Anoka County Model

Chase Baggett

Before Looking at the Data

I am creating a holdout set of 20% of the data for testing my model, which is not viewed until the end.

Exploratory Data Analysis

I looked at only the training set, and very quickly I learned the data was very right skewed, so I log transformed both the predictors and the response. I generated scatter plots of the response by each of the continuous predictors faceted by the Use. It appeared evident to me that the slope was consideribly different for the continuous predictors by Use.

Model Selection

Using the transformed data, I was able to use added variable plots and p values to identiy that FTE was the strongest variable. Adding in additional ones and looking for significance, I only found FTE to be significant.

I suspected, however, that there might be some complex interractions with Use, but lacking any strong intuition about which businesses might generate waste, I relied on an automated selection technique. I have a holdout set I can test it on at the end, and will be competing it against my simpler model.

I built a model with every 2,3 and 4 way interaction with our continuous variables, and an interraction with each of the Uses with the continuous variables. I did traditional backwards stepwise removal using BIC. I passed the results of this into a second layer of backward selection that uses cross-validation at each step of the stepwise function.

For each step of the backward selection, I trained and tested the model using 80/20 train/test sampling five times. If the mean MSE of the five random test did not improve over the original model, the term was removed. I restored the marginality principle when my technique broke it.

Testing the Two Models on the Holdout Set

Simple Model $E(log(Wst)|Use, FTE) = log(FTE) + Use$

```
## [1] "Holdout MSE: 0.870820221244127"
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.62	0.55	1.04	8	0	6	-40.49	94.99	104.8	26.12	24

Automatically Selected Model $E(log(Wst)|Use, FTE) = Use.5 + log(LandV) + log(ImprV) + log(FTE) + log(Size) + Use.4 : log(ImprV) + Use.5 : log(ImprV) + Use.2 : log(FTE) + Use.2 : log(Size) + Use.5 : log(Size) + FTE : log(Size) + LandV : log(Size) + ImprV : log(Size)$

```
## [1] "Holdout MSE: 0.539914738720824"
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.77	0.6	0.98	4.68	0	13	-33.32	94.65	114.26	16.2	17