STAT 400 Data Analysis Nov. 30, 2005

Review:

I Cleveland:

Ch 1 - Intro
Shortcomings of Ordinary Histograms and Scatter Plots for Data Analysis
Value of visualizations for data analysis

Ch 2 - Univariate Data

- Singer Data – describe structure of it – why approp here?
- Multi-panel graphs
- Quantiles and Quantile Plots, interpolation
- Q-Q plots and Normal Q-Q Plots – just like scatterplots?
- m-d plot – what does it show?
- box plot – when is it most useful?
- Fits and Residuals – Understanding the process
- Monotone Spread – why is it worth noticing?
- Additive and Multiplicative shifts
- Power Transformations and Log Transformation – when needed
  - how to choose a good one, and why needed?
- r-f plot, s-l plot – what do they show?

Ch 3 – Bivariate Data

- Banking to 45˚ - aspect ratio – why? how?
- Loess Details – choice of alpha, weight function, grid (role of grid?)
- Weighted Least Squares – why?
- Bisquare – robust fitting – role of weighting – why? how?
- Density estimation – kernel estimation – how and how to explain?
  - weight function role, grid
- Jittering (when useful?)
- Slicing, choosing overlapping slicing intervals (regression connection?)
- Role of variables – prediction vs fitting the data
- Time Series examples
  - Iterative residual analysis
  - Cut-and-stack plots
  - Cycle Plots
- Brushing and Labeling

Ch 4 – Trivariate Data

- Matrix Plots
- Coplots of Z on X_1 and X_2
- Interaction of X_1 and X_2 on Z - two coplots
- Brushing again
- Coplots of Fitted Surfaces – rubber data, ethanol data
- parametric vs non-parametric approaches (as in ethanol exericise)
- Cropping – why? – only for smoothed surfaces?
- Level Plots – compare with contour plots
- Coplots vs Level Plots – when is one better than the other?
- Improvisation (using context to adapt methods)
- Contour Plots – method of producing from grid
    - use of colour
    - choice of number of contours
    - effect of smoothing
- Level Plots of (Fitted, usually) Surfaces
- Wireframe Plots
    - varying perspective

Ch 5 Hypervariate Data

Role of Variables
Matrix Plots and Visual Linking (Brushing)
Four Variable Coplots (1 dept + 3 indept)

- Problem with equal count algorithm (p 133 and p 278)
- Looking for interactions in columns, rows and along diagonal of plot matrix
- Use of 3 coplots for same data
- Verbalizing the visualization

Cropping – simple and complex – needed?
Coplots of hypervariate surfaces – how different from coplots of hypervariate data
s-l plots of residuals – purpose and consequences
q-q plots of residuals – purpose and consequences
Simplification using Indexes (e.g. Iris Elongation p 300)

Ch 6 Multiway Data

Categorical Variables and Counts
Transformations of Counts (positive measurements are often skewed right)
Making use of unspecified characteristics of a graph (like order of labels, order of panels)
Median ranking for ordering – robustness
Modeling Counts – additive and multiplicative fits – use for residual analysis
Spatial (or geographic) location conditioned on frequency (Fig 6.11 p 319)

Sections 6.3 can be omitted. 6.4 relates to barley data in book intro.

II Non-Cleveland Data Sets:

Mercedes – nonparametric smoothing (loess and moving averages)
periodicity and seasonality
bias vs imprecision – choice of degree of smoothing
residual analysis
role of graphics

Bimbo – modeling and simulation to assess missing data
comparison of distributions with ecdf
maximization by evaluation on a grid
role of graphics

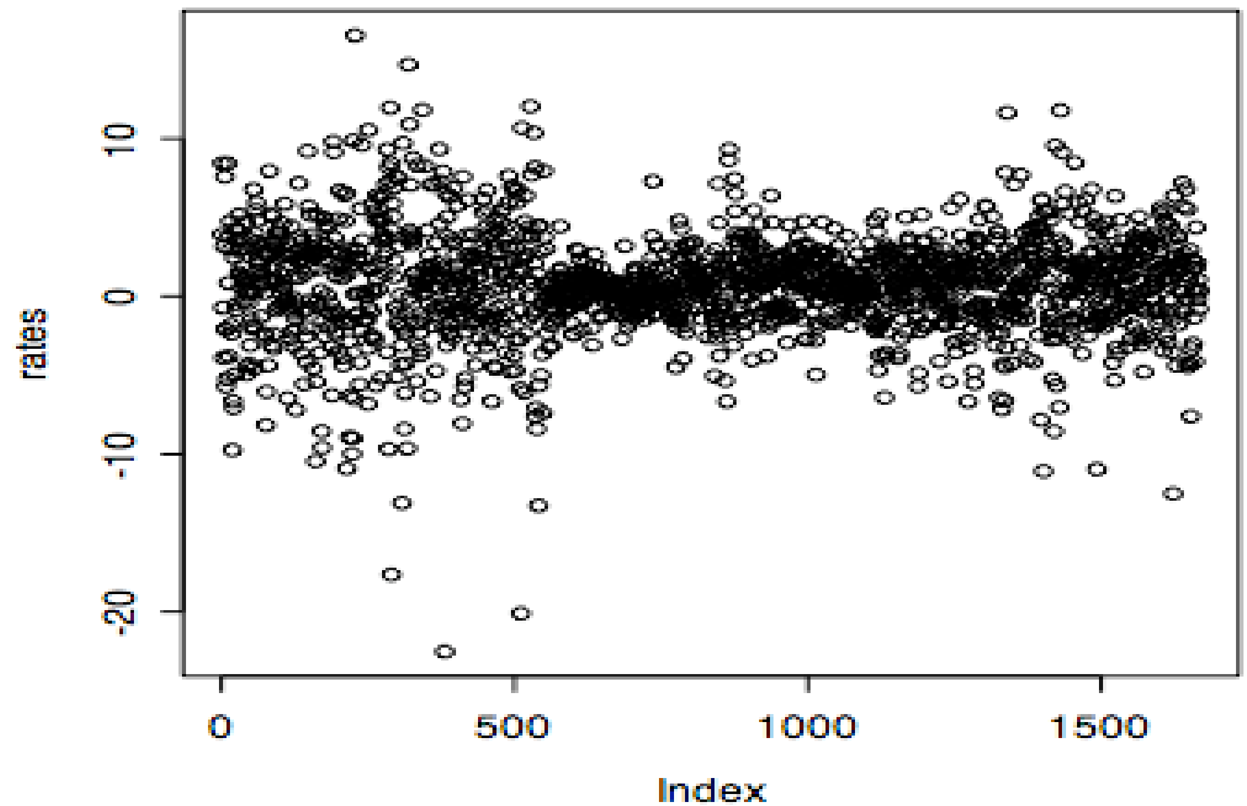
Men – predictive vs explanatory models
role of graphics and residual analysis
use and abuse of stepwise methods
the bootstrap for assessing variability
and applicability to complex procedures
parametric approaches – strengths and limitations

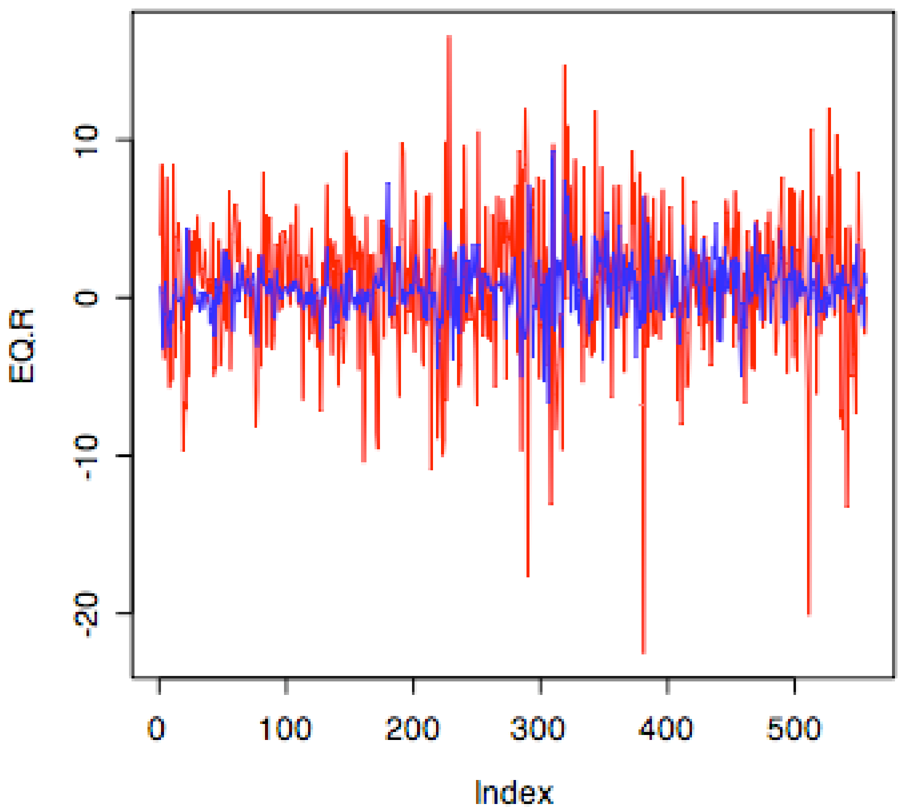
Ubi – preprocessing of data with index formation
confounding between experimental conditions
hazards of one-at-a-time adjustment
predictive vs explanatory models
dummy variables
transformations
use of dummy variables
linear spaces as approximations
role of graphics and residual analysis
compromising to obtain partial information

TSE - variability as a characteristic of interest
time series serial correlation – symmetric random walk
logarithms and percentage changes

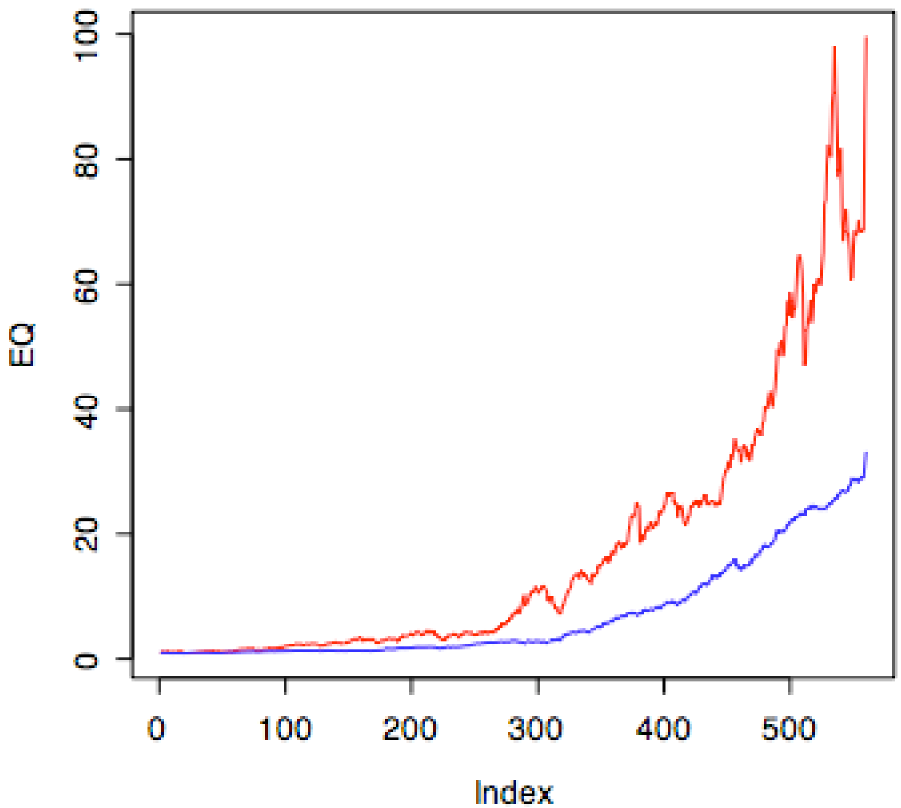
TSE data: Equities and Bonds 1956-2002

```
rates=c(EQ.R,BD.R,MIX.R)
> plot(rates)
> plot(EQ.R, type="l",col="red")
> lines(BD.R,col="blue")
```

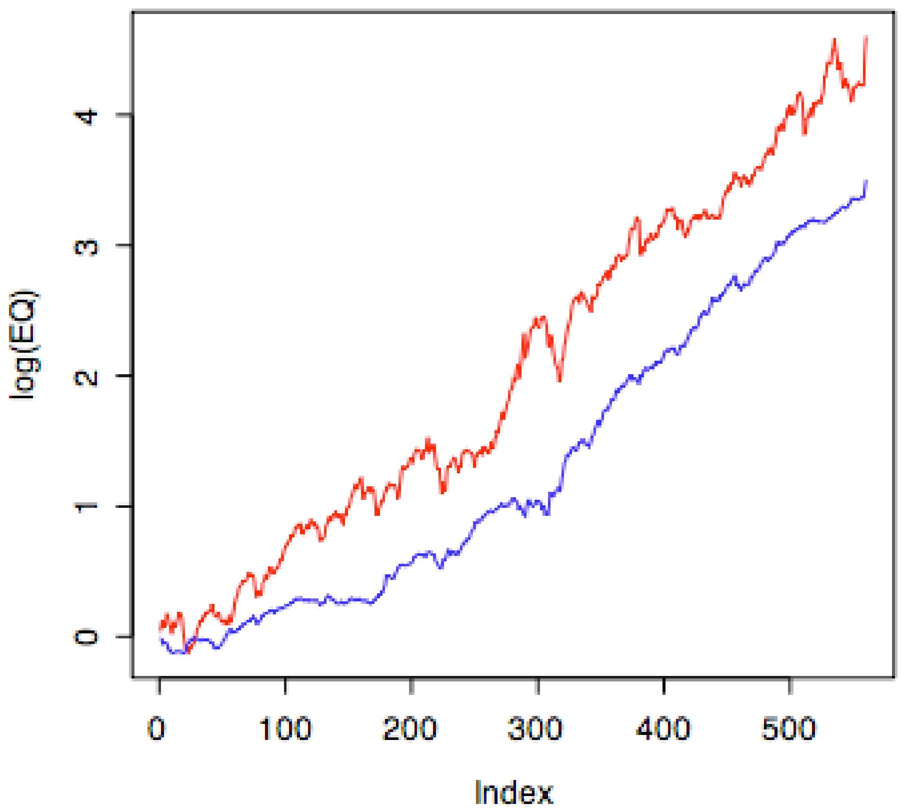





```
plot(EQ,type="l",col="red")
> lines(BD,col="blue")
```



Now take logs ....(why?)



SD is used as a measure of "risk". Is this right?