第03章 放宽假设

3.1 多重共线性(Multi-collinearity)问题

3.2 异方差(hetro-scadasticity)问题

3.3 自相关(auto-correlation)问题

3.4 内生自变量(endogeneity-variable)问题

3.1 多重共线性(Multi-collinearity)问题

3.1.1 重共线性的定义和来源

3.1.2 多重共线性的影响和后果

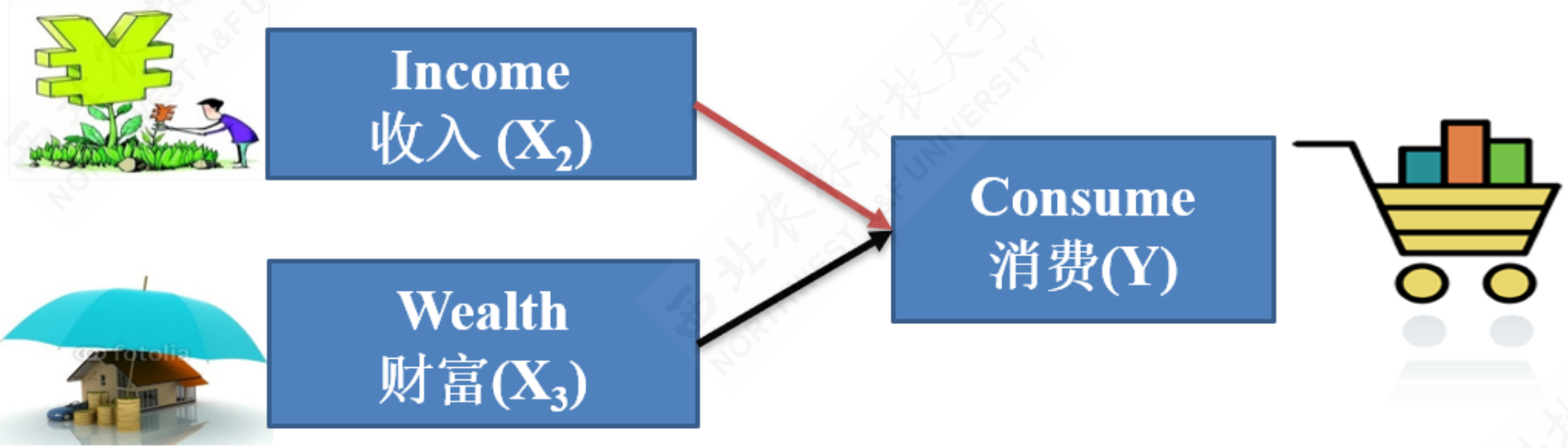
3.1.3 多重共线性问题的诊断

3.1.4 多重共线性问题的矫正

3.1.1 重共线性的定义和来源

消费案例：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$



- Y_i表示消费
- X_{2i}表示收入； X_{3i}表示财富。

引子2

慈善捐款案例：

回归模型：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$

- Y_i表示慈善捐款
- X_{2i}表示城市总人口； X_{3i}表示城市GDP。

现实情况：城市GDP=城市总人口*人均GDP

多重共线性来源：自变量间表现为某种因果关系

引子3

消费者忠诚度案例：

回归模型：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$

- Y_i表示消费者忠诚度
- X_{2i}表示产品满意度； X_{3i}体验满意度。

现实情况：收入水平与满意度有关（奢侈品、拼多多）

多重共线性来源：自变量背后有共同的潜在因素

什么是模型多重共线性？

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad \text{(PRM)}$$
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + e_i \qquad \text{(SRM)}$$

多重共线性：在多元线性回归模型中，各解释变量$\{X_2, X_3, \cdots, X_k\}$之间有交互相关，但又非完全相关的现象。正式地：

$$\lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} + v_i = 0$$

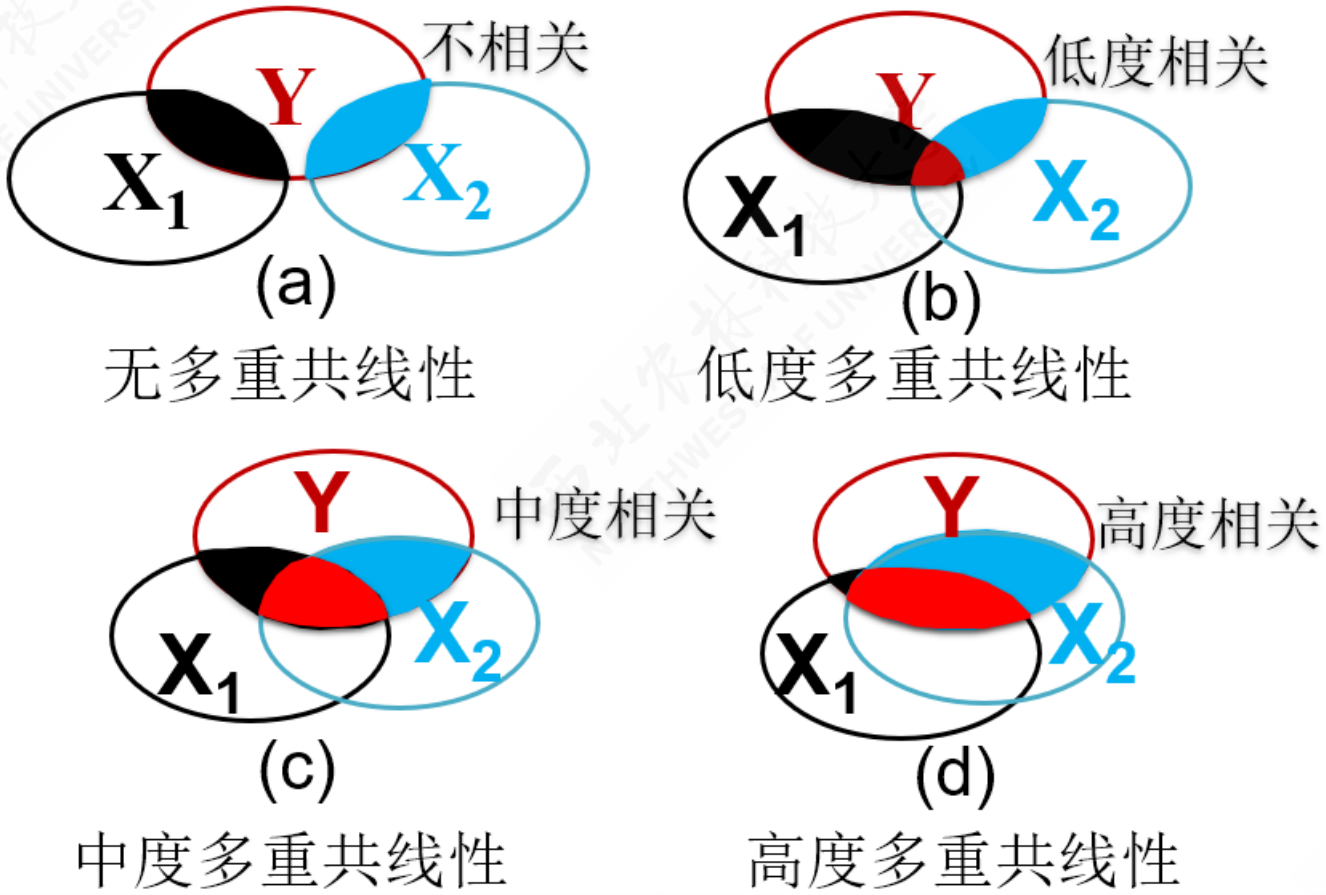
其中，v_i为随机误差项。

我们称总体回归模型存在**多重共线性**，此时$\lambda_1; \lambda_2; \cdots; \lambda_k$不全为0，且$v_i \neq 0$。

我们称总体回归模型存在**完全多重共线性**，此时$\lambda_1; \lambda_2; \cdots; \lambda_k$不全为0，且$v_i = 0$。

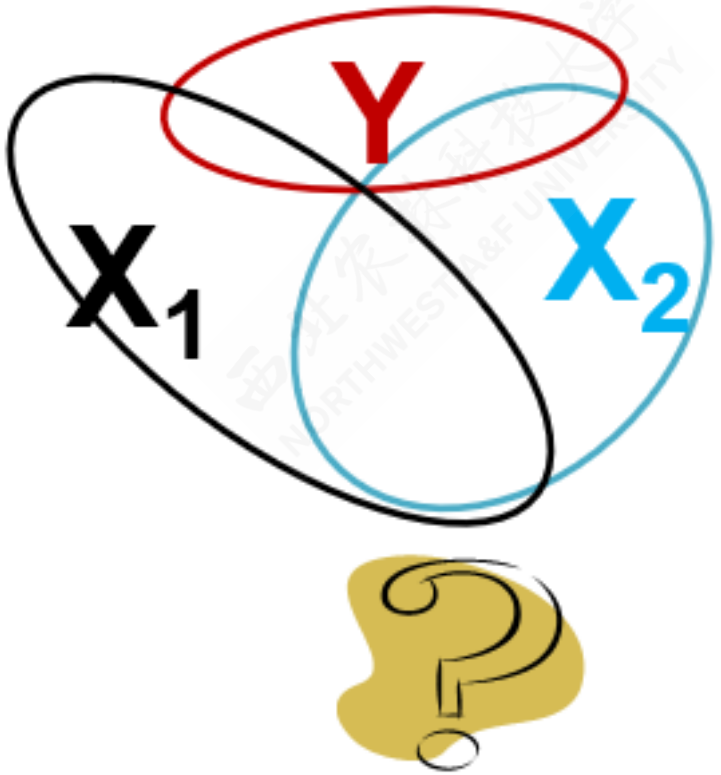
什么是模型多重共线性？

下面用一个直观图进行说明：



什么是模型多重共线性？

- 简单相关r(Pearson Relationship)
- 多重共线性(Multicollinearity)



引起多重共线性的原因

- 数据采集所用的方法: 回归元限于一个范围
- 模型或从中取样的总体受到约束：如做电力消费（Y）对收入（X2）和住房面积（X3）回归时，可能X2高的X3也大
- 模型设定: 如在模型中加入多项式；数据范围小等
- 一个过度决定的模型：回归元个数大于观测次数。医学研究中可能只有少数病人
- 相同的时间趋势。消费支出(Y)对收入、财富和人口的回归

完全多重共线性时的估计问题

在完全多重共线性的情况下，回归系数是不确定的。

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} - \hat{\beta}_3 \overline{X}_3$$

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

完全多重共线性时的估计问题

在完全多重共线性的情况下，回归系数是不确定的。

假设**完全共线性**情形下，$X_{3i}=\lambda X_{2i}$，则容易发现：

$$\hat{\beta}_2=\frac{\left(\sum y_i x_{2i}\right)\left(\lambda^2\sum x_{2i}^2\right)-\left(\lambda\sum y_i x_{2i}\right)\left(\lambda\sum x_{2i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\lambda^2\sum x_{2i}^2\right)-\lambda^2\left(\sum x_{2i}^2\right)^2}=\frac{0}{0}$$

完全多重共线性时的估计问题

在完全多重共线性的情况下，回归系数是不确定的。

假定**不完全多重共线性**下，$x_{3i}=\lambda x_{2i}+v_i$，则有：

$$\hat{\beta}_2=\frac{\sum(y_ix_{2i})\left(\lambda^2\sum x_{2i}^2+\sum v_i^2\right)-\left(\lambda\sum y_ix_{2i}+\sum y_iv_i\right)\left(\lambda\sum x_{2i}^2\right)}{\sum x_{2i}^2\left(\lambda^2\sum x_{2i}^2+\sum v_i^2\right)-\left(\lambda\sum x_{2i}^2\right)^2}$$

如果v_i足够小，以至于接近于零，则上式将表示**完全共线性**情形。

3.1.2 多重共线性的影响和后果

多重共线性的理论后果

如果模型出现多重共线性问题，在N-CLRM假设下，OLS估计量仍然是最优线性无偏估计量（BLUE）：

- 只要不是**完全共线性**，在近似多重共线性的情形下，OLS估计量仍然是无偏的
- 只要不是**完全共线性**，在近似多重共线性的情形下，OLS估计量的方差一定是小的
- 多重共线性本质上是一种样本现象。即使总体中X变量间不存在共线性，由于抽样方法或小样本问题，也可能带来多重共线性问题

多重共线性的实际后果

实际后果1：更大的方差和协方差，估计精度大大下降。$\hat{\beta}_2$和$\hat{\beta}_3$的**真实方差**分别为：

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_{2i}^2 \left(1 - r_{23}^2\right)} \equiv \frac{\sigma^2}{\sum x_{2i}^2 \cdot TOL} \equiv \frac{\sigma^2}{\sum x_{2i}^2} \cdot VIF$$

$$\sigma^2_{\hat{\beta}_3} = \frac{\sigma^2}{\sum x_{3i}^2 \left(1 - r_{23}^2\right)} \equiv \frac{\sigma^2}{\sum x_{3i}^2 \cdot TOL} \equiv \frac{\sigma^2}{\sum x_{3i}^2} \cdot VIF$$

$$\operatorname{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) = \frac{-r_{23}\sigma^2}{\left(1 - r_{23}^2\right)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}} \quad \leftarrow \left[r_{23}^2 = \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}\right]$$

$$r_{23}^2 \equiv \frac{\left(\sum x_{2i}x_{3i}\right)^2}{\sum x_{2i}^2 \sum x_{3i}^2}; \quad TOL \equiv \left(1 - r_{23}^2\right); \quad VIF \equiv \frac{1}{\left(1 - r_{23}^2\right)}$$

随着r_{23}的增大，方差和协方差的绝对值也增大。

多重共线性的实际后果

方差增大的速度用**方差膨胀因子**(VIF, variance-inflating factor)衡量：

$$VIF = \frac{1}{(1 - r_{23}^2)}$$

容忍度(tolerance , TOL)定义为VIF的倒数：

$$\text{TOL}_\text{j} = \frac{1}{VIF_\text{j}} = 1 - R_j^2$$

R_j^2表示 X_j对其余(k-2)个回归元进行回归的判定系数

- 当 $R_j^2 = 1$，即完全共线性时， $TOL_j = 0$;
- 当 $R_j^2 = 0$，即不存在共线性时， $TOL_j = 1$;
- 由于VIF 和TOL 之间有密切关系，所以可以将它们互换使用。

> 注意：（古扎拉蒂）在k个变量的回归模型中有是k-1个回归元。

多重共线性的实际后果

实际后果2：置信区间变宽，系数的检验倾向于不显著！即更倾向接受原假设H0，认为系数为零。

- 标准误增大，则有关总体参数的置信区间随之变大。

$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-k) \cdot S_{\hat{\beta}_1}$$
$$\hat{\beta}_2 \pm t_{1-\alpha/2}(n-k) \cdot S_{\hat{\beta}_2}$$
$$\hat{\beta}_3 \pm t_{1-\alpha/2}(n-k) \cdot S_{\hat{\beta}_3}$$

多重共线性的实际后果

实际后果3： 系数的t值倾向于统计上不显著，但 R^2 却会很高。

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

$$t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$

$$t^*_{\hat{\beta}_3} = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}}$$

将 $t^*_{\hat{\beta}_i}$ 值和临界t值相比较。高度共线性使估计的标准误增加很快，t值迅速变小。

因而，在高度多重共线性的情形下，增加了接受错误假设的概率（第二类错误）

多重共线性的实际后果

实际后果3：系数的t值倾向于统计上不显著，但 R^2 却会很高。

- 在高度共线性情形中，有可能会发现一个或多个偏斜率系数基于t检验不是个别统计显著的，然而这时 R^2 却高达(比如说)0.9以上，从而根据F检验，可令人信服地拒绝原假设：

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

- 但是，个别偏回归系数的t检验可能并不显著——这就是多重共线性的一个信号
- 这里的真正问题在于估计量之间的协方差，而这些协方差是同回归元之间的相关性有关系的。

多重共线性的实际后果

实际后果4：OLS估计量及其标准误对数据的微小变化非常敏感。

说明性例子（数据）

Year	Y	X2	X3	X4
1947	976.4	1035.2	5166.815	-10.35094
1948	998.1	1090	5280.757	-4.719804
1949	1025.3	1095.6	5607.351	1.044063
1950	1090.9	1192.7	5759.515	0.407346
1951	1107.1	1227	6086.056	-5.283152
1952	1142.4	1266.8	6243.864	-0.277011
1953	1197.2	1327.5	6355.613	0.561137
1954	1221.9	1344	6797.027	-0.138476

Showing 1 to 8 of 54 entries

Previous 1 2 3 4 5 6 7 Next

消费支出与收入和财富的数据(n=54)

- Y_i表示人均消费支出
- X_{2i}表示人均收入； X_{3i}表示财富； X_{4i}表示利率

说明性例子（精简分析报告）

我们可以构建如下的回归模型：

$$log(Y) = + \hat{\beta}_1 + \hat{\beta}_2 log(X2) + \hat{\beta}_3 log(X3) + \hat{\beta}_4 X4 + e_i$$

计算并整理回归分析结果如下：

$$\begin{array}{lllll}
\widehat{log(Y)} = -0.47 & & +0.80log(X2) + 0.20log(X3) - 0.00X4 & & \\
(\mathrm{t}) & (-10.9334) & (45.9984) & (11.4406) & (-3.5293) \\
(\mathrm{se}) & (0.0428) & (0.0175) & (0.0176) & (0.0008) \\
(\mathrm{fitness}) & R^2 = 0.9996; & \bar{R}^2 = 0.9995 & & \\
 & F^* = 37832.61; p = 0.0000 & & &
\end{array}$$

说明性例子（EViews软件报告）

Dependent Variable: LOG(Y)
Method: Least Squares
Date: Time: 15:24
Sample: 1947 2000
Included observations: 54

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.467711	0.042778	-10.93344	0.0000
LOG(X2)	0.804873	0.017498	45.99837	0.0000
LOG(X3)	0.201270	0.017593	11.44061	0.0000
X4	-0.002689	0.000762	-3.529269	0.0009

R-squared	0.999560	Mean dependent var	7.826093
Adjusted R-squared	0.999533	S.D. dependent var	0.552368
S.E. of regression	0.011934	Akaike info criterion	-5.947704
Sum squared resid	0.007121	Schwarz criterion	-5.800372
Log likelihood	164.5880	Hannan-Quinn criter.	-5.890884
F-statistic	37832.61	Durbin-Watson stat	1.289219
Prob(F-statistic)	0.000000		

说明性例子（R软件报告）

利用R软件给出更为详细的分析报告如下：

```
Call:
lm(formula = mod_mat, data = data_income)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018441 -0.010001  0.000337  0.007039  0.032578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4677111  0.0427780 -10.933 7.33e-15 ***
log(X2)      0.8048729  0.0174979  45.998  < 2e-16 ***
log(X3)      0.2012700  0.0175926  11.441 1.43e-15 ***
X4          -0.0026891  0.0007619  -3.529 0.000905 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01193 on 50 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic: 3.783e+04 on 3 and 50 DF,  p-value: < 2.2e-16
```

说明性例子（ANOVA）

变异来源	平方和符号SS	平方和计算公式	自由度df	均方和符号MSS	均方和计算公式
回归平方和	ESS	$\sum \hat{y}_i^2$	3	MSS_{ESS}	$\hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y} - n\bar{Y}^2$ =16.1637
残差平方和	RSS	$\sum e_i^2$	50	MSS_{RSS}	$\boldsymbol{y}\boldsymbol{y}' - \hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{y}$=0.0071
总平方和	TSS	$\sum y_i^2$	53	MSS_{TSS}	$\boldsymbol{y}' \boldsymbol{y} - n\bar{Y}^2$=16.1709

说明性例子（F检验）

根据方差分析表ANOVA和样本F统计量计算公式，可以得到：

$$F^* = \frac{ESS_U/df_{ESS_U}}{RSS_U/df_{RSS_U}} = \frac{\left(\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{Y}^2\right)/(k-1)}{\left(\boldsymbol{y}\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}\right)/(n-k)} = \frac{16.1637/3}{0.0071/50} = 37832.6142$$

得到显著性检验的判断结论。因为 $F^* = 37832.6142$ 大于 $F_{1-\alpha}(k-1, n-k) = F_{0.95}(3,50)$=2.7900，所以模型整体显著性的F检验结果**显著**。

说明性例子（其他模型）

此外，我们可以分别构建如下回归模型，并得到回归分析结果：

$$
\begin{array}{lll}
\widehat{log(X3)} = +1.65 & +0.99log(X2) \\
(\mathrm{t}) \quad (8.8575) & (42.0585) \\
(\mathrm{se}) \quad (0.1868) & (0.0235) \\
(\mathrm{fitness}) \quad R^2 = 0.9714; & \bar{R}^2 = 0.9709 \\
F^* = 1768.92; p = 0.0000
\end{array}
$$

$$
\begin{array}{lll}
\widehat{log(Y)} = -0.07 & +1.00log(X2) \\
(\mathrm{t}) \quad (-1.6471) & (178.3010) \\
(\mathrm{se}) \quad (0.0444) & (0.0056) \\
(\mathrm{fitness}) \quad R^2 = 0.9984; & \bar{R}^2 = 0.9983 \\
F^* = 31791.26; p = 0.0000
\end{array}
$$

$$
\begin{array}{lll}
\widehat{log(X2)} = -1.40 & +0.98log(X3) \\
(\mathrm{t}) \quad (-6.2996) & (42.0585) \\
(\mathrm{se}) \quad (0.2223) & (0.0234) \\
(\mathrm{fitness}) \quad R^2 = 0.9714; & \bar{R}^2 = 0.9709 \\
F^* = 1768.92; p = 0.0000
\end{array}
$$

$$
\begin{array}{lll}
\widehat{log(Y)} = -1.52 & +0.98log(X3) \\
(\mathrm{t}) \quad (-8.3771) & (51.5957) \\
(\mathrm{se}) \quad (0.1814) & (0.0191) \\
(\mathrm{fitness}) \quad R^2 = 0.9808; & \bar{R}^2 = 0.9805 \\
F^* = 2662.12; p = 0.0000
\end{array}
$$

3.1.3 多重共线性问题的诊断

多重共线性诊断：郎利案例

obs	Year	Y	X2	X3	X4	X5	X6	X7
1947	1947	60323	830	234289	2356	1590	107608	1
1948	1948	61122	885	259426	2325	1456	108632	2
1949	1949	60171	882	258054	3682	1616	109773	3
1950	1950	61187	895	284599	3351	1650	110929	4
1951	1951	63221	962	328975	2099	3099	112075	5
1952	1952	63639	981	346999	1932	3594	113270	6

Showing 1 to 6 of 16 entries

Previous 1 2 3 Next

美国就业情况的郎利数据(n=16)

- Y_i表示就业人数；
- X_{2i}表示消费价格指数（CPI）； X_{3i}表示国民生产总值（GNP，以百万美元计)；X_{4i}表示失业人数(以千人计)；X_{5i}表示军队中的人数；X_{6i}表示14 岁以上的非机构人口数；X_{7i}表示时间（$t = 1, 2, \cdots, n$）

相关性分析诊断法

利用样本数据绘制图形和图表，分析自变量之间是否存在明显相关关系。如果有，则表明模型很可能会产生**多重共线性**问题。

判断依据：

- **相关系数矩阵**发现高度线性相关（相关系数大于0.8）
- **散点图矩阵**发现高度线性相关的数据分布模式

相关性分析诊断法（案例）

郎利案例中，相关系数矩阵表计算如下：

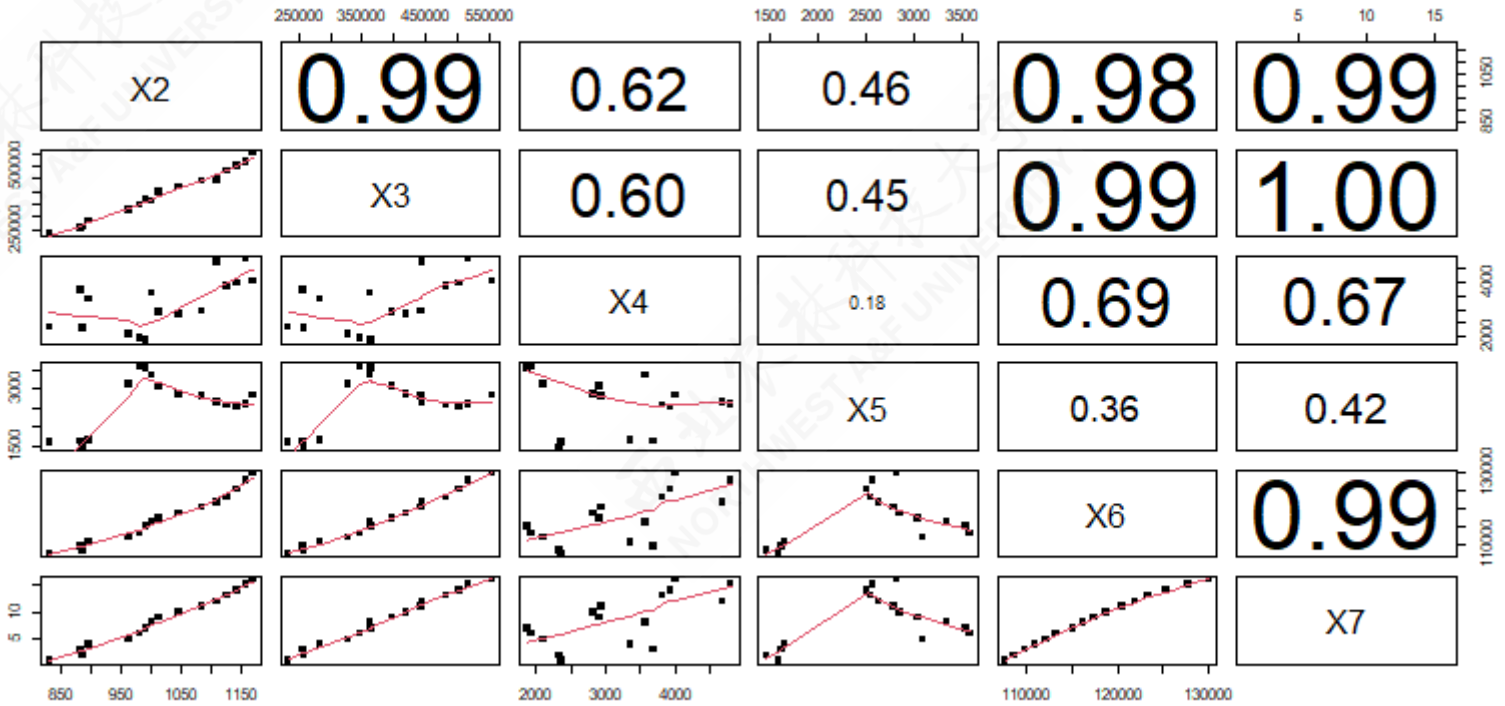
X2	X3	X4	X5	X6	X7
1.0000	0.9916	0.6206	0.4647	0.9792	0.9911
0.9916	1.0000	0.6043	0.4464	0.9911	0.9953
0.6206	0.6043	1.0000	-0.1774	0.6866	0.6683
0.4647	0.4464	-0.1774	1.0000	0.3644	0.4172
0.9792	0.9911	0.6866	0.3644	1.0000	0.9940
0.9911	0.9953	0.6683	0.4172	0.9940	1.0000

相关系数矩阵表

> Y_i表示就业人数； X_{2i}表示消费价格指数（CPI）； X_{3i}表示国民生产总值（GNP，以百万美元计); X_{4i}表示失业人数(以千人计) ; X_{5i}表示军队中的人数; X_{6i}表示14 岁以上的非机构人口数; X_{7i}表示时间（$t = 1, 2, \cdots, n$）

相关性分析诊断法（案例）

郎利案例中，散点矩阵图如下：



主回归方程诊断法

如果主回归方程分析报告结果异常，则可能认为存在多重共线性问题。

诊断依据：

- 主回归分析报告的 R^2值高（大于0.8）
- 分析报告 F^*检验显著
- 不显著的 t^*检验较多（多于回归系数个数的一半及以上）

主回归方程诊断法（案例简要报告）

郎利案例中，我们构建如下主回归模型：

$$\begin{aligned} Y = &+\hat{\beta}_1 &+\hat{\beta}_2 X2 + \hat{\beta}_3 X3 + \hat{\beta}_4 X4 \\ (\text{cont.}) &+\hat{\beta}_5 X5 + \hat{\beta}_6 X6 + \hat{\beta}_7 X7 + e_i \end{aligned}$$

主回归模型的回归分析结果如下：

$$\begin{array}{lllll} \hat{Y} = & +77270.12 & +1.51X2 & -0.04X3 & -2.02X4 \\ (\text{t}) & (3.4332) & (0.1774) & (-1.0695) & (-4.1364) \\ (\text{se}) & (22506.7070) & (8.4915) & (0.0335) & (0.4884) \\ (\text{cont.}) & -1.03X5 & -0.05X6 & +1829.15X7 & \\ (\text{t}) & (-4.8220) & (-0.2261) & (4.0159) & \\ (\text{se}) & (0.2143) & (0.2261) & (455.4785) & \\ (\text{fitness}) & R^2 = 0.9955; \bar{R}^2 = 0.9925 & & & \\ & F^* = 330.29; p = 0.0000 & & & \end{array}$$

主回归方程诊断法（案例R报告）

```
Call:
lm(formula = mods$main, data = data_longley)

Residuals:
    Min      1Q  Median      3Q     Max
-410.11 -157.67  -28.16  101.55  455.39

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.727e+04  2.251e+04   3.433 0.007470 **
X2           1.506e+00  8.491e+00   0.177 0.863141
X3          -3.582e-02  3.349e-02  -1.070 0.312681
X4          -2.020e+00  4.884e-01  -4.136 0.002535 **
X5          -1.033e+00  2.143e-01  -4.822 0.000944 ***
X6          -5.110e-02  2.261e-01  -0.226 0.826212
X7           1.829e+03  4.555e+02   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 304.9 on 9 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

辅助回归方程诊断法（判定系数）

辅助回归方程：为侦察多元回归模型是否存在多重共线性，构建关于自变量之间的一类线性回归方程。

正式地，对于样本回归模型，我们可以构建得到如下辅助回归方程：

$$
\begin{aligned}
X_{2i} &= \hat{\alpha}_1 + \cdots + \hat{\alpha}_j X_{ji} + \cdots + \hat{\alpha}_k X_{ki} + \epsilon_{2i} \\
&\vdots \\
X_{ji} &= \hat{\alpha}_1 + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_k X_{ki} + \epsilon_{ji} \\
&\vdots \\
X_{ki} &= \hat{\alpha}_1 + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_j X_{ji} + \cdots + \epsilon_{ki}
\end{aligned}
$$

辅助回归方程诊断法（判定系数）

对于样本回归模型，在普通最小二乘法下，我们可以证明：

$$S^2_{\hat{\beta}_j} = \frac{\hat{\sigma}^2}{(n-1)S^2_{X_j}} \cdot \frac{1}{1-R^2_j}$$

其中 R^2_j为辅助回归方程的判定系数。 $S^2_{X_j}$为变量 X_j的样本方差。

辅助回归方程诊断法（判定系数）

克莱因经验法则(Klein‘s rule of thumb)：

- 当来自一个辅助回归的 R_j^2**2**大于得自主回归中的 R^2值时，多重共线性才算是一个麻烦的问题。

辅助回归方程诊断法（辅助方程X2）

郎利案例中，辅助回归方程（X_2）的简要结果如下：

$$X2 = +\hat{\beta}_1 + \hat{\beta}_2 X3 + \hat{\beta}_3 X4 + \hat{\beta}_4 X5 + \hat{\beta}_5 X6 + \hat{\beta}_6 X7 + e_i$$

$$\begin{array}{lccccccc}
\widehat{X2} = & +2044.58 & +0.00X3 & +0.03X4 & +0.01X5 & -0.02X6 & -9.99X7 \\
(\mathrm{t}) & (3.8333) & (2.7006) & (2.1098) & (1.1770) & (-2.7720) & (-0.5996) \\
(\mathrm{se}) & (533.3698) & (0.0009) & (0.0151) & (0.0075) & (0.0063) & (16.6654)
\end{array}$$

$$(\text{fitness}) R^2 = 0.9926; \bar{R}^2 = 0.9889$$

$$F^* = 269.06; p = 0.0000$$

辅助回归方程诊断法（辅助方程X3）

郎利案例中，辅助回归方程（X_3）的简要结果如下：

$$X3 = +\hat{\beta}_1 + \hat{\beta}_2 X2 + \hat{\beta}_3 X4 + \hat{\beta}_4 X5 + \hat{\beta}_5 X6 + \hat{\beta}_6 X7 + e_i$$

$$\begin{array}{lccccccc}
\widehat{X3} = & -480986.04 & +164.66X2 & -13.79X4 & -3.00X5 & +5.62X6 & +10902.88X7 \\
(\mathrm{t}) & (-3.2408) & (2.7006) & (-9.1921) & (-1.6774) & (4.7649) & (4.2411) \\
(\mathrm{se}) & (148413.7872) & (60.9699) & (1.5002) & (1.7873) & (1.1804) & (2570.7562)
\end{array}$$

$$(\text{fitness}) R^2 = 0.9994;\ \bar{R}^2 = 0.9992$$

$$F^* = 3575.03; p = 0.0000$$

辅助回归方程诊断法（辅助方程X4）

郎利案例中，辅助回归方程（X_4）的简要结果如下：

$$X4 = +\hat{\beta}_1 + \hat{\beta}_2 X2 + \hat{\beta}_3 X3 + \hat{\beta}_4 X5 + \hat{\beta}_5 X6 + \hat{\beta}_6 X7 + e_i$$

$$\begin{array}{lccccccc}
\widehat{X4} = & -28518.24 & +9.65X2 & -0.06X3 & -0.27X5 & +0.35X6 & +768.55X7 \\
(\mathrm{t}) & (-2.4914) & (2.1098) & (-9.1921) & (-2.4895) & (3.6779) & (4.6007) \\
(\mathrm{se}) & (11446.8866) & (4.5736) & (0.0071) & (0.1090) & (0.0954) & (167.0507)
\end{array}$$

$$(\text{fitness}) R^2 = 0.9703; \bar{R}^2 = 0.9554$$

$$F^* = 65.24;\ p = 0.0000$$

辅助回归方程诊断法（辅助方程X5）

郎利案例中，辅助回归方程（X_5）的简要结果如下：

$$X5 = +\hat{\beta}_1 + \hat{\beta}_2 X2 + \hat{\beta}_3 X3 + \hat{\beta}_4 X4 + \hat{\beta}_5 X6 + \hat{\beta}_6 X7 + e_i$$

$$\begin{array}{lllllll}
\widehat{X5} = & -11881.24 & +13.82X2 & -0.07X3 & -1.41X4 & +0.20X6 & +1167.78X7 \\
(\mathrm{t}) & (-0.3600) & (1.1770) & (-1.6774) & (-2.4895) & (0.6084) & (2.0791) \\
(\mathrm{se}) & (33002.4231) & (11.7447) & (0.0437) & (0.5663) & (0.3276) & (561.6770)
\end{array}$$

$$(\text{fitness}) R^2 = 0.7214; \bar{R}^2 = 0.5820$$

$$F^* = 5.18; \quad p = 0.0133$$

辅助回归方程诊断法（辅助方程X6）

郎利案例中，辅助回归方程（X_6）的简要结果如下：

$$X6 = +\hat{\beta}_1 + \hat{\beta}_2 X2 + \hat{\beta}_3 X3 + \hat{\beta}_4 X4 + \hat{\beta}_5 X5 + \hat{\beta}_6 X7 + e_i$$

$$\begin{array}{lccccccc}
\widehat{X6} = & +95694.37 & -24.76X2 & +0.12X3 & +1.64X4 & +0.18X5 & -782.04X7 \\
(\mathrm{t}) & (11.0221) & (-2.7720) & (4.7649) & (3.6779) & (0.6084) & (-1.3319) \\
(\mathrm{se}) & (8682.0335) & (8.9319) & (0.0259) & (0.4454) & (0.2943) & (587.1614)
\end{array}$$

$$(\text{fitness}) R^2 = 0.9975; \bar{R}^2 = 0.9962$$

$$F^* = 796.30; p = 0.0000$$

辅助回归方程诊断法（辅助方程X7）

郎利案例中，辅助回归方程（X_7）的简要结果如下：

$$X7 = +\hat{\beta}_1 + \hat{\beta}_2 X2 + \hat{\beta}_3 X3 + \hat{\beta}_4 X4 + \hat{\beta}_5 X5 + \hat{\beta}_6 X6 + e_i$$

$$\begin{array}{lllllll}
\widehat{X7} = & +8.31 & -0.00X2 & +0.00X3 & +0.00X4 & +0.00X5 & -0.00X6 \\
(\mathrm{t}) & (0.5392) & (-0.5996) & (4.2411) & (4.6007) & (2.0791) & (-1.3319) \\
(\mathrm{se}) & (15.4036) & (0.0058) & (0.0000) & (0.0002) & (0.0001) & (0.0001)
\end{array}$$

$$(\text{fitness}) R^2 = 0.9987;\ \bar{R}^2 = 0.9980$$

$$F^* = 1515.96; p = 0.0000$$

辅助回归方程诊断法（判定系数比较）

主回归模型的判定系数为0.9955，辅助回归的判定系数见下表：

辅助回归方程的判定系数

models	R2	判定系数诊断结论
X2	0.9926	...
X3	0.9994	严重问题
X4	0.9703	...
X5	0.7214	...
X6	0.9975	严重问题
X7	0.9987	严重问题

辅助回归方程诊断法（VIF）

辅助回归方程方差膨胀因子的理论计算公式为

$$VIF_j = \frac{1}{1-R_j^2}, (j = 1, 2, \cdots, k-1)$$

诊断依据：

- 辅助回归方程的方差膨胀因子中如果 $VIF_j \in [10, 100]$表明中度多重共线性；
- 如果 $VIF_j \geq 100$表明严重多重共线性

辅助回归方程诊断法（VIF）

郎利案例中，辅助回归方程的VIF诊断结果为：

方差膨胀因子分析

models	R2	VIF	VIF诊断结论
X2	0.9926	135.5324	非常严重
X3	0.9994	1788.5135	非常严重
X4	0.9703	33.6189	比较严重
X5	0.7214	3.5889	...
X6	0.9975	399.1510	非常严重
X7	0.9987	758.9806	非常严重

根据计算结果汇总，可以认为主模型存在较严重的多重共线性问题。其中VIF值大于100的系数就包括X2、X3、X6、X7。

辅助回归方程诊断法（VIF）

EViews软件给出的VIF结果：

Table: TAB_VIF Workfile: LONGLEY::employee\

View | Proc | Object | Print | Name | Edit+/- | CellFmt | Grid+/- | Title | Comments+/-

	A	B	C	D
1	Variance Inflation Factors			
2	Date: Time:			
3	Sample: 1 16			
4	Included observations: 16			
5				
6	Variable	Coefficient Variance	Uncentered VIF	Centered VIF
7				
8				
9	C	5.07E+08	87208.72	NA
10	X1	72.10545	12970.23	135.5324
11	X2	0.001122	30814.07	1788.513
12	X3	0.238534	452.3831	33.61889
13	X4	0.045913	57.29873	3.588930
14	X5	0.051109	121723.5	399.1510
15	X6	207460.7	3339.515	758.9806
16				
17				

辅助回归方程诊断法（TOL）

辅助回归方程容忍度的理论计算公式为：

$$TOL_j = 1 - R_j^2 = \frac{1}{VIF_j}, (j = 1, 2, \cdots, k-1)$$

诊断依据：

- 如果辅助回归方程的容忍度 $TOL_j \in [0.01, 0.1]$表明中度多重共线性；
- 如果辅助回归方程的容忍度 $TOL_j \leq 0.01$表明存在严重的多重共线性

辅助回归方程诊断法（TOL）

郎利案例中，辅助回归方程的TOL诊断结果为：

容忍度分析结果

models	R2	VIF	TOL	TOL结论
X2	0.9926	135.5324	0.0074	非常严重
X3	0.9994	1788.5135	0.0006	非常严重
X4	0.9703	33.6189	0.0297	...
X5	0.7214	3.5889	0.2786	比较严重
X6	0.9975	399.1510	0.0025	非常严重
X7	0.9987	758.9806	0.0013	非常严重

回归系数方差分解诊断法(CVD)

回归系数方差分解法（Coefficient Variance Decomposition）：通过计算特征值(eigenvalues)，进而得到病态数(K_j)和方差分解比率 VDP_j，最后做出多重共线性的诊断结论：

对于k变量回归模型：

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e} \qquad \text{(SRM)}$$

可以得到OLS下参数估计的方差协方差矩阵：

$$var - cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}'$$

其中，$\mathbf{D}$是含有矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$的**特征值**（eigenvalues） E_m（ $m \in 1, 2, \cdots, k$）的一个对角矩阵，而 $\mathbf{V}$是由相应**特征向量**构成的一个矩阵。

回归系数方差分解诊断法(CVD)

病态数（condition number）采用微分的方法，考察引入一个变量对（多重共线性）模型结果恶化情形出现的相对改变数，一般记为K，并正式定义为：

$$K_j = \frac{min(E_m)}{E_j}$$

方差分解比率（variance-decomposition proportion），一般记为 VDP_{ji}，并正式定义为：

$$\phi_{ij} = \frac{v_{ij}^2}{E_j}; \quad VDP_{ji} = \frac{\phi_{ij}}{\phi_i}$$

其中 v_{ij}为矩阵 $\mathbf{V}$的第 (i,j)个元素。

回归系数方差分解诊断法(CVD)

病态数和方差分解比率：用来诊断多元线性回归模型的多重共线性问题严重程度的指标。诊断依据为：

- 若发现至少一个病态数 $K_j \leq 0.001$，则表明存在严重多重共线性；
- 观察病态数最小时所对应的方差分解比率，如果有多个斜率系数的 $VDP_j \geq 0.5$，则表明存在严重的多重共线性
- 系数方差分解诊断方法由Belsley, Kuh and Welsch (BKW) 2004提出，具体细节可以参考Eviews帮助文档，网址为：
  http://www.eviews.com/help/helpintro.html#page/content/testing-Coefficient_Diagnostics.html]。
- Eviews软件中病态数的计算是基于矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$，而不是基于矩阵 $\mathbf{X}$

回归系数方差分解诊断法(CVD)

郎利案例中，主回归模型系数方差分解的多重共线性EViews诊断结果为：

Coefficient Variance Decomposition
Date: Time:
Sample: 1 16
Included observations: 16

Eigenvalues	5.07E+08	201598.2	28.96906	0.037097	0.008008	1.32E-05	3.36E-08
Condition	6.63E-17	1.67E-13	1.16E-09	9.05E-07	4.19E-06	0.002542	1.000000

Variance Decomposition Proportions

	Associated Eigenvalue						
Variable	1	2	3	4	5	6	7
C	1.000000	4.61E-09	4.72E-15	5.83E-22	1.50E-21	3.04E-24	3.64E-28
X1	0.595056	0.003296	0.401648	1.07E-07	2.73E-09	6.97E-12	2.77E-15
X2	0.512335	0.477718	0.008255	0.000714	3.46E-05	0.000915	2.76E-05
X3	0.383038	0.533542	0.003177	0.059528	0.020715	5.86E-09	8.51E-12
X4	0.012803	0.289451	0.132728	0.498229	0.066790	5.18E-08	2.87E-11
X5	0.923977	0.052811	0.022710	0.000263	2.17E-08	0.000238	5.10E-08
X6	0.028271	0.971729	1.67E-10	8.26E-14	8.55E-15	9.26E-19	8.13E-23

Eigenvectors

	Associated Eigenvalue						
Variable	1	2	3	4	5	6	7
C	0.999994	-0.003403	0.000287	2.82E-06	-9.73E-06	-1.08E-05	-2.34E-06
X1	-0.000291	-0.001086	0.999861	-0.014433	-0.004961	-0.006168	-0.002438
X2	1.07E-06	5.16E-05	-0.000565	0.004645	-0.002202	0.278758	-0.960347
X3	1.34E-05	0.000795	-0.005115	-0.618683	0.785518	-0.010288	-0.007777
X4	1.08E-06	0.000257	-0.014504	-0.785263	-0.618815	-0.013425	-0.006268
X5	-9.66E-06	-0.000116	-0.006330	0.019049	-0.000372	-0.960193	-0.278617
X6	-0.003403	-0.999993	-0.001094	-0.000679	0.000471	0.000121	-2.24E-05

3.1.4 多重共线性问题的矫正

删除变量法

一旦发现模型存在比较严重的多重共线性问题，则需要对模型进行修正处理，具体方法可参考：

简单剔除变量法：

- 依据经济学和实践经验观察，进行变量甄选或变量变换。利用先验信息（成为研究领域的专家！）酌情删除特定变量，减弱模型的多重共线性问题。那怎样才能获得先验信息呢？它往往源自经验研究工作或者有关基础理论。

> 怎样获得先验信息呢?它可以经验研究工作或者有关基础理论。例如，在柯布-道格拉斯生产函数中，如果人们预期规模报酬不变成立，则有 $\beta_2 + \beta_3 = 1$。如果劳动和资本之间存在共线性，这一变换就减轻或消除了共线性问题。

- 变量变换法，进行变量处理。具体又包括差分变换法、比率变换法

删除变量法（案例）

郎利案例中，模型各变量的含义：

- Y_i表示就业人数；
- X_{2i}表示消费价格指数（CPI）；X_{3i}表示国民生产总值（GNP，以百万美元计)；X_{4i}表示失业人数(以千人计)；X_{5i}表示军队中的人数；X_{6i}表示14岁以上的非机构人口数；X_{7i}表示时间（$t = 1, 2, \cdots, n$）。

简单删除的依据：

- 不用名义GNP，改用真实GNP。将名义GNP（X3）除以价格指数CPI（X2），得到**实际GNP**（X3/X2）。
- 留下X5（军队中的人数），去掉X6（14 岁以上非机构人口数）。因为X6（14 岁以上非机构人口数）随时间（X7）不断增长，它与时间变量X7高度相。
- 去掉变量X4（失业人数）。可能失业率是劳动力市场状况的一个更好的度量指标，但我们没有这方面的数据，而失业人数X4也没有充分的理由包括进来。

删除变量法（案例R报告）

运用删除变量法，调整后的回归模型为：

$$Y = +\hat{\beta}_1 + \hat{\beta}_2 I(X3/X2) + \hat{\beta}_3 X5 + \hat{\beta}_4 X6 + e_i$$

回归结果为：

$$\begin{array}{lllll}
\widehat{Y} = & +65720.37 & +97.36I(X3/X2) & -0.69X5 & -0.30X6 \\
(\mathrm{t}) & (6.1856) & (5.4347) & (-2.1350) & (-2.1130) \\
(\mathrm{se}) & (10624.8077) & (17.9155) & (0.3222) & (0.1418) \\
(\mathrm{fitness}) & R^2 = 0.9814; \bar{R}^2 = 0.9768 & & & \\
 & F^* = 211.10; p = 0.0000 & & &
\end{array}$$

变量变换法（一阶差分法）

面对严重的共线性，最简单的方法就是去掉某些变量，但剔除变量会导致设定误差。实际中需要权衡利弊。

一阶差分法(first difference form)巧妙删除变量：模型中两个解释变量 $X_{k,i}$和 $X_{w,i}$可能导致高度多重共线性，但是分别对二者进行一阶差分，再进行回归建模，新模型可能的多重共线性问题很可能大大缓解！具体变换如下:

$$
\begin{aligned}
Y_t &= \beta_1 + \beta_2 X_{2,t} + \beta_3 X_{3,t} + u_t && \text{原模型} \\
Y_{t-1} &= \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} && \text{滞后1阶变量模型} \\
Y_t - Y_{t-1} &= \beta_2 (X_{2,t} - X_{2,t-1}) + \beta_3 (X_{3,t} - X_{3,t-1}) + (u_t - u_{t-1}) && \text{一阶差分模型} \\
Y_t^* &= \beta_2 X_{2,t}^* + \beta_3 X_{3,t}^* + v_t && \text{精简化模型}
\end{aligned}
$$

> 需要注意的是，“按下葫芦浮起瓢”，治疗比疾病更糟糕？
>
> 差分变换 Y_{t-1}减少了自由度；同时 $v_t = (u_t - u_{t-1})$可能带来异方差问题。

变量变换法（比率变换法）

比率变换法(ratio transformation)巧妙删除变量：模型中两个解释变量 $X_{k,i}$和 $X_{w,i}$ 可能导致高度多重共线性，如果可以用其中的一个变量同时对模型其他变量进行比率变换，而且如果变换后的所有变量还能具有经济学含义，那么理论上将至少消掉一个回归元，从而大大缓解甚至消除多重共线性问题！

以**消费支出决定**为例：

Y_t 为以真实价格表示的消费支出， $X_{2,t}$表示GDP， $X_{3,t}$表示总人口。

$$Y_t = \beta_1 + \beta_2 X_{2,t} + \beta_3 X_{3,t} + u_t \qquad \text{原模型}$$

$$\frac{Y_t}{X_{3,t}} = \frac{\beta_1}{X_{3,t}} + \beta_2 \frac{X_{2,t}}{X_{3,t}} + \frac{u_t}{X_{3,t}} \qquad \text{比率变换模型}$$

$$Y_t^* = \beta_1^* + \beta_2^* X_{2,t}^* + v_t \qquad \text{精简化模型}$$

> 同样需要注意的是，“按下葫芦浮起瓢”，治疗比疾病更糟糕？
>
> $v_t = \frac{u_t}{X_{3,t}}$ 可能带来异方差问题。

逐步回归法

逐步最小二乘回归法（Stepwise Least Squares Regression）通过多个统计标准，可以自动判断模型该引入还是删除某些自变量X。这些统计标准主要包括分析引入新变量对回归平方和ESS的贡献大小，及F检验等。

- **前向逐步回归法**（Stepwise-Forwards），是从一个简化模型（很少X变量）开始，再逐步引入新的X变量，直至达到某个统计标准（主要是p值标准）
- **后向逐步回归法**（Stepwise-Backwards），是从一个完全模型（全部X变量）开始，对模型逐步删除某些X变量，直至剩余变量都达到某个统计标准（主要是p值标准）

逐步回归法

逐步最小二乘回归法一般采用如下标准来自动筛选变量：

- p值判别法： $p \in [0.1, 0.05)$(比较显著)； $p \in [0.05, 0.01)$(比较显著)； $p \leq 0.01$(极其显著)
- t^*值判别法：2t法则
- **AIC**信息准则：越小越好

R统计软件下采用**后向逐步回归法**（Stepwise-Backwards）自动删除变量，最终回归结果报告见下页-->

后向逐步回归法分析结果

```
Call:
lm(formula = Y ~ X3 + X4 + X5 + X7, data = data_longley)

Residuals:
    Min      1Q  Median      3Q     Max
-421.65 -124.57  -24.16   83.69  452.68

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.417e+04  4.252e+03  17.445 2.30e-09 ***
X3          -4.019e-02  1.647e-02  -2.440 0.032833 *
X4          -2.088e+00  2.900e-01  -7.202 1.75e-05 ***
X5          -1.015e+00  1.837e-01  -5.522 0.000180 ***
X7           1.887e+03  3.828e+02   4.931 0.000449 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 279.4 on 11 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.9937
F-statistic: 589.8 on 4 and 11 DF,  p-value: 9.5e-13
```

*后向逐步回归的具体细节，请参看附录7-A给出的详细报告

逐步回归法（案例报告）

经过**后向逐步回归法**的标准筛选变量，得到的新模型为：

$$Y = +\hat{\beta}_1 + \hat{\beta}_2 X3 + \hat{\beta}_3 X4 + \hat{\beta}_4 X5 + \hat{\beta}_5 X7 + e_i$$

逐步回归法矫正后回归模型的简要回归报告如下：

$$\begin{array}{lllllll}
\widehat{Y} = & & +74169.53 & -0.04X3 & -2.09X4 & -1.01X5 & +1887.41X7 \\
(\mathrm{t}) & & (17.4451) & (-2.4398) & (-7.2021) & (-5.5223) & (4.9310) \\
(\mathrm{se}) & & (4251.5849) & (0.0165) & (0.2900) & (0.1837) & (382.7665) \\
(\mathrm{fitness}) & R^2 = 0.9954; & \bar{R}^2 = 0.9937 & & & & \\
& F^* = 589.76; & p = 0.0000 & & & &
\end{array}$$

补充新数据法（有时候有用！）

由于多重共线性是一个样本特性，故有可能在关于同样变量的另一样本中共线性没有第一个样本那么严重

在三变量回归中，有：

$$\operatorname{var}\left(\hat{\beta}_2\right)=\frac{\sigma^2}{\sum x_{2i}^2\left(1-r_{23}^2\right)}$$

- 随着样本增加，$\sum x_{2i}^2$一般地都会增加（为什么？）
- 因此，对任何给定的 r_{23}^2，$\hat{\beta}_2$的方差将会减小，从而减低 $\hat{\beta}_2$的标准误，使我们能够更准确地估计 β_2。

多项式回归法（原理）

多项式回归模型中离差形式或正交多项式(orthogonal polynomials)可以降低共线性的影响。

- 多项式回归模型的一个特点是解释变量以不同的幂出现，从而容易导致多重共线性
- 处理办法：离差形式或正交多项式(orthogonal polynomials)方法

相关资料可以参考宾夕法尼亚大学的Reducing Structural Multicollinearity

多项式回归法（案例数据）

obs	Y	X	X^2	y	x	x^2
1	881	34.6	1,197.16	-676.63	-16.04	257.17
2	1290	45	2,025.00	-267.63	-5.64	31.77
3	2147	62.3	3,881.29	589.37	11.66	136.03
4	1909	58.9	3,469.21	351.37	8.26	68.28
5	1282	42.5	1,806.25	-275.63	-8.14	66.21
6	1530	44.3	1,962.49	-27.63	-6.34	40.15
7	2067	67.9	4,610.41	509.37	17.26	298.02
8	1982	58.5	3,422.25	424.37	7.86	61.83

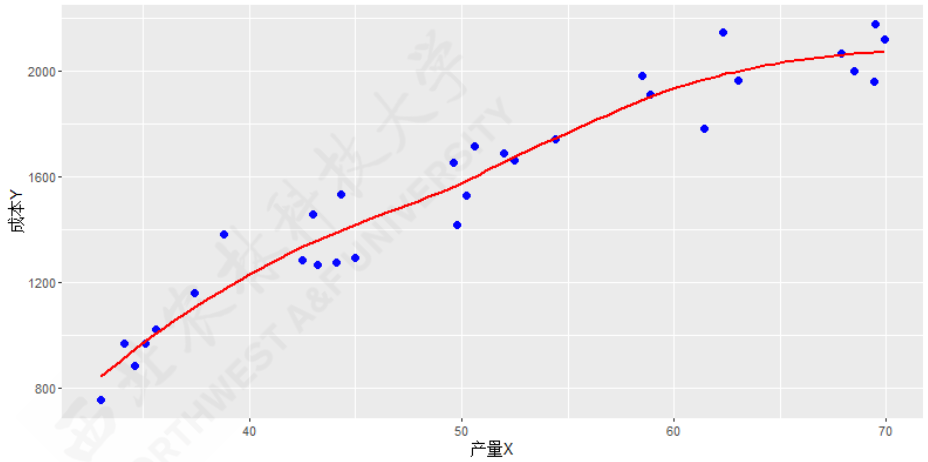
Showing 1 to 8 of 30 entries

Previous 1 2 3 4 Next

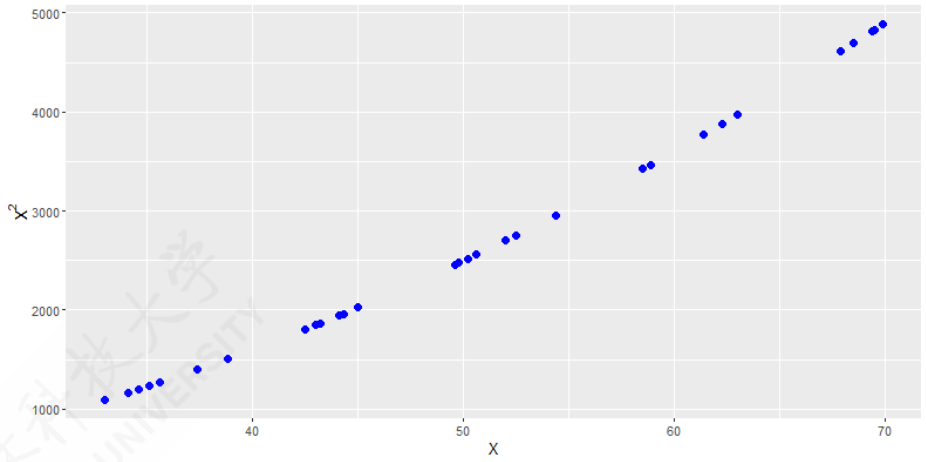
成本与产量数据

其中：Y表示成本（`Cost`）；X表示产品产量（`Output`）；X表示产品产量的平方；y表示成本的离差；x表示产量的离差；x^2表示产量离差的平方。

多项式回归法（案例绘图1）



产量和成本的散点图



产量和产量平方的散点图

实际上产量 X和产量平方 X^2之间的相关系数为： $r_{(X,X^2)} = 0.9949846$。人们可能执意构建如下的**二次多项式抛物线模型1**：

$$Y = + \hat{\beta}_1 + \hat{\beta}_2 X + \hat{\beta}_3 I(X^2) + e_i$$

上述二次多项式抛物线模型1回归分析结果如下：

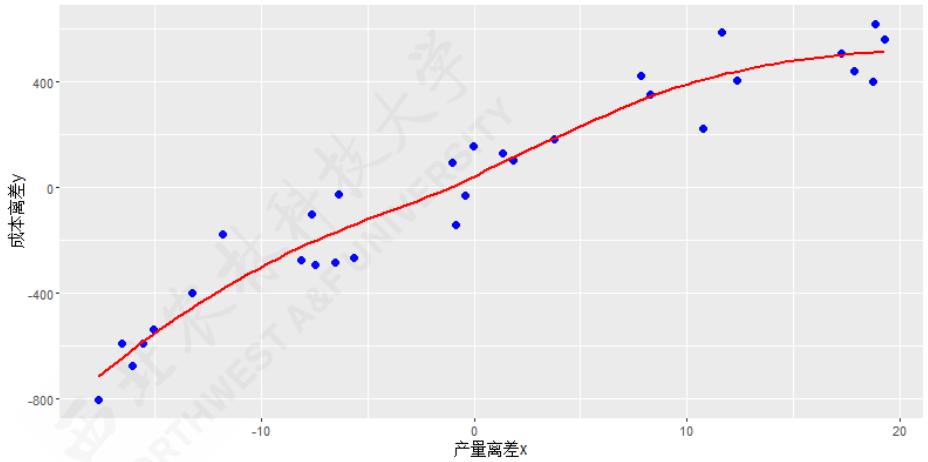
$$
\begin{aligned}
\widehat{Y} = \quad & -1464.40 \quad +88.31X \quad -0.54I(X^2) \\
(\mathrm{t}) \quad & (-3.5596) \quad (5.3606) \quad (-3.3900) \\
(\mathrm{se}) \quad & (411.4012) \quad (16.4735) \quad (0.1582) \\
(\mathrm{fitness}) & R^2 = 0.9377; \bar{R}^2 = 0.9331 \\
& F^* = 203.16; p = 0.0000
\end{aligned}
$$

对以上模型进行方差膨胀因子（VIF）分析，结果为：

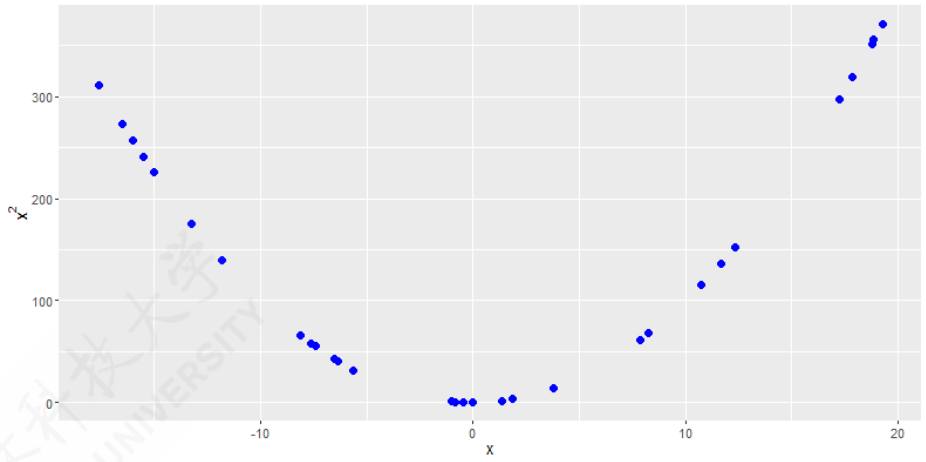
自变量方差膨胀因子分析结果

terms	VIF
X	99.9426
X^2	99.9426

多项式回归法（案例绘图2）



产量离差和成本离差的散点图



产量离差和产量离差平方的散点图

此时产量离差 x和产量离差平方 x^2之间的相关系数为： $r_{(x,x^2)} = 0.2195179$。我们再构建如下的**二次多项式抛物线模型2**：

$$y = + \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 I(x^2) + e_i$$

二次多项式抛物线模型2的回归分析结果为：

$$
\begin{array}{lllll}
\hat{y} = & +74.56 & +34.00x & -0.54I(x^2) \\
(\mathrm{t}) & (2.5406) & (20.1297) & (-3.3900) \\
(\mathrm{se}) & (29.3486) & (1.6890) & (0.1582) \\
(\mathrm{fitness}) & R^2 = 0.9377; \bar{R}^2 = 0.9331 & & \\
 & F^* = 203.16; p = 0.0000 & &
\end{array}
$$

对该模型进行方差膨胀因子（VIF）分析，结果为：

terms	VIF
x	1.05
x^2	1.05

岭回归法（基本原理）

岭回归法(ridge regression)，也被称为脊回归，常被用来"解决"多重共线性问题。可惜这些技术都要利用矩阵代数才便于讨论*。

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}\in\mathbb{R}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \lambda\|\boldsymbol{B}\|_2^2$$

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

其中：$\|\boldsymbol{B}\|_2 = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_k^2}$表示向量范数（vector norm）运算

[*] 岭回归的具体细节，请参看附录材料1 Ridge Regression for Better Usage、材料2 和材料3 Ridge Regression

岭回归法（R软件分析）

利用R软件进行岭回归分析，结果报告如下：

```
Coefficients:
             Estimate Scaled estimate Std. Error (scaled)
(Intercept) 40423.1812              NA                 NA
X2              8.6257       3605.1454           768.4962
X3              0.0115       4432.7298           330.6156
X4             -0.8884      -3215.1838           675.8977
X5             -0.3554       -957.9782           593.0804
X6              0.1136       3060.7725           638.2902
X7            244.8853       4515.4627           398.9649
            t value (scaled)  Pr(>|t|)
(Intercept)               NA        NA
X2                      4.69 0.0000027 ***
X3                     13.41   < 2e-16 ***
X4                      4.76 0.0000020 ***
X5                      1.62      0.11
X6                      4.80 0.0000016 ***
X7                     11.32   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.0481, chosen automatically, computed using 2 PCs

Degrees of freedom: model 3.05 , variance 2.61 , residual 3.49
```

主成分分析法（原理）

主成分分析法（Principal components regression，PCR），其核心思想是**降维**（Dimensionality reduction），也即通过一定的统计方法把**原来**个数较多的自变量，转换为个数相对较少的**新的**自变量，然后再利用较少的自变量进行回归，从而达到消除多重共线性的目的。

- 先根据主成分分析确定主成分个数（看累积解释百分比）
- 再用**主成分变量**进行回归分析

技术细节可以参看Performing Principal Components Regression (PCR) in R，以及Principal Component Regression in R

主成分分析法（案例）

郎利案例中，R软件下主成分分析结果为：

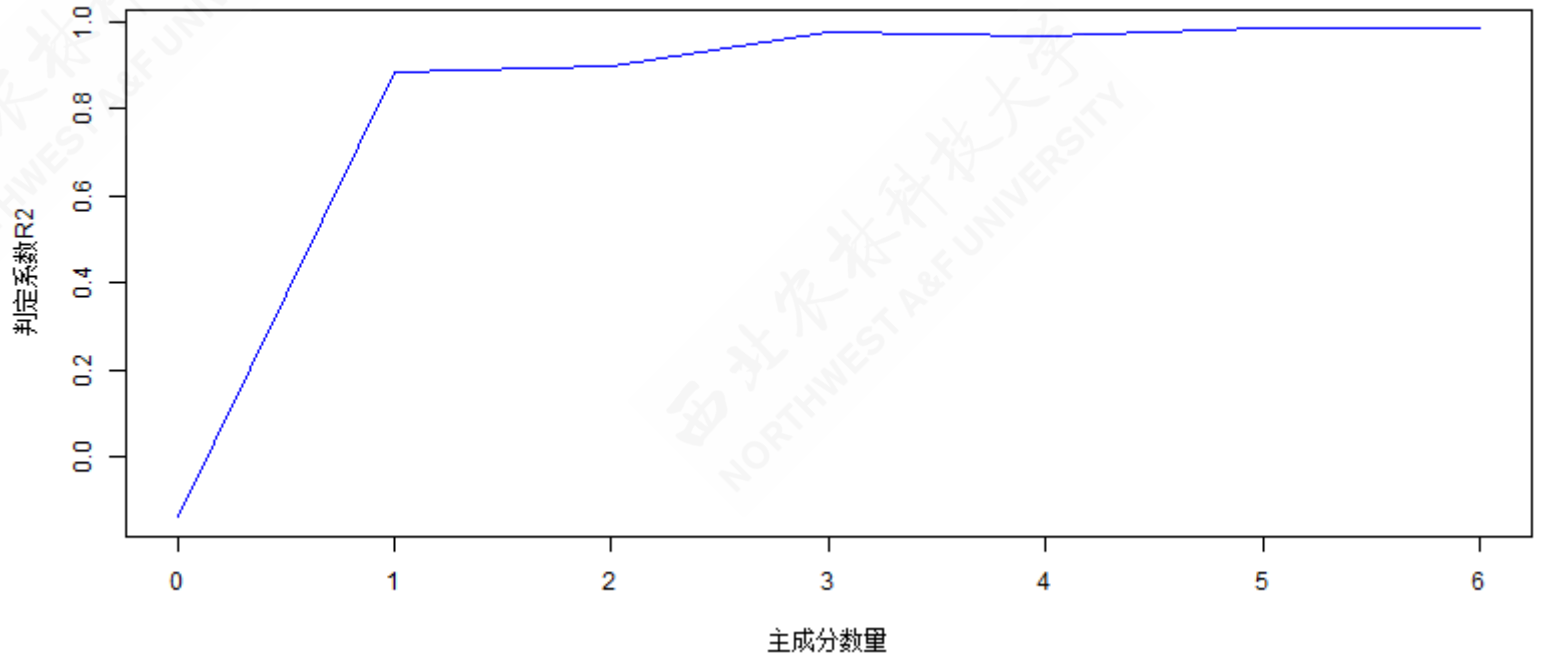
```
Data:     X dimension: 16 6
    Y dimension: 16 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV            3627     1152     1085    508.1    616.9    432.0    424.5
adjCV         3627     1141     1073    502.1    607.1    423.8    414.7

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X    76.72    96.31     99.7    99.95    99.99   100.00
Y    91.43    92.89     98.6    98.61    99.40    99.55
```

主成分分析法（案例：碎石图）

主成分分析的判定系数与主成分数量的变化情况如下：



从上图可以看出，主成分个数为3个时，判定系数基本就达到很高的水平了。

主成分分析法（案例：载荷矩阵）

此时，我们可以看看主成分的载荷矩阵：

主成分分析的载荷矩阵

vars	PC1	PC2	PC3	PC4	PC5	PC6
X2	0.4618	-0.0578	0.1491	0.7929	-0.3379	0.1352
X3	0.4615	-0.0532	0.2777	-0.1216	0.1496	-0.8185
X4	0.3213	0.5955	-0.7283	0.0076	-0.0092	-0.1075
X5	0.2015	-0.7982	-0.5616	-0.0773	-0.0243	-0.0180
X6	0.4623	0.0455	0.1960	-0.5897	-0.5486	0.3116
X7	0.4649	-0.0006	0.1281	-0.0523	0.7495	0.4504

请注意前三个主成分`PC1`、`PC2`、`PC3`，以及它们在自变量上的载荷分布。

主成分分析法（案例：因子得分及新数据）

我们把因变量进行标准化变换 $Y_i^* = \frac{Y_i - \bar{Y}}{S_{Y_i}}$，同时利用以上主成分分析的因子得分，可以得到**主成分降维后的新数据**：

Y.std	PC1	PC2	PC3
-1.4220	-3.4789	0.7515	0.3079
-1.1945	-3.0105	0.8490	0.6422
-1.4653	-2.3433	1.5400	-0.4934
-1.1760	-2.0939	1.2763	-0.1113
-0.5968	-1.4382	-1.2358	-0.0291

Showing 1 to 5 of 16 entries

Previous 1 2 3 4 Next

因子得分及新数据

主成分分析法（案例：新模型回归）

利用以上降维后的新数据，可以构造如下新的模型：

$$Y.std = +\hat{\beta}_1 + \hat{\beta}_2 PC1 + \hat{\beta}_3 PC2 + \hat{\beta}_4 PC3 + e_i$$

$$\begin{array}{lllll}
\widehat{Y.std} = -0.00 & & +0.45PC1 & -0.11PC2 & +0.53PC3 \\
(\mathrm{t}) & (-0.0000) & (27.9607) & (-3.5371) & (6.9867) \\
(\mathrm{se}) & (0.0331) & (0.0159) & (0.0315) & (0.0758) \\
(\text{fitness}) & R^2 = 0.9860; \bar{R}^2 = 0.9825 & & & \\
 & F^* = 281.04; p = 0.0000 & & &
\end{array}$$

对该模型进行方差膨胀因子（VIF）分析，VIF都比较小，多重共线性问题得到解决。

新模型的方差膨胀因子分析结果

terms	VIF
PC1	1
PC2	1
PC3	1

多重共线性一定是坏事吗?

如果回归分析的唯一目的是预测或预报，则多重共线性就不是一个严重的问题。因为 R^2值越高，预测越准。

无为而治：多重共线性是普遍存在的，它并不是OLS 或其他一般性统计方法引起的问题

本章结束