Lasso Ridge Regularization

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [3]:
```
from sklearn import preprocessing
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split
```

In [4]:
```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.metrics import r2_score
```

In [5]:
```
data=pd.read_csv(r'C:\Users\yogay\OneDrive\Desktop\Yogita_Yadav\Data Science\8th\TASK-22_LASSO,RIDGE\car-mpg.csv'
```

In [6]:
```
data.head()
```

Out[6]:

	mpg	cyl	disp	hp	wt	acc	yr	origin	car_type	car_name
0	18.0	8	307.0	130	3504	12.0	70	1	0	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	0	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	0	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	0	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	0	ford torino

In [7]:
```
data = data.drop(['car_name'], axis = 1)
data['origin'] = data['origin'].replace({1: 'america', 2: 'europe', 3: 'asia'})
data = pd.get_dummies(data,columns = ['origin'])
data = data.replace('?', np.nan)
data = data.apply(lambda x: x.fillna(x.median()), axis = 0)
```

In [8]:
```
data.head()
```

Out[8]:

	mpg	cyl	disp	hp	wt	acc	yr	car_type	origin_america	origin_asia	origin_europe
0	18.0	8	307.0	130	3504	12.0	70	0	1	0	0
1	15.0	8	350.0	165	3693	11.5	70	0	1	0	0
2	18.0	8	318.0	150	3436	11.0	70	0	1	0	0
3	16.0	8	304.0	150	3433	12.0	70	0	1	0	0
4	17.0	8	302.0	140	3449	10.5	70	0	1	0	0

In [9]:
```
X = data.drop(['mpg'], axis = 1) # independent variable
y = data[['mpg']] #dependent variable
```

In [10]:
```
X_s = preprocessing.scale(X)
X_s = pd.DataFrame(X_s, columns = X.columns)

y_s = preprocessing.scale(y)
y_s = pd.DataFrame(y_s, columns = y.columns)
```

In [11]:
```
X_train, X_test, y_train,y_test = train_test_split(X_s, y_s, test_size = 0.30, random_state = 1)
X_train.shape
```

Out[11]: (278, 10)

In [12]:
```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)

for idx, col_name in enumerate(X_train.columns):
    print('The coefficient for {} is {}'.format(col_name, regression_model.coef_[0][idx]))

intercept = regression_model.intercept_[0]
print('The intercept is {}'.format(intercept))
```

```
The coefficient for cyl is 0.3210223856916103
The coefficient for disp is 0.3248343091848394
The coefficient for hp is -0.22916950059437657
The coefficient for wt is -0.7112101905072294
The coefficient for acc is 0.01471368276419114
The coefficient for yr is 0.37558119495107434
The coefficient for car_type is 0.3814769484233101
The coefficient for origin_america is -0.07472247547584179
The coefficient for origin_asia is 0.044515252035678
The coefficient for origin_europe is 0.04834854953945406
The intercept is 0.019284116103639722
```

In [13]:
```
ridge_model = Ridge(alpha = 0.3)
ridge_model.fit(X_train, y_train)

print('Ridge model coef: {}'.format(ridge_model.coef_))
```

```
Ridge model coef: [[ 0.31649043  0.31320707 -0.22876025 -0.70109447  0.01295851  0.37447352
   0.37725608 -0.07423624  0.04441039  0.04784031]]
```

In [14]:
```
lasso_model = Lasso(alpha = 0.1)
lasso_model.fit(X_train, y_train)
print('Lasso model coef: {}'.format(lasso_model.coef_))
```

```
Lasso model coef: [-0.         -0.         -0.01690287 -0.51890013  0.          0.28138241
  0.1278489  -0.01642647  0.          0.        ]
```

In [15]:

```
#Model score - r^2 or coeff of determinant
#r^2 = 1-(RSS/TSS) = Regression error/TSS


#Simple Linear Model
print(regression_model.score(X_train, y_train))
print(regression_model.score(X_test, y_test))

print('*************************')
#Ridge
print(ridge_model.score(X_train, y_train))
print(ridge_model.score(X_test, y_test))

print('*************************')
#Lasso
print(lasso_model.score(X_train, y_train))
print(lasso_model.score(X_test, y_test))
```

```
0.8343770256960538
0.8513421387780066
*************************
0.8343617931312617
0.8518882171608501
*************************
0.7938010766228453
0.8375229615977084
```

In [16]:
```
data_train_test = pd.concat([X_train, y_train], axis =1)
data_train_test.head()
```

Out[16]:

	cyl	disp	hp	wt	acc	yr	car_type	origin_america	origin_asia	origin_europe	mpg
350	-0.856321	-0.849116	-1.081977	-0.893172	-0.242570	1.351199	0.941412	0.773559	-0.497643	-0.461968	1.432898
59	-0.856321	-0.925936	-1.317736	-0.847061	2.879909	-1.085858	0.941412	-1.292726	-0.497643	2.164651	-0.065919
120	-0.856321	-0.695475	0.201600	-0.121101	-0.024722	-0.815074	0.941412	-1.292726	-0.497643	2.164651	-0.578335
12	1.498191	1.983643	1.197027	0.934732	-2.203196	-1.627426	-1.062235	0.773559	-0.497643	-0.461968	-1.090751
349	-0.856321	-0.983552	-0.951000	-1.165111	0.156817	1.351199	0.941412	-1.292726	2.009471	-0.461968	1.356035

In [17]:
```
import statsmodels.formula.api as smf
ols1 = smf.ols(formula = 'mpg ~ cyl+disp+hp+wt+acc+yr+car_type+origin_america+origin_europe+origin_asia', data = 
ols1.params
```

Out[17]:
```
Intercept         0.019284
cyl               0.321022
disp              0.324834
hp               -0.229170
wt               -0.711210
acc               0.014714
yr                0.375581
car_type          0.381477
origin_america   -0.074722
origin_europe     0.048349
origin_asia       0.044515
dtype: float64
```

In [18]:
```
print(ols1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     150.0
Date:                Wed, 08 Nov 2023   Prob (F-statistic):           3.12e-99
Time:                        23:26:54   Log-Likelihood:                -146.89
No. Observations:                 278   AIC:                             313.8
Df Residuals:                     268   BIC:                             350.1
Df Model:                           9
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          0.0193      0.025      0.765      0.445      -0.030       0.069
cyl                0.3210      0.112      2.856      0.005       0.100       0.542
disp               0.3248      0.128      2.544      0.012       0.073       0.576
hp                -0.2292      0.079     -2.915      0.004      -0.384      -0.074
wt                -0.7112      0.088     -8.118      0.000      -0.884      -0.539
acc                0.0147      0.039      0.373      0.709      -0.063       0.092
yr                 0.3756      0.029     13.088      0.000       0.319       0.432
car_type           0.3815      0.067      5.728      0.000       0.250       0.513
origin_america    -0.0747      0.020     -3.723      0.000      -0.114      -0.035
origin_europe      0.0483      0.021      2.270      0.024       0.006       0.090
origin_asia        0.0445      0.020      2.175      0.031       0.004       0.085
==============================================================================
Omnibus:                       22.678   Durbin-Watson:                   2.105
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.139
Skew:                           0.513   Prob(JB):                     1.42e-08
Kurtosis:                       4.438   Cond. No.                     1.27e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.72e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

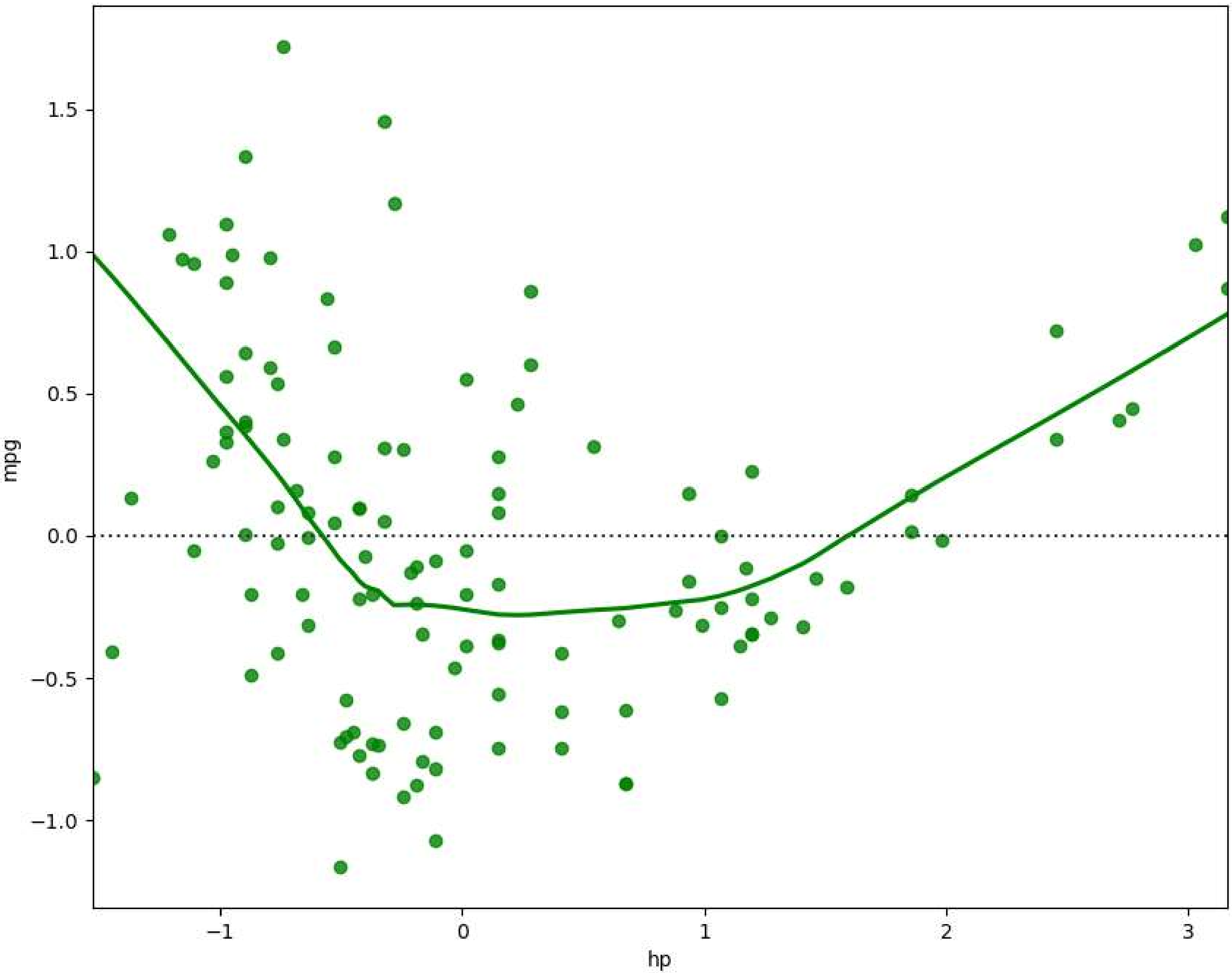
In [19]:
```
mse  = np.mean((regression_model.predict(X_test)-y_test)**2)
import math
rmse = math.sqrt(mse)
print('Root Mean Squared Error: {}'.format(rmse))
```

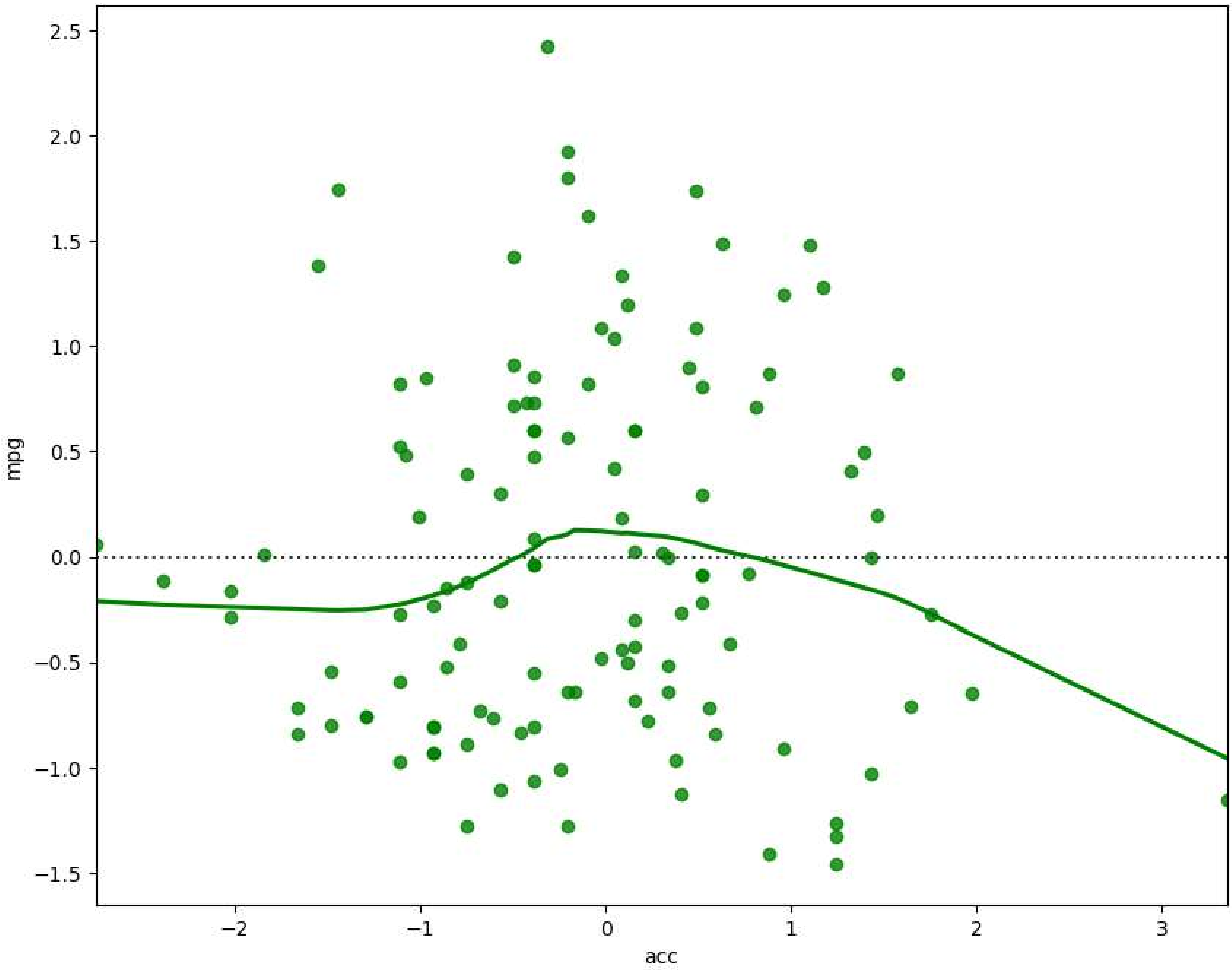
```
Root Mean Squared Error: 0.3776693425408784

C:\Users\yogay\anaconda3\lib\site-packages\numpy\core\fromnumeric.py:3430: FutureWarning: In a future version, D
ataFrame.mean(axis=None) will return a scalar mean over the entire DataFrame. To retain the old behavior, use 'f
rame.mean(axis=0)' or just 'frame.mean()'
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

In [20]:
```
fig = plt.figure(figsize=(10,8))
sns.residplot(x= X_test['hp'], y= y_test['mpg'], color='green', lowess=True )
fig = plt.figure(figsize=(10,8))
sns.residplot(x= X_test['acc'], y= y_test['mpg'], color='green', lowess=True )
```

Out[20]:
```
<Axes: xlabel='acc', ylabel='mpg'>
```


1.5
1.0
0.5
0.0
−0.5
−1.0
mpg
−1
0
1
2
3
hp


2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
mpg
−2
−1
0
1
2
3
acc

In [21]:
```
y_pred = regression_model.predict(X_test)
plt.scatter(y_test['mpg'], y_pred,c='red')
```

Out[21]: <matplotlib.collections.PathCollection at 0x1a0f63e2ef0>