Multivariate Linear Regression and the problem of multicollinearity

October 7, 2014

Bivariate linear regression fits usin sklearn/statsmodels return a very high standard error on the parameter estimates a_j, a_y and c where

$$L_{max} = a_j * t2_j + a_y * t2_y + c \tag{1}$$

moreover, the y-band slope is -0.0072. A *negative* slope implies a lower L_{max} for a higher $t_2(Y)$ which is clearly not what the data suggests. The high standard error also means that the combined fit yields very high errors on L_{max}. This points to the problem of multicollinearity (high correlation between predictors). We collect evidence to substantiate this and how to solve the problem

Evidence for multicollinearity

- Insigficant t-stat for the parameters, despite a hgih F-stat
- high r^2 between x_1 and x_2
- high standard error and opposite direction of the parameters
- VIF (variance inflation factor) $>>$ 10

where VIF is used as a 'rule of thumb' statistic to test the collinearity of predictor variables It is related to the pearsonr coefficient as

$$VIF = \frac{1}{1 - r^2} \tag{2}$$

In order to verify the parameter estimates, we looked at the output parameters from different least squares packages, which all yielded a negative slope for the $t_2(Y)$

Possible solutions

The above criteria for testing the presence of multicollinearity as seen in our dataset. We,therefore look at the possible solutions for this condition, without dropping a variable (that is another possibility which is recommended in such cases, but since the aim is to see how much better than a J-band only fit we can do, we keep this as a last resort).

1. Partial Least Squares: Using the linear model package in scikit learn, we then look at the parameter estimates from a partial least squares. The problem of directionality still remains

2. Ridge Regression: This method uses an l2 regularization with a linear least squares. It returns the parameter estimates. The errors are calculated by bootstrap sampling. For small positive values of α, the estimates are identical to the other methods. If we use a high α, we get lower errors on the slope and intercept, however, the resulting

We find that none of the given solutions are adequate to give the desired parameter estimates. A PCA is an alternative, but it would reduce the dataset to a lower dimensionality, an equivalent of dropping one of the correlated regressors.