# STAT511HW7

*Ben Straub*

*November 18, 2015*

Severe wildfires in southern California in 2009 present a rare opportunity to study which factors influence whether a house near a forest is burned in a wildfire. I will examine the data set, Fire, paying particular attention to six predictor variables that are under the control of the homeowner: planted, buildings, perc.woody, perc.cleared, distance2bush and distance2tree. I will attemtp to make recommendations to homeowners, based on my analysis of the Fire data, to mitigate their chance of their home being burned down.

# (1) EXPLORATORY AND VARIABLE SELECTION

## EXPLORATORY DATA ANALYSIS

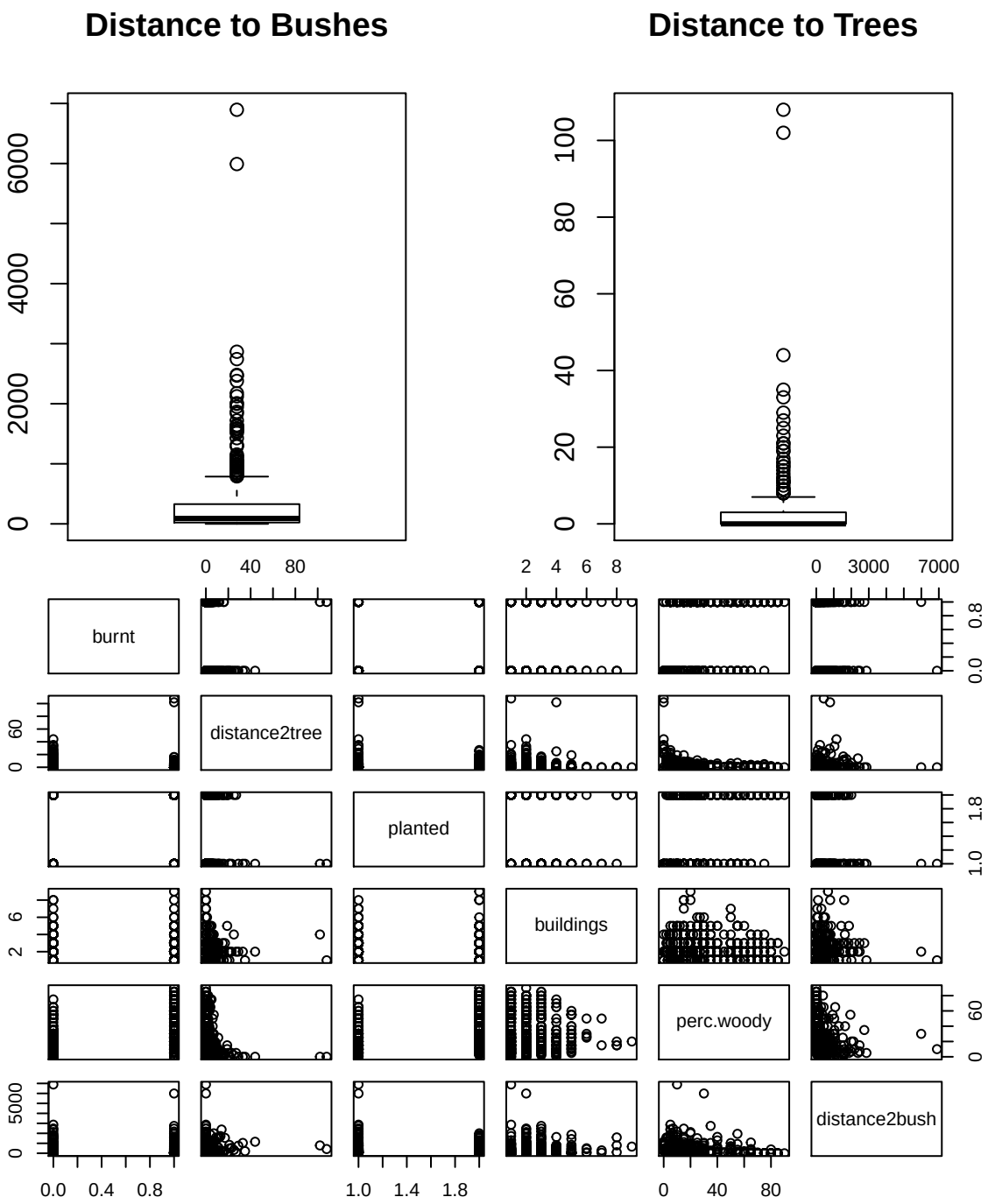
*GENERAL OBSERVATIONS OF FIRE DATA*

- The Total Data Set has 487 observations with 21 variables
- There are 274 homes that burned from this Survey
- The Variables planted, adj.for.type are factor variables with two levels and 3 levels

*OBSERVATIONS OF VARIABLES TO BE ADDRESSED IN STUDY*

- The homeowners have control over the following variables for this study: planted, buildings, perc.woody, perc.cleared, distance2bush and distance2tree.

- These six variables will receive the most attention in the following analysis.
- I looked at boxplots for each of the six variables, but found only Distance to Bushes and Distance to Trees to have any interesting features.
- The boxplots show that the houses have a lot of bushes and trees around them, i.e. the distance is close to zero, but there are a significant amount of houses that have outliers associated with them.
- I believe there is some sort of quadratic relationship between some of the variables in the data sets. I am making this ovservations form subsetting the data and looking at its pairs plots.




# MODEL SELECTION USING AIC

Table 1: Full Data Set

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.0707155 | 0.6195446 | -4.9564077 | 0.0000007 |
| ffdi | 0.0069344 | 0.0029740 | 2.3316683 | 0.0197181 |
| topo | -0.1186014 | 0.0807015 | -1.4696310 | 0.1416617 |

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| perc.cleared | 0.0081901 | 0.0054250 | 1.5096833 | 0.1311243 |
| perc.burntless5yrs | 0.0350144 | 0.0192007 | 1.8236024 | 0.0682122 |
| amt.not.NP | 0.0000337 | 0.0000181 | 1.8612503 | 0.0627088 |
| adj.for.typeOPEN FOREST | -0.1825636 | 0.4076048 | -0.4478936 | 0.6542300 |
| adj.for.typeWOODLAND | 0.4763813 | 0.4462029 | 1.0676337 | 0.2856858 |
| distance2tree | 0.0221118 | 0.0130313 | 1.6968279 | 0.0897292 |
| plantedr | 1.8977666 | 0.2444425 | 7.7636516 | 0.0000000 |
| perc.woody | 0.0604026 | 0.0077366 | 7.8073677 | 0.0000000 |

Table 2: Subsetted Data

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.2490079 | 0.2452104 | -9.171746 | 0 |
| plantedr | 1.8649140 | 0.2318778 | 8.042657 | 0 |
| perc.woody | 0.0539866 | 0.0070599 | 7.646980 | 0 |

_OBSERVATIONS ON AIC MODEL SELECTION

- The first table shows the AIC technique on the entire data set and it produces the following model.
- Model Equation:

$$burnt = ffdi + slope + topo + distance2tree + planted + perc.woody$$

- The only coefficients in this model that are significant are plantedr and perc.woody
- The second table shows the results from running the AIC technique on subsetted data that only includes the variables that homeowners control.
- Model Equation:

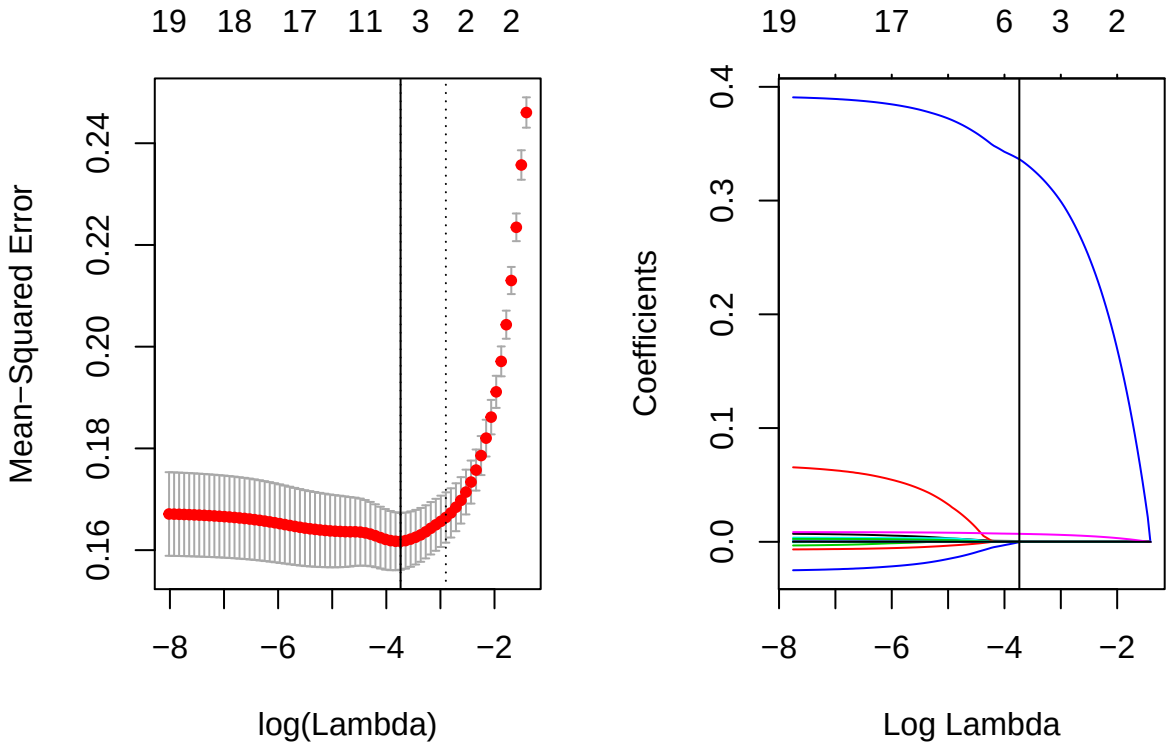
$$burnt = planted + perc.woody$$

- The 2nd table shows that plantedr and perc.woody as being significant, i.e. the likelihood of your house being burnt increases by the log-odds of plantedr and perc.woody. Still unclear how to best interpret my cofficients with GLM.

*CONCLUSION:* Using the AIC techinque over the entire data set or the subsetted data set produces models with the same coefficients being significant, i.e. plantedr and perc.wood. I decided that subsetting the data was a bad move and will not doing it for the following model selection techniques: LASSO, Ridge and p-value.

# MODEL SELECTION USING LASSO

```
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-2
```




*OBSERVATIONS:*

- Using the LASSSO technique I found the following two variables to be of significance: planted and perc.woody. The two graphs show how the tuning parameter lamba zeroes out everything excpet for planted and perc.woody.
- Model Equation:

$$burnt = planted + perc.woody$$

# MODEL SELECTION USING P-VALUE BASED VARIABLE SELECTION

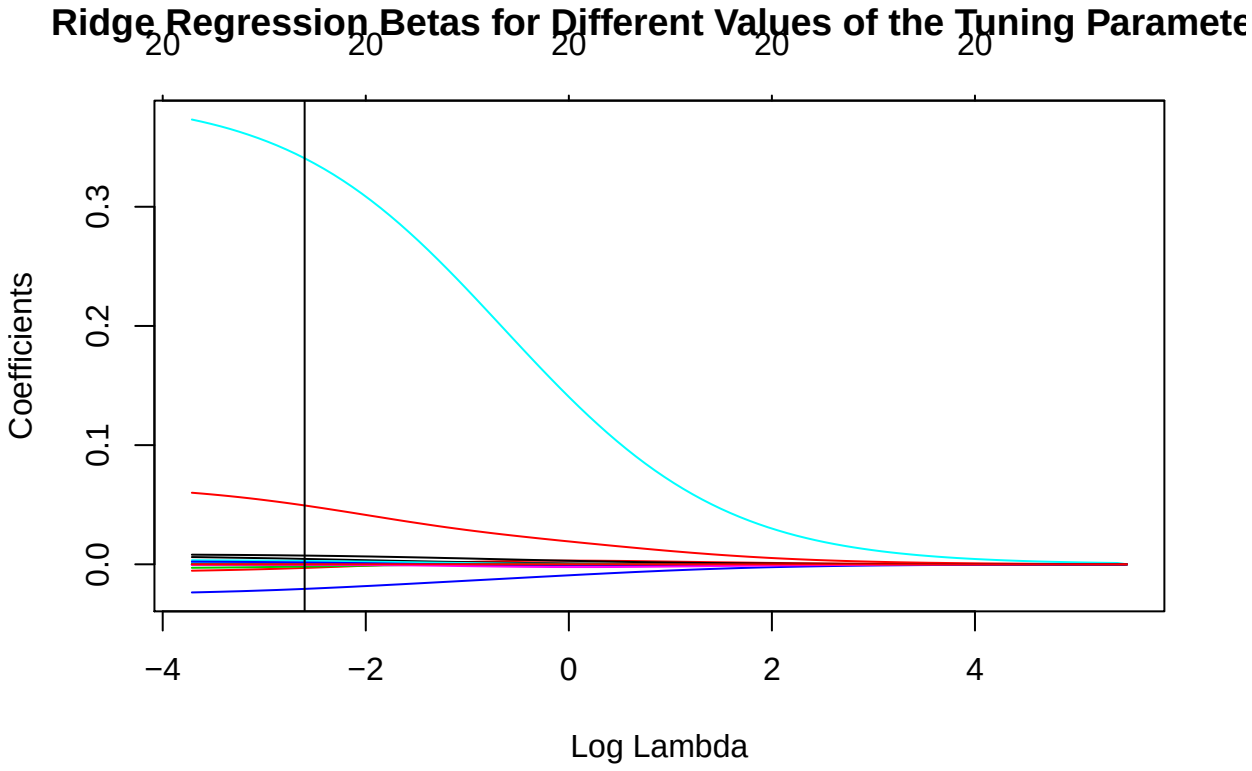
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.7235281 | 0.8067310 | -3.3760052 | 0.0007355 |
| ffdi | 0.0077290 | 0.0031355 | 2.4650002 | 0.0137013 |
| slope | -0.0181923 | 0.0331793 | -0.5483018 | 0.5834847 |
| aspect | -0.0002497 | 0.0016381 | -0.1524095 | 0.8788640 |
| topo | -0.1629255 | 0.0867009 | -1.8791660 | 0.0602218 |
| perc.cleared | 0.0073718 | 0.0060430 | 1.2198868 | 0.2225078 |
| amt.burntless5yrs | 0.0000220 | 0.0000295 | 0.7474128 | 0.4548144 |
| perc.burntless5yrs | 0.0423145 | 0.0285653 | 1.4813246 | 0.1385201 |
| amt.not.burnt5to10yrs | -0.0000258 | 0.0000276 | -0.9340735 | 0.3502660 |
| perc.burnt5to10yrs | -0.0382417 | 0.0512884 | -0.7456193 | 0.4558974 |
| amt.unlogged | -0.0000170 | 0.0000292 | -0.5815504 | 0.5608696 |
| perc.logged | 0.0232785 | 0.0373098 | 0.6239263 | 0.5326760 |
| amt.not.NP | 0.0000287 | 0.0000268 | 1.0678320 | 0.2855963 |
| amt.not.SF | -0.0000071 | 0.0000321 | -0.2219233 | 0.8243736 |
| adj.for.typeOPEN FOREST | -0.0731776 | 0.4530346 | -0.1615276 | 0.8716778 |

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| adj.for.typeWOODLAND | 0.5733010 | 0.4952105 | 1.1576916 | 0.2469899 |
| edge | 0.0000126 | 0.0000299 | 0.4222116 | 0.6728706 |
| distance2tree | 0.0208118 | 0.0441411 | 0.4714839 | 0.6372952 |
| plantedr | 2.2105287 | 0.2730149 | 8.0967331 | 0.0000000 |
| buildings | -0.0362674 | 0.0924090 | -0.3924660 | 0.6947140 |
| perc.woody | 0.0545751 | 0.0080568 | 6.7737525 | 0.0000000 |
| distance2bush | 0.0002001 | 0.0002049 | 0.9766909 | 0.3287222 |
| poly(distance2tree, 2)2 | 12.5454263 | 5.8810531 | 2.1331938 | 0.0329088 |

*OBSERVATIONS:*

- In my exploratory data analysis I thought I noticed some of the data having a quadratic pattern, particularly the distance2tree variable. I decided to try a GLM with a distance2tree as a polynomial and it turned out to be significant.

*CONCLUSION:* The glm model has plantedr, perc.woody and distance2tree as being significant. I think trees are very important! If the tree is tall and it is burning, then its integrity could be at issue and fall and destroy/burn the hosue.




- *OBSERVATION:* I'm unsure how to pull out the ridge regression coefficients, but I was able to calculate the mspe and cvmspe to make comparison of all the models.

# (2) MODEL PREDICTION

| | CVMSPE-Train | MSPE-Test |
|---|---|---|
| AIC | 0.1658912 | 3.3076013 |
| RR | 0.1693127 | 0.1534715 |
| Lasso | 0.1617225 | 0.1583821 |
| GLM | 0.1654877 | 0.1532960 |

- I compared four techniques: Lasso, AIC, Ridge and p-value selection, to arrive at proper parameters. Running through the Test and Train sets I found that the Lasso Model had the best predicting power of the fours models. I will use my Lasso model with perc.woody and planted as my predictor variables to investigate the three proposals to deal with mitigating houses lost to fires.

# (3) THREE PROPOSALS

### *Proposal A*

Remove all Trees within 10 meters of the house

### *Method:*

I took the entire fire data set and found the variable distance2tree and recoded anything that was less than 10 as 10. I ran the lasso through the cross validation again with the new fire data set

### *Results:*

I found the mean squared prediction error to be 0.1854609, which is slighly worse than our previous model's prediction error of 0.1583821. The new Proposal tells us that 218 Houses will burn and 269 will not burn. Therefore, we will save 56 houses from wildfire with this Proposal.

### *Proposal B:*

Require any home with at least 50% woody vegetation within 40m of the house to remove vegetation until they only have 50% woody vegetation within 40m of their house.

### *Method:*

I took the entire fire data set and found the variable perc.woody and recoded anything that had perc.woody greater than 50% as just 50%. I then ran the lasso model through the cross validation again with the new fire data set.

### *Results:*

I found the mean squared prediction error to be 0.1906117, which is slighly worse than our previous model's prediction error of 0.1583821. The new Proposal tells us that 217 Houses will burn and 270 will not burn. Therefore, we will save 57 houses from wildfire with this Proposal.

### *Proposal C:*

Replant any "remnant" vegetation within 40m of all houses with identical "planted" vegetation.

### *Method:*

I took the entire fire data set and found the variable planted and recoded anything that had plnated as r and replaced with a p. I then ran the lasso model through the cross validation again with the new fire data set.

### *Results:*

I found the mean squared prediction error to be 0.1854609, which is much worse than our previous model's prediction error of 0.1583821. The new Proposal tells us that 162 Houses will burn and 325 will not burn. Therefore, we will save 112 houses from wildfire with this Proposal.

| | Proposal A | Proposal B | Proposal C |
|---|---|---|---|
| Houses Burned | 228 | 227 | 53 |
| Houses Not Burned | 259 | 260 | 434 |
| Houses Saved | 46 | 47 | 221 |

*CONCLUSION:* I found the Lasso model with perc.woody and planted as predictor variables to be the best predictive model when compared to other models using the AIC technique, Ridge Regression and p-value variable selection. I used the Lasso model to make predictions based on the 3 Proposals. The above table gives us number of houses saved based on the Proposal A, B, C. I found that Proposal C to have the most number of houses saved. Therefore, I recommend that California implement Proposal C and remove any 'remnant' vegetation around houses.

- I think there is something wrong with my MSPE for each proposal. I thought I would get less error with the changing of the data set.