浙江大学数学科学学院

前沿数学专题讨论 PR04

9 Nov 2022

Submitted To:
张南松老师
22-23 秋冬学期

Submitted By :
吴凡
农业工程 + 统计学

Contents

1 背景

1.1 线性模型的最小二乘估计

考虑线性模型：

$$y = X\beta + \epsilon, E(\epsilon) = 0, Cov(\epsilon) = \sigma^2 I \tag{1}$$

的参数 β 和 σ^2 的估计问题，y 为 n*1 观测向量，X 为 n*p 的设计矩阵，β 是 p*1 未知参数向量，ϵ 为随机误差，σ^2 为误差的方差。

传统回归分析估计向量 β 的基本方法为最小二乘法，其思想是使得误差向量 $\epsilon = y - X\beta$ 尽可能的小, 使

$$Q(\beta) = ||\epsilon||^2 = ||y - X\beta||^2 = (y - X\beta)'(y - X\beta) \tag{2}$$

达到最小，解得

$$\hat{\beta} = (X'X)^{-1}X'y \tag{3}$$

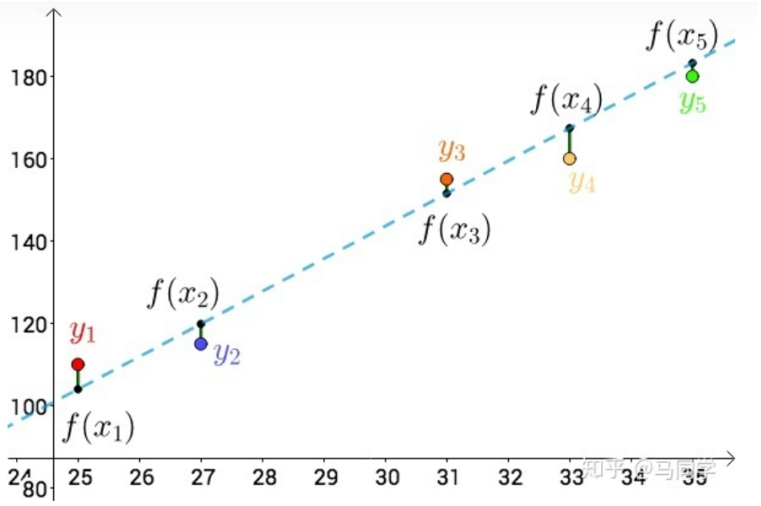
在传统的数据分析场景里，我们通常接触的数据大部分都是 n>p，即样本数大与变量数。此时，若 rank(X)=p, 则 X'X 可逆，这时 $\hat{\beta}$ 是 β 的无偏估计。

考察参数估计量的统计性质是衡量估计量好坏的主要准则。一个用于考察总体的估计量，可以从六个方面考察其优劣性。

估计量的统计性质	具体含义
线性	其是否是另一个随机变量的线性函数。
无偏性	其均值或期望是否等于总体的真实值。
有效性	其发生在所有的线性无偏估计量中具有最小方差。
渐近无偏性	样本容量趋于无穷大时，其均值序列趋于总体的真值。
一致性	样本容量趋于无穷大时，其是否依概率收敛于总体的真值。
渐近有效性	样本容量趋于无穷大时，其在所有的一致估计量中具有最小的渐近方差。

最小二乘估计有许多优良性质：线性、无偏性、最小方差性。

著名的高斯-马尔可夫 (Gauss-Markov) 定理：在古典假设条件下，最小二乘估计是最佳线性无偏估计量。



1.2 岭估计的提出

随着大数据的兴起，数据采集能力的指数级提升，大量的数据集出现了变量数多余样本数的情况，即 $p > n$。此时矩阵 X 出现多重共线性的情况，X'X 不可逆，因而没法用传统的 OLS（最小二乘估计）方法。

为了解决这一问题，就有了岭回归（Ridge Regression）方法，简单来说，岭回归就是在前面最小化目标函数 $Q(\beta)$ 的后面加了一个 2-范数的平方：

$$Q(\beta) = ||y - X\beta||^2 + \lambda||\beta||_2^2 \tag{4}$$

$$\Leftrightarrow argmin||y - X\beta||^2 \quad s.t. \sum \beta_j^2 \leq s \tag{5}$$

上式求解可得到 β 的岭估计:

$$\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1}X'y \tag{6}$$

从上面我们可以明显看到参数 λ 保证了 $X'X + \lambda I$ 满秩、可逆，当然也由于其加入，此时的 $\hat{\beta}(\lambda)$ 是一个有偏估计。

岭估计有以下性质：

1. 岭估计 $\hat{\beta}(\lambda)$ 是 β 的有偏估计.
2. 岭估计 $\hat{\beta}(\lambda)$ 是最小二乘估计 $\hat{\beta}$ 的一个线性变换.

3. $\forall k > 0$, 若 $||\hat{\beta}|| \neq 0 \Longrightarrow ||\hat{\beta}(\lambda)|| < ||\hat{\beta}||$, 岭估计是一个压缩的有偏估计。

4. 存在 $\lambda > 0$, 使得在均方误差意义下, 岭估计优于最小二乘估计, 即 $MSE(\hat{\beta}(\lambda)) < MSE(\hat{\beta})$

关于 λ 的选择，有 Horel-Kennard 公式、岭迹法、交叉验证法.....

Horel-Kennard 公式:

Hoerl 和 Kennard 提出的选择岭参数 k 的公式为

$$\hat{k} = \frac{\hat{\sigma}^2}{\max\limits_i \hat{\alpha}_i^2} .$$

注意到，理论上的最优岭参数是下列方程的解：令 $f'(k) = 0$，则有

$$f'(k) = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} = 2$$
$$\sum_{i=1}^{p} \frac{\lambda_i \left(k\alpha_i^2 - \sigma^2\right)}{(\lambda_i + k)^3} = 0 .$$

若 $k\alpha_i^2 - \sigma^2 < 0$ 对 $i = 1, 2, \cdots, p$ 都成立，则 $f'(k) < 0$，于是取

$$k^* = \frac{\hat{\sigma}^2}{\max\limits_i \hat{\alpha}_i^2} ,$$

当 $0 < k < k^*$ 时，$f'(k) < 0$ 恒成立，因而 $f(k)$ 在 $(0, k^*)$ 上是单调递减函数。再由 $f(k)$ 是 $(0, k^*)$ 上的连续函数得到 $f(k^*) < f(0)$。用 $\hat{\alpha}_i$ 和 $\hat{\sigma}^2$ 代替 α_i 和 σ^2 即可得到我们需要的岭参数 $\hat{k}$。

2 LASSO

LASSO 全名 least absolute shrinkage and seletion operator, 最小收缩算子法。他最大的特点就是引入了惩罚项 ，这个参数可以对模型变量进一步筛选，使模型不至于过于复杂，从而提高其泛化能力。

与岭回归类似，Lasso 就是在目标函数 $Q(\beta)$ 后面加了一个 1-范数

$$Q(\beta) = ||y - X\beta||^2 + \lambda||\beta||_1 \tag{7}$$
$$\Leftrightarrow argmin||y - X\beta||^2 \quad s.t. \sum |\beta_j| \leq s \tag{8}$$

三种估计量的比较：

$$\text{Linear Regression}: \sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$$

$$\text{Lasso Regression}: \sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

$$\text{Ridge Regression}: \sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}{\beta_j}^2$$

关于 LSAAO 在岭估计之后提出，为什么加 1-范数的 LASSO 没有加 2-范数的岭回归早，可能的原因是 1-范数作为绝对值之和不方便求导。

2.1 Lasso 的优势

为什么至今 Lasso 仍长青不老?

因为它可以解决现在高维数据一个普遍问题——稀疏性。高维数据即 p>n 的情况，现在随着数据采集能力的提高，变量（也叫特征）数采集的多，但是其中可能有很多特征是不重要的，系数很小，如果用岭回归，可能这个不重要的变量也给你估出来了，而且可能还不小，而用 Lasso 方法，就可以把这些不重要变量的系数压缩为 0，既实现了较为准确的参数估计，也实现了变量选择（降维）。

以 Lasso 始祖 Tibshiran 在其著作中举的 p=2 的情形为例：

图中青色部分表示两种方法的约束。

从图中可以看出，Lasso 方法与之相交的地方恰为 $(0,\beta_2)$，而从图中也可以看出 $\hat{\beta}$ 所处的位置本就是 β_2 大，β_1 小，我们取 Lasso 的结果，意味着 β_1 的系数被压缩到了 0。

2.2 Lasso 最优解

Lasso 因为其约束条件（也有叫损失函数的）不是连续可导的，因此常规的解法如梯度下降法、牛顿法、就没法用了。目前常用的方法有：坐标轴下降法与最小角回归法。

坐标轴下降法:

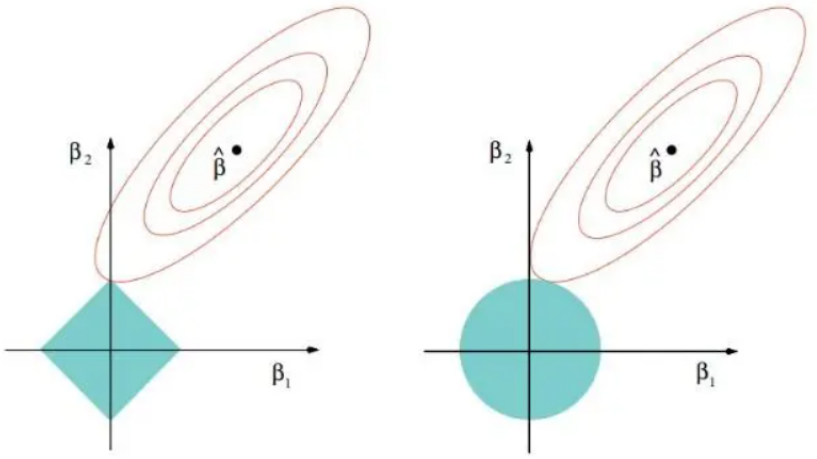
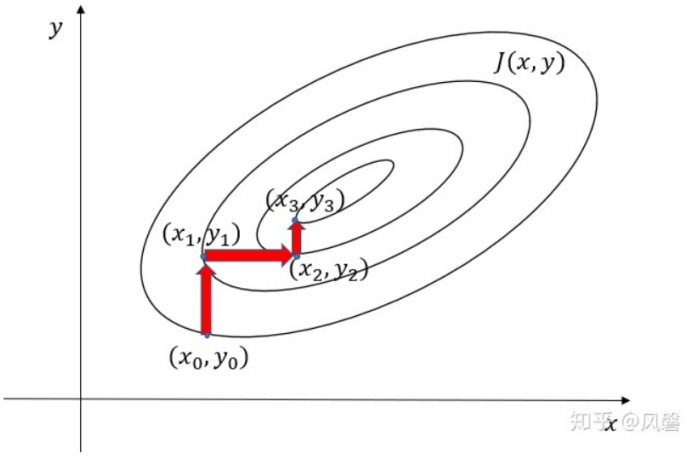


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

坐标轴下降法是一种迭代算法，与梯度下降法利用目标函数的导数来确定搜索方向不同，坐标轴下降法是在当前坐标轴上搜索函数最小值，不需要求目标函数的导数。

以 2 维为例，设损失函数为凸函数 J(x,y)，在初始点固定 x_0，找使得 J(y) 达到最小的 y_1，然后固定 y_1，找使得 J(x) 达到最小的 x_2，这样一直迭代下去，因为 J(x,y) 是凸的，所以一定可以找到使得 J(x,y) 达到最小的点 (x_k, y_k)。



2.3 ElasticNet 模型

ElasticNet 模型，弹性网络模型，是结合了 lasso 和 ridge regression 的模型。

ElasticNet 模型的目标函数 $Q(\beta)$

$$Q(\beta) = ||y - X\beta||^2 + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] \tag{9}$$

弹性网络惩罚由 α 控制，当 $\alpha = 1$ 时为 Lasso 模型，当 $\alpha = 0$ 时为 Ridge 模型.

弹性网络在很多特征互相联系的情况下是非常有用的。Lasso 很可能只随机考虑这些特征中的一个，而弹性网络更倾向于选择两个。在实践中，Lasso 和 Ridge 之间权衡的一个优势是它允许在循环过程（Under rotate）中继承 Ridge 的稳定性。

3 算法实例

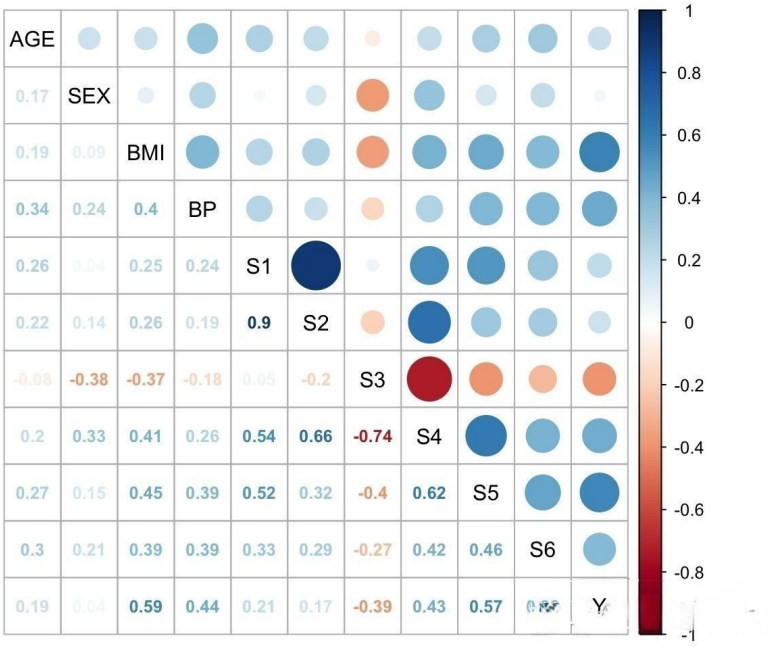
3.1 使用 R 进行 Ridge 回归

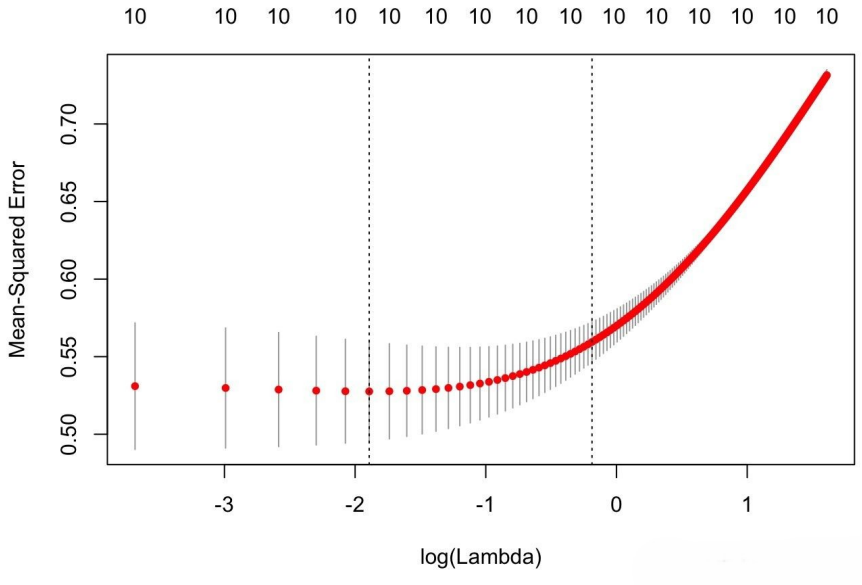
使用的数据集为糖尿病病情数据集（diabetes.csv）。糖尿病病情数据集包含从 442 例糖尿病患者中获得的十个变量：年龄（AGE）、性别（SEX）、体重指数（BMI）、平均血压（BP）和六个血清测量值（S1-S6），以及一个我们感兴趣的因变量 Y。得到的可视化相关系数如图 1 所示：

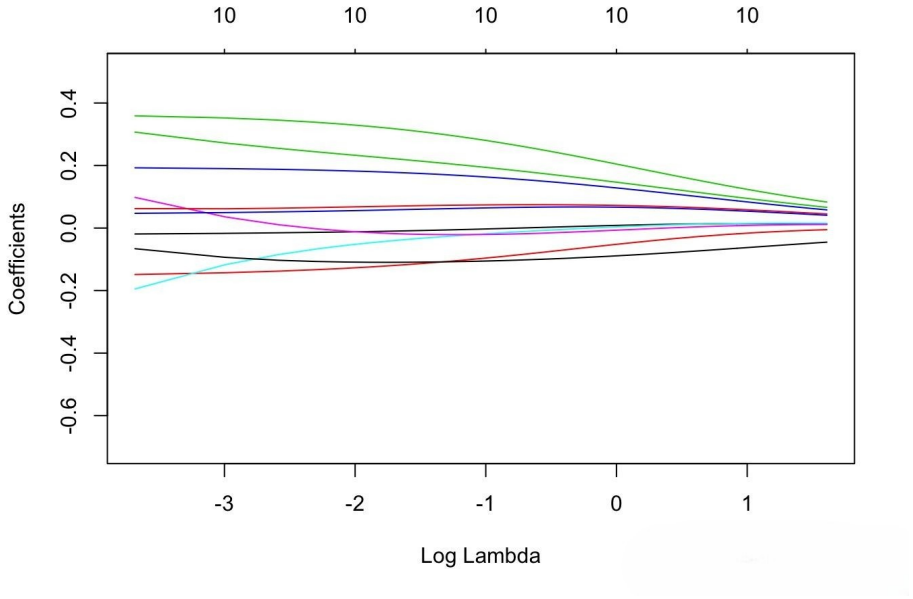
数据相关性分析结果表明，因变量 Y 和 AGE、SEX、S1、S2 四个变量的相关系数较小，和 S3 是负相关，与其余变量的相关系数较大，且都是正相关；S1 和 S2 之间的正相关性较强，S3 和 S4 的负相关性较强。

在 Ridge 回归模型中，需要指定一个合适的参数 lambda，该参数为施加在回归系数上的惩罚系数，不同的参数 lambda 可以得到不同的 Ridge 回归模型。glmnet 包中的 cv.glmnet() 可以通过交叉验证的方式，来分析在不同的参数 lambda 下回归模型的效果。下面使用 cv.glmnet() 函数和训练数集分析模型参数的影响。

通过建立 Ridge 回归，在不同的 lambda 下，变量合数、各个自变量回归系数的变化如下图所示：


AGE
SEX
BMI
BP
S1
S2
S3
S4
S5
S6
Y
0.17
0.19 0.09
0.34 0.24 0.4
0.26 0.25 0.24
0.22 0.14 0.26 0.19 0.9
-0.08 -0.38 -0.37 -0.18 -0.2
0.2 0.33 0.41 0.26 0.54 0.66 -0.74
0.27 0.15 0.45 0.39 0.52 0.32 -0.4 0.62
0.3 0.21 0.39 0.39 0.33 0.29 -0.27 0.42 0.46
0.19 0.59 0.44 0.21 0.17 -0.39 0.43 0.57
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
Mean-Squared Error
0.50
0.55
0.60
0.65
0.70
-3
-2
-1
0
1
log(Lambda)



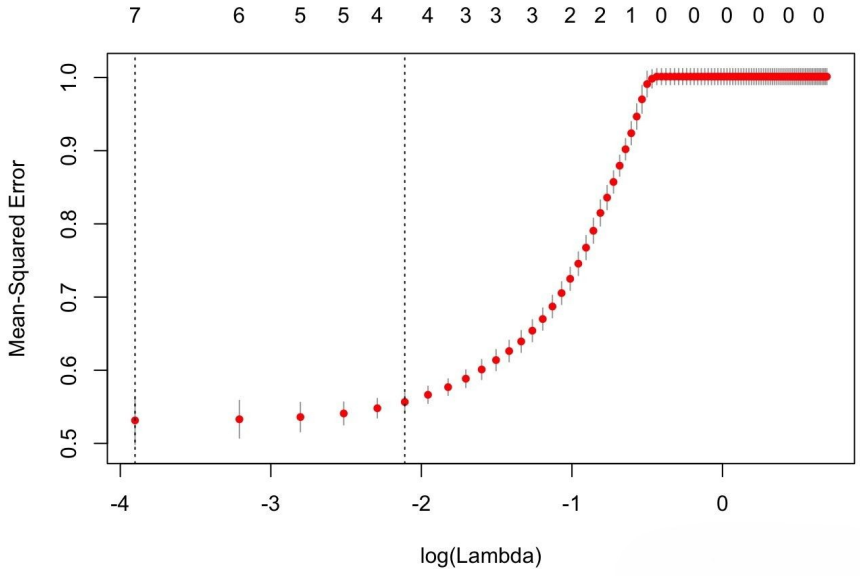
结果表明，不存在系数为 0 的自变量，这是因为 Ridge 模型不容易把变量的系数压缩到 0，Ridge 回归模型并没有自动进行变量选择的能力。为了验证 Ridge 回归模型的预测效果，可在测试集上进行预测，并计算出平均绝对值误差的大小。通过 predict() 函数得到测试集的预测值后，再使用 Metrics 包的 mae() 函数计算平均绝对值误差。通过将预测值进行逆标准化变换，与标准化前的因变量进行比较发现，Ridge 回归模型的预测平均绝对值误差为 45.99。

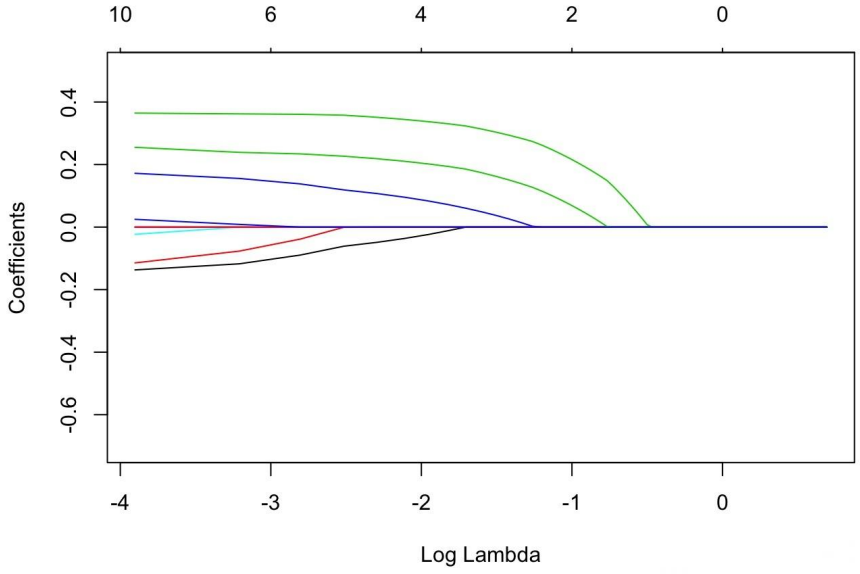
3.2 使用 R 进行 Lasso 回归

使用同样的数据集进行 Lasso 回归分析。

使用 cv.glmnet() 函数，利用交叉验证的方式，分析不同的参数 lambda 下 Lasso 回归模型的效果。在得到交叉验证后的模型 lasso model 后，使用 plot() 函数可视化不同参数下的均方误差以及各个自变量的回归系数变化情况，结果下图所示。

随着参数 lambda 值的增加，Lasso 回归使用的自变量数目在减少，同时模型的预测误差在增大。从 Lasso 回归模型的系数可以发现，AGE、S2、S4 三个自变量的回归系数等于 0，说明模型将这 3 个对因变量影响不显著的特征剔除了。


7 6 5 5 4 4 3 3 3 2 2 1 0 0 0 0 0 0
Mean-Squared Error
0.5 0.6 0.7 0.8 0.9 1.0
-4 -3 -2 -1 0
log(Lambda)


10 6 4 2 0
Coefficients
-0.6 -0.4 -0.2 0.0 0.2 0.4
-4 -3 -2 -1 0
Log Lambda

通过 predict() 函数得到测试集的预测值后，再使用 Metrics 包的 mae() 函数计算平均绝对值误差。通过将预测值进行逆标准化变换，与标准化前的因变量进行比较发现，Lasso 回归模型的预测平均绝对值误差为 46.22。结果发现，Lasso 回归预测的误差比 Ridge 回归大了一点，这是因为 Lasso 回归使用了更少的自变量来建立回归模型。虽然误差稍微变大，但是 Lasso 使用更少的特征，使模型更稳定，降低了模型的复杂度。

3.3 ElasticNet 模型

ElasticNet 模型，弹性网络模型，是结合了 lasso 和 ridge regression 的模型。弹性网络惩罚由 α 控制，当 $\alpha = 1$ 时为 Lasso 模型，当 $\alpha = 0$ 时为 Ridge 模型.