Leigh Murdick
NEAS Regression Analysis Project
Spring 2011

Stellar Distances

Introduction

Accurately measuring the distances to stars has always been a challenge for astronomers. However, I am going to attempt to use existing data to develop a formula that even an amateur astronomer with basic equipment could use to calculate the distance to certain stars. There are several complicating factors for the regression analysis, which I will discuss later. I am going to start with the well-known distance modulus formula:

$m - M = 5 \log(d/10)$

m = apparent magnitude of the star, defined as the brightness as viewed from earth
M = absolute magnitude of the star, defined as the brightness from a distance of 10 parsecs
d = distance to the star in parsecs

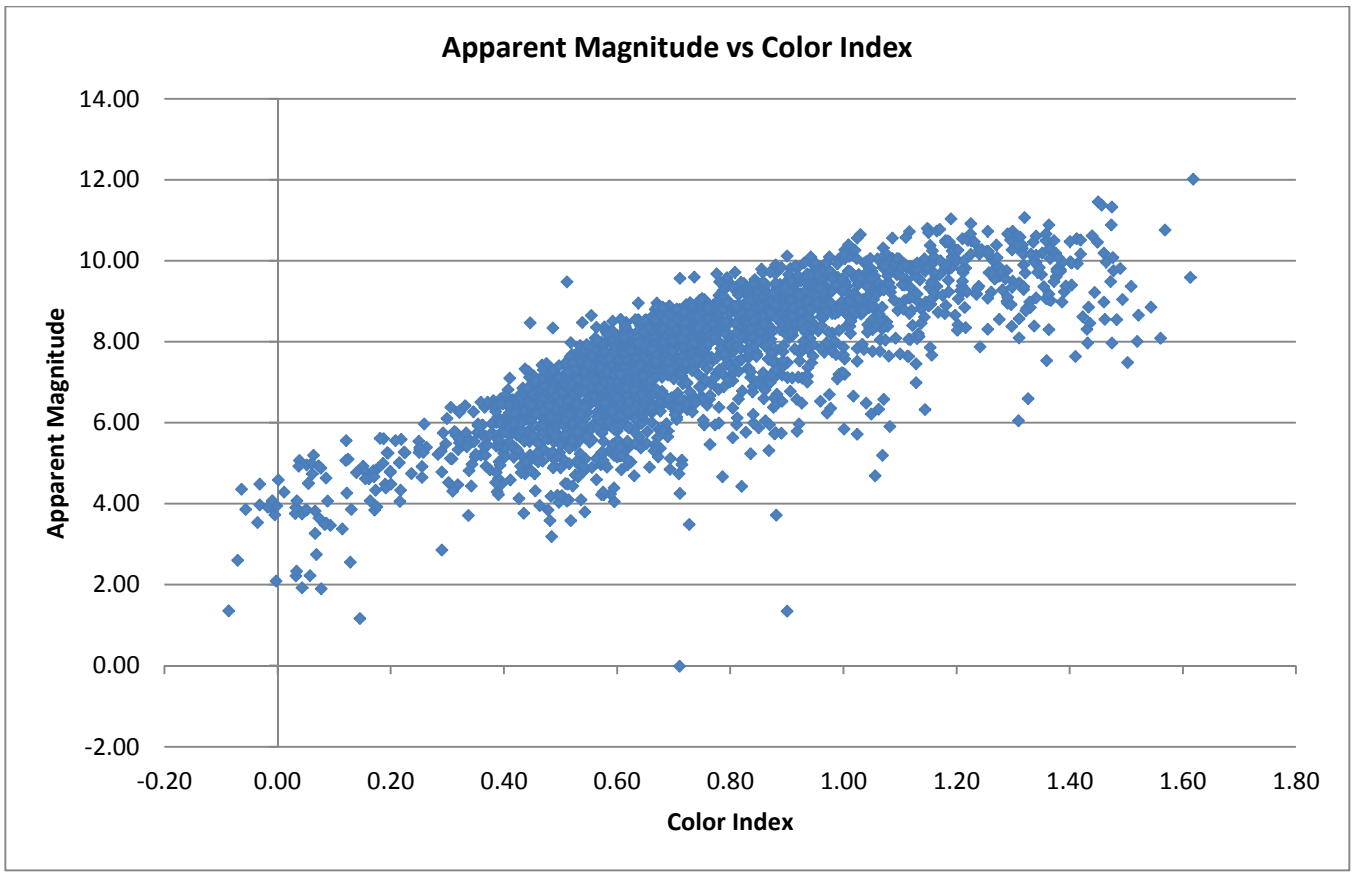
It is worth noting that magnitude itself is a logarithmic scale and that lower magnitudes correspond to higher brightness.

Although distance is the variable I ultimately wish to solve for, apparent magnitude must be the response variable in the regression as there is no physical explanation for a correlation between distance and absolute magnitude. Rewriting the formula, it is clear that there is a nice linear relationship using apparent magnitude as the response variable:

$m = M + 5 \log(d) - 5$

Unfortunately, absolute magnitude cannot be directly measured, so a substitution is needed. For "main sequence" stars – normal stars like the sun in the main part of their lifespan, there is a strong relationship between temperature and absolute magnitude. There exists a standard measure of temperature called the color index, which compares the magnitude of a star at two specific wavelength bands. The color index is also known as B-V and can be obtained simply by measuring the magnitude of the star through two different colored filters. Below, absolute magnitude is plotted against the color index for the stars in my sample. The relationship is not perfectly linear and includes other factors, but it is roughly linear over a large part of the range. This substitution of color index (CI) for absolute magnitude is what creates the need for a regression, which will have the form of:

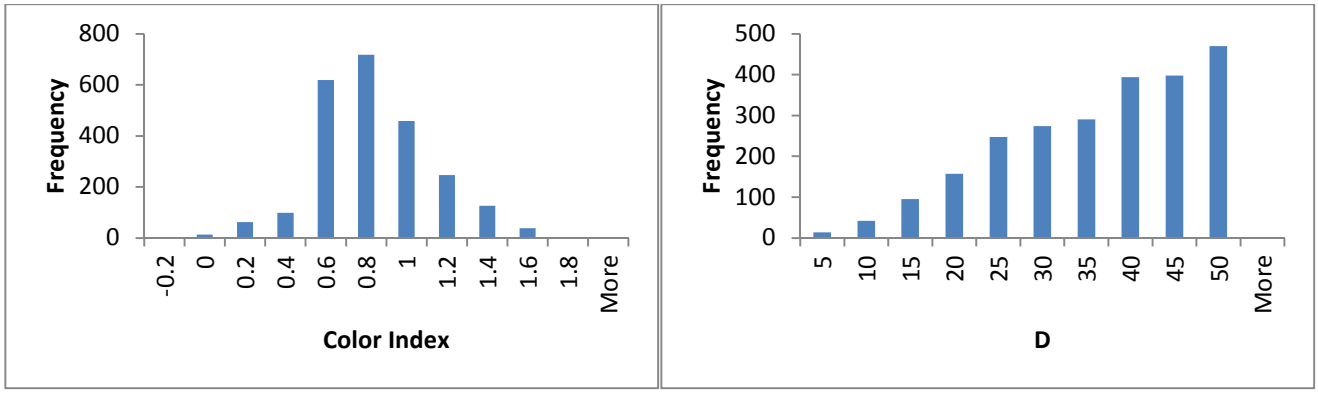
$m = \alpha + \beta_1 * \log(d) + \beta_2 * CI$

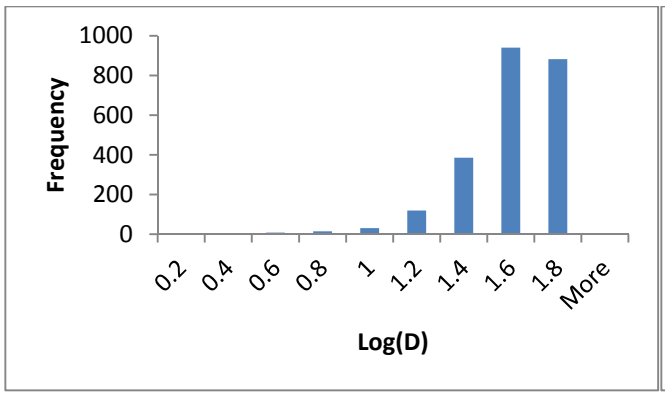


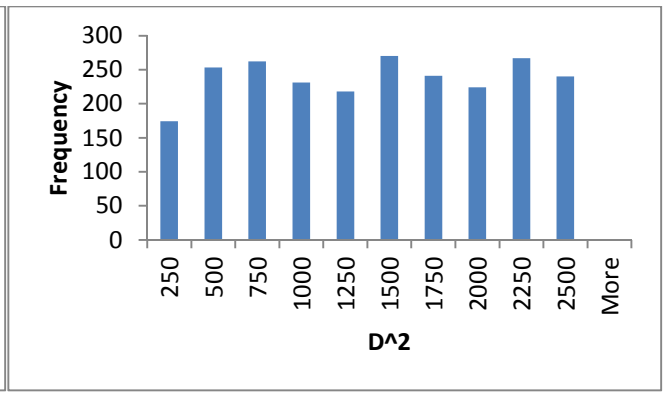
Data Analysis

I obtained the stellar data from a compiled database available at http://www.astronexus.com/node/34. I then edited the data to remove all non-main-sequence stars, all stars further than 50 parsecs, and all stars with missing or obviously erroneous data. The reason for the distance edit will be explained later. After the edits, 2381 stars remain in the sample.

Binning the explanatory variables and plotting histograms, it is apparent that the distribution of color index is somewhat normal while distance is negatively skewed. The log(d) transformation the formula suggests is even more skewed. It can be seen that d^2 has a roughly uniform distribution, but that transformation takes the relationship even further from linearity. Despite its less than ideal distribution, the log(d) transformation must be made to create a linear relationship.





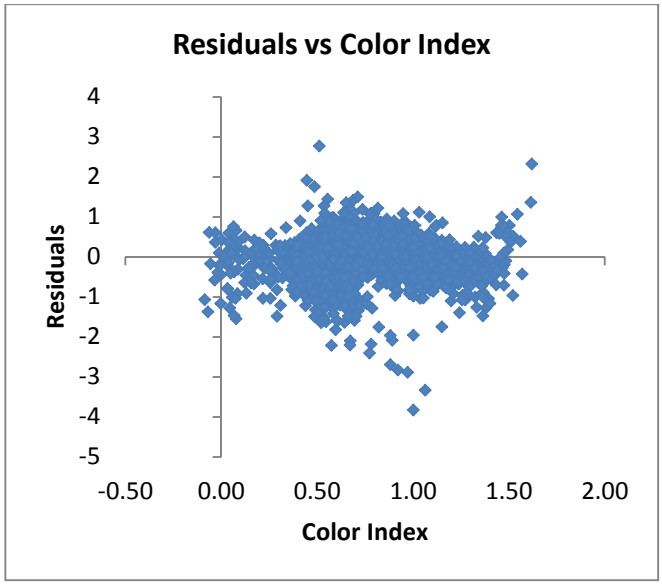


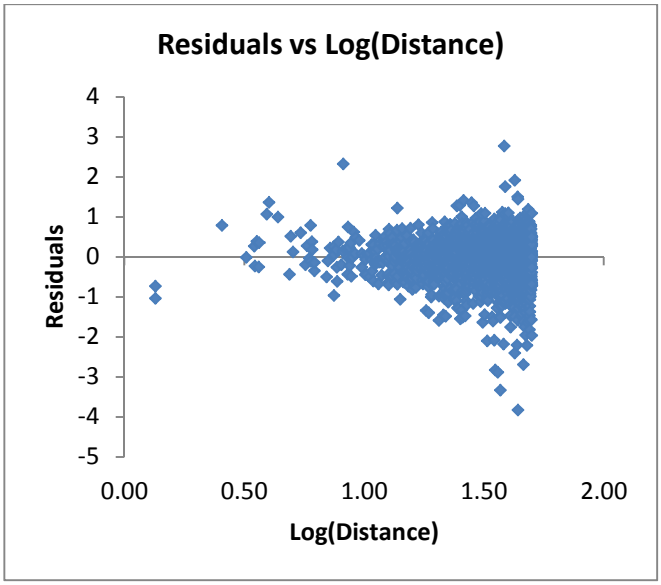
Another key assumption that must be violated is that all explanatory variables are measured without error. This is effectively true for color index, but distance measurements have considerable error, and error variance that increases with the distance. I will try to determine whether this fact significantly degrades the analysis or not.

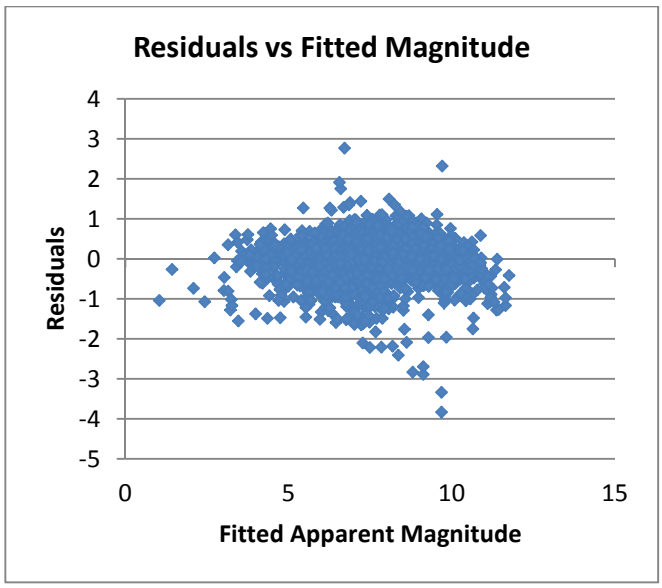
Plotting apparent magnitude vs color index and apparent magnitude vs log(d), the relationship to color index is clear to see, but surprisingly the relationship to distance is much harder to make out. A bivariate regression analysis confirms this with a shockingly low coefficient of determination: R^2 = 0.054. The heavily skewed distribution and measurement error could be coming into play here.

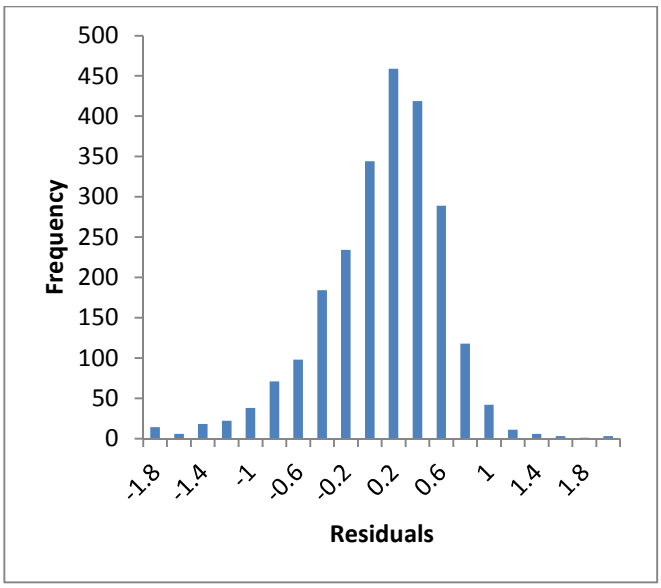
Regression Analysis

A multiple regression analysis was performed that produced the following residual plots: Residuals vs color index, residuals vs log(d), and residuals vs fitted apparent magnitude. As predicted, the error variance increases with the distance and the not-quite-linearity of color index can be seen. The residual distribution appears to be somewhat normal although the does appear to be some correlation to the fitted values.









The results of the regression can be seen below. Despite all the shortcomings, the fit turned out pretty well, with a multiple R^2 value of 0.89 and fairly tight 95% confidence intervals for each of the regression coefficients.

Regression Statistics	
Multiple R	0.9452
R Square	0.8935
Adjusted R Square	0.8934
Standard Error	0.5249
Observations	2381

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5495.0156	2747.5078	9972.6878	0
Residual	2378	655.1467	0.2755		
Total	2380	6150.1623			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.4990	0.1036	-33.7628	0.0000	-3.7023	-3.2958	-3.7023	-3.2958
Log(Distance)	4.6595	0.0596	78.1357	0.0000	4.5426	4.7765	4.5426	4.7765
Color Index	5.5313	0.0404	136.8550	0.0000	5.4521	5.6106	5.4521	5.6106

The formula resulting from the regression is:

$m = -3.50 + 4.66*\log(d) + 5.53*CI$

The measurement error and increasing error variance associated with distance did prove to be detrimental to the regression analysis. I tried alternate regressions extending the sample out to longer distances and the regressions got progressively worse. However, distance did prove to be essential to

the regression as theory suggests. A bivariate regression of apparent magnitude on color index yields and R^2 value of 0.62, significantly lower than that of the full regression.

Conclusion

The regression appears to be successful despite the highly skewed log(distance) distribution and the use of color index instead of the non-measurable absolute magnitude. Stellar distances are difficult to measure directly, even with precision equipment. This regression analysis provides a formula to very roughly estimate the distances to certain stars using very basic equipment, which I think is pretty neat. Solving for distance, the final formula:

d = 10^((m + 3.50 – 5.53*CI)/4.66)