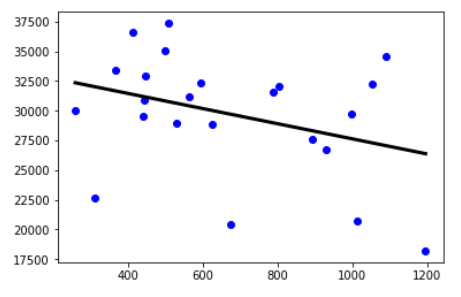
# **Machine Learning**

The Python code for this section is in the Jupyter notebook titled, “03\_Machine Learning.” With a continous variable for a target, machine learning for this project is limited to regression.

## *Linear regression with a single feature, the overall winter severity index*

The revised hypothesis – that the weather index is negatively predictive of admissions for alcoholism – was tested by conducting regression using only the AWSSI and overall alcoholism admissions. The plot of that regression line, Graphic 19, shows that the AWSSI data are sparse and not especially predictive of the target.



*Graphic 19: Regression line for AWSSI and alcoholsim admissions.*

The *R*2, at 0.12, indicates that with linear regression using only the primary feature, the overall winter severity index, very little variance in the target is predicted by the feature. A train/test split analysis computed a test set score of only -0.01, a harbinger of all subsequent analyses. The conclusion for this first analysis is that the winter severity index alone has no value for developing a predictive algorithm for alcoholism admissions.

For this and subsequent regression analyses, linear regression coefficients (slopes) and intercepts were computed. Since there is a negative relationship between all features (winter weather severity index, temperature score, snow score, length, start scale, and end scale) and the target and its larger subgroups (overall alcoholism admissions and demographic breakdowns of the overall count), most of the regression coefficients were negative; exceptions are noted.

## *Ordinary least squares regression using all available features indiscriminately*

* 1. *Large RMSE*

The six features are: AWSSI, TempScore, SnowScore, Length, start\_scale and end\_scale. This regression yielded worse results: An R2 of -0.03. Additionally, the root mean squared error, the standard deviation of the differences between predicted values and observed values in the same units as the response variable (in this case, in units of admissions for alcoholism), was quite high. It was 4,712, for a target with values ranging from 18,000 to 35,000.

* 1. *Improved training score due to overfit*

This analysis generated a training set score of 0.34 and a test set score of -0.14. Compared to to linear regression using only one feature, using all available features increases the training set score, as would be expected with a more complex model that more accurately describes the specific characteristics of the training sample, but does poorly on another sample, i.e., the test data. This is because the model overfits to the training set, and is not generalizable to the test set, or other data.

* 1. *Small coefficients reflect minor effects of most features.*

The coefficients are shown below. Most were very small, reflecting small effects on the target, and the positive slopes were the smallest in terms of absolute value. These small coefficients reflect the minimal effect on the target these features have when combined with all available features in this regression analysis. Notably, in this analysis, the features for length of winter and the start of winter were larger by a factor of 10 relative to all but one of the other coefficients, and by a factor of 6 for that last one.

AWSSI -6

TempScore -11.61

SnowScore 5.61

Length -70.94

Start\_scale -76.93

End\_scale 6

The intercept is: 48716.86

## *Ridge regression*

Considering the small coefficients calculated in the linear regression analysis above, ridge regression would be expected to produce similar train/test scores. Indeed, they’re identical: training set score of 0.34, test set score of -0.14.

That’s because ridge regression coefficients are chosen to be as small as possible, as well as predict on the training data. The features used for this project have such small predictive strength as it is – which means the coefficients are small – the addition of that as a coefficient selection criteria in the ridge regression analysis made little difference.

## *Lasso regression*

Using an alpha of 1.0 and increasing the maximum iterations to 100,000 produced a model that converged and showed the features for length of winter and start of winter to be the two most predictive under the regression model that constrains some coefficients to zero, i.e., ignores them entirely. Obviously, the outcome from the lasso model, indicating the relative predictive strength of winter length and the start date, is just what ordinary least squares regression indicated. These similarities reflect the minimal predictive power of the features in this analysis; the results all come down to the same meaningless regression outcomes.

Conducting ordinary least squares regression using the two most predictive features identified by the lasso regression resulted in the coefficients and intercept below, shown with the coefficients and intercept resulting from OLS that included all features (2, above):

Two Features Only With Other Four Features

Length -104.85 -70.94

Start\_scale -59.81 -76.93

Intercept: 47749.63 48716.86

Similar intercepts and slightly more variance associated with length and start of winter for the regression using these features alone is to be expected. The variance associated from the other for features that the lasso regression model reduced to zero is now associated with length and start scale.

All results:

lr.coef\_ -104.85, -59.81  
lr.intercept\_ 47749.63  
Training set score 0.27  
Test set score -0.05  
R^2 -0.05  
Root Mean Squared Error 6315.72

The RMSE for this model, with only two of the six features, is almost 50% greater than the RMSE for the regression model using all available features. This is explained by the overfit of the all-feature model to the train data set. This model was more generalizable, with a R^2 that was greater than the overfitting all-feature model. However, this difference was only .02, and the two models are both beyond useless for predictive purposes, so the difference is essentially meaningless, and likely well with the error associated with the parameters.