Data Analysis for Final exam

Amal Agarwal

STAT 511

Pennsylvania State University

Supervisor: Prof. Ephraim Hanks

December 2014

Contents

Chapter 1

Data description

1.1 House Loss to Wildfires

In 2009, Southern California was struck by a series of wildfires that burned more than 404,601 acres of land from early February through late November, destroying hundreds of structures, injuring 134 people, and killing two. The wildfires caused a lot of capital damage. In particular the month of August was especially notable for several very large fires which burned in Southern California, despite being outside of the normal fire season for that region.

The dataset "fire.RData" contains data from 487 randomly selected housesin neighbourhoods judged to be "at risk" to forest fires. Many characteristics of these houses were obtained from satellite images and other remote geospatial data sources. In particular, this includes 6 "characteristics near the home", that can be controlled by homeowners.

1.1.1 General problem of interest

The State of California is interested in using this data to make recomendations to homeowners on how to maintain their property to minimize the risk of fire damage to their home. The effects of following 6 "characteristics near the home" are of particular interest to analyze since these are the only items that home owners can control and minimize the risk of fire damage to their homes.

- "planted": This predictor provides a visual assessment of whether woody vegetation within a circle with a radius of 40m from the centroid of each house was predominantly planted ("p") or remnant ("r") using the prefire imagery.
- "buidings": This is the number of buildings within a radius of 40m from the centroid of each house.
- "perc.woody": This is the percentage of mapped woody vegetation within a circle with a radius of 40m from the house
- "perc.cleared": This is the visual estimate of the percentage of land without woody vegetation within a circle with a radius of 40m from the centroid of each house using the pre-fire imagery.
- "distance2bush": This is the distance from each house to nearest bush.
- "distance2tree": This is the distance from each house to nearest tree or shrub (m) visible.

Chapter 2

Data analysis

2.1 Exploratory Data Analysis

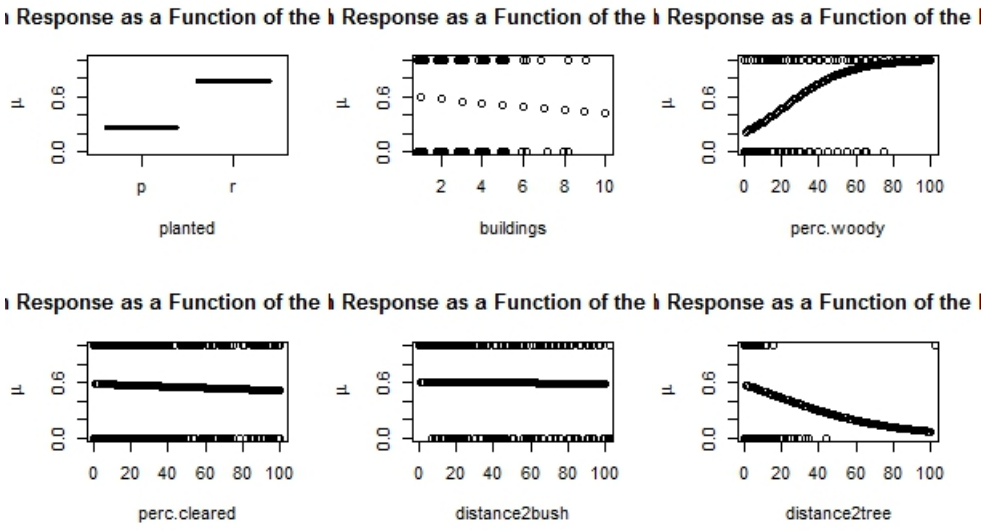
The response variable "burnt" is an indicator variable that takes values "1" if the house was burnt and "0" if the house was not burnt. The pairwise scatter plots of all the six "characteristics near the home" were analyzed to get a rough idea of the relationship between the response and the predictors.

In particular a total of 274 homes were burnt out of 487 observations that were made.

2.2 Answer 1

2.2.1 Part 1

To understand the relationship between the "burnt" response and 6 "characteristics near the home", logistic models were fitted with one predictor at a time. Note that the results from this kind of analysis can't be relied upon completely since there may be interaction terms that can affect the relationship under true model. But this is just exploratory data analysis and we assume for simplicity that there are no interactions to get a rough idea of what's happening.



These figures roughly demonstrate that the probability that a home is burnt in a wildfire is positively correlated with the percentage of woody vegetation within a circle of radius 40 m. Also it is higher for homes with planted status "r" i.e. those with nearby vegetation as a remnant of surrounding vegetation compared to those with "p" i.e. when the nearby vegetation is planted by homeowners. It is highly negatively correlated with the the distance from the nearest tree or shrub and slightly negatively correlated with the number of buildings and percentage of land without woody vegetation within a radius of 40 m. It appears to be not depending on the distance from the nearest bush.

2.2.2 Part 2

Now to find the difference in the fire risk of houses with similar nearby tree characteristics (i.e. perc.woody and distance2tree), but different planted status, we divided the whole dataset in two parts. One part contains those houses with planted status "p" and the other part contains those houses with planted status "r". Then we formed pairs taking one observation from each group and checking if the two houses in the pair have similar nearby tree characteristics i.e. if they satisfy the following two conditions:

- The difference between the percentage of woody vegetation within a circle of radius 40 m is less than 2.
- The difference between the distance to the nearest tree or shrub is less than 5.

We segregated those pairs satisfying the above conditions and calculated the difference in response among these pairs individually. Finally we get the mean of the difference of response as 0.074 which suggests that the houses with planted status as "r" have a 0.074 higher probability of fire damage compared to those with planted status "p" when they have similar nearby tree characteristics.

2.3 Answer 2

2.3.1 Part 1: Model Selection

First we started with the following simple GLM logistic model with all the predictors.

2.3.2 Model 1

$$\text{burnt}_i \sim Bern(p_i) \tag{2.1}$$

with

$$g(E(Y_i)) = \eta_i = X_i\beta \tag{2.2}$$

Here g is the canonical logit link in bernoulli model, i.e.

$$g(E(Y_i)) = g(p_i) = log\left(\frac{p_i}{1-p_i}\right) = X_i\beta \tag{2.3}$$

and $\eta_i = X_i\beta$ is the linear predictor.

There were no problems of multicollinearity with this model since all the variance inflation factors were well within 10. There were no influential points also according to both the leverage criterion and Cook's distance criterion. There were a couple of outliers as indicated by the studentized residuals plot. But since they are not influential, it's probably fine to leave them in.

The parameter estimates and high p values indicate that "4 out of 6 characteristics" don't matter in explaining the response viz. perc.cleared, buildings, distance2bush and distance2tree.

For model diagnostics, data was simulated from this model and the resulting residuals plots and QQ plots match with the ones obtained with original data. This verifies the model to some extent. However AIC of this model was 483.435 which may not be the lowest.

Various interaction terms were tried taking combinations of the 6 "characteristics near the home". It was found that the interaction term corresponding to percy.woody and planted was very much significant (low p value). Model 1 was modified by including this interaction term in the linear predictor.

Following this, multiple approaches for estimating the regression parameters viz. stepwise AIC, Ridge regression and LASSO regression were employed. The dataset was divided in two parts training and test in the ratio of $2:1$ (approx.). GLM modified model 1 was then fitted (with all the predictors) on the training set only.

2.3.3 Model 2 (Best AIC)

Stepwise AIC was performed on the above fit to get the best AIC model as:

$$\text{burnt}_i \sim Bern(p_i) \tag{2.4}$$

with

$$g(E(Y_i)) = \eta_i = X_i\beta \tag{2.5}$$

Here g is the canonical logit link in bernoulli model, i.e.

$$g(E(Y_i)) = g(p_i) = log(\frac{p_i}{1-p_i}) = X_i\beta \tag{2.6}$$

and $\eta_i = X_i\beta$ is the linear predictor as

$$\eta_i = \beta_0 + \beta_1 ffdi + \beta_2 topo + \beta_3 perc.looged + \beta_4 distance2tree + \tag{2.7}$$

$$\beta_5 planted + \beta_6 perc.woody + \beta_7 planted * perc.woody$$

The stepwise AIC results indicate that no predictor needs to be added or substracted in modified model 1 for attaining minimum AIC.

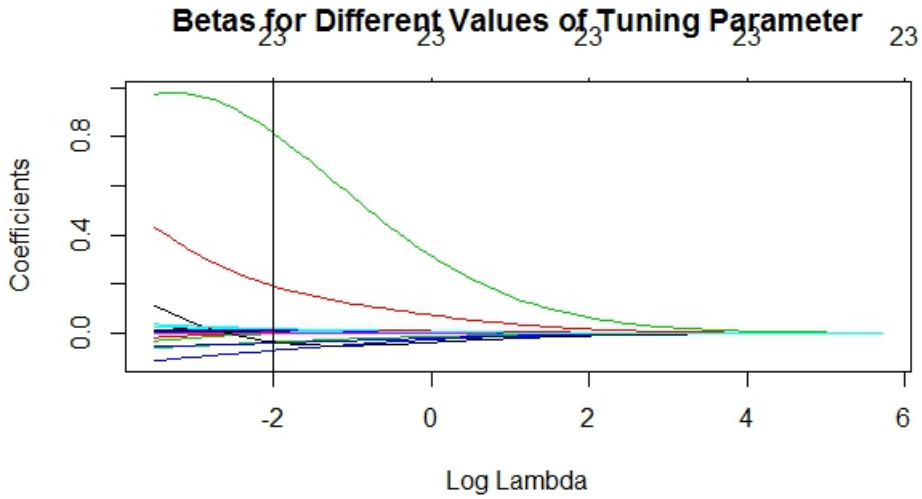
Prediction on test dataset using this best AIC model was performed. Results for MSPE are tabulated later.

2.3.4 Model 3 (Ridge Regression)

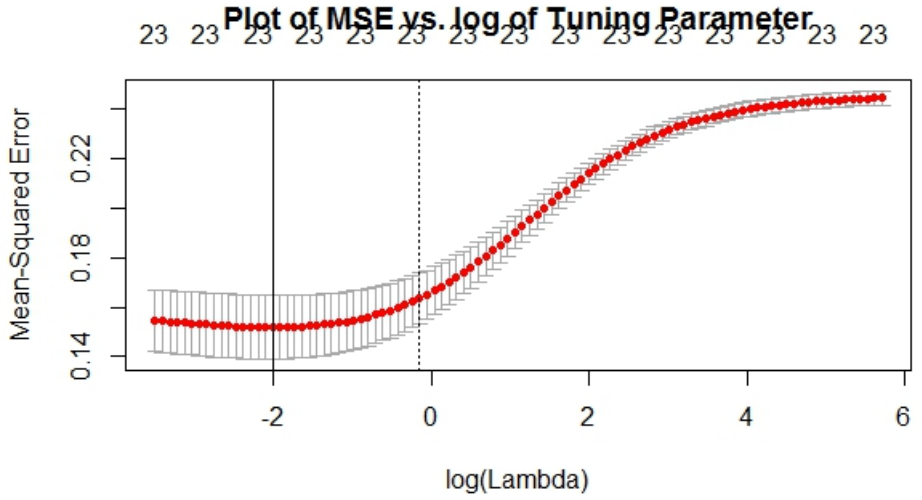
Ridge Regression is generally useful to alleviate multicollinearity. Variance inflation factors for the model 1 fitted on training dataset were calculated. It was found that the "edge" predictor had VIF as $14.45 > 10$. This suggests that Ridge regression might be useful. Ridge regression can be performed by adding the penalty term to the negative log-likelihood and minimizing the resultant function as:

$$\hat{\beta} = argmin\left(-l(Y|\beta) + \sum_{i=1}^{p}\beta_i^2\right) \tag{2.8}$$

Following are some relevant plots:



The above plot shows the different regrssion parameter estimates as functions of log of tuning parameter. The best value of lambda using 10 fold cross-validation was obtained as 0.133. The log of this value is also indicated as a vertical line.



The above plot shows the variation of Mean squared error with the log of tuning parameter. The curve slightly goes down initially attaining a minima at the best value of log of tuning parameter, increases between 0 to 2 and then flattens out after that.

Prediction on test dataset using this ridge regression model was performed. Results for MSPE are tabulated later.

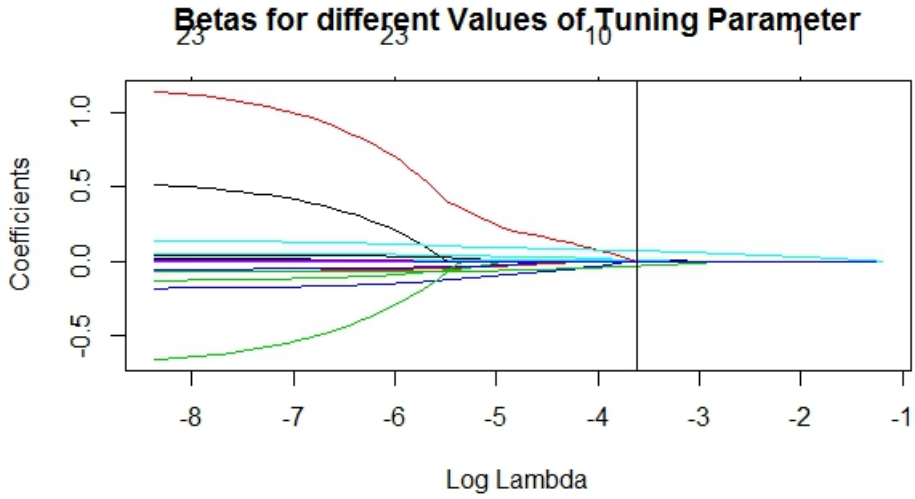
2.3.5 Model 4 (Lasso Regression)

Lasso is particularly useful for variable selection since it shrinks out some parameters to zero. Lasso regression can be performed by adding the penalty term to the negative log-likelihood and minimizing the

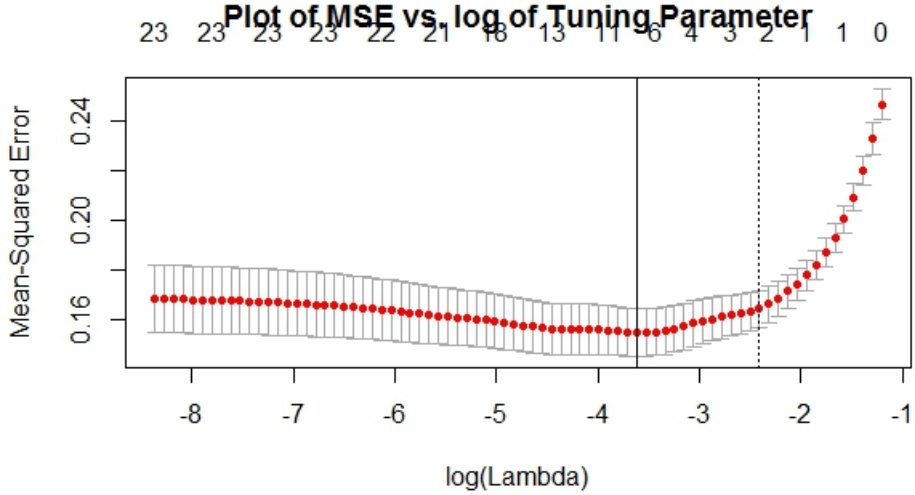
resultant function as:

$$\hat{\beta} = argmin\left(-l(Y|\beta) + \sum_{i=1}^{p} |\beta_i|\right) \tag{2.9}$$

Following are some relevant plots in this context:



The above plot shows the different regrssion parameter estimates as functions of log of tuning parameter. The best value of lambda using 10 fold cross-validation was obtained as 0.0268. The log of this value is also indicated as a vertical line.



The above plot shows the variation of Mean squared error with the log of tuning parameter. The curve slightly goes down initially attaining a minima at the best value of log of tuning parameter and then increases rapidly after -2.

Prediction on test dataset using this Lasso regression model was performed. Results for MSPE are tabulated later.

2.3.6 V fold cross validation

The training dataset was divided equally in 10 parts and cross validation was done for model 1 and model 2 (best AIC). The Mean Squared Prediction Error (MSPE) and Cross Validation- Mean Squared Prediction Error (CVMSPE) for all the models are tabulated below:

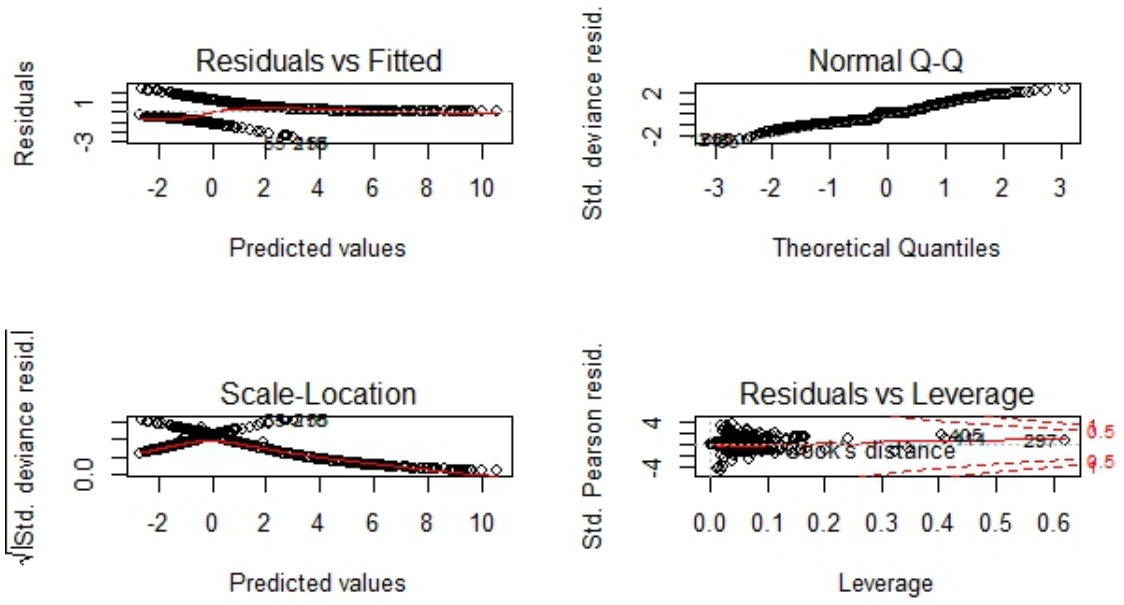
Summary	CVMSPE-Train	MSPE-Test
Model 1	13.754636	0.1630435
AIC	11.5944556	0.1594203
RR	0.1519503	0.192029
Lasso	0.1548702	0.1847826

From the above table, it can be concluded that the best predictive model is Model 2, the one obtained from stepwise AIC since it has the lowest MSPE on test set.

2.3.7 Part 2: Fitting with final model (AIC)

Since it has been established that the Model 2 obtained from stepwise AIC is the best predictive model, it can be now fitted over the whole dataset. The estimated parameter values indicate that the log of odds of a house suffering fire damage is positively correlated with the distance from the nearest tree. Since there is an interaction term present, if the planted status is "r", the log of odds increases with the percentage of woody vegetation within a circle of radius 40 m while if it is "p", the log of odds decrease.

A thorough model diagnostics was conducted for this final model. All the variance inflation factors are well below 10 and there are no problems of multicollinearity. There are no influential points according to leverage as well as the Cook's distance criteria. There are a couple of outliers but since they are not influential, they can probably be ignored. Following figure shows the residual plots, QQ plots and leverage & Cook's distance plots.



Note that the QQ plot looks very good. Further the leverage and Cook's distance plot confirms that there are no influential points.

The residual plots look similar to the those obtained for the simulated data under this model. This verifies the final model to a good extent.

Based on the data, it was found that this model correctly predicts whether the house will suffer fire damage or not, 80.493% of the time. This looks pretty good.

2.4 Answer 3

2.4.1 Part (a)

To find the effect of removing all trees within 10 m of all the houses, the subset of observations where there are any trees within 10 meters was obtained. Subsequently the trees were removed within 10 meters of all houses in this subset by making the "distance2tree" predictor i.e. the distance from each house to nearest tree or shrub equal to 10. With this modification in the design matrix, the response was predicted with our best predictive model 2 from lasso regression for this subset. The difference between the predicted response and the actual response was then averaged over all the observations in this subset. It was found that the probabilty of getting a house burned decreases by 0.0479.

2.4.2 Part (b)

To find effect of reducing all the woody vegetation which is above 50% to just 50%, similar steps as in part (a) were performed. It was found that the probabilty of getting a house burned decreases by 0.0489.

2.4.3 Part (c)

To find the effect of replanting any remnant vegetation within 40m of all houses with identical planted vegetation, similar steps as in parts (a) and (b) were performed. It was found that the probabilty of getting a house burned decreases by 0.68. This is a far greater decrease than earlier management practices. Thus this management practice is the most beneficial and most recommended.

If this management practice had been implemented before 2009, the number of houses that would have been burned would be 75 which is much less than the number of houses actually burned i.e. 274.

Chapter 3

Bibliography

- http://en.wikipedia.org/wiki/2009_California_wildfires