上机附加题部分

韩硕
2018302060281

班 级: 18 级统计

2020 年 12 月 12 日

1 附加题

1.1 前头的废话

因为当时小组直接一股脑把所有题目一起合作完成，并且保证每个方法每个人每个部分都有涉及，于是我们小组附加题部分思路差不多。都是给定真值然后用 R 语言来模拟出具有明显线性关系的一组数据，并用岭回归，最小二乘法，主成分分析法进行比较。

1.2 模拟数据的思路

使用 r 语言随机生成一组数据用于模拟：考虑模拟四个变量的模型，其中 X_1 是取自（1,3）的均匀分布的随机变量，X_2 是取自（2,4）的均匀分布的随机变量，而 X_3 是在 X_1 上附加一个微小扰动后生成的数据变量，由此进一步确定出了共线性；X_4 是取自（4,6）的均匀分布的随机变量，Y 是由 $Y_{real} = 13 + 3X_1 + 4X_2 + 3X_3 - 2X_4 + \varepsilon$ 生成的模拟因变量，其中 ε 是从标准正态分布中取出的随机数。
设置了不同的种子后模拟出的一组数据如下：

Y	X_1	X_2	X_3	X_4
19.97	**1.5**	**2.4**	**1.3**	**5.2**
29.58	**1.7**	**3.4**	**1.9**	**4.0**
27.66	**2.1**	**3.1**	**2.0**	**4.6**
31.33	**2.8**	**2.3**	**2.7**	**4.6**
26.76	**1.4**	**3.9**	**1.5**	**5.6**
35.87	**2.8**	**3.9**	**2.9**	**4.5**
28.97	**2.9**	**2.3**	**2.7**	**5.4**
30.08	**2.3**	**3.7**	**2.2**	**5.8**
27.09	**2.3**	**2.9**	**2.3**	**5.9**
23.76	**1.1**	**3.1**	**1.2**	**4.1**

图 1: The Testing Data

1.3 最小二乘法估计

直接导入数据，用自带 LM 函数来进行最小二乘估计，得到的值为：

$$\beta_0 = 11.6892$$

$$\beta_1 = -0.8351; \beta_2 = 3.1123$$

$$\beta_3 = 7.2840; \beta_4 = -1.3275$$

与模拟的真值进行比较发现相差巨大，这是意料中的事情，因为设置变量中存在明显相关性。

1.4 主成分处理

首先考虑主成分分析对此数据的改善

1.4.1 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。

1.4.2 Geographic Definition

The rotation transformation allows the original data distributed on the X_1, X_2 axis convert to the coordinate system represented by the F1 and F_2 axes. Through which the data can be scattered on the F_1 axis to the greatest extent. And meanwhile the F_2 axis can be ignored.In short, most of the information the data contains can be expressed only through the F_1 axis, so as to achieve the purpose of dimensional reduction.

1.4.3 Main clue

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

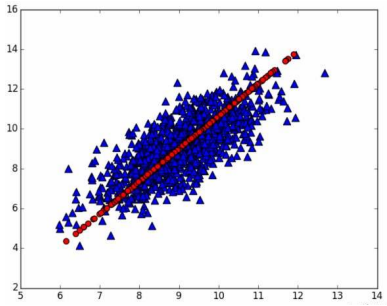

图 2: A example

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。

1.4.4 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

	Comp.1	Comp.2	Comp.3	Comp.4
x1	0.691		0.110	0.715
x2	-0.201	-0.723	0.652	0.105
x3	0.667		0.283	-0.688
x4	0.193	-0.690	-0.694	

图 3: 主成分提取结果

四个主成分:

$Z_1 = 0.691X_1 - 0.201X_2 + 0.667X_3 + 0.193X_4; Z_2 = -0.723X_2 - 0.6909X_4$

$Z_3 = 0.11X_1 - 0.652X_2 + 0.28X_3 - 0.694X_4; Z_4 = 0.72X_1 + 0.105X_2 - 0.688X_3$

	Comp.1	Comp.2	Comp.3
Standard deviation	1.4368462	0.9926661	0.9707028
Proportion of Variance	0.5161318	0.2463465	0.2355660
Cumulative Proportion	0.5161318	0.7624783	0.9980442

	Comp.4
Standard deviation	0.088447872
Proportion of Variance	0.001955757
Cumulative Proportion	1.000000000

图 4: 各个主成分占比结果

故可以看出前 3 个主成分的贡献率已经足够高，于是选取 $Z_1; Z_2; Z_3$

1.5 反解出方程

还原变量得到回归方程

```
Coefficients:
            Estimate
(Intercept)  28.0722
z1            1.8831
z2           -0.8629
z3            2.9408
```

图 5: 主成分回归系数

利用前 3 个变量可得到

$$Y = 28.0722 + 1.8831Z_1 - 0.8629Z_2 + 2.9408Z_3$$

又因为

$$Z_1 = 0.691X_1 - 0.201X_2 + 0.667X_3 + 0.193X_4$$

$$Z_2 = -0.723X_2 - 0.6909X_4$$

$$Z_3 = 0.11X_1 - 0.652X_2 + 0.28X_3 - 0.694X_4$$

带入关于 Z_1,Z_2,Z_3 的方程得到

```
(Intercept)          x1          x2          x3          x4
  11.730222    2.651082    3.635649    3.705617   -1.637811
```

图 6: 最终回归方程

所以最后还原标准化的主成分回归方程为

$$Y = 11.730222 + 2.651082X_1 + 3.635649X_2 + 3.705617X_3 - 1.637811X_4$$

1.6 承上转下

可以发现相比最小二乘，结果已经得到充分改善，下面用岭回归分析，先给出定理介绍

1.7 岭回归解释与相关定理

1.7.1 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 个变量的离差和。则可知，C_{jj} 越大，对参数估计的偏差就越大。另有定理：

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 $\mathrm{VIF_j}$，可知 $\mathrm{VIF_j} \geqslant 1$。

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。

1.7.2 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。

1.7.3 岭回归理论

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（β（k））<MSE（$\hat{\beta}$）。

1.7.4 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

1.8 运行结果

1.8.1 绘制岭迹图并解释

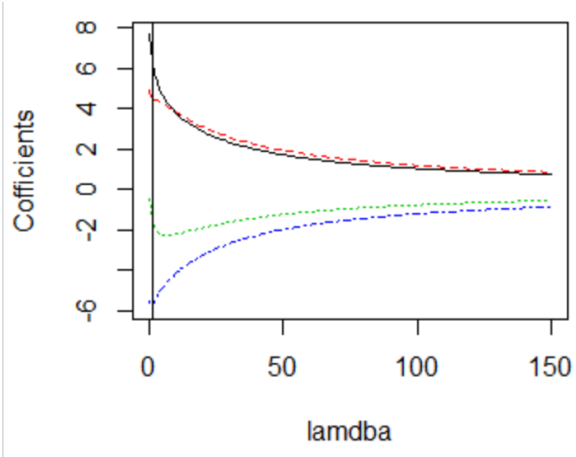

图 7: 不同参数的岭迹图

图中，β_1（0）（蓝色）为负数，但当 k 增大时，β_1（k）（绿色）迅速上升且稳定为正值。从最小二乘回归看，X_1 对因变量的有“负”的影响，从岭回归的观点看，X_1 对因变量有”正”的影响。同样，β_3 变化显著，但是 β_1 和 β_3 的总和几乎不变，这暗示着 X_1 和 X_3 变量之间有较强的线性相关。这与模拟数据中的构造相符合。从上图看，λ 的选择并不是那么重要，只要不离 $lambda = 0$ 太近就没有多大差别。并且越往后，各个 β（k）已大体上趋于稳定，因此，在这区间上取一个 λ 值作岭回归可能得到较好的效果。下面我们用 ridge 中的包中的 linearRidge() 函数进行自动选择岭回归参数来取一个较为合理的 λ 参数

1.8.2

```
Coefficients:
            Estimate Scaled estimate Std. Error (scaled)
(Intercept)  84.6663              NA                   NA
X1            1.3193         25.8786               3.7466
X2            0.2882         15.5365               4.1957
X3           -0.1539         -3.4143               3.7137
X4           -0.3820        -22.1485               4.2062
            t value (scaled) Pr(>|t|)
(Intercept)              NA       NA
X1                    6.907 4.94e-12 ***
X2                    3.703 0.000213 ***
X3                    0.919 0.357896
X4                    5.266 1.40e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.01527505, chosen automatically, computed using 2 PC

Degrees of freedom: model 3.022 , variance 2.833 , residual 3.21
```

图 8: linearRidge() 函数进行自动选择岭回归参数

1.8.3 计算出岭回归参数

此处取 k=0.15，代入计算结果如下：

$$\beta_0 = 8.98; \beta_1 = 2.78$$

$$\beta_2 = 3.93; \beta_3 = 3.34$$

$$\beta_4 = -1.68$$

故岭回归方程为

$$Y_{ridge} = 8.98 + 2.78X_1 + 3.93X_2 + 3.34X_3 - 1.68X_4$$

而使用最小二乘估计得到的方程是

$$Y_{LSE} = 11.67 - 8.31X_1 + 3.11X_2 + 7.28X_3 - 1.32X_4$$

而原模型的模型真值为：$Y_{real} = 13 + 3X_1 + 4X_2 + 3X_3 - 2X_4 + \varepsilon$ 可见原模型和最小二乘法的估计有很大差别，这是意料之中的。而且通过选择 k=0.13 的岭回归有效的减少了系数 β_i 上的差别。除了 β_0, 其他的几个系数估计都更加准确了。