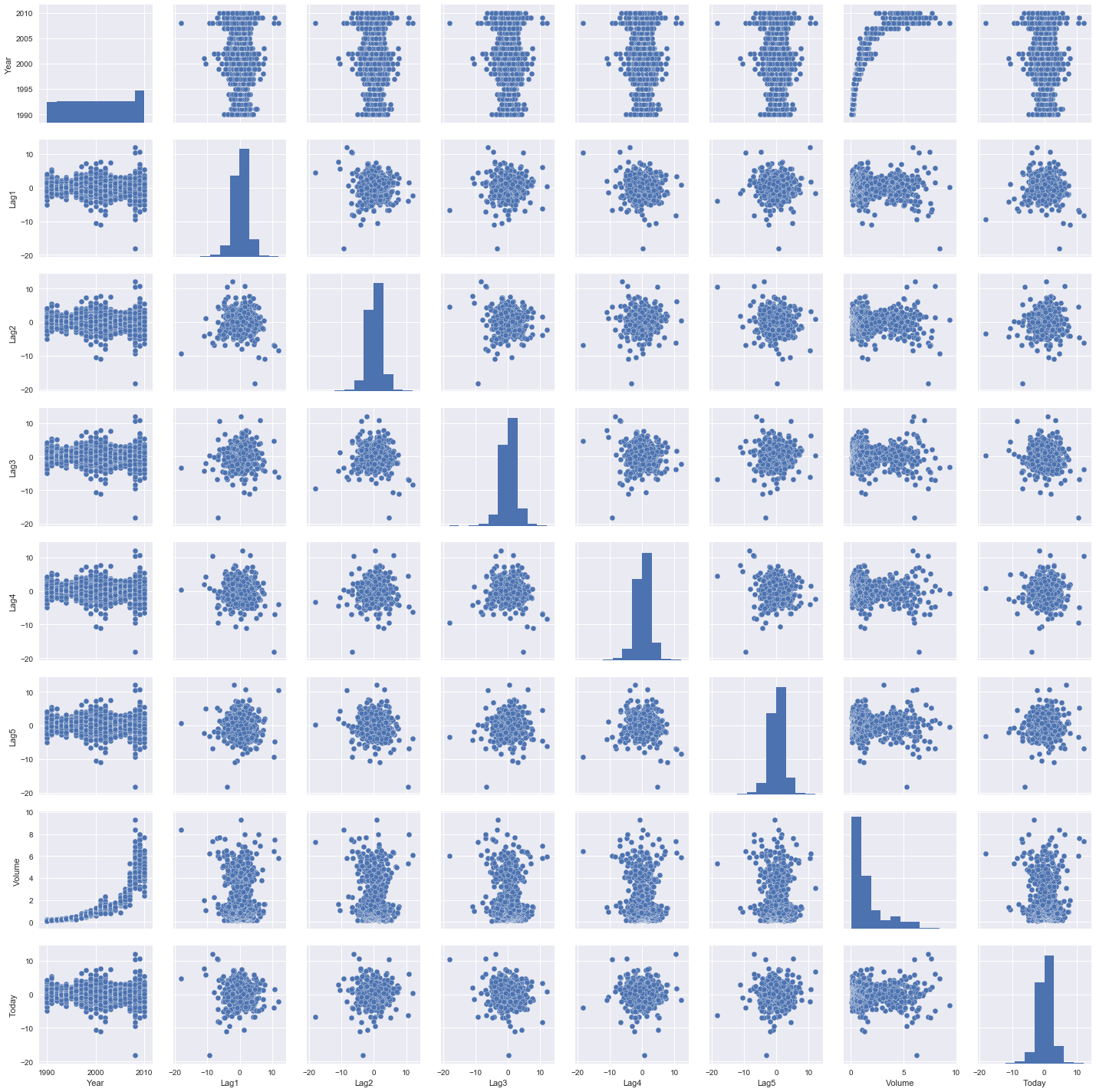
**LAB 5 WRITE UP**

1. (Optional)
2. (Optional)
3. **Chapter 4, Problem 10 (see problem3.ipynb attached for full code)**
   1. Yes, there do seem to be some patterns in the data. Most interesting is that Volume looks to have an exponential relationship with Year. In addition, all the lag variables pairwise scattered against each other seem to form fairly normal distributions, and they also seem to remain mostly static over the years. See problem3a.png (below, attached) and probem3a\_numerical\_summary.txt (attached) for details.



* 1. None of the variables appear statistically significant as far as we can tell.

[[ 55 429]

[ 47 558]]

% Correct Predictions: 56.290174472%

The covariance matrix indicates that the Logistic Regression model is pretty good at correctly predicting weeks when the market truly goes up, but it is poor at predicting weeks that it will go down.

[[ 9 34]

[ 5 56]]

% Correct Predictions: 62.5%

[[ 9 34]

[ 5 56]]

% Correct Predictions: 62.5%

[[ 0 43]

[ 0 61]]

% Correct Predictions: 58.6538461538%

[[21 22]

[31 30]]

% Correct Predictions: 49.0384615385%

* 1. Logistic Regression and LDA perform much better than the other methods and have similar test error.
  2. We experimented with KNN for k = 10 and k = 100, whose confusion matrices are shown below. However, Logistic Regression and LDA from parts d and e respectively still performed better than anything we tried.

KNN with k = 10:

[[22 21]

[24 37]]

% Correct Predictions: 56.7307692308%

KNN with k = 100:

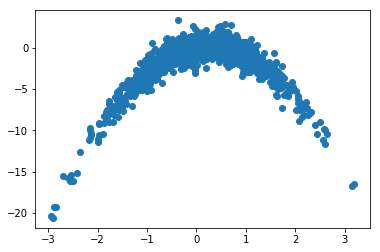
[[11 32]

[14 47]]

% Correct Predictions: 55.7692307692%

1. (Optional)
2. **Chapter 5, Problem 5 (see problem5.ipynb in directory for complete code):**
   1. See code: problem5.ipynb
   2. Initial validation set error: 0.0254
   3. Performing 3 times produced these results:
      1. 0.0254
      2. 0.0248
      3. 0.0238
   4. Using student dummy variable produced: 0.0256. Therefore, you can see that using a dummy variable did not affect the error.
   5. n = 100, p = 2

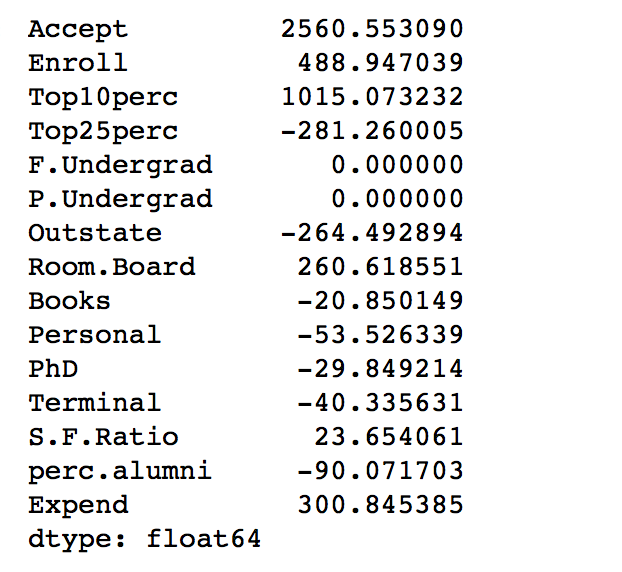
Model: x – 2X2 +noise

* 1. The scatter plot seems to indicate a quadratic relationship between X and Y. See problem6b.png, attached and shown below:
     1. Error with degree-1 polynomial: 2712.71581957
     2. Error with degree-2 polynomial: 182050.20655988
     3. Error with degree-3 polynomial: 12144160.59967541
     4. Error with degree-4 polynomial: 9.55137253e+08
     5. Error with degree-1 polynomial: 16214660.07227357
     6. Error with degree-2 polynomial: 8.66576373e+10
     7. Error with degree-3 polynomial: 1.61773224e+15
     8. Error with degree-4 polynomial: 6.85023463e+19

The results are exactly the same as in part c because LOOCV evaluates all n possible folds of a single observation.

* 1. The model that had the smallest error was the quadratic polynomial (and by that I mean that’s what should have happened, but we wrestled with understanding what we were supposed to be doing for so long that once we finally understood, we didn’t have a chance to fully correct our code), which is to be expected given that the data’s true form (seen in part b) is quadratic.
  2. The values of linear and quadratic beta terms appear statistically significant, which makes sense with our results from part c taken into consideration because of the quadratic nature of the data.

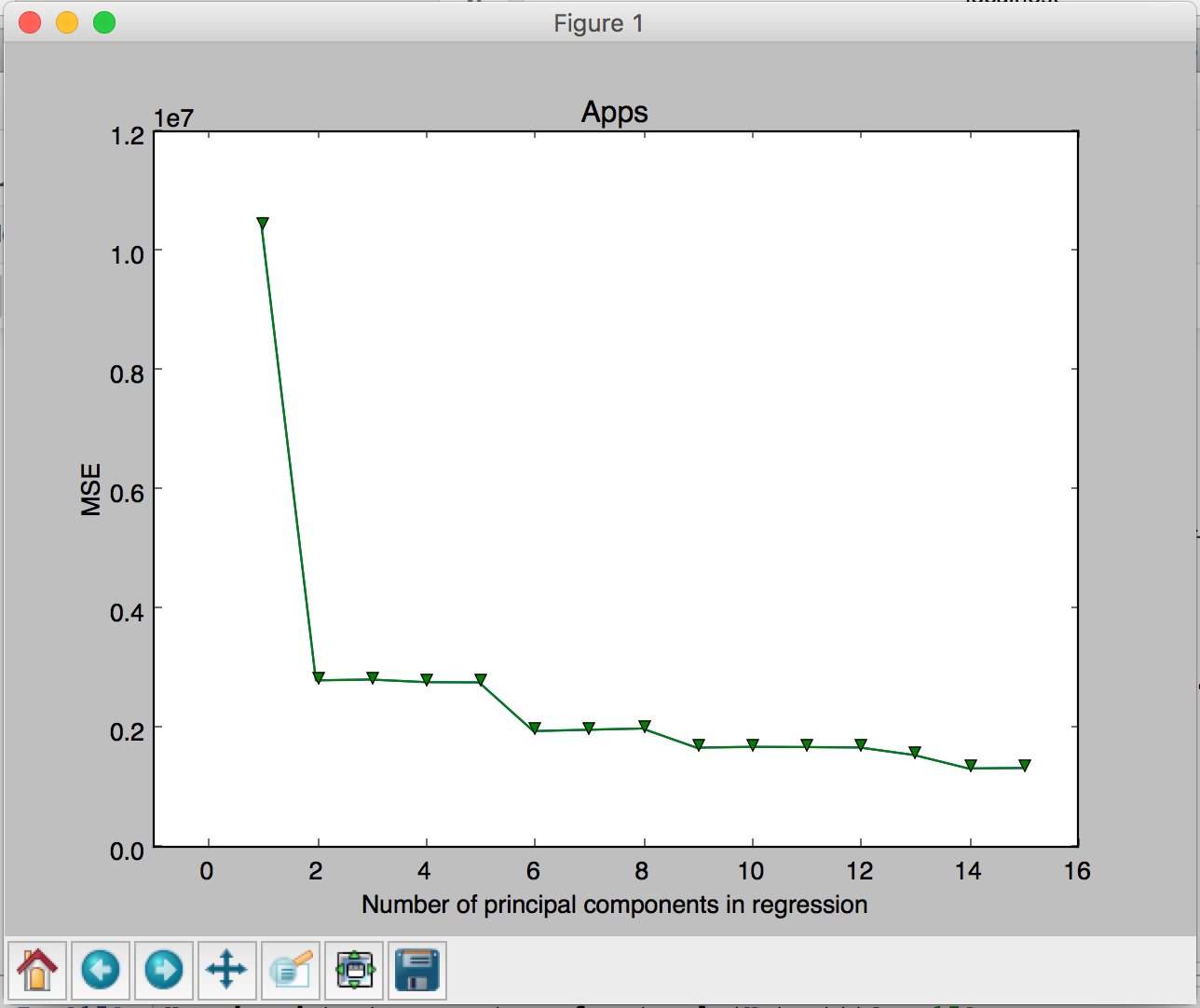
1. **Chapter 6, Problem 9 (see problem7.ipynb in directory for complete code):**
   1. See code: problem7.ipyn
   2. For linear model using least squares test RSS was: 1461087
   3. For ridge regression model, test RSS was: 2269279
   4. For lasso model, test RSS was: 2149904



so 13 non-zero coefficients.

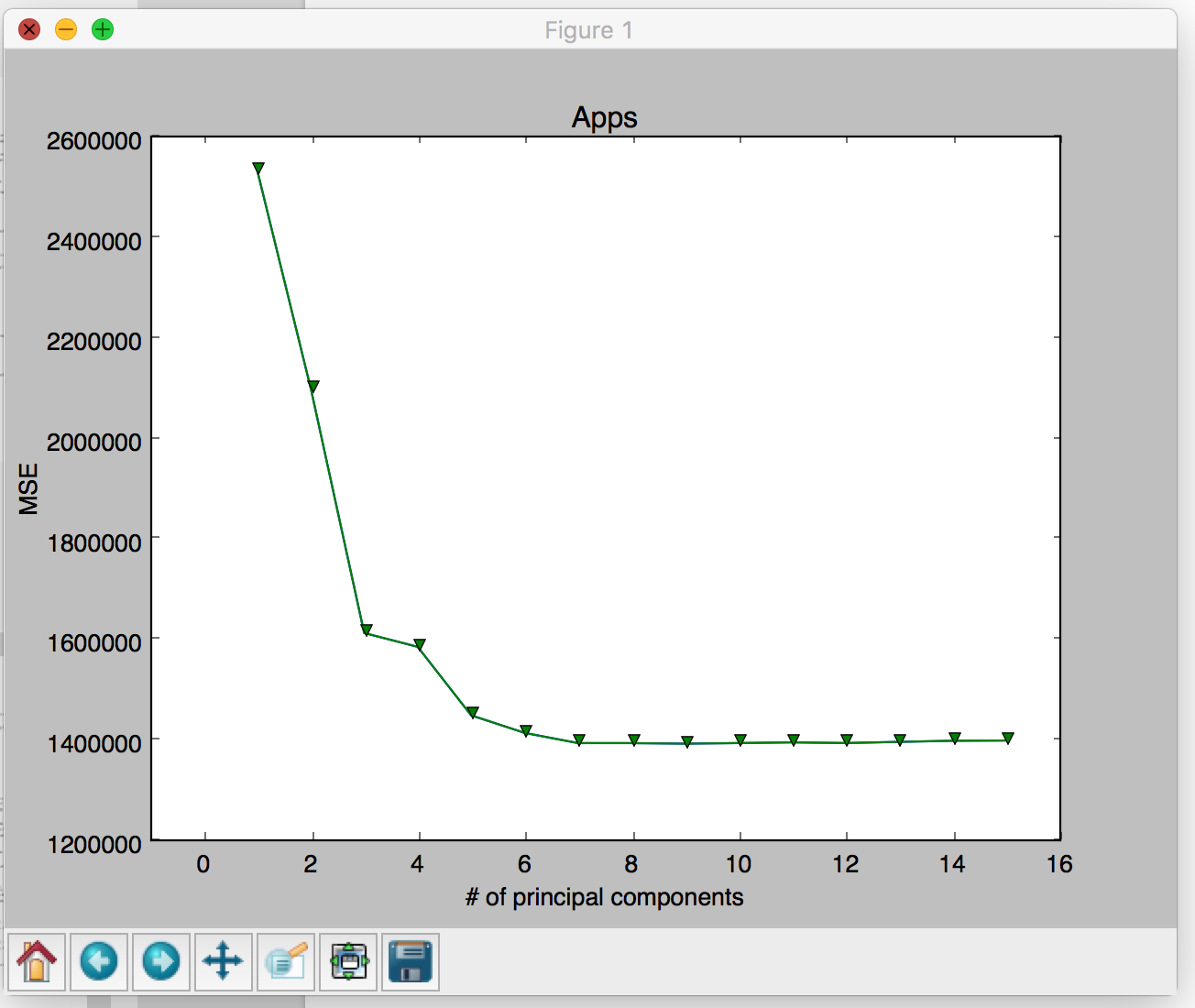
* 1. For PCR, test RSS was: 2176819

MSE Graph for PCR:



* 1. For PLS, test RSS was: 3672607

MSE Graph for PLS:



* 1. The differences between the different models were not very significant. The most significant difference was seen in the Linear Model using least squares mean and PLS model.

1. **Chapter 6, Problem 11 (see problem8.ipynb attached for complete code)**
   1. From one particular split of the data, we got the following results:

Lasso score: 0.473486403448

Ridge score: 0.453813155377

PCR score (1 components): 0.45919580706

PCR score (2 components): 0.396326136743

PCR score (3 components): 0.399295472784

PCR score (4 components): 0.399907825387

PCR score (5 components): 0.413375026782

PCR score (6 components): 0.418992830481

PCR score (7 components): 0.443451163793

PCR score (8 components): 0.453709100546

PCR score (9 components): 0.460451390705

PCR score (10 components): 0.453283370737

PCR score (11 components): 0.455853127185

PCR score (12 components): 0.458447790115

PCR score (13 components): 0.426165918575

Lasso score (10 components): 0.47109398332

Lasso Regression seems to perform the best overall. PCR seems to show the best performance is when 9 or 10 features of the data are used as predictors for crime rate.

* 1. Lasso Regression models using 13 features of the data seem to perform best. This is because it yielded better results than any other model, and despite our PCR results seem to suggest that using 9 or 10 features as predictors is better than using all 13, performing Lasso regression with all 13 still yields better results than with 9 or 10 features. For validation set error for this model, see part A or run attached code (problem8.ipynb).
  2. Our chosen model (Lasso Regression) uses all 13 features in the data set, for the simple reason that it consistently yields the lowest validation error despite PCR suggesting that using 9 or 10 features would be better.