Linear-based and other applicable methods for riboflavin dataset

Chenkai Wang

Abstract

This paper investigates the use of regression techniques for solving a current problem. We begin by performing data preprocessing and then introduce a range of linear-based methods. We also examine dimensionality reduction techniques, including PCR and the ensemble method, random forest. Based on the findings from the various model, we propose and evaluate the best linear model. The assessment and final models are given in the end.

Keywords: Linear-based methods, Random Forest, PCR, Cross Validation

SUSTech Southern University of Science and Technology

Contents

1 Introduction of Data Set and Problem

The data set is about riboflavin production by Bacillus subtilis, which contains $n = 71$ observations of $p = 4088$ predictors (gene expressions) and a one-dimensional response variable(riboflavin production) available at [1]. In the rest of this paper, we use the terms predictors and features interchangeably. Let Y_i and $\hat{Y}_i$ denote the response value and estimated value of $i's$ observation, respectively and let $X_{i,j}$ be the $j's$ component of observation i, $i = 1, \cdots, n, j = 1, \cdots, p$. The main purpose of this essay is to estimate a function m

$$\hat{Y}_i = m(X_{i,1}, X_{i,2}, \cdots, X_{i,p}) + \epsilon_i$$

such that minimizes

$$\sum_i \left(Y_i - \hat{Y}_i\right)^2$$

where

$$\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2) \text{ and } i = 1, 2, \cdots, n.$$

On the one hand, based on the work of other scholars[5], it is believed that $m(\bullet)$ be well approximated by a linear model; on the other hand, we can fit with a linear model first, and then perform residual analysis to verify the rationality of the linear model. In addition, we also used dimensionality reduction methods and other machine learning algorithms to try to fit $m(\bullet)$.

In the following sections of this paper, we present a comprehensive analysis of regression techniques for solving the current problem. After performing data preprocessing in section two, we introduce various linear-based methods in section three. We then explore PCR, a dimensionality reduction technique, and the ensemble method, random forest. In section five, we demonstrate the use of these methods and present our proposed model, the best linear model, based on the findings. Finally, in section six, we evaluate and select the best model, summarizing the various models generated throughout the study.

2 Data Preprocessing

We first draw the box plot, histogram plot, and the fitted curve of riboflavin production(q_RIBFLV), which are shown below.

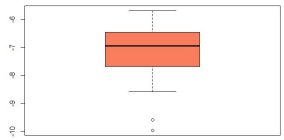

Figure 1: box plot

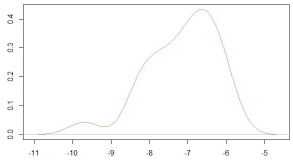

Figure 2: fitted curve

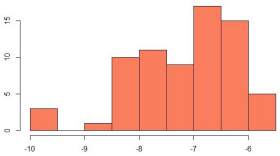

Figure 3: histogram

The data shows that the median value of riboflavin production is -7 and approximately follows a skewed distribution. Then we draw the mean of $p = 4088$ features.

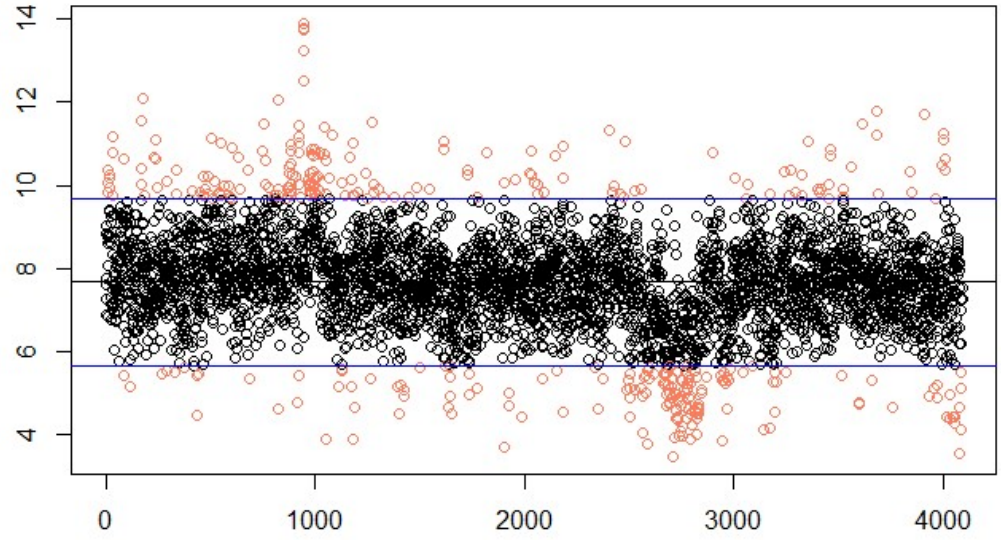

Figure 4: mean of features

The mean of all features $M \approx 7.67$, Additionally, most features are in $[M-1, M+1]$.

To apply linear-based methods, we standardize the data by

$$X_{i,j}^{\text{new}} = \frac{X_{i,j} - \mu_j}{s_j}$$

where

$$\mu_j = \frac{\sum_{i=1}^{n} X_{i,j}}{n} \text{ and } s_j = \sqrt{\frac{\sum_{i=1}^{n} (X_{i,j} - \mu_j)}{n}}$$

Meanwhile, we standardize Y_i similarly.

3 Linear Regression With Varies Penalties

We start with linear regression. Suppose there exists a linear relationship between the response variable and predictors. Let

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The model can be written as

$$Y = X\beta + \epsilon.$$

Let $Q_1(\beta)$ denote the loss function of linear regression, then

$$Q_1(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta)$$

Minimize the equation above by taking the derivative of Q_1 with respect to β yields $\hat{\beta}$, which is

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

The problem is, when $p >> n$, $X'X$ is not invertible, which makes $\hat{\beta}_{\text{LSE}}$ incalculable. We will demonstrate different linear-based techniques to fix it in the rest of this section.

3.1 Lasso Regression

For simplicity, compared with linear regression, lasso regression[6] add an additional L_1 norm penalty term in the loss function. More specifically, let $Q_2(\beta)$ be the loss function of lasso regression, then

$$Q_2(\beta) = \|Y - X\beta\|^2 + \lambda_{lasso}\|\beta\|_1$$

which is equivalent to $\arg\min \|Y - X\beta\|^2$ s.t. $\sum_{j=1}^{p} |\beta_j| \leq t_{lasso}$

Minimizing $Q_2(\beta)$, we can get

$$\hat{\beta}_{\text{lasso}} = \begin{cases} \hat{\beta}_{LSE} - \tilde{\lambda}, & \text{if } \hat{\beta}_{LSE} \geq \tilde{\lambda} \\ 0, & \text{if } -\tilde{\lambda} \leq \hat{\beta}_{LSE} < \tilde{\lambda} \\ \hat{\beta}_{LSE} + \tilde{\lambda}, & \text{if } \hat{\beta}_{LSE} \leq -\tilde{\lambda} \end{cases}$$

where $\tilde{\lambda} = \frac{\lambda_{lasso}}{2}$.

Based on the result above, we can see that lasso regression can select variables by setting the smaller to 0. The hyperparameter λ_{lasso} can be obtained by cross validation, which is explained comprehensively in section five.

3.2 Ridge Regression

The loss function $Q_3(\beta)$ for ridge regression[4] is

$$Q_3(\beta) = \|Y - X\beta\|^2 + \lambda_{ridge}\|\beta\|_2^2$$

which is equivalent to $\arg\min \|Y - X\beta\|^2$ s.t. $\sum_{j=1}^{p} \beta_j^2 \leq t_{ridge}$. Minimizing $Q_3(\beta)$, we can get

$$\hat{\beta}_{\text{ridge}} = \left(X'X + \lambda_{ridge} I\right)^{-1} X'Y$$

We will also use cross validation method to determine the hyperparameter λ_{ridge} in section five.

3.3 Elastic Net

Elastic net[8] is a combination of ridge and lasso. More specifically, the loss function $Q_4(\beta)$ is

$$Q_4(\beta) = \|Y - X\beta\|^2 + \lambda_{net}\left\{0.5(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right\}$$

where $0 \leq \alpha \leq 1$; if $\alpha = 1$, it is lasso and $\alpha = 0$ ridge.

3.4 Least Angle Regression

Least angle regression(LARS)is an algorithm for fitting linear regression models to high-dimensional data[3]. It is a type of forward stepwise regression,

which starts with no variables in the model and adds variables one at a time based on the strength of their association with the response variable.

The specific steps for implementation are as follows:

- Step1: Initialize the set of variables to be empty, and set the current response variable to be the original response variable minus the mean of the response variable:

$$S = \emptyset$$
$$Y^{(current)} = Y - \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  where S is the set of variables, $Y^{(current)}$ is the current response variable, y is the original response variable, and n is the number of observations.

- Step2: Calculate the correlation between each explanatory variable and the current response variable:

$$\rho_j = \frac{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_{i,j} - \overline{X}_j)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

  where ρ_j is the correlation between the jth feature and the current response variable.

- Step3: Select the explanatory variable with the largest correlation with the current response variable and add it to the set of variables:

$$j^* = \arg\max_{j \in S} |\rho_j|$$
$$S = S \cup j^*$$

  Where j^* is the index of the explanatory variable with the largest correlation with the current response variable, and S is the set of variables.

- Step4: Update the current response variable to be the original response variable minus the mean of the current model's predicted values:

$$Y^{(current)} = Y - \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$$

where $Y^{(current)}$ is the current response variable, Y is the original response variable, $\hat{Y}_i$ is the predicted value for the ith observation, and n is the number of observations.

- Step5: Repeat steps 2-4 until all explanatory variables have been added to the set of variables or the absolute value of the current response variable is less than a threshold: while $|Y^{(curr)}| >$ threshold and $|S| < p$
- Step6: Use the set of variables to fit a linear regression model:

$$\hat{\beta}_{\text{LARS}} = (X_S^T X_S)^{-1} X_S^T Y$$

where $\hat{\beta}_{\text{LARS}}$ is the vector of estimated coefficients, X_S is the matrix of explanatory variables in the set S, and Y is the response variable.

LARS has the advantage of handling high-dimensional data, particularly in cases where $n << p$, and it does not require the explanatory variables to be orthogonal (uncorrelated). It can also mitigate the effects of multicollinearity, making it a valuable tool for regression analysis.

4 Other applicable Methods

4.1 Principal Component Regression

Principal Component Regression, abbreviated as PCR, is closely related to principal component analysis(PCA). Here we only provide a general introduction and overview of the method, and the technical details can be found here[7]. PCA is a statistical technique for reducing the dimension of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data.

Let $Z_1, \cdots, Z_K$ represent a linear combination of original features, i.e.

$$Z_k = \sum_{j=1}^{p} \phi_{jk} X_j$$

where $\phi_{1m}, \cdots, \phi_{1m}$ are elaborately designed constants and $X_j = [X_{j,1}, \cdots, X_{j,p}]'$. These Z_ks are called principal components, which can be used to fit a linear

model:

$$\hat{Y}_i = \beta_0 + \sum_{k=1}^{K} \beta_k Z_{i,k} + \epsilon_i,\ i = 1, \cdots, n$$

Selecting a suitable handmade hyperparameter K can effectively address the issue of $n >> p$.

In short, the procedure of PCR is concluded in the following graph.

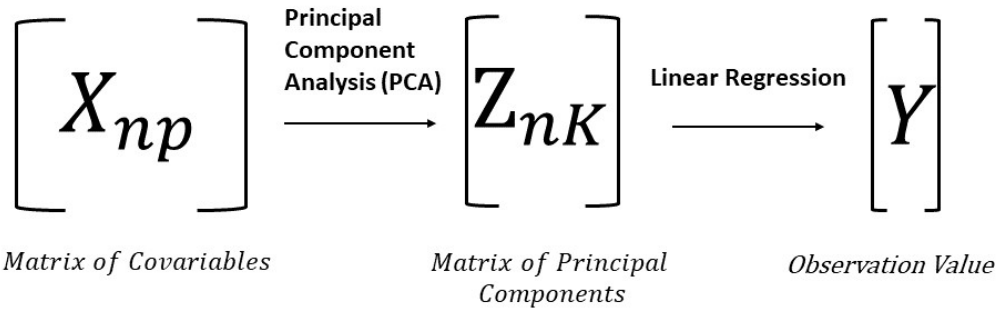

Figure 5: The procedure of PCR

4.2 Random Forest

Random forest[2] is a machine-learning algorithm that can be used for classification and regression tasks. They are an ensemble method, which means they are made up of multiple individual models that work together to make a final prediction.

In the case of random forests, the individual models are decision trees. A decision tree is a flowchart-like tree structure that makes predictions based on the values of input features. A random forest builds many decision trees and aggregates their predictions to make a final prediction.

One of the critical features of random forests is that they use a random subset of the features at each split in the decision tree. This means that each decision tree in the forest is slightly different, making different predictions. The final prediction is made by averaging the predictions of all the individual decision trees.

Random forests are widely used because they are relatively simple to implement and can handle high-dimensional and multicollinear data well. They are also resistant to overfitting, which means that they generally do not perform as well on the training data as on test data. This makes them a good choice for many types of machine-learning tasks.

To sum up, the random forest construct procedure is shown in the following figure.

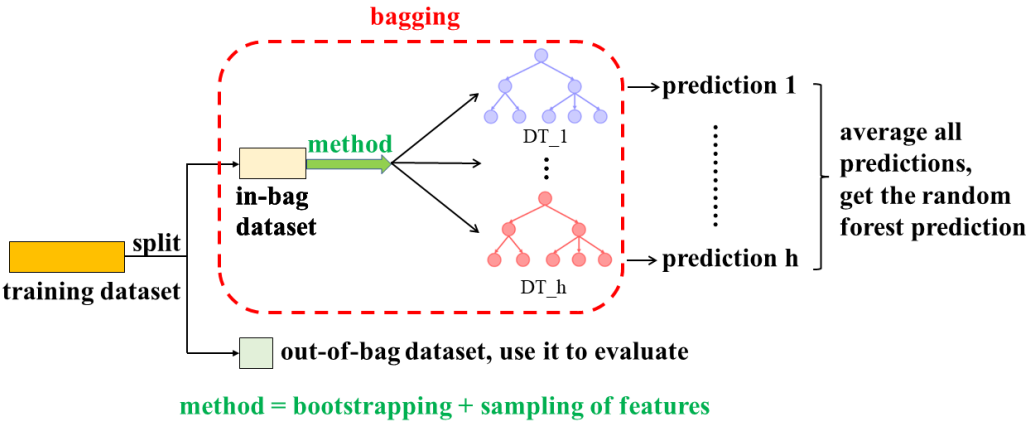

Figure 6: The procedure of random forest

5 Core Procedures and Linear Regression Analysis

5.1 Core Procedure

The core procedures to fit the model are shown below.

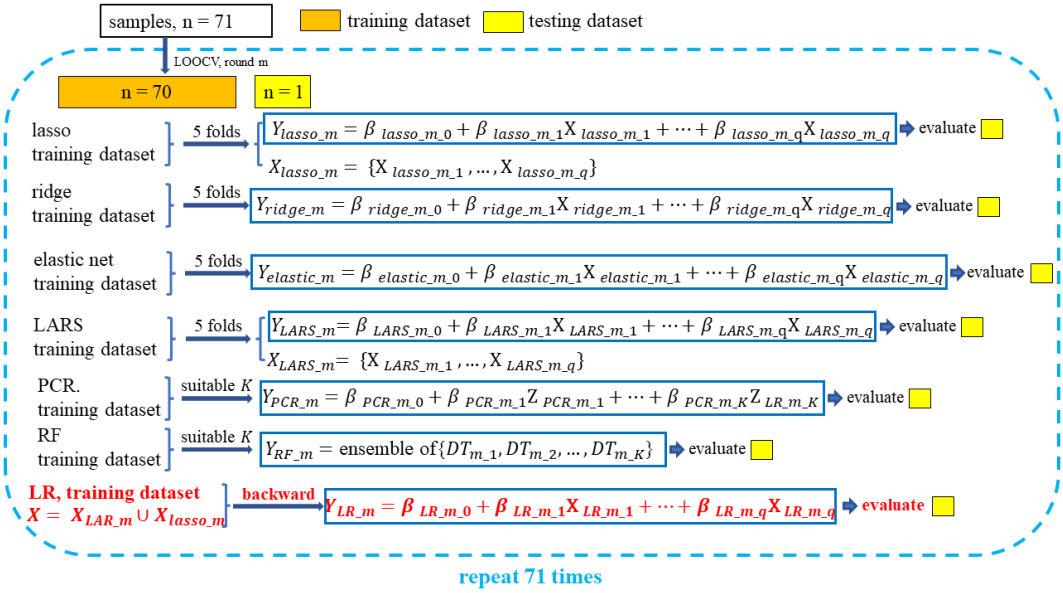

Figure 7: core procedure to fit the model

In short, the core procedures can be summarized in three steps:

- Step1: Use leave one out cross validation(LOOCV) to separate training dataset and testing dataset
- Step2: For each training dataset, apply the methods in section three to fit the specific model and evaluate it on the testing dataset
- Step3: Use the specific features to fit and evaluate a linear model

These procedures are implemented primarily through the "scikit-learn" package in python. The rest of this section is a detailed explanation of steps two and three. In step two, we need to fit different models based on different methods. In round m, we have specific training and testing dataset. Take the lasso as an example. We run it on the training set via five folds cross validation, which yields the best λ_{lasso} and the model.

$$Y_{\text{lasso\_m}} = \beta_{\text{lasso\_m\_0}} + \beta_{\text{lasso\_m\_1}} X_{\text{lasso\_m\_1}} + \cdots + \beta_{\text{lasso\_m\_q}} X_{\text{lasso\_m\_q}} + \epsilon$$

Similarly, we do it for ridge, elastic net, least angle regression, respectively. Moreover, we need to find the specific feature for linear regression in step 3. We know that lasso and least angle regression can select important features. Based on this property, we use set

$$X_{\text{lasso\_m}} = \{X_{\text{lasso\_m\_1}}, \cdots, X_{\text{lasso\_m\_q}}\}$$

and

$$X_{\text{LARS\_m}} = \{X_{\text{LARS\_m\_1}}, \cdots, X_{\text{LARS\_m\_q}}\}$$

to record the features respectively.

In step 3, we fit the linear regression model based on the feature

$$X_{\text{LR\_m}} = X_{\text{lasso\_m}} \cup X_{\text{LARS\_m}}$$

By applying the backward selection method, we can get the linear regression model

$$Y_{\text{LR\_m}} = \beta_{\text{LR\_m\_0}} + \beta_{\text{LR\_m\_1}} X_{\text{LR\_m\_1}} + \cdots + \beta_{\text{LR\_m\_q}} X_{\text{LR\_m\_q}}$$

and we call it the best linear regression model.

5.2 Best Linear Regression Model Analysis

This section will analyze the best linear model in a randomly chosen round m. The feature sets of lasso and least angle regression are

$$X_{\text{lasso\_m}} = \{X_{72}, X_{243}, X_{314}, X_{414}, X_{623}, X_{858}, X_{1100}, X_{1130}, X_{1206}, X_{1209}, X_{1278}, X_{1363}, X_{1424}, X_{1477}, X_{1515}, X_{1523}, X_{1527}, X_{1635}, X_{1638}, X_{1761}, X_{1826}, X_{1848}, X_{1854}, X_{2026}, X_{2033}, X_{2241}, X_{2344}, X_{2461}, X_{2483}, X_{2563}, X_{2738}, X_{2873}, X_{3103}, X_{3104}, X_{3225}, X_{3310}, X_{3464}, X_{3513}, X_{4002}, X_{4003}, X_{4044}, X_{4074}\}$$

$$X_{\text{LARS\_m}} = \{X_{72}, X_{314}, X_{414}, X_{623}, X_{826}, X_{1122}, X_{1130}, X_{1278}, X_{1363}, X_{1474}, X_{1477}, X_{1515}, X_{1523}, X_{1527}, X_{1638}, X_{1761}, X_{1819}, X_{1826}, X_{1848}, X_{1854}, X_{1856}, X_{2026}, X_{2241}, X_{2563}, X_{2738}, X_{3104}, X_{3225}, X_{3310}, X_{3513}, X_{4002}, X_{4003}, X_{4074}\}$$

Then we have $X_{\text{LR\_m}} = X_{\text{lasso\_m}} \cup X_{\text{LAR\_m}}$ and $|X_{\text{lasso\_m}}| = 42$, $|X_{\text{LARS\_m}}| = 32$, $|X_{\text{LR\_m}}| = 46$. Using the backward selection method, we remove the least significant variable from the set $X_{\text{LR\_m}}$ at each step and eventually obtain the following model.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.961
Model:	OLS	Adj. R-squared:	0.950
Method:	Least Squares	F-statistic:	88.17
Date:	Mon, 18 Dec 2022	Prob (F-statistic):	2.05e-32
Time:	20:07:37	Log-Likelihood:	19.982
No. Observations:	70	AIC:	-7.964
Df Residuals:	55	BIC:	28.01
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0('ARGF_at', 72)	-0.2223	0.034	-6.575	0.000	-0.290	-0.154
1('GAPB_at', 414)	0.0962	0.039	2.465	0.017	0.018	0.174
2('LYSC_at', 623)	-0.3798	0.166	-2.285	0.026	-0.713	-0.047
3('SPOVAA_at', 1130)	0.5040	0.132	3.826	0.000	0.240	0.768
4('XHLB_at', 1278)	0.1508	0.043	3.526	0.001	0.065	0.237
5('YEBC_at', 1761)	-0.7963	0.151	-5.289	0.000	-1.098	-0.494
6('YFHE_r_at', 1826)	0.2406	0.094	2.555	0.013	0.052	0.429
7('YFIO_at', 1854)	0.4870	0.139	3.491	0.001	0.207	0.767
8('YHDS_r_at', 2026)	0.2436	0.073	3.332	0.002	0.097	0.390
9('YKBA_at', 2241)	0.2808	0.080	3.502	0.001	0.120	0.441
10('YOAB_at', 2563)	-0.8496	0.145	-5.840	0.000	-1.141	-0.558
11('YQJU_at', 3104)	0.6532	0.158	4.123	0.000	0.336	0.971
12('YTGB_at', 3310)	-0.0950	0.044	-2.149	0.036	-0.184	-0.006
13('YXLD_at', 4002)	-0.3553	0.033	-10.836	0.000	-0.421	-0.290
14('YYDA_at', 4074)	-0.2826	0.087	-3.244	0.002	-0.457	-0.108

Omnibus:	1.722	Durbin-Watson:	2.144
Prob(Omnibus):	0.423	Jarque-Bera (JB):	1.068
Skew:	-0.091	Prob(JB):	0.586
Kurtosis:	3.577	Cond. No.	3.57e+03

Figure 8: The result of the best linear model in round m

Meanwhile, we provide the residual plots:

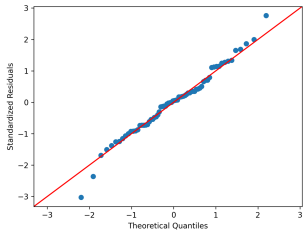

Figure 9: Q_Q plot

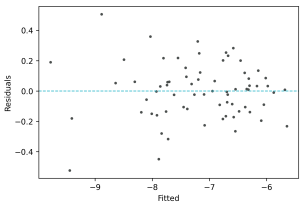

Figure 10: residual plot

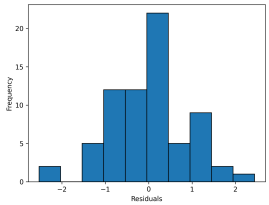

Figure 11: residual histogram

Recall that the assumptions for linear models are:

- normality: residuals are normally distributed,
- independence: residuals are independent of each other,
- homoscedasticity: The variance of residuals is homogeneous.

Based on the results from figure 8 to 11, we verify the assumption one by one with $\alpha = 0.05$:

- normality: from figure 9, 11, and the value of skew and kurtosis, we believe that residuals are normally distributed,
- independence: from the value of Durbin-Watson statistics, we believe that residuals are independent of each other,
- homoscedasticity: from figure 10, we believe that the mean of residuals is 0. However, we cannot conclude that the variance of residuals is homogeneous.

To solve heteroscedasticity, we can apply a proper transformation. However, considering that the pattern can be eliminated by removing outliers and all other assumptions are satisfied, and both R^2 and *adjust* R^2 values are fantastic, we believe this linear model is reasonable.

6 Assessment Criteria and Final Models

6.1 Assessment Criteria and Analysis

We propose four assessment criteria, MSE_{train}, ER_{train} for the training sets, and MSE_{test}, ER_{test} for the testing set where ER represents the error range. The formula for these criteria is as follows:

$$MSE_{train} = \sum_{m=1}^{round=71} \frac{1}{round} \sum_{m\_train=1}^{|training\ set|=70} \frac{(Y_{m\_train} - \hat{Y}_{\text{m\_train}})^2}{|training\ set|}$$

$$MSE_{test} = \sum_{m=1}^{round=71} \frac{1}{round} \sum_{m\_test} (Y_{m\_test} - \hat{Y}_{\text{m\_test}})^2$$

$$ER_{train} = \left[\min_{round} \left(\min_{train} \left(Y_j - \hat{Y}_j \right) \right), \max_{round} \left(\max_{train} \left(Y_j - \hat{Y}_j \right) \right) \right]$$

$$ER_{test} = \left[\min_{round} \left(\min_{test} \left(Y_j - \hat{Y}_j \right) \right), \max_{round} \left(\max_{test} \left(Y_j - \hat{Y}_j \right) \right) \right]$$

Here m_train and m_train represent the data in the training dataset and the testing dataset for round m, respectively. Also,Y and $\hat{Y}$ represent the actual value and estimated value by a specific method, respectively. The table below records the values of different methods under these criteria, where "overfit LR" represents the linear model fitted with all features, "best LR" represents the best linear regression, and "RF" represents random forest. The parentheses following each method in the first column represent the run time of the method.

criteria / method	MSE_{train}	MSE_{test}	ER_{train}	ER_{test}
overfit LR(4.978s)	2.4542e-29	0.2657	[-3.3750,2.9310]	[-1.6054,1.1532]
lasso(831.2s)	0.03308	0.2020	[-0.4523,0.9383]	[-1.5853,0.7661]
ridge(6.988s)	4.4030e-5	0.2649	[-0.0183,0.0296]	[-1.6144,1.1497]
elastic net(433.6s)	0.03359	0.2209	[-0.4974,0.9089]	[-1.7289,0.7718]
LARS(354.3s)	0.04833	0.1955	[-0.5273,1.0854]	[-1.6902,0.7919]
PCR(59.36s)	0.1238	0.2247	[-0.9759,1.5454]	[-1.7926,0.9869]
RF(565.9s)	0.05855	0.4568	[-0.6246,1.0796]	[-2.0450,1.2706]
best LR(15.44s)	0.03211	0.1792	[-0.5328,0.6925]	[-1.1666,0.9580]

Table 1: The assessment criteria for different methods

In fact, compared to the test set, we are not particularly concerned with the performance on the training set because, in the case of $p >> n$, the model is very prone to overfitting and cannot predict for a new observation. A typical example is "overfit LR". This model performs exceptionally well on the training set, but its performance on the test set is poor. Therefore, we need to focus on the model's performance on the test set. As we can see, "best LR" has the smallest MSE and ER on the test set. Hence we consider the proposed "best LR" model is the best.

6.2 Final Models

Ensemble methods are a set of techniques that combine the predictions of multiple models to produce a more accurate and stable prediction of a single model by reducing overfitting and increasing generalization.

We will use the ensemble idea to present the final model based on different methods. More specifically,

Final lasso model Y_{lasso}: ensemble of $\{Y_{\text{lasso\_1}}, \cdots, Y_{\text{lasso\_71}}\}$
Final ridge model Y_{ridge}: ensemble of $\{Y_{\text{ridge\_1}}, \cdots, Y_{\text{ridge\_71}}\}$
Final elastic net model $Y_{elastic}$: ensemble of $\{Y_{\text{elastic\_1}}, \cdots, Y_{\text{elastic\_71}}\}$
Final LARS model Y_{elas}: ensemble of $\{Y_{\text{LARS\_1}}, \cdots, Y_{\text{LARS\_71}}\}$
Final PCR model Y_{PCR}: ensemble of $\{Y_{\text{PCR\_1}}, \cdots, Y_{\text{PCR\_71}}\}$
Final random forest model Y_{RF}: ensemble of $\{Y_{\text{RF\_1}}, \cdots, Y_{\text{RF\_71}}\}$
Final best linear regression model Y_{LR}: ensemble of $\{Y_{\text{LR\_1}}, \cdots, Y_{\text{LR\_71}}\}$

In addition to the given 71 samples, if there is a new observation, then the prediction value of the specific model generated by a certain method for this new sample is the simple average of the prediction values of the 71 models from the ensemble set.

References

[1] https://rdrr.io/rforge/hdi/man/riboflavin.html.

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[4] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[5] Juming Pan. "Improved two-stage model averaging for high-dimensional linear regression, with application to Riboflavin data analysis". In: *BMC bioinformatics* 22.1 (2021), pp. 1–17.

[6] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[7] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[8] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.