STA 371H Notes
Monday, March 31 2014

Agenda for this week:

Probability
Risk Modeling
Decision Analysis

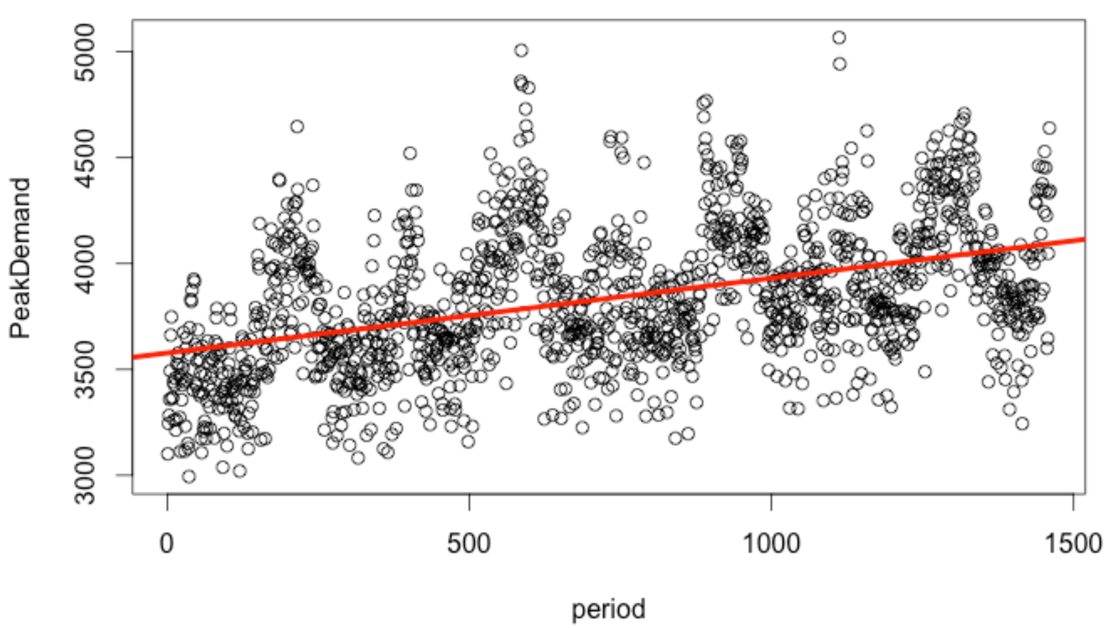
Peak Demand Data Set (first part of class)

Model choice on time series and forecasting
Data Set: peakdemand.csv
Goal: build a good forecasting model for peak demand.

Findings:

- Month: clear seasonal demand
  - More energy used during the summer and winter months
- Daily Temperature: parabolic relationship
  - We model this by using temp, and temp^2 as predictors
- Weekends: less demand for energy than on weekdays
  - People are home, therefore they use more energy
- Average Peak demand: grows over time
  - Model a time series by regressing on a *time index* (different column that counts the number of periods)

Regressing Peak Demand on Time Index



Question is: how do we develop a model that best predicts peak demand? What factors do we consider to account for these findings?

Critiques:

1. Not using all available information:

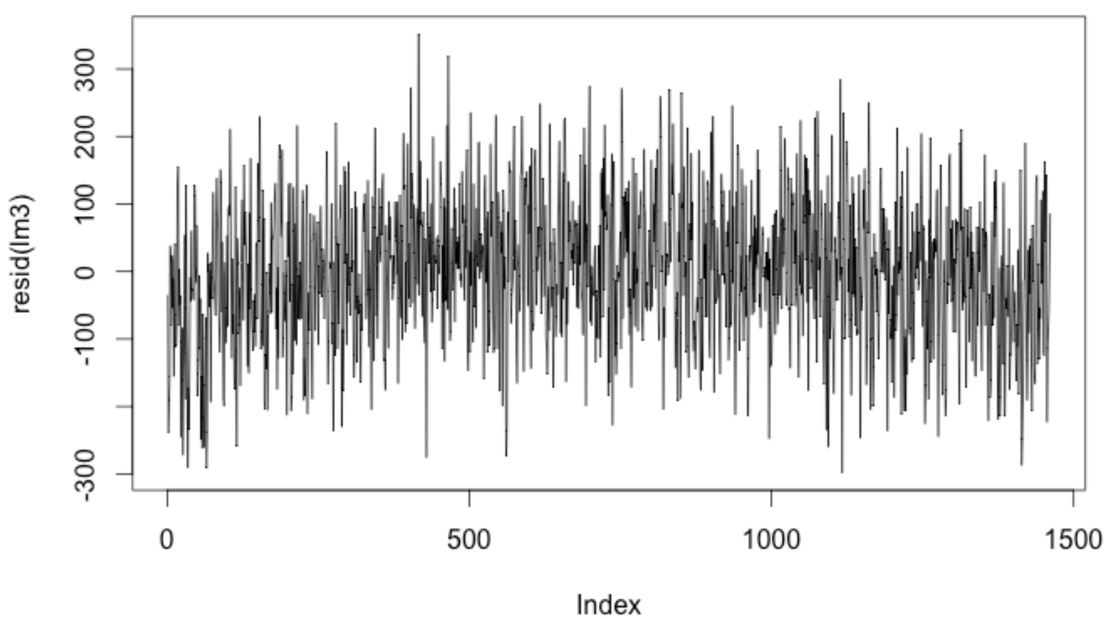
    If you're trying to forecast peak demand of energy today, what other information can we use? Peak demand of yesterday.

    We could build a forecasting model that uses the previous date to predict the following date. → using *lag predictors*

2. Some bowing of residuals, which indicate you probably didn't take all the x-ness out of y.

    Linear trend not perfect for predicting

    It's ok because we only have an 80% model, lag predictors and other factors take into account the other 20%



Big ideas:
Anova: Run an analysis of variance to see if any one variable is marginally & comparatively related. Specifically, look at Sum Sq to see if the variables are comparatively the same. If not, then get rid of the variable. In this case, the Sum Sq is relatively large for all variables and therefore we don't get rid of any.

```
> anova(lm3)
Analysis of Variance Table

Response: PeakDemand
                  Df   Sum Sq  Mean Sq   F value    Pr(>F)
period             1 32523200 32523200 2894.7026 < 2.2e-16 ***
DailyTemp          1 12015848 12015848 1069.4614 < 2.2e-16 ***
I(DailyTemp^2)     1 81481983 81481983 7252.2418 < 2.2e-16 ***
Sat                1   585949   585949   52.1519  8.28e-13 ***
Sun                1 23356438 23356438 2078.8219 < 2.2e-16 ***
factor(Month)     11   326603    29691    2.6426  0.002378 **
Residuals       1444 16223947    11235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In case we aren't' sure whether a predictor is significant or not, we can compare t statistics. T statistic- signal to noise ratio

> If higher than 2, you can predict pretty closely. Lm3 gives us massive t statistics for individual predictive variables. Lm2 and lm3 have r^2 pretty close, which shows us that although the models may be good, they don't really show us whether we can get rid of a variable.

```
> summary(lm3)

Call:
lm(formula = PeakDemand ~ period + DailyTemp + I(DailyTemp^2) +
    Sat + Sun + factor(Month), data = peakdemand)

Residuals:
    Min      1Q  Median      3Q     Max
-298.20  -69.38    2.11   70.23  351.84

Coefficients:
                 Estimate Std. Error t value
(Intercept)     6.744e+03  5.090e+01 132.485
period          3.269e-01  6.803e-03  48.057
DailyTemp      -1.180e+02  1.876e+00 -62.907
I(DailyTemp^2)  1.045e+00  1.667e-02  62.670
Sat            -1.179e+02  8.052e+00 -14.645
Sun            -3.665e+02  8.035e+00 -45.614
```

Do a permutation test or AIC test to see which is better. Remember lowest AIC is better, so if you drop month, AIC drops. Step(lm3)

```
> lmstep = step(lm3, direction = 'backward') #notice you
Start:  AIC=13643.39
PeakDemand ~ period + DailyTemp + I(DailyTemp^2) + Sat +
    factor(Month)

                 Df Sum of Sq      RSS   AIC
<none>                         16223947 13643
- factor(Month)  11    326603 16550549 13650
- Sat             1   2409609 18633556 13844
- Sun             1  23376336 39600282 14945
- period          1  25947689 42171636 15037
- I(DailyTemp^2)  1  44128014 60351960 15561
- DailyTemp       1  44461622 60685568 15569
```

This Occam's Razor simulation started with all predictors. Notice how the AIC increases when we drop the predictor *Month.* Therefore, we keep our original predictors in the model.

Exploratory data analysis (trying to use individual variables to get to the larger predictive model) benefits: (as opposed to just doing step wise and working backwards)

1. People like visual evidence for understanding how each variable works
2. Avoid easy pitfalls
    a. Could have easily put in month as a predictor, not knowing month should be a categorical variable

Probability

Notes in [Notes on Probability]() link of course pack
"Probability and Risk"

Basic rules (Kolmogorov's Axioms/Rules)

1. Probabilities (P) sum to 1
    a. Mutually exclusive events
2. P of disjoint events add together
    a. Mutually exclusive
    b. Students in Texas *or* Oklahoma
3. P must always be between 0 and 1
4. Connection between hockey and probability? Russians dominate both
    a. Steven asked a really good question here that made the whole class pause in awe.

More Complex Rules

1. *Addition Rule or Union Rule*
    a. P(A U B) = P(A) + P(B) – P(A,B)
        i. Read: "probability of A or B," "probability of A union B"
        ii. P(A,B) – takes both, so want to avoid double counting
            1. Joint probability: probability of A and B at once, P(A & B)

2. *Multiplication Rule*
    a. P(A,B) = P(A)*P(B| A)
        i. P (B|A) = "probability of B given A", "conditional upon"
        ii. = P (B, A)
        iii. = P (B)*P(A|B)

"But what does it mean?"
What doe sit mean to have a (joint) probability of 60%?

Two interpretations:
1. *Frequency interpretation*:
    a. P (A) = number of times A happens/ number of opportunities A was given to happen
        i. It's the limiting case
        ii. If you had a hypothetically infinite case, what limit would the ratio approach
2. *Degree-of-belief interpretation or "Fair Value" on a bet interpretation*
    a. Subjective (e.g. chance of rain) (Vegas)
        i. Denominator is 1; there's only one chance of something happening
        ii. Starting on page 5
    b. Fair value on a $100 bet (Wall Street)
        i. If there's a $100 contract on a bet, how much are you willing to pay to have someone holding that contract
            1. E.g. I bet there's a 45% chance of rain, therefore I won't bet more than $45. I won't profit if I bet more.

Bayes' Rule

Proof:
P(A|B)*P(B)= P(B|A)*P(A)
P(A|B) = P(B|A)*P(A)/ P(B)

Ex. 1
A and B are events that are familiar (e.g. ordering a fun book on Amazon)
P(A): Zach guessed 50% Dana ordered Hunger Games, class "ooohed"
P (B): Gets new piece of information: Dana ordered *Fault in our Stars*
How does Zach's guess change? Well, he drops down to 5%. Bummer.

Posterior Probability: P(A|B)
Prior Probability: P(A)
Update Factor: P(B|A) / P(B)

Ex. 2
Event G: accused person is guilty

Presumption is person is innocent, then new piece of evidence is presented. How does this change the situation?

Event D: accused person's DNA matched (their DNA matches the crime scene DNA gathered)

Next time: How to calculate this

Coding

```
library(mosaic)

#load peakdemand.csv

#Time Series
N = nrow(peakdemand)
peakdemand$period = 1:N
head(peakdemand)

plot(PeakDemand ~ period, data=peakdemand)
lm1 = lm(PeakDemand~ period, data=peakdemand)
abline(lm1, col='red', lwd=4)

# Adding in the other factors
lm2 = lm(PeakDemand~ period+ DailyTemp + I(DailyTemp^2)+ Sat+Sun,
data=peakdemand)
plot(resid(lm2),type='l')

lm3 = lm(PeakDemand~ period+ DailyTemp + I(DailyTemp^2)+
Sat+Sun+factor(Month), data=peakdemand)
plot(resid(lm3),type='l')

anova(lm3)
lmstep = step(lm3, direction = 'backward') #notice you don't want to delete
anything
```