Scribing Monday: 4/06/15

Overview:

Finish up logistic regression (logit)

ACL Case study: explore the data to find patters of correlation

Time Series Analysis: Introduction and Mauna Loa walkthrough then Peak Demand from homework

Homework due Tuesday 04/14/15

Logistic regression: useful for predicting binary outcomes

$P(y_i=1)= e^\wedge(\psi_i)/1+ e^\wedge(\psi_i)$ – this is the link function: it binds values to (0,1)

$\psi_i=B_0+B_1x_{i2}+B_2x_{i2}+\ldots+B_nx_{in}$ – this is the linear predictor

for ψ_i in r, use glm(response~predictor, data=”” ,family=binomial)

Further explanation of math in Chapter 9 of course packet

Reason through a logistic model the same as an ordinary lm:

Look through the data, make plots, examine coefficients, table the binary outputs- a table is graphic description of binary data (think titanic data)

Interpret coefficients in the context of the link function

- They should make sense as predictors
- They can be used to make forecasts for archetypal models
- They can be hard to understand; meaningful in their predictive value

Things from ordinary lm that translate to logit:

- Fitted values
- Generalizing single to multi variable models
- Coding dummy and interaction terms
- Stepwise selection if black box model

think of logit as a souped up lm that predicts binary output

Question on when to use Stepwise selection:

Stepwise selection is good if making a black box model or wanting to prune down a large model (if P predictors 2^P models possible). Step( ) uses an optimization algorithm to find the lowest AIC. AIC shows predictive value, not model robustness in the context of a desired coefficient or variables.

Think of stepwise selection as a machete used to hack through a jungle of data. Your brain and other tests, such as a permutation test, act as more of a fine scalpel when selecting the best model in the desired context

ACL fest: what other festivals predict participation in ACL?

10-15 minutes working in class then discussion

Notable highlights from data:

- Bonnaroo/HL had a negative coefficient, which decreased ACL chance
- Outsidelands/HL was the largest predictior with a coeffieient of 1.52
- If you input 0 for all values x_i to find the intercept of the link function, the output is 0.05896 (the probability of appearing only at ACL)

Time Series Analysis & Forecasting:

Analyzing data over time. You usually observe the same thing over time in the same units. Ex: y_t = population of x in month t.

This is a broad topic so the focus is on two main issues:

Trends: long term trend of the time series, does the data generally go up or down?

If so, how much? If not, why? Ex: Global ocean temperatures have risen since 1850.

Seasonalities: cycles in data, relative highs and lows.

Ex: traffic density on I-35 fluctuates between rush hours and 3:30 a.m.

Use dataset "Mauna Loa Atmospheric CO_2" – walkthrough available here

Trends:

Plot the data- the upward trend is easy to see

Key strategy for tends: regress on time index

Time index counts periods since beginning t

Time series $y_t=B_0+B_1t+ \varepsilon_t$

t is the time index, think of time itself as a predictor

B_1 is the linear trend, it gives the average per period growth or decay in y_t

Eyeballed $B_1=\Delta Y/ \Delta X =380-315/550-1=\sim0.12$ parts per million per month.
To check eyeballed estimate, add a time index to be used as a variable (see walkthrough for r specifics). Run coef( ); $B_1=0.1134$, which is close to eyeball

The linear trend may not be a perfect predictor when looking at the global trend.
If more accuracy is required, a non-linear trend may be a better fit.

Seasonalities:
Plot the data-the annual cycles are easy to see

Key strategy for seasonalities: use seasonal dummy variables to define repeating patterns

Boxplot residuals by month and order them chronologically

mean monthly differences vary

Make the model with time index and month as predictors.

$y_t=B_0+B_1t+B_{month}+ \varepsilon_t$

you can predict missing values in the data by calling predict() based on existing data

Peak demand for electricity- data in Homework from Duke energy

The goal is to build the best predictive model using trends and seasonal variation.

Useful for knowing when to build a new power plant