Modeling the Relationship between Two Variables

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 6, 2016

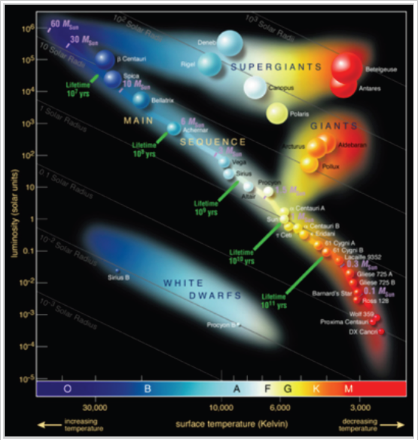
What did we think about last time?

- Relationships between Variables
- Visualizing Duplicated Values
- Piecewise Constant Joint Distributions: 2D Histogram (e.g.)
- 2-D KDE: Kernels, Bandwidths, Computational Issues
- High and low frequency areas; level sets, contours
- Visualizing matrices

Now thinking about the modeling/learning the relationship between variables

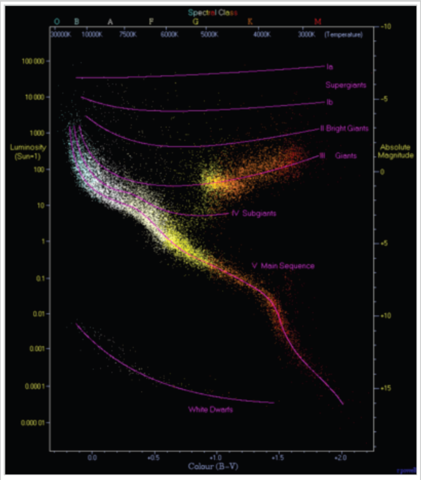
Visualizing Stars with Hertzsprung-Russell Diagram

Looking at Colors (Temperature) and Luminosities/Brightness



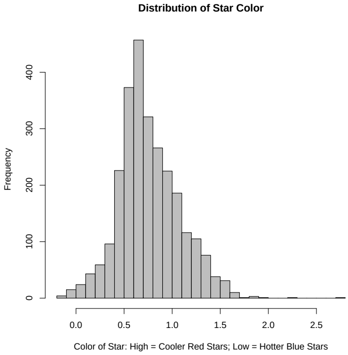
Hipparcos Stars

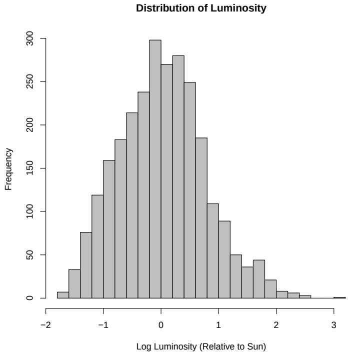
European Space Agency launched the Hipparcos satellite in the 1990s with higher measurement precision for about 100,000 stars



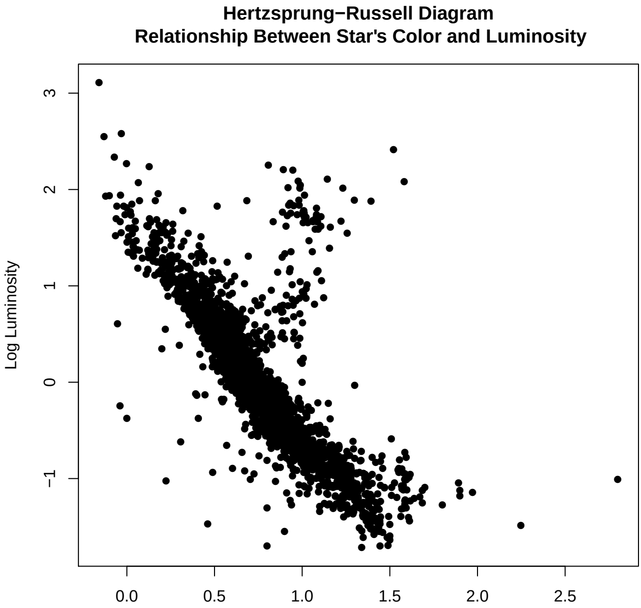
Star's Color and Luminosity

About 2700 Hipparcos stars mostly from the Hyades cluster



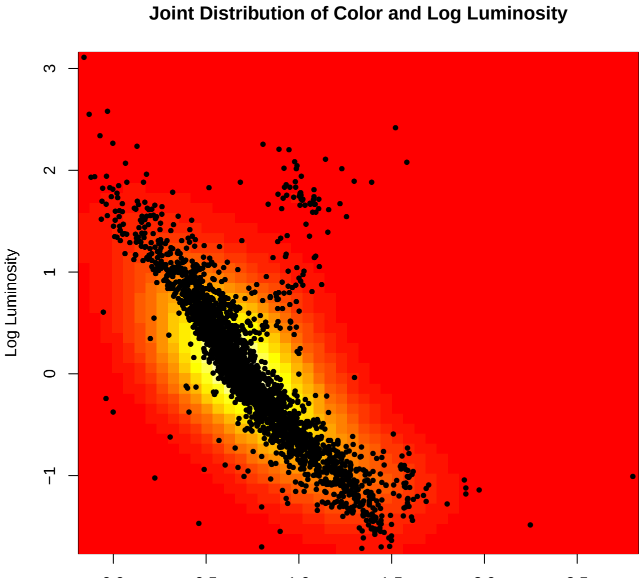


The Hipparcos Stars

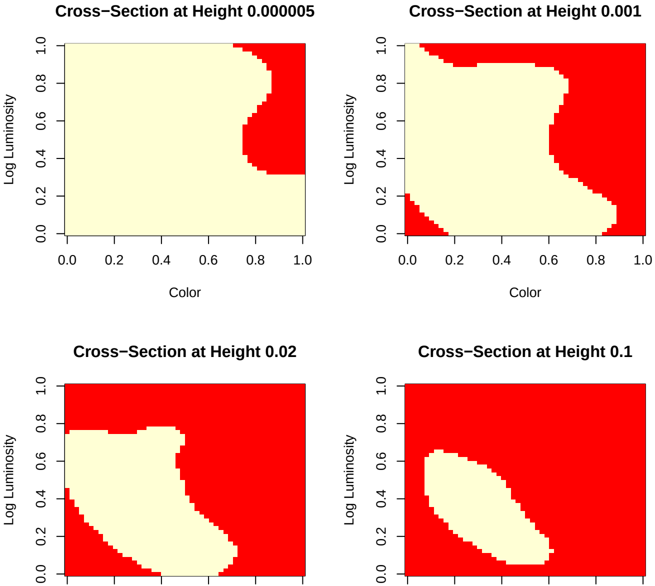


The Hipparcos Stars

2-D KDE (default kernel, bw = {1,1}, 50 bins each dim)



Cross-Sections/Level Sets at heights $\lambda = 0.000005, 0.001, 0.02, 0.1$



Linear Regression: A Least Squares Fit

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\text{where } E[\epsilon_i] = 0; Var[\epsilon_i] = \sigma^2, \epsilon_i, \epsilon_j \text{ uncorrelated}$$

- β_0: $E[Y_i]$ when $X_i = 0$
- β_1: change in $E[Y_i]$ associated with one unit increase in X_i

Can estimate using least squares:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Linear Regression: Normal Errors

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2), \ \epsilon_i \text{ independent}$$

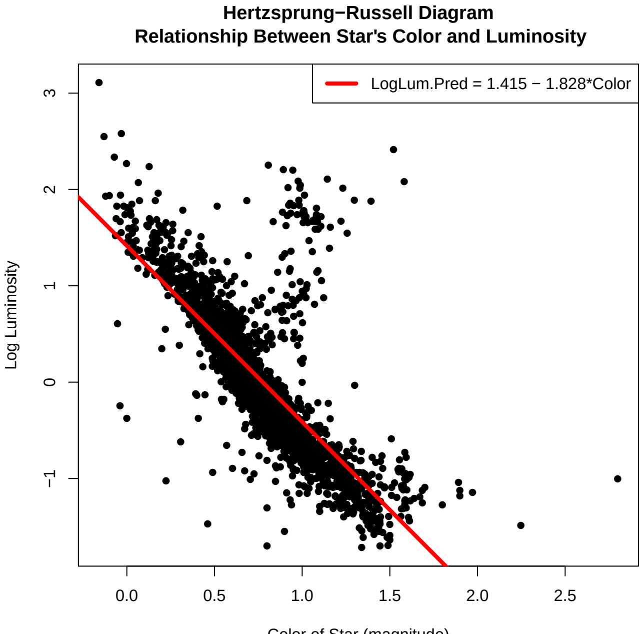
Assumptions:

- Linear relationship between Y and X
- Errors are normally distributed
- Errors have expectation zero, constant variance
- Errors are independent

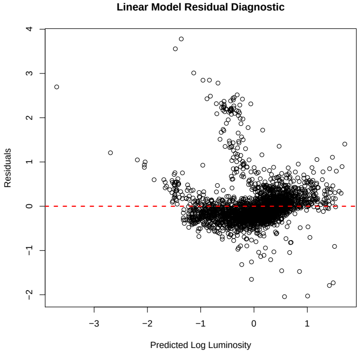
Can estimate with Maximum Likelihood Estimation

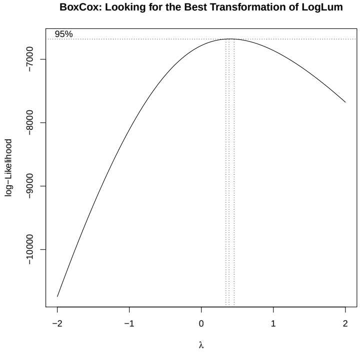
$$L(Y|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$$

Fitting the Line



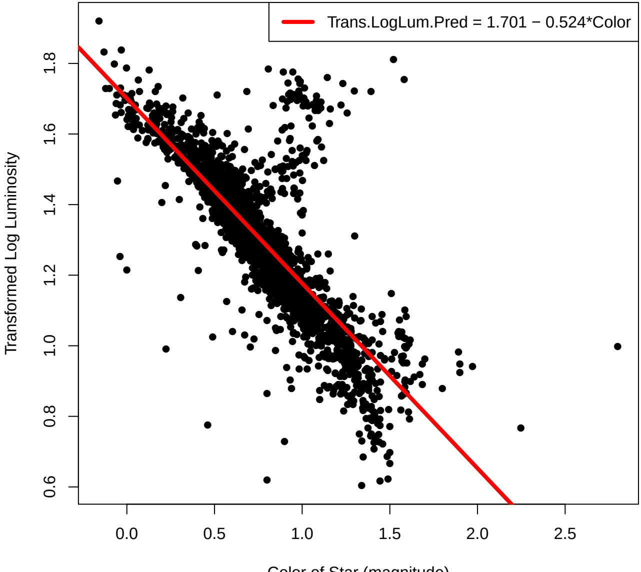
Looking at Diagnostics



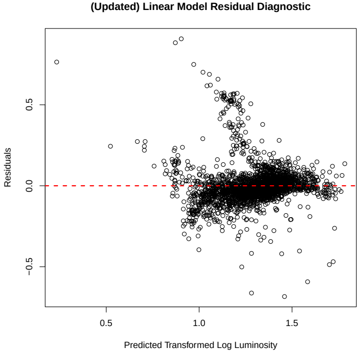


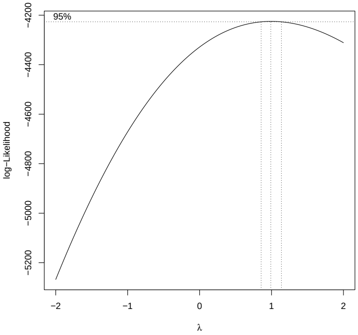
Box-Cox suggests λ in a range around [0.33, 0.50]

After Transforming Log Luminosity



Double-checking Diagnostics





Using a Smoother (For Next Time)

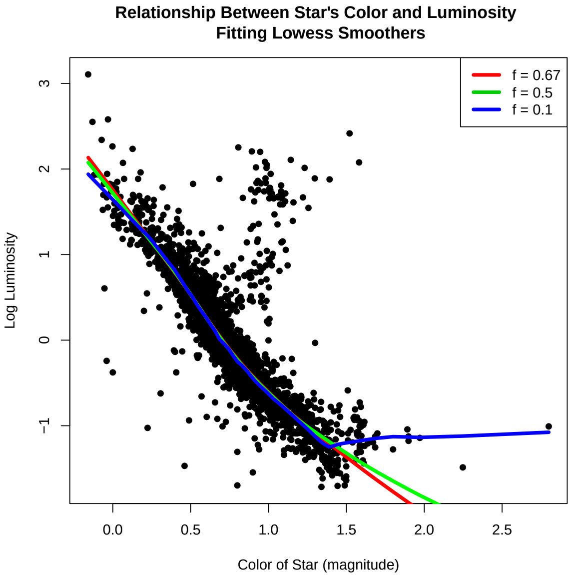
One common tool is the Lowess Smoother

- Locally-weighted polynomial regression

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \ldots\ldots$$

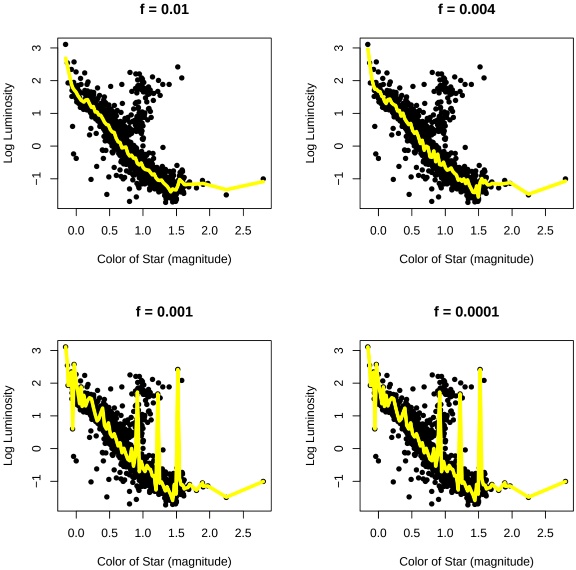
  Can choose degree or use default
- Weights are related to closeness of points to the estimation location (close points, heavy weight)
- "Sliding window" across the data
- Parameter = size of window: wide, global; small, local

Smoothing with Different Windows



How could we head toward the white dwarfs and/or the gas giants?

Smoothing with Different Windows



In summary: What did we think about?