Multivariate Regression (Trees)

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

PCMI Undergraduate Summer School 2016

July 8, 2016

What did we think about last time?

- Linear Model Assumptions
- Hypothesis Tests for Variable Relationships
- Nonparametric LOWESS Smoothers
- Cubic polynomial splines

Now we'll try

- adding more variables into the mix
- regression with partitioning

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
- Assumptions
  - Linear relationship between Y and X
  - $\epsilon_i \sim N(0, \sigma^2)$,
  - ϵ_i, ϵ_j independent
- Use t-tests to determine significance of the estimated relationship between Y_i and X_i

  $H_0 : \beta_1 = 0$ (no relationship); $H_a : \beta_1 \neq 0$ (relationship)
  use p-value to reject/fail to reject the null hypothesis

Extending to Multivariate Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

Same assumptions:

- Linear relationship $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}$
- $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i, ϵ_j independent

Interpretation changes slightly:

- β_0 : $E[Y_i]$ when all X_i are zero
- β_j: change in $E[Y_i]$ associated with a one unit increase in X_j holding all other variables fixed

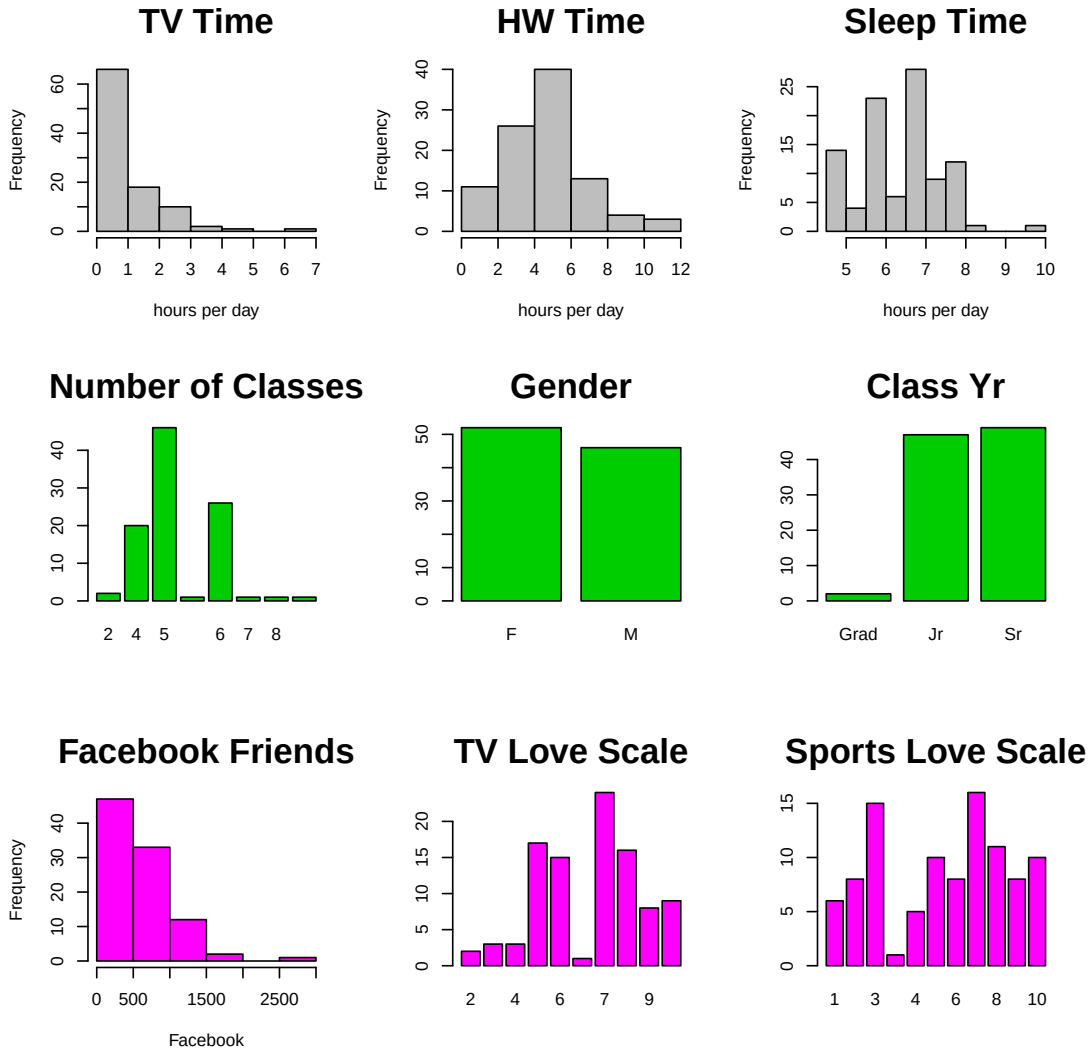
- Significance tests for $H_0 : \beta_j = 0$, $H_a : \beta_j \neq 0$ now use the t_{p-1} distribution
- Can still use residual diagnostics and Box-Cox plots

Some CMU Undergraduate Data

Surveyed about 100 students in Regression course;
interested in how much time spent watching TV/movies

- time spent watching TV/movies
- time spent doing HW
- how many classes
- gender
- major
- class year
- Facebook friends
- Netflix account
- how much do you love TV/movies (scale 1-10)
- average sleep per night
- device used to watch TV/movies
- how much you like sports (scale 1-10)
- favorite TV show

EDA



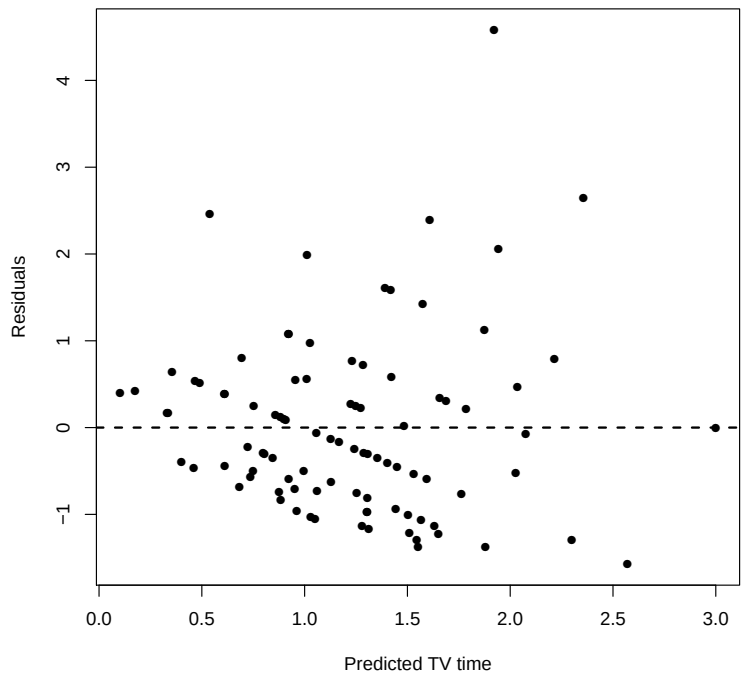
44% had a Netflix Account;
Computers (36.7%) and Laptops (44.9%) most common devices

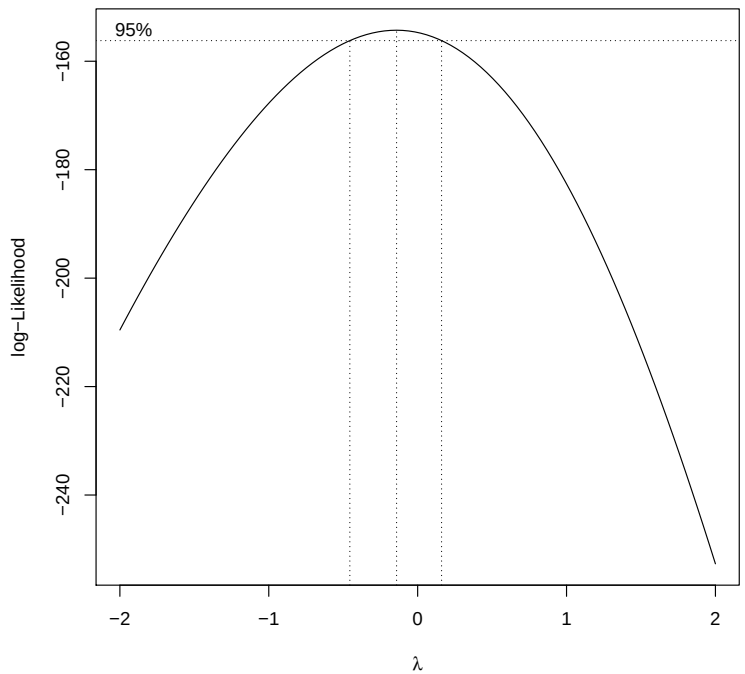
Example Regression Model

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7758104  1.2584975   1.411  0.16215
HWTime       -0.0478706  0.0530619  -0.902  0.36971
Classes      -0.2276657  0.1299174  -1.752  0.08359 .
GenderM      -0.7712372  0.2809823  -2.745  0.00749 **
grad         -1.5655970  0.9918993  -1.578  0.11847
senior       -0.3320023  0.2465508  -1.347  0.18196
Facebook     -0.0003707  0.0002604  -1.424  0.15849
NetflixYes    0.1012115  0.2548795   0.397  0.69237
TVLove        0.0486660  0.0656859   0.741  0.46096
AvgSleep      0.0306497  0.1257534   0.244  0.80807
DeviceLaptop  0.3163228  0.2713850   1.166  0.24729
DevicePhone   1.1435347  1.2031553   0.950  0.34478
DeviceTablet  0.4335289  0.5282366   0.821  0.41428
DeviceTV      0.2712325  0.4518562   0.600  0.55005
SportsLove    0.1468576  0.0495706   2.963  0.00403 **

Multiple R-squared:  0.2161, Adjusted R-squared:  0.07714
F-statistic: 1.555 on 14 and 79 DF,  p-value: 0.1112
```

Diagnostics





Updated Model

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5284629  0.6571497   0.804  0.42371
HWTime       -0.0294770  0.0277074  -1.064  0.29063
Classes      -0.0978578  0.0678390  -1.443  0.15311
GenderM      -0.3756239  0.1467206  -2.560  0.01237 *
grad         -0.5911255  0.5179401  -1.141  0.25719
senior       -0.1053054  0.1287414  -0.818  0.41584
Facebook     -0.0002036  0.0001360  -1.497  0.13827
NetflixYes    0.1184639  0.1330904   0.890  0.37611
TVLove        0.0362302  0.0342992   1.056  0.29405
AvgSleep     -0.0020085  0.0656646  -0.031  0.97568
DeviceLaptop  0.1248748  0.1417091   0.881  0.38088
DevicePhone   0.4352216  0.6282517   0.693  0.49050
DeviceTablet  0.3169247  0.2758293   1.149  0.25403
DeviceTV      0.1274085  0.2359458   0.540  0.59072
SportsLove    0.0800158  0.0258843   3.091  0.00275 **

Multiple R-squared:  0.2157, Adjusted R-squared:  0.07671
F-statistic: 1.552 on 14 and 79 DF,  p-value: 0.1123
```

Regression Trees

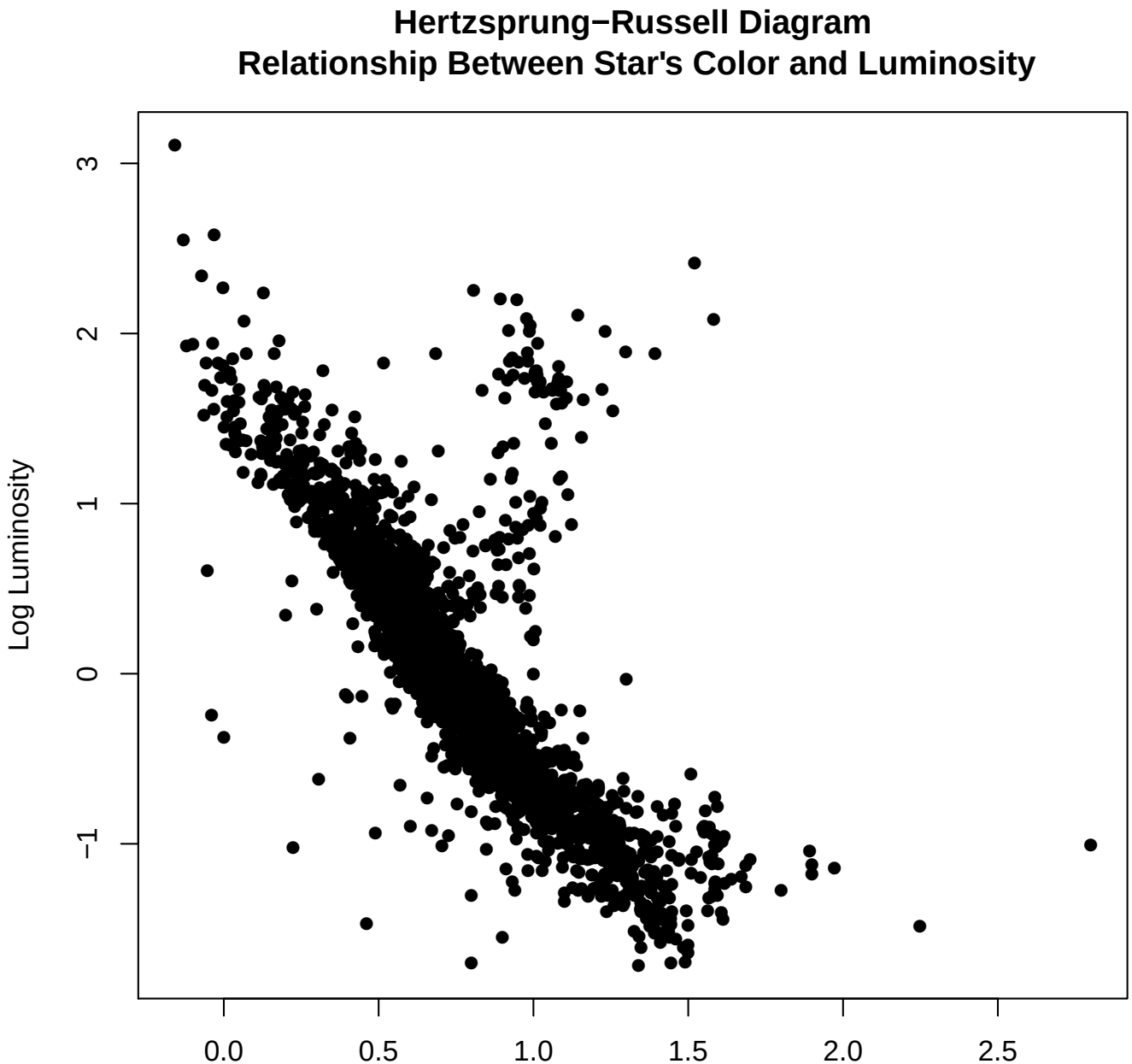
Instead of fitting a $p-1$-dim response surface and then predict a different value for each observations, we might partition the observations into similar subgroups and assign them a group value
Want to answer the questions:

- What variables are useful for prediction and subgroup separation?
- What values are useful "cutoffs"?

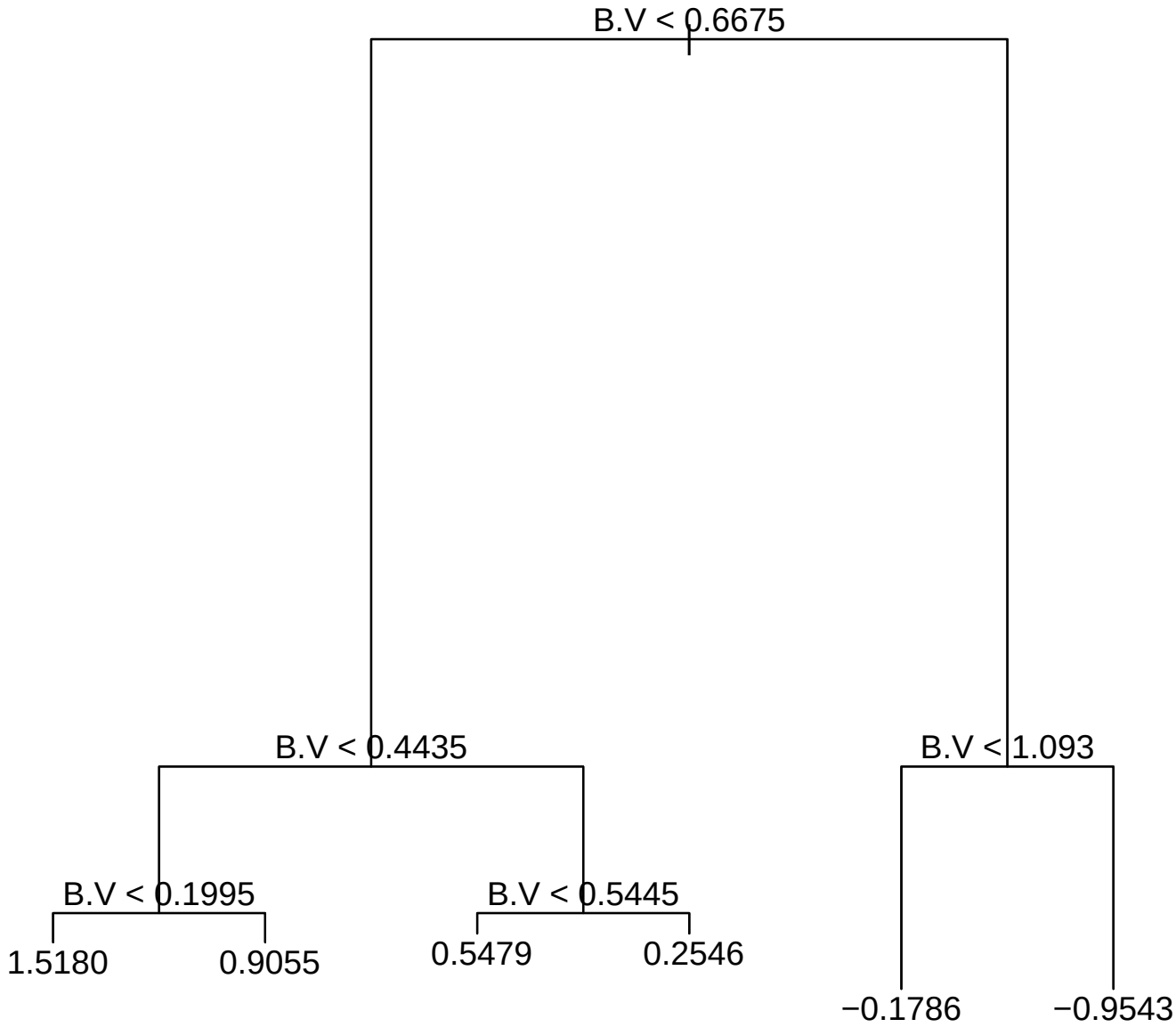
Regression Trees: recursively partition the feature space using hyper-rectangles to find groups of similar observations

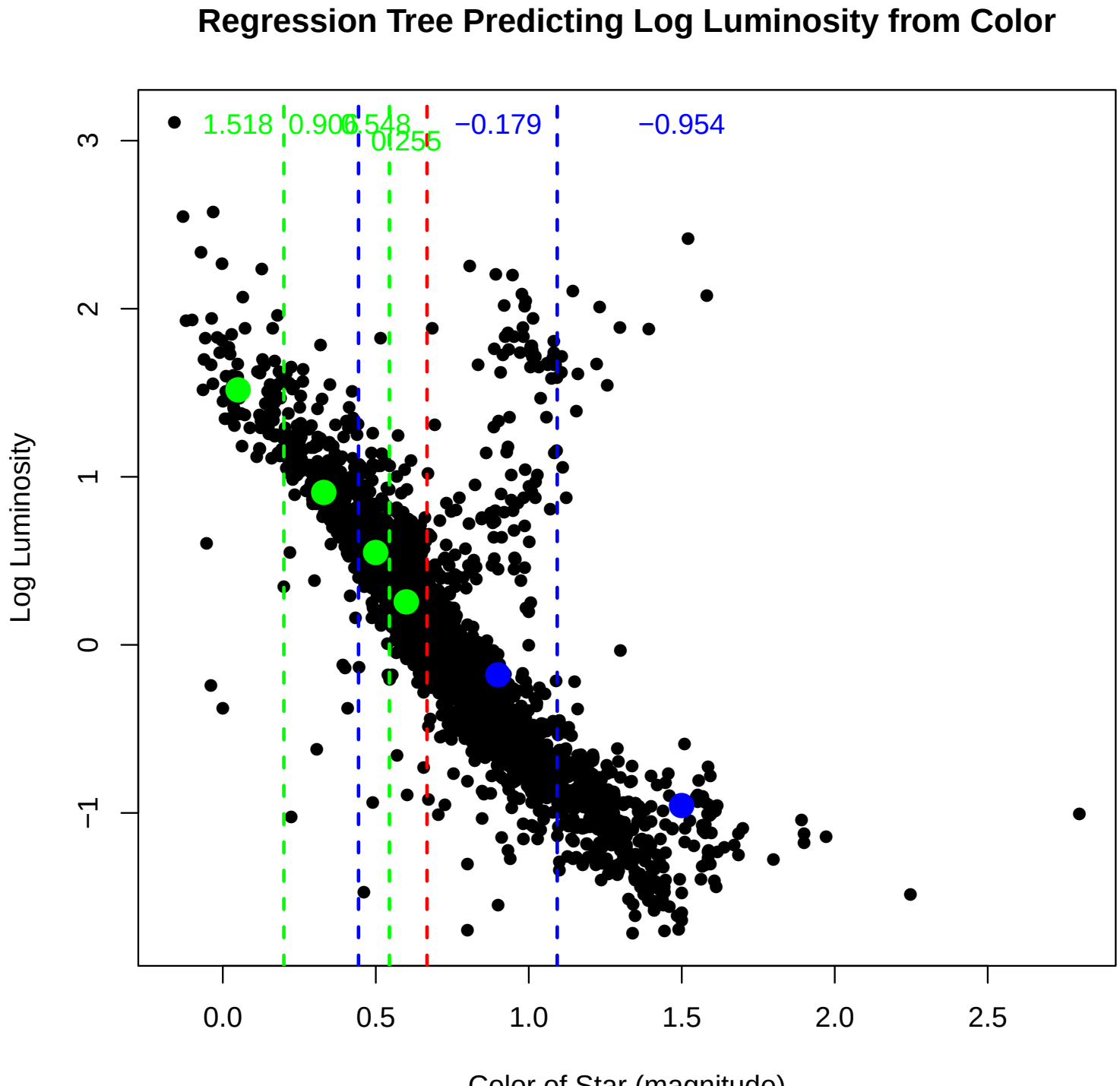
- Search for "best separation" in Y's
- "Closeness" measured in deviance: $D = \sum(Y_i - \mu_i)^2$
- Stored in hierarchical binary tree structure
- Stop splitting when size threshold or splitting criteria met
- Groups assigned their average Y value

Reminder of our Hipparcos Stars

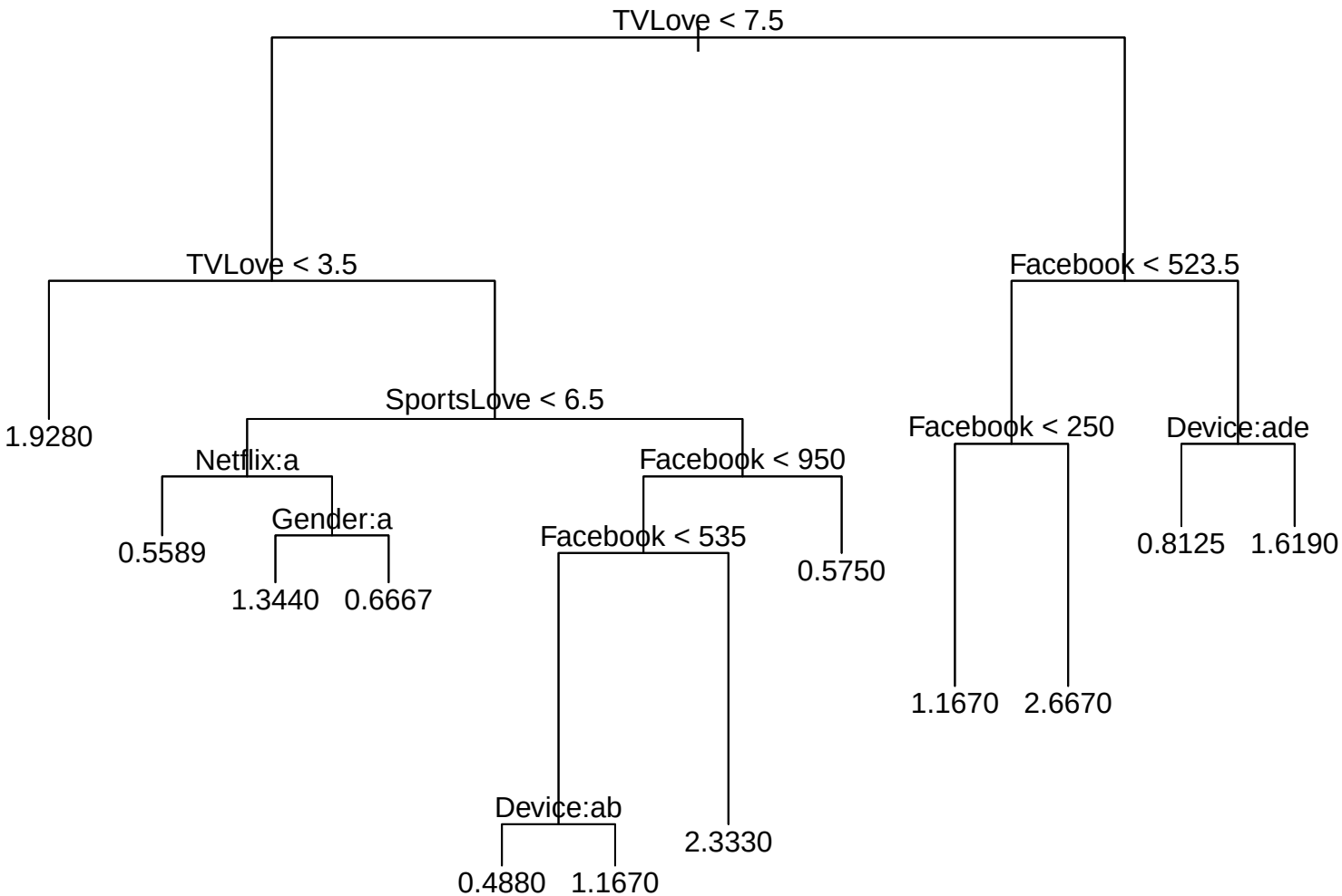


Regression Tree Predicting Log Luminosity




Regression Tree Predicting Log Luminosity from Color
1.518
0.906
0.548
0.255
−0.179
−0.954
Log Luminosity
3
2
1
0
−1
0.0
0.5
1.0
1.5
2.0
2.5
Color of Star (magnitude)

Back to our CMU Undergrads



Downside of Regression Tree

Often criticized for ability to overfit; can be controlled through parameter selection and paying attention

Another Option: *Random Forests*

- Ensemble/Collection of Regression Trees
- Each tree: random subset of variables and observations
- Final predictions are aggregated across the set of trees
- More stable; nice, theoretical properties

Check out "The Elements of Statistical Learning" by Hastie, Tibshirani, Friedman; it's all kinds of free on the web

Another good one is "An Introduction to Statistical Learning" by James, Witten, Hastie, Tibshirani. Not so much free.