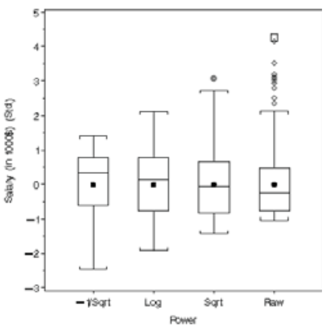
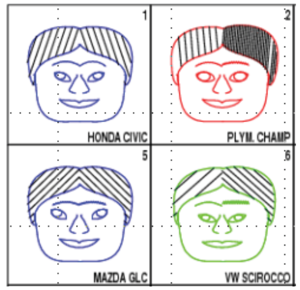
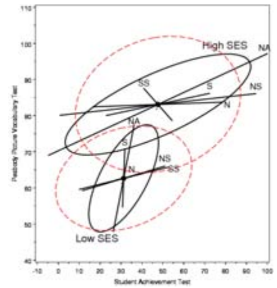






Graphical Methods for Data Analysis & Multivariate Statistics

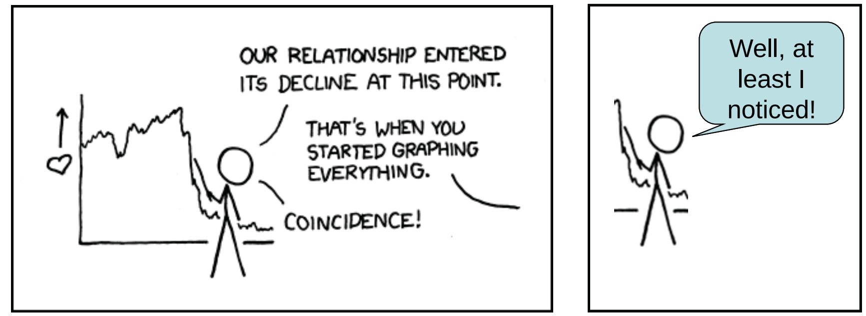
Michael Friendly
Psychology 6140

Why plot your data?

Graphs help us to see

patterns, trends, anomalies and other features

not otherwise easily apparent from numerical summaries.

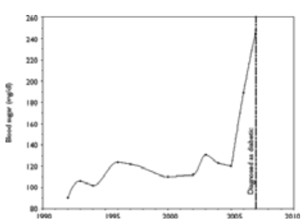


Source: http://xkcd.com/523/

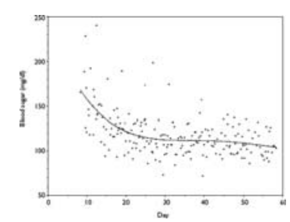
How graphs can change your life (n=1)

Personal analytics

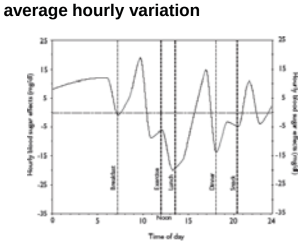
15 yr. blood sugar, pre-diagnosis



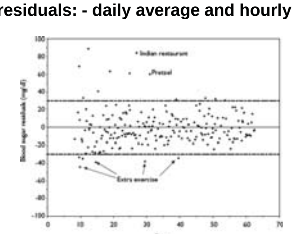
daily average, after diagnosis



average hourly variation



residuals: - daily average and hourly



A statistician contracts diabetes, and uses graphs to monitor his blood sugar.

→ Visual feedback on diet & exercise reinforce behavioral change

→ Residual plots show unexplained events, possibly important

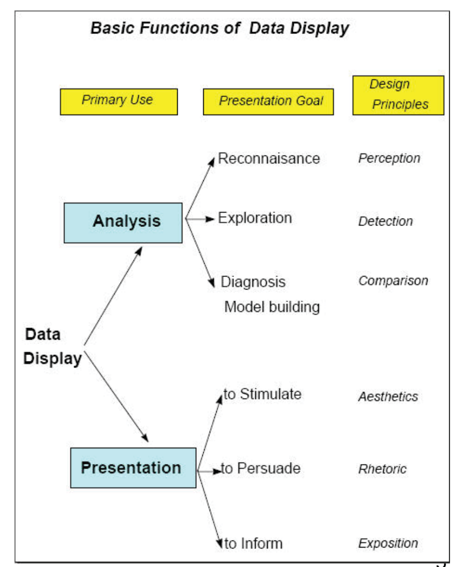
Ref: Wainer & Velleman, Looking at blood sugar, *Chance*, 2008, 21(4), 56-61

Different graphs for different purposes

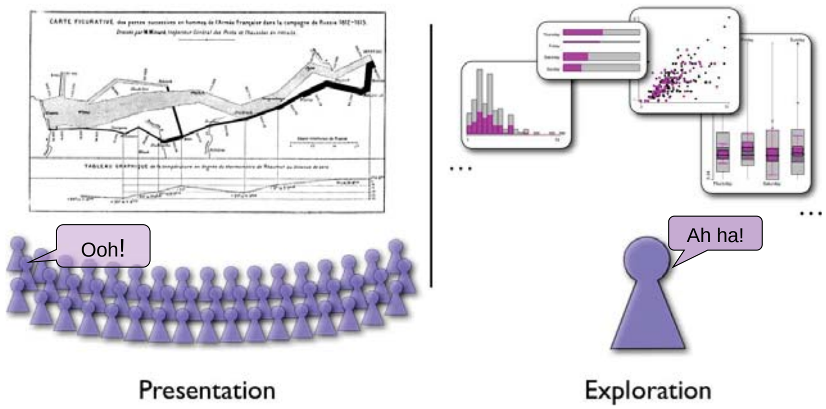
Graphs (& tables) as communication:
- What audience?
- What message?

•**Analysis graphs**: design to see patterns, trends, aid the process of data description, interpretation

•**Presentation graphs**: design to attract attention, make a point, illustrate a conclusion



Different graphs for different purposes

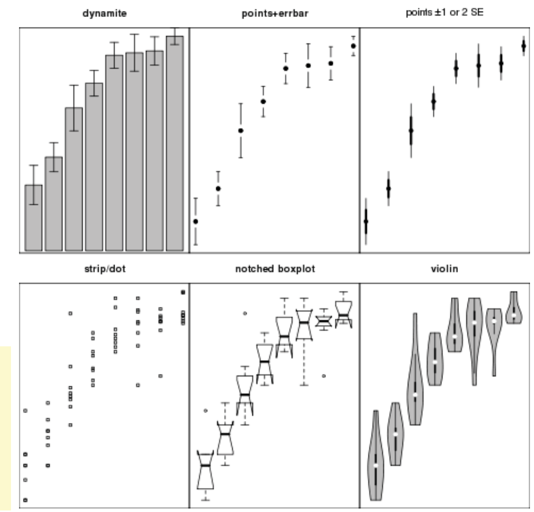


Presentation graphs: single image for a large audience
Exploratory graphs: many images for a narrow audience (you!)

Comparing groups: Analysis vs. Presentation graphs

Six different graphs for comparing groups in a one-way design

- which group means differ?
- equal variability?
- distribution shape?
- what do error bars mean?
- unusual observations?



Never use dynamite plots

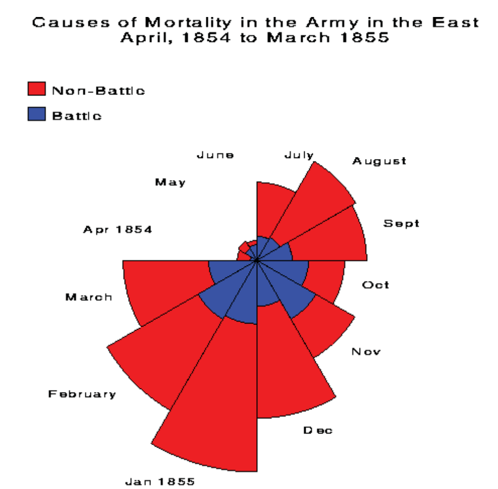
Always explain what error bars mean

Consider tradeoff between summarization & exposure

Presentation graph: Nightingale's coxcomb

Florence Nightingale: Deaths in the Crimean war from battle vs. other causes (disease, wounds)

She used this to argue for better field hospitals (MASH units)



From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858

The best presentation graphs pass the **Interocular Traumatic Test**: The message hits you between the eyes!

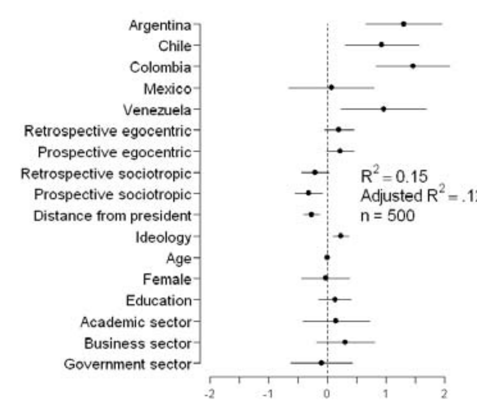
Presentation: Turning tables into graphs

Table 2 from Stevens (2006): Determinants of Authoritarian Aggression

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
Countries	
Argentina	1.31 (.33)$^{\#\#\#\ B,M}$
Chile	.93 (.32)$^{\#\#\#\ B,M}$
Colombia	1.46 (.32)$^{\#\#\#\ B,M}$
Mexico	.07 (.32)^{A,CH,CO,V}
Venezuela	.96 (.37)$^{\#\#\ B,M}$
Threat	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12)$^{\#}$
Retrospective sociotropic economic perceptions	-.21 (.12)$^{\#}$
Prospective sociotropic economic perceptions	-.32 (.12)$^{\#\#}$
Ideological Distance from president	
Ideology	
Ideology	.23 (.07)$^{\#\#\#}$
Individual Differences	
Age	.00 (.01)
Female	-.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	-.10 (.27)
R^2	.15
Adjusted R^2	.12
n	500

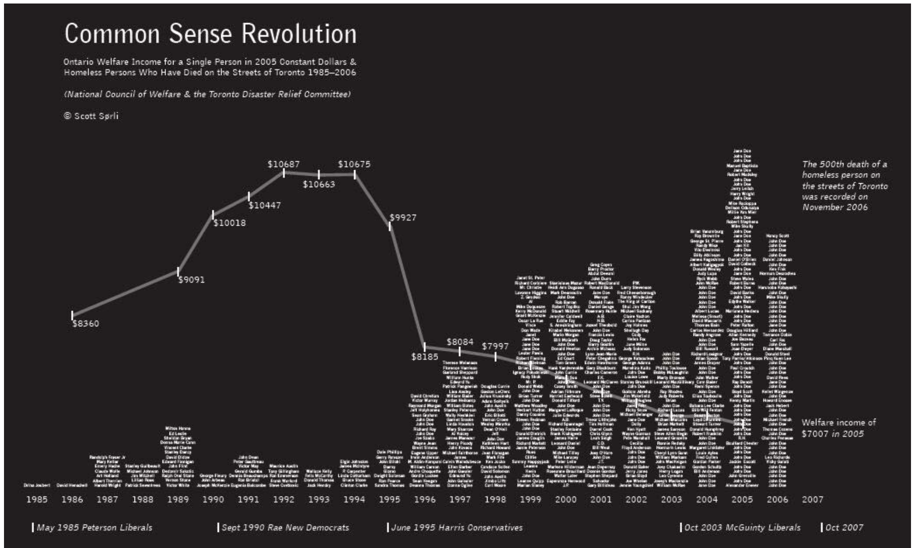
$^{\#\#\#}p < .01$, $^{\#\#}p < .05$, $^{\#}p < .10$ (two-tailed)
[A] Coefficient is significantly different from Argentina's at p < .05;
[B] Coefficient is significantly different from Brazil's at p < .05;
[CH] Coefficient is significantly different from Chile's at p < .05;
[CO] Coefficient is significantly different from Colombia's at p < .05;
[M] Coefficient is significantly different from Mexico's at p < .05;
[V] Coefficient is significantly different from Venzeluela's at p < .05

Graphs of model coefficients are often clearer than tables



Source: tables2graphs.com

Rhetorical graph: Common Sense Revolution

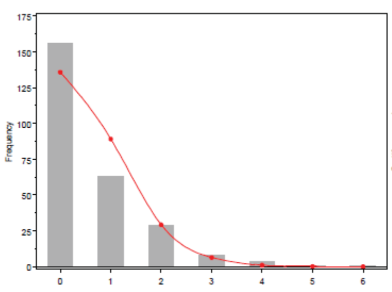


Effective data display

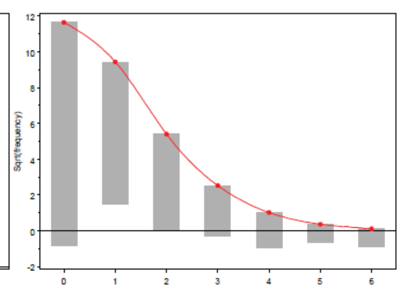
- Make the data stand out
  - Fill the data region (axes, ranges)
  - Use visually distinct symbols (shape, color) for different groups
  - Avoid chart junk, heavy grid lines that detract from the data
- Facilitate comparison
  - Emphasize the important comparisons visually
  - Side-by-side easier than in separate panels
  - “data” vs. a “standard” easier against a horizontal line
  - Show uncertainty where possible

Comparisons— Make visual comparisons easy

- Visual grouping— connect with lines, make key comparisons contiguous
- Baselines— compare *data* to *model* against a line, preferably horizontal
- Frequencies often better plotted on a square-root scale



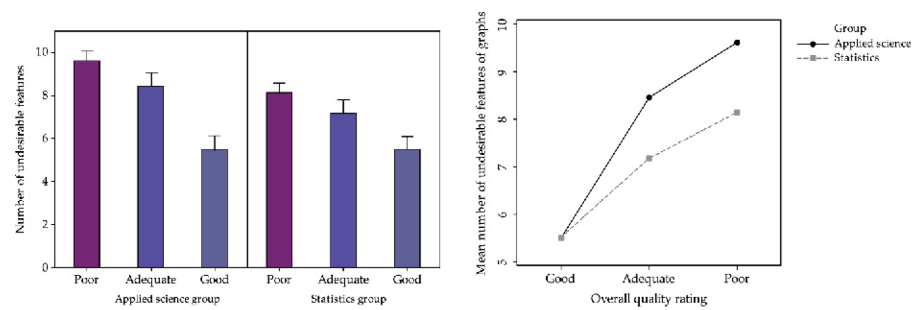
Standard histogram with fit



Suspended rootogram

Make comparisons *direct*

- Points not bars
- Connect similar by lines
- Same panel rather than different panels

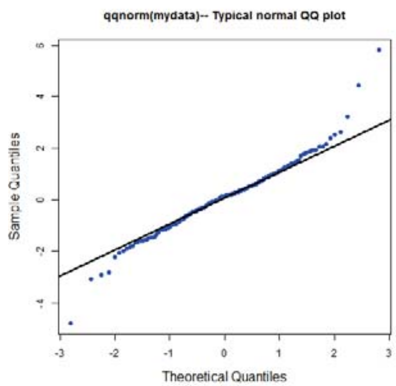


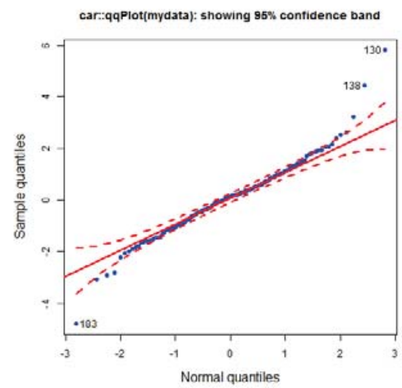
Published in: Ian Gordon; Sue Finch; *Journal of Computational and Graphical Statistics* **2015,** 24, 1210-1229.
DOI: 10.1080/10618600.2014.989324

Showing uncertainty

- Standard plots of observed vs. predicted lack a basis for assessment of uncertainty
- Confidence envelopes indicate extent of deviation
- Identify "noteworthy" observations to track them down

Example: Normal QQ plots used to assess normality of data



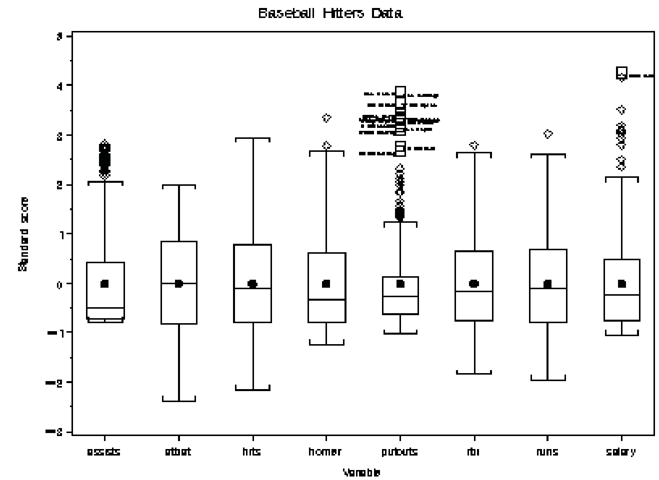


Analysis graph: Screening

Side-by-side boxplots of variables in the baseball data show the shapes of distributions --- aid to transformation

- Each variable is standardized to allow comparison.
- Plot is produced by **datachk** macro.

See:
http://datavis.ca/sasmac/datachk.html



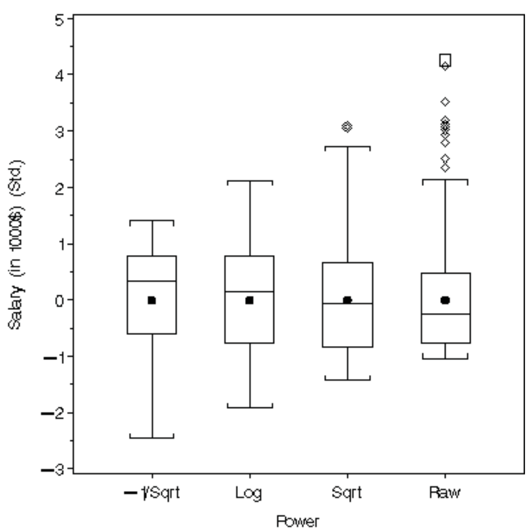
Exploratory graphs: Transformations

Data often needs to be transformed to meet analysis assumptions:

- Symmetry (~ Normality)
- Linear relations
- Constant variance

For symmetry, a **symbox** plot shows a variable transformed to various powers.

SAS: symbox macro
R: car package: symbox()



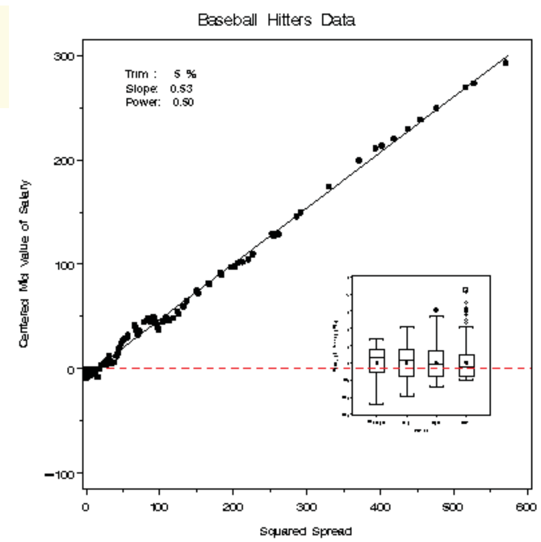
Diagnostic graphs: Transformations

Diagnostic plots can be used to suggest corrective action, often by a power transformation: $y \rightarrow y^p$

Symmetry transformation plot:

- Constructed so symmetric data plots as horizontal line
- Slope (b) of data line → power: p = 1 – b → $y^p = y^{(1-b)}$

Other diagnostic plots use the same idea: slope (b) → $y^{(1-b)}$



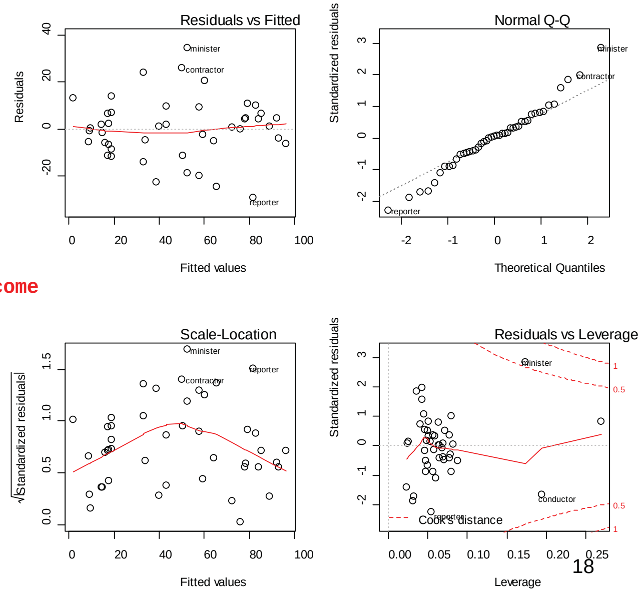
Model diagnosis: regression quartet

Statistical software should make it easy to get informative diagnostic plots

In R, plotting a **lm** model object → the "regression quartet" of plots

```
> model <- lm(prestige ~ income
            + education)
> plot(model)
```

(SAS has similar, using ODS graphics)

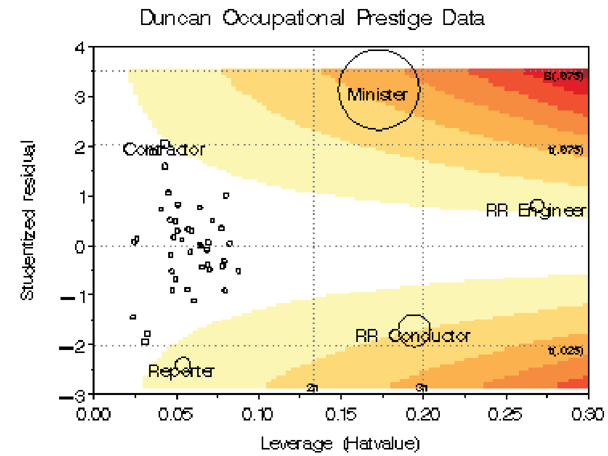


Model diagnosis: Influence in regression

Multiple regression model: prestige ~ income + education

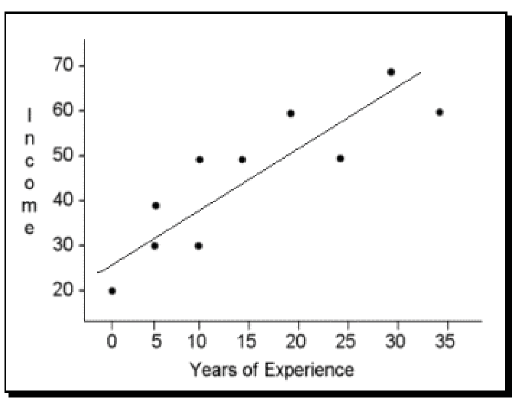
Influence plots can show:

- model residual
- leverage (potential impact)
- influence ~ residual x leverage (Cook D statistic)
- contour map of influence



Scatterplots: A basic workhorse for quantitative data

- Show the relation between two Q variables (ignoring all others!)
- More useful when enhanced to show visual summaries
- Vary point color/shape to show strata/groups
- Combine in multi-panel displays to show more
  - Scatter plot matrix: all pairs
  - Conditional relations: Y vs. X stratified by Group

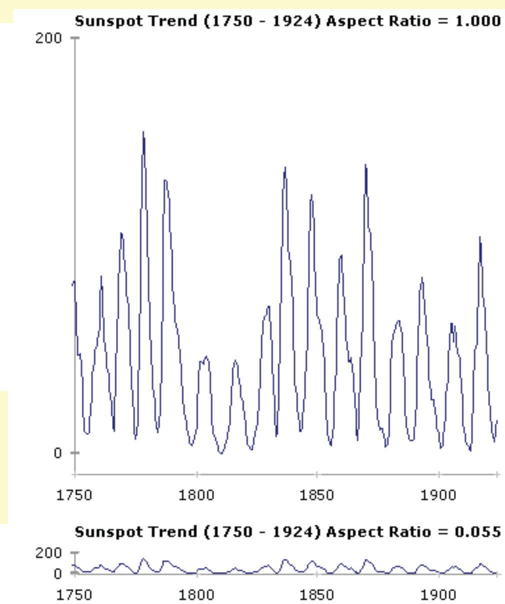


Scatterplots: Scales matter

Computer plots are usually generated with a given *aspect ratio*, to conform to the page or screen.

A better idea is to scale the plot so that slopes of lines or curves average ~ 45 degrees.

In the rescaled version, we can see that, within each cycle, sunspots tend to increase more quickly than they decline.

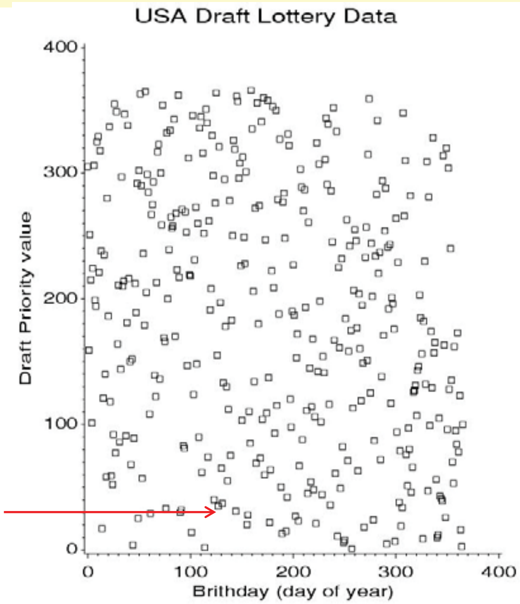


Scatterplots: Annotations enhance perception

Data from the US draft lottery, 1970

- Birth dates were drawn at random to assign a “draft priority value” (1=bad)
- Can you see any pattern or trend?

Me (May 7):
127 → priority = 35

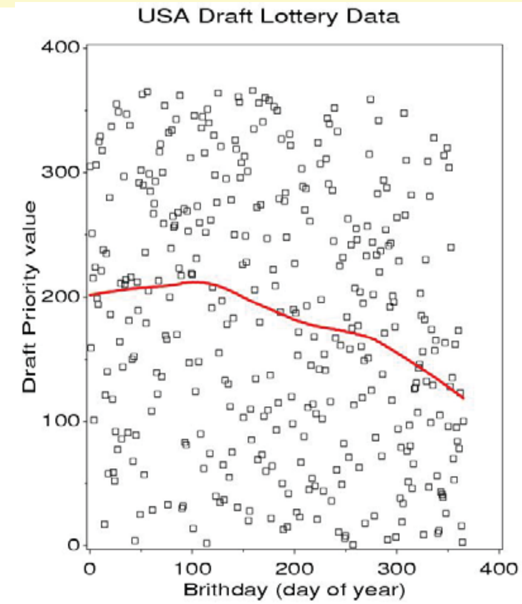


Scatterplots: Annotations enhance perception

Drawing a smooth curve shows a systematic decrease toward the end of the year.

- The smooth curve is fit by **loess**, a form of non-parametric regression.

Visual explanation:

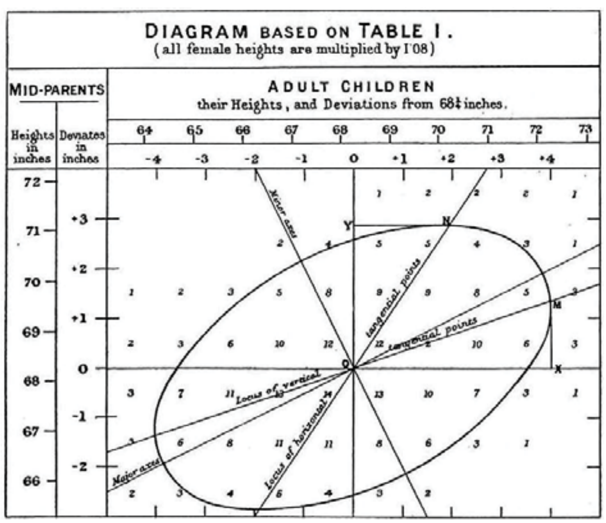


Scatterplots: Data ellipses

Galton’s (1886) semi-graphic table, showing relation of mid-parent’s height to children’s height.

As shown:

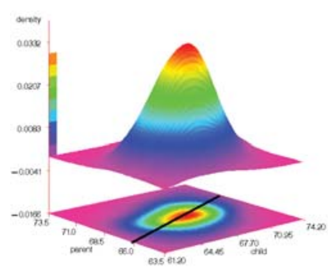
- Contours of equal frequency formed ellipses
- Regression lines of Y on X and X on Y are the loci of vertical and horizontal tangents
- Major/minor axes are the principal components

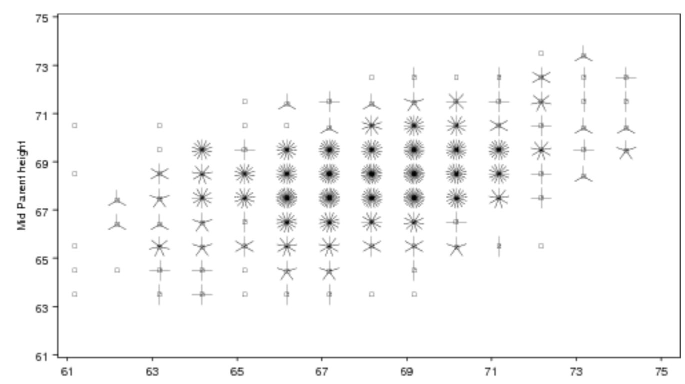


Scatterplots: Data ellipses

Galton’s data on child & mid-parent heights, shown as a sunflower plot: each sunflower symbol shows the number of observations in the (x, y) cell.

2D density estimate of bivariate surface





Scatterplots: Data ellipses

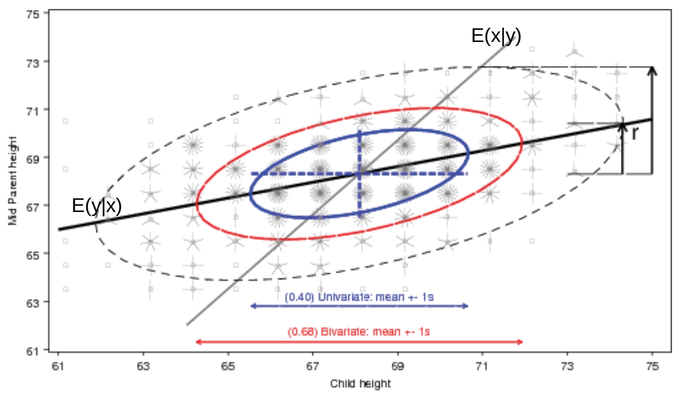
Any scatterplot can be summarized by data ellipses (assuming normality). These show: means, standard deviations, and allow correlations & regression lines to be visually estimated.

Data ellipse:

$$D^2(y) \approx \chi^2_p(1-\alpha)$$

Galton data, 40%, 68% & 95% data ellipses. Sizes are:

- X^2 (0.40) = 1.0
- X^2 (0.68) = 2.28
- X^2 (0.95) = 6.0



Visualizing multivariate data

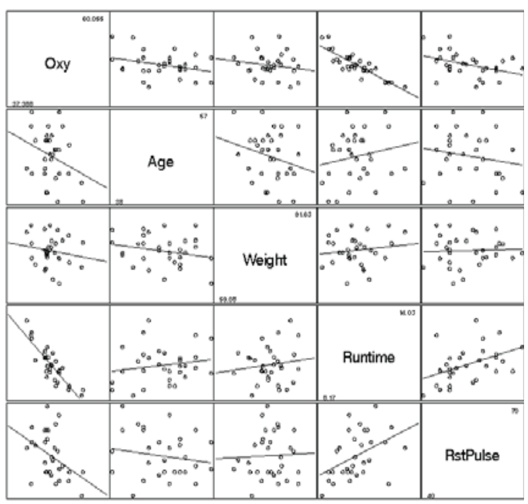
Showing relations among 3 or more variables:

- Scatter plot matrices (enhance with visual summaries, thin for many variables)
- Conditional plots: Y ~ X | (Z, Group)
- Seeing multivariate profiles, clusters:
  - Star plots, face plots, parallel coordinates
- Biplots: project data into low-D view

Scatterplot matrix

- **Fitness data**:

  Oxy ~ Age + Weight + Runtime + Rstpulse
- Each panel shows row var vs. col var
- Reg line shows *linear* relation

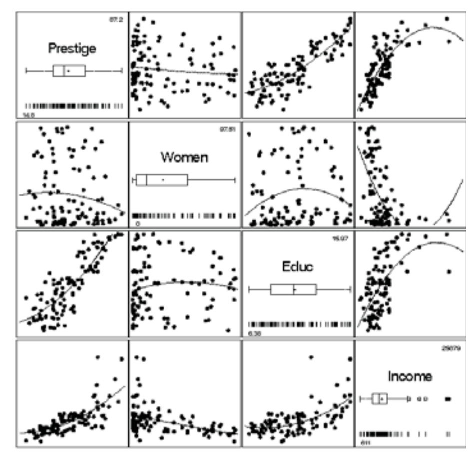


Questions:
- What is the best predictor of Oxy?
- Which two predictors are most highly correlated?

Scatterplot matrix

- **Occ. prestige**:

  Prestige ~ %women + Educ + Income
- Box, rug plots show univar. distributions
- Quadratic regressions show linear/non-linear relations (loess would be better)



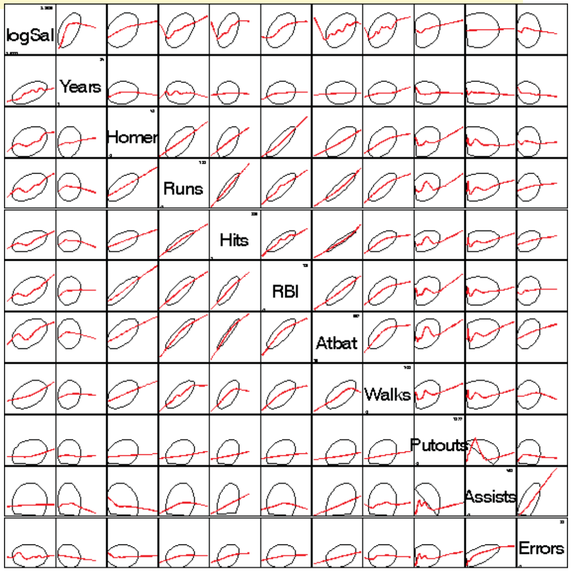
Questions:
- How should Educ be modeled?
- How should Income be modeled?

Larger data sets: Visual thinning

Baseball data: log(Salary) ~ performance variables

- Too much data to show individual points
- Each scatterplot is summarized by a loess smoothed curve and a data ellipse

Questions:
- Which variables are most strongly related to logSal?
- Which relations are strongly nonlinear?
- Which predictors are too highly correlated?



Larger data sets: Corrgrams

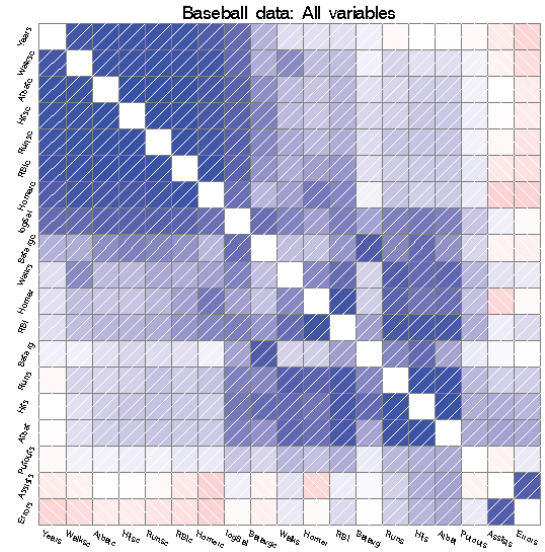
Correlation diagram shows **pattern** of correlations for many variables.

Variables are re-ordered to make the groupings most visually apparent.

This graphic assumes that all relations are linear, not necessarily always true

Graph using SAS **corrgram** macro,
http://datavis.ca/sasmac/corrgram.html

R: corrgram package

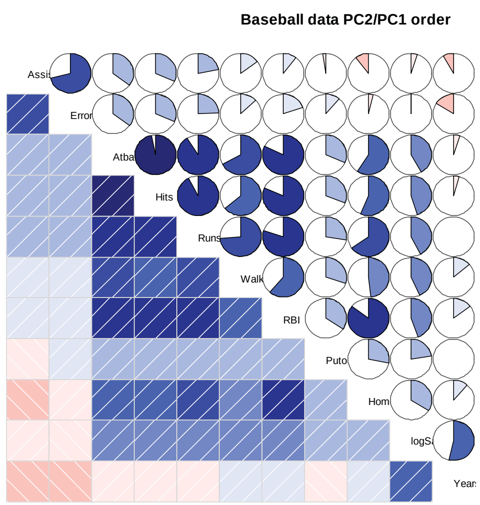


Corrgrams: Different renderings

The value of a correlation may be rendered in different ways, with different visual impact.

- Shading levels: help detect similar values
- Pie symbols: make it easier to compare for larger/smaller

Graph using R **corrgram** package



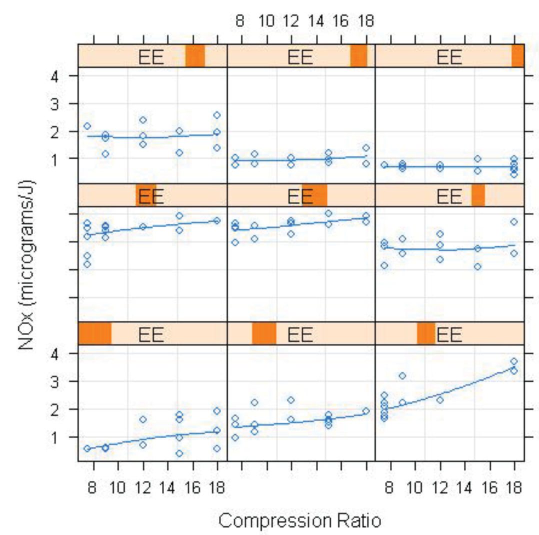
Conditional plots: Y ~ X | Z

Often want to explore how the relation between Y and X depends on/ varies with some other variable(s) Z.

- Moderator variables
- Interactions

Emission of NOx from ethanol in relation to engine compression ratio and richness of air/ethanol mixture (EE)

Graph using R **lattice** package



Conditional plots: Y ~ X | Z

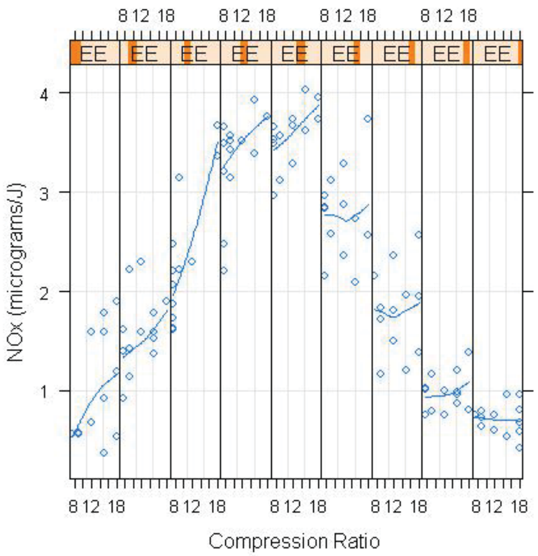
The same data is shown in a different format, with

• loess smooth curves

• curves banked to ~ 45º

The joint dependence on CR and EE is now much clearer

(These are examples of **lattice plots**, produced using R software.)



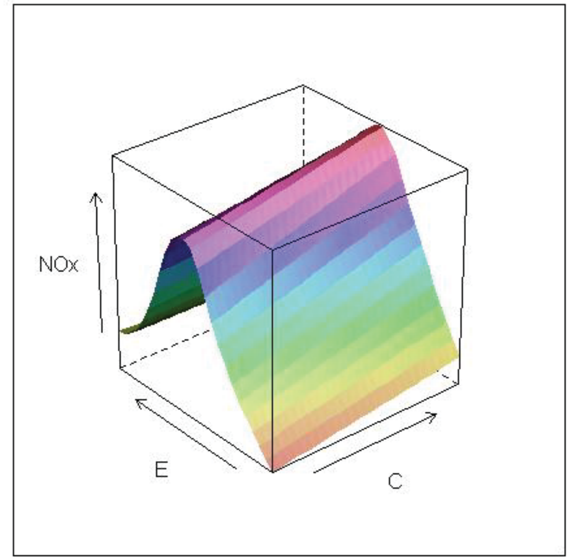
3D plots

Often not useful, unless done with great care.

This plot shows the loess **smoothed** predicted values of NOx in relation to EE and CR. (But, raw data not shown.)

Color is used to show the predicted NOx, using a "heatmap" color scale.

The interpretation is simple!



3D plots

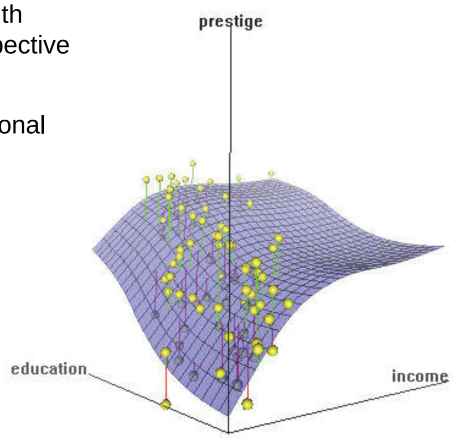
3D plots can be enormously useful with dynamic, interactive software & perspective

This plot shows a relation of occupational prestige to income & education.

• points are shown in perspective, connected to the fitted surface

• the fitted surface (linear, quadratic, smoothed) can be changed interactively

• the plot can be rotated dynamically to see other views



Seeing multivariate clusters: face plot

A faces plot assigns variables to facial features, to show **configural patterns** of many variables.

Pros: Easy to see similar patterns in large data sets.

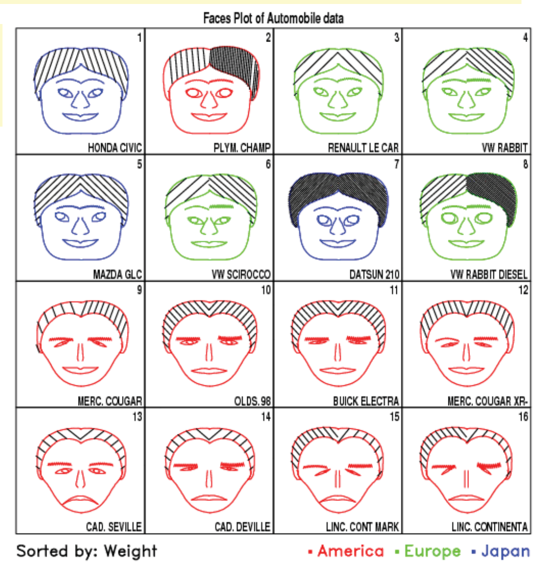
Cons:

• Hard to connect features to variables for interpretation

• No good rules/ideas for assigning variables to features.

Graph using SAS **faces** macro,

http://datavis.ca/sasmac/faces.html

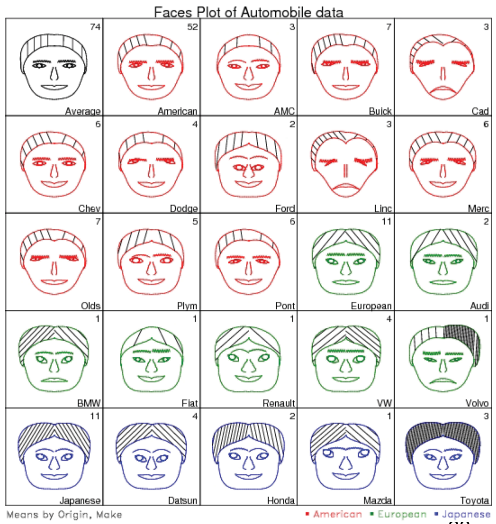


Seeing multivariate clusters: face plot

Variable assignment key

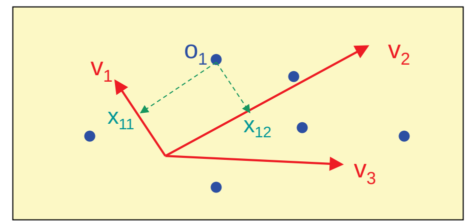
Parameter	Left Side Variable	Right Side Variable
Eye size	mpg	mpg
Pupil size	mpg	mpg
Pupil position	turn	turn
Eye slant	turn	turn
Eye X position	hroom	hroom
Eye Y position	hroom	hroom
Eyebrow curvature	rseat	rseat
Density of eyebrow	rseat	rseat
Eyebrow X position	displa	displa
Eyebrow Y position	length	length
Upper hair line	rep77	rep78
Lower hair line	weight	weight
Face line	weight	weight
Hair darkness	rep77	rep78
Hair shading slant	gratio	gratio
Nose line	length	length
Mouth size	price	price
Mouth curvature	price	price

Means, by make & origin

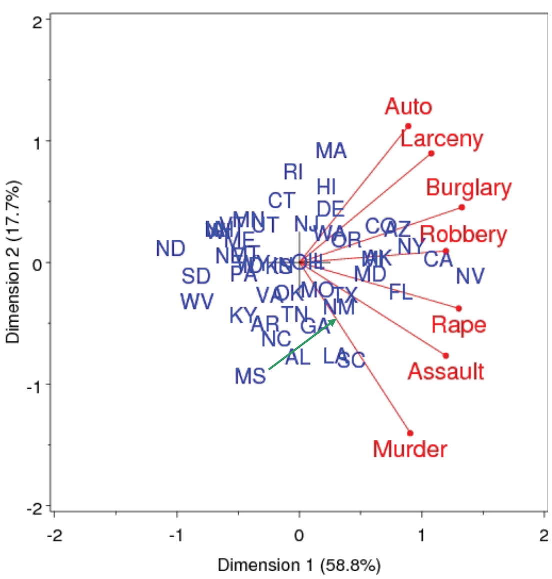


Biplots: variables and obs. in low-D View

- Based on PCA: data is shown in 2D (3D) view that accounts for greatest variance
- Observations: plotted as points
- Variables: vectors from origin (=mean)
- Angles between vectors ~ correlations
- Projection of point on vector ~ score



Biplot: US crime rates



Dim1: ~ Overall crime rate

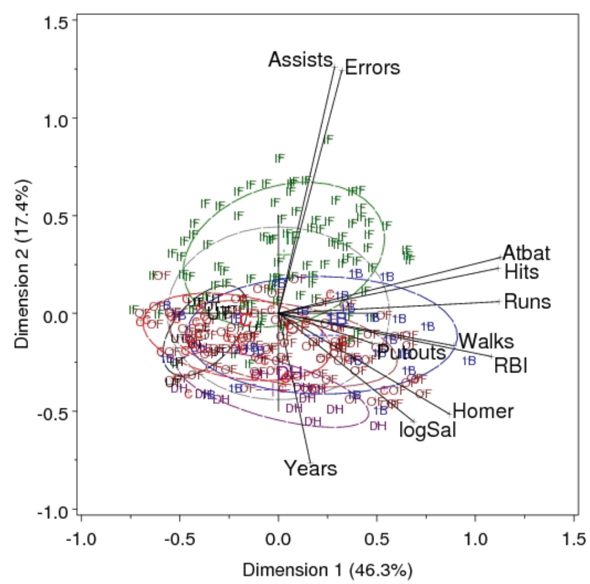
Dim2: Property vs. personal

Note: clusters of southern, New England, western states

This 2D biplot only shows 76.5% of total variance.

Still, it gives a useful summary of 9 variables and 50 observations.

Biplot: Baseball data



Baseball hitters' data:
- Dim2: fielding, -years
- Dim1: batting performance

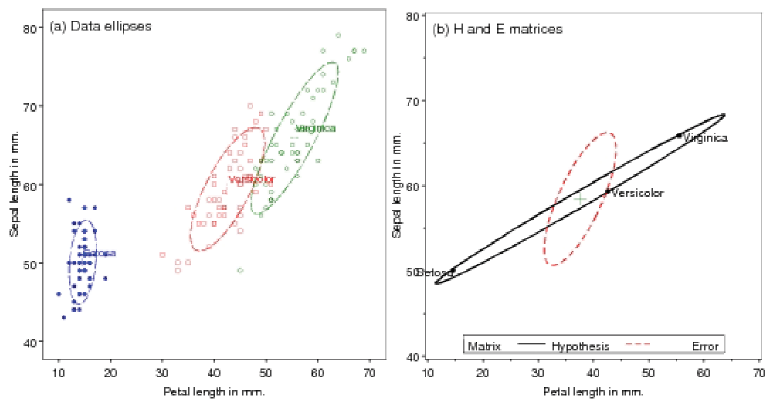
Players identified by position, with data ellipses for each
- IF: more assists, errors
- DH: older

This 2D biplot only shows 63.7% of total variance.

HE plots for MANOVA, MMReg

HE plots provide a way to visualize hypothesis tests in MANOVA and multivariate multiple regression, using data ellipses for fitted (H) and residual (E) co-variances.

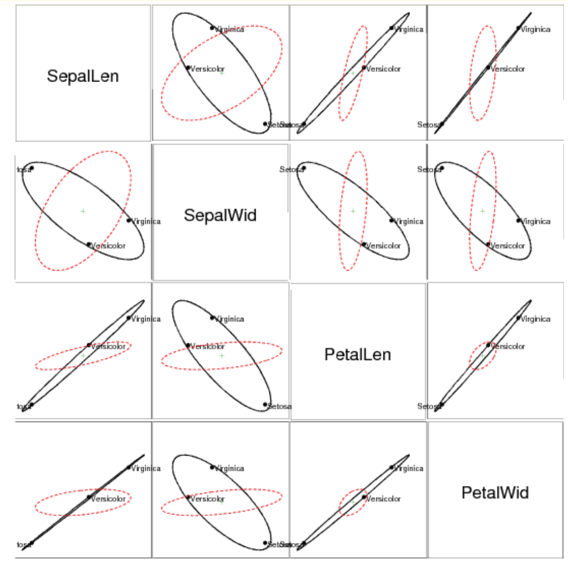
Graphic ideas: (a) Data ellipses summarize H & E (co)variation; (b) Scale H ellipse so it projects outside E ellipse *iff* effect is significant (Roy test)



HE plot matrices

HE plots in a scatterplot matrix show effects for all pairs of responses.

For the iris data, the Species means are highly correlated on all variables except Sepal length.

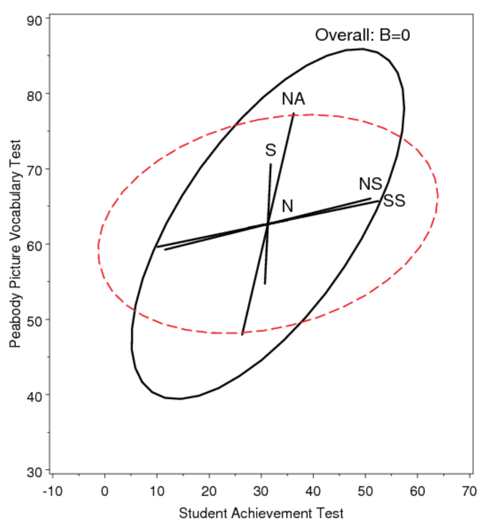


HE plots: MMRA

Rohwer data: Cognitive ability and PA tests: n=37, Low SES group

(SAT, PPVT, Raven) ~ n + s + ns + na + ss

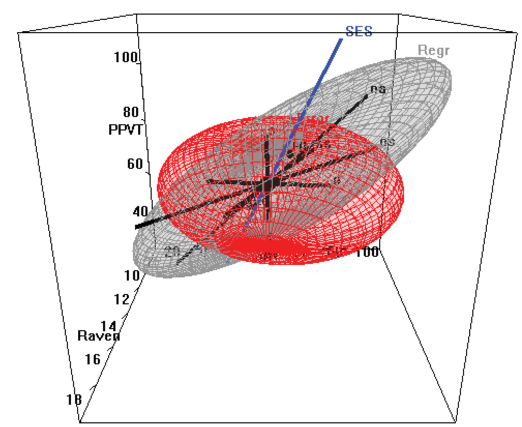
- Only one predictor, NA, is (barely) significant
- Yet, overall multivariate test: H_0: $\mathbf{B} = \mathbf{0}$ is highly so!



HE plots: MMRA & MANCOVA

Rohwer data: Low SES & Hi SES groups

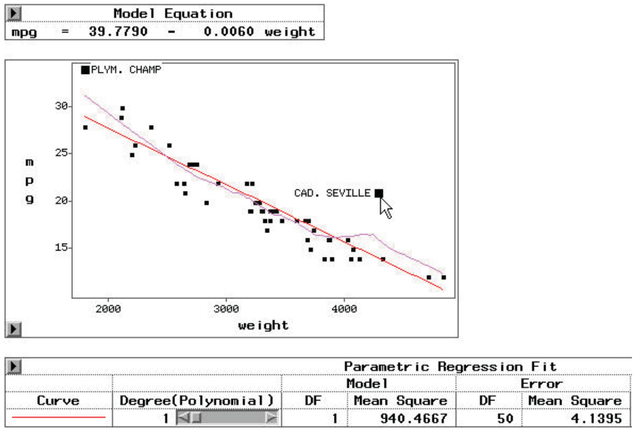
(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss



Dynamic, interactive graphics

Interactive graphics & data analysis provides:

- Identifying points
- Model & display controls



SAS/Insight: mpg ~ weight, **linear** fit

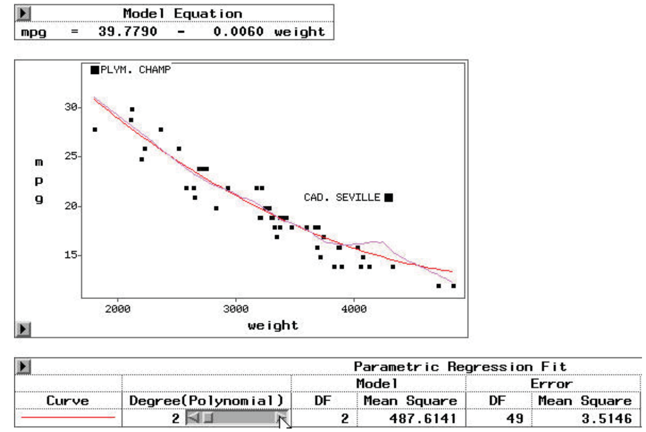
Dynamic, interactive graphics

Interactive graphics & data analysis provides:

- Identifying points
- Model & display controls

These methods are much more highly developed in R

- googleVis
- shiny
- ggvis
- ggobi -> rggobi



SAS/Insight: mpg ~ weight, **quadratic** fit

Dynamic, interactive graphics

Dynamic graphics provide multiple, linked views of a data set

Selecting points, regions in one plot ("brushing") selects the same observations in all other plots

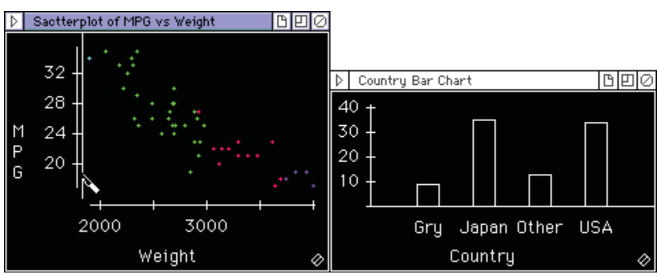
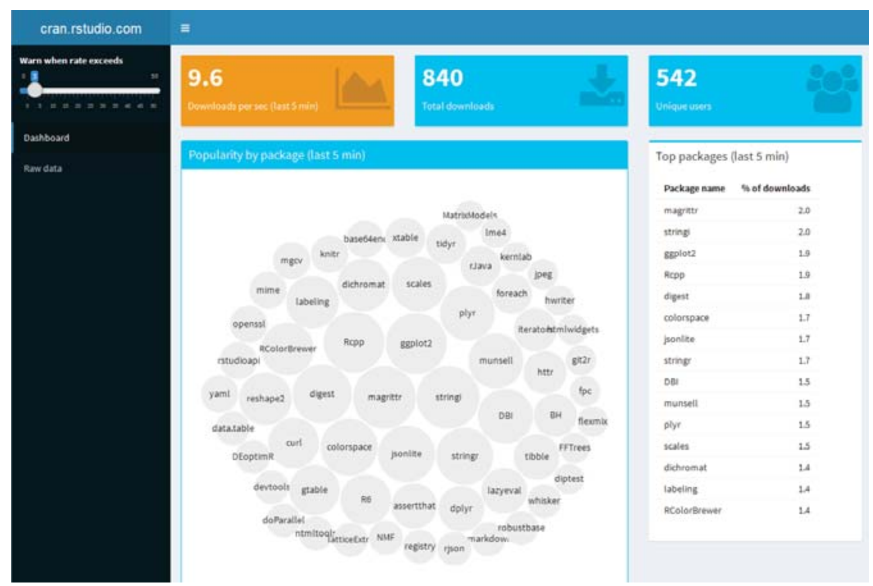


Image source: Data Desk (Paul Velleman)

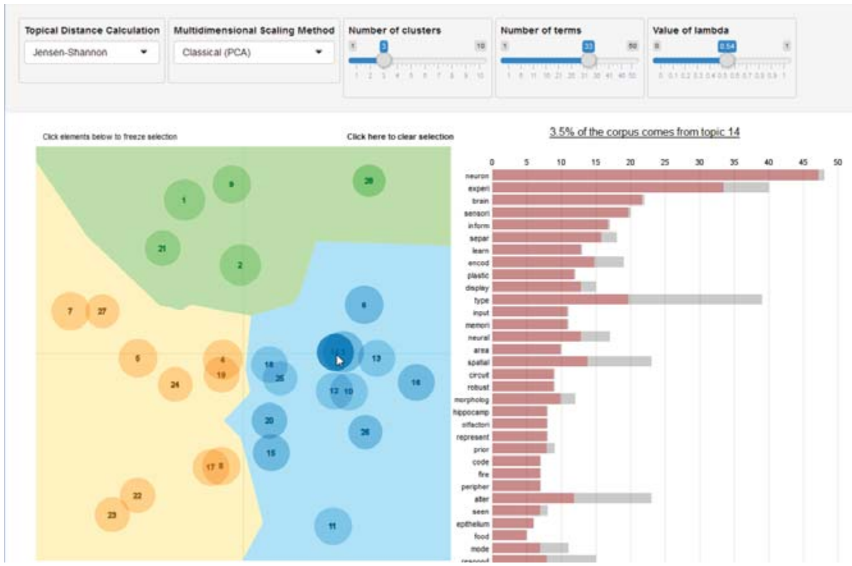
See: http://www.activstats.com/products/mediadx/custom/lessonbook/nyheart.shtml

shiny: dynamic app showing downloads of R packages
https://gallery.shinyapps.io/087-crandash/



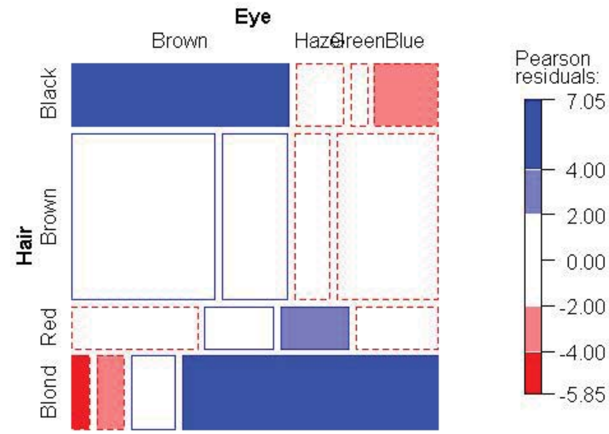
Text mining: Latent distance analysis of a corpus of research papers https://gallery.shinyapps.io/LDAelife/

Uses MDS to find a 2D space from distances among terms



Multivariate frequency data: mosaic plots

Two-way table: [Hair][Eye]



A contingency table can be visualized by tiles whose area ~ cell frequency.

Shading: ~ Pearson residual,

$$d_{ij} = (O_{ij} - E_{ij}) / \sqrt{E_{ij}}$$

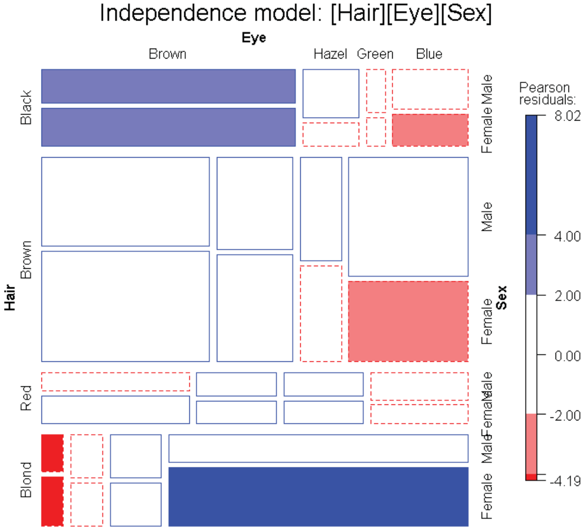
Color:

- blue: $O_{ij} > E_{ij}$; red: $O_{ij} > E_{ij}$

Interp: + association (dark hair, dark eyes), (light hair, light eyes)

N-way tables

Independence model: [Hair][Eye][Sex]



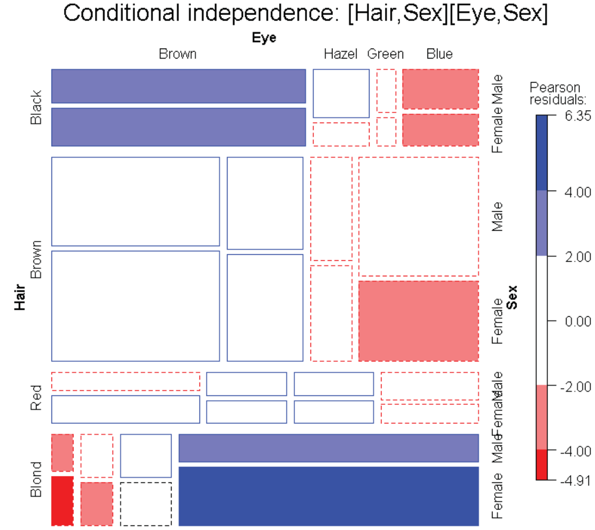
3+ way tables: split each tile ~ conditional proportions of the next variable.

Now, there are several different models that can be fit.

- Mutual independence: [H][E][S] → all vars unassociated
- Residuals: show associations not acct'd for by the model

N-way tables

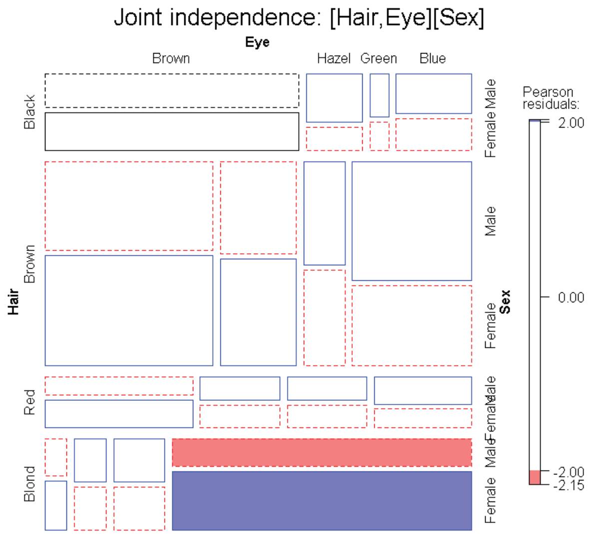
Conditional independence: [Hair,Sex][Eye,Sex]



All models fit to the same table have **same**-sized tiles (O_{ijk}), but **different** residuals.

This model of conditional independence, [HS][ES] → H, E independent *given* Sex.

N-way tables



The model of joint independence, [HE][S] allows Hair, Eye color association, but → [HE] assoc. is independent of Sex.

This model obviously fits much better, except for blue-eyed blonds, where females are more prevalent than the model allows.

Summary

- Goal of statistical analysis: summarization
- Goals of graphical analysis: exposure!
  - Often more useful when enhanced with visual summaries (fitted curve, data ellipse)
- Different graphs for different purposes:
  - Reconnaisance (overview)
  - Exploration (detecting patterns, trends)
  - Model diagnosis (assumptions, outliers)

Summary

- Multivariate data requires novel graphs to display increasing # of variables
  - Enhanced scatterplot matrices
  - Visual thinning: less is often more
  - Low-D views (biplots / MDS)
  - HE plots to visualize multivariate tests
  - Mosaic plots to visualize *n*-way frequency tables.