Advances in Applied Mathematics 应用数学进展, 2023, 12(9), 3771-3784
Published Online September 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.129371

Hans汉斯

关于混凝土抗压强度的线性回归分析

黄　梅

云南财经大学统计与数学学院，云南　昆明

收稿日期：2023年7月30日；录用日期：2023年8月23日；发布日期：2023年8月31日

摘　要

本文研究了混凝土抗压强度与龄期和混凝土成分之间的关系，使用最小二乘法、最优子集选择、岭回归和Lasso回归建立了5个回归模型对混凝土抗压强度进行建模。通过建立训练集与测试集，计算测试集上的均方误差来评价模型的预测效果。结果显示，用交叉验证做最优子集选择拟合出的模型的均方误差是最小的，其次加权最小二乘回归的效果也比较好，它可以有较好的拟合优度，它的预测误差和其他几个模型的差距也不是很大。同时，模型也还有进一步改进的空间，可以考虑各特征之间的特性来构造出最合适的模型。

关键词

最小二乘回归，最优子集选择，岭回归，Lasso回归

Linear Regression Analysis on the Compressive Strength of Concrete

Mei Huang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Jul. 30th, 2023; accepted: Aug. 23rd, 2023; published: Aug. 31st, 2023

Abstract

In this paper, the relationship between concrete compressive strength and age and concrete composition was investigated and five regression models were developed to model concrete compressive strength using least squares, optimal subset selection, ridge regression and Lasso regression. The prediction effectiveness of the models was evaluated by establishing training and test sets and calculating the mean square error on the test set. The results show that the mean square error of the model fitted by using cross validation as the optimal subset selection is the smallest, followed by the weighted least squares regression, which can have a better goodness of fit, and the difference between its prediction error and several other models is not very large. At the same time,

文章引用: 黄梅. 关于混凝土抗压强度的线性回归分析[J]. 应用数学进展, 2023, 12(9): 3771-3784.
DOI: 10.12677/aam.2023.129371

there is still room for further improvement of the model, and the characteristics between the features can be considered to construct the most suitable model.

Keywords

Least Squares Regression, Optimal Subset Selection, Ridge Regression, Lasso Regression

1. 引言

混凝土强度是混凝土质量控制的核心内容，是结构设计和施工的重要依据，是混凝土最重要的性能之一。混凝土的质量，直接关系到分项工程、分部工程以及单位工程的评定验收。混凝土的原材料处理常用的水泥以外，新出现了球状水泥、调粒水泥、活化水泥等，而矿渣、粉煤灰、沸石等超细掺合料及硅粉等，成为高强度混凝土中不可缺少的组分[1]。超强塑化剂、高效减水剂等外加剂的应用也越来越普遍。所有这些使得混凝土的组成更加复杂，性能更为优越，适应性更强，应用范围更广泛[2]。

一般来说，混凝土抗压强度是对混凝土试件进行 28 d 标准养护后，通过测试获得的，但混凝土强度的测试过程非常复杂且费时，即使试验结果不能满足规定强度，混凝土强度也不能提高。因此，混凝土抗压强度早期预测具有重要研究意义[3]。本文主要研究混凝土与水泥、高炉矿渣、粉煤灰、水、高效减水剂、粗骨料和细骨料以及龄期之间的线性关系，当然线性模型只是最简单和基础的模型，BP 神经网络在建立非线性模型方面有非常好的效果[4]，但依据本学期学习的内容，本文只研究了线性关系，也取得了比较好的效果。

2. 模型理论

本文主要做的是线性回归分析，标准的线性回归模型是：

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\varepsilon$$

通常用于描述响应变量 Y 与预测变量 $X_1,X_2,\cdots,X_p$ 之间的线性关系，$\beta_0,\beta_1,\cdots,\beta_p$ 是需要估计的参数。下面介绍拟合此模型的方法。

2.1. 最小二乘回归

最小二乘法是求解线性回归模型最经典的方法，它的主要通过最小化 RSS，

$$\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip}\right)^2$$

从而得到 $\hat{\beta}$。

对于 n 组观测数据，则需要拟合下列的线性回归方程组：

$$\begin{cases}y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1\\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2\\ \cdots\\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n\end{cases}$$

它可以表示成矩阵的形式： $y=X\beta+\varepsilon$，对于求解出最小的 RSS，即关于求解 $\hat{\beta}$ 值，一般采用梯度下降法或者牛顿法。可以得到它的算法是：

$$\hat{\beta}=\left(X'X\right)^{-1}X'Y$$

其中 $X'X$ 可逆，则要求 $n\geq p+1$ 且特征不冗余。

2.2. 最优子集选择

最优子集选择是对 p 各预测变量的所有可能线性组合分别使用最小二乘回归进行拟合。对含有一个预测变量的模型，拟合 p 个模型；对含有两个预测变量的模型，拟合 $\left(\frac{p}{2}\right)\geq p\left(p-1\right)/2$ 个模型，以此类推。最后在所有可能的模型中选择最优模型[5]。它的算法过程是：对于 $k=1,2,\cdots,p$；k 取每一个值都会生成相应数量 C_p^k 个模型，从中挑出 RSS 最小或 R 方最大的模型，记为 $M_0,M_1,\cdots,M_p$；再根据交叉验证预测误差、$C_p\left(\text{AIC}\right)$ (残差平方和)、BIC 或者调整的 R 方找出最优的模型。其中

$$C_p=\frac{1}{n}\left(\text{RSS}+2d\hat{\sigma}^2\right)$$

$$\text{AIC}=\frac{1}{n\hat{\sigma}^2}\left(\text{RSS}+2d\hat{\sigma}^2\right)$$

$$\text{BIC}=\frac{1}{n}\left(\text{RSS}+\log\left(n\right)d\hat{\sigma}^2\right)$$

d 表示预测变量个数。

C_p 统计量是测试均方误差的无偏估计，所以测试误差较低的模型 C_p 取值也比较低，因此，可以通过选择具有较低 C_p 的模型作为最优模型；AIC 是赤池信息准则，它适用于许多极大似然法进行拟合的模型，对于最小二乘模型 C_p 与 AIC 是成正比的；BIC 是从贝叶斯信息准则，它通常给包含多个变量的模型施以较重的惩罚，所以与 C_p 相比，得到的模型规模更小[5]。以上三个统计量都是选择值最小的模型作为最优模型。

调整的 R^2 是另一种对一系列具有不同变量个数的模型进行选择的方法，它定义为：

$$\text{调整的}R^2=1-\frac{\text{RSS}/\left(n-d-1\right)}{\text{TSS}/\left(n-1\right)}$$

其中 $\text{TSS}=\sum\left(y_i-\bar{y}\right)^2$。它的使用准则是调整的 R^2 值越大，模型测试误差越小，因此选择值最大的模型作为最优模型。

2.3. 岭回归

岭回归系数估计值 $\hat{\beta}$ 通过最小化下式得到：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_jx_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\text{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

其中 $\lambda\sum_{j=1}^{p}\beta_j^2$ 是压缩惩罚，在机器学习中又叫做正则项，可以通过控制 λ 的大小来调节模型的复杂度，当 $\lambda=0$ 时，惩罚项不产生作用，岭回归与最小二乘估计结果相同，随着 $\lambda\to\infty$，压缩惩罚项的影响力增加，岭回归系数估计值越来越接近于 0。

岭回归的系数估计问题等价于求解以下问题：

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\},\ \sum_{j=1}^{p}\beta_j^2\le s$$

得到的模型估计是：

$$\hat{\beta}(k)=\left(X'X+kI\right)^{-1}X'y$$

与最小二乘相比，岭回归的优势在于它综合权衡了误差与方差。随着 λ 的增加，岭回归拟合效果的光滑度降低，虽然方差降低，但偏差在增大。而岭回归的劣势在于它无法做子集选择，它的最终模型包含全部的变量。

2.4. Lasso 回归

Lasso 的系数估计值 $\hat{\beta}$ 通过求解下式的最小值得到：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

岭回归的惩罚项采用的是二范数，而 Lasso 的惩罚项采用的 1 范数，它定义为 $\|\beta\|_1=\sum\left|\beta_j\right|$。Lasso 估计也将系数估计值往 0 的方向进行缩减，而且，当参数 λ 足够大时，惩罚项具有将其中一些系数估计值强制设定为 0 的作用。所以 Lasso 可以完成变量选择，它得到了稀疏模型——只包含所有变量的一个子集的模型[5]。

Lasso 的系数估计也等价于求解以下问题：

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\},\ \sum_{j=1}^{p}\left|\beta_j\right|\le s$$

当 $n=p$ 时，Lasso 估计形式如下：

$$\hat{\beta}_j=\begin{cases}y_j-\dfrac{\lambda}{2}, & y_j>\dfrac{\lambda}{2}\\ y_j+\dfrac{\lambda}{2}, & y_j<-\dfrac{\lambda}{2}\\ 0, & \left|y_j\right|\le\dfrac{\lambda}{2}\end{cases}$$

3. 实证分析

3.1. 数据介绍

3.1.1. 数据来源说明

本文所用的数据集来源于机器学习存储库(UIC)，由台湾中华大学土木工程系叶毅成教授捐赠的混凝土抗压强度数据集。数据是由实验室测定的在特定龄期(天)下给定混合物的实际混凝土抗压强度(MPa)，混凝土的成分包括水泥、高炉矿渣、粉煤灰、水、高效减水剂、粗骨料和细骨料。数据为原始形式(未缩放)。本文使用 R 软件对数据进行分析。

3.1.2. 数据变量说明

该数据集一共有 9 个变量，其中预测变量 8 个，响应变量 1 个，样本有 1030 个，没有缺失值，而且所有变量都是定量变量，具体变量信息见表 1。

Table 1. Variable description
表 1. 变量描述

变量分类	变量	变量名称	变量含义
预测变量	*X*1	Cement	水泥：kg/m^3 混合物
	*X*2	Blast Furnace Slag	高炉矿渣：kg/m^3 混合物
	*X*3	Fly Ash	粉煤灰：kg/m^3 混合物
	*X*4	Water	水：kg/m^3 混合物
	*X*5	Superplasticizer	高效减水剂：kg/m^3 混合物
	*X*6	Coarse Aggregate	粗骨料：kg/m^3 混合物
	*X*7	Fine Aggregate	细骨料：kg/m^3 混合物
	*X*8	Age (day)	龄期：天
响应变量	*Y*	Concrete compressive strength	抗压强度：MPa

3.1.3. 描述性分析

对每个变量做描述性统计分析，得到的表 2 是各变量的最小值、第一四分位数、中位数、均值、第三四分位数、最大值。最小值和最大值体现了数据的波动范围，而它们的差值大小可以初步判断数据的波动大小，均值和中位数分别是在做参数估计和非参数估计使用的重要统计量，而分位数也是我们在估计位置参数时常用到的统计量。

Table 2. Descriptive statistics
表 2. 描述统计

	Min	1st Qu	Median	Mean	3rd Qu	Max
*X*1	102.0	192.4	272.9	281.2	350.0	540.0
*X*2	0.0	0.0	22.0	73.9	142.9	359.0
*X*3	0.00	0.00	0.00	54.19	118.30	200.10
*X*4	121.8	164.9	185.0	181.6	192.0	247.0
*X*5	0.000	0.000	6.400	6.205	10.200	32.200
*X*6	801.0	932.0	968.0	972.9	1029.4	1145.0
*X*7	594.0	731.0	779.5	773.6	824.0	992.6
*X*8	1.00	7.00	28.00	45.66	56.00	365.00
Y	2.33	23.71	34.45	35.82	46.13	82.60

相关系数矩阵热图用来判断两个变量之间的线性相关性，蓝色程度越深表示正相关性越强，红色程度越深表示负相关性越强。从图 1 可以看出 *X*4 和 *X*5 两个变量，即混合物中水的含量和高效减水剂的含量之间有较强的负相关性，在后续做线性回归时，可以考虑将其中一个用另一个线性表出。变量 *X*3 和 *X*6，变量 *X*8 和 *X*6 之间的相关系数接近于 0，即在建模的时候可以考虑混合物中粉煤灰的含量与粗骨料的含量相互独立，粗骨灰的含量与混凝土的龄期相互独立。

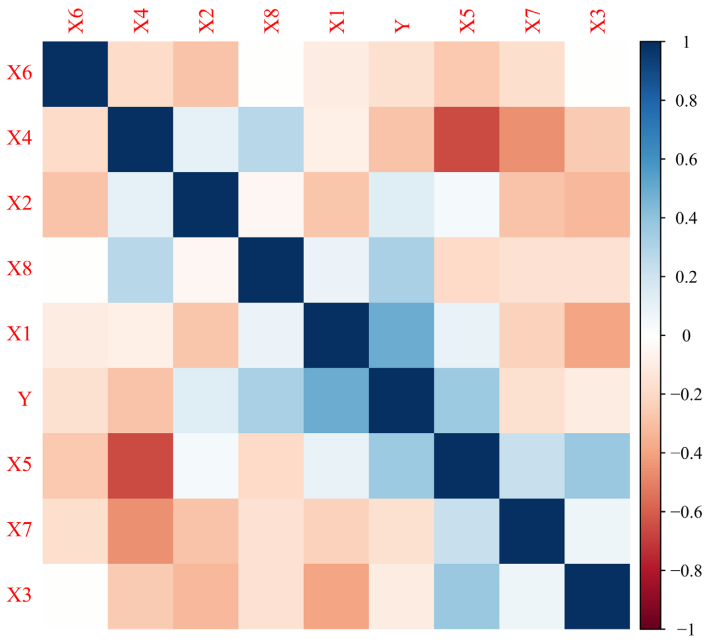

Figure 1. Heat map of correlation coefficient matrix
图 1. 相关系数矩阵热图

在回归模型建立过程中，我们先划分数据的训练集和测试集，在训练集上拟合模型，在测试集上做预测，然后用均方误差评价模型效果。本文将 2/3 的数据作为训练集，1/3 的数据作为测试集，见表 3。

Table 3. Experimental data set
表 3. 试验数据集

数据类型	预测变量	响应变量
训练集	xtrain 687 项	ytrain 687 项
测试集	xtest 343 项	ytest 343 项

3.2. 最小二乘法

做简单的线性回归分析，先用最小二乘法拟合多元线性模型，首先得到的是表 4 模型的描述统计数据，分别是残差的最小值、分位数、最大值。

Table 4. Residual data for model 1
表 4. 模型 1 的残差数据

Min	1Q	Median	3Q	Max
−31.653	−6.147	−0.678	6.839	34.804

表 5 是初始模型拟合结果，Estimate 是回归参数的估计，它包括了常数项系数和预测变量系数，所以初始模型为：

$$\text{Model 1: } Y = -39.707456 + 0.121677 * X1 + 0.104744 * X2 + 0.089102 * X3$$
$$-0.13217 * X4 + 0.33244 * X5 + 0.022714 * X6 + 0.029404 * X7 + 0.128748 * X8$$

Std.Error 为回归参数的标准误，t value 表示检验 t 值，Pr(>|t|)是 t 检验的 P 值。“***”说明极为显著，“**”说明高度显著，“*”说明显著，“.”说明不太显著，没有记号为不显著。所以，从表 5 可以看出拟合出的模型所有变量的系数都是显著的。

Table 5. Estimation of regression parameters for model 1
表 5. 模型 1 的回归参数估计

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−39.707456	30.968586	−1.282	0.20022
*X*1	0.121677	0.010171	11.963	<2e−16***
*X*2	0.104744	0.012182	8.599	<2e−16***
*X*3	0.089102	0.015261	5.839	8.16e−09***
*X*4	−0.132170	0.046732	−2.828	0.00482**
*X*5	0.332440	0.115643	2.875	0.00417**
*X*6	0.022714	0.010974	2.070	0.03885*
*X*7	0.029404	0.012651	2.324	0.02041*
*X*8	0.128748	0.007202	17.876	<2e−16***

表 6 是关于模型拟合效果的估计，虽然从检验的 P 值看出该模型估计是显著的，但通过 R^2 和调整的 R^2 可以看出，模型的拟合度并不高，只达到了 0.62，模型还有望调整，使得模型具有更高的拟合度和预测精度。

Table 6. The effect of the fit of model 1
表 6. 模型 1 的拟合效果

Residual standard error	10.36 on 678 degrees of freedom
Multiple *R*-squared (R^2)	0.6258
Adjusted *R*-squared (调整的 R^2)	0.6214
F-statistic (F 统计量)	141.7on 8 and 678 DF
P-value (P 值)	<2.2e−16

绘制模型的拟合效果图可以观察到很多改进方向，图 3 的第一行第一列是残差拟合图，它是模型拟合值与残差之间的散点图，红色线条表示二者关系的平滑曲线。若满足线性假设，二者应该不存在任何趋势性的关系，即红色线条应该与 y = 0 基本重合。从图中观察，红色线条大致与 y = 0 重合，但是散点图呈现倒漏斗形状，并不是均匀地分布在 y = 0 周围，可以考虑误差项的异方差性。

Q-Q 图它可以检验数据序列是否满足某种概率分布，它将对应分布的概率分位数作为横坐标、数据序列的分位数作为纵坐标作成散点图，若该数据序列满足该概率分布，则散点的趋势线应该与某条直线基本重合。图 2 的第一行第二列是模型 1 的正态 Q-Q 图，可以看出散点的趋势与 y = x 是基本重合的，认为总体上看，模型 1 的残差大致服从标准正态分布。

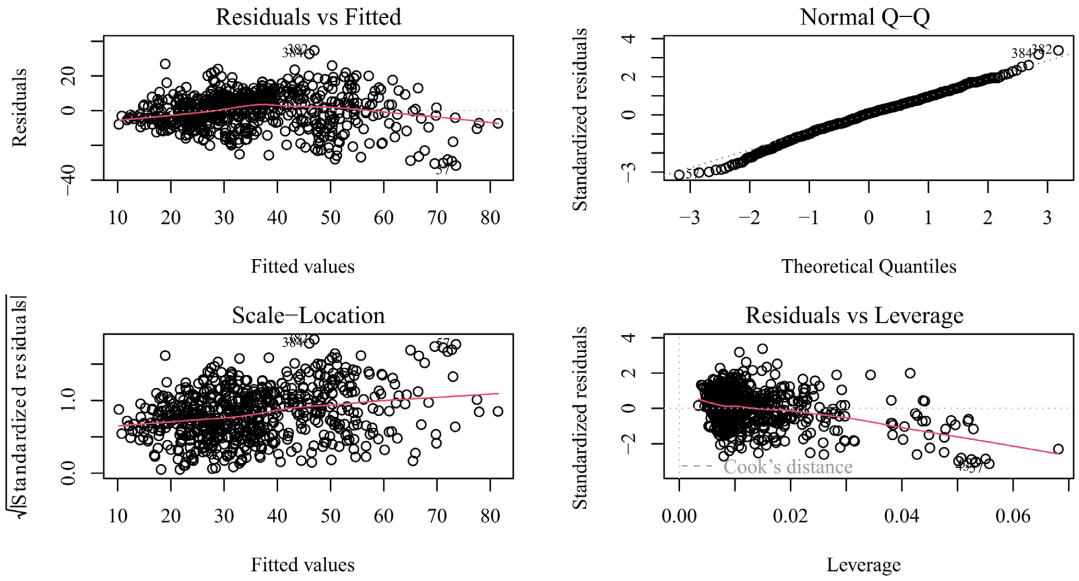

Figure 2. Residual fit plot for model 1
图 2. 模型 1 的拟合效果图

若模型满足同方差假设，则拟合值处于不同位置时残差的分布范围应该基本相同，即没有明显的集聚性或离散性，这一点可以通过观察残差－拟合图散点的离散程度，但更直观地是通过观察尺度－位置图。该图的纵坐标为标准化残差绝对值的平方根。若满足假设，则趋势线应该基本呈水平状。图 2 的第二行第一列是模型 1 的尺度－位置图，可以看出趋势线是呈缓慢上升状的，得到与残差拟合图相似的结论，即模型的误差项存在异方差性。

图形只能作为我们猜想的一个依据，在下结论之前我们需要做严格的假设检验。做 Shapiro-Wilk 检验：得到的 P 值是 0.01341，在 0.05 的置信水平下拒绝原假设，认为残差服从正态分布。做异方差检验：得到的 P 值 $< 2.22*10^{-16}$，在 0.05 的置信水平下拒绝原假设，认为模型是存在异方差性的。Durbin-Watson 检验可以判断残差是否存在一阶自相关，即检验模型是否满足独立性假设，得到的 P 值 < 0.05，在 0.05 的置信水平下拒绝原假设，认为模型的残差是不存在一阶自相关性的。方差膨胀系数(VIF)可以用来检验自变量之间是否具有多重共线性。自变量对应的 VIF 越大，说明其越有可能与其他自变量存在多重共线性。如果自变量之间存在多重共线性，即某个特征可以有其他自变量线性表出，这会增大解的不确定性，可能导致最后与真实的 β 差别很大，或解的取值范围变大。我们通过上一节特征之间的相关矩阵发现 $X4$ 与 $X5$ 是存在线性相关的可能性的。从另一个方面，我们查看特征的方差膨胀系数(VIF)，其中

$$\mathrm{VIF}\left(\hat{\beta}_j\right)=\frac{1}{1-R^2_{X_j|X_{-j}}}$$

结果如表 7 所示，发现膨胀系数最高的是 $X1$，而剩下的变量 $X2$、$X3$、$X4$、$X7$ 都存在较高的方差膨胀系数。

Table 7. Characteristic expansion factor for model 1
表 7. 模型 1 的特征膨胀因子

$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$
7.15	7.01	6.05	6.54	3.09	4.78	6.64	1.09

线性回归中一个重要的假设是误差项的方差是恒定的，即 $Var(\varepsilon_i)=\sigma^2$。线性模型中的假设检验和标准误差、置信区间计算都依赖于这一假设，但我们从前面的分析已经知道，误差项是存在异方差性的。因此考虑用凹函数对响应变量做变换，取响应变量为 log(*Y*)，得到的模型结果显示 R^2 降低到了 0.5615，说明取对数并没有使得模型的拟合效果更好。另一个消除异方差性的方法是使用加权最小二乘法拟合模型，加权最小二乘的基本思想是对原模型进行加权，即对较小的残差平方赋予较大的权数，对较大的残差平方赋予较小的权数，使之变成一个新的不存在异方差性的模型，然后再使用不同的最小二乘法估计模型参数。用加权最小二乘法拟合模型得到了模型 2，它有一个非常理想的结果，如表 8 中的数据所示，模型显著，且拟合度达到了 0.9052。表 9 是各变量和常数项的系数估计。

Table 8. The effect of the fit of model 2
表 8. 模型 2 的拟合效果

Residual standard error	2.858 on 678 degrees of freedom
Multiple R-squared (R^2)	0.9063
Adjusted R-squared (调整的 R^2)	0.9052
F-statistic (F 统计量)	819.9on 8 and 678 DF
P-value (P 值)	<2.2e−16

Table 9. Values of fitted coefficients for model 2
表 9. 模型 2 的拟合系数值

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−49.225666	15.446142	−3.187	0.0015**
*X*1	0.126639	0.005268	24.039	<2e−16***
*X*2	0.108608	0.006117	17.755	<2e−16***
*X*3	0.095798	0.007087	13.517	<2e−16***
*X*4	−0.116527	0.023477	−4.963	8.77e−07***
*X*5	0.320649	0.062235	5.152	3.38e−07***
*X*6	0.024671	0.005396	4.572	5.73e−06***
*X*7	0.032837	0.006165	5.327	1.36e−07***
*X*8	0.137415	0.004471	30.736	<2e−16***

我们在测试集上对上面两个模型进行预测，计算模型预测值与真实值之间的均方误差(MSE)，得到的结果如表 10 所示，可以看出模型 1 的均方误差要更小，即预测误差更小，这说明不能只用拟合效果来判断模型的好坏，这可能会出现过拟合的情况。

Table 10. Mean square error values for the two models
表 10. 两个模型的均方误差值

	Model 1	Model 2
MSE	112.7072	116.0546

3.3. 最优子集选择

用最优子集选择法估计线性回归模型参数具有更好的解释性，它实现了最优预测变量子集的筛选，得到了表 11 的结果，第 1 个模型只包含变量 $X1$，第 2 个模型包含变量 $X1$、$X5$，第 3 个模型包含变量 $X1$、$X5$、$X8$，第 4 个模型包含变量 $X1$、$X2$、$X4$、$X8$，第 5 个模型包含变量 $X1$、$X2$、$X3$、$X4$、$X8$，第 6 个模型包含变量 $X1$、$X2$、$X3$、$X4$、$X5$、$X8$，第 7 个模型包含变量 $X1$、$X2$、$X3$、$X4$、$X5$、$X6$、$X8$，第 8 个模型包含所有变量。

Table 11. Models for optimal subset selection
表 11. 最优子集选择的模型

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$
1	“*”	“ ”	“ ”	“ ”	“ ”	“ ”	“ ”	“ ”
2	“*”	“ ”	“ ”	“ ”	“*”	“ ”	“ ”	“ ”
3	“*”	“ ”	“ ”	“ ”	“*”	“ ”	“ ”	“*”
4	“*”	“*”	“ ”	“*”	“ ”	“ ”	“ ”	“*”
5	“*”	“*”	“*”	“*”	“ ”	“ ”	“ ”	“*”
6	“*”	“*”	“*”	“*”	“*”	“ ”	“ ”	“*”
7	“*”	“*”	“*”	“*”	“*”	“*”	“ ”	“*”
8	“*”	“*”	“*”	“*”	“*”	“*”	“*”	“*”

评价模型的指标是 R^2，RSS，调整的 R^2，C_p(AIC)或者 BIC，其中 R^2 和调整的 R^2 越大，模型拟合度越高，RSS，C_p，BIC 越小，模型拟合度越高。从表 12 随着变量个数的增加，各统计量都得到改善，依照每一个统计量的数据选择了最优的子集个数，图 3 中的红色点表示各统计量选择的最优结果。结果显示选择 6 个特征是最优的。

Table 12. Model statistics for optimal subset selection
表 12. 最优子集选择的模型统计量

	R^2	RSS	调整的 R^2	C_p	BIC
1	0.2478	216002.7	0.2471	971.395	−279.471
2	0.3512	186326.8	0.3499	698.980	−424.756
3	0.4818	148827.4	0.4802	354.220	−649.275
4	0.5578	127003.2	0.5560	154.410	−805.670
5	0.6110	111718.1	0.6091	15.067	−930.813
6	0.6140	110843.2	0.6118	8.977	−931.974
7	0.6142	110798.1	0.6115	10.559	−925.456
8	0.6155	110413.2	0.6125	9.000	−922.103

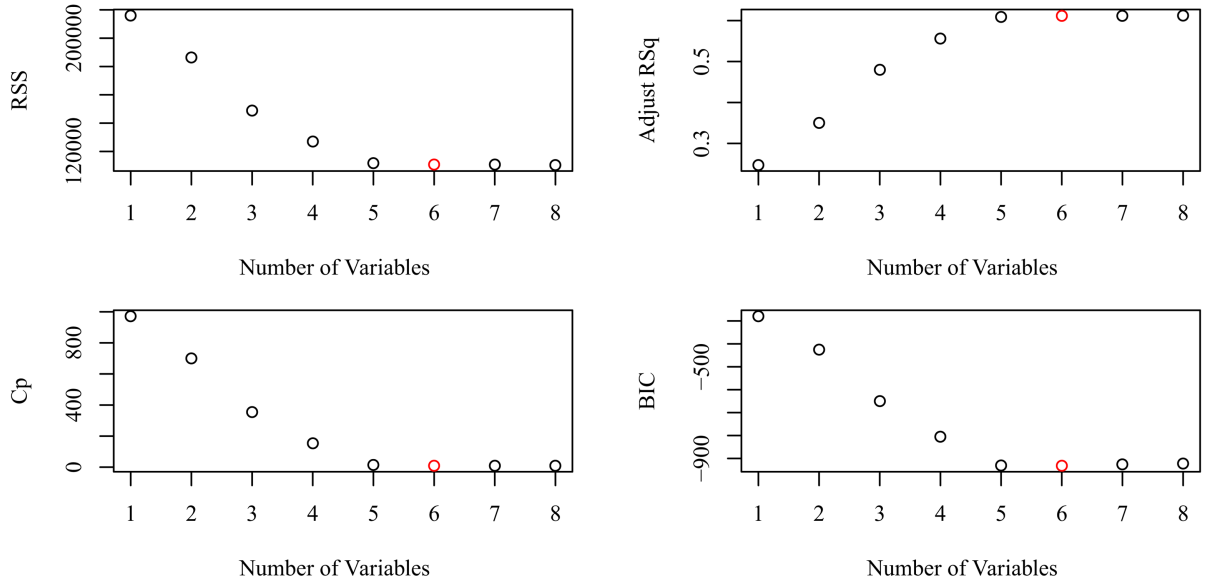

Figure 3. Model selection results for each statistic
图 3. 各统计量的模型选择结果

K 折交叉验证法能够准确估计检验误差，它的基本原理是将观测数据随机地分为 K 个大小一致的组，或者说折。第一折作为测试集，然后剩下的 K-1 折上拟合模型，重复这个步骤 K 次，每一次在不同的 K-1 折上拟合模型，每一次拟合都会得到一个均方误差，K 折的 CV 估计就是这些均方误差再求均值得到。用 K 折交叉验证法做最优子集选择，得到的结果是选择前 6 个特征的均方误差为 109.1655，是最优的模型。

对整个数据集用最优子集选择，获得 5 个变量的参数估计结果汇总在表 13 中，所以我们通过最优子集法得到的模型是：

$$\text{Model } 3: Y = 28.9930 + 0.1054 * X1 + 0.0865 * X2 + 0.0687 * X3 - 0.2181 * X4 + 0.2403 * X5 + 0.1135 * X8$$

Table 13. Coefficients of the variables of the optimal subset method
表 13. 最优子集法的各变量系数

Intercept	*X*1	*X*2	*X*3	*X*4	*X*5	*X*8
28.9930	0.1054	0.0865	0.0687	-0.2181	0.2403	0.1135

3.4. 压缩估计

3.4.1. 岭回归

在做岭回归时，同样也是考虑使用交叉验证来调节参数 λ，要注意的是使用岭回归要先对数据进行标准化，这里使用 glmnet()函数在默认设置下，已经将所有变量进行了标准化。图 4 是 10 组子样本的交叉验证绘制 CV 曲线图寻找最佳的惩罚函数系数 λ。采用最小 MSE 取值 113.5488 对应的 $\lambda_{\min} = 0.8129$，对应 8 个解释变量。

用最优的 λ 建立岭回归模型，得到的变量系数见表 14，所以我们建立第 4 个模型：

$$\text{Model } 4: Y = 65.3414 + 0.0830 * X1 + 0.0606 * X2 + 0.0354 * X3 - 0.2360 * X4 + 0.3640 * X5 - 0.0106 * X6 - 0.0170 * X7 + 0.1051 * X8$$

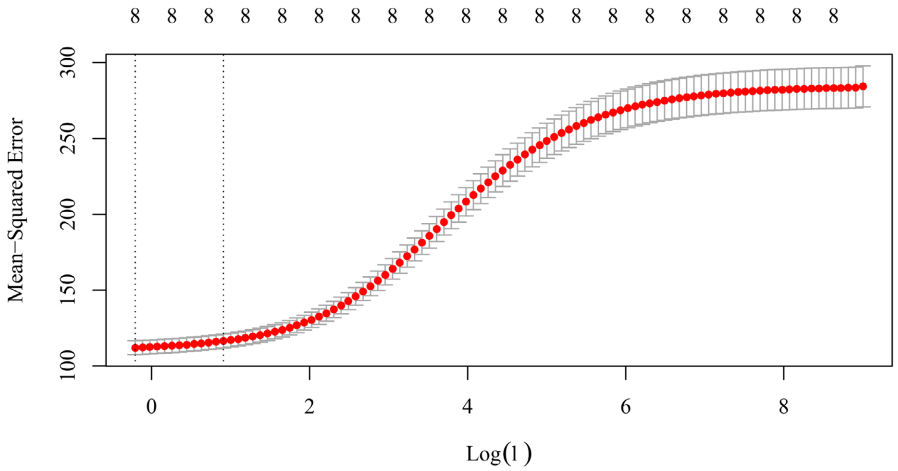

Figure 4. CV cross-validation plot for ridge regression
图 4. 岭回归的 CV 交叉验证图

Table 14. Ridge regression model coefficients
表 14. 岭回归模型系数

Intercept	*X*1	*X*2	*X*3	*X*4	*X*5	*X*6	*X*7	*X*8
65.3414	0.0830	0.0606	0.0354	−0.2360	0.3640	−0.0106	−0.0170	0.1051

3.4.2. Lasso 回归

与岭回归相比，当最小二乘回归出现较大方差时，Lasso 以牺牲偏差为代价去降低方差，从而得到更为精确的预测结果。Lasso 的优势是：Lasso 的系数估计是稀疏的，它可以实现特征选择，因此得到的模型更容易解释。我们同样是使用 10 折交叉验证来调节最优的参数 λ，图 5 是 10 组子样本的交叉验证绘制 CV 曲线图寻找最佳的惩罚函数系数 λ。可以看到模型实现了特征选择，在只包含 6 个变量时模型就可以达到最优了。最小的均方误差是 112.6013，它对应的 λ 是 $\lambda_{\min} = 0.0076$。

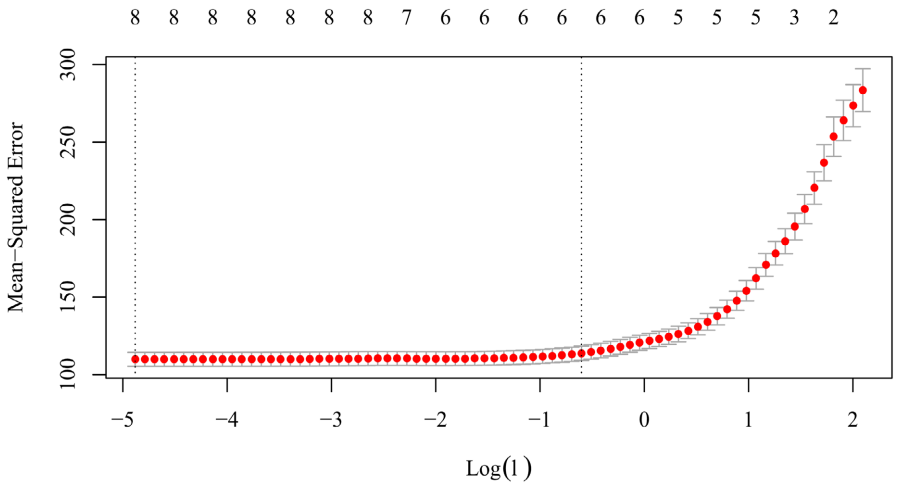

Figure 5. Lasso's cross-validation plot
图 5. Lasso 的交叉验证图

用最优的 λ 建立岭回归模型，得到的变量系数见表 15，所以我们建立第 6 个模型：

$$\text{Model } 5: Y = -13.1773 + 0.1166 * X1 + X2 + 0.0834 * X3 - 0.1625 * X4 + 0.2867 * X5 + 0.0147 * X6$$

Table 15. Ridge regression model coefficients
表 15. 岭回归模型系数

Intercept	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
−13.1773	0.1166	1.0000	0.0834	−0.1625	0.2867	0.0147

3.5. 模型的选择

综合上述的 5 个模型，表 16 是各模型的均方误差对比结果，可以看到选择第三个模型是最优的。

Table 16. Comparison of mean square error of models
表 16. 模型的均方误差对比

	MSE
Model 1	112.7072
Model 2	116.0546
Model 3	109.1655
Model 4	113.5488
Model 5	112.6013

4. 结论

本文对混凝土的抗压数据做线性回归分析，首先对数据集进行划分，然后采用多种参数估计方法来估计模型的系数，最后通过预测均方误差来选择最优的模型。

在用最经典的最小二乘法估计系数时发现模型的拟合度并不高，因此对多元线性回归模型进行模型诊断，首先考虑了误差项的异方差性，第一种解决办法是对因变量取对数后再进行简单的最小二乘回归方法进行拟合，发现模型的拟合效果和预测效果都没有得到提升；第二种解决办法是使用加权最小二乘回归方法进行拟合，得到了很高的拟合度，预测效果并不如简单的最小二乘回归方法，但是相差不是很大，所以相比较之下也可以考虑使用模型 2 来拟合数据。

考虑各变量之间可能存在共线性，解决办法是做最优子集选择。最优子集选择是通过选择最优或者说对模型有很大贡献率的特征子集来提高模型的拟合度的，选择最优模型参考的统计量是 C_p，AIC，BIC 和调整的 R^2，在训练模型是时候使用交叉验证法提高模型的准确度。得到的结果是：选择 6 个变量作为最优的模型的预测变量，计算模型的均方误差发现模型的预测效果还是有所提升的。

岭回归和 Lasso 都属于压缩估计的方法，它们通过对系数进行约束或加惩罚项的技巧对模型进行拟合，就是说将系数的估计值往 0 的方向压缩，这会提升拟合效果，Lasso 的另一个优势是可以做特征选择，拟合结果显示，模型的预测均方误差并没有下降非常的多。

综合所有方法得到的结论是压缩估计都没有很好地提高模型的预测效果，最优的模型是做子集选择。

当然，模型还有很多改进的空间，由于本文使用的数据集的特征个数并不算多，做数据降维的效果也并不理想，没办法达到降维的效果。但是各特征之间的关系还是可以再进一步考虑的，比如相关性很强的两个特征可以通过线性组合等方式把一个两个特征变成一个新的特征。

参考文献

[1] 桂许兰, 詹微微, 龙行航. 水泥基超高性能混凝土强度影响因素综述[J]. 湖南交通科技, 2020, 46(3): 12-15.

[2] 杨占昱. 混凝土强度影响因素的分析与试验[J]. 四川建材, 2018, 44(6): 9-10.

[3] 贺慧芳. 轻质高强混凝土之强度影响因素[J]. 黑龙江科技信息, 2016(32): 243.

[4] YehI-Cheng. 利用实验和神经网络设计混凝土强度分析[J]. 土木工程材料学报, 2006, 18(4): 597-604.

[5] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7