南開大學

第五讲 回归分析

提　纲

1. 回归分析

【问题】

- 如何拟合和解释回归模型、鉴别模型潜在问题的方法、并解决这些问题。
- 如何选择变量，变量不是越多越好，有一些变量并不起作用，有些两者间存在共线性。
- 模型在现实世界中的表现如何。
- 预测(自)变量的重要程度。

1. 回归分析

【主要解决的问题】

包括：

- 一元线性回归
- 多元线性回归
- 回归诊断
- 异常值观测

1. 回归分析

- 线性回归

【问题提出】

1、数据集women提供了15个年龄在30~39岁间女性的身高和体重信息，我们希望通过身高来预测体重，获得一个等式可以帮助我们分辨出那些过重或过瘦的个体。

2、针对state.x77数据集，我们想探究一个州的犯罪率和其他因素的关系，包括人口、文盲率、平均收入和结霜天数（温度在冰点以下的平均天数）。

1. 回归分析

【基本方法】OLS回归

OLS（ordinary least squares）回归，全称为普通最小二乘回归，通过一系列的预测变量（自）来预测响应变量（因）。OLS回归拟合模型的形式为：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}, \quad i = 1, \ldots, n$$

其中，

n	观测的数目
k	预测变量的数目
$\hat{Y}_i$	第i次观测对应的因变量的预测值（具体来讲，它是在已知预测变量值的条件下，对Y估计的均值）
X_{ji}	第i次观测对应的第j个预测变量值
$\hat{\beta}_0$	截距项（当所有的预测变量都为0时，Y的预测值）
$\hat{\beta}_j$	预测变量j的回归系数（斜率表示X_j改变一个单位所引起的Y的改变量）

1. 回归分析

通过减少响应变量的真实值与预测值的差值来获得模型参数（截距项和斜率），即使得残差平方和最小:

$$\sum_{1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{1}^{n}(Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki})^2 = \sum_{1}^{n}\varepsilon^2$$

1. 回归分析

为了能够恰当地解释OLS模型的系数，数据必须满足以下统计假设：

- 正态性　对于固定的自变量值，因变量值成正态分布。
- 独立性　Y_i值之间相互独立。
- 线　　性　因变量与自变量之间为线性相关。
- 同方差性　因变量的方差不随自变量的不同而变化。也可称作不变方差。

如果违背了以上假设，统计显著性检验结果和所得的置信区间很可能就不精确。注意，OLS回归还假定自变量是固定的且测量无误差，但在实践中通常都放松了这个假设。

1. 回归分析

- 一元线性回归

一元线性回归的研究对象包含两个变量，是只有一个自变量的回归分析。假设变量y和变量x满足：

$$y = a + bx$$

那么我们称$y = a + bx$为因变量y和自变量x的回归函数，**a**表示截距，**b**称为回归系数。

1. 回归分析

- 多元线性回归

在实际问题中，影响因变量y的自变量往往不止一个，如果p个自变量$x_1, x_2, \ldots, x_p$与因变量y之间存在着相关关系，通常就意味当$x_1, x_2, \ldots, x_p$变量取定值后，y 便有相应的值与之对应。当自变量不止一个时，简单线性回归就变成了多元线性回归。线性回归模型表示为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, i = 1, \ldots, n$$

1. 回归分析

- 回归诊断

回归诊断主要用于检验关于回归假设是否成立，以及检验模型形式是否错误，否则我们通过最小二乘法求得的回归方程就缺乏理论依据。这些检验主要探究的问题为：

1）残差是否为随机性、是否为正态性、是否不为异方差；

2）高度相关的自变量是否引起了共线性；

3）模型的函数形式是否错误或在模型中是否缺少重要的自变量；

4）样本数据中是否存在异常值。

1. 回归分析

- 异常值观测

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。

◆ 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。落在置信区间带外的点即可被认为是离群点。另外一个粗糙的判断准则是标准化残差值大于2或者小于-2的点可能是离群点，需要特别关注。

1. 回归分析

◆ 高杠杆值点

高杠杆值观测点是与其他自变量有关的离群点。换句话说，它们是由许多异常的自变量值组合起来的，与因变量值没有关系。高杠杆值的观测点可通过帽子统计量（hat statistic）判断。

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}.$$

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，即可以认定为高杠杆值点。高杠杆值点可能会是强影响点，也可能不是，这要看它们是否是离群点。

1. 回归分析

◆ 强影响点

强影响点是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，那么你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是自变量数目。

1. 回归分析

【使用R的具体回归过程及结果解读】

在R中，拟合线性模型基本的函数就是lm()，格式为：

myfit <- lm(formula, data)

其中，formula指要拟合的模型形式:

Y ~ X1+X2+…+Xk

~左边为因变量，右边为各个自变量，自变量之间用+符号分隔。表8-2中的符号可以不同方式修改这一表达式。

data是一个数据框，包含了用于拟合模型的数据。

结果对象（本例中是myfit）存储在一个列表中，包含了所拟合模型的大量信息。

1. 回归分析

表8-2　R表达式中常用的符号

符　号	用　途
~	分隔符号，左边为响应变量，右边为解释变量。例如，要通过x、z和w预测y，代码为y ~ x + z + w
+	分隔预测变量
:	表示预测变量的交互项。例如，要通过x、z及x与z的交互项预测y，代码为y ~ x + z + x:z
*	表示所有可能交互项的简洁方式。代码y~ x * z * w可展开为y ~ x + z + w + x:z + x:w + z:w + x:z:w
^	表示交互项达到某个次数。代码y ~ (x + z + w)^2可展开为y ~ x + z + w + x:z + x:w + z:w
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量x、y、z和w，代码y~.可展开为y ~ x + z + w
-	减号，表示从等式中移除某个变量。例如，y ~ (x + z + w)^2 - x:w可展开为y ~ x + z + w + x:z + z:w
-1	删除截距项。例如，表达式y ~ x - 1拟合y在x上的回归，并强制直线通过原点
I()	从算术的角度来解释括号中的元素。例如，y~x+(z+w)^2将展开为y~x+z+w+z:w。相反,代码y ~ x + I((z + w)^2)将展开为y ~ x + h，h是一个由z和w的平方和创建的新变量
function	可以在表达式中用的数学函数。例如，log(y) ~ x + z + w表示通过x、z和w来预测log(y)

1. 回归分析

【提示】表8-3 对拟合线性模型非常有用的其他函数

表8-3　对拟合线性模型非常有用的其他函数

函　数	用　途
summary()	展示拟合模型的详细结果
coefficients()	列出拟合模型的模型参数（截距项和斜率）
confint()	提供模型参数的置信区间（默认95%）
fitted()	列出拟合模型的预测值
residuals()	列出拟合模型的残差值
anova()	生成一个拟合模型的方差分析表，或者比较两个或更多拟合模型的方差分析表
vcov()	列出模型参数的协方差矩阵
AIC()	输出赤池信息统计量
plot()	生成评价拟合模型的诊断图
predict()	用拟合模型对新的数据集预测响应变量值

1. 回归分析

- 一元线性回归

```
# save current graphical parameters
opar <- par(no.readonly = TRUE)
women
fit <- lm(weight ~ height, data = women) #拟合模型
summary(fit)  #模型参数
women$weight   #真实值
fitted(fit)   #预测值
residuals(fit)  # 残差值
```

1. 回归分析

```
Estimate Std. Error t value Pr(>|t|)
 (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
  height       3.45000    0.09114   37.85 1.09e-14   ***
Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

通过输出结果，可以得到预测等式：

$$\widehat{Weight} = -87.52 + 3.45 \times Height$$

因为身高不可能为**0**，没必要给截距项一个物理解释，它仅仅是一个常量调整项。在$\Pr(>|t|)$栏，可以看到回归系数（**3.45**）显著不为**0**（**p<0.001**），表明身高每增高**1**英寸，体重将预期增加**3.45**磅。残差标准误差（**1.53 lbs**）则可认为是模型用身高预测体重的平均误差。**R**平方项（**0.991**）表明模型可以解释体重**99.1%**的方差，它也是实际和预测值之间的相关系数（$R^2 = r^2_{\hat{Y}Y}$）。**F**统计量检验所有的自变量预测因变量是否都在某个几率水平之上。由于简单回归只有一个自变量，此处**F**检验等同于身高回归系数的**t**检验。

1. 回归分析

- T检验：检验解释变量的显著性；
- R-squared：查看方程拟合程度；
- F检验：是检验方程整体显著性；

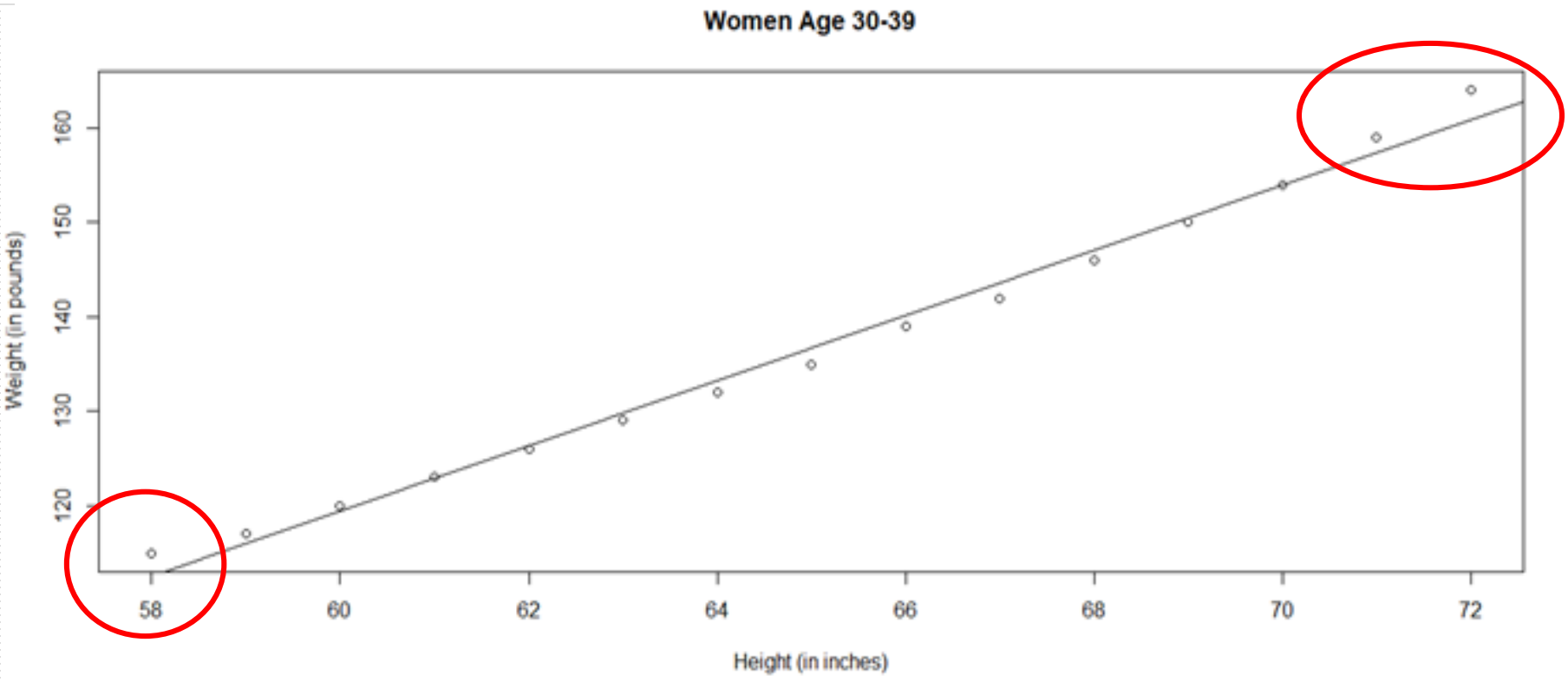
对拟合模型的解读，参考：

https://blog.csdn.net/qq_27586341/article/details/92636671

1. 回归分析

```
# scatter plot of height by weight
plot(women$height, women$weight, main =
"Women Age 30-39",
    xlab = "Height (in inches)", ylab = "Weight (in
pounds)")
# add the line of best fit
abline(fit)
par(opar)
```

1. 回归分析



从图中可以看出，最大的残差值在身高矮和身高高的地方出现，同时可以用含一个弯曲的曲线来提高预测的精度。比如，模型$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$就能更好地拟合数据，即多项式回归。

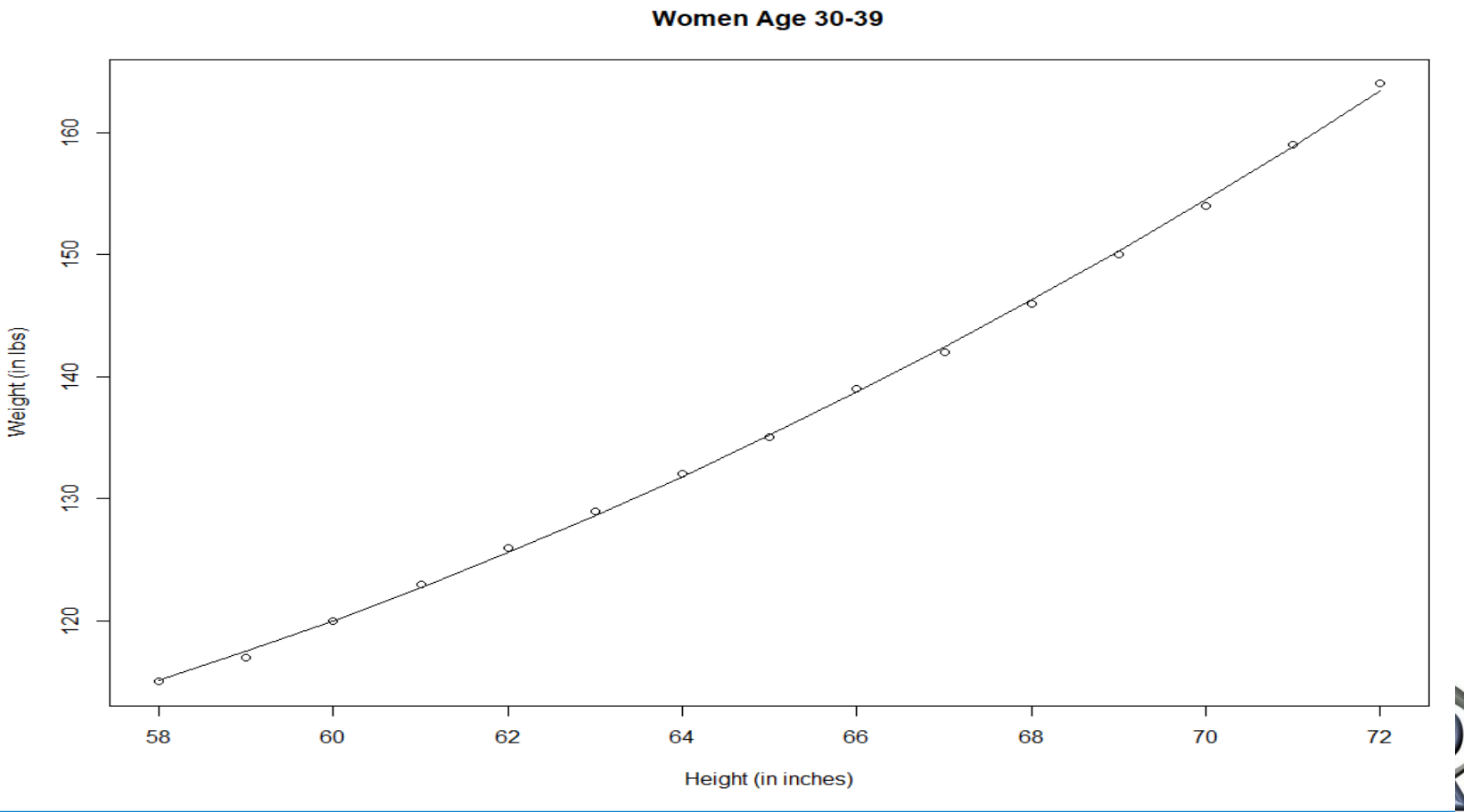
1. 回归分析

- 多项式回归

```
fit2 <- lm(weight ~ height + I(height^2), data = women)
summary(fit2)
plot(women$height, women$weight, main = "Women Age
30-39",  xlab = "Height (in inches)", ylab = "Weight (in lbs)")
lines(women$height, fitted(fit2))
```

1. 回归分析

在p<0.001水平下，回归系数都非常显著。模型的方差解释率已经增加到了99.9%。二次项的显著 性（t = 13.89，p<0.001）表明包含二次项提高了模型的拟合度。



1. 回归分析

#用scatterplot很容易、方便地绘制 二元关系图

```
install.packages("car")
library(car)
scatterplot(weight ~ height, data = women, spread =
FALSE,  lty.smooth = 2, pch = 19, main = "Women Age 30-
39", xlab = "Height (inches)",
    ylab = "Weight (lbs.)")
```

1. 回归分析

【提示】

线性模型与非线性模型

多项式等式仍可认为是线性回归模型，因为等式仍是预测变量的加权和形式（本例中是身高和身高的平方）。即使这样的模型：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \log X_1 + \hat{\beta}_2 \times \sin X_2$$

仍可认为是线性模型（参数项是线性的），能用这样的表达式进行拟合：

$$Y \sim \log(X_1) + \sin(X_2)$$

相反，下面的例子才能算是真正的非线性模型：

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 e^{\frac{x}{\beta_2}}$$

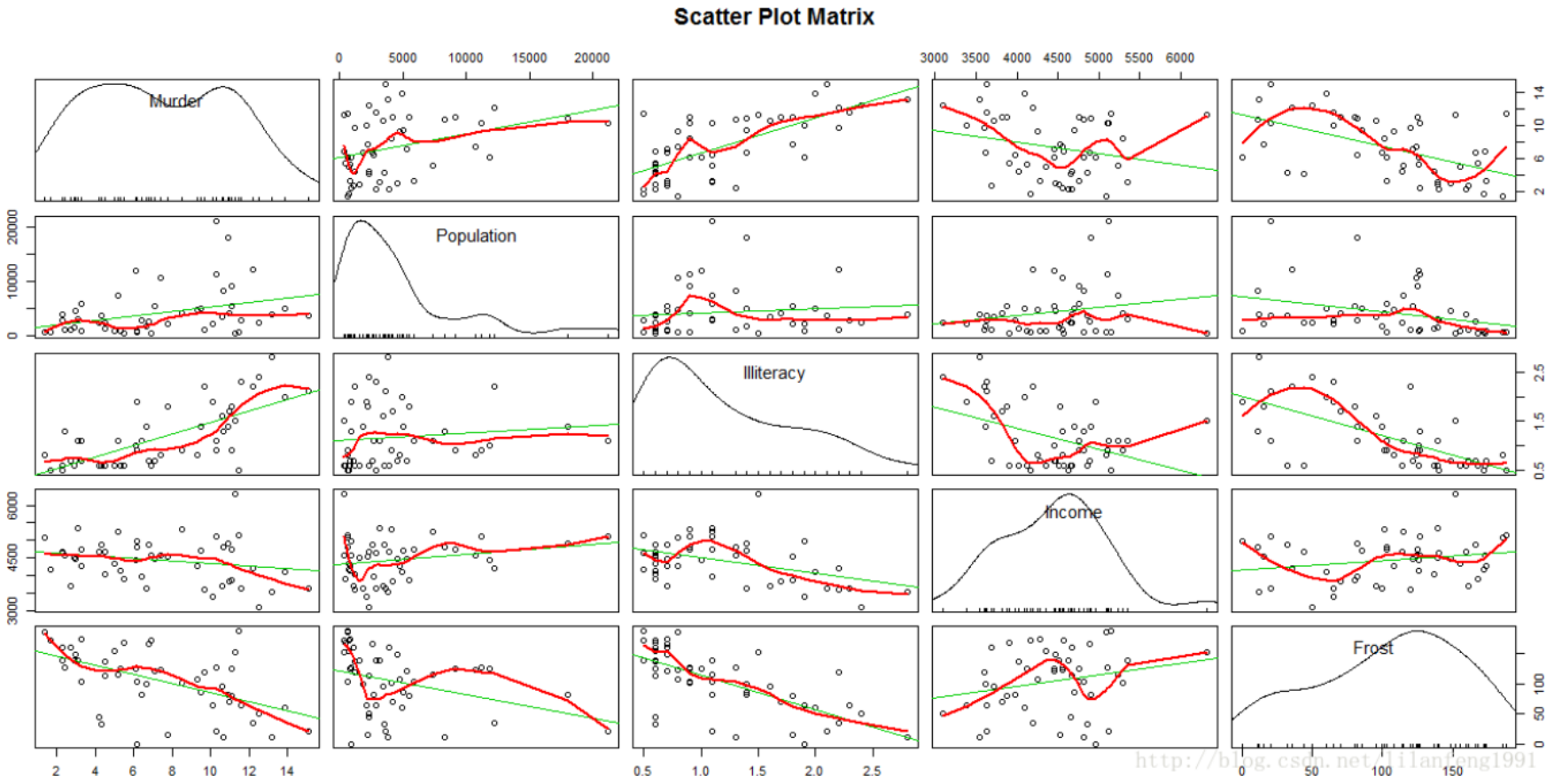
这种非线性模型可用nls()函数进行拟合。

1. 回归分析

- 多元线性回归

```
# save current graphical parameters
opar <- par(no.readonly = TRUE)
states <- as.data.frame(state.x77[,c("Murder", "Population",  "Illiteracy", "Income", "Frost")])
states
cor(states)  # 计算二变量之间的相关系数
library(car)
scatterplotMatrix(states,smooth=list(spread=FALSE,lty=2), main="Scatter Plot Matrix")
```

1. 回归分析



scatterplotMatrix（）函数默认在非对角线区域绘制变量间的散点图，并添加平滑（loess）和线性拟合曲线

1. 回归分析

```
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states) #多元线性回归模型拟合
summary(fit) #模型参数
par(opar)
```

1. 回归分析

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.235e+00	3.866e+00	0.319	0.7510	
Population	2.237e-04	9.052e-05	2.471	0.0173	*
Illiteracy	4.143e+00	8.744e-01	4.738	2.19e-05	***
***Income	6.442e-05	6.837e-04	0.094	0.9253	
Frost	5.813e-04	1.005e-02	0.058	0.9541	

Multiple R-squared: 0.567, Adjusted R-squared: 0.5285

当自变量不止一个时，回归系数的含义为，一个自变量增加一个单位，其他自变量保持不变时，因变量将要增加的数量。例如，文盲率的回归系数为4.14，表示控制人口、收入和温度不变时，文盲率上升1%，谋杀率将会上升4.14%，它的系数在p<0.001的水平下显著不为0。相反，Frost的系数没有显著不为0（p=0.954），表明当控制其他变量不变时，Frost与Murder不呈线性相关。总体来看，所有的自变量解释了各州谋杀率57%的方差。

1. 回归分析

```
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data=states)
confint(fit) #获得一个模型的参数的置信区间
                    2.5 %        97.5 %
(Intercept) -6.552191e+00  9.0213182149
Population   4.136397e-05  0.0004059867
Illiteracy   2.381799e+00  5.9038743192
Income      -1.312611e-03  0.0014414600
Frost       -1.966781e-02  0.0208304170
```

文盲率改变1%，谋杀率就在95%的置信区间[2.38,5.90]中变化。另外，因为Frost的置信区间包含0，可以得出结论说，当其他变量不变时，温度的改变与谋杀率无关。不过，对这些结果的认同，都只建立在数据满足统计假设的前提之上。

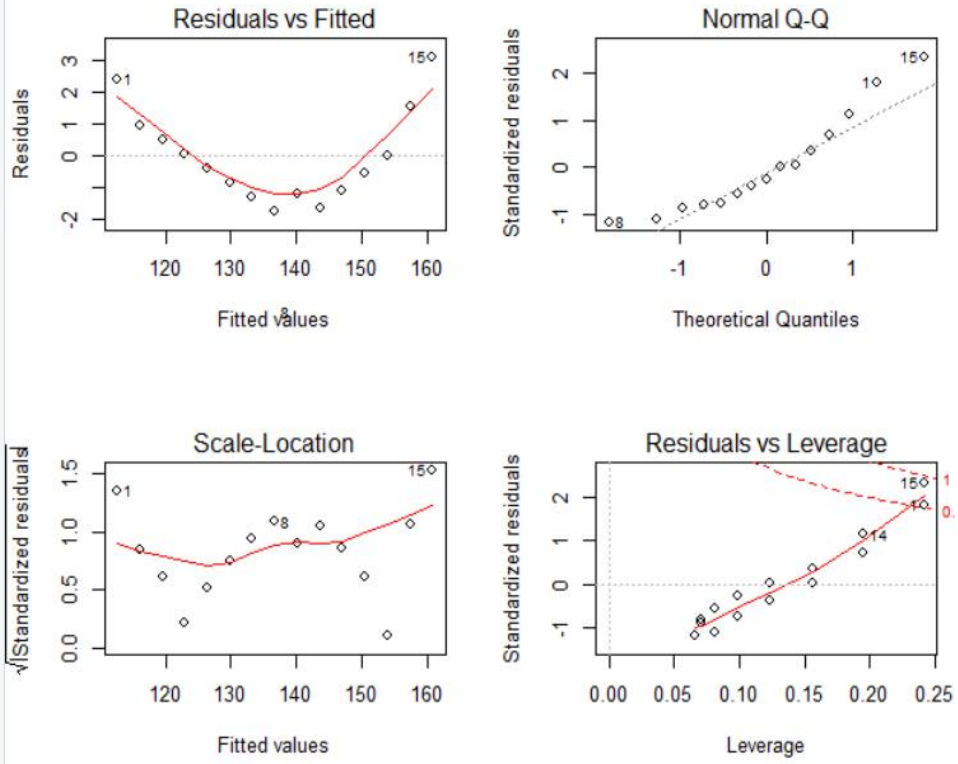
1. 回归分析

- 回归诊断的标准方法

R基础安装中提供了大量检验回归分析中统计假设的方法。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。

```
fit <- lm(weight ~ height, data=women)
par(mfrow=c(2,2))
plot(fit)
```

1. 回归分析



线性 在“残差图与拟合图”（左上）中可以清楚的看到一个曲线关系，这暗示着你可能需要对回归模型加上一个二次项。

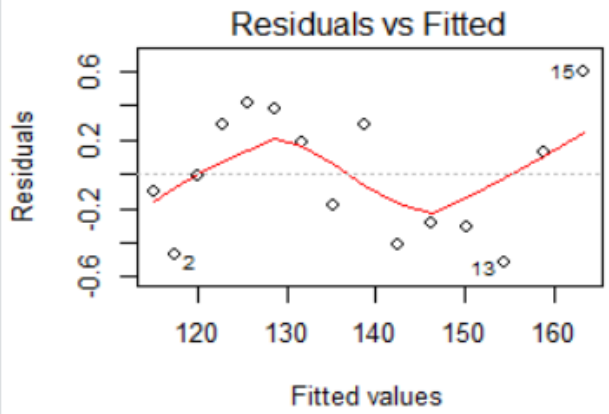
正态性 正态**Q-Q**图（右上）是在正态分布对应的值下，标准化残差的概率图。若满足正态假设，那么图上的点应该落在呈**45**度角的直线上；否则就 违反了正态性的假设。

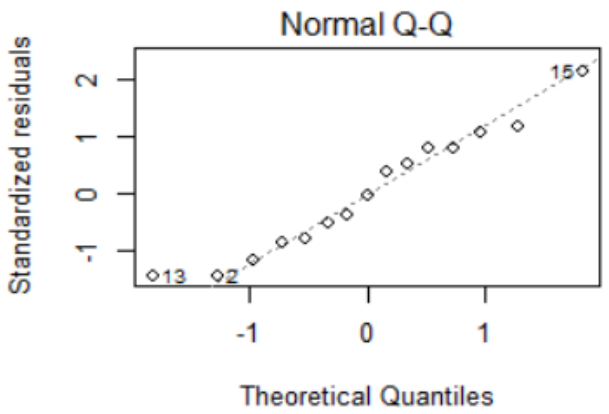
同方差性 若满足不变方差假设，那么在位置尺度图（左下）中， 水平线周围的点应该随机分布。该图似乎满足此假设。

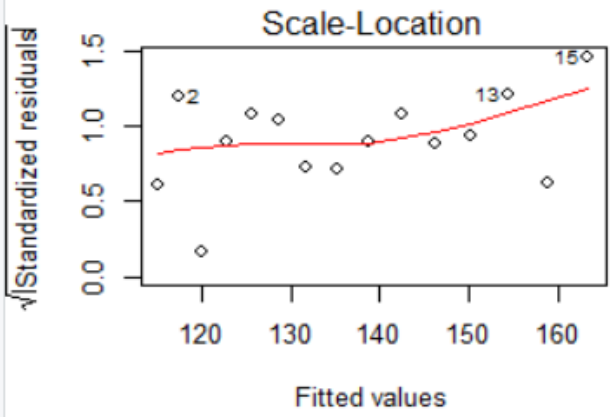
“残差与杠杆图”（**Residuals vs Leverage**，右下）提供了可能关注的单个观测点的信息。从图形可以鉴别出离群点、高杠杆值点和强影响点。

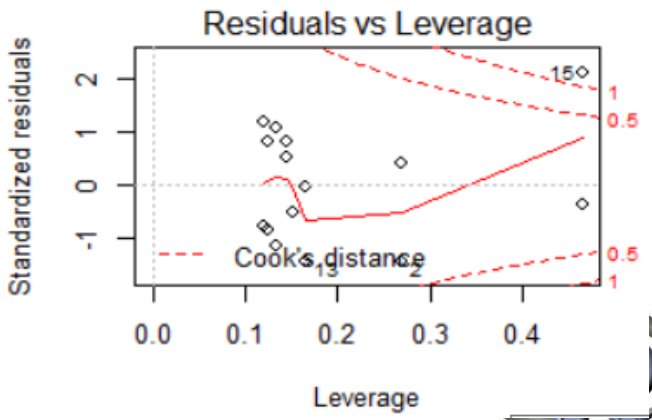
1. 回归分析

```
fit2<-lm(weight~height+I(height^2),data=women)
par(mfrow=c(2,2))
plot(fit2)
```

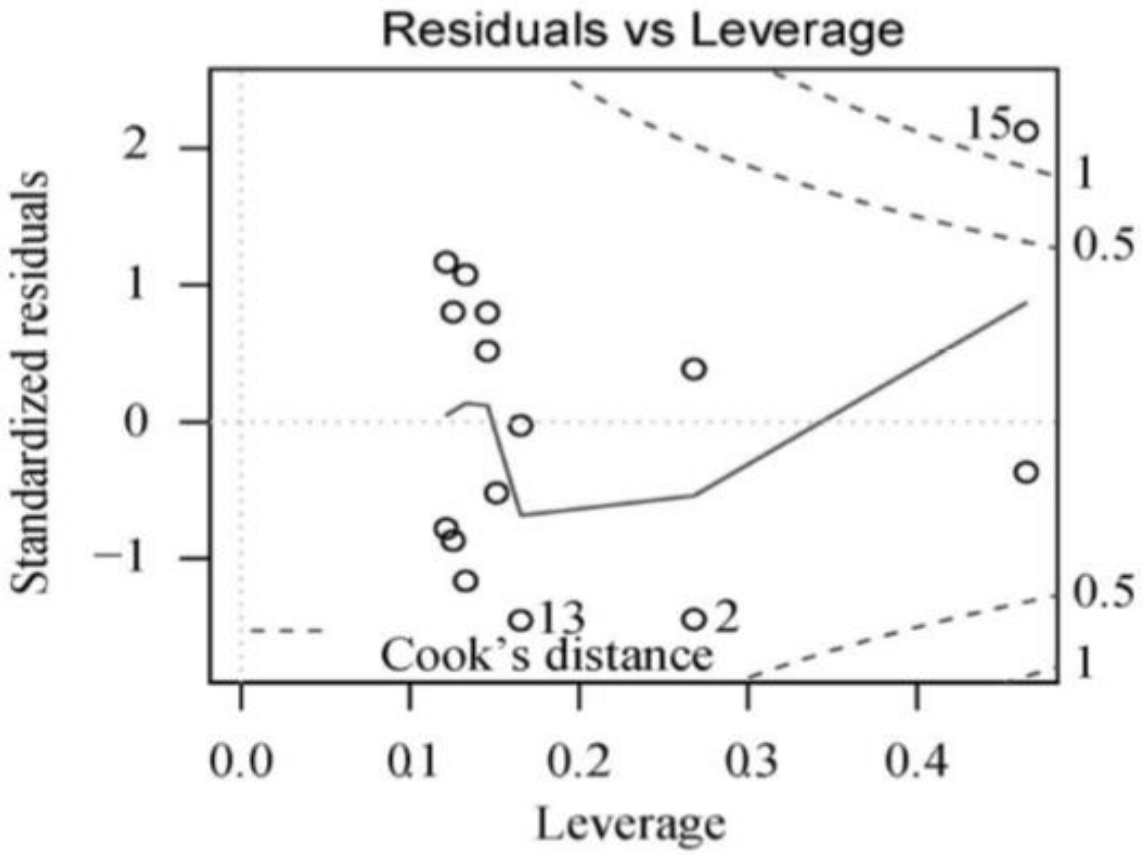








1. 回归分析

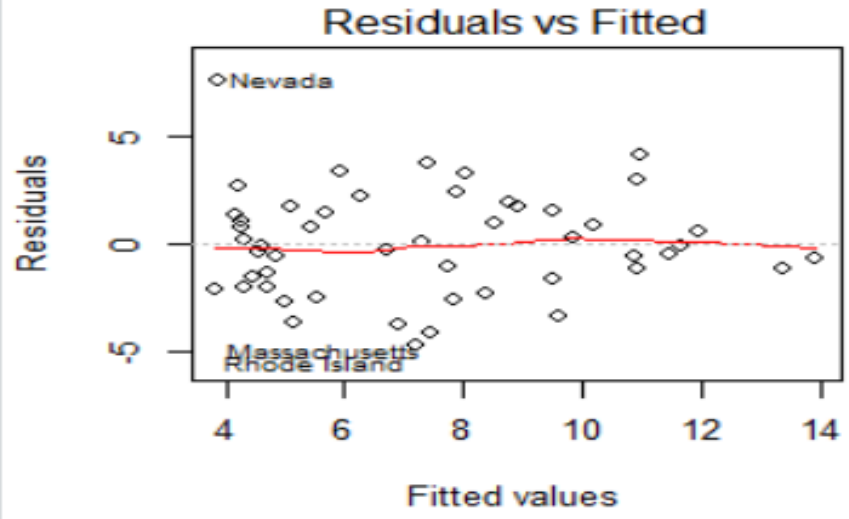


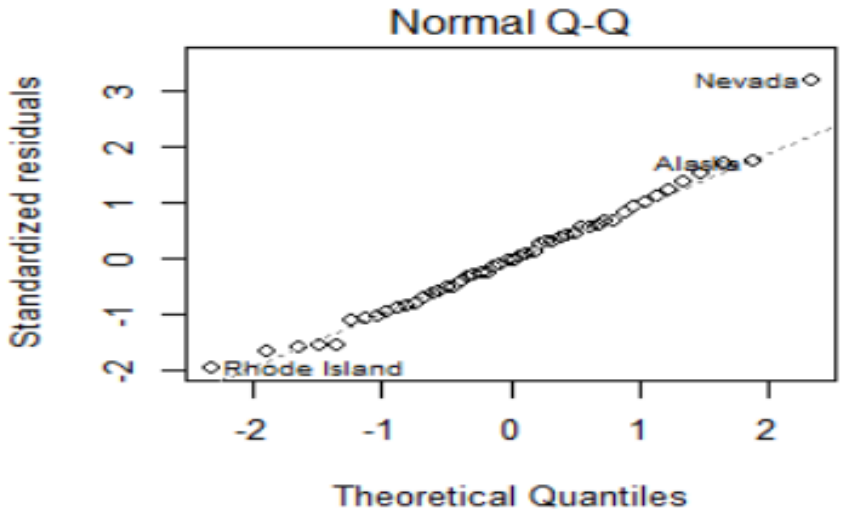
上图是二次拟合的“残差与杠杆图”，观测点15看起来像是强影响点（根据是它有较大的 Cook距离值）。

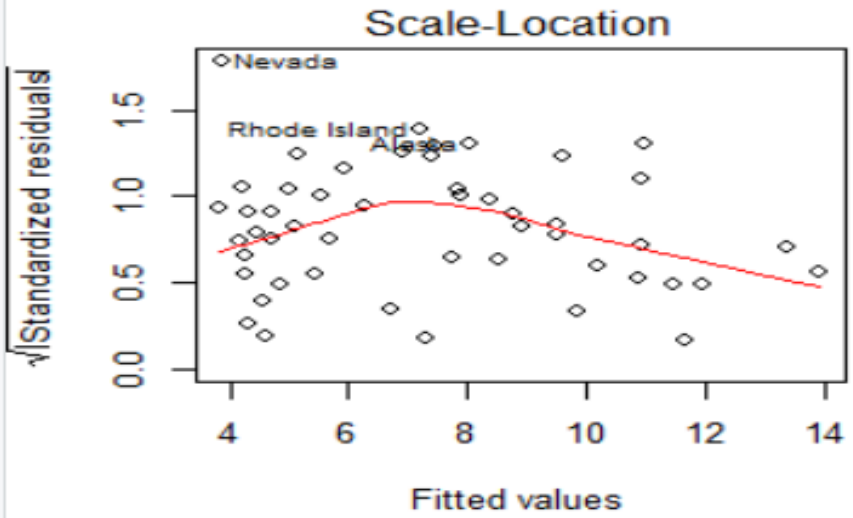
1. 回归分析

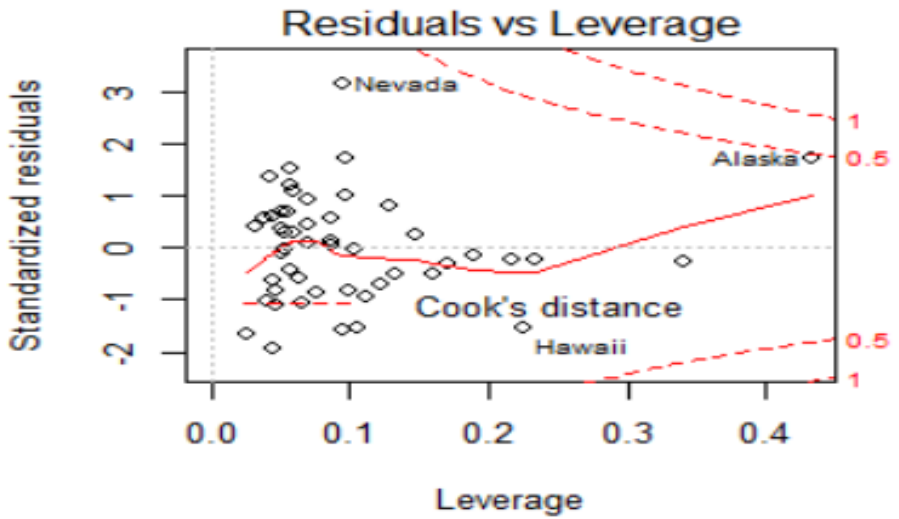
```
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data = states)# states的多元回归问题
par(mfrow=c(2,2))
plot(fit)#
par(mfrow=c(1,1))
```

1. 回归分析


Residuals vs Fitted
Residuals
Fitted values
Nevada
Massachusetts
Rhode Island


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Nevada
Alaska
Rhode Island


Scale-Location
√|Standardized residuals|
Fitted values
Nevada
Rhode Island
Alaska


Residuals vs Leverage
Standardized residuals
Leverage
Nevada
Alaska
Hawaii
Cook's distance

1. 回归分析

- 异常值观测——离群点

```
library(car)
outlierTest(fit)
```

```
rstudent unadjusted p-value Bonferonni
pNevada 3.542929       0.00095088    0.047544
```

可以看到Nevada被判定为离群点（p=0.048）。注意，该函数只是根据单个最大（或正或负）残差值的显著性来判断是否有离群点。若不显著，则说明数据集中没有离群点；若显著，则你必须删除该离群点，然后再检验是否还有其他离群点存在。

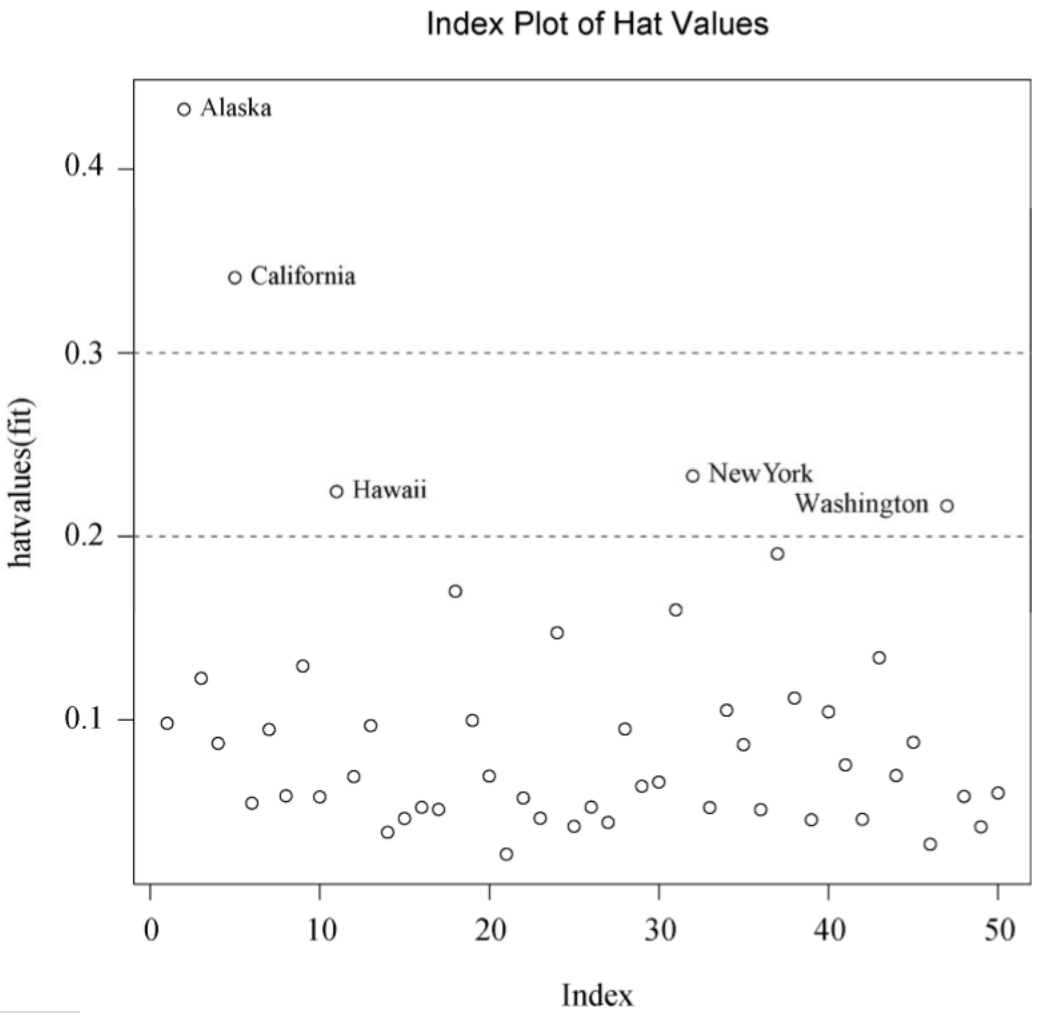
【提示】car包提供了大量函数，大大增强了拟合和评价回归模型的能力（参见表8-4）。

1. 回归分析

- 异常值观测——高杠杆值点

```
# Index plot of hat values
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(fit) # use the mouse to identify points interactively
```

1. 回归分析



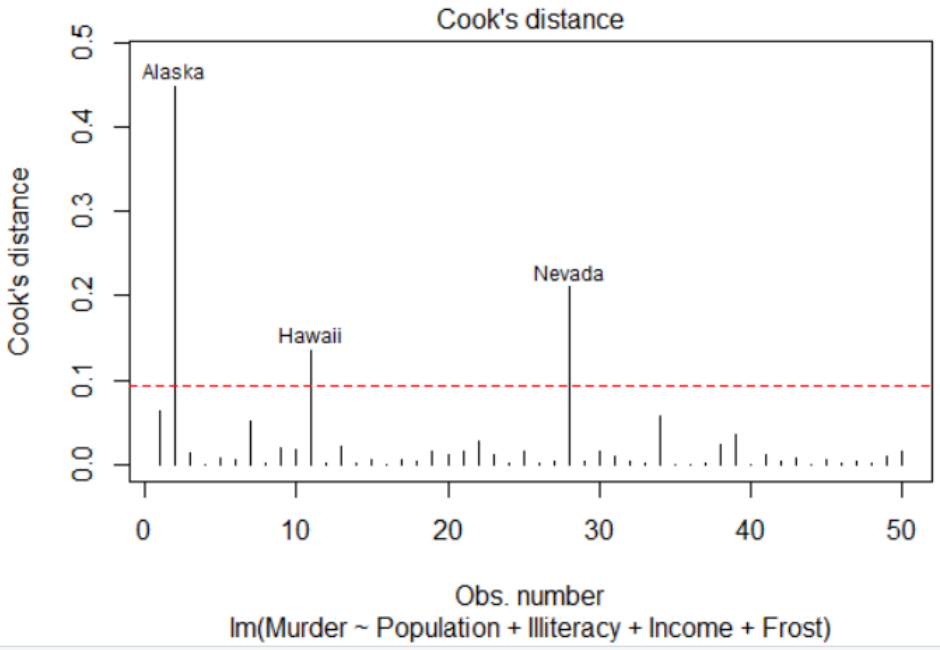
水平线标注的即帽子均值2倍和3倍的位置。定位函数（locator function）能以交互模式绘图： 单击感兴趣的点进行标注，停止交互时，用户可按Esc键退出。

此图中，可以看到Alaska和California非常异常，查看它们的预测变量值， 与其他48个州进行比较发现：Alaska收入比其他州高得多，而人口和温度却很低；California人口比其他州府多得多，但收入和温度也很高。

1. 回归分析

- 异常值观测——强影响点

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



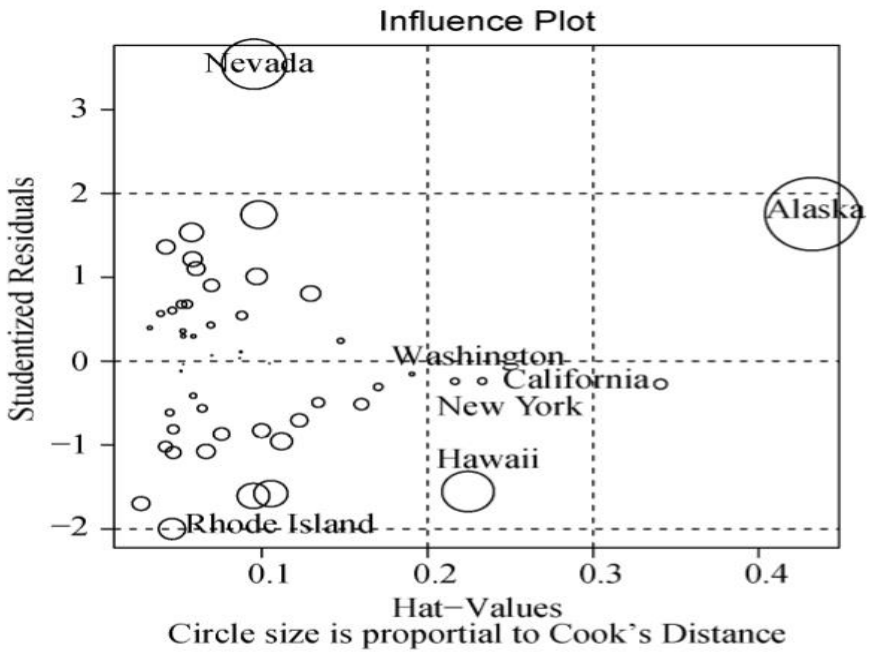
通过图形可以判断Alaska、Hawaii和Nevada是强影响点。若删除这些点，将会导致回归模型截距项和斜率发生显著变化。注意，虽然该图对搜寻强影响点很有用，但我逐渐发现以1为分割点比4/(n-k-1)更具一般性。若设定D=1为判别标准，则数据集中没有点看起来像是强影响点。

1. 回归分析

- 异常值观测

利用car包中的influencePlot()函数，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(fit, id =list(method="identify"), main = "Influence Plot", sub = "Circle size is proportial to Cook's Distance")



纵坐标超过+2或小于-2的州可被认为是离群点，Nevada和Rhode Island是离群点。水平轴超过0.2或0.3的州有高杠杆值（通常为预测值的组合），New York、California、Hawaii和Washington有高杠杆值。圆圈大小与影响成比例，圆圈很大的点可能是对模型参数的估计造成的不成比例影响的强影响点，Nevada、Alaska和Hawaii为强影响点。

1. 回归分析

- 多重共线性分析

回归系数测量的是当其他自变量不变时，某个自变量对因变量的影响。如果有假定出生日期不变，然后测量握力与年龄的关系， 这种问题就称作多重共线性（**multicollinearity**）。它会导致模型参数的置信区间过大，使单系数解释起来很困难。

多重共线性可用统计量**VIF**（**Variance Inflation Factor**，方差膨胀因子）进行检测。**VIF**的平方根表示变量回归参数的置信区间能膨胀为与模型无关的自变量的程度（因此而得名）。

car 包中的**vif()**函数提供**VIF**值。一般原则下，

sqrt(vif) >2就表明存在多重共线性问题。

1. 回归分析

```
library(car)
vif(fit)
Population   Illiteracy    Income     Frost
1.245282      2.165848   1.345822    2.082547
sqrt(vif(fit)) > 2
Population Illiteracy Income  Frost
FALSE       FALSE    FALSE  FALSE
```

结果表明自变量不存在多重共线性问题。

1. 回归分析

◆ 改进的措施

（1）删除观测点

删除观测点可提高数据集对于正态假设的拟合度。强影响点会干扰结果，通常也会被删除。删除最大的离群点或强影响点，模型需要重新拟合，若离群点或强影响点仍然存在，重复以上过程直到获得比较满意的拟合。

对删除观测点应持谨慎态度。

1. 回归分析

（2）变量变换

当模型不符合正态性、线性或同方差性假设时，一个或多个变量的变换通常可以改善或调整模型效果。当模型违反了正态假设时，通常可以对因变量尝试某种变换。

1. 回归分析

（3）增删变量

改变模型的变量会影响模型的拟合度，增加或删除变量解决多重共线问题：岭回归

1. 回归分析

参考：

R语言回归篇：

https://blog.csdn.net/lilanfeng1991/article/details/29627405

R语言与回归分析学习笔记（应用回归小结）

https://blog.csdn.net/yujunbeta/article/details/9252883

2. 回归分析练习

R基础安装中的state.x77数据集，它提供了美国50个州在1977年的人口、收入、文盲率、预期寿命、谋杀率和高中毕业率数据。

基于这些数据，对你感兴趣的话题进行回归分析：

- 基于相关分析结果进行回归分析；
- 针对拟合的回归模型进行回归诊断；
- 分析数据是否存在异常点，提出合理的解决办法并进行讨论（例如8.5 改进措施 ）。
- 使用部分数据建模，部分数据验证你的模型（8.7.1 交叉验证）