# 《线性回归》

杨　　瑛

清华大学　数学科学系
Email: yyang@math.tsinghua.edu.cn
Tel: 62796887

2019.04.22

# Outline

# Outline

## 主要研究的问题：

0. 本节内容主要来自于Oliver, pp. Chapter 3 和Draper and Smith, Chapter 16. 以及Seber and Lee (2003), Section 9.7

1. 什么是多重共线性(multicollinearity)？

2. 多重共线性对LSE的影响是什么？

3. 如何判断数据中存在多重共线性？

## 多重共线性的定义

- 在实际问题的解决中，通常会遇到多个变量的回归问题，但是估计结果不理想，表现为：
- 某些回归系数的估计的绝对值异常的大
- 系数的LSE与问题的实际背景相违背
- 问题的原因在于： 回归自变量之间存在着近似线性关系，称之为多重共线性(multicolinearity)

- 多重共线性：简单情形

  设$x_1$和$x_2$为两个自变量，其$n$ 次观测数据分别为：

  $\mathbf{x}_1 = (x_{11}, \cdots, x_{n1})$,

  $\mathbf{x}_2 = (x_{12}, \cdots, x_{n2})$.

  $\mathbf{x}_1$ 和$\mathbf{x}_2$ 的样本相关系数的平方：

$$r^2 = \frac{[\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)]^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2} = \frac{s_{12}^2}{s_{11}s_{22}} \tag{1}$$

  其中，$\bar{x}_1$ 和$\bar{x}_2$ 为$\mathbf{x}_1$ 和$\mathbf{x}_2$ 的样本均值.

将自变量数据$\mathbf{x}_i$ 标准化：

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{(s_{jj})^{1/2}}, j = 1, 2, \tag{2}$$

$\mathbf{x}_j^* = (x_{1j}^*, \cdots, x_{nj}^*)^T$ 称为$\mathbf{x}_j$的标准化，$j = 1, 2.$

这样以来，(1) 中的$r^2$相当于$x_1^*$ 和$x_2^*$ 之间的“样本”相关系数。

- 结论：若$|r| = 1$，则存在$a \neq 0, b \neq 0$ 使

$$a\mathbf{x}_1^* + b\mathbf{x}_2^* = \mathbf{0}$$

因为

$$\begin{aligned} \|a\mathbf{x}_1^* + b\mathbf{x}_2^*\|^2 &= (a\mathbf{x}_1^* + b\mathbf{x}_2^*)^T(a\mathbf{x}_1^* + b\mathbf{x}_2^*) \\ &= a^2 + 2abr + b^2 \end{aligned} \tag{3}$$

令

$$a = \frac{1}{\sqrt{2}}, \qquad b = -\frac{1}{\sqrt{2}}\text{sgn}(r), \tag{4}$$

其中

$$\text{sgn}(r) = \begin{cases} 1, & \text{当} r \geq 0 \\ -1, & \text{当} r < 0 \end{cases}$$

由(3) 得：

$$\|a\mathbf{x}_1^* + b\mathbf{x}_2^*\|^2 = 1 - |r| \tag{5}$$

故$|r| = 1 \Rightarrow \quad a\mathbf{x}_1^* + b\mathbf{x}_2^* = \mathbf{0}$. 这就是所谓的“(严)共线性”。

## 统计中的共线性：

$$\|a\mathbf{x}_1^* + b\mathbf{x}_2^*\| \approx 0. \tag{6}$$

其中$a$和$b$满足$a^2 + b^2 = 1$(为什么？ ).

- 考虑线性模型

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, 1 \leq i \leq n. \tag{7}$$

其中$\mathbf{x}_1 = (x_{11}, \cdots, x_{n1})^T$ 和$\mathbf{x}_2 = (x_{12}, \cdots, x_{n2})^T$ 共线性。

设计矩阵为$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{x2} \end{pmatrix}$,

$$|\mathbf{X}^T\mathbf{X}| = n\left[s_{11}s_{22} - s_{12}^2\right] = ns_{11}s_{22}(1 - r^2),$$

其中

$$\begin{aligned} s_{11} &= \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2, \qquad s_{22} = \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2, \\ s_{12} &= \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2). \end{aligned}$$

## 结论

由此可以看出：

⋆ 如果$|r| = 1$, 则$(\mathbf{X}^T\mathbf{X})$ 的逆矩阵不存在。
⋆ 线性模型(7) 没有形如

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

的LSE。

⋆ 但是正规方程组：

$$(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{Y}$$

仍然可能有(无穷多个)解。

如果$\mathbf{x}_1$和$\mathbf{x}_2$是统计共线性的，则

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \frac{\sigma^2}{s_{11}s_{22}(1-r^2)} \\ &\times \begin{pmatrix} A_{11} & \bar{x}_2 s_{12} - \bar{x}_1 s_{22} & \bar{x}_1 s_{12} - \bar{x}_2 s_{11} \\ \bar{x}_2 s_{12} - \bar{x}_1 s_{22} & s_{22} & -s_{12} \\ \bar{x}_1 s_{12} - \bar{x}_2 s_{11} & -s_{12} & s_{11} \end{pmatrix}\end{aligned}$$

其中$A_{11} = s_{11}s_{22} + s_{11}\bar{x}_2^2 + \bar{x}_1^2 s_{22} - s_{12}^2 - 2s_{12}\bar{x}_1\bar{x}_2$.

- $\hat{\beta}_i$ 的方差将会趋于正无穷，当$|r| \to 1$.

当只有两个自变量时，用样本相关系数$r$的绝对值或者平方的大小可以判断两个变量共线性的强弱程度，但不适合多个自变量的情形。

## 方差膨胀因子(variance inflation factor)

- 可以利用所谓的“方差膨胀因子”来判定变量之间的多重共线性
- 两个自变量情形：

$$\mathrm{VIF}_j = \frac{1}{1 - r^2}, \quad j = 1, 2 \tag{8}$$

## 方差膨胀因子(variance inflation factor)(续)

- 多个自变量情形：

$$\mathrm{VIF}_j = s^{jj*} = \frac{1}{1 - R_j^2}, \quad j = 1, \cdots, k, \tag{9}$$

其中$R_j^2$是把第$j$个自变量看作因变量，用其余$k-1$个变量做线性回归所得到的决定系数。

- $\mathrm{VIF}_j$ 可以度量每个变量所受到的多共线性的影响的大小
- 相关系数度量共线性简单、直观，但只适用于有两个自变量的情形；
- 方差膨胀因子适用于多个自变量的情形。

## 共线性的危害

$|r| \to 1$将会导致：

- LSE 的方差增大；
- 估计的性质不稳定；
- 置信区间的长度增加；
- 在假设检验中，将会导致对因变量有显著影响的自变量判定为无显著影响【为什么？】

评价估计量的优良性的标准：

## 均方误差(mean squared errors, MSE)

设$\beta \in R^p$是未知的参数向量，$\widehat{\beta}$ 是$\beta$ 的一个估计，则$\widehat{\beta}$ 的均方误差定义为

$$\text{MSE}(\widehat{\beta}) = E\|\widehat{\beta} - \beta\|^2 = E(\widehat{\beta} - \beta)^T(\widehat{\beta} - \beta).$$

- MSE 的意义：度量了估计量$\widehat{\beta}$ 与未知参数$\beta$ 的平均偏离的大小。
- 好的估计量应该有比较小的均方误差。

**Theorem**

在一定的正则性条件下，

$$MSE(\widehat{\beta}) = trace(Cov(\widehat{\beta})) + \|E\widehat{\beta} - \beta\|^2. \tag{10}$$

- $\widehat{\beta}$ 的均方误差可分解为：
  $\widehat{\beta}$ 的各分量的方差之和；
  $\widehat{\beta}$ 的各分量的偏差之和。
- 估计的MSE由其各分量的方差和偏差所确定。
- 好的估计量应该有较小的方差和较小的偏差。

**Proof.**

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\beta}) &= E(\widehat{\beta}-\beta)^T(\widehat{\beta}-\beta) \\
&= E([\widehat{\beta}-E\widehat{\beta}]+[E\widehat{\beta}-\beta])^T([\widehat{\beta}-E\widehat{\beta}]+[E\widehat{\beta}-\beta]^T) \\
&= E(\widehat{\beta}-E\widehat{\beta})^T(\widehat{\beta}-E\widehat{\beta})+(E\widehat{\beta}-\beta)^T(E\widehat{\beta}-\beta) \\
&= E\mathrm{trace}([\widehat{\beta}-E\widehat{\beta}]^T[\widehat{\beta}-E\widehat{\beta}])+\|E\widehat{\beta}-\beta\|^2 \\
&= E\mathrm{trace}([\widehat{\beta}-E\widehat{\beta}][\widehat{\beta}-E\widehat{\beta}]^T)+\|E\widehat{\beta}-\beta\|^2 \\
&= \mathrm{trace}E([\widehat{\beta}-E\widehat{\beta}][\widehat{\beta}-E\widehat{\beta}]^T)+\|E\widehat{\beta}-\beta\|^2 \\
&= \mathrm{trace}(\mathrm{Cov}(\widehat{\beta}))+\|E\widehat{\beta}-\beta\|^2.
\end{aligned}
$$

□

- 由于$\beta$ 的LSE 为
$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
是$\beta$的无偏估计，所以上述定理(10) 式中的最后一项等于零。
- 于是：
$$\mathrm{MSE}(\widehat{\beta}) = \sigma^2\mathrm{trace}((\mathbf{X}^T\mathbf{X})^{-1}). \tag{11}$$
- 设$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ 为$\mathbf{X}^T\mathbf{X}$ 的特征根，因为$\mathbf{X}^T\mathbf{X}$ 可逆，所以$(\mathbf{X}^T\mathbf{X})^{-1}$ 的特征根为
$$\lambda_1^{-1}, \cdots, \lambda_p^{-1}$$
所以(11) 变为
$$\mathrm{MSE}(\widehat{\beta}) = \sigma^2 \sum_{k=1}^{p} \frac{1}{\lambda_k}. \tag{12}$$

- 由此可以看出：
  - 如果$\mathbf{X}^T\mathbf{X}$ 至少有一个特征根特别小且接近于零，
  则MSE$(\widehat{\beta})$ 就会非常大。
  - 从均方误差的标准来看，这时的LSE就不是一个好估计。
- 由于

$$\text{MSE}(\widehat{\beta}) = E(\widehat{\beta}-\beta)^T(\widehat{\beta}-\beta) = E\|\widehat{\beta}\|^2 - \beta^T\beta,$$

从而，

$$E\|\widehat{\beta}\|^2 = \|\beta\|^2 + \text{MSE}(\widehat{\beta}) = \|\beta\|^2 + \sigma^2\sum_{k=1}^{p}\frac{1}{\lambda_k}.$$

- 即，自变量的数据相关矩阵$\mathbf{X}^T\mathbf{X}$ 至少有一个特征根特别小且接近于零时，就会导致：
  - $\widehat{\beta}$ 的长度要比真正的未知向量$\beta$的长度长得多；
  - $\widehat{\beta}$ 的某些分量的绝对值非常之大。

- 总而言之，$\mathbf{X}^T\mathbf{X}$ 极小的特征根就会导致LSE $\widehat{\beta}$ 不再是一个好的估计。

## 进一步的讨论：

- 设$x_{(k)}$ 为$\mathbf{X}$ 的第$k$ 列，$\lambda$ 为$\mathbf{X}^T\mathbf{X}$ 的一个特征根，$\eta = (\eta_1, \cdots, \eta_p)^T$ 为相应的单位特征向量，如果$\lambda \approx 0$, 则

$$\mathbf{X}^T\mathbf{X}\eta = \lambda\eta \approx 0.$$

- 用$\eta^T$ 左乘上式两端，

$$\eta^T\mathbf{X}^T\mathbf{X}\eta = \lambda\eta^T\eta \approx 0.$$

- 于是

$$\mathbf{X}\eta \approx 0.$$

即

$$\eta_1 x_{(1)} + \cdots + \eta_p x_{(p)} \approx 0 \tag{13}$$

- 从(13) 可以看出，对于$n$组数据有

$$\eta_1 \mathbf{X}_{(1)} + \cdots + \eta_p \mathbf{X}_{(p)} \approx 0. \tag{14}$$

这就是所谓的线性回归模型存在共线性。在有些教科书也称为设计矩阵是病态的(ill-conditioning)。

## 条件数:

- $\mathbf{X}^T\mathbf{X}$ 有几个很小的特征根，设计阵就可能有几个共线性关系存在;

- 方阵$\mathbf{X}^T\mathbf{X}$ 的条件数定义为

$$k = \frac{\lambda_1}{\lambda_p}$$

条件数刻画了$\mathbf{X}^T\mathbf{X}$ 的特征根的散布程度。可以用来判断共线性是否存在以及共线性的严重程度。

## 条件数的经验值：

在实际应用中，

- 若$k < 100$, 共线性的程度很小；
- 若$100 \leq k \leq 1000$, 共线性的程度中等或者较强；
- 若$k > 1000$, 共线性的程度非常严重。

**Example**

下面的表中给出了有6 个协变量的回归问题的原始数据。

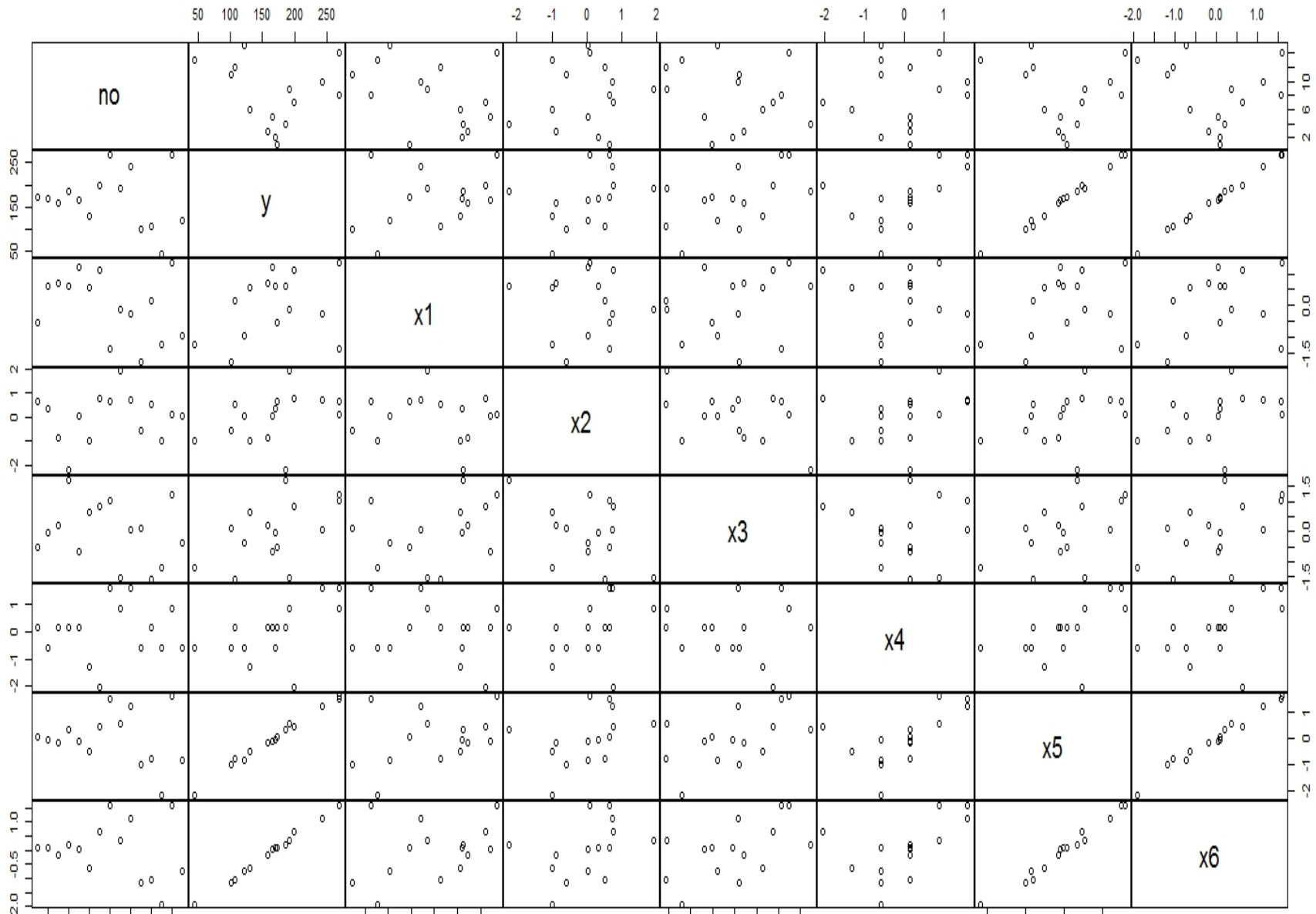
# 原始数据

| no | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|---|
| 1 | 172.480 | -0.399 | 3.057 | 2.298 | 3 | 8.888 | 17.23 |
| 2 | 169.085 | 0.690 | 2.438 | 3.355 | 2 | 8.547 | 17.11 |
| 3 | 158.530 | 0.815 | 0.156 | 3.944 | 3 | 8.180 | 15.46 |
| 4 | 185.481 | 0.711 | -2.341 | 7.330 | 3 | 9.700 | 17.84 |
| 5 | 165.873 | 1.290 | 1.881 | 1.947 | 3 | 8.330 | 16.89 |
| 6 | 129.689 | 0.668 | -0.021 | 4.869 | 1 | 7.015 | 12.58 |
| 7 | 198.504 | 1.190 | 3.228 | 5.397 | 0 | 10.107 | 20.70 |
| 8 | 268.064 | -1.202 | 3.015 | 5.822 | 5 | 13.308 | 26.51 |
| 9 | 192.043 | -0.019 | 5.384 | 0.023 | 4 | 10.346 | 18.91 |
| 10 | 242.086 | -0.156 | 3.182 | 3.636 | 5 | 12.428 | 23.82 |
| 11 | 101.139 | -1.604 | 0.712 | 3.713 | 2 | 5.488 | 9.12 |
| 12 | 107.732 | 0.257 | 2.760 | -0.023 | 3 | 6.125 | 9.91 |
| 13 | 44.947 | -1.056 | -0.018 | 0.773 | 2 | 1.794 | 4.48 |
| 14 | 268.694 | 1.415 | 1.961 | 6.246 | 4 | 13.637 | 26.74 |
| 15 | 121.015 | -0.805 | 1.903 | 2.605 | 2 | 5.992 | 12.02 |

## 探索性数据分析

- $Y$与$x_1, x_2, \cdots, x_6$ 之间是否有关系？
- 如果有，是什么关系？
- 看图学统计！
- 请看下面的散点图：

## 例



- 尝试线性模型
- 利用lm 命令建立线性回归模型

$$\begin{aligned} Y_i &= \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 \\ &\quad + x_6\beta_6 + \epsilon_i, \\ &\quad 1 \le i \le 15. \end{aligned} \tag{15}$$

回归系数等结果如下：

Call: lm(formula $= y \sim x1 + x2 + x3 + x4 + x5 + x6$)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -0.13383 | -0.01544 | 0.00290 | 0.03692 | 0.09700 |

Coefficients:

| | Estimate | Std. Error | t value | Pr($>\|t\|$) |
|---|---|---|---|---|
| (Intercept) | -0.11329 | 0.06824 | -1.66 | 0.135 |
| x1 | 0.97912 | 0.07589 | 12.90 | 1.23e-06 |
| x2 | 2.04989 | 0.08741 | 23.45 | 1.16e-08 |
| x3 | 3.03878 | 0.08884 | 34.20 | 5.83e-10 |
| x4 | 4.02942 | 0.08784 | 45.87 | 5.63e-11 |
| x5 | 4.95964 | 0.07462 | 66.47 | 2.92e-12 |
| x6 | 6.00878 | 0.02742 | 219.16 | $<$ 2e-16 |

Residual standard error: 0.08256 on 8 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 1.332e+06 on 6 and 8 DF, p-value: $<$ 2.2e-16

## 例

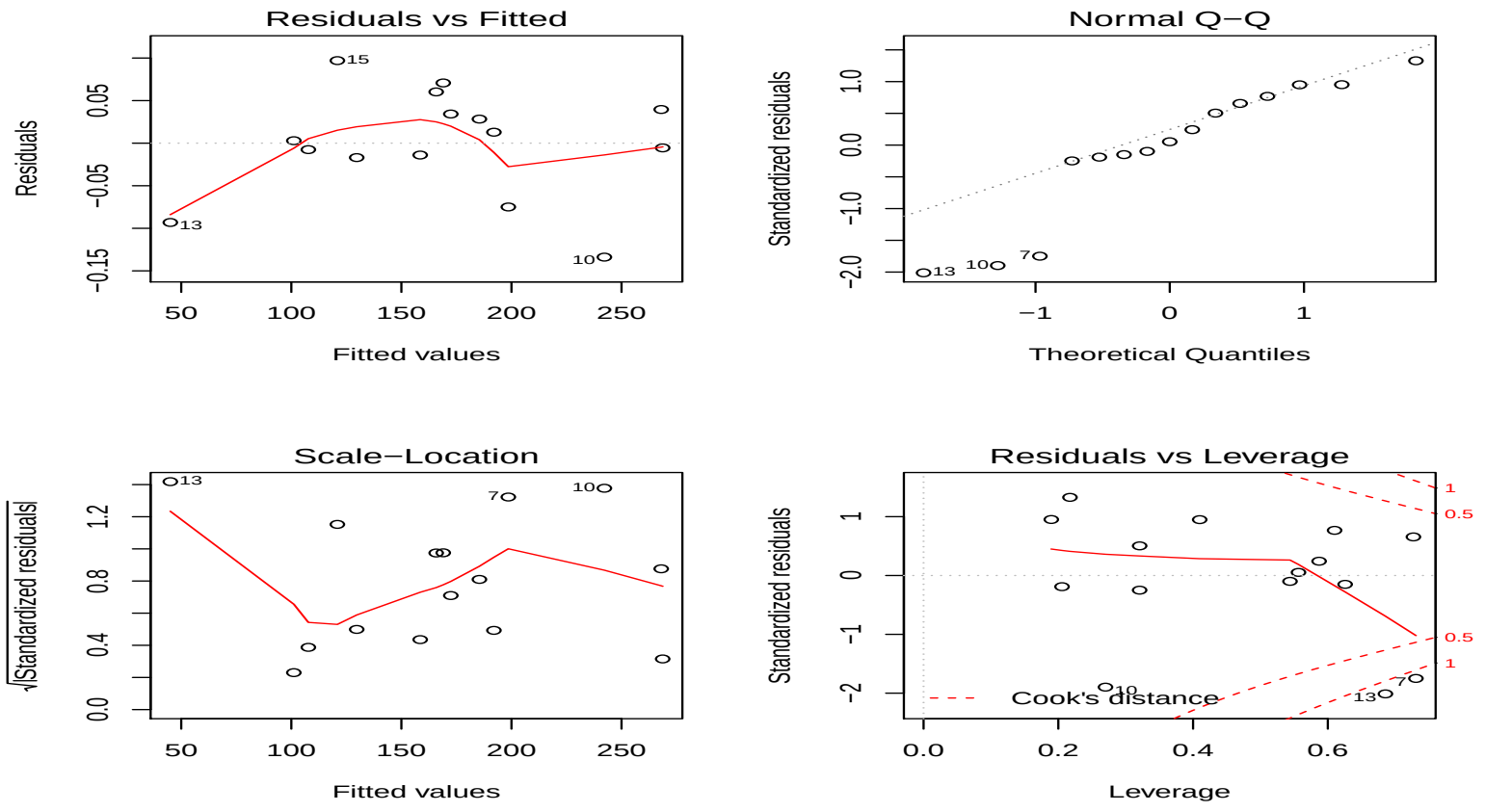



**Figure:** Model 0123456 with intercept

## 分析 1：

从上面的结果可以看出：

- intercept 项是不显著的。
- 除此之外，其余的$t$ 检验、$F$检验都是显著的。
- $R^2 = 1.$
- 一切看上去都很好。
- 是否可以断定$y$ 与$x1,\cdots,x6$ 的之间的关系是线性关系？
- 是否所有的变量都是必要的？
- 下面将截距项去掉作线性回归，结果如下：

Call:

$$\text{lm(formula} = y \sim 0 + x1 + x2 + x3 + x4 + x5 + x6)$$

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -0.1767 | -0.0432 | 0.0197 | 0.0310 | 0.0955 |

Coefficients:

| | Estimate | Std. Error | t value | Pr($>\|t\|$) |
|---|---|---|---|---|
| x1 | 0.98438 | 0.08290 | 11.88 | 8.42e-07 |
| x2 | 2.04354 | 0.09547 | 21.41 | 4.98e-09 |
| x3 | 3.03015 | 0.09696 | 31.25 | 1.72e-10 |
| x4 | 4.02093 | 0.09586 | 41.94 | 1.24e-11 |
| x5 | 4.94186 | 0.08073 | 61.21 | 4.18e-13 |
| x6 | 6.01578 | 0.02962 | 203.12 | $<$ 2e-16 |

Residual standard error: 0.09026 on 9 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 9.813e+06 on 6 and 9 DF, p-value: $<$ 2.2e-16

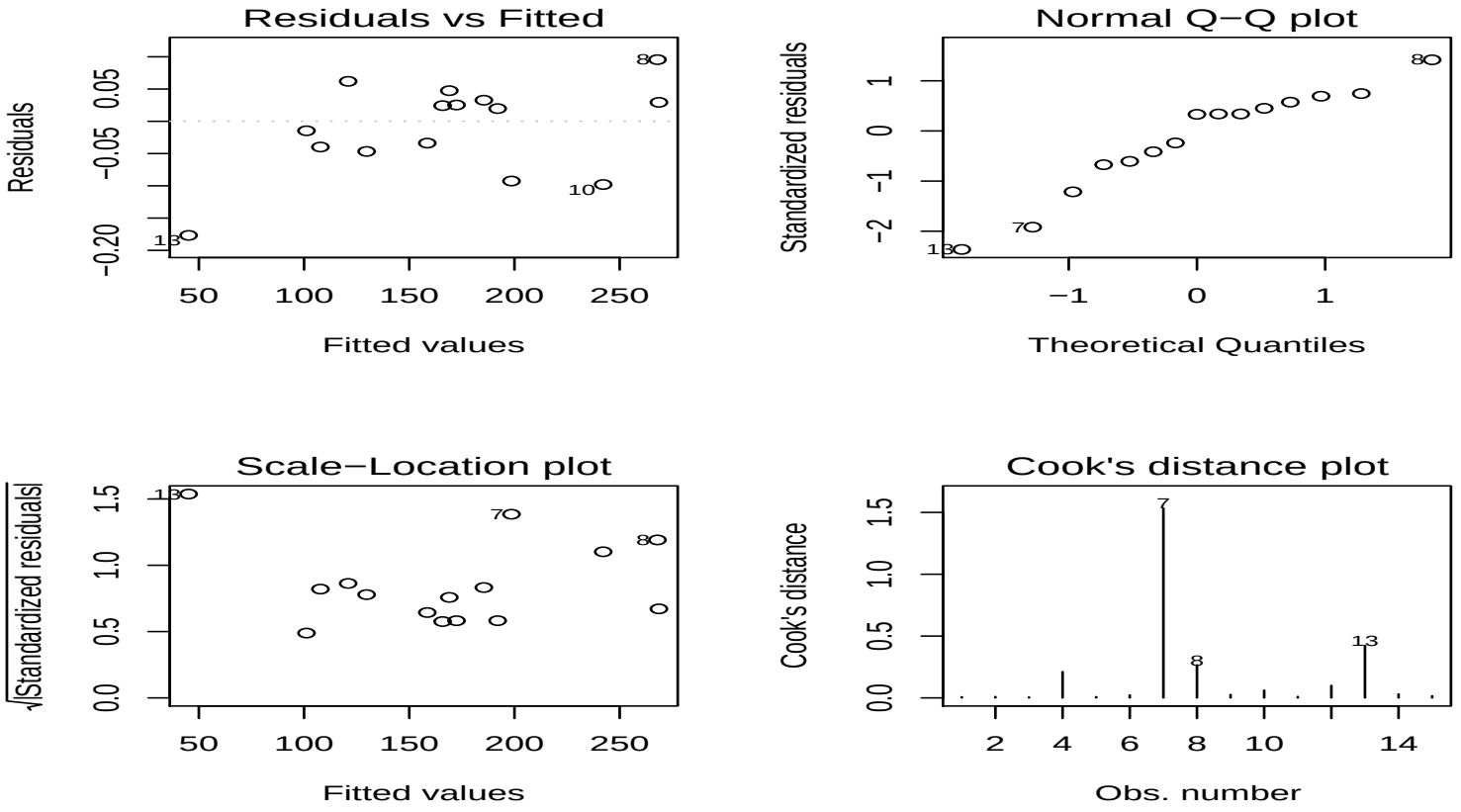



**Figure:** Model 123456 without intercept

## 例

- 从上面的结果可以看出，去掉截距项之后，一切似乎表现的很完美(t, F, $R^2$ 等)
- 是否存在共线性？
- $\mathbf{X}^T\mathbf{X}$ 的六个特征根为

$$\begin{aligned} \lambda_1 &= 6353.4031059 \quad \lambda_2 = 86.3911570 \\ \lambda_3 &= 23.7276929 \quad \lambda_4 = 11.8731137 \\ \lambda_5 &= 2.6619385 \quad \lambda_6 = 0.2155923 \end{aligned}$$

【$\mathbf{X}$ 标准化之后的特征根为40.61645975, 23.16515389, 13.29941369 6.71022405, 0.15467628, 0.05407234】

例

- 条件数为

$$k = \frac{6353.4031059}{0.2155923} = 29469.52 >> 1000$$

【标准化情形:
$k = 40.61645975/0.05407234 = 751.1504 > 100$】

- 条件数远远大于1000【标准化情形：：大于100】，根据前面的标准，模型(15) 的设计阵存在严重(或者较严重的)的共线性，因为$\lambda_6 = 0.2155923$ 相对很小, 其对应的特征向量为

$$\eta^T = (0.408, 0.482, 0.492, 0.486, -0.343, -0.0611)$$

【标准化情形：$(-0.192, -0.446, -0.546, -0.330, 0.575, 0.156)$】

## 例

- 因而回归自变量之间有如下的的共线性关系：

$$0.408x_1 + 0.482x_2 + 0.492x_3 + 0.486x_4 - 0.343x_5 - 0.0611x_6 \approx 0$$

【标准化情形：$-0.192x_1 - 0.446x_2 - 0.546x_3 - 0.330x_4 + 0.575x_5 + 0.156x_6 \approx 0$】

- 上式中$x_6$和$x_5$ 系数的符号相同，可以将系数较小的一个变量舍弃。得到

$$0.40896584x_1 + 0.48240293x_2 + 0.49203282x_3 + 0.48603711x_4 - 0.34345798x_5 \approx 0$$

- 由此推断$x_1, x_2, x_3, x_4$ 和$x_5$ 又有共线性关系。
利用回归R: lm($x_5 \sim 0 + x_1 + x_2 + x_3 + x_4$) 可检验之。

例

- 利用变量$x_1, x_2, x_3$ 和$x_4$ 作回归分析(有截距项)，结果如下：
- Call: lm(formula $= y \sim x1 + x2 + x3 + x4$).
- Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -11.7813 | -2.2789 | 0.4507 | 3.8036 | 8.0045 |

- Coefficients:

| | Estimate | Std. Error | t value | Pr($>\|t\|$) |
|---|---|---|---|---|
| (Intercept) | -1.5305 | 5.2127 | -0.294 | 0.775 |
| x1 | 15.4653 | 1.8960 | 8.157 | 9.93e-06 |
| x2 | 19.9389 | 1.0905 | 18.284 | 5.15e-09 |
| x3 | 21.1995 | 0.9061 | 23.396 | 4.61e-10 |
| x4 | 20.8403 | 1.3469 | 15.472 | 2.59e-08 |

- Residual standard error: 6.426 on 10 degrees of freedom
- Multiple R-Squared: 0.9924, Adjusted R-squared: 0.9894
- F-statistic: 327.3 on 4 and 10 DF, p-value: 1.491e-10

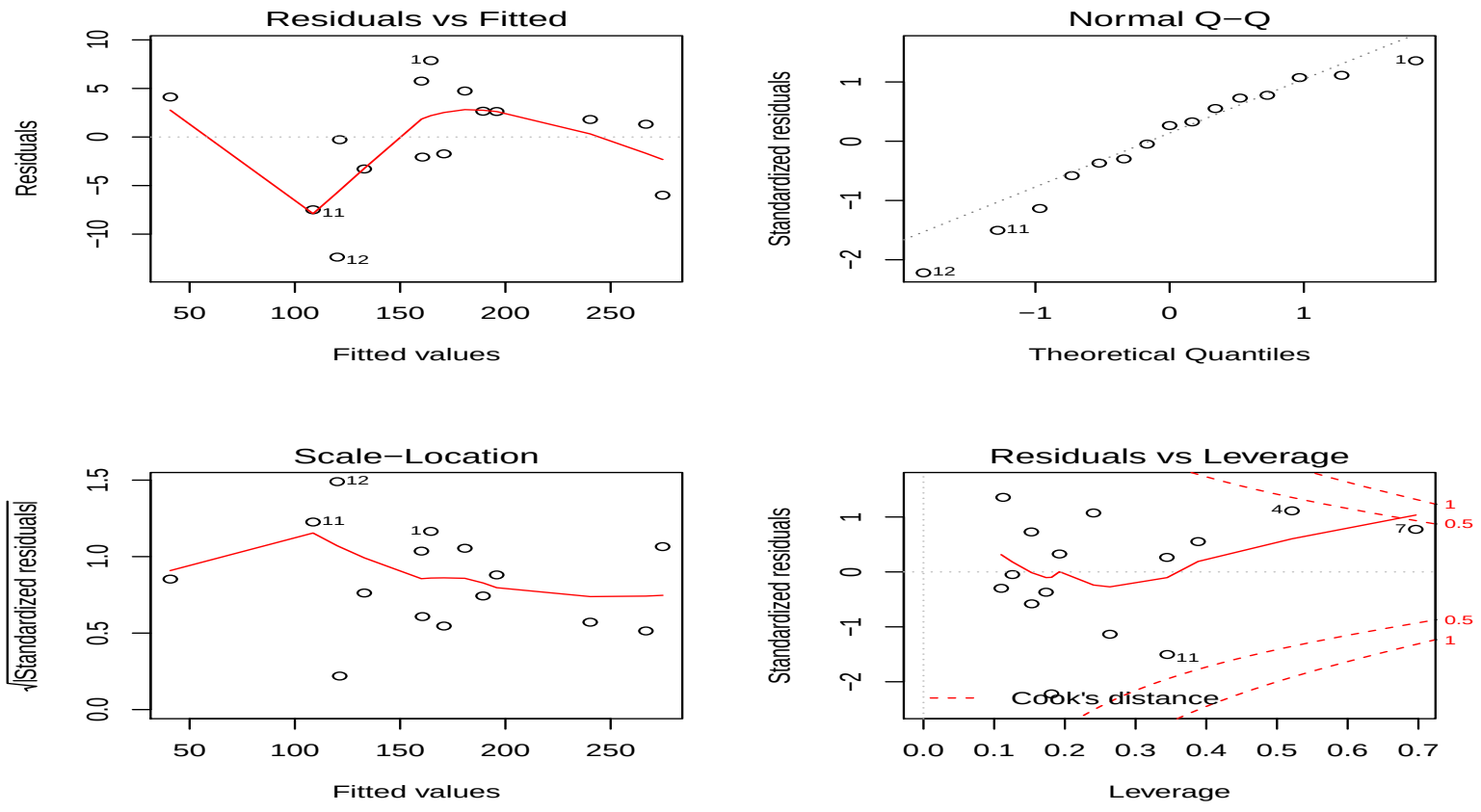



**Figure:** Model 01234 with intercept

例

- 利用变量$x_1, x_2, x_3$ 和$x_4$ 作回归分析(没有截距项)，结果如下：
- Call: lm(formula $= y \sim 0 + x1 + x2 + x3 + x4$).
- Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -12.366 | -2.679 | 1.317 | 3.389 | 7.869 |

- Coefficients:

| | Estimate | Std. Error | t value | Pr($>|t|$) |
|---|---|---|---|---|
| x1 | 15.4888 | 1.8139 | 8.539 | 3.50e-06 |
| x2 | 19.8055 | 0.9493 | 20.862 | 3.40e-10 |
| x3 | 21.0300 | 0.6688 | 31.443 | 4.00e-12 |
| x4 | 20.6443 | 1.1203 | 18.427 | 1.28e-09 |

- Residual standard error: 6.154 on 11 degrees of freedom
- Multiple R-Squared: 0.9991, Adjusted R-squared: 0.9988
- F-statistic: 3164 on 4 and 11 DF, p-value: $<$ 2.2e-16

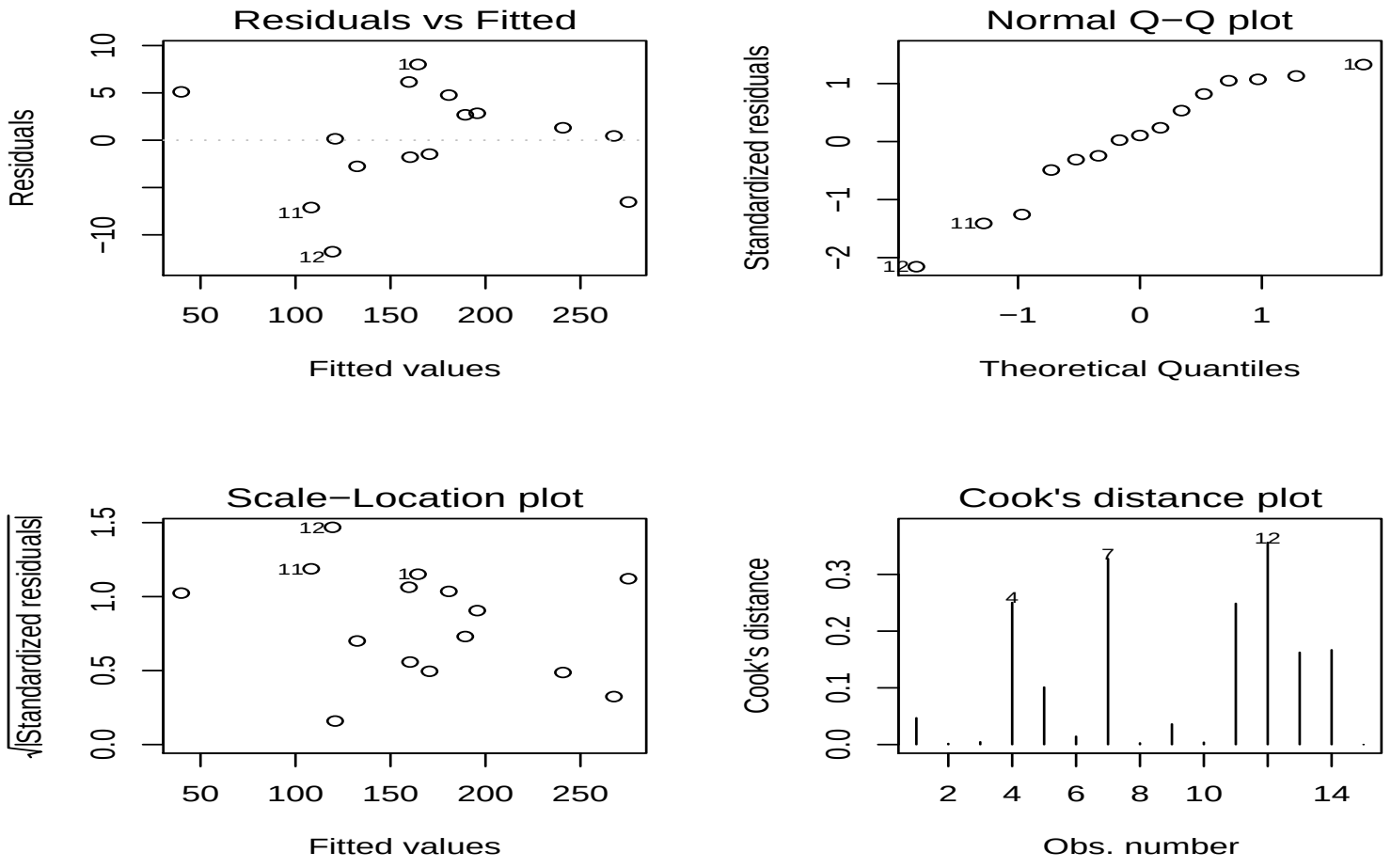



**Figure:** Model 01234 without intercept

## 例

- 最后确定的可用的回归模型为(注意，不是唯一的模型！)

$$\hat{y} = 15.4888x_1 + 19.8055x_2 + 21.0300x_3 + 20.6443x_4.$$

## 计算方差膨胀因子

方差的膨胀系数

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}, \ j = 1, \cdots, k. \tag{16}$$

$\mathrm{VIF}_j$ 可用来检查每一个协变量所受到多重共线性的影响大小。可将特别大的$VIF_j$ 对应的协变量$x_j$剔除之，以消除共线性。

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $R_j^2$ | 0.909 | 0.9817 | 0.9877 | 0.9665 | 0.9913 | 0.9839 |
| $\mathrm{VIF}_j$ | 10.989 | 54.644 | 81.300 | 29.850 | 114.942 | 62.111 |

说明变量$x_5$与其余五个自变量呈高度线性相关，可以把$x_5$从线性模型$Y = \sum_{i=1}^{6} \beta_i x_i$中剔除！

将自变量$x_5$剔除之后进一步的分析：

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $R_j^2$ | 0.8033 | 0.9601 | 0.9722 | 0.9128 | × | 0.9833 |
| $\text{VIF}_j$ | 5.083 | 25.062 | 35.971 | 11.467 | × | 59.880 |

说明变量$x_6$与其余四个自变量呈高度线性相关，可以把$x_4$从线性模型中剔除！

进一步将自变量$x_6$剔除之后做方差膨胀因子分析：

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $R_j^2$ | 0.1162 | 0.2894 | 0.2807 | 0.1377 | × | ×× |
| $\text{VIF}_j$ | 1.131 | 1.407 | 1.390 | 1.159 | × | ×× |

这时方差膨胀因子达到了合理的水平！可以利用$x_1, x_2, x_3$ 和$x_4$与$y$ 建立线性模型！

- 如果自变量的个数较多时，诊断共线性是一个反复的过程，可能需要多步！
- 事实上，我们可以利用方差膨胀因子去选择模型/变量！
- 方差膨胀因子也可以调用R 的packages ‘car’快速得到：
  library('car') # Companion to Applied Regression
  $vif(lm(y \sim x1 + x2 + x3 + x4 + x5 + x6, data = dat))$
  得到！
- 【了解‘car’的功能！】

## 关于方差膨胀因子的说明：

- 方差膨胀因子是通过确定性系数定义的，膨胀系数越大，$R_j^2$越接近于1，因此共线性越强；
- 方差膨胀因子可以度量整体的共线性；
- 方差膨胀因子同时继承了确定性系数的缺点：它不能甄别变量之间可能共存的其它相关性.
- 更进一步的内容请阅读：
  - Chatterjee, S. and B. Price (1977) *Regression Analysis by Example*, John Wiley and Sons: New York.
  - Chatterjee, S. and A. S. Hadi (2006) *Regression Analysis by Example*, 4th Ed. John Wiley and Sons: New Jersey. (有中译本)

## 关于回归诊断的参考书

- David A. Belsley, Edwin Kuh, Roy E. Welsch (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley-Interscience
- Peter J. Rousseeuw, Annick M. Leroy (1987) Robust Regression and Outlier Detection. Wiley.
- Samprit Chatterjee, Ali S. Hadi (1988). Sensitivity Analysis in Linear Regression. Wiley

## 共线性产生的原因：

- 数据收集的局限性；
- 自变量客观上有近似的线性关系；
- 数据的再加工,……

- 注意：在许多大型的回归问题中，由于人们对于自变量之间的关系缺乏足够的认识，很有可能把有共线性关系的变量引入回归方程，可能会导致LSE的性质不理想和不稳定。
- 甄别和消除共线性在回归模型的实际应用中是非常重要的。

## 关于多重共线性进一步的说明

- 完全共线性: 如果$|X^TX| = 0$,(通常需要先标准化数据，再做分析)
- 不完全共线性: 如果$|X^TX| \approx 0$,
- 前面在考察共线性时, 只是从设计矩阵是否列相关程度上考虑的.
- 事实上, 把每一列向量看作是一个随机变量的实现时, 利用相关系数也可以判断两个变量之间是否存在共线性. 在实际中,若两个变量的相关系数(的绝对值)较大时, 例如,大于0.8, 则认为变量之间存在共线性.[相当于数据标准化之后再做共线性判断!]

## 共线性的相关系数检验法

- 首先将$X$的列向量$x_j = (x_{1j}, \cdots, x_{nj})^T$ 标准化:

$$x_j^* = (x_j - \bar{x}_j)/\mathsf{SS}_{jj}, j = 1, \cdots, p$$

其中$\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$, $\mathsf{SS}_{jj} = \left[\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2\right]^{1/2}$.

- 以$x_j^*(j = 1, \cdots, p)$为列向量形成新的关联矩阵$X^*$, 则

$$X^{*T}X^* = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

其中$r_{kj}$ 是$x_k^*$ 和$x_j$ 之间的相关系数.

## 共线性的相关系数检验法(续)

- 从$r_{kj} \approx 1$可以看出两个变量共线性的;
- 当要判断多个变量之间的共线性时,只须判断$X^{*T}X^*$ 的最小特征根是否接近或者等于零即可.

## 如何克服共线性

- 在可控的试验中,事先设计好自变量的观测值, 预先避免观测变量之间的共线性或者降低共线性;[思考：如何设计可达到此要求？]
- 当自变量是不可控制时, 当在数据中已经发生较为严重的共线性时, 要消除多于的变量。

## 消除共线性的方法: 岭估计(ridge regression)

- 牺牲估计的无偏性, 以降低估计的方差.【以MSE 为评价标准。】
- 通常的LS估计为: $\widehat{\beta} = (X^T X)^{-1} X^T Y$,
- 岭估计为: $\widehat{\beta}_{(k)} = (X^T X + kI_p)^{-1} X^T Y$, 其中$k$是适当选取的正数, 使得

$$\mathrm{MSE}(\widehat{\beta}_{(k)}) \leq \mathrm{Var}(\widehat{\beta}).$$

- 记$S = X^T X$, $S$的特征根为$\lambda_1 \geq \cdots \geq \lambda_p > 0$, 则可以证明:

$$\mathrm{MSE}(\widehat{\beta}) = \sum_{j=1}^{p} \mathrm{Var}(\widehat{\beta}_j) = \sigma^2 \mathrm{trace}(S^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

## 消除共线性的方法: 岭估计(ridge regression)(续)

- 当存在比较严重的共线性时, 有部分的$\lambda_j$ 的值会非常之小且接近于0, 从而: 使得MSE$(\widehat{\beta})$ 变得非常之大.
- 另外对于岭估计$\widehat{\beta}_{(k)}$, 有

$$\text{MSE}(\widehat{\beta}_{(k)}) = \sum_{j=1}^{p} \frac{\sigma^2 \lambda_j + k^2 \beta_j^2}{(\lambda_j + k)^2}.$$

- 可以适当的选择$k > 0$ 使得:

$$\text{MSE}(\widehat{\beta}_k) \leq \text{Var}(\widehat{\beta}),$$

这就是ridge 回归的基本思想。

## 关于判断共线性的进一步说明：

- 共线性通常只对连续性的变量来判断。
- 若自变量中既有连续变量，又有分类变量的话。正常情况下连续变量不会被分类变量线性表示，分类变量应该不参与共线性的判断。
- 如果有多个分类变量，需要对分类变量判别共线性。多个分类变量如果不经过人工选择，有可能会有某个因为逻辑关系、归属关系等，被别的分类变量确定。比如，同时有省名与省编码。
- 如果变量中有多个连续变量和多个分类变量，则需要量两类变量分开，然后分别判断是否有共线性。
- 如果一个变量时连续型的，另外一个变量是分类变量，Pearson 相关系数并不是合适的度量相关性的指标（个人观点）。
- 哑变量通常不需要标准化。

## 作业

1. 阅读教材Seber and Lee (2003), p. 249-263
2. 阅读教材Draper and Smith (1998), p. 369-386 以及p. 387-400
3. 在阅读教材的基础上，试评价两本书上关于共线性论述的异同，你有什么看法？
4. 安装R的package ‘car’，熟悉vif的使用方法。利用data()查看R中自带的数据集，自行选择一个数据集，建立线性模型，并判断共线性是否存在。