

# Homework 3 Linear Model Selection and Regularization

by Chu Zhuang

```
# import relevant packages
import random
import math
import itertools
import numpy as np
import pandas as pd
import seaborn
import sklearn
import matplotlib as mpl
import matplotlib.pyplot as plt
```

## Conceptual Exercises

1. Generate a data set with  $p=20$  features, 1000 observations, and an associated response vector by  $Y=X*\beta+error$

```
#generate a data set
n_train, n_test=100,900
n_feature=20

#generate X 20-features, 1000-observations
np.random.seed(123)
X= np.random.rand(n_train+n_test,n_feature)*10

#generate Beta 20-coefficient, the last 8 are set to be zero
beta_coef=np.random.rand(n_feature)
beta_coef[12:]=0.0
beta_coef=beta_coef.reshape(20,1)

#calculate Y accordingly with errors
y_error=np.random.randn(n_train+n_test,1)
Y0=np.dot(X,beta_coef)
Y=np.dot(X,beta_coef)+y_error

#organize into a dataframe
df_cp=pd.DataFrame({'X':list(X), 'Y':list(Y), 'y_error':list(y_error)})
```

```
df_cp
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	X	Y	y_error
0	[6.964691855978616, 2.8613933495037944, 2.2685...	[27.62564196099788]	[0.8092276895919737]
1	[6.344009585513211, 8.494317940777895, 7.24455...	[27.269392593843648]	[0.7010854594935072]
2	[6.239529517921111, 1.1561839507929572, 3.1728...	[30.386079332341446]	[2.9152554205081374]
3	[6.693137829622723, 5.859365525622129, 6.24903...	[25.71139760261844]	[0.8081492140769071]
4	[3.1876642638187636, 6.919702955318197, 5.5438...	[31.090624072617736]	[1.4912948137199595]
...	...	...	...
995	[0.7544400694446007, 3.465663070787391, 2.1228...	[20.40875753816898]	[0.011585083014903465]
996	[0.4658994178681364, 5.325660638187052, 0.8139...	[29.59283384861341]	[0.3646261309540851]
997	[5.148211758684324, 1.5583108257630751, 8.4394...	[31.303558721465993]	[-1.2090071180126118]
998	[7.349337943650992, 3.54780816851492, 4.160529...	[27.674115275673262]	[1.337151062369074]
999	[8.435495618748645, 2.070012716323065, 3.86273...	[18.433197918147233]	[-0.6840920422376263]

1000 rows × 3 columns

2. Split data into training and test set: **900-Training, 100-Test**

```
X_train, X_test=X[:n_train],X[n_train:]
Y_train, Y_test=Y[:n_train],Y[n_train:]
Y_train.shape
```

```
(100, 1)
```

3. Perform best subset selection on the training set and plot MSE of training set-best model of each size.

```
#Linear regression model
from sklearn import linear_model
model_lr= linear_model.LinearRegression()

#construct list to save MSE
bestmodel_MSE=[]
bestmodel_fsubset=[]
bestmodel_beta=[]
bestmodel_fnum=[]

for i in range(20):
    #size of each subset
    n=i+1
    #combination: select n features from 20 features
    feature_set=list(itertools.combinations(list(range(20)),n))

    #define list to store result of each model
    MSE_subset=[]
    p_beta_subset=[]

    #iterate over each model
    for sub_set in feature_set:
        #selec X feature and fit in model
        X_train_subset=X_train[:,list(sub_set)]
        model_lr.fit(X_train_subset,Y_train)
        #Predict Y based on Training dataset
        p_Y=model_lr.predict(X_train_subset)

        #calculate MSE and beta
        p_beta=model_lr.coef_
        y_diff=(Y_train-p_Y).reshape(100)
        model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
        #store MSE/beta for each model
        MSE_subset.append(float(model_MSE))
        p_beta_subset.append(p_beta)

    #find the model of least MSE, for each subset
    #and store the subset feature and beta accordingly
    min_MSE=min(MSE_subset)
    min_model=MSE_subset.index(min_MSE)
    min_subset=feature_set[min_model]
    min_beta=p_beta_subset[min_model]

    #store MSE/beta/feature of best model of each subset
    bestmodel_MSE.append(min_MSE)
    bestmodel_fsubset.append(min_subset)
    bestmodel_fnum.append(n)
    bestmodel_beta.append(min_beta)
```

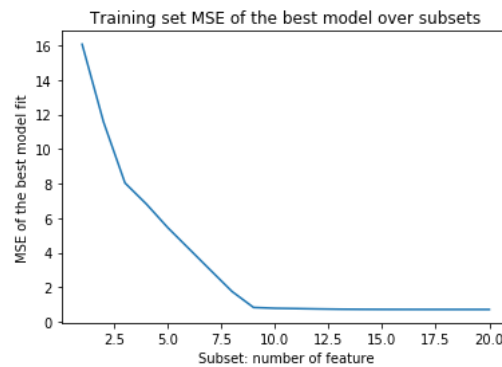
Plot the training set MSE over the number of subset features:

- **Size num 20:** As we can see from the figure below, MSE decreases monotonically with the number of features selected. With the feature size of 20, model has the minimum MSE 0.7049;
- **Floor around 12 features:** while we could see from the figure below that around 12 features, MSE decreases more slowly and gradually touches the 'floor'. Actually in the model, there are only 12 effective features, based on the true parameter and construction of Y.

```
#find the model size with minnum MSE
min_MSE_all=min(bestmodel_MSE)
bestmodel_index=bestmodel_MSE.index(min_MSE_all)
bestmodel_fnum_all=bestmodel_fnum[bestmodel_index]
print('minnum MSE:',round(min_MSE_all,4))
print('model size:',bestmodel_fnum_all)
```

```
minnum MSE: 0.7049
model size: 20
```

```
#plot the Training set MSE
plt.plot(bestmodel_fnum,bestmodel_MSE);
plt.xlabel('Subset: number of feature');
plt.ylabel('MSE of the best model fit');
plt.title('Training set MSE of the best model over subsets');
```



4/5. Plot the **Test dataset** MSE associated with the best model of each size.

- First of all, change the pipeline to select the best model based on MSE from **Test** dataset:

```
#construct list to save MSE
bestmodel_MSE_t=[]
bestmodel_fsubset_t=[]
bestmodel_beta_t=[]
bestmodel_fnum_t=[]

for i in range(20):
    #size of each subset
    n=i+1
    #combination: n features from 20 features
    feature_set=list(itertools.combinations(list(range(20)),n))

    #list to store the result of each model
    MSE_subset_t=[]
    p_beta_subset_t=[]

    #iterate over each model
    for sub_set in feature_set:
        #selec X feature and fit in model
        X_train_subset=X_train[:,list(sub_set)]
        X_test_subset=X_test[:,list(sub_set)]
        model_lr.fit(X_train_subset,Y_train)

        #Predict y based on test dataset
        p_Y=model_lr.predict(X_test_subset)
        #calculate MSE and beta
        p_beta=model_lr.coef_
        y_diff=(Y_test-p_Y).reshape(p_Y.shape[0])
        model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
        #store MSE/beta for each model
        MSE_subset_t.append(float(model_MSE))
        p_beta_subset_t.append(p_beta)

    #find the model of least MSE
    min_MSE_t=min(MSE_subset_t)
    min_model_t=MSE_subset_t.index(min_MSE_t)
    min_subset_t=feature_set[min_model_t]
    min_beta_t=p_beta_subset_t[min_model_t]

    #store info of best model of each subset of feature
    bestmodel_MSE_t.append(min_MSE_t)
    bestmodel_fsubset_t.append(min_subset_t)
    bestmodel_fnum_t.append(n)
    bestmodel_beta_t.append(min_beta_t)
```

Plot the Test set MSE over the number of subset features, and its associated parameters (feature size, feature set and predicted beta):

```
#find the model size with minmum MSE
min_MSE_all_t=min(bestmodel_MSE_t)
bestmodel_index_t=bestmodel_MSE_t.index(min_MSE_all_t)
bestmodel_fnum_all_t=bestmodel_fnum_t[bestmodel_index_t]
print('minimum MSE:',round(min_MSE_all_t,4))
print('model size:',bestmodel_fnum_all_t)
print('feature number:',bestmodel_fsubset_all_t)
print('beta coefficient:', bestmodel_beta_all_t[0])
```

```

minimum MSE: 1.1589
model size: 16
feature number: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 18, 20]
beta coefficient: [ 0.05231662  0.61695955  0.46036422  0.67380485  0.48166209  0.91526139
 0.00600486  0.75717369  0.46510931  0.34258971  0.0476581  0.40346762
-0.00171955  0.02314112  0.00462924  0.03789629]

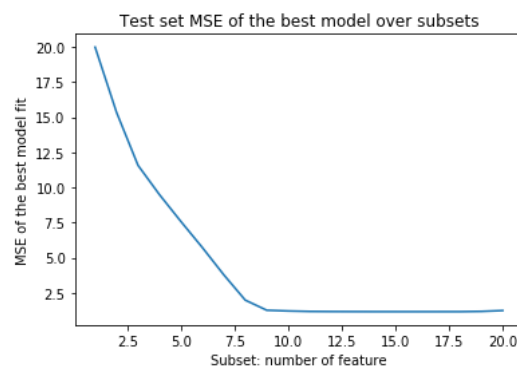
```

- **Size num 16:** As we can see from the figure below, MSE also decreases dramatically with the number of features selected in test dataset. However, for test dataset, the best model is of size 16--**16 features, instead of 20**, compared with the best model from training dataset. The minimum MSE is about 1.16, a little bit higher than the best model from Training dataset (0.7049).
- As for the feature selected, we could see that the best model with minimum MSE selected all the first **12 'true' features** (the features who have non-zero contribution in the original model) and the other 4 features; however, for these 4 features, the coefficients predicted is relatively low (-0.0017-0.037), compared with the first 12 'true' features (0.006-0.915), which means the extra 4 features contribute less to the prediction, which is reasonable.

```

#plot the test set MSE of the best model
plt.plot(bestmodel_fnum_t,bestmodel_MSE_t);
plt.xlabel('Subset: number of feature');
plt.ylabel('MSE of the best model fit');
plt.title('Test set MSE of the best model over subsets');

```



6. Compare the Coefficient between true model and best model from Test MSE:

- As we could see from the graph below, the predicted coefficients (in Red) from the best Test model with least MSE, are very close to the true model coefficients (in Blue). Even coefficients for some features are nearly identical across the true model and best predicted model, with very small difference in several other features. As for the extra features selected in the predicted model, their coefficients are very low and even close to zero, which indicates the overall similarity between the best fit and the true model.

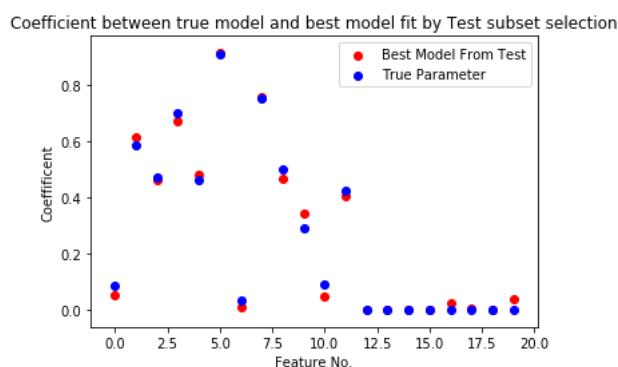
```

#plot coefficients from the both models
colors=['r','b']
labels=['Best Model From Test','True Parameter']

x=range(20)
y1=[0]*20
feature=bestmodel_fsubset_t[bestmodel_index_t]
beta=bestmodel_beta_all_t[0]
for i in range(len(y1)):
    if i in feature:
        y1[i]=beta[feature.index(i)]
plt.scatter(x,y1,c=colors[0],label=labels[0]);

y2=list(beta_coef)
plt.scatter(x1,y2,c=colors[1],label=labels[1]);
plt.xlabel('Feature No. ');
plt.ylabel('Coefficient');
plt.legend();
plt.title('Coefficient between true model and best model fit by Test subset selection');

```



7. Plot squared root difference of coefficients between best model from test data subset and the true model:

- The squared root difference of coefficient (Left) is very similar to MSE (right) across feature size. However, for squared root diff of coefficient measurement, it shows larger difference at the tail-generally worse performance with large feature size, which is inline with our true model that the extra features are

irrelevant and could overfit the model and generate worse performance on test dataset. Squared root difference evaluation could better capture this **overfit problem**, by directly comparing the difference between true model and predicted model.

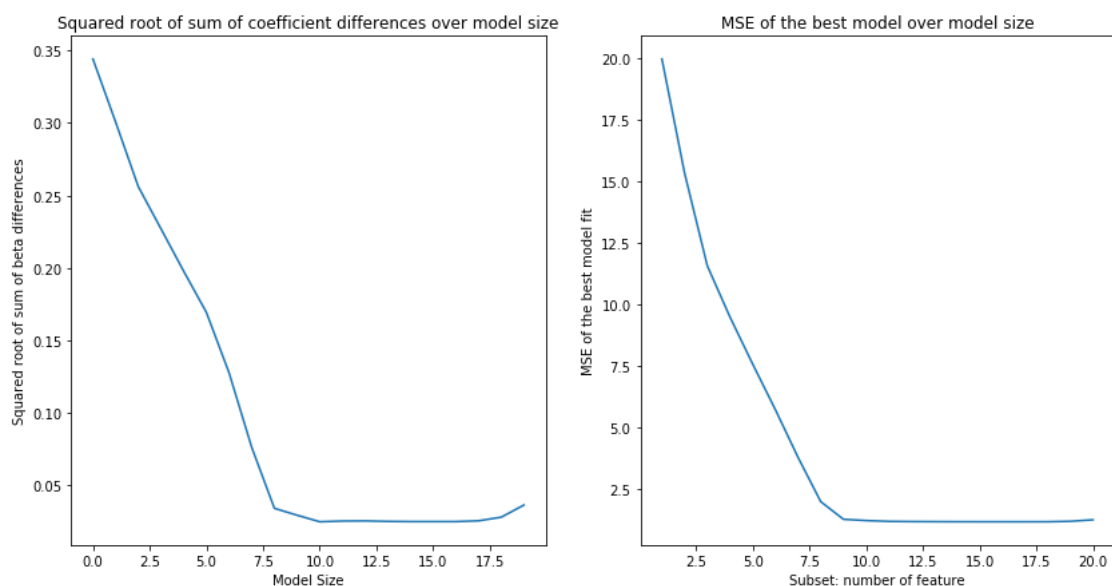
```
#calculate the squared root of coefficient difference
sum_coef_diff=[]
for i in range(20):
    coef_diff0=[]
    #extract the feature-set/beta-info from the best model of each subset
    model_subsetf=list(bestmodel_fsubset_t[i])
    model_beta=list(list(bestmodel_beta_t[i])[0])

    #iterate across 20 features
    for j in range(20):
        if j in model_subsetf:
            #if there is a prediction in the feature,subtract
            beta0=model_beta[model_subsetf.index(j)]
            beta_diff=beta_coef[j]-beta0
        else:
            #if not, equal to the original beta
            beta_diff=beta_coef[j]
        coef_diff0.append(beta_diff*beta_diff)

    #sum and take squared root for each feature size model
    r_coef_diff=math.sqrt(np.mean(np.array(coef_diff0)))
    sum_coef_diff.append(r_coef_diff)
```

```
#plot the Squared root difference of beta and MSE together
plt.figure(figsize=(14,7))
plt.subplot(121)
x=range(20)
plt.plot(x,sum_coef_diff);
plt.xlabel('Model Size');
plt.ylabel('Squared root of sum of beta differences');
plt.title('Squared root of sum of coefficient differences over model size');

plt.subplot(122)
plt.plot(bestmodel_fnum_t,bestmodel_MSE_t);
plt.xlabel('Subset: number of feature');
plt.ylabel('MSE of the best model fit');
plt.title('MSE of the best model over model size');
```



## Application exercise--Predict Egalitarianism

Read the data into DataFrame:

```
df_gss_train=pd.read_csv('data/gss_train.csv')
df_gss_test=pd.read_csv('data/gss_test.csv')
df_gss_train
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	age	attend	authoritarianism	black	born	childs	colath	colrac	colcom	colmil	...	zodiac_GEMINI	zodiac_CANCE
0	21	0	4	0	0	0	1	1	0	1	...	0	0
1	42	0	4	0	0	2	0	1	1	0	...	0	0
2	70	1	1	1	0	3	0	1	1	0	...	0	0
3	35	3	2	0	0	2	0	1	0	1	...	0	0
4	24	3	6	0	1	3	1	1	0	0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1476	61	8	6	0	0	0	1	1	0	0	...	0	0
1477	53	7	6	0	0	0	1	1	0	1	...	0	0
1478	48	7	3	0	0	2	1	1	0	1	...	0	0
1479	37	5	1	0	0	8	0	1	1	0	...	0	0
1480	22	0	0	0	0	0	0	1	1	1	...	0	0

1481 rows × 78 columns

```
#organize the data to fit in model, for feature and predict value
#drop the predict value from the feature set
df_gss_train0=df_gss_train.drop('egalit_scale',axis=1)
df_gss_test0=df_gss_test.drop('egalit_scale',axis=1)

np_gss_train_feature=df_gss_train0.values
np_gss_train_y=df_gss_train['egalit_scale'].values

np_gss_test_feature=df_gss_test0.values
np_gss_test_y=df_gss_test['egalit_scale'].values
```

1. Fit a least squares linear model on training set, and report MSE on test dataset:

```
#generate linear regression model
from sklearn import linear_model
model_lr= linear_model.LinearRegression()
model_lr.fit(np_gss_train_feature,np_gss_train_y)
lg_p_beta=model_lr.coef_

#predict y based on test dataset
p_Y=model_lr.predict(np_gss_test_feature)
lg_p_Y=p_Y

#calculate MSE
y_diff=(np_gss_test_y-p_Y).reshape(p_Y.shape[0])
model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
lg_MSE=model_MSE
```

```
#record all MSE results
MSE_all=[]
method=[]
lamda_all=[]
MSE_all.append(lg_MSE)
lamda_all.append([])
method.append('Linear Regression')
```

```
from sklearn.metrics import mean_squared_error
mean_squared_error(p_Y,np_gss_test_y)
```

63.213629623014995

```
print('test MSE:',round(model_MSE,4))
```

test MSE: 63.2136

2. Fit Ridge Regression on the training set, with lamda chosen by 10-fold cross validation and report test MSE:

```
#generate ridge regression model
from sklearn.linear_model import Ridge,RidgeCV
alphas=np.logspace(-5, 5, 100) #set a range of lammda
ridge_cv = RidgeCV(alphas = alphas, normalize= True, cv = 10) #10-cross validation
```

```
ridge_cv.fit(np_gss_train_feature,np_gss_train_y)
ridge_p_beta=ridge_cv.coef_
ridge_best_alpha = ridge_cv.alpha_
```

```
#predict Y based on test dataset
p_Y=ridge_cv.predict(np_gss_test_feature)
ridge_p_Y=p_Y

#calculate MSE
y_diff=(np_gss_test_y-p_Y).reshape(p_Y.shape[0])
model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
ridge_MSE=model_MSE
```

```
#record all MSE results
MSE_all.append(ridge_MSE)
lamda_all.append(ridge_best_alpha)
method.append('Ridge Regression')
```

```
print('Ridge Test MSE:',round(model_MSE,4))
print('Ridge best Alpha/Lamnda:',round(ridge_best_alpha,4))
```

```
Ridge Test MSE: 61.074
Ridge best Alpha/Lamnda: 0.2205
```

3. Fit Lasso Regression on the training set, with lamnda chosen by 10-fold cross validation and report test MSE and number of non-zero coefficient estimates.

```
#generate lasso regression model
from sklearn.linear_model import Lasso,LassoCV
alphas=np.logspace(-5, 5, 100) #set a range of lammda
lasso_cv = LassoCV(alphas = alphas, normalize= True, cv = 10) #10-cross validation
lasso_cv.fit(np_gss_train_feature,np_gss_train_y)
lasso_p_beta=lasso_cv.coef_
lasso_best_alpha = lasso_cv.alpha_
```

```
#find the non-zero coefficients in Lasso:
lasso_p_beta0=[(i,p) for i,p in enumerate(lasso_p_beta) if p>0]
```

```
#predict Y based on test dataset
p_Y=lasso_cv.predict(np_gss_test_feature)
lasso_p_Y=p_Y
```

```
#calculate MSE
y_diff=(np_gss_test_y-p_Y).reshape(p_Y.shape[0])
model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
lasso_MSE=model_MSE
```

```
#record all MSE results
MSE_all.append(lasso_MSE)
lamda_all.append(lasso_best_alpha)
method.append('Lasso Regression')
```

```
print('Lasso Test MSE:',round(model_MSE,4))
print('Lasso best Alpha/Lamnda:',round(lasso_best_alpha,4))
print('Lasso Num of Non-zero Coefficicents:',len(lasso_p_beta0))
print('Lasso Coefficicents:',lasso_p_beta0)
```

```
Lasso Test MSE: 61.349
Lasso best Alpha/Lamnda: 0.0067
Lasso Num of Non-zero Coefficicents: 12
Lasso Coefficicents: [(3, 0.971654455572674), (5, 0.13891656388208276), (15, 0.19252916745788745), (20, 0.700600502807097), (27, 0.9415469553119209), (28, 0.0935048639324311), (33, 0.17289475760719566), (47, 0.05832206358190714), (64, 0.0030775798812500595), (65, 1.200418232637711), (70, 0.2649203259820222), (75, 0.0855343306801017)]
```

4. Fit Elastic Net Regression on the training set, with alpha and lamnda chosen by 10-fold cross validation and report test MSE and number of non-zero coefficient estimates:

```
#generate Elastic Net regression model
from sklearn.linear_model import ElasticNet,ElasticNetCV
alphas=np.logspace(-5, 5, 100) #set a range of lammda
l1_ratio=np.linspace(0,1,11) #set a range of l1_ratio
en_cv = ElasticNetCV(l1_ratio=l1_ratio, alphas = alphas, normalize= True, cv = 10) #10-cross validation
en_cv.fit(np_gss_train_feature,np_gss_train_y)
```

```
#record the parameters: coefficient, alpha, l1_ratio
en_p_beta=en_cv.coef_
en_best_alpha = en_cv.alpha_
en_best_l1_ratio = en_cv.l1_ratio_
```

```
#find the non-zero coefficients in Lasso:
en_p_beta0=[(i,p) for i,p in enumerate(en_p_beta) if p>0]
```

```
#predict Y based on test dataset
p_Y=en_cv.predict(np_gss_test_feature)
en_p_Y=p_Y

#calculate MSE
y_diff=(np_gss_test_y-p_Y).reshape(p_Y.shape[0])
model_MSE=np.dot(y_diff,y_diff)/y_diff.shape[0]
en_MSE=model_MSE
```

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

```

    tol, rng, random, positive)
C:\Users\zhuangchu\Anaconda3\lib\site-packages\sklearn\linear_model\coordinate_descent.py:471: UserWarning: Coordinate descent with alpha=0
may lead to unexpected results and is discouraged.
    tol, rng, random, positive)
C:\Users\zhuangchu\Anaconda3\lib\site-packages\sklearn\linear_model\coordinate_descent.py:471: ConvergenceWarning: Objective did not
converge. You might want to increase the number of iterations. Duality gap: 37139.20812323103, tolerance: 12.506681020255064
    tol, rng, random, positive)

```

```

#record all MSE results
MSE_all.append(en_MSE)
lamda_all.append(en_best_alpha)
method.append('Elastic Net Regression')

```

```

print('Elastic Net Test MSE:',round(model_MSE,4))
print('Elastic Net Alpha/Lamnda:',round(en_best_alpha,4))
print('Elastic Net L1_ratio:',round(en_best_l1_ratio,4))
print('Elastic Net Num of Non-zero Coefficients:',len(en_p_beta0))
print('Elastic Net Coefficients:',en_p_beta0)

```

```

Elastic Net Test MSE: 61.349
Elastic Net Alpha/Lamnda: 0.0067
Elastic Net L1_ratio: 1.0
Elastic Net Num of Non-zero Coefficients: 12
Elastic Net Coefficients: [(3, 0.971654455572674), (5, 0.13891656388208276), (15, 0.19252916745788745), (20, 0.700600502807097), (27,
0.9415469553119209), (28, 0.0935048639324311), (33, 0.17289475760719566), (47, 0.05832206358190714), (64, 0.0030775798812500595), (65,
1.200418232637711), (70, 0.2649203259820222), (75, 0.0855343306801017)]

```

## 5. Model Comparisons:

From the table below, we could see that Ridge Regression model generally has the best performance, with the smallest MSE score,least square linear regression performs worst, with the highest MSE, while the difference between all 4 methods are relatively small, which might indicate the general stability/confidence of prediction across these modeling techniques.

While with a slightly higher MSE because of smaller number of variables, **Lasso regression** generates the simplest model, only select 12 variables (including,black,num of childs, happy,owngun, spend\_liberal as significant predictors; the spend\_liberal has ths highest coefficient for egalitarianism, which is reasonable; then comes to black people have higher attitudes towards egalitarianism.).

For the **Elastic Net Regression**, from the parameters generated, we could see that l1\_ratio is equal to 1, which means the Elastic Net Regression model generated here is identical to Lasso Regression. However, I am a little bit confused since when L1\_ratio=0-ridge regression, the MSE of which will be smaller than that of Lasso Regression, while Lasso Regression model is selected. Might it because that the algorithm in Sklearn already takes the **1-SE rule** into account? since the difference of MSE is relatively small between these two models, while Lasso could generate the simplest one with L1\_ratio=1.

```

df_MSE_all=pd.DataFrame({'MSE':MSE_all, 'Lamda':lamda_all, 'Method':method})
df_MSE_all

```

```

.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}

```

	MSE	Lamda	Method
0	63.213630	[]	Linear Regression
1	61.074015	0.220513	Ridge Regression
2	61.348952	0.00673415	Lasso Regression
3	61.348952	0.00673415	Elastic Net Regression

```

#plot a bar graph to show MSE
label=method
index=np.arange(len(label))
width=0.3

plt.bar(index,MSE_all,width,color='SkyBlue',label='MSE');
plt.xlabel('Regression Model');
plt.ylabel('MSE');
plt.xticks(index+width/2,label,FontSize=9);
plt.legend();
plt.title('MSE of 4 Regression Models');

```

