实用回归实验报告

韩硕
2018302060281
张铭清
2018302091006
马铸秋
2018302100209

班 级: 18 级统计

2020 年 12 月 10 日

1 第六题

1.1 题目重述

对某地 18 年某种消费品销售数据 Y 进行调查，并考虑有关的三个因素：X_1-居民可支配收入；X_2-该类消费品的价格指数；X_3-其它消费品平均价格指数。对该地区 18 年某种消费品销售数据试用 RMS_q、C_p 和 AIC 准则，建立子集回归模型。

序号	Y	X_1	X_2	X_3
1	7.8	81.2	85.0	87.0
2	8.4	82.9	92.0	94.0
3	8.7	83.2	91.5	95.0
4	9.0	85.9	92.9	95.5
5	9.6	88.0	93.0	96.0
6	10.3	99.0	96.0	97.0
7	10.6	102.0	95.0	97.5
8	10.9	105.3	95.6	98.0
9	11.3	117.7	98.9	101.2
10	12.3	126.4	101.5	102.5
11	13.5	131.2	102.0	104.0
12	14.2	148.0	105.0	105.9
13	14.9	153.0	106.0	109.5
14	15.9	161.0	109.0	111.0
15	18.5	170.0	112.0	110.0
16	19.5	174.0	112.5	112.0
17	19.5	185.0	113.0	112.3
18	20.5	189.0	114.0	113.0

图 1: 题 6 数据表

1.2 解题思路

在应用回归分析去处理实际问题时，回归自变量的选择是首先要解决的重要问题。通常，在做回归分析时，人们根据问题本身的专业理论及有关经验，常常把各种与因变量有关或可能有关的自变量引进回归模型，这样的话不但计算量大而且估计和预测的精读都会下降。因此，在应用回归分析时，对进入模型的自变量作精心的选择是十分必要的。

- **全模型**

  这里为 y 观测向量，X 为 n*p 的列满秩矩阵，我们约定 X 的第一列元素皆为 1

  假设我们根据某些自变量选择准则，剔除了全模型中一些对因变量影响较小的自变量，不妨假设剔除了后 p-q 个自变量 $X_q, X_{q+1}...X_{p-1}$ 记

$$y = X\beta + e, \quad E(e) = 0, \quad Cov(e) = \sigma^2 I$$

X'=$\begin{pmatrix} X_q \\ X_t \end{pmatrix}$,β'=$\begin{pmatrix} \beta_q' \\ \beta_t' \end{pmatrix}$
则我们得到了一个新模型

- **选模型**

$$y = X\beta + e, \quad E(e) = 0, \quad Cov(e) = \sigma^2 I$$

- **主要结论**
  1) 即使全模型正确，剔除一部分自变量之后，可使得剩余的那部分自变量的回归系数的 LS 估计的方差减小，但此时的估计一般为有偏估计。若被剔除的自变量对因变量影响较小，则可使得剩余的那部分自变量的回归系数的 LS 估计的精度提高。
  2) 当全模型正确时，用选模型做预测，预测一般是有偏的，但预测偏差的方差减小。若被剔除的自变量对因变量影响较小，则剔除掉这些变量后可使得预测的精度提高。
  3) 当全模型为真时，$E(\tilde{\sigma_q^2}) \geq E(\check{\sigma}^2)$ 仅当 $\beta^t = 0$ 时等号成立。即全模型的残差平方和总达到最小

1.2.1 平均残差平方和准则 (RMS_q)

残差平方和 SS_e 的大小刻画了数据与模型的拟合程度，SS_e 愈小，拟合得愈好，但是“SS_e 愈小愈好”却不能作为回归自变量的选择准则，因为它将导致全部自变量的入选。
为了防止选取过多的自变量，一种常见的做法是在残差平方和 SS_e 上添加对增加变量的惩罚因子。平均残差平方和 RMS_q 就是其中的一例，平均残差平方 RMS_q 和定义为

$$RMS_q = \frac{SS_{eq}}{n-q}$$

这里为 q 选模型设计阵 X_q 的列数。实际上 RMS_q 就是选模型下误差方差的 LS 估计。因子 $(n-1)^{-1}$ 随自变量的个数增加而变大，它体现了对变量个数的增加所施加的惩罚。依 RMS_q 准则，按 “RMS_q 愈小愈好” 选择自变量子集。

1.2.2 C_p 准则

对于选模型，C_p 统计量定义为

$$C_p = \frac{SS_{eq}}{\tilde{\sigma}^2} - (n-2q)$$

这里 SS_{eq} 为选模型下的残差平方和，$\tilde{\sigma^2}$ 为全模型下 σ^2 的得 LS 估计，q 为选模型设计阵的列数。依 C_q 准则，按 “C_q 愈小愈好” 选择自变量子集

1.2.3 AIC 准则

对于一般的统计模型，按 $Y_1, Y_2 ... Y_k$ 为一组样本，如果它们服从某个含 k 个参数的模型，对应的似然函数的最大值记为 $L_k(X_1, X_2 ... X_k)$ ，则 AIC 准则是选择使 AIC 统计量达到最小的模型

$$AIC = lnL_k(X_1, X_2 ... X_k) - k$$

等价地，可以取

$$AIC = ln(SS_{eq}) + 2q$$

1.3 解题过程

1.3.1 第一步

先导入 csv 文件中的数据，再对所有的全子集回归模型进行拟合

1.3.2 第二步

分别计算全子集回归模型 LS 估计及 C_p、RMS_q 和 AIC 值，并且分别三种准则从大到小进行排序

1.3.3 第三步

分别求出在 C_p、RMS_q 和 AIC 准则下，最优子集回归

```
The result for RSSE
[[1]]
[[1]]$lm123

Call:
lm(formula = y ~ x1 + x2 + x3, data = data)

Coefficients:
(Intercept)          x1          x2          x3
    -6.5564      0.0763      0.4262     -0.3225
```

图 2: RMS_q 准则下的结果

```
The result for cp
[[1]]
[[1]]$lm123

Call:
lm(formula = y ~ x1 + x2 + x3, data = data)

Coefficients:
(Intercept)          x1          x2          x3
    -6.5564      0.0763      0.4262     -0.3225
```

图 3: C_p 准则下的结果

```
The result for AIC
[[1]]
[[1]]$lm123

Call:
lm(formula = y ~ x1 + x2 + x3, data = data)

Coefficients:
(Intercept)          x1          x2          x3
    -6.5564      0.0763      0.4262     -0.3225
```

图 4: AIC 准则下的结果

1.4 结果解释

此问题有三个变量，共有 7 个不同的自变量子集。这 7 个变量自己的 LS 估计和 RMS_q，C_p 和 AIC 在上述程序中已经从小到大排序。根据如上所示结果可知，子集 (X_1, X_2, X_3) 对应的 RMS_q，C_p 和 AIC 值都达到最小，因此若没有别的附加考虑，在 RMS_q，C_p 或 AIC 准则下，最优子集回归为

$$y = -6.554 + 0.0763X_1 + 0.04262 - 0.3225X_3$$

2 第七题

2.1 题目重述

在林业工程中，研究树干的体积 Y 与离地面一定高度的树干直径 X1 树干高度 X2 之间的关系具有重要的意义，因为这种关系使我们能够用简单的方法从 X1 和 X2 的值去估计一棵树的体积，进而估计一片森林的木材储量。下表是一组观测数据：

X_1	X_2	Y	X_1	X_2	Y
8.3	70	10.3	12.9	85	33.8
8.6	65	10.3	13.3	86	27.4
8.8	63	10.2	13.7	71	25.7
10.5	72	16.4	13.8	64	24.9
10.7	81	18.8	14.0	78	34.5
10.8	83	19.7	14.2	80	31.7
11.0	66	15.6	14.5	74	36.3
11.0	75	18.2	13.0	72	38.3
11.1	80	22.6	16.3	77	42.6
11.2	75	19.9	17.3	81	55.4
11.3	79	24.2	17.5	82	55.7
11.4	76	21.0	17.9	80	58.3
11.4	76	21.4	18.0	80	51.5
11.7	69	21.3	18.0	80	51.0
12.0	75	19.1	20.6	87	77.0
12.0	74	22.2			

图 5: 题 7 数据表

2.2 解题思路

一般来讲，$\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n$ 是相关的，且它们的方差不等。所以不宜直接用 $\hat{e}_i$ 作比较来判别异常点。

引进学生化残差

称：

$$r_i=\frac{\hat{e}_i}{\sqrt{1-h_{ii}}s}\qquad(i=1,2,\cdots,n)$$

为学生化残差（或标准化残差），其中：h_{ii} 为 H 矩阵的第 i 个对角元，

$s=\sqrt{\frac{SSE}{n-p-1}}$。

一般说来，由于 $\hat{e}_i$ 与 s 是不独立的，r_i 的确切分布很难求得，但在模型假设成立时，$r_1,r_2,\cdots,r_n$ 近似独立，且近似服从 $N(0,1)$，可以近似认为 $r_1,r_2,\cdots,r_n$ 是来自 $N(0,1)$ 的随机子样（r_i 的分布可见陈希孺、王松桂《近代回归分析》，1987）。依据标准化残差 $r_1,r_2,\cdots,r_n$ 近似服从 N(0, 1)近似相互独立这一结论，常用残差图对模型假设的合理性进行检验。

对观测得到的实验数据集 (x_i',y_i),(i=1...n) 若经过回归诊断后得知，它们不满足 Gauss-Markov 条件，我们就要对数据采取“治疗”措施，实践证明，数据变幻时处理有问题数据的一种好方法。数据变换有很多种，本题采用 Box-Cox 变换 λ

$$Y^{(\lambda)}=\begin{cases}\frac{Y^{\lambda}-1}{\lambda} & ,\quad \lambda\neq 0\\ lnY & ,\quad \lambda=0\end{cases}$$

我们要确定变换参数 λ，使得 $y^{(\lambda)}$ 满足

$$y^{(\lambda)}=X\beta+e$$

$$Y^{(\lambda)} = \begin{pmatrix} y_1^{\lambda} \\ y_2^{\lambda} \\ \vdots \\ y_n^{\lambda} \end{pmatrix} \sim N_n(X\beta, \sigma^2 I_n)$$

Box-Cox 变换不仅在误差非正态场合可用，而且在回归函数非线性，误差方差非齐性，观测值间不独立时均可选用，因为通过这个变换，要求线性回归中四条假定都满足。

我们用极大似然方法来确定 λ

$y^{(\lambda)}$ 的似然函数为:

$$LL(\beta,\sigma^2,\lambda) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(Y^{(\lambda)} - X\beta\right)'\left(Y^{(\lambda)} - X\beta\right)\right\}$$

利用变换的形式可知 Y 的似然函数为：

$$L(\beta,\sigma^2,\lambda) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(Y^{(\lambda)} - X\beta\right)'\left(Y^{(\lambda)} - X\beta\right)\right\} \cdot |J|$$

其中：$|J| = \prod_{i=1}^{n}\left|\frac{dy_i^{(\lambda)}}{dy_i}\right| = \prod_{i=1}^{n}\left|y_i^{\lambda-1}\right|$

对固定的 λ，可得 β,σ^2 的有极大似然估计为：

$$\hat{\beta}_{\lambda} = (X'X)^{-1}X'Y^{(\lambda)}$$

$$\hat{\sigma}_{\lambda}^2 = \frac{1}{n}Y^{(\lambda)'}[1 - X(X'X)^{-1}X']Y^{(\lambda)} \overset{\Delta}{=} \frac{1}{n}SSE(\lambda, y)$$

对固定的 λ，有：

$$L_{\max}(\lambda) = \max_{\beta,\sigma^2} L(\beta,\sigma^2,\lambda)$$

$$= L(\hat{\beta}_{\lambda},\hat{\sigma}_{\lambda}^2,\lambda) = (2\pi\hat{\sigma}_{\lambda}^2)^{-\frac{n}{2}} e^{-\frac{n}{2}} \cdot |J|$$

$$\ln L_{\max}(\lambda) = -\frac{n}{2}\ln[SSE(\lambda, y)] + \ln|J| = -\frac{n}{2}\ln\left\{\frac{Y^{(\lambda)'}}{|J|^{1/n}}[1 - X(X'X)^{-1}X']\frac{Y^{(\lambda)}}{|J|^{1/n}}\right\}$$

$Z^{(\lambda)}=\dfrac{Y^{(\lambda)}}{\left|J^{1/n}\right|}$，则有：

$$z_i^{(\lambda)}=\frac{y_i^{(\lambda)}}{|J|^{1/n}}=\begin{cases} y_i^{(\lambda)}\Big/(\prod_{i=1}^{n}y_i)^{\frac{\lambda-1}{n}}, & \lambda\neq 0 \\ (\ln y_i)(\prod_{i=1}^{n}y_i)^{\frac{\lambda-1}{n}}, & \lambda=0 \end{cases} \qquad i=1,2,\cdots,n$$

$$SSE(\lambda,Z)=Z^{(\lambda)'}[1-X(X'X)^{-1}X']Z^{(\lambda)}$$

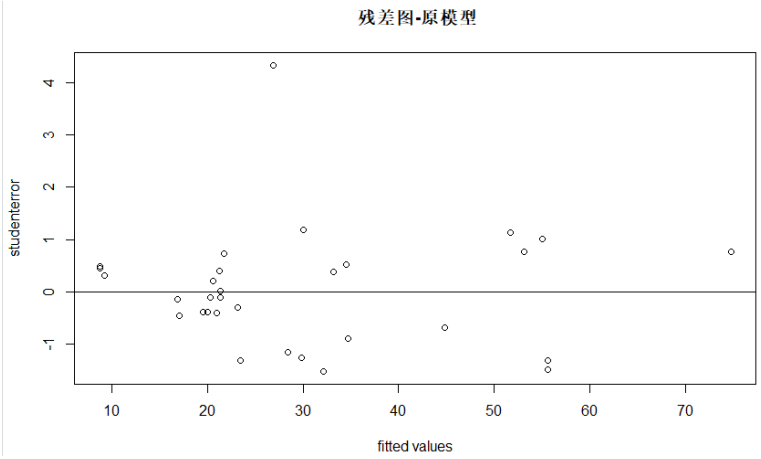
$$L_{\max}(\lambda)=-\frac{n}{2}\ln[SSE(\lambda,Z)]$$

为找出 λ 使 $L_{\max}(\lambda)$ 达到最大，只要使 $SSE(\lambda,Z)$ 达到最小即可。其解析解是比较难找的，通常的做法是给出一系列的 λ 值，画出 λ 与 $SSE(\lambda,Z)$ 的曲线，从图上找出使 $SSE(\lambda,Z)$ 达到最小的近似 $\hat{\lambda}$。并用此 $\hat{\lambda}$ 作Box-Cox变换并求出回归方程。

2.3 解题过程

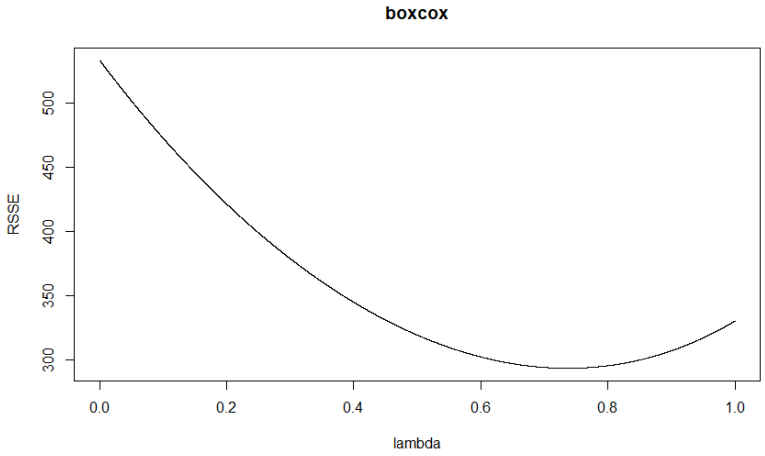
2.3.1 第一步

先导入 csv 文件中的数据，利用实验数据进行线性回归拟合，假设 Y 与 X_1 和 X_2 有如下线性回归关系做最小二乘分析 $Y=\alpha+\beta_1X_1^2+\beta_2X_2$，并做相应的残差图，试计算 Box-Cox 变换参数



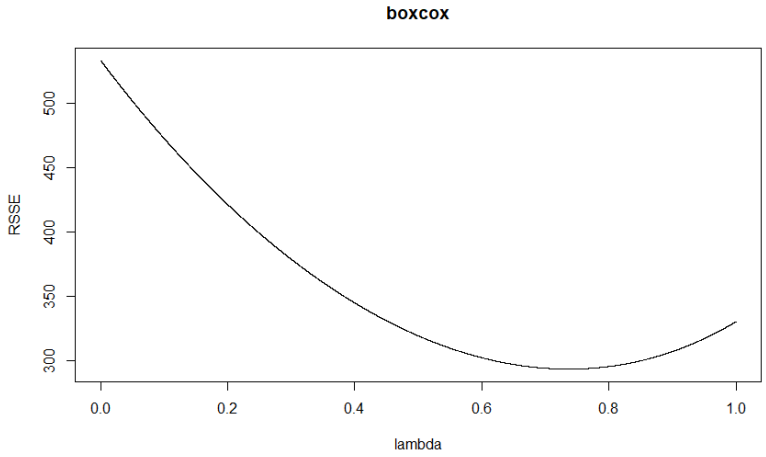
2.3.2 第二步

给出一系列 λ 的值，画出 $lambda$ 与 $SSE(\lambda, Z)$ 的曲线，从图上找出使 $SSE(\lambda, Z)$ 达到最小的 $\hat{\lambda}$, 并用此 $\hat{\lambda}$ 作 Box-Cox 变换并求出回归方程



2.3.3 第三步

出对所求 $\hat{\lambda}$ 值，做相应的 Box-Cox 变换，并对变换后的因变量做对 X_1 和 X_2 的最小二乘回归，并做残差图



2.4 结果分析

第一小问假设 Y 与 X_1 和 X_2 有如下线性回归关系 $Y = \alpha+\beta_1 X_1^2+\beta_2 X_2$ 做最小二乘分析，并做其相应的残差图，发现 Y 的估计值越小学生化残差越小，残差图呈发散趋势，Y 的学生化残差之间可能不独立。因此考虑对 Y 进行 Box-Cox 变换，并作出 Box-Cox 图，图中给出了区间 [0,1] 中不同的值所对应的残差平方和 SSE(λ,Z)，由程序得到 λ=0.737 时，残差平方和 SSE(λ,Z) 达到最小，因此我们可以近似认为 0.737 就是变换参数 λ 的最优选择

第二问再考虑 $Y^{0.737} = \alpha + \beta_1 X_1^2 + \beta_2 X_2$ 的最小二乘分析，并作出相应的残差图，可发现 Y 的学生化残差之间独立性加强，因此该模型成立。

3 第 11 题

3.1 题目重述

做了 10 次试验得观测数据如下：

(1) 若以 X_1，X_2 为回归自变量，问它们之间是否存在复共线关系?

(2) 试用岭迹法求 Y 关于 X_1，X_2 的岭回归方程，并画出岭迹图。

y	x1	x2
16.3	1	1.1
16.8	1.4	1.5
19.2	1.7	1.8
18	1.7	1.7
19.5	1.8	1.9
20.9	1.8	1.8
21.1	1.9	1.8
20.9	2	2.1
20.3	2.3	2.4
22	2.4	2.5

图 6: 题 11 数据表

- **解题思路**

  欲寻找 Y 与 X_1,X_2 之间合适的回归方程。观察可看出 X_1 随着 X_2 增减而增减，故需要先考虑两者之间存在复共线关系。由于若存在复共线性，则对参数的估计会产生较大的误差，为了减少误差（均方误差），在得出具体结果后，考虑采用岭估计法并使用岭迹图寻找合适的参数 k，代入得到相应的岭回归方程，并以岭估计的结果作为参数的估计值。

3.2 解题过程

首先从两个方面测算变量之间多重共线性的程度。

3.2.1 方差扩大因子法理论基础

有定理

$$\mathrm{var}(\hat{\beta}_j) = c_{jj}\sigma^2 / L_{jj}$$

其中，C_{jj} 是样本相关阵的逆阵的第 j 个对角元，L_{jj} 为第 j 个变量的离差

记 $R^2_{j|1,\cdots,j-1,j+1,\cdots,p}$ 为自变量 x_j 对其余 p−1 个自变量的复判定系数，可以证明

$$c_{jj}=\frac{1}{1-R^2_{j|1,\cdots,j-1,j+1,\cdots,p}} \quad （练习）$$

c_{jj} 定义为方差扩大因子 VIF_j，可知 $VIF_j \geqslant 1$。

和。则可知，C_{jj} 越大，对参数估计的偏差就越大。另有定理：
可见 c_{jj} 确实反映了样本存在多重共线性的程度。而且 C_{jj} 越大，多重共线性越强，参数估计的误差越大。因此将 C_{jj} 定义为方差扩大因子 VIF_j。故计算 VIF_j，当 VIF_j 大于 10，可以认为自变量 X_j 与其余变量之间存在严重的多重共线性。还可用 VIF 的平均值来考虑样本之间的多重共线性，若其均值远大于 1，则样本间存在严重的多重共线性。

3.2.2 特征根判定法

可以证明，当矩阵 X’ X 至少有一个特征根近似为零时，X 的列向量间必存在复共线性。故当 X’ X 特征值为 $\lambda_1 \leq \lambda_2 \leq .. \leq \lambda_p$ 时，考虑条件数 $k=\frac{\lambda_1}{\lambda_p}$。在应用经验中
0<k<100 时，设计矩阵 X 没有多重共线性；
100<k<1000 时，认为 X 存在较强的多重共线性；
当 k>1000 时，则认为存在严重的多重共线性。

3.2.3 第一部分小结

比较两种方法可见，特征根判定法判定矩阵要附加上常数项，而 VIF 的计算只与样本相关阵有关。另外，特征根判定法算法简单，结果清晰，但是相应的并不能给出一个变量是否可以被其他变量近似线性表示，而 VIF 法就能告诉我们这一点信息。

3.3 采用岭估计法

给出不同的 k 值时，相应估计参数的值，并绘出岭迹图。

3.3.1 岭回归的提出

当自变量间存在复共线性时，$|X'X| \approx 0$，我们设想给 X’X 加上一个正常数矩阵 kI（k>0），那么 X’X+kI 接近奇异的程度就会比 X’X 接近奇异的程度小。我们称 $\beta(k)=(X'X+kI)^{-1}X'Y$ 为 $\hat{\beta}$ 的岭估计，其中 k 称为岭参数。则有如下重要性质：(1)β（k）是回归参数 $\hat{\beta}$ 的有偏估计。(2) 对任

意 k>0，$||\beta||\neq 0$，总有 $||\beta(k)||<||\hat{\beta}||$。(3) 以 MSE 表示估计向量的均方误差，则存在 k>0，使得 MSE（$\beta(k)$）<MSE（$\hat{\beta}$）。

3.3.2 采用岭迹法选取 k 值原则

各回归系数的岭估计基本稳定
用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理
回归系数没有不合乎意义的绝对值
残差平方和增大不太多

3.4 结果展示与结果解释

3.4.1 多重共线性判断

方差扩大因子计算结果如下：

```
> C
          [,1]      [,2]
[1,]  34.23439 -33.73068
[2,] -33.73068  34.23439
```

图 7: 方差扩大因子

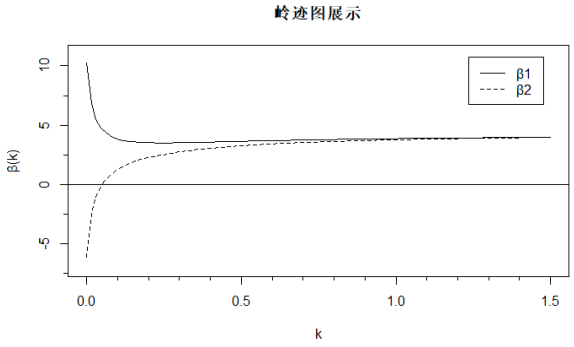
可见 VIF1=c11，VIF2=c22 较大，即 X1 与 X2 间有明显多重共线性。

```
> lamd
[1] 79.58548926  0.37258320  0.02192753
> k
[1] 3629.477
```

图 8: 特征值与条件数计算

可见变量间有严重的多重共线性

3.4.2 绘制岭迹图并解释



图中，β_2（0）比较大，且为负数，但当 k 增大时，β_2（k）迅速上升且稳定为正值。从最小二乘回归看，X_2 对因变量的有“负”的影响，从岭回归的观点看，X_2 对因变量有”正”的影响。从整体上看，当 k 达到 0.5—0.7 的范围时，各个 β_2（k）已大体上趋于稳定，因此，在这区间上取一个 k 值作岭回归可能得到较好的效果。

3.4.3 计算出岭回归参数

此处取 k=0.7，代入计算结果如下：

```
> b
        [,1]
β0 5.745425
β1 3.758205
β2 3.541735
```

故岭回归方程为 $Y = 5.75 + 3.76X_1 + 3.54X_2$
而使用最小二乘估计得到的方程是 $Y = 12.4 + 10.2X_1 - 6.1X_2$
可见有较大差别。

4 第 12 题

4.1 概述

社会上存在着许多因素，某种商品的销售量也和其外因和内因有关，例如：外因有人们的收入，其他商品的价格；内因有该商品的平均价格指数，该商品的社会拥有量。以下对题目中数据表进行分析以得到可信服的结论。
关键词：数据标准化，主成分分析法，R

- **问题重述**
  对某种商品的销售量 Y 进行调查，并考虑有关的四个因素：X_1: 居民可支配收入；X_2: 该商品的平均价格数；X_3: 该商品的社会拥有；量 X_4: 其他消费品的平均价格指数
  下面是调查数据：

序号	X_1	X_2	X_3	X_4	Y
1	82.9	92	17	94	8.4
2	88	93	21.3	96	9.6
3	99.9	96	25.1	97	10.4
4	105.3	94	29	97	11.4
5	117.7	100	34	100	12.2
6	131.0	101	40	101	14.2
7	148.2	105	44	104	15.8
8	161.8	112	49	109	17.9
9	174.2	112	51	111	19.6
10	184.7	112	53	111	20.8

图 9: 12 题数据表

- **主成分分析法的基本原理**
  主成分分析的原理是设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原来变量的信息的统计方法，它也是数学上处理降维的一种方法。主成分分析是设法将原来众多具有一定相关性（比如 P 个指标），重新组合成一组新的互相无关的综合指标来代替原来的指标。通常数学上的处理就是将原来 P 个指标作线性组合，作为新的综合指标。最常用的做法就是用 F_1（选取的第一个线性组合，即第一个综合指标）的方差来表达，即 Var（F_1）越大，表示 F_1 包含的信息越多。因此在所有的线性组合中选取的 F_1 应该是方差最大的，故称 F_1 为第一主成分。如果第一主成分不足以代表原来 P 个指标的信息，再考虑选取 F_2 即选第二个线性组合，为了有效地反映原来信息，F_1 已有的信息就不需要再出现在 F_2 中，用数学语言表达就是要求 Cov

(F_1,F_2) =0，则称 F_2 为第二主成分，依此类推可以构造出第三、第四，……，第 P 个主成分。

- **解题思路**

  由标准化的数据得到相关系数矩阵，对其进行变量的提取。找到主要的变量从而完成对变量的降维。通过对得到的主要变量进行线性回归，再将此主成分变量回归方程代回原变量得到关于原变量的回归方程。

4.2 主成分提取

利用相关系数矩阵进行主成分分析并提取变量 四个主成分：

	Comp.1	Comp.2	Comp.3	Comp.4
x1	0.502	0.234	0.582	0.596
x2	0.500	-0.492	-0.610	0.369
x3	0.498	0.709	-0.368	-0.338
x4	0.501	-0.448	0.392	-0.628

图 10: 主成分提取结果

$$Z_1 = 0.502X_1^* + 0.500X_2^* + 0.498X_3^* + 0.501X_4^*$$

$$Z_2 = 0.234X_1^* - 0.492X_2^* + 0.709X_3^* - 0.448X_4^*$$

$$Z_3 = 0.582X_1^* - 0.610X_2^* - 0.368X_3^* + 0.392X_4^*$$

$$Z_4 = 0.596X_1^* + 0.396X_2^* - 0.338X_3^* - 0.628X_4^*$$

```
Importance of components:
                          Comp.1
Standard deviation     1.9858342
Proportion of Variance 0.9858844
Cumulative Proportion  0.9858844
                           Comp.2
Standard deviation     0.20041009
Proportion of Variance 0.01004105
Cumulative Proportion  0.99592547
```

图 11: 各个主成分占比结果

故可以看出前两个主成分的贡献率为 0.9959, 则采用这两个变量

4.3 反解出方程

还原变量得到回归方程

```
Coefficients:
            Estimate
(Intercept) 14.03000
z1           2.06126
z2           0.61382
```

图 12: 主成分回归系数

利用前两个变量可得到

$$Y = 14.0300 + 2.06126Z_1 + 0.68162Z_2$$

又因为

$$Z_1 = 0.502X_1 + 0.500X_2 + 0.498X_3 + 0.501X_4$$

$$Z_2 = 0.234X_1 - 0.492X2 + 0.709X3 - 0.448X4$$

带入关于 Z_1,Z_2 的方程得到

```
 (Intercept)          x1          x2          x3          x4
-16.88460655  0.03420968  0.09376460  0.11954881  0.12360237
```

图 13: 最终回归方程

所以最后还原标准化的主成分回归方程为

$$Y = -16.8846 + 0.03408X_1 + 0.0937X_2 + 0.1189X_3 + 0.1236X_4$$