Problem 7: Celestial Neighbourhood

Assigned: 3 November
Due: 18 November
Maximum Mark: 25 Points
Maximum Submission Length: 8 pages

The file stars.csv contains measurements of the basic properties of the 101 closest catalogued stars to Earth. These include: distance from the Sun (units of parsecs), luminosity (how much visible light the star produces, units of Solar luminosity), and colour index (a measure of how blue or red the object appears; larger values mean a redder star).

1. Examine the distributions of each of the three quantitative variables above (distance, luminosity, colour index). Then, perform an appropriate mathematical transformation on the luminosity to create a new variable. Confirm (visually and/or with a test) that the distribution of this transformed variable has a more regular distribution than un-transformed luminosity.

2. Calculate the Spearman correlation coefficient, and check for significant correlations, for the following three variable pairs:
   (a) Distance and luminosity.
   (b) Distance and colour index.
   (c) Colour index and luminosity.

Provide a p-value for each pair. Does it matter whether you use the transformed or un-transformed luminosity? Explain why or why not.

3. Make a scatter plot of colour index versus transformed luminosity, colour-coded by "EvolType".

4. Next, consider only stars with an evoltype of "WD". Fit a simple linear model to the relationship between colour index (independent variable) and transformed luminosity (dependent variable) for these stars. Does the model indicate a statistically significant linear relationship between these variables for this group? (If not, does it rule one out?)

5. Next, consider only stars with an evoltype of "MS". Fit a simple linear model to the relationship between colour index (independent variable) and transformed luminosity (dependent variable) for these stars. Does the model indicate a statistically significant linear relationship between these variables for this group? (If not, does it rule one out?)

6. Write down two equations, describing the model fit to MS stars in Part 5:
   (a) the equation of the line above as fit/shown using the transformed variable.
   (b) the equivalent relationship between the *untransformed* variables.

7. Produce the following four plots, based on your linear model from Part 5:
   (a) a residuals vs. fitted plot
   (b) a normal quantile-quantile plot
   (c) a scale-location plot
   (d) a residuals vs. leverage plot.

Summarize in **one sentence per plot** what the implications of each are. Are the statistical assumptions of linear regression satisfied under this model?

8. Calculate, using your data and your model predictions, the following quantities:
   (a) SSY, the sum of squares of data values minus the mean.
   (b) SSM, the sum of squares of model predictions minus the mean
   (aka SSR in textbook)
   (c) SSE, the sum of squares of error terms.

9. Fit a more complicated parametric model of your choice to the MS stars (same two variables as before). The model can be linear or nonlinear, but should provide a meaningful improvement in the quality of the fit (assessed visually or in terms of SSE). Plot the data with the model curve overplotted, and calculate SSM and SSE for this model.

10. Fit a nonparametric model of your choice to the MS stars. Plot the data with the model curve overplotted, and calculate SSM and SSE for this model.