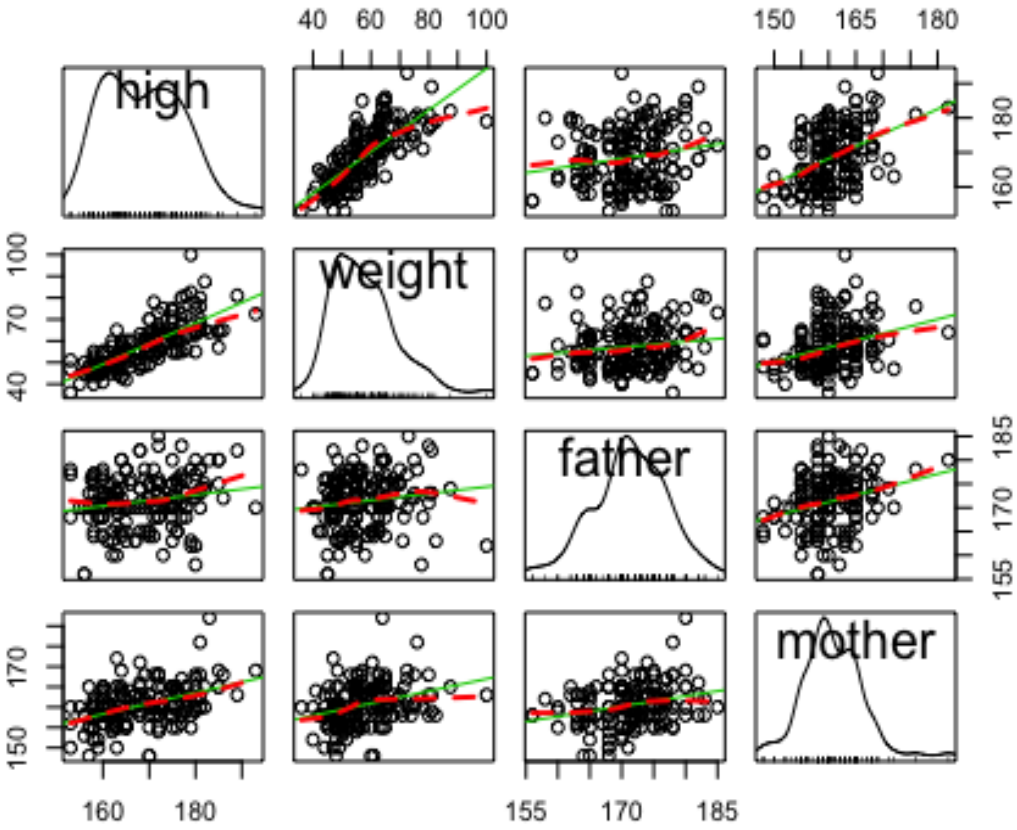
```
ggplot(df,aes(x=sex,y=BMI,fill=sex))+
  geom_boxplot()+
  geom_hline(yintercept =c(18.5,24,27))+
```

```
  labs(title="BMI 分布箱线图",x="性别",y="BMI")+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))
```



一般认为，受到遗传因素的影响，子女的身高应该和父母的身高有关，但从散点图看这种相关特征似乎并不明显。在散点图矩阵中可以看到，变量 high 没有随着变量 father 增加而增加的趋势，随变量 mother 增加的趋势也不明显，子女身高与父母身高间的相关系数也很小，这显然有悖常理。究其原因，是我们忽略了性别差异。男女生的身高分布具有明显差别，在研究子女与父母间的身高关系时，应该分开观察。我们以不同颜色表示男、女生的身高，重新绘制散图，子女身高和父母身高间的关系便清晰呈现。

```
scatterplotMatrix(df[2:5],spread = FALSE,smoother.args = list(lty=2))
```



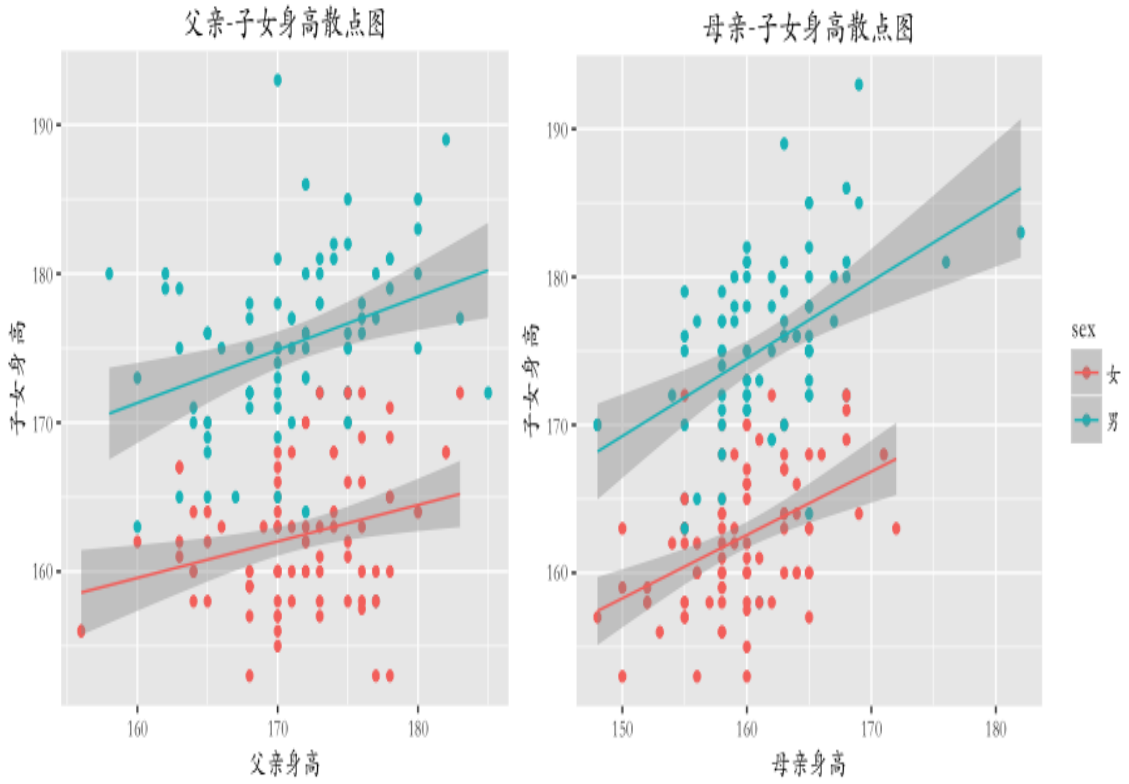

```
cor(df[2:5])
```

```
    ##         high weight father mother
## high   1.000  0.759  0.181  0.434
## weight 0.759  1.000  0.137  0.313
## father 0.181  0.137  1.000  0.272
## mother 0.434  0.313  0.272  1.000
```

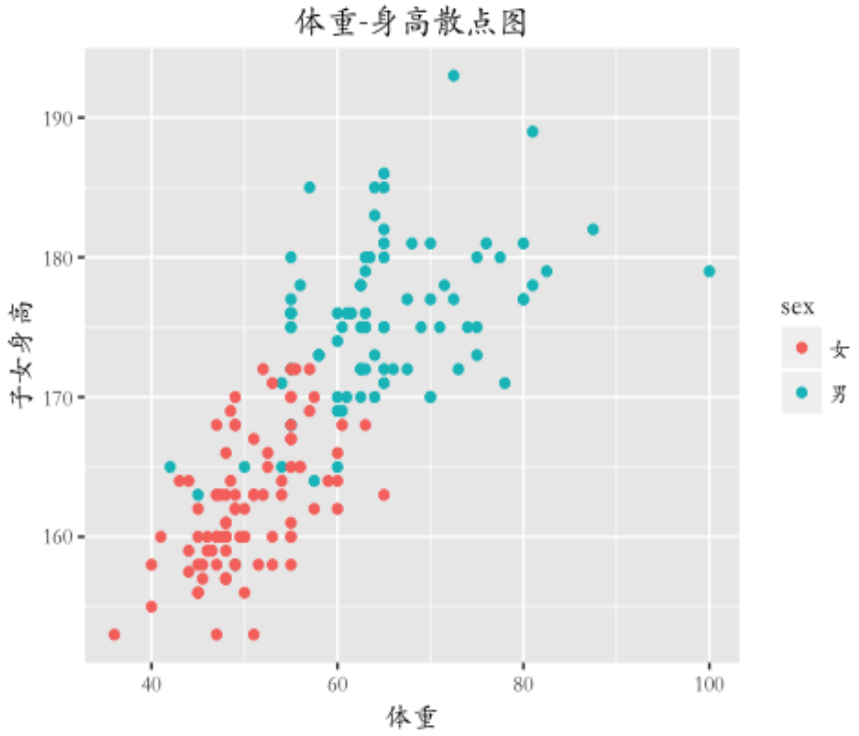
```
ggplot(df,aes(x=father,y=high,color=sex))+
  geom_point()+
  labs(title="父亲-子女身高散点图",x="父亲身高",y="子女身高")+
  geom_smooth(method = lm,size=0.5)+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))

ggplot(df,aes(x=mother,y=high,color=sex))+
  geom_point()+
  labs(title="母亲-子女身高散点图",x="母亲身高",y="子女身高")+
  geom_smooth(method = lm,size=0.5)+
```

```
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))
```




```
ggplot(df,aes(x=weight,y=high,color=sex))+
  geom_point()+
  labs(title="体重-身高散点图",x="体重",y="子女身高")+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(family="STKaiti"))
```



3 预测子女身高

3.1 回归分析

通过描述性分析我们可以发现，子女的身高明显和自身体重以及父母身高有关，性别差异使得子女身高和父（母）亲之间的拟合直线具有不同的截距，但斜率大致一致，因此我们可以用包含性别虚拟变量的多元线性回归来探究变量间的线性关系。为了检验模型的预测效果，我们将数据集的 80%作为训练集，20%作为测试集。

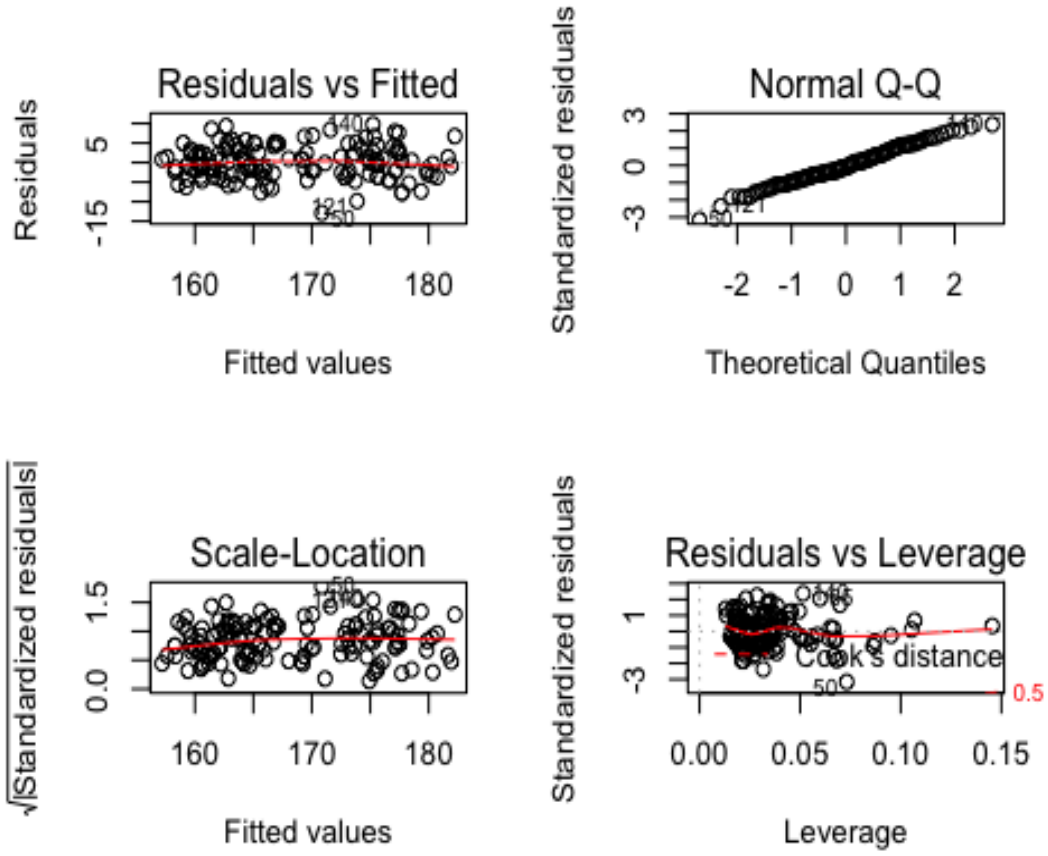
```
#划分训练集
set.seed(1234)
n <- sample(nrow(df),0.8*nrow(df),replace = FALSE)
training <- df[n,]
testing <- df[-n,]
#回归
fit <- lm(high~sex+weight+father+mother,data=training)
summary(fit)
## Call:
## lm(formula = high ~ sex + weight + father + mother, data = training)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.859  -2.712  -0.417   2.701   9.772
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.6620    13.6790    5.75  5.6e-08 ***
```

```
## sex 男          7.3822     1.0578    6.98  1.2e-10 ***
## weight         0.3155     0.0549    5.75  5.7e-08 ***
## father         0.1484     0.0677    2.19  0.03000 *
## mother         0.2677     0.0747    3.58  0.00047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.22 on 136 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.729
## F-statistic:   95 on 4 and 136 DF,  p-value: <2e-16
```

不出所料，线性回归的结果很理想，模型的R^2达到了 72.9%，我们使用的解释变量性别、父亲身高和母亲身高均显著。变量 sex 的回归系数为 7.38，表明在其他变量相同的情况下，男生的身高可能会比女生高 7.38cm 左右。母亲身高对子女身高的影响要比父亲身高的影响更大。

但是，我们对模型参数推断的信心依赖于它在多大程度上满足 OLS 模型统计假设，在用于预测之前我们需要进行回归诊断。

```
par(mfrow=c(2,2))
plot(fit)
```



在残差与拟合图（左上）中，几乎没有系统关联，呈现均匀分布趋势，模型满足线性假定；位置尺度图中的点随机分布在水平线周围，满足正态性假定；QQ 图

（右上）除了点 50 外，其余基本分布在 45 度线上，并且点 50 的残差较大，可以将此点作为离群点删去。因此我们将 50 号样本出去后重新进行回归，并用拟合模型对预测集进行预测，对测试集预测的均方误差为 17.5。

```
fit.lm <- lm(high~sex+weight+father+mother,data=df[-50,])
summary(fit.lm)
```

```
##
## Call:
## lm(formula = high ~ sex + weight + father + mother, data = df[-50,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.690  -2.769  -0.377   2.422  13.386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.3326    12.9261    5.60  8.6e-08 ***
## sex 男        8.4331     0.9010    9.36  < 2e-16 ***
## weight        0.2650     0.0442    6.00  1.2e-08 ***
## father        0.1379     0.0617    2.24    0.027 *
## mother        0.3325     0.0692    4.81  3.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.25 on 171 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.741
## F-statistic:  126 on 4 and 171 DF,  p-value: <2e-16
```

```
lm.pred <- predict(fit.lm,testing)
lm.MSE <- mean((testing$high-lm.pred)^2)
lm.MSE
```

```
## [1] 17.5
```

3.2 支持向量机

支持向量机（SVM）是一类可用于分类和回归的有监督机器学习模型，虽然不能像线性回归一样解释和表述变量间的关系及影响大小，但如果建立了一个成功的模型，在新样本预测时的准确度很高。我们运用这种方法来预测子女身高，并和回归的预测效果进行对比。

```
library(e1071)
set.seed(1234)


tuned <tune.svm(high~sex+weight+father+mother,data=training,gamma=10^(-
6:1),cost = 10^(-5:5))
tuned
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  gamma  cost
##  1e-04 10000
##
## - best performance: 18.4
```

```
fit.svm <- svm(high~sex+weight+father+mother,data=training,gamma=1e-
4,cost=10000)
fit.svm
```

```
##
## Call:
## svm(formula = high ~ sex + weight + father + mother, data =
training,
##     gamma = 1e-04, cost = 10000)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  10000
##       gamma:  1e-04
##     epsilon:  0.1
##
##
## Number of Support Vectors:  125
```

```
svm.pred <- predict(fit.svm,testing)
svm.MSE <- mean((testing$high-svm.pred)^2)
svm.MSE
```

```
## [1] 19.3
```

我们尝试了不同的 gamma 和成本参数 cost，训练集中 10 折交叉验证误差最小的模型所对应的参数为 gamma=0.0001，cost=10000。基于这一参数结果，我们训练了 SVM 模型，并对测试集进行预测，测试集的预测均方误差为 19.3，明显大于回归模型的均方误差 17.5，因此针对这一调查数据的预测中，线性回归模型的预测效果较好。

4 结论

本文运用主要运用描述统计、线性回归以及支持向量机的方法对身高调查数据进行了研究，主要结论有以下几三点：

随着人们生活水平的提高和生活方式的改变，子女身高的平均水平显著高于父母，这一趋势应该会持续，但增速可能会逐渐放缓。

在大学生人群中，65%左右的同学 BMI 值处于正常范围内，这一比例并不高，女生中体重过轻问题严重，大学生应当注重更加健康的生活方式，避免过瘦或肥胖带来的健康问题。

子女身高受到父母的身高的显著影响，研究结果表明母亲的影响更大一些，另外在预测身高时不应忽略性别的影响，平均来说，即使父母身高和自身体重都相同，男生的身高会比女生高 7cm 左右。