Notes:

1. OLS of full data set shown NO significant predictors, even if alcohol is legal or not in a country has no predictive value

2. Countries with >80% Islamic faith were removed. This makes the model of little predictive value for nations with high % Islamic faith, but provides better correlations for all other countries.

3. The data set cannot be split into two sets (training and test). After splitting half and half, the training set has 5 significant predictors, the test set only 1. There is not enough data, and the relationships too weak, to split the data up.

4. Using the full data set (<80% ISLAMPCT), OLS gives R-squared fo 0.23 and unemployment, higen, and sunni, and shia are now all significant.

5. BoxCox suggests square root

6. Best subset with sqrt target. BIC 2 predictors, Unemployment and latitude

Cp has more including JDPCT, ISGEN PCT, HIGENPCT, SHIApct, and NOREPPCT.

7. Best subset with 10fold cross validation shows best model with 5 or 6 predictors.

5 predictors are:

(Intercept) Avg\_Per\_Unemp ISGENPCT HIGENPCT ShiaPCT LATITUDE

-5.204170e-18 1.058967e+00 9.162962e-03 1.725234e-02 4.784845e-03 1.478791e+00

8. Lasso with 10fold cross validation give only the intercept.

9. 10 fold cross validation is most reliable, as we have no test set, the cross validation is important. Fitting of this 5 predictor model gives a n adusted R2 or 0.1.

(Intercept) Avg\_Per\_Unemp ISGENPCT HIGENPCT ShiaPCT LATITUDE

3.327275e-17 1.058967e+00 9.162962e-03 1.725234e-02 4.784845e-03 1.478791e+00

10. The QQplot and residuals versus fitted values look okay. I think there is no need to do the sqrt transformation, although the difference is very slight.

Coefficients: ## NOT SQRT, not standardized and centered.

Estimate Std. Error t value Pr(>|t|)

(Intercept) 27.05561 2.10919 12.828 <2e-16 \*\*\*

Avg\_Per\_Unemp 0.43407 0.18397 2.359 0.0198 \*

ISGENPCT 7.42811 7.00753 1.060 0.2911

HIGENPCT 23.46109 9.48779 2.473 0.0147 \*

ShiaPCT 103.72644 46.61358 2.225 0.0278 \*

LATITUDE 0.03060 0.04415 0.693 0.4895

**Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Predicted | High | Low |
| Miami | 30.07782 | 55.31779 | 4.837855 |
| Detriot | 42.09314 | 68.30498 | 15.881304 |
| Omaha City | 30.48065 | 55.78614 | 5.175171 |
| SanFrancisco | 33.84767 | 59.13495 | 8.560395 |
| Albuquerque | 31.29539 | 56.55462 | 6.03616 |
| Boston | 32.23274 | 57.53157 | 6.933911 |
| NewYorkCity | 34.39789 | 59.7027 | 9.09308 |
| Munich | 31.71961 | 57.08642 | 6.352799 |
| Beijing | 31.02369 | 56.44064 | 5.606745 |
| Mumbai | 66.60228 | 98.33549 | 34.869072 |
| Vladivostock | 31.23637 | 56.54388 | 5.928854 |

11. See predictions CSV for details. Taking ratio of consumption relative to NY, only Mumbai, at a 70% CI, can be stated to drink more than NYC. At a 95% CI, all ratios may be greater or less than 1. No city can be said to definitively drink less than NYC.