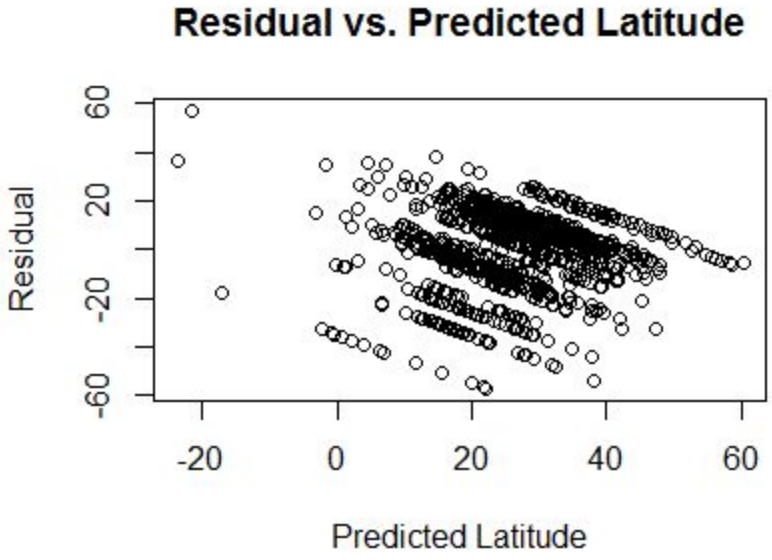
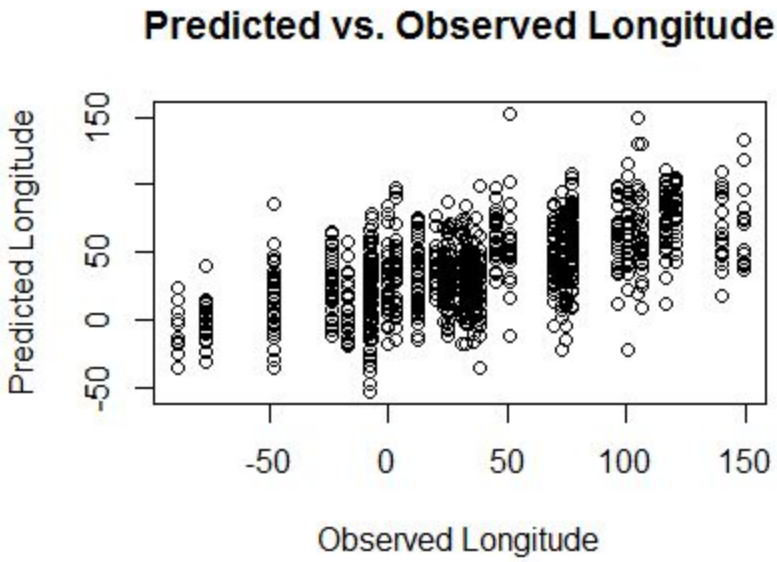
Assignment 6 Report
Harrison Kiang hkiang2, Umberto Ravaioli urjav2, Annlin Sheih sheih2

Q1.

1) Build a straightforward linear regression of latitude (resp. longitude) against features. What is the R-squared? Plot a graph evaluating each regression. (see regression.R)

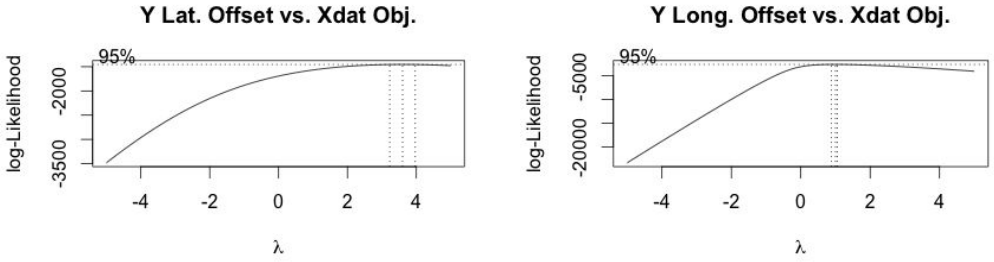
   R-squared: 0.3645767





2) Does a Box-Cox transformation improve the regressions? **Notice that the dependent variable has some negative values, which Box-Cox doesn't like. You can deal with this by remembering that these are angles, so you get to choose the origin.** Why do you say so? For the rest of the exercise, use the transformation if it does improve things, otherwise, use the raw data. (see regression.R)

Offset of 90 was chosen to generate the below graphs:



$\lambda_{max\_log\_likelihood} = 3.6$ $\lambda_{max\_log\_likelihood} = 1$

Box-Cox transformation does not improve the regressions. Box-Cox produced values $R^2_{lat} = 0.248$ and $R^2_{long} = 0.365$, which are not higher than the values produced by the previously performed linear regression.

3) Use glmnet to produce… (see q1.R) Note: a 70/30 split was used

Unregularized regression

Longitude mean square error: 279.5661

Latitude mean square error: 1978.276

a) A regression regularized by L2 (equivalently, a ridge regression). You should estimate the regularization coefficient that produces the minimum error. Is the regularized regression better than the unregularized regression?

Latitude

Alpha	Best Regularization Coefficient	Mean Square Error
0	4.1074558	279.8826
0.1	3.2550885	281.7320
0.2	0.9313406	281.2472

Longitude

Alpha	Best Regularization Coefficient	Mean Square Error
0	9.587934	1977.381
0.1	4.348012	2014.856
0.2	1.980873	2015.120

A ridge regression with α = 0 is comparable with the unregularized regression.

b) A regression regularized by L1 (equivalently, a lasso regression). You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?

Latitude

Alpha	Best Regularization Coefficient	Mean Square Error
0.8	0.3378040	309.1046
0.9	0.6936631	306.0854
1	0.5183021	306.7261

Longitude

Alpha	Best Regularization Coefficient	Mean Square Error
0.8	0.4952183	2060.811
0.9	0.5302153	2061.518
1	0.4348012	2060.375

A lasso regression with α = 1 is worse than the unregularized regression.

c) A regression regularized by elastic net (equivalently, a regression regularized by a convex combination of L1 and L2). Try three values of alpha, the weight setting how big L1 and L2 are. You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?

Latitude

Alpha	Best Regularization Coefficient	Mean Square Error
0.4	1.2957553	306.3296
0.5	1.0366043	306.4010
0.6	0.8638369	306.4822

Longitude

Alpha	Best Regularization Coefficient	Mean Square Error
0.4	1.4369538	2074.148
0.5	0.7219591	2061.794
0.6	0.9579692	2070.515

A regression regularized by elastic net with $\alpha = 0.5$ is worse than the unregularized regression.

Q2.

The unregularized regression ended up yielding the best accuracy result. This is likely due to the fact that outlier points were not removed, potentially interfering with the ability of the regularization to better model the data.

Lasso regression outperformed ridge regression with this dataset. The elastic net regression did not vary too greatly with different values of alpha. The small differences peaked at an alpha value of 0.7

Without removing outlier points, unregularized regression is the best strategy for this problem.

Accuracy for each regularization scheme (80-20 train-test split) in classification

Unregularized:

- Train: 81.0%
- Test: 81.1%

Lasso:

- Regularization Constant: 0.000668381
- Train: 78.95%
- Test: 79.22%

Ridge:

- Regularization Constant: 0.01473867
- Train: 78.7%
- Test: 78.95%

Elastic Net:

- alpha = 0.1:
  - Regularization Coef: 0.0034849438
  - Train: 78.888%
  - Test: 79.2%
- alpha = 0.2:
  - Regularization Coef: 0.0027745080
  - Train: 78.892%
  - Test: 79.2%
- alpha = 0.3:
  - Regularization Coef: 0.0020300128
  - Train: 78.921%
  - Test: 79.2%
- alpha = 0.4:
  - Regularization Coef: 0.0018338682
  - Train: 78.904
  - Test: 79.2%
- alpha = 0.5:
  - Regularization Coef: 0.0009213781
  - Train: 78.971%
  - Test: 79.233%
- alpha = 0.6:
  - Regularization Coef: 0.0011139683
  - Train: 78.933%
  - Test: 79.2%
- alpha = 0.7:
  - Regularization Coef: 0.0005463888

Train: 78.996%
Test: 79.233%

alpha = 0.8:

Regularization Coef: 0.0010063340
Train: 78.925%
Test: 79.2%

alpha = 0.9:

Regularization Coef: 0.0008150525
Train: 78.938%
Test: 79.183%