---
title: "PSTAT126 Group Project Step 4"
author: "Hanya Ansari, Carina Yuen, Daren Aguilera"
output:
html_document: default
word_document: default
pdf_document: default
date: "2023-12-10"

Introduction:

Wine Quality Based on Physicochemical Tests from UCI Machine Learning Repository https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/code?datasetId=4458&searchQuery=R

No Missing Attribute Values: 0

Number of Instances: red wine: 1599

Number of Variables: 12 total, 11 continuous, 1 discrete (fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol and 1 integer output variable: quality score between 0 and 10)

```
knitr::opts_chunk$set(echo = FALSE,
                      message = F,
                      warning  = F,
                      fig.width = 4,
                      fig.height = 3,
                      fig.align = 'center',
                      fig.pos = 'H')
```

```
## [1] 0.06613643
```

```
## [1] "BIC"
```

```
## [1] -114034.5
```

```
##   (Intercept)           x_f           x_v           x_c           x_r
##  1.136554e-15 -1.914169e-17  1.000000e+00  1.451585e-17  2.303403e-19
##          x_ch          x_fs          x_ts           x_d           x_p
##  1.284814e-17  3.272747e-20 -2.433415e-20 -7.527894e-16  2.313397e-17
##           x_s           x_a
## -3.607185e-18 -8.201320e-19
```

```
##
## Call:
## lm(formula = volatile.acidity ~ x_f + x_v + x_c + x_r + x_ch +
##     x_fs + x_ts + x_d + x_p + x_s + x_a, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.033e-15 -2.780e-18  2.190e-18  6.610e-18  7.159e-17
##
## Coefficients:
```

```
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.137e-15  2.517e-15  4.520e-01    0.652
## x_f         -1.914e-17  3.082e-18 -6.211e+00  6.7e-10 ***
## x_v          1.000e+00  1.438e-17  6.953e+16  < 2e-16 ***
## x_c          1.452e-17  1.748e-17  8.300e-01    0.406
## x_r          2.303e-19  1.782e-18  1.290e-01    0.897
## x_ch         1.285e-17  4.980e-17  2.580e-01    0.796
## x_fs         3.273e-20  2.579e-19  1.270e-01    0.899
## x_ts        -2.433e-20  8.655e-20 -2.810e-01    0.779
## x_d         -7.528e-16  2.569e-15 -2.930e-01    0.770
## x_p          2.313e-17  2.275e-17  1.017e+00    0.309
## x_s         -3.607e-18  1.358e-17 -2.660e-01    0.791
## x_a         -8.201e-19  3.145e-18 -2.610e-01    0.794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.696e-17 on 1587 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.864e+32 on 11 and 1587 DF,  p-value: < 2.2e-16
```

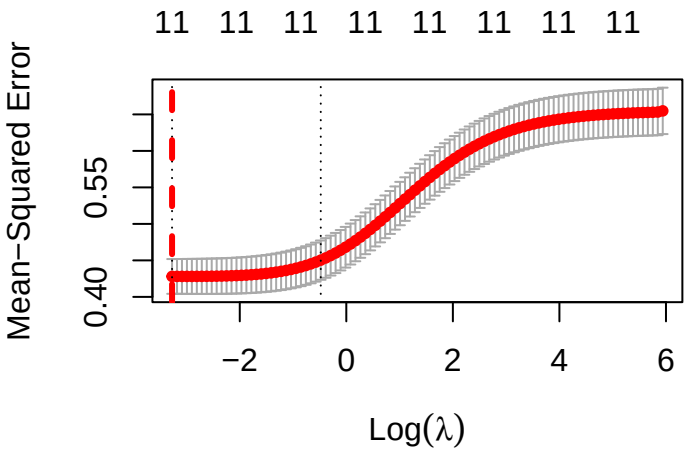
Shrinkage Methods: Lasso and Ridge Regression with continuous response variable

From our previous Project Step, our best model was fit_normal3, a linear model that has the response variable as volatile acidity, and the predictors (exlcuding itself) as the whole model.

Ridge Regression

Ridge regression shrinks the coefficients towards zero. Upon using coefficients (best_model), we got the following estimates: 10.42 for the intercept, 0.987 for alcohol, -0.0025 for free sulfur dioxide, and -0.0043 for total sulfur dioxide. Using cv.glmnet, we found that the optimal lambda is 0.03105, as a dotted vertical line marks the value of Log() that minimizes the MSE value.

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           33.730618813
## fixed.acidity          0.046331467
## volatile.acidity      -1.012462407
## citric.acid            0.004948314
## residual.sugar         0.017663385
## chlorides             -1.626540061
## free.sulfur.dioxide    0.004699672
## total.sulfur.dioxide  -0.002976293
## density              -30.454702348
## pH                    -0.176906287
## sulphates              0.898470288
## alcohol                0.246254306
```




```
##  [1]  33.485383311   0.045876581  -1.011644247   0.005783178   0.017558752
##  [6]  -1.629820799   0.004711746  -0.002981510 -30.198089295  -0.179643510
## [11]   0.898523824   0.246466206            NA            NA            NA
## [16]            NA            NA            NA            NA            NA
```

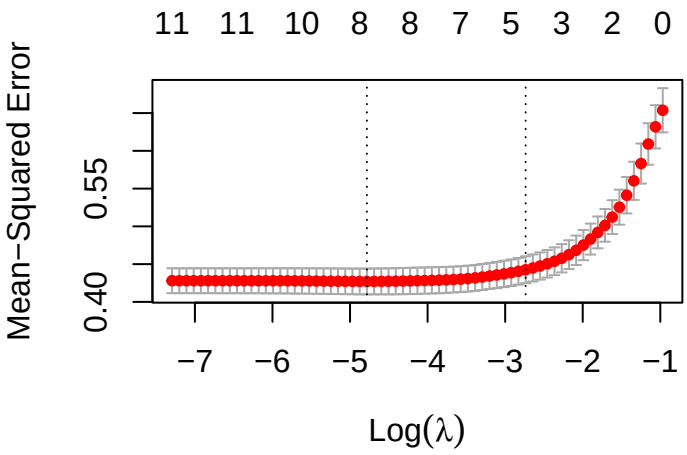
Lasso Regression

We decided to fit the other predictors to our response variable, volatile acidity. Using Lasso Regression, we obtained the following estimates for the intercept, and coefficients of fixed acidity, and density, respectively from the best model : 10.4230, 0.8485, -1.0751). Fixed acidity and density were highlighted because they were relatively large in magnitude, compared to the mean predictor magnitude (excluding the intercept) is about 0.3245. The optimal lambda (to minimize test MSE) was calculated to be about 0.001139. The small eigenvalue indicates multicollinearity.

Lasso Regression (Least Absolute Shrinkage and Selection Operator): Upon using Lasso Regression, we noted that all the predictors except alcohol are dropped from the model. In the output matrix, the coefficients are "empty" values. Since there is only one "non-zero" coefficient of alcohol (besides the intercept), it indicates the regularization parameter lambda might be causing too much regularization. Here, we can see the limitations of using certain criteria for determining the calculated best_lambda is 0.03105506. The MSE plot for the Lasso Regression looks similar to that of Ridge Regression, with the vertical dotted line being plotted slightly more to the left at -3.75 instead of -3.5.

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           3.557162153
## fixed.acidity         0.019918790
## volatile.acidity     -1.062756353
## citric.acid           .
## residual.sugar        .
## chlorides            -1.437641812
## free.sulfur.dioxide   0.003065161
## total.sulfur.dioxide -0.002380744
## density               .
```

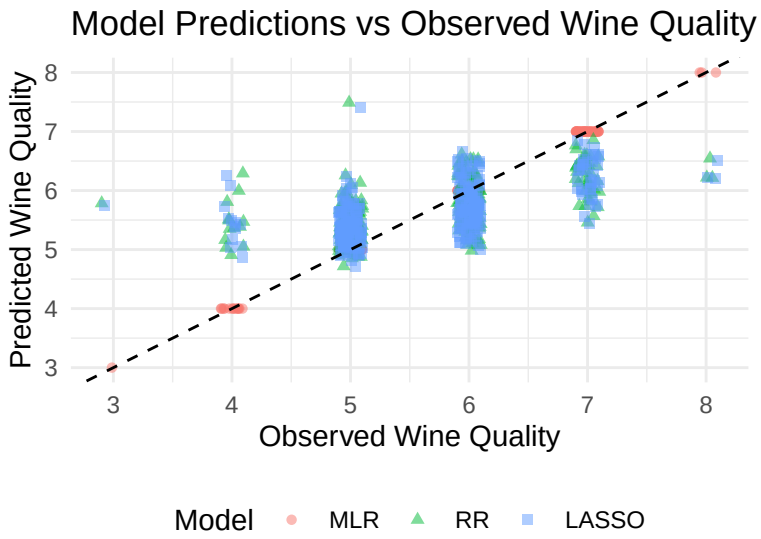
```
## pH                   -0.243790804
## sulphates             0.821890228
## alcohol               0.281239758
```




```
##                         Estimate   Std. Error       t value     Pr(>|t|)
## (Intercept)           2.911886e-14 2.034891e-14  1.430979e+00 1.526827e-01
## fixed.acidity         5.029839e-17 2.519780e-17  1.996142e+00 4.613198e-02
## volatile.acidity     -1.049432e-15 1.163763e-16 -9.017574e+00 6.926179e-19
## citric.acid          -1.022384e-16 1.415465e-16 -7.222951e-01 4.702462e-01
## residual.sugar        1.475905e-17 1.519323e-17  9.714224e-01 3.315232e-01
## chlorides            -1.449937e-15 4.092305e-16 -3.543081e+00 4.098448e-04
## free.sulfur.dioxide   5.191228e-18 2.100162e-18  2.471822e+00 1.357315e-02
## total.sulfur.dioxide -2.781734e-18 7.279006e-19 -3.821585e+00 1.390236e-04
## density              -2.663185e-14 2.078264e-14 -1.281447e+00 2.002710e-01
## pH                   -1.975087e-16 1.862827e-16 -1.060263e+00 2.892269e-01
## sulphates             8.598359e-16 1.125855e-16  7.637179e+00 4.352808e-14
## alcohol               2.475233e-16 2.645374e-17  9.356837e+00 3.603119e-20
## quality               1.000000e+00 2.396087e-17  4.173471e+16 0.000000e+00

## $rmse_ridge
## [1] 0.6516314
##
## $rmse_lasso
## [1] 0.6514806
##
## $rmse_mlr
## [1] 6.723887e-16
```

The plot below measures the prediction using three regression models. Comparing the points to the dotted line, we see that the LASSO regression method has clustered values of the predicted response at observed response 4-7.



Weighted Least Squares

Innovation: New Regression Method with Weighted Least Squares Regression (WLS) In addition to the Ordinary Least Squares Regression, we looked into the new method, Weighted Least Squares Regression. WLS Regression adjusts for the violation in the assumption because each weight is made inversely proportional to error variance, so our results weigh observations with lower variance heavier. The intuition behind this weight The WLS Regression coefficients for the variables volatile acidity, free sulfur dioxide, total sulfur dioxide, and alcohol are 1, -8.507*10-20, and 2.815*10-18, respectively.

We chose this as our new method because it helps when the data violates the homoscedasticity assumption. From our Scale Location plot of the full linear model with the response being volatile acidity, we can see there is a defined pattern in the graph (see page 5). . While we were able to mostly address this by using BIC to pick our best model, using the WLS method.

```
##
## Call:
## lm(formula = quality_data ~ x_f + poly(x_v, 3, raw = T) + x_c +
##     x_r + poly(x_ch, 3, raw = T) + poly(x_fs, 3, raw = T) + poly(x_ts,
##     3, raw = T) + x_d + poly(x_p, 3, raw = T) + poly(x_s, 3,
##     raw = T) + poly(x_a, 3, raw = T), data = data, weights = wt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4480 -0.8116 -0.0729  0.8684  4.0658
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.379e+01  5.554e+01   0.608  0.54302
## x_f                        3.344e-02  2.675e-02   1.250  0.21151
## poly(x_v, 3, raw = T)1    -2.194e+00  9.595e-01  -2.286  0.02236 *
## poly(x_v, 3, raw = T)2     2.323e+00  1.331e+00   1.746  0.08108 .
## poly(x_v, 3, raw = T)3    -1.217e+00  5.658e-01  -2.151  0.03161 *
## x_c                       -2.590e-01  1.476e-01  -1.755  0.07946 .
## x_r                        1.367e-02  1.524e-02   0.897  0.36988
```

```
## poly(x_ch, 3, raw = T)1 -4.778e+00  2.251e+00  -2.123  0.03391 *
## poly(x_ch, 3, raw = T)2  1.665e+01  9.980e+00   1.668  0.09547 .
## poly(x_ch, 3, raw = T)3 -2.016e+01  1.194e+01  -1.689  0.09150 .
## poly(x_fs, 3, raw = T)1  2.528e-02  1.234e-02   2.049  0.04060 *
## poly(x_fs, 3, raw = T)2 -9.583e-04  4.511e-04  -2.124  0.03382 *
## poly(x_fs, 3, raw = T)3  1.058e-05  4.889e-06   2.163  0.03066 *
## poly(x_ts, 3, raw = T)1  3.247e-03  3.395e-03   0.956  0.33898
## poly(x_ts, 3, raw = T)2 -7.384e-05  3.399e-05  -2.173  0.02996 *
## poly(x_ts, 3, raw = T)3  2.463e-07  9.750e-08   2.527  0.01161 *
## x_d                     -3.662e+01  2.310e+01  -1.585  0.11308
## poly(x_p, 3, raw = T)1   3.286e+01  4.227e+01   0.777  0.43709
## poly(x_p, 3, raw = T)2  -9.034e+00  1.258e+01  -0.718  0.47262
## poly(x_p, 3, raw = T)3   7.998e-01  1.245e+00   0.642  0.52065
## poly(x_s, 3, raw = T)1   9.217e+00  1.462e+00   6.304 3.76e-10 ***
## poly(x_s, 3, raw = T)2  -7.752e+00  1.542e+00  -5.027 5.56e-07 ***
## poly(x_s, 3, raw = T)3   2.016e+00  4.825e-01   4.179 3.09e-05 ***
## poly(x_a, 3, raw = T)1  -9.474e+00  3.034e+00  -3.122  0.00183 **
## poly(x_a, 3, raw = T)2   8.762e-01  2.744e-01   3.193  0.00144 **
## poly(x_a, 3, raw = T)3  -2.608e-02  8.216e-03  -3.175  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.303 on 1573 degrees of freedom
## Multiple R-squared:  0.3934, Adjusted R-squared:  0.3838
## F-statistic: 40.81 on 25 and 1573 DF,  p-value: < 2.2e-16
```