Project 2: Linear Regression

36-600

Fall 2024

The goal of this project is to use linear regression to model the temperatures of stars given their brightness, their relative motion across the sky, and their location in the sky.

Your `HTML` file (generated by knitting an `R Markdown` source file) should be uploaded to Canvas by Friday, October 25th, at 11:59 PM. As was the case for Project 1, there is no limit on the length here, but conciseness is a virtue. (There will be no lab on Thursday the 24th; this will help provide time to complete the project.)

Data

You will examine the dataset `stellar_temperature.csv`, which is downloadable from the Project 2 Canvas module page.

The response variable is `teff`, the effective temperature of a star in degrees Kelvin. Your goal is inference, but what that really means is that you will at least check for multicollinearity, compute variance-inflation factors, do best-subset selection, and look at PCA and PC regression. You will want to output test-set mean-squared errors for at least a full linear model, and the best-subset-selection linear model. More "expectations" are listed below.

The predictor variables are the following:

name	description
`ra_x`	right ascension (celestial longitude)
`dec_x`	declination (celestial latitude)
`parallax`	stellar parallax (inversely proportional to distance)
`pmra`	proper motion along longitude line
`pmdec`	proper motion along latitude line
`[g,b,r]_mag`	magnitudes in green, blue, and red bands
`br_col`	the B-R color
`L`	galactic longitude
`B`	galactic latitude

Expectations

Your report should include the following elements:

- A description of the data (sample size, number of variables), and a list of removed variables, if there are any.
- Concise EDA, including the identification and removal of proposed outliers (if there are any) and, potentially, the transformation of the response variable (but see below first). (Note: always try transformations first before identifying and removing outliers…that lonely data point in a skew distribution may look fine after a transformation is performed.) You can, if you wish, transform any predictor variables that are highly skew, but this is not required. Also, create a correlation plot and comment on the possibility of multicollinearity in the predictor variables.

- A description of how you split the data into training and test sets.

- An analysis of the full dataset with linear regression, with comments on the output (does the fit appear to be good?). Provide the mean-squared error (and the root mean-squared error…interpret this number!) and the predicted response vs. observed response diagnostic plot (being sure to make the limits the same along each axis and being sure to overplot a diagonal line with slope 1).

- Comment on the value of adjusted R^2: is the linear model useful, or might other models perform better?

- Realize that the response variable needs to be transformed if and only if (a) the *residuals* of the fit are not normally distributed *and* (b) you wish to do inference using the hypothesis test output from `lm()`. To plot the residuals, find the difference between the observed test-set response values and the predicted test-set response values, and make a histogram: does this plot look normal? If you perform a hypothesis test, do you reject the null hypothesis that the residual data are normal?

- Does a plot of the residuals versus the predicted test-set response values suggest that the variance around the regression line is σ^2?

- Computation of variance-inflation factors. Comment if it appears that there is substantial multicollinearity, or not. Identify the potentially problematic variables (just list them). Do not remove variables with high vif factors, if there are any…just focus on variable selection via best-subset selection.

- A best-subset-selection analysis of the dataset. Which predictor variables are important for predicting stellar temperatures? Compute the MSE for the `BestModel` and compare it to the MSE for the full set of predictors. (And remember: you need to change the name of the response variable to `y` before running BSS!) Do you expect to see a large number of variables removed? A small number? Do the results match your expectation?

  ```
  w  <- which(names(df)=="teff")
  y  <- df[,w]
  df <- df[,-w]
  df <- data.frame(df,"y"=y)
  ```

- Do a PCA analysis on the predictors with the goal of determining the "true" dimensionality of the predictor variables. Select some number of PCs to keep (while being sure to indicate what criterion you used). Run linear regression on these PCs and compute the test-set MSE, and show how it compares with the full-set MSE and the BSS MSE. (The expectation, given the example in the PCA notes, is that the MSE will increase, perhaps substantially, if we do not include all the PCs in our PC regression model.) There is no need to comment on *how* the original predictors map to PC space…just pick a suitable number of PCs and proceed directly to regression.

- *NOTE*: as there are no factor variables in the dataset, you will not pursue random effects analyses in this project.