

ISYE6501 HW5

Keh-Harn Feng

June 16, 2017

0.1 Preface

This is a reproducible report with most of the codes doing the heavy lifting hidden in the background. You can download the source code of the report by [clicking here](#). All of the information you need as a reviewer is in the report, not in the code. You should NEVER run any R scripts from an untrusted source on your own computer.

1 Question 1

Using the crime data set from Homework 3, build a regression model using: 1. Stepwise regression 2. Lasso 3. Elastic net For Parts 2 and 3, remember to scale the data first – otherwise, the regression coefficients will be on different scales and the constraint won't have the desired effect. For Parts 2 and 3, use the `glmnet` function in R.

Variable `So` is converted to a 2-level factor as it is categorical. The data is then splitted into a training (~90%) and testing (~10%) sets as usual. Predictor variables from the training set are transformed using Box-Cox transformation then centered and scaled (see my [HW3 Report](#)). The response `Crime` is not preprocessed.

1.1 Stepwise Regression

Stepwise regression in both directions is carried out using the `step()` function and AIC as the selection metric. The initial starting point is an OLS model using all features.

The final model selected is shown below:

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Pop + NW + U1 + U2 + Wealth +
##      Ineq + Prob, data = q1.train.preproc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -319.2 -106.9  -16.6   92.2  368.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    904.6      27.9    32.42 < 2e-16 ***
## M               96.2      44.9     2.15  0.03960 *
## Ed             162.5      55.9     2.91  0.00657 **
## Po1            260.7      71.5     3.64  0.00094 ***
## Pop           -112.6      43.4    -2.60  0.01411 *
## NW              95.0      54.6     1.74  0.09168 .
## U1            -70.9      55.6    -1.28  0.21088
## U2            130.5      59.8     2.18  0.03654 *
## Wealth         148.7      91.5     1.63  0.11397
## Ineq           307.4      72.8     4.22  0.00019 ***
```

```
## Prob          -116.7          38.5    -3.03  0.00480 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 183 on 32 degrees of freedom
## Multiple R-squared:  0.834, Adjusted R-squared:  0.782
## F-statistic: 16.1 on 10 and 32 DF,  p-value: 8.11e-10
```

While the process cut down the number of predictors from 15 to 10, features `NW`, `U1`, and `Wealth` show p-values higher than 0.05. It should be noted that model statistics from stepwise regression often [do not mean what they are supposed to mean](#). It is generally frowned upon as a naive method for feature selection.

1.2 LASSO Regression

A LASSO regression is fitted using `glmnet` ($\alpha = 0$). 10-fold cross-validation is used to generate estimated MSE to select the optimal regularization parameter λ . The resulting LASSO model contains the following coefficients:

name	coefficient
(Intercept)	904.581
M	67.912
Ed	37.455
Po1	271.232
Po2	6.735
M.F	93.077
NW	66.459
U2	46.723
Ineq	112.696
Prob	-91.500

9 features out of 15 are selected for the LASSO model. Notice that from HW3 we know `Po1` and `Po2` are highly correlated. In fact, `Po1` was removed during exploratory analysis due to high colinearity check in my [HW3 Report](#). Here LASSO does not remove either predictor. Instead `Po1` is assigned a much heavier influence compared to `Po2`. This is one of the common criticisms about LASSO: one of the highly correlated variables tends to be selected and assigned heavy influence while others are penalized seemingly randomly.

1.3 Elastic Net

Before we move on to elastic net, Figure 1 shows the effect of α on model MSE (all CV folds are fixed to the folds used to build the LASSO model).

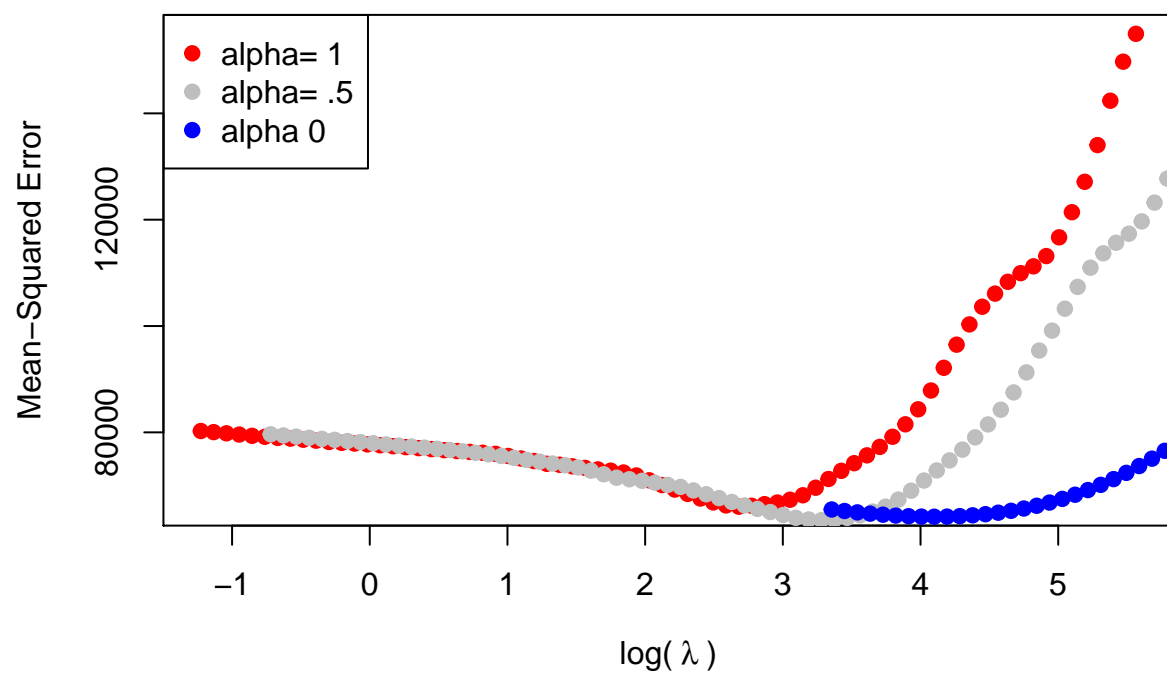


Figure 1: 10-CV estimated model MSE vs $\log(\lambda)$.

It seems that $\alpha = 0.5$ achieves the smallest MSE compared to both LASSO ($\alpha = 1$) and ridge regression ($\alpha = 0$). A grid-search is done to find the alpha value that corresponds to the minimum estimated MSE using the same CV folds.

The minimum MSE found vs corresponding alpha values is shown in Figure 2. With this in mind, an elastic net model is constructed using the alpha value that corresponds to the smallest MSE, $\alpha = 0.202$.

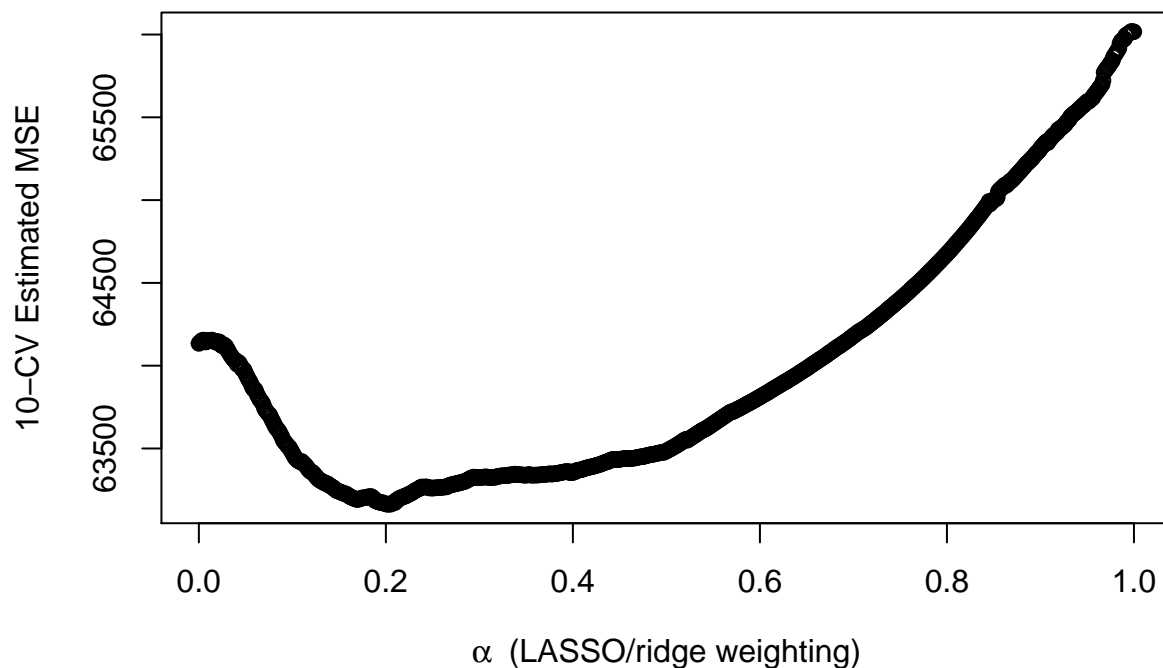


Figure 2: Min MSE vs alpha

The model coefficients are shown below.

name	coefficient
(Intercept)	904.58
M	66.79
Ed	42.91
Po1	133.88
Po2	116.30
LF	11.90
M.F	86.19
NW	78.51
U1	-4.59
U2	60.67
Ineq	92.44
Prob	-93.13

With 11 out of 15 features selected the model is perhaps in danger of overfitting. However, in contrast to LASSO the colinear predictors Po1 and Po2 are assigned roughly equal weights. This is an expected behavior

of elastic net regression. The estimated MSE as a function of $\log(\lambda)$ from both LASSO and elastic net is shown in Figure 3. The minimum MSE achieved by LASSO is 6.6016×10^4 while that of the elastic net is 6.3163×10^4 .

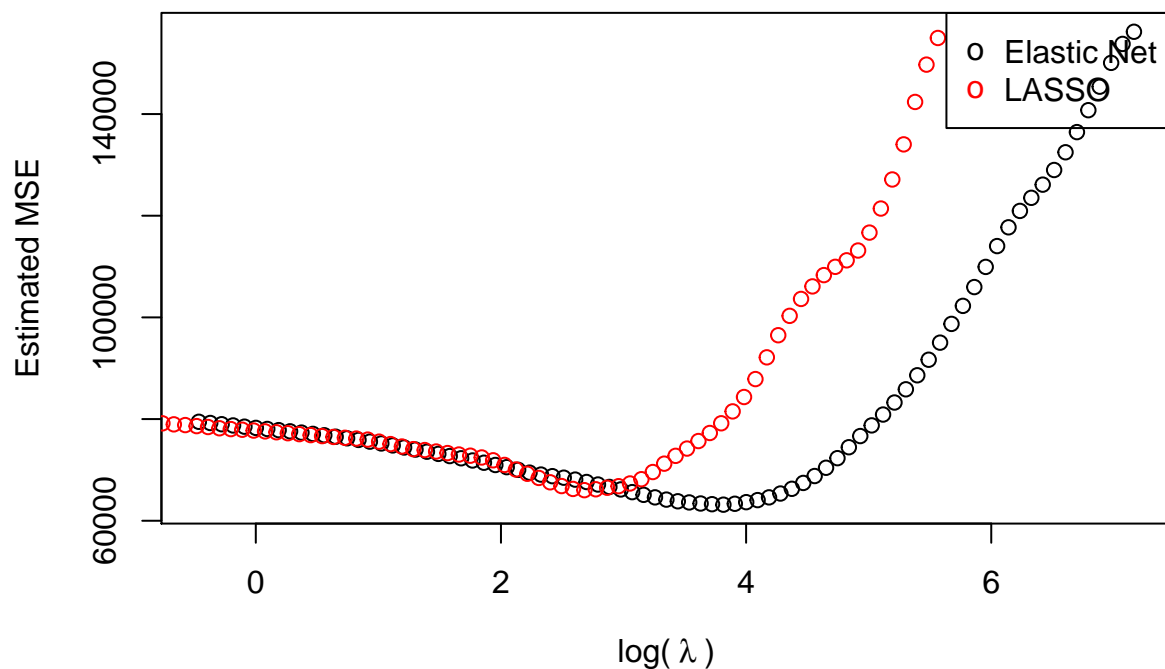


Figure 3: 10-CV estimated MSE vs $\log(\lambda)$ between LASSO and elastic net ($\alpha \sim 0.202$).

1.4 Model Performance

Model performance is evaluated using the test set with the RMSE as the metric. The result is shown below.

Model	RMSE
Stepwise	181.7
LASSO	290.6
Elastic Net	303.5

Surprisingly stepwise regression is able to achieve the best test set RMSE by far. Both LASSO and elastic net are roughly equal and significantly worse than stepwise regression. I do not have any satisfactory explanation for this besides that the sample size is very small to begin with and that the result could easily be due to any one of the model hitting a fortunate (or unfotunate) jackpot.

2 Question 2

Describe a situation or problem from your job, everyday life, current events, etc., for which a design of experiments approach would be appropriate.

I was the acting manager and coach of an amateur softball team a couple of years ago. One of my responsibilities was to oversee the design, manufacturing and distribution of the team jerseys. With over 20 players of vastly different backgrounds and ages it was not easy to find an optimal design that satisfied everyone. People have different preferences when it came to material, color, font, and graphics. If I am given the same task again, once the team administrators and designers figure out a limited set of choices for the different elements on the jersey full/fractional factorial design can be utilized to generate different virtual samples using computer graphics. These can then be used in a survey for the team members in order to reach a consensus for an optimal choice.

3 Question 3

To determine the value of 10 different yes/no features to the market value of a house (large yard, solar roof, etc.), a real estate agent plans to survey 50 potential buyers, showing a fictitious house with different combinations of features. To reduce the survey size, the agent wants to show just 16 fictitious houses. Use R's FrF2 function (in the FrF2 package) to find a fractional factorial design for this experiment: what set of features should each of the 16 fictitious houses? Note: the output of FrF2 is "1" (include) or "-1" (don't include) for each feature.

The 50 buyers are assumed to hold different preferences and weightings regarding different features (ie: the same design may elicit different responses to different buyers). Therefore it is important that each buyer browses all 16 designs in order to eliminate potential bias from assigning designs to smaller groups. The order of browsing is assumed to be inconsequential. This can be described as designing an experiment with 10 two-level features that will be run 16 times, each time involving a replicated survey of 50 people and each run with a different combination of feature levels. Since all buyers will browse all 16 designs and order doesn't matter the number of buyers is actually not important. The design matrix generated by `FrF2()` is shown below:

Feat.1	Feat.2	Feat.3	Feat.4	Feat.5	Feat.6	Feat.7	Feat.8	Feat.9	Feat.10
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	Yes	No	Yes	No	Yes	No	No
No	No	Yes	Yes	Yes	No	No	No	No	Yes
No	Yes	No	No	No	Yes	No	Yes	Yes	No
Yes	No	No	No	No	No	Yes	No	No	No
Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No
Yes	Yes	No	No	Yes	No	No	No	Yes	Yes
Yes	No	Yes	No	No	Yes	No	No	Yes	Yes
No	Yes	Yes	Yes	No	No	Yes	No	Yes	No
Yes	Yes	No	Yes	Yes	No	No	Yes	No	No
No	No	Yes	No	Yes	No	No	Yes	Yes	No
No	No	No	No	Yes	Yes	Yes	Yes	No	Yes
No	Yes	Yes	No	No	No	Yes	Yes	No	Yes
Yes	No	No	Yes	No	No	Yes	Yes	Yes	Yes
No	Yes	No	Yes	No	Yes	No	No	No	Yes
No	No	No	Yes	Yes	Yes	Yes	No	Yes	No

Digression: If the assumption above doesn't hold and the 50 buyers contain features of their own that may affect the experiment outcome (ie: genders, age, ethnicity etc) then these factors must be accounted for as well (possibly through blocking). If the order for which example houses are shown to the buyers first matters

(ie: seeing a great house first negatively affects the outcome of a mediocre house shown right after, vice versa) then the buyers must be carefully divided into groups and with the houses shown in randomized orders to control its effect.

4 Question 4

For each of the following distributions, give an example of data that you would expect to follow this distribution (besides the examples already discussed in class).

As someone with a MS in physics I am going to put on my *Big Bang Theory* hat and troll a bit:

- a. **Binomial** The number of female partners I got randomly assigned to during my third year undergraduate physics labs. FYI, the expected value as I recall was just about zero (fine let's call it 0.5). Each class has a total of 10 lab sessions.
- b. **Geometric** The probability distribution of the number of times I need to roll a d20 dice before landing a critical hit during a game of Dungeons & Dragons. You land a critical hit when you roll a 20 (usually).
- c. **Poisson** The number of times some engineers got snubbed asking physicists questions in the last week. Don't worry, mathematicians tend to turn the table on us.
- d. **Exponential** The probability distribution for the amount of time that passes before another engineer gets snubbed by a physicist. On behalf of my fellow physicists, I apologize.
- e. **Weibull** The probability of people switching out of undergraduate physics as a function of their year of study. Believe me, by the time you get to the higher years those who stay either know exactly what they are doing, don't mind double espressos at 1AM or a bit of both.