Modeling the Relationship between Two Variables, Part 2

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

PCMI Undergraduate Summer School 2016

July 7, 2016

What did we think about last time?

- Relationships between Variables
- Level Set/Cross-Section Images
- Linear Relationships
- Competing Sets of Assumptions
- Potential Transformations (Box-Cox)
- Diagnostic Visuals
- Whether or not to continue to date someone if they throw away data
- Why you shouldn't play Adult Kickball

Now continuing to think about learning the relationship between two variables

Reminder of our Linear Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

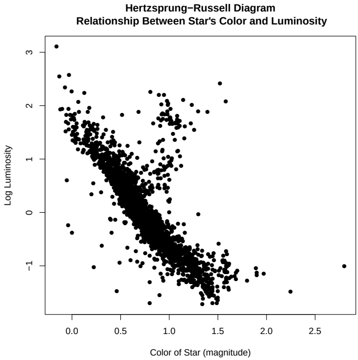
- β_0: $E[Y_i]$ when $X_i = 0$ (might be out of scope)
- β_1: change in $E[Y_i]$ associated with one unit increase in X_i
- Two sets of possible assumptions
  - Linear relationship between Y and X
    $E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$, ϵ_i, ϵ_j uncorrelated
    Fit using least squares criterion

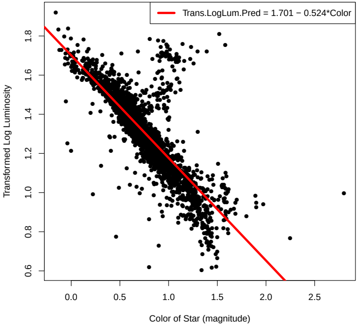
$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  - Linear relationship between Y and X
    $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i, ϵ_j independent
    Fit using Maximum Likelihood Estimation

Reminder of our Hipparcos Stars

Original Data vs Modeled, Transformed Data





FIRST CHECK THAT YOUR ASSUMPTIONS ARE MET

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.68397 -0.04943 -0.01225  0.02818  0.90742

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.701053   0.006029  282.15   <2e-16 ***
B.V         -0.524138   0.007305  -71.75   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1203 on 2676 degrees of freedom
Multiple R-squared:  0.658, Adjusted R-squared:  0.6579
F-statistic:  5148 on 1 and 2676 DF,  p-value: < 2.2e-16
```

Using a Smoother

One option: LOWESS (Locally Weighted Scatterplot Smoother)

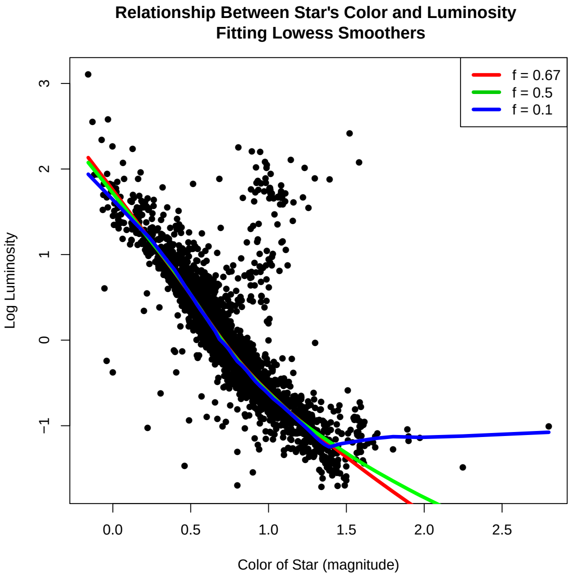
- Nonparametric; doesn't estimate parameters, focuses on fit
- Locally-weighted polynomial regression

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

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

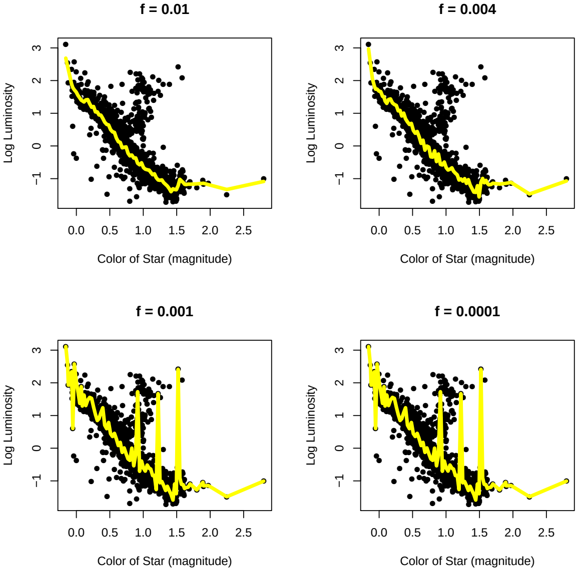
  $$\propto (1 - |\frac{x - x_i}{max\ dist\ in\ window}|^3)^3$$

Smoothing with Different Windows



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

Smoothing with Different Windows



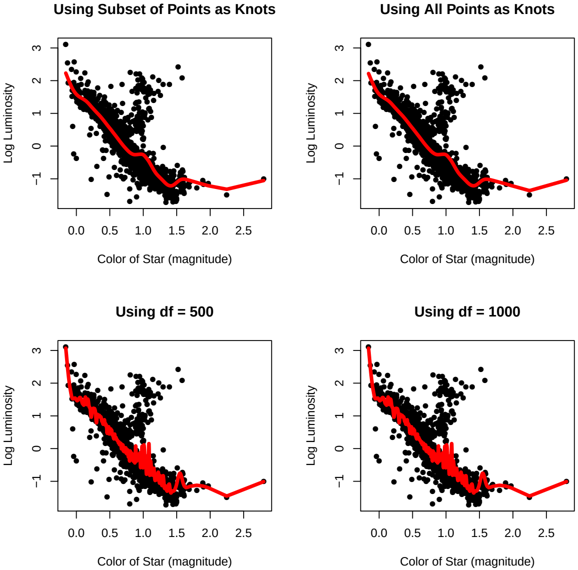
Another Smoothing Option: Splines

Trying to minimize:

$$\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2 + \lambda \int_{range_x} \hat{f}''(X)^2 dx$$

- kth order spline is piecewise polynomial function of degree k
- has derivatives up to order $k-1$ at its knot points
- knot points?
  locations spread throughout the space
  could be all the data points
  could be a subset
  could be an evenly spaced grid
- Cubics are probably the most popular;
  difficult to even see knot locations

Experimenting with Different Splines



In summary: What did we think about?

- Significance of our Linear Model
- Nonparametric smoothers
- Window parameter: global vs local
- Smoothness "penalties"
- Knot points: reduce computation; want derivative smoothness