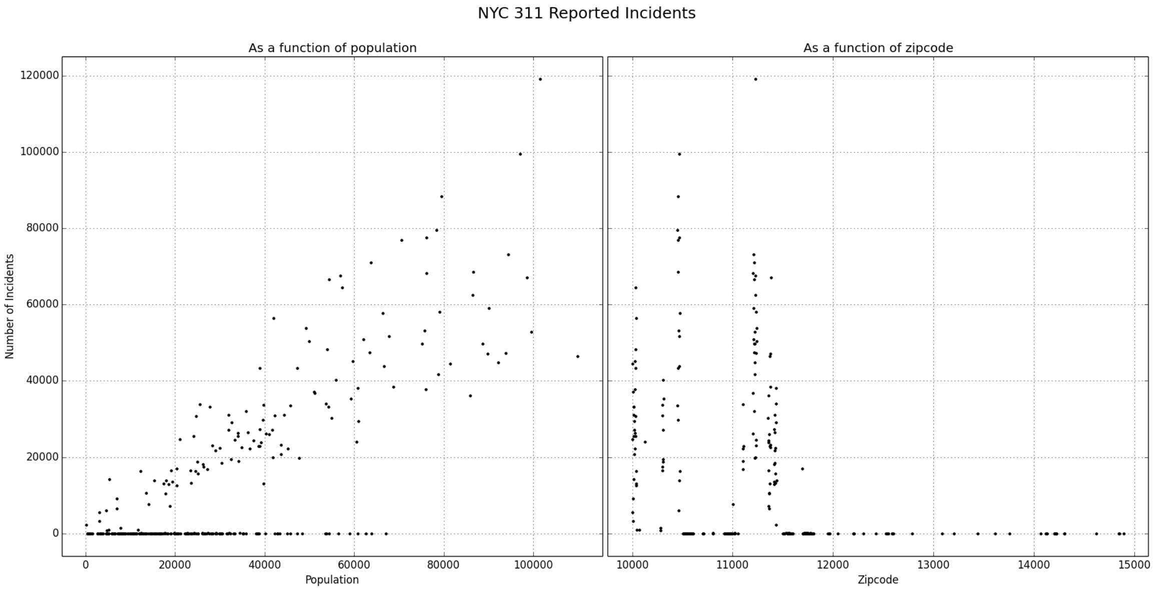
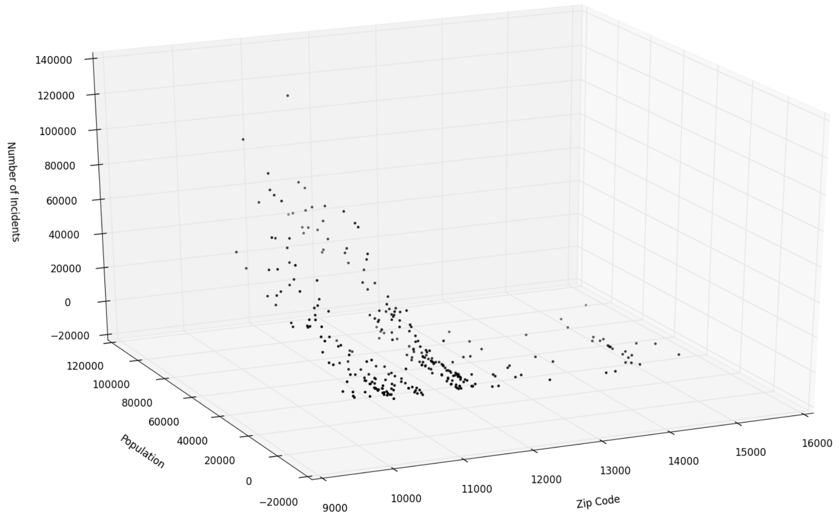
Michael Musick
CUSP 5003
Assignment 6

Problem_a:





The data seems semi-correlated between population and the number of incidents reported. Visually, this appears as though it will likely be a linear relationship. However, I fear the large number of zip codes, which only have a single incident to report, will skew the data. It would seem as though it will be necessary to find a data feature that could explain those low incident level zip codes.

I am not sure if the zip code numbers themselves will prove useful in providing a feature that increases the predictive power of the final models.

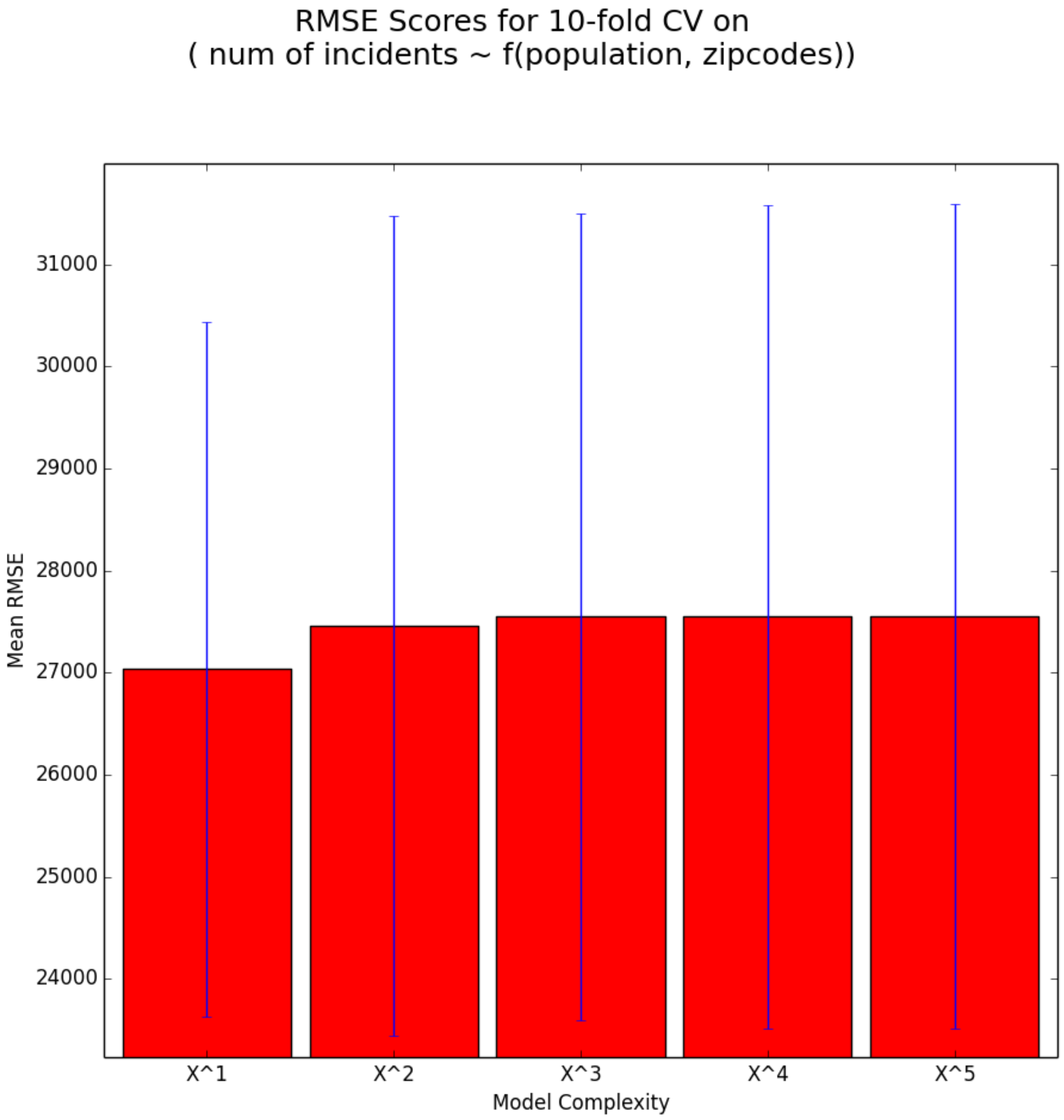
Problem_b:
When the OLS models are run through polynomials of increasing complexity between 1st and 5th, with only "number of incidents" being a predicted function of population, my code converges towards a 2nd degree polynomial being the best option.

```
Degree      RMSE        R^2
   1        13470.3     0.556027428638
   2        13111.8     0.578497376421
   3        13008.1     0.572377430657
   4        13030.9     0.570643937527
   5        13139.8     0.56627416975
```

Above is an example output from my code.

As you can see, the degree 2 fit produces the best R^2 and lowest RMSE values.

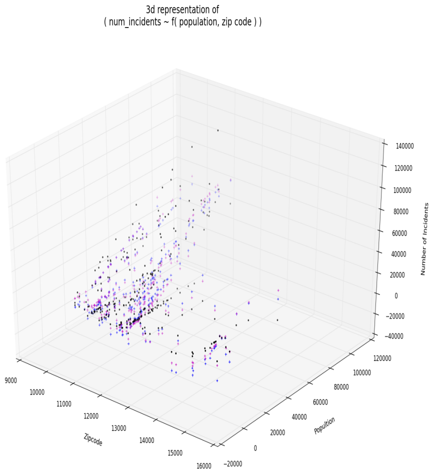
Problem_c:
I recently became slightly confused as to what was expected of this problem. So I will present how I originally interpreted it and my results. I believed that the request to use the "whole training set" meant to include the zip code as a feature for potential predictions. So that is how I have developed my code for this problem. The system is trained with a 10-fold cross validation method, using both zip code and population as predictive features.

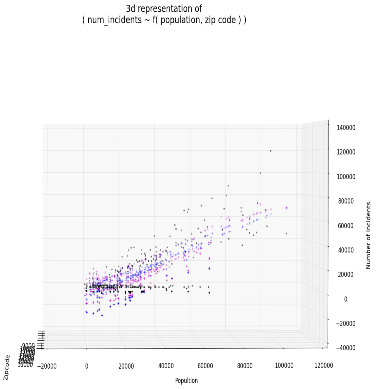


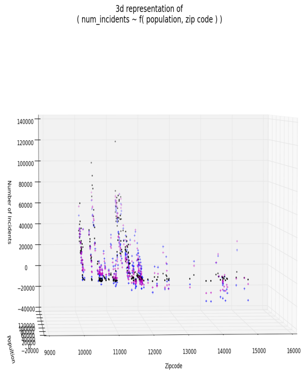
The bar chart showing the RMSE scores of each complexity level consistently returns degree 1 as the best fitting OLS model, with the least amount of standard error present. It would seem that the addition of zip codes as a predictive feature makes fitting a multi-dimensional OLS model more difficult for the system.

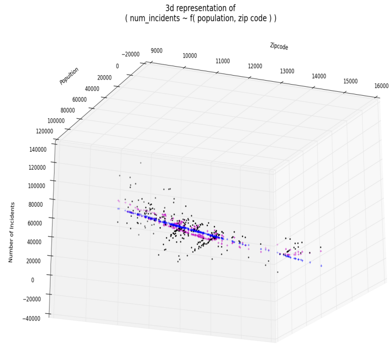
As mentioned I did use zip code as a predictor. I feel justified in this because zip codes in the NYC area do roughly correlate to geographical space and serve as a general predictor of location relationships.

Below are 3d representations of this same run, visualizing the final cross-validation set. The black dots represent the labeled data. The blue '+' are the X^1 OLS predictions and the purple '+' are the X^5 predictions. The first image is from a general angle trying to visualize all three features. The second image is oriented so that data is presented as a function of population. The third image is oriented so that the data is presented as a function of zip codes. The final image is oriented so that you are able to see the predicted plane created from the X^1 model as it slices through both predictive data sets. This final image also shows the relationship of the X^5 model going slightly above and below the X^1 model.









Problem_d:

My final model uses the population and zip code features supplied for the problem set. Additionally, a data set of “Adjusted Gross Income” (AGI) per zip code, was added. This data set was collected from the IRS’ website for Statistics of Income page (SOI).[1] This data set was reduced by hand in Microsoft Excel so as to not take up a large amount of space in the GIT repository. It was also saved as a .txt file because of strange encodings that made it difficult to import to python. This file is saved as

[1] http://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-Free-ZIP-Code-data-(SOI)

‘ny_state_AGI_2.txt’ in my directory. The function for importing to python is located near the end of my python code.

Based on my problem_c results I decided to use a X^1 OLS model.

Here are my predictions:

```
zipcode 11692 predicted # of incidents: 5237.62375358
zipcode 11520 predicted # of incidents: 22302.9746274
zipcode 11372 predicted # of incidents: 39399.2587484
zipcode 11801 predicted # of incidents: 17731.1874597
zipcode 10566 predicted # of incidents: 14095.9756027
zipcode 10003 predicted # of incidents: 30874.1831761
zipcode 11361 predicted # of incidents: 13089.5412288
zipcode 11005 predicted # of incidents: -2925.57479538
zipcode 11716 predicted # of incidents: -740.940726633
zipcode 14710 predicted # of incidents: -21252.3212314
zipcode 11023 predicted # of incidents: 1131.13543464
zipcode 14450 predicted # of incidents: 4084.46160581
zipcode 11221 predicted # of incidents: 49686.1877029
zipcode 13790 predicted # of incidents: -5744.89165847
zipcode 11426 predicted # of incidents: 5495.31274182
zipcode 11416 predicted # of incidents: 11053.3672698
zipcode 10504 predicted # of incidents: 486.314865639
zipcode 10030 predicted # of incidents: 19854.5905677
zipcode 11565 predicted # of incidents: -1465.39496219
zipcode 10452 predicted # of incidents: 51388.5988016
zipcode 11766 predicted # of incidents: -151.227780185
zipcode 11434 predicted # of incidents: 34008.285907
zipcode 14209 predicted # of incidents: -15652.8708534
zipcode 10017 predicted # of incidents: 6103.27555555
zipcode 11364 predicted # of incidents: 16834.3544426
zipcode 11419 predicted # of incidents: 26460.9011776
zipcode 11003 predicted # of incidents: 23704.8736351
zipcode 10031 predicted # of incidents: 40028.1420443
zipcode 10952 predicted # of incidents: 22296.115815
zipcode 10471 predicted # of incidents: 13218.366019
zipcode 11558 predicted # of incidents: -1405.06074997
zipcode 10162 predicted # of incidents: 1445.30046548
zipcode 11233 predicted # of incidents: 41293.8298015
zipcode 10461 predicted # of incidents: 33157.6186077
zipcode 11206 predicted # of incidents: 51761.3238773
zipcode 10307 predicted # of incidents: 8958.77621361
zipcode 11234 predicted # of incidents: 53182.1877176
zipcode 10460 predicted # of incidents: 38839.061882
zipcode 14203 predicted # of incidents: -19841.3920452
zipcode 11720 predicted # of incidents: 11399.1495525
zipcode 11421 predicted # of incidents: 20594.8221413
zipcode 10473 predicted # of incidents: 38982.8063735
zipcode 11428 predicted # of incidents: 6790.29537931
zipcode 11559 predicted # of incidents: -2742.02498756
zipcode 11693 predicted # of incidents: 476.592643677
zipcode 14051 predicted # of incidents: -8005.53169433
zipcode 11096 predicted # of incidents: 1327.5152162
zipcode 14467 predicted # of incidents: -16292.1723616
zipcode 11964 predicted # of incidents: -7857.26508193
zipcode 10512 predicted # of incidents: 15141.8226316
zipcode 11768 predicted # of incidents: 4977.63170425
zipcode 10010 predicted # of incidents: 17881.2403797
zipcode 10308 predicted # of incidents: 17635.8012015
zipcode 11577 predicted # of incidents: 134.857749107
zipcode 10502 predicted # of incidents: 1977.21609453
zipcode 10583 predicted # of incidents: 13505.4464468
zipcode 11103 predicted # of incidents: 21769.5075354
zipcode 11209 predicted # of incidents: 40079.8132161
zipcode 11797 predicted # of incidents: -3654.36145659
zipcode 10941 predicted # of incidents: 5587.16718681
zipcode 11239 predicted # of incidents: 4002.54489415
zipcode 11557 predicted # of incidents: -2721.36294676
zipcode 11581 predicted # of incidents: 6361.62741416
zipcode 11697 predicted # of incidents: -5029.41068875
zipcode 11030 predicted # of incidents: 3210.47556924
zipcode 11229 predicted # of incidents: 48725.0063317
zipcode 11772 predicted # of incidents: 21796.3694511
zipcode 10530 predicted # of incidents: 6139.13936093
zipcode 10019 predicted # of incidents: 22294.4672353
zipcode 10314 predicted # of incidents: 56130.4123543
zipcode 11373 predicted # of incidents: 63658.3679577
zipcode 11948 predicted # of incidents: -8208.61439378
zipcode 10550 predicted # of incidents: 23971.3043646
zipcode 11040 predicted # of incidents: 21578.2437233
zipcode 14052 predicted # of incidents: -8950.6523267
zipcode 10553 predicted # of incidents: 5306.91366037
zipcode 10603 predicted # of incidents: 9118.38668251
zipcode 11207 predicted # of incidents: 59819.2887364
zipcode 11703 predicted # of incidents: 3085.23304063
zipcode 10580 predicted # of incidents: 3299.70776691
zipcode 10598 predicted # of incidents: 15993.2019105
zipcode 12553 predicted # of incidents: 3984.79540767
zipcode 10025 predicted # of incidents: 58197.4033724
zipcode 11412 predicted # of incidents: 17620.3502107
zipcode 13057 predicted # of incidents: -4841.85483706
zipcode 11001 predicted # of incidents: 13392.0008339
zipcode 10024 predicted # of incidents: 27455.7181804
zipcode 10570 predicted # of incidents: 5725.34917332
zipcode 10463 predicted # of incidents: 44110.9008975
zipcode 11378 predicted # of incidents: 17687.7501729
zipcode 10701 predicted # of incidents: 40823.9173267
zipcode 10921 predicted # of incidents: -760.941708176
zipcode 14009 predicted # of incidents: -16054.1208984
zipcode 11691 predicted # of incidents: 33980.9571948
zipcode 10465 predicted # of incidents: 27205.3820294
zipcode 10475 predicted # of incidents: 26392.5165528
zipcode 11714 predicted # of incidents: 6995.85911837
zipcode 11741 predicted # of incidents: 10065.3316244
zipcode 14614 predicted # of incidents: -22243.8910649
zipcode 14701 predicted # of incidents: 4317.1644582
zipcode 11212 predicted # of incidents: 53618.6174283
zipcode 14228 predicted # of incidents: -7072.07828056
zipcode 11365 predicted # of incidents: 22411.6604144
zipcode 10801 predicted # of incidents: 24336.2572006
zipcode 10005 predicted # of incidents: 490.929130841
zipcode 11042 predicted # of incidents: -3936.81592376
zipcode 11222 predicted # of incidents: 19636.7197579
zipcode 11204 predicted # of incidents: 48232.2624186
zipcode 11789 predicted # of incidents: -2866.5000016
zipcode 10454 predicted # of incidents: 25159.4427305
zipcode 10009 predicted # of incidents: 41436.636961
zipcode 11580 predicted # of incidents: 19425.7358078
zipcode 11561 predicted # of incidents: 16609.5579396
zipcode 11590 predicted # of incidents: 23072.5657484
```

Problem_e

My expected average RMSE for this OLS model is 12246.3. After further testing, the data set I ultimately pulled in (AGI) seems too highly correlated with population to be of much use in training my model. It would have been more useful to explore additional training sets that showed low covariance with the population feature. Additionally, I should have found a better feature to represent geographical location, such as actual lat/lng values for the center of each zip code. What slowed me down was the collection of this AGI data set. Originally, I had sought out to collect a median income per zip code data set. However, this data only seemed to exist in relation to Census tract numbers, not zip codes. Therefore the re-mapping would have been incredibly time consuming. As mentioned previously, the data set I did find proved very difficult to import directly, and also required a large amount of “massaging” in order to even use.

The other features that I would have liked to explore included racial, and education data. I feel as though these might have proved more useful in the training of my model. To summarize these points, I really needed to spend more time collecting data and searching for sets that described these trends in unique ways, with low correlation to the population data.