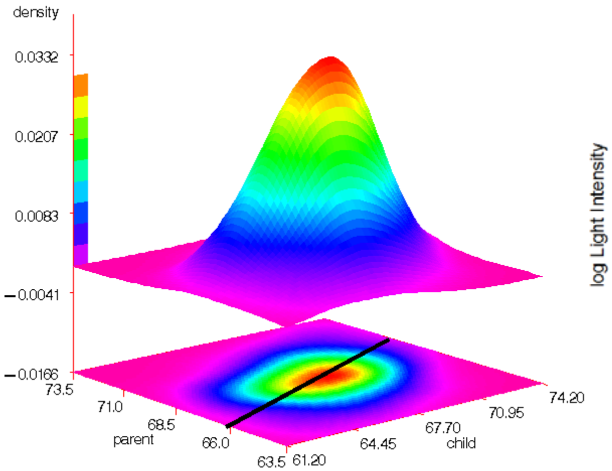
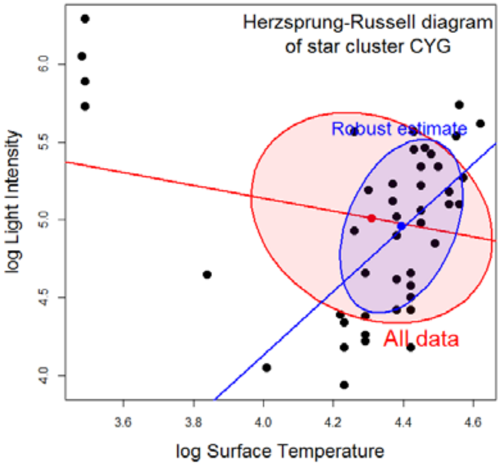
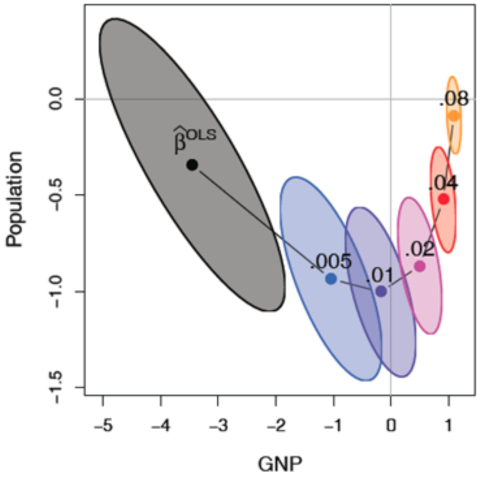






Elliptical Insights: Geometric Travels in Multivariate Data Visualization

Statistical Consulting Service

Michael Friendly, York University

TORA-SABA Data Visualization Workshop

May 5, 2017

Slides: http://datavis.ca/papers/EllipticalInsights-2x2.pdf

YORK
UNIVERSITÉ
UNIVERSITY

Introducing: me & co-conspirators

Statistical graphics and data visualization

moi

John Fox

Georges Monette

David Meyer

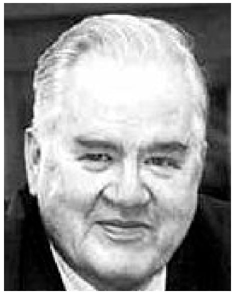
Forrest Young

History of data visualization: Les Chevaliers & inspirators

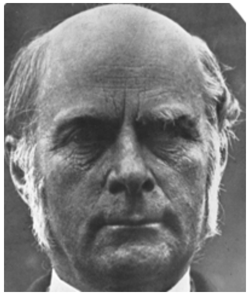
Ian Spence

Howard Wainer

Antoine de Falguerolles



John W. Tukey



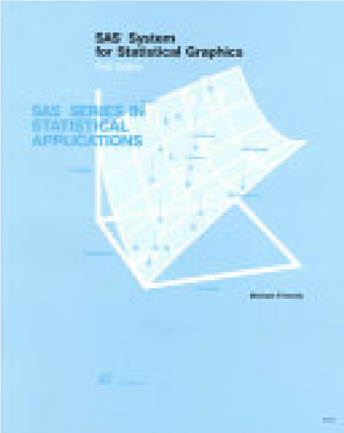
Francis Galton

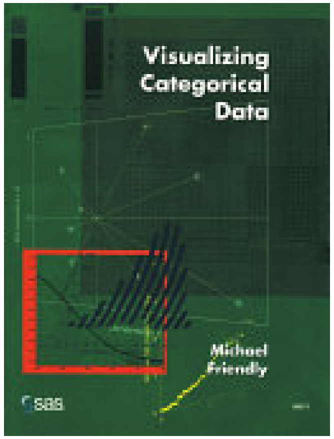
Florence Nightingale

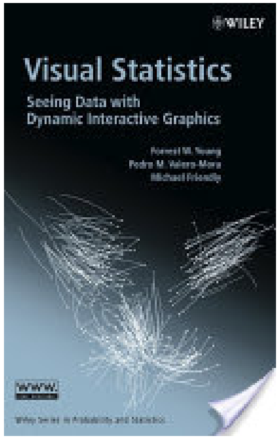
Books: Theory → Practice

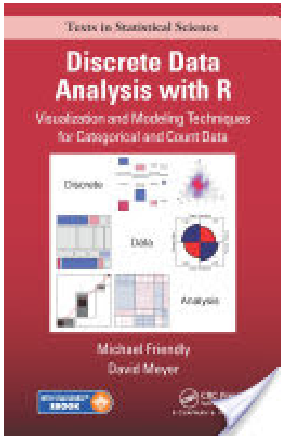
Tukey’s maxim (Tukey, 1959):

The practical power of any statistical method =
Statistical power × Probability anyone will use it






WILEY
Visual Statistics
Seeing Data with Dynamic Interactive Graphics
Forrest W. Young
Pedro M. Valero-Mora
Michael Friendly
WWW
Wiley Series in Probability and Statistics


Texts in Statistical Science
Discrete Data Analysis with R
Visualization and Modeling Techniques for Categorical and Count Data
Discrete
Data
Analysis
Michael Friendly
David Meyer
CRC Press

http://ddar.datavis.ca

Current project: Friendly & Wainer, *The Origin of Graphical Species*,
Harvard Univ. Press, 201?

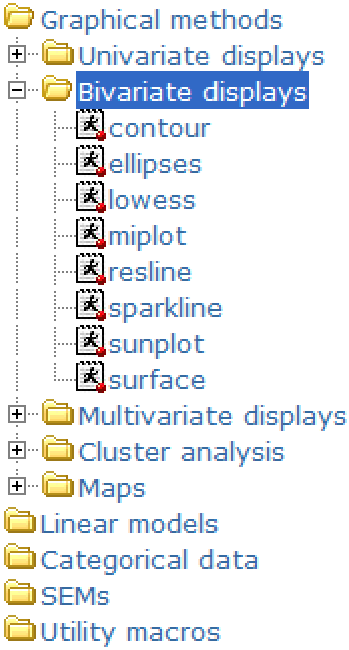
Software: R packages on CRAN

- LM & GLM visualization
  - car : extends graphical methods (John Fox)
  - effects : visualize effects of terms in a complex model (John Fox)
  - genridge : generalized ridge regression / shinkage methods
  - matlib : teaching package for linear algebra and 2D, 3D visualization
  - vcd & vcdExtra: mosaic plots & others for loglinear/logistic regression
- MLMs
  - heplots : HE plots & related methods for MLMs
  - candisc : Analyze/view MLMs in low-D space
  - mvinfluence : measures and new plots for multivariate influence
- Largely data
  - Lahman : Everything you ever wanted to know about baseball statistics
  - HistData: Data sets from the history of statistics & data vis

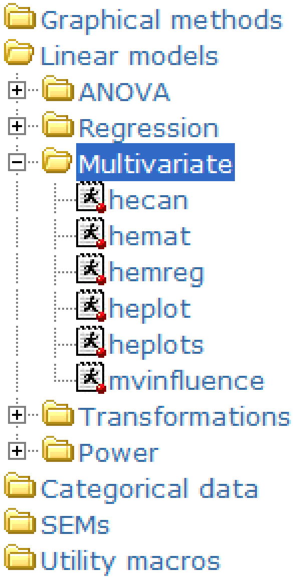
Easy install: source(“http://friendly.apps01.yorku.ca/psy6140/R/install-hebb-pkgs.R”)

Software: SAS macros

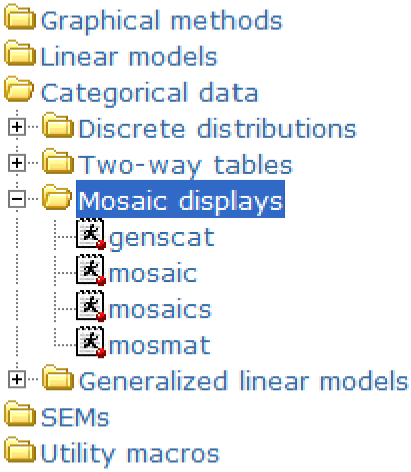
- All use SAS/Graph; some use SAS/IML; some incorporated into SAS
  - Fair warning: I no longer actively maintain or develop these
- Available at:
  - http://datavis.ca/sasmac/ (with documentation)
  - http://friendly.apps01.yorku.ca/psy6140/psy6140.zip (entire collection)



- Graphical methods
  - Univariate displays
  - Bivariate displays
    - contour
    - ellipses
    - lowess
    - miplot
    - resline
    - sparkline
    - sunplot
    - surface
  - Multivariate displays
  - Cluster analysis
  - Maps
- Linear models
- Categorical data
- SEMs
- Utility macros



- Graphical methods
- Linear models
  - ANOVA
  - Regression
  - Multivariate
    - hecan
    - hemat
    - hemreg
    - heplot
    - heplots
    - mvinfluence
  - Transformations
  - Power
- Categorical data
- SEMs
- Utility macros



- Graphical methods
- Linear models
- Categorical data
  - Discrete distributions
  - Two-way tables
  - Mosaic displays
    - genscat
    - mosaic
    - mosaics
    - mosmat
  - Generalized linear models
- SEMs
- Utility macros

Today's topic: Ellipses everywhere

"Once you tune into ellipses, you will begin to see them everywhere ..."
-- James McMullan, https://opinionator.blogs.nytimes.com/2010/09/23/the-frisbee-of-art/

Marcel Duchamp, *Bicycle Wheel*, 1913

"In 1913 I had the happy idea to fasten a bicycle wheel to a kitchen stool and watch it turn." (apropos of the "readymade" art movement)

Rotation transforms the circle to an ellipse to a line and back again:

- Hey, a line is just a degenerate ellipse!
- In 3D, it sweeps out a special ellipsoid, called a "sphere"

Animation: https://www.youtube.com/watch?v=L7t3sUTCtZQ

Today’s topic: Ellipses everywhere

“The ellipse is the Frisbee of art, the circle freed from its flatness that sails out into imagined space tilting this way and that and ending up on the top of the soup bowl and silver cup in Jean-Baptiste Chardin’s still life…” -- James McMullan

More to the point:
The ellipse is the happy intersection of statistics, data vis & geometry

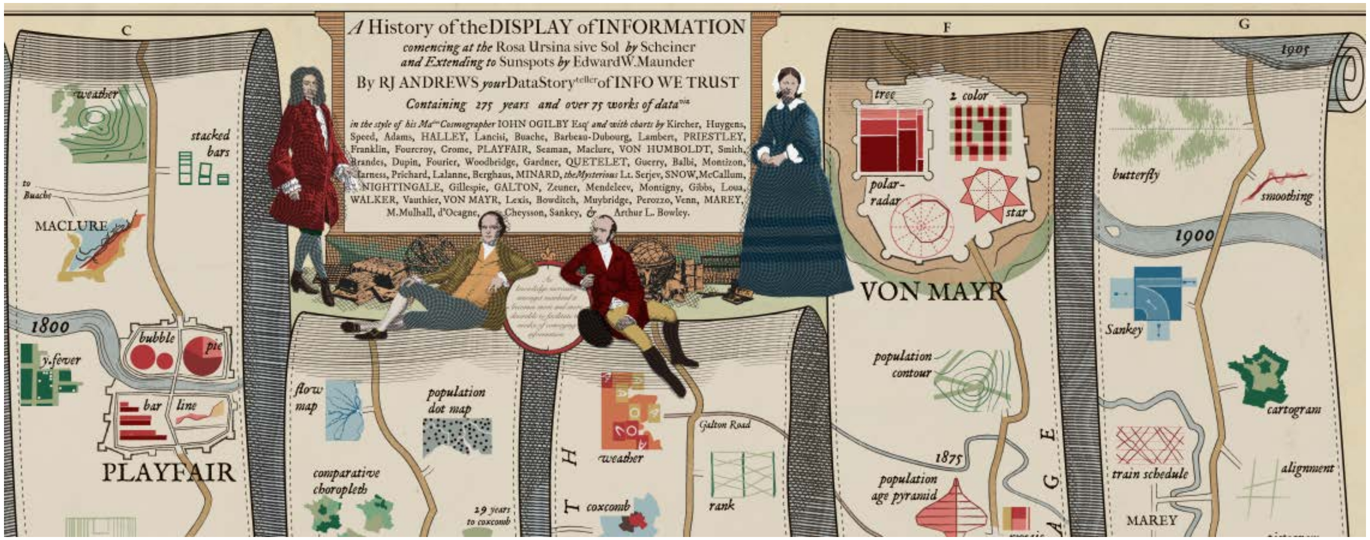
“Whatever relates to extent and quantity may be represented by geometrical figures.

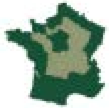
Statistical projections which speak to the senses without fatiguing the mind, posses the advantage of fixing attention on a great number of important facts”

-- Alexander von Humboldt (1811)

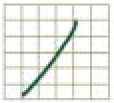
Outline

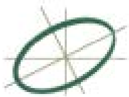
- Introduction: A whirlwind tour of History of DataVis
- Data ellipsoids
- The HE plot framework
- Understanding ridge regression & shrinkage methods


C
F
G
A History of the DISPLAY of INFORMATION
comencing at the Rosa Ursina sive Sol by Scheiner
and Extending to Sunspots by Edward W. Maunder
By RJ ANDREWS your DataStoryteller of INFO WE TRUST
Containing 275 years and over 75 works of data
in the style of his Maties Cosmographer IOHN OGILBY Esq and with charts by Kircher, Huygens, Speed, Adams, HALLEY, Lancisi, Buache, Barbeau-Dubourg, Lambert, PRIESTLEY, Franklin, Fourcroy, Crome, PLAYFAIR, Seaman, Maclure, VON HUMBOLDT, Smith, Brandes, Dupin, Fourier, Woodbridge, Gardner, QUETELET, Guerry, Balbi, Montizon, Harness, Prichard, Lalanne, Berghaus, MINARD, the Mysterious Lt. Serjev, SNOW, McCallum, NIGHTINGALE, Gillespie, GALTON, Zeuner, Mendeleev, Montigny, Gibbs, Loua, WALKER, Vauthier, VON MAYR, Lexis, Bowditch, Muybridge, Perozzo, Venn, MAREY, M.Mulhall, d'Ocagne, Cheysson, Sankey, & Arthur L. Bowley.
weather
stacked bars
to Buache
MACLURE
1800
y.fever
bubble
pie
bar
line
PLAYFAIR
flow map
population dot map
comparative choropleth
29 years to coxcomb
weather
coxcomb
Galton Road
rank
tree
2 color
polar-radar
star
VON MAYR
population contour
1875
population age pyramid
1905
butterfly
smoothing
1900
Sankey
cartogram
train schedule
alignment
MAREY



- Prelude: the birth of data
- Moral statistics: the birth of modern social science
- William Farr on cholera
- JFW Herschel & the 1st scatterplot
- Galton: the birth of modern statistical methods





Introduction: A whirlwind tour of the history of Data Visualization

Images: RJ Andrews, http://infowetrust.com/history/

Prelude to data visualization: The birth of data

- Mrs. Isabella Beeton's (~ 1860) recipe for rabbit stew: “First catch a rabbit”
  - Data vis: First get some data; now make sense of it.
- When was the idea of “data” invented?
- A longer story, but I’ll start in the early 1800s
- Social problems, demanding policy solutions:
  - France: Upheaval following Napoleon’s defeat: migration, crime, suicide, prostitution, …
  - England: Outbreaks of cholera, poverty, “poor laws”, debtor prisons, …
- Suddenly, an avalanche of data crying for explanation!
  - But where was data vis?

France: Guerry, *La Statistique Morale*

- In France, widespread, national data collection on social issues began ~ 1810—1825
  - literacy: % of army conscripts who could read and write
  - crime: Ministry of Justice launches the *Compte Générale*
    - every criminal charge recorded, with all details: age, sex, occupation, date, court outcome
    - mandated quarterly reports to Paris
- Suddenly, one could attempt to answer important questions using data rather than philosophy
  - Is greater literacy related to less crime?
  - Do more priests lead to less crime, suicide or prostitution?
- Moral statistics: the beginning of modern social science
  - Social data could lead to “social laws” à la “physical laws”

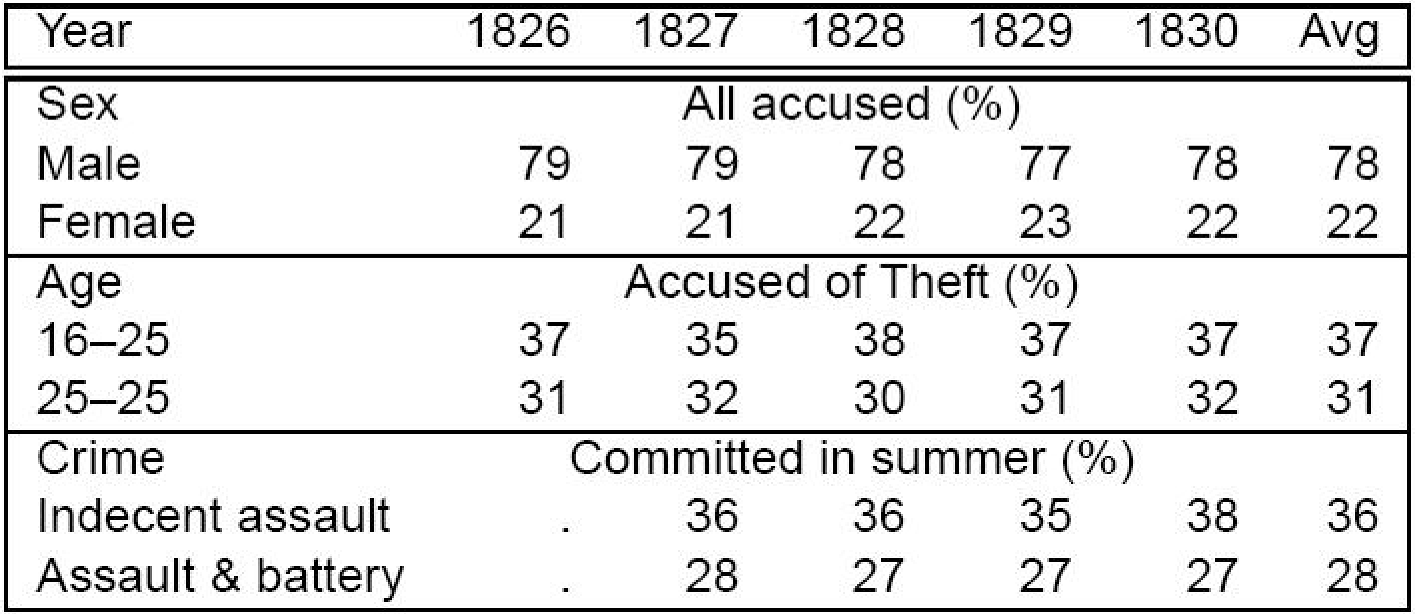
See: Friendly (2007) A.-M. Guerry's *Moral Statistics of France*: Challenges for Multivariable Spatial Analysis *Statistical Science, 22*, 368-399

The discovery of “social facts”

Stability and Variation

Guerry’s results were both compelling and startling:

- Rates of crime and suicide remained remarkably invariant over time, yet varied sytematically by region, sex of accused, type of crime, etc.
- In any given French city or department, almost the same number committed suicide, stole, gave birth out of wedlock, etc.

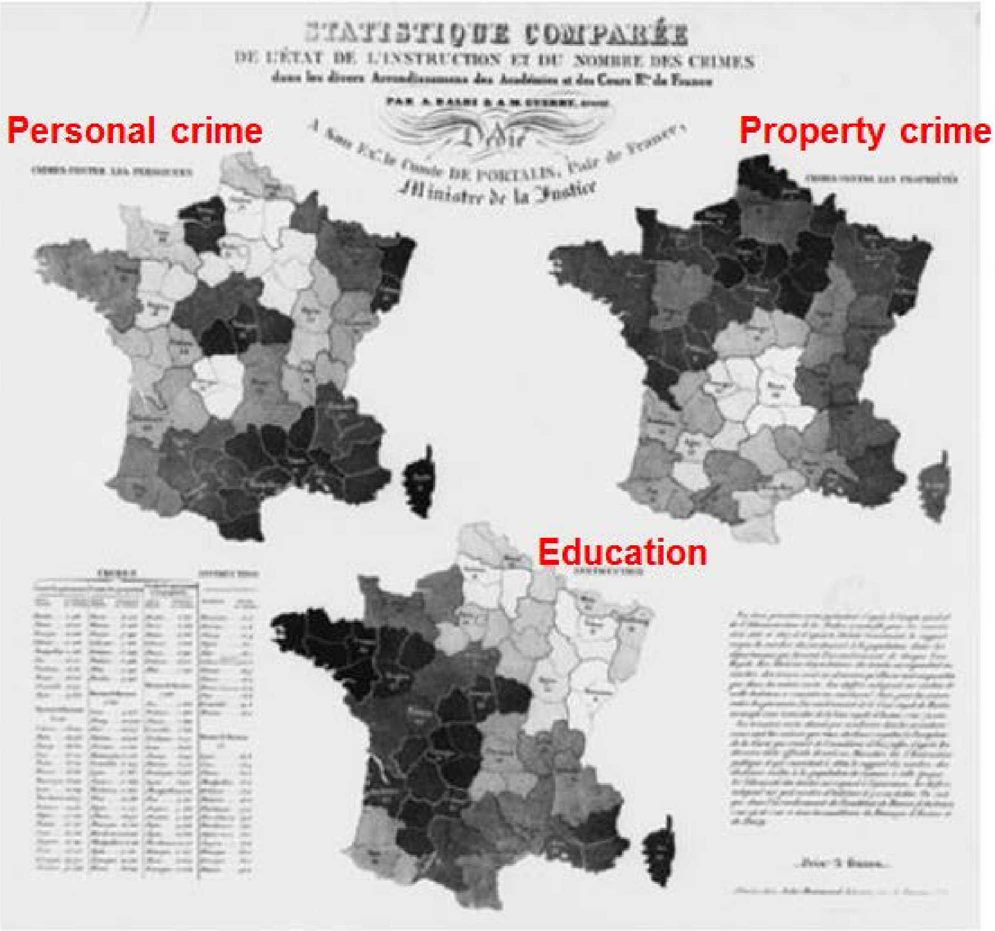


Year	1826	1827	1828	1829	1830	Avg
Sex	All accused (%)					
Male	79	79	78	77	78	78
Female	21	21	22	23	22	22
Age	Accused of Theft (%)					
16–25	37	35	38	37	37	37
25–25	31	32	30	31	32	31
Crime	Committed in summer (%)					
Indecent assault	.	36	36	35	38	36
Assault & battery	.	28	27	27	27	28

“We are forced to conclude that the facts of the moral order are subject, like those of the physical order to invariable laws.” (Guerry, 1833, p14)

Guerry & Balbi (1829): Comparative statistics of crime & education

- First shaded thematic maps of crime data
- First comparative maps of social data
- ↦ crime against persons seemed inversely related to crime against property!
- Instruction: ↦ *France obscure* and *France éclairée* (Dupin, 1826)
- North of France highest in education, but also in property crime!


STATISTIQUE COMPARÉE
DE L'ÉTAT DE L'INSTRUCTION ET DU NOMBRE DES CRIMES
dans les divers Arrondissemens des Académies et des Cours Rles de France
PAR A. BALBI & A. M. GUERRY
A Son Exce le Comte DE PORTALIS, Pair de France, Ministre de la Justice
Personal crime
Property crime
Education

What is missing: (a) idea of plotting Y vs. X; (b) measures of co-relation

Guerry (1864): General causes of crime

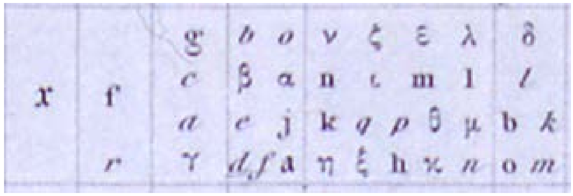
Plate XVII: Guerry's magnum opus

Goal:

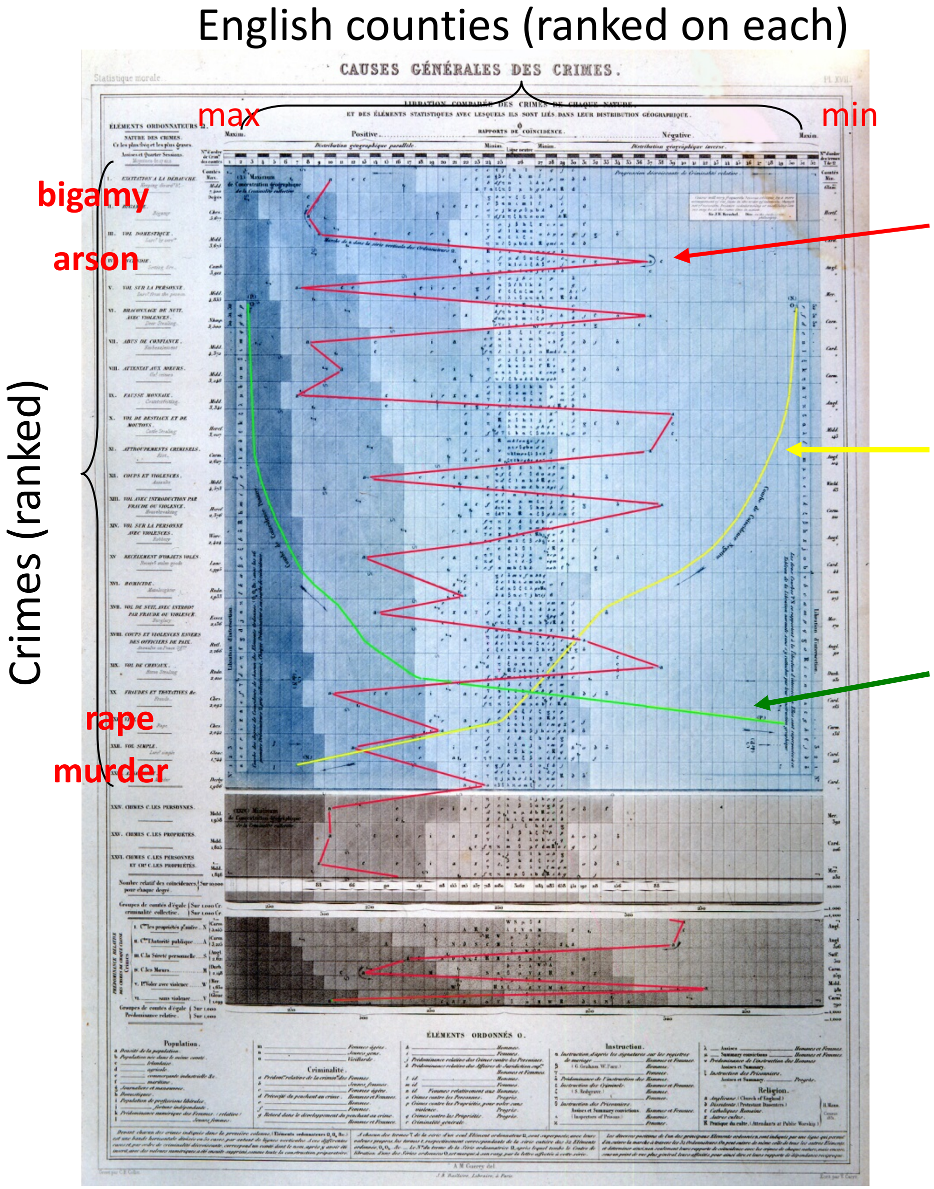
- Show multivariate factors associated with distribution of crimes of various type
- Before invention of correlation

Entries: Codes for factors

- Pop: (% Irish, domestics, …)
- Criminality: (male, young, …)
- Religion (Anglicans, dissenters, …)


g b o v ζ ε λ δ
x f c β a n ι m l l
a e j k q p ϑ μ b k
r γ d f a η ξ h ϰ n o m

Guerry, A.-M. (1864). *Statistique morale de l'Angleterre comparée avec la statistique morale de la France*



High pop. density

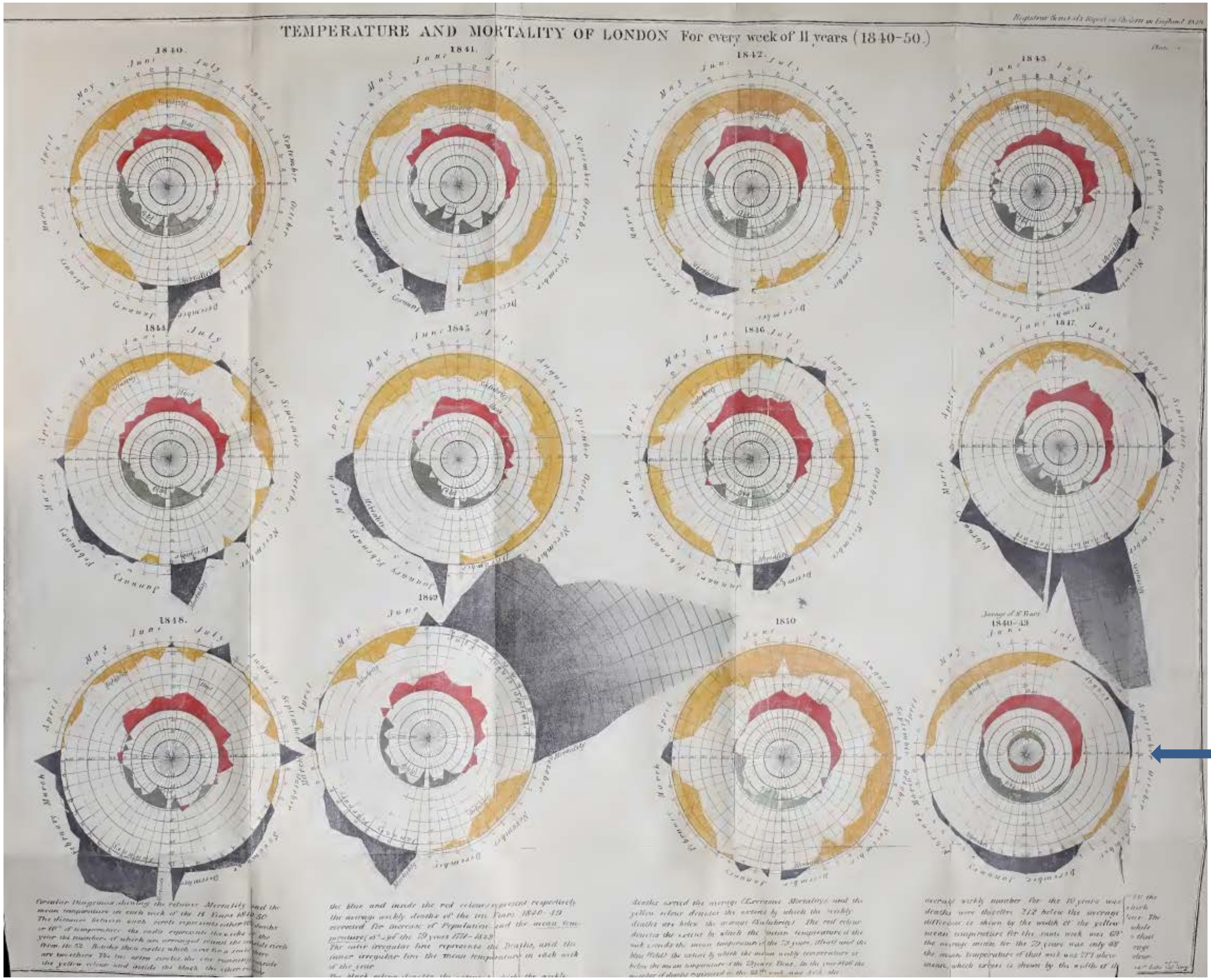
Curve of neg. association

Curve of pos. association

England: William Farr on cholera

- General Register Office (GRO), 1836
  - Record every birth, death in England & Wales
  - A universal data base of the entire population
- William Farr [1807—1883]: 1st official UK statistician
  - Institutes recording of causes of mortality & circumstances
  - Idea of identifying "risk factors" by tabulating deaths in relation to potential causes (poverty, environmental, …)
- Cholera outbreaks
  - India 1820s → UK 1831—1832; by 1837, greatest worldwide pandemic of 19th C; returns in 1848, 1852
  - Miasmatic hypothesis: bad air ("the big stink")
  - Test: Mortality ~ temperature, season, elevation, …

Radial diagram of temperature and mortality in London, by week, for 1840—1850. From: Farr (1852), *Report on the mortality from cholera …* Plate IV



Seasonal effects on mortality?

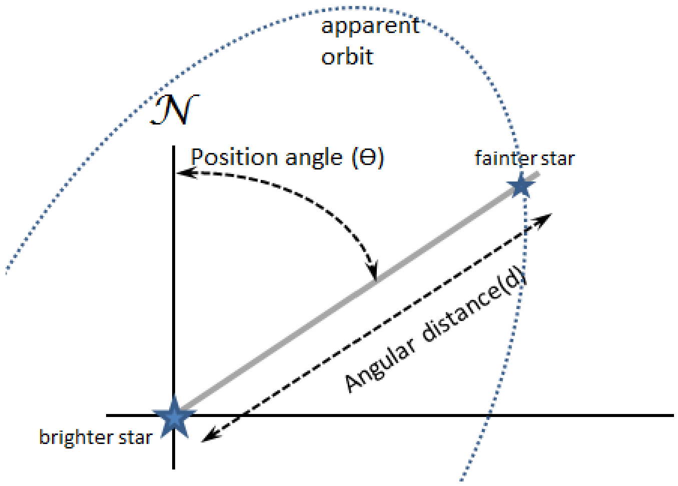
Relation to temperature?

Still no idea of Y vs. X

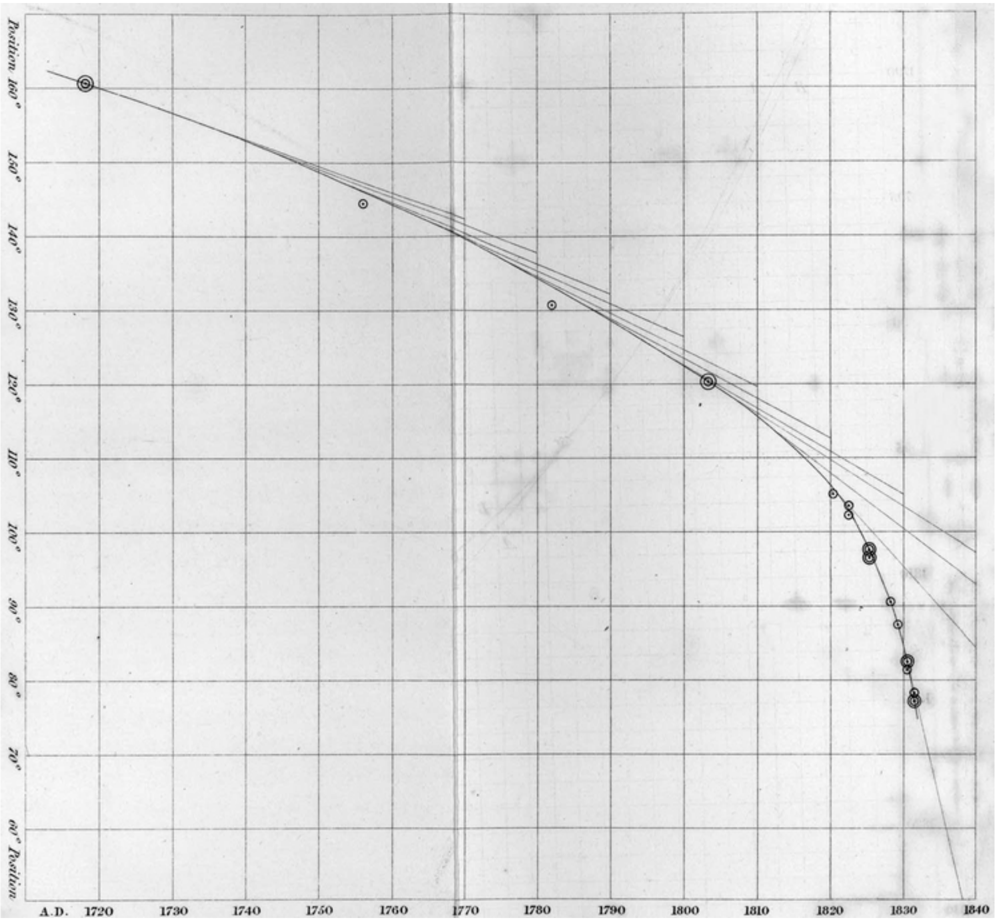
Avg. over 11 years

JFW Herschel: The 1st scatterplot

- Problem: determine the orbits of twin stars from observations of measured angles and apparent distances
  - Observations (*n*=14) only recorded over long periods of time
  - Theory: elliptical orbit → 7 equations in 7 unknowns, easy since Gauss
  - But: data subject to "extravagant errors"



The process by which I propose to accomplish this is one essentially graphical; by which term I understand not a mere substitution of geometrical construction and measurement for numerical calculation, but one which has for its object to perform that which no system of calculation can possibly do, by bringing in the aid of the eye and hand to guide the judgment, in a case where judgment only, and not calculation, can be of any avail. (Herschel, 1833, p. 178)

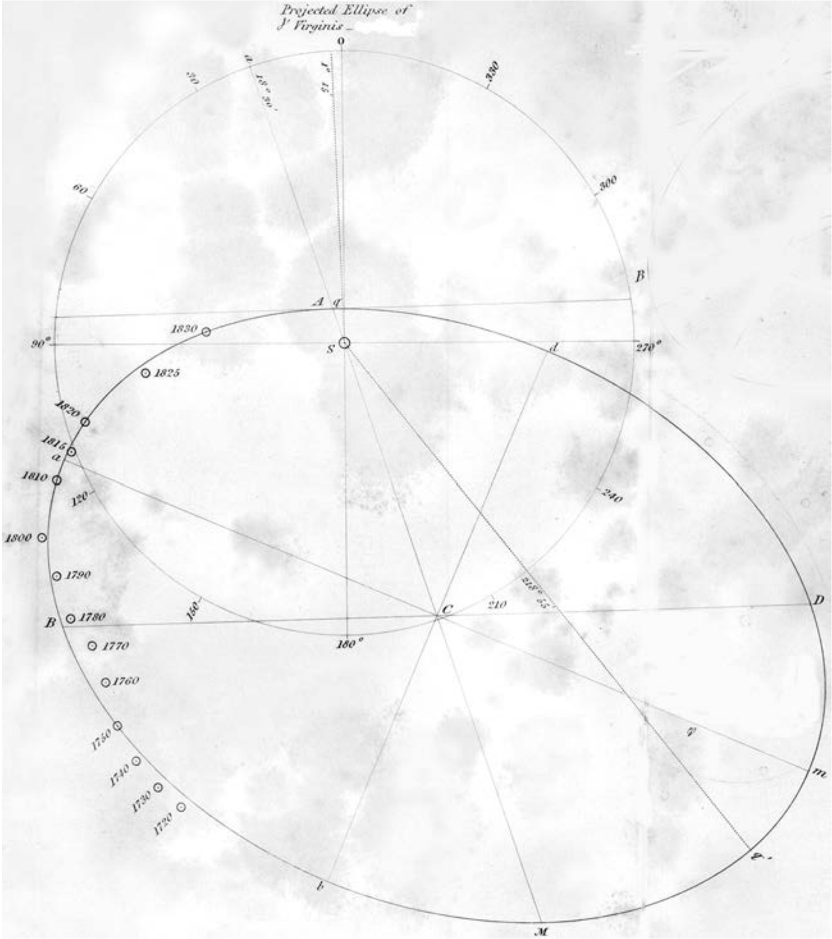


Herschel's (1833) graphical method, applied to the data on the double star γ *Virginis.* Image from: Hankins (2006), Fig. 2

Hershel plots data on position angle (Y) over time (X)

adds an eye-fit smoothed curve that respects the relative error in the 14 observations

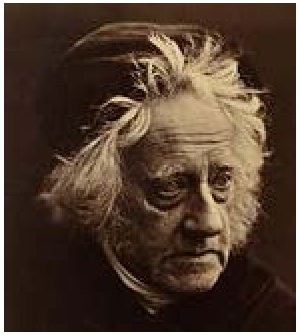
uses the fitted curve to calculate angular velocity --- the slopes of tangents to the curve



Herschel’s geometric construction of the apparent elliptical orbit of γ *Virginis* from the calculations based on his smoothed scatterplot. Image from: Hankins (2006), Fig. 3.

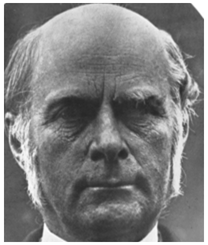
Thus was born:

- The idea that plotting Y vs. X could be used for something more
- Smoothing is often crucial to see a pattern or calculate a trend



Galton: Heredity → Regression

- Francis Galton, in work on heritability of traits, introduces the idea of “reversion” (later: “regression”) toward the mean



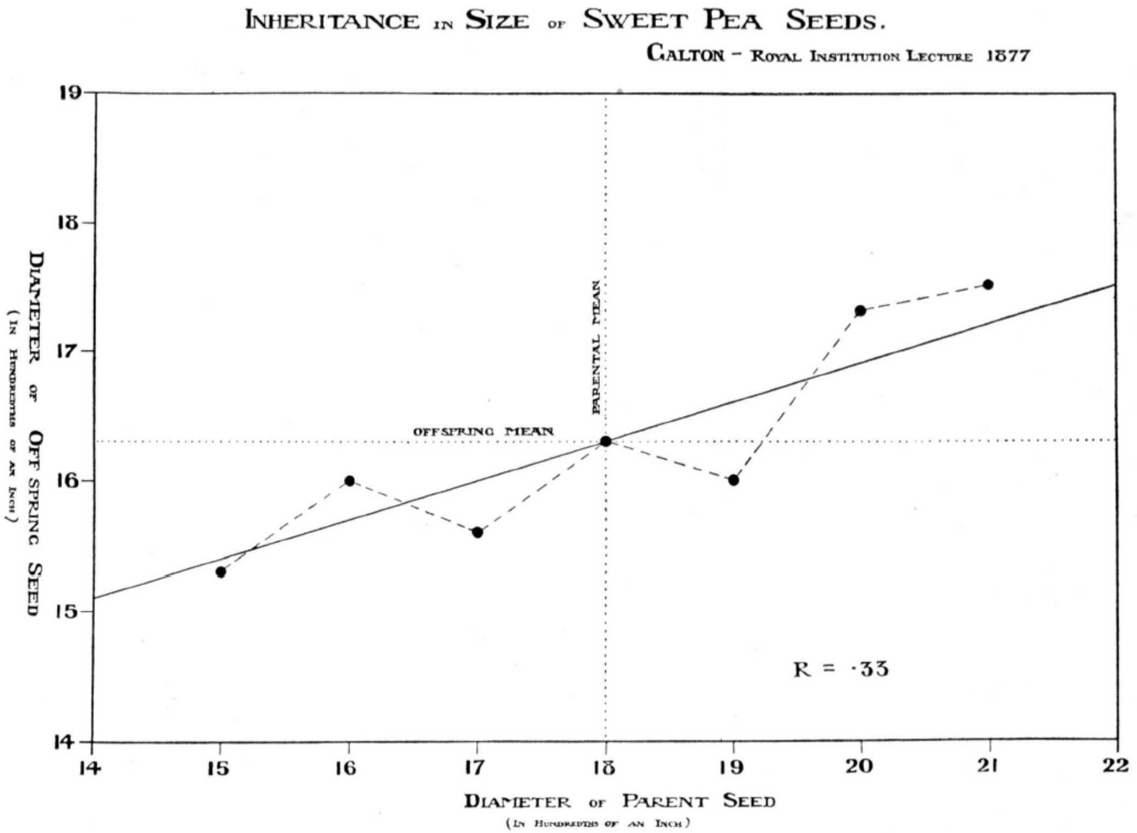

Inheritance in Size of Sweet Pea Seeds.
Galton – Royal Institution Lecture 1877
Diameter of Off spring Seed
(in Hundredths of an Inch)
Diameter of Parent Seed
(in Hundredths of an Inch)
Parental Mean
Offspring Mean
R = ·33
19
18
17
16
15
14
14
15
16
17
18
19
20
21
22

Fig. 1. The first “Regression Line.”

An early crowd-sourced experiment:

- packets of 10 seeds of 7 given sizes sent to 7 friends
- “Please grow these & return the offspring”

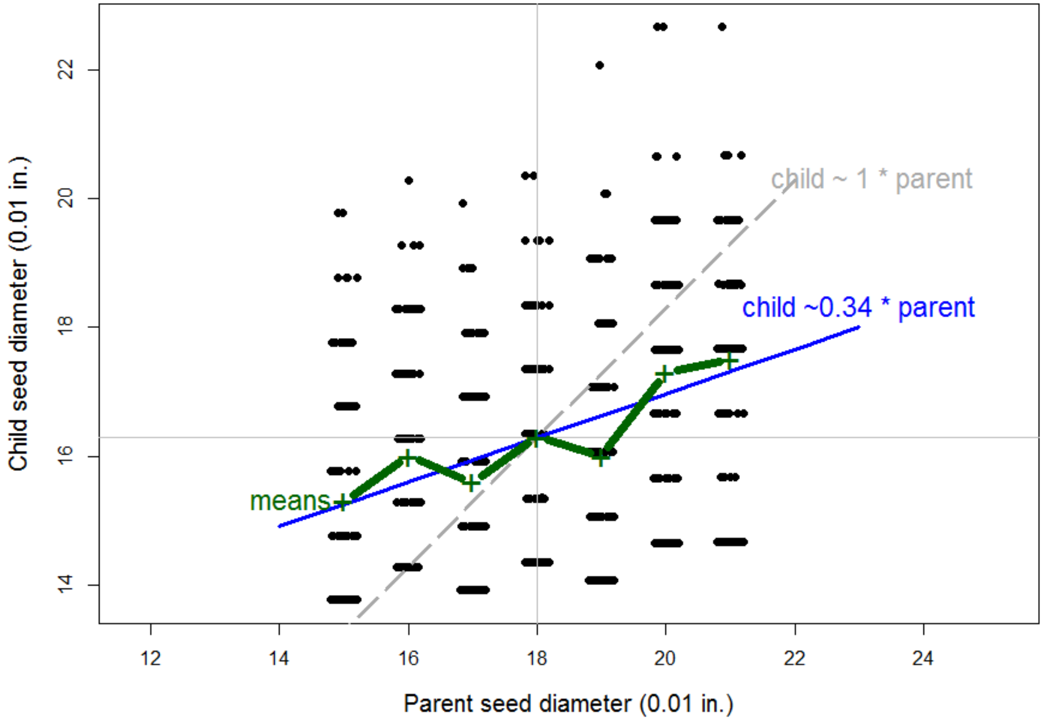
Graph:

- plot the means,
- draw a line,
- calculate the slope (“R”)
- → a theoretical conclusion!

Image: K. Pearson, *The Life, Letters and Labours of Francis Galton*, v. 3A, Ch 14, Fig. 1

Galton: Heredity → Regression

- Galton's argument made explicit:
  - slope < 1 → regression toward the mean



"... offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small ..." (Galton 1886)

Galton’s visual discovery:

Bivariate normal correlation surface (1886)

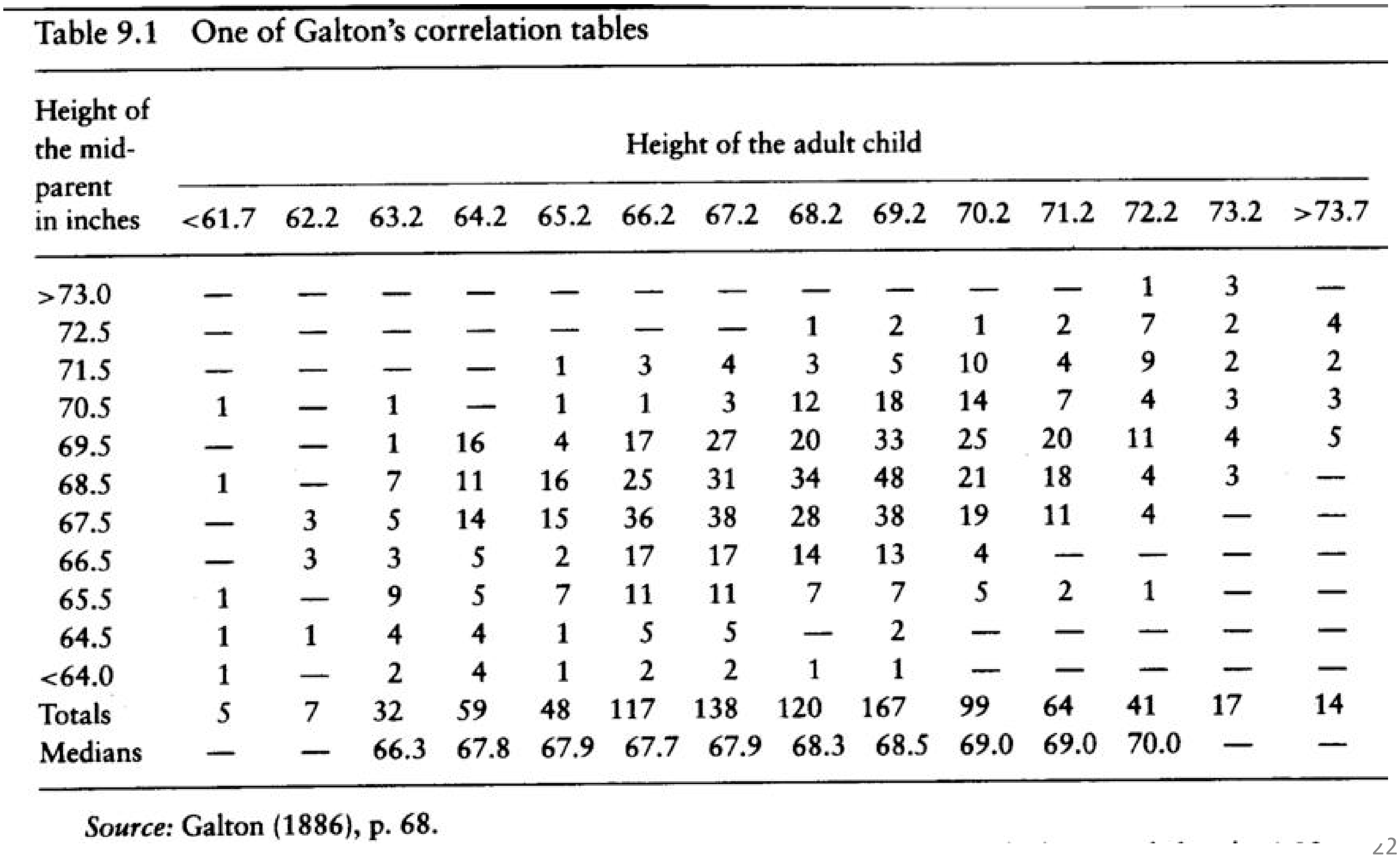


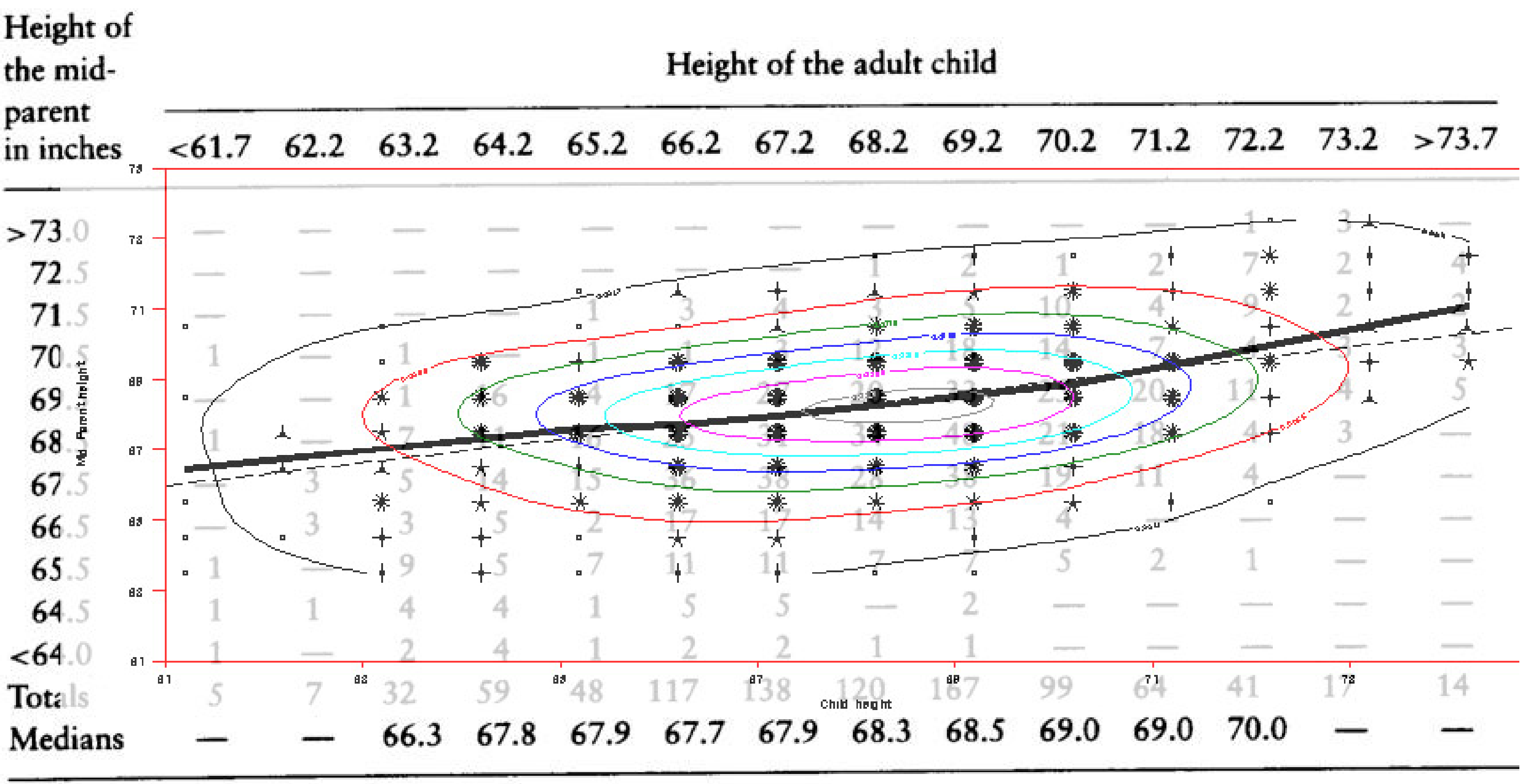
Table 9.1 One of Galton’s correlation tables

Height of the mid-parent in inches	Height of the adult child													
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—

Source: Galton (1886), p. 68.

Visual smoothing → Insight

Table 9.1 One of Galton's correlation tables

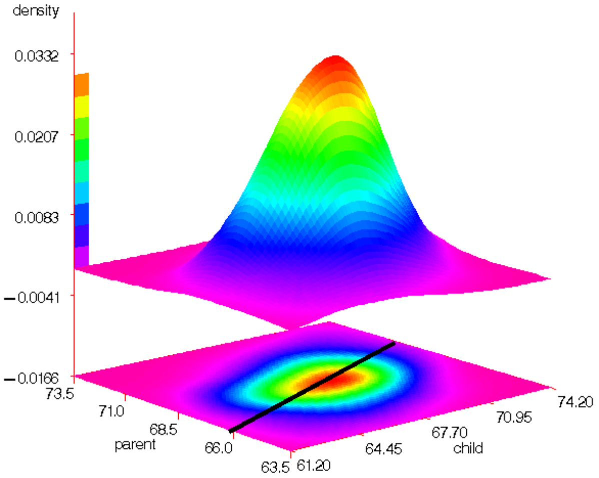

Mid. Parent height
Child height

Height of the mid-parent in inches	Height of the adult child													
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—

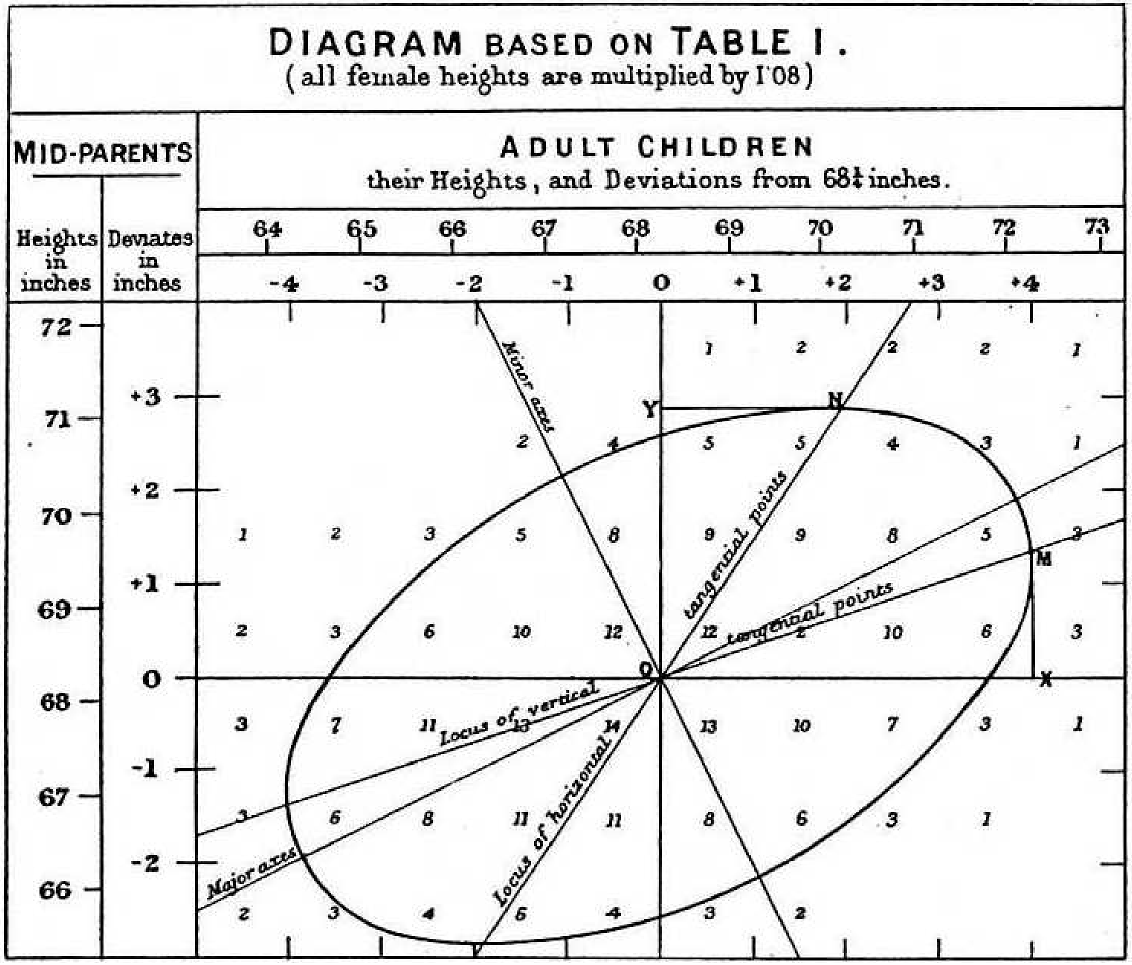
Source: Galton (1886), p. 68.

Visual insight → Theory (the OMG! moment)

- Level curves are **ellipses**
- Regression lines are loci of conjugate **tangents**



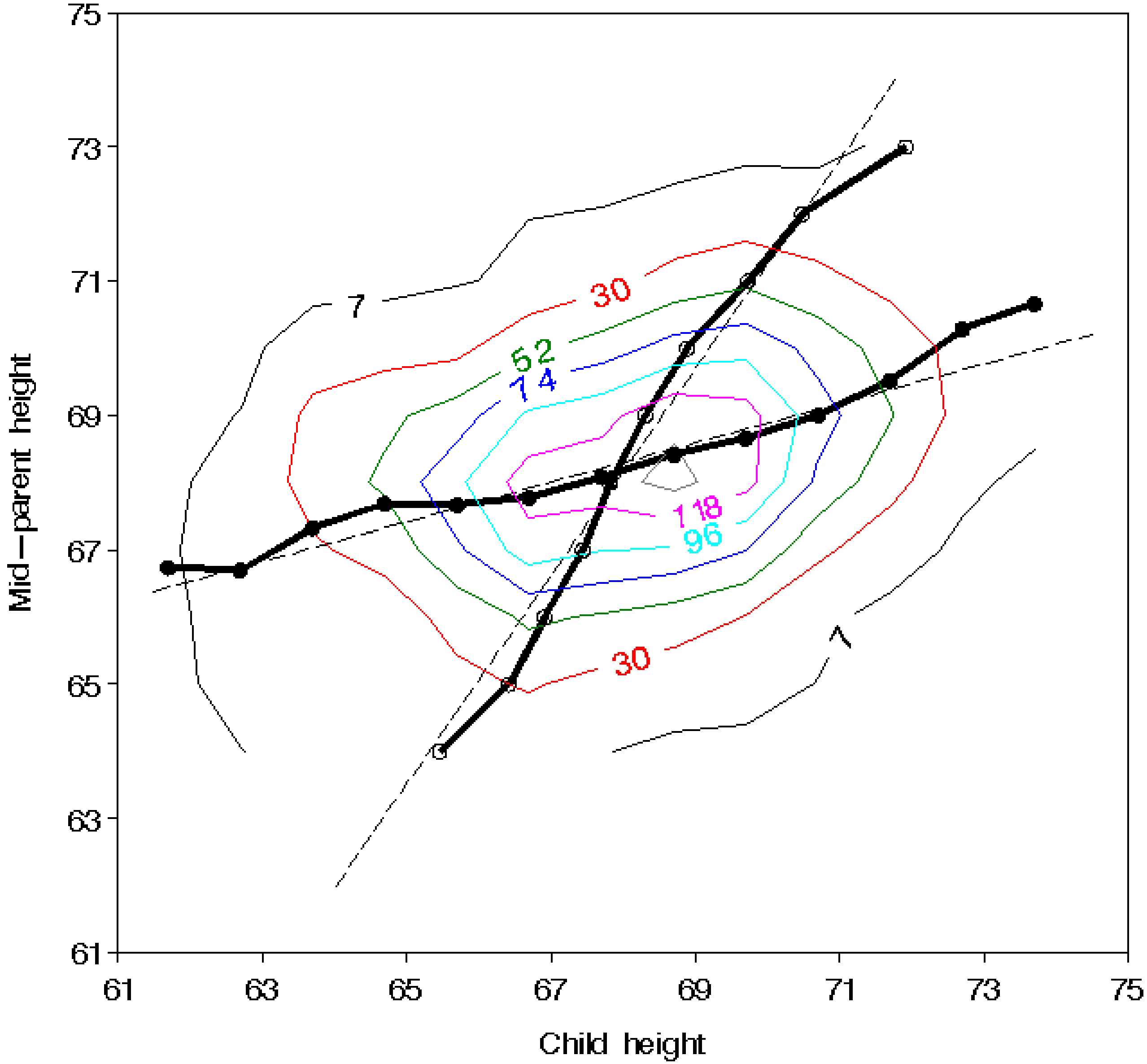
… that Galton should have evolved all this … is to my mind one of the most note-worthy scientific discoveries arising from analysis of pure observation (Pearson 1920, p37)



Galton (1886, Pl X): Smoothed contours of heights of parents and children

How did Galton reach this conclusion?

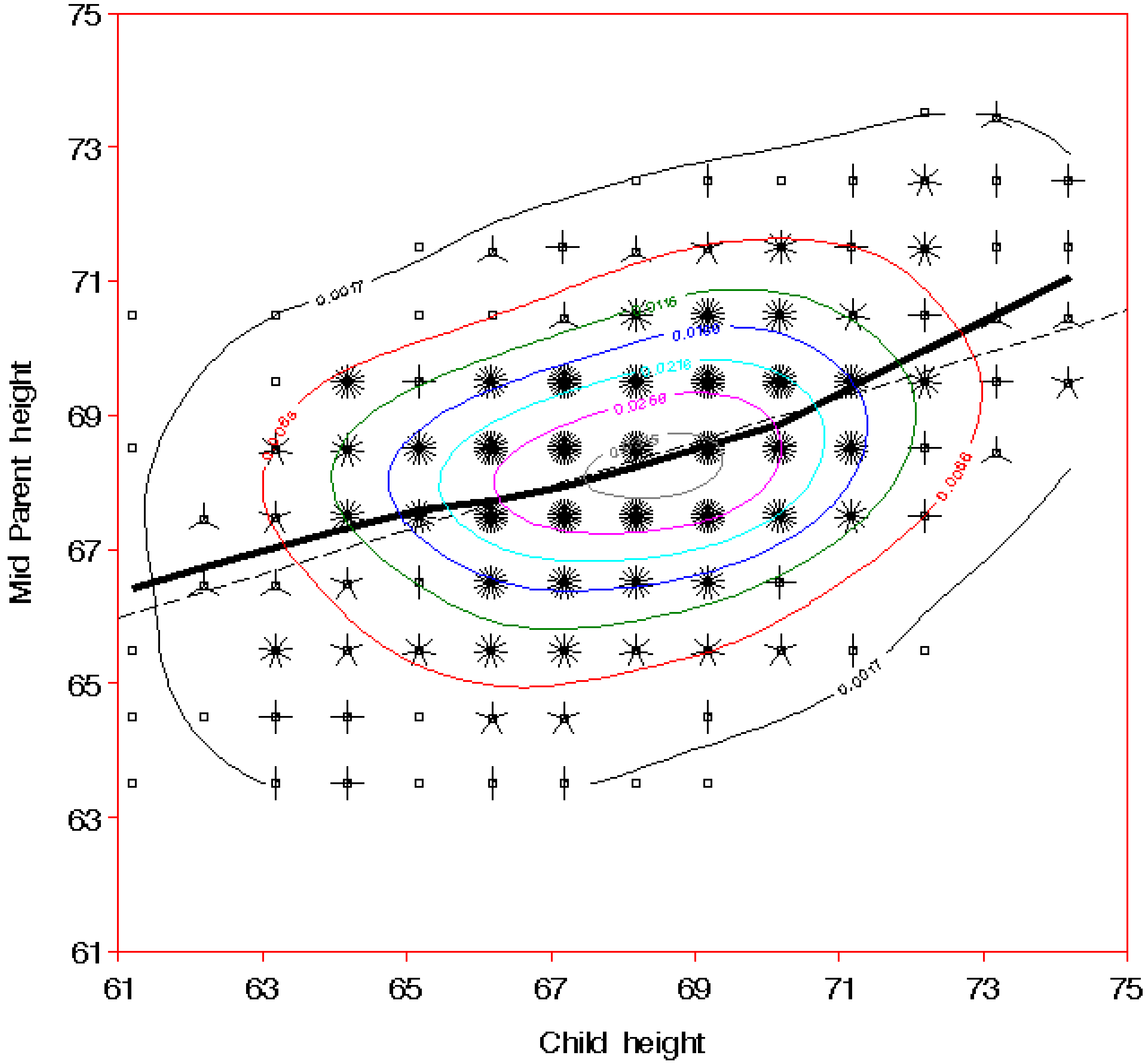
Literal application of Galton's smoothing method only vaguely suggests "concentric ellipses" or lines of means as conjugate axes



How did Galton reach this conclusion?

Modern smoothing methods (kernel density estimate) suggests that Galton:

- smoothed by ‘eye & brain’
- was probably looking for ellipses



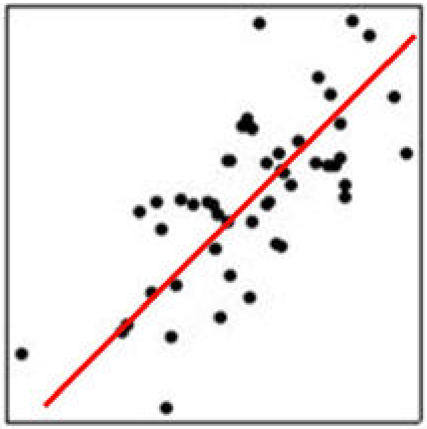
- The LM family & friends
- Geometrical ellipsoids
- The data ellipse

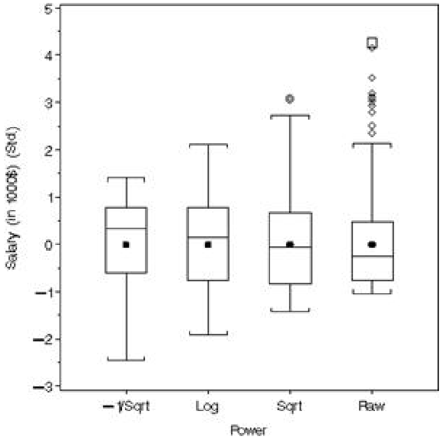
Data Ellipsoids

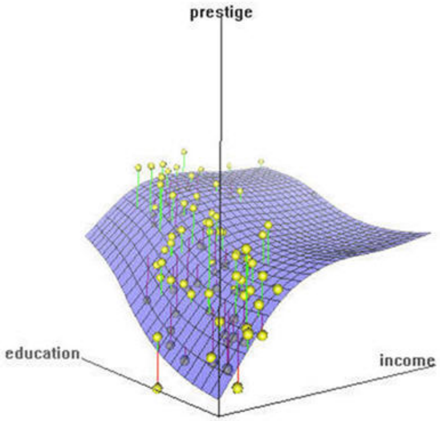
The LM family & friends

Models, graphical methods & opportunities

# of response variables	Classical linear models	Generalized linear models
1	LM family: $E(\mathbf{y})=\mathbf{X}\beta$, $V(y\vert\mathbf{X})=\sigma^2\mathbf{I}$ ANOVA, regression, … Many graphical methods: effect plots, spread-leverage, influence, …	GLM: $E(\mathbf{y})=g^{-1}(\mathbf{X}\beta)$, $V=V[g^{-1}(\mathbf{X}\beta)]$ poisson, logistic, loglinear, … Some graphical methods: mosaic plots, 4fold plots, diagnostic plots, …
2+	MLM: $E(\mathbf{Y})=\mathbf{X}\beta$, $V(\mathbf{Y}\vert\mathbf{X})=\mathbf{I}\otimes\mathbf{\Sigma}$ MANOVA, MMReg, … Graphical methods: ???	MGLM: ??? Graphical methods: ???



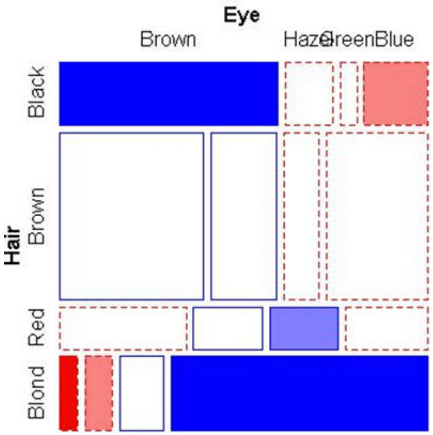


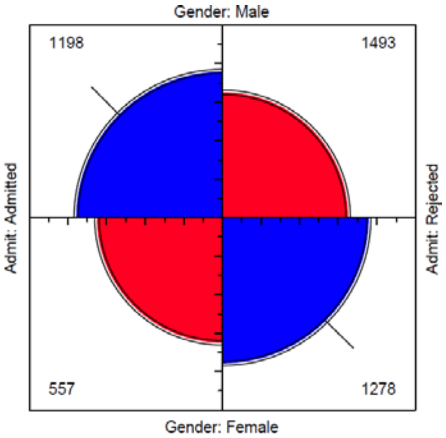


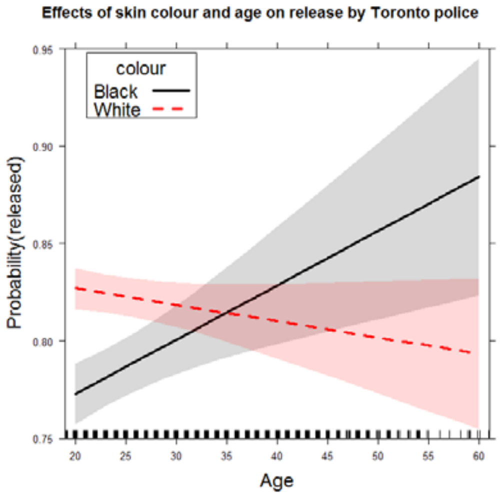
The LM family & friends

Models, graphical methods & opportunities

# of response variables	Classical linear models	Generalized linear models
1	LM family: $E(\mathbf{y})=\mathbf{X}\beta$, $V(y \mid \mathbf{X})=\sigma^2\mathbf{I}$ ANOVA, regression, … Many graphical methods: effect plots, spread-leverage, influence, …	GLM: $E(\mathbf{y})=g^{-1}(\mathbf{X}\beta)$, $V=V[g^{-1}(\mathbf{X}\beta)]$ poisson, logistic, loglinear, … Some graphical methods: mosaic plots, 4fold plots, diagnostic plots, …
2+	MLM: $E(\mathbf{Y})=\mathbf{X}\beta$, $V(\mathbf{Y} \mid \mathbf{X})=\mathbf{I}\otimes\boldsymbol{\Sigma}$ MANOVA, MMReg, … Graphical methods: ???	MGLM: ??? Graphical methods: ???

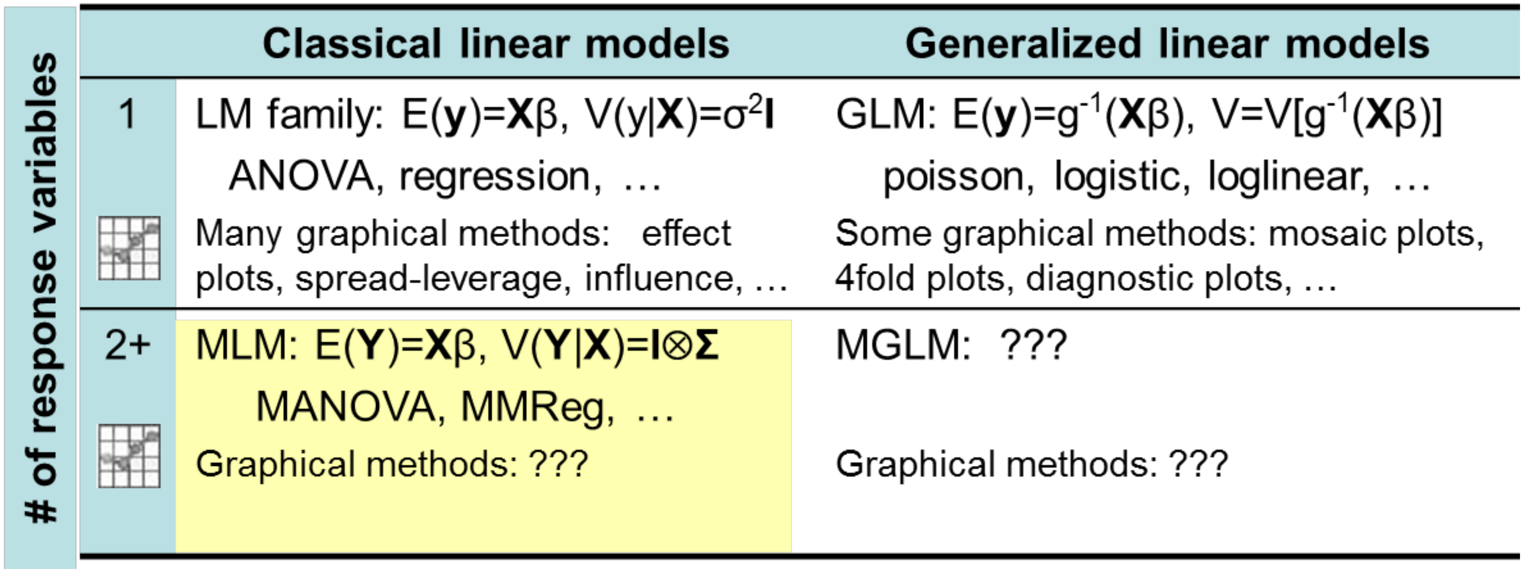






The LM family & friends

Models, graphical methods & opportunities

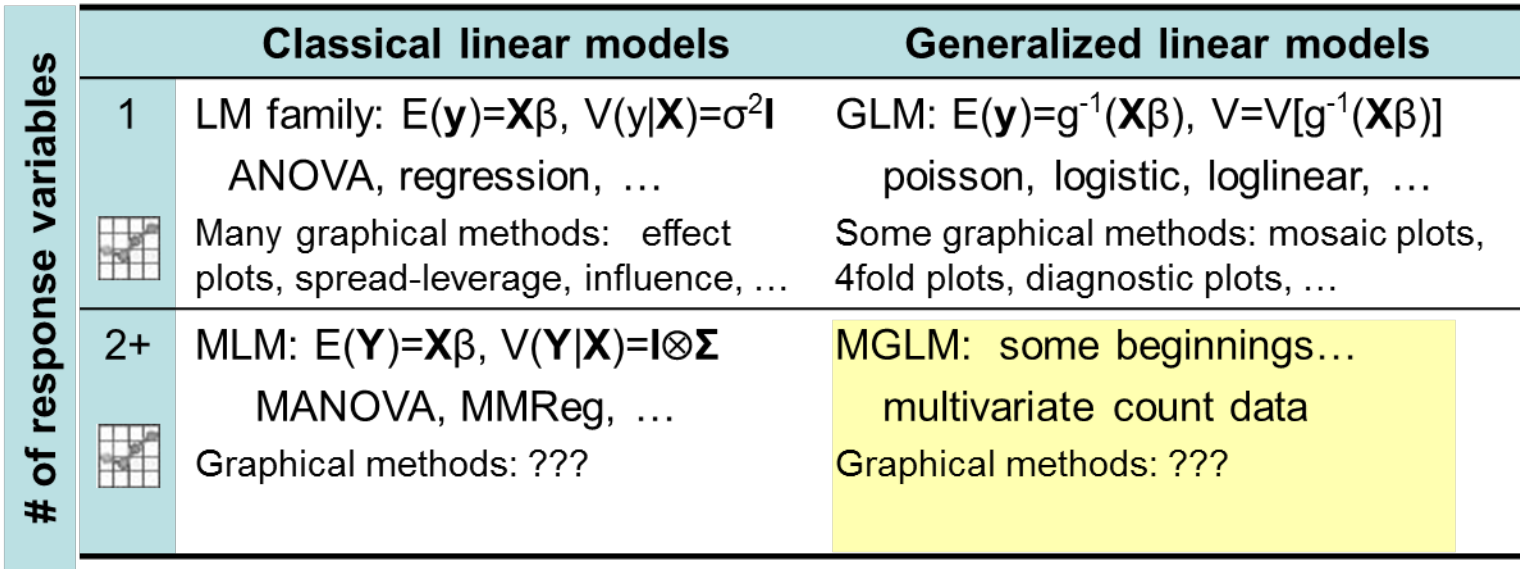


# of response variables		Classical linear models	Generalized linear models
	1	LM family: $E(\mathbf{y})=\mathbf{X}\beta$, $V(y\|\mathbf{X})=\sigma^2\mathbf{I}$ ANOVA, regression, … Many graphical methods: effect plots, spread-leverage, influence, …	GLM: $E(\mathbf{y})=g^{-1}(\mathbf{X}\beta)$, $V=V[g^{-1}(\mathbf{X}\beta)]$ poisson, logistic, loglinear, … Some graphical methods: mosaic plots, 4fold plots, diagnostic plots, …
	2+	MLM: $E(\mathbf{Y})=\mathbf{X}\beta$, $V(\mathbf{Y}\|\mathbf{X})=\mathbf{I}\otimes\mathbf{\Sigma}$ MANOVA, MMReg, … Graphical methods: ???	MGLM: ??? Graphical methods: ???

Today: HE plots & related methods

The LM family & friends

Models, graphical methods & opportunities



# of response variables	Classical linear models	Generalized linear models
1	LM family: $E(\mathbf{y})=\mathbf{X}\beta$, $V(y\|\mathbf{X})=\sigma^2\mathbf{I}$ ANOVA, regression, … Many graphical methods: effect plots, spread-leverage, influence, …	GLM: $E(\mathbf{y})=g^{-1}(\mathbf{X}\beta)$, $V=V[g^{-1}(\mathbf{X}\beta)]$ poisson, logistic, loglinear, … Some graphical methods: mosaic plots, 4fold plots, diagnostic plots, …
2+	MLM: $E(\mathbf{Y})=\mathbf{X}\beta$, $V(\mathbf{Y}\|\mathbf{X})=\mathbf{I}\otimes\mathbf{\Sigma}$ MANOVA, MMReg, … Graphical methods: ???	MGLM: some beginnings… multivariate count data Graphical methods: ???

Tomorrow: Someone's PhD thesis (better models)

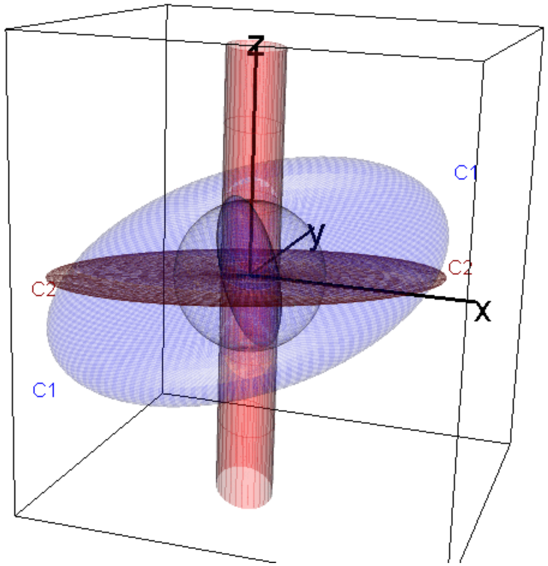
Applications: big data, genomics, … beg for better graphical methods

Geometric ellipsoids

- Ellipsoids in *p* dimensional space
  - proper (“fat”)
  - improper (“thin”) – rank(C) < p
  - unbounded – infinite eigenvalue(s)

$$\mathcal{E} := \{\mathbf{x}, \quad \text{such that} \quad \mathbf{x}^T \mathbf{C} \mathbf{x} \le 1\}$$

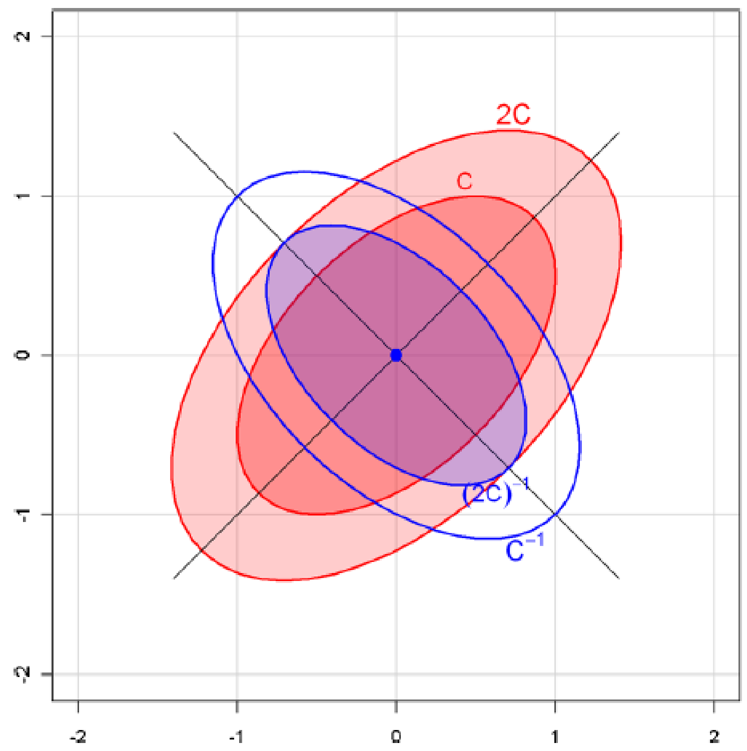
$$\mathbf{C}_1 = \begin{bmatrix} 6 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}, \quad \mathbf{C}_2 = \begin{bmatrix} 6 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$



C_1 (blue): proper & fat; C_1^{-1} is also fat, but in orthogonal directions

C_2 (red): improper & thin; C_2^{-1} is an unbounded cylinder of elliptical x-section

Geometric ellipsoids



- In 2D, ellipses are easily seen as determined by their principal axes– eigenvectors of **C**
  - Eigenvalues, λ_i, determine the sizes
  - In applications, this is often called "data space", using **C** = $\mathbf{X}^T\mathbf{X}$ or a multiple
- There is also a **dual space**, that of C^{-1}
  - Same axes, but sizes ~ $1/\lambda_i$
  - In applications, this is often called "**β** space" or parameter space, using $\mathbf{C}^{-1}=(\mathbf{X}^T\mathbf{X})^{-1}$ or a multiple
- This is a powerful idea that can be exploited in data visualization
  - Galton recognized the first point; Hotelling (1933) made it precise
  - The idea of the dual space comes from Dempster (1969); Monette (1990) explained why it mattered.

Data ellipsoids

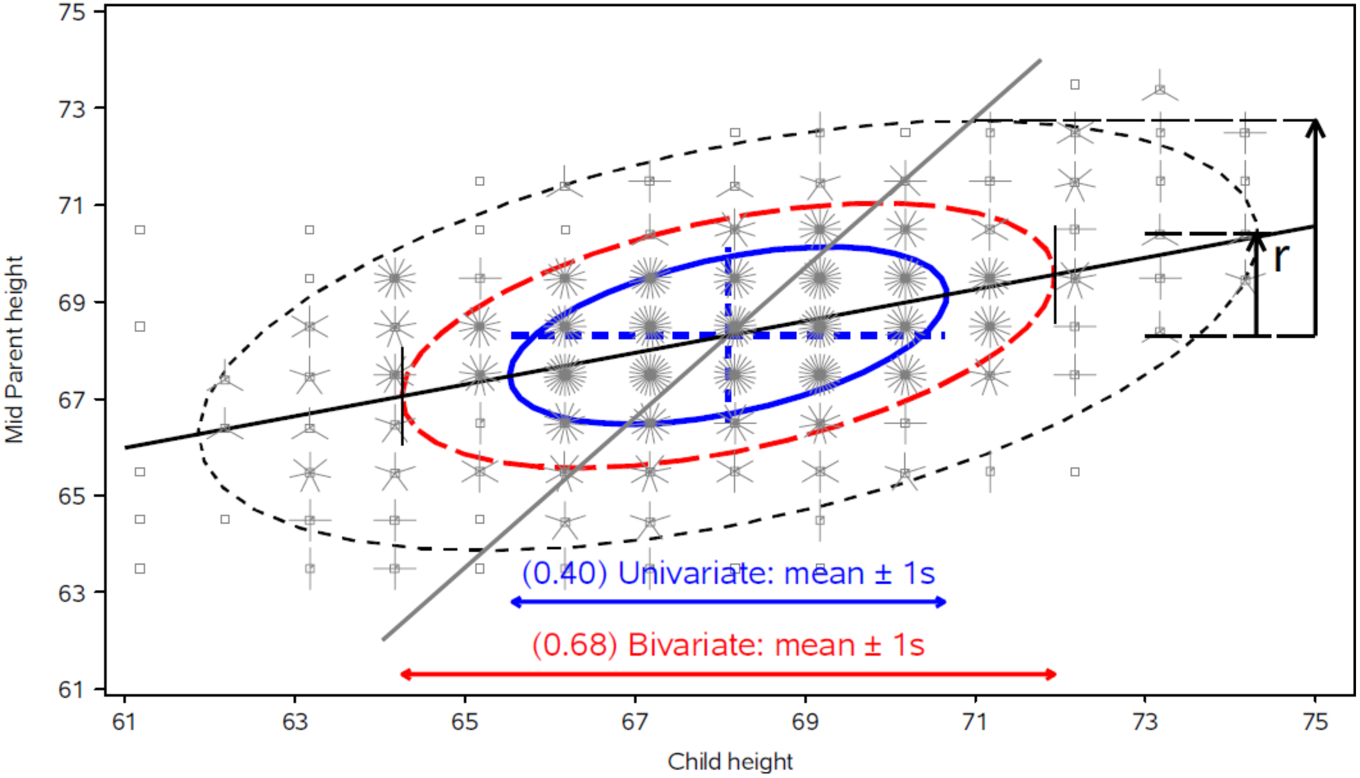
- For a *p*-dimensional multivariate sample, $\mathbf{Y}_{N \times p}$, the sample mean vector, $\bar{\mathbf{y}}$, and sample covariance matrix, **S**, are minimally sufficient statistics under classical (gaussian) assumptions.
- These can be represented visually by the *p*-dimensional data ellipsoid, $\mathcal{E}_c$ of size ("radius") *c*,

$$\mathcal{E}_c(\bar{\mathbf{y}}, \mathbf{S}) := \{\mathbf{y} : (\mathbf{y} - \bar{\mathbf{y}})^T \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \le c^2\}$$

- $\rightarrow$ an ellipsoid centered at the means whose size & shape reflects variances & covariances
- We consider this a minimally sufficient visual summary of multivariate location and scatter.

Data ellipsoids: properties

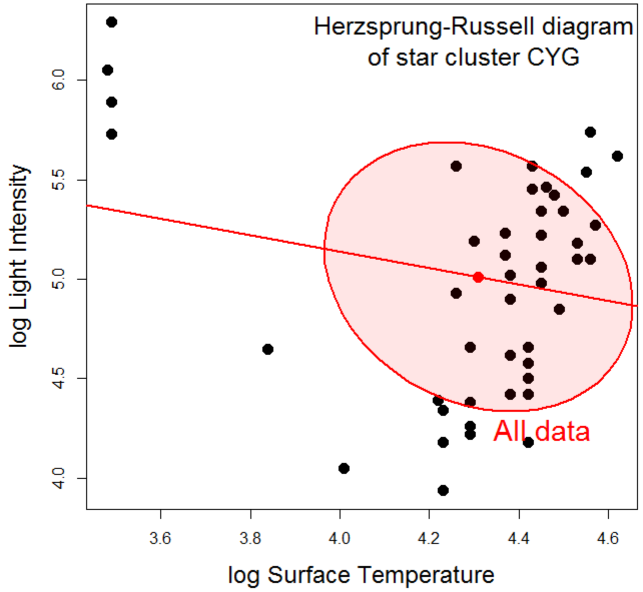
- Ellipsoid boundary: Mahalanobis $D_M^2(y_i) \sim \chi_p^2$
  - p=2: shadows generalize univariate confidence intervals
  - eccentricity: precision; visual estimate of correlation



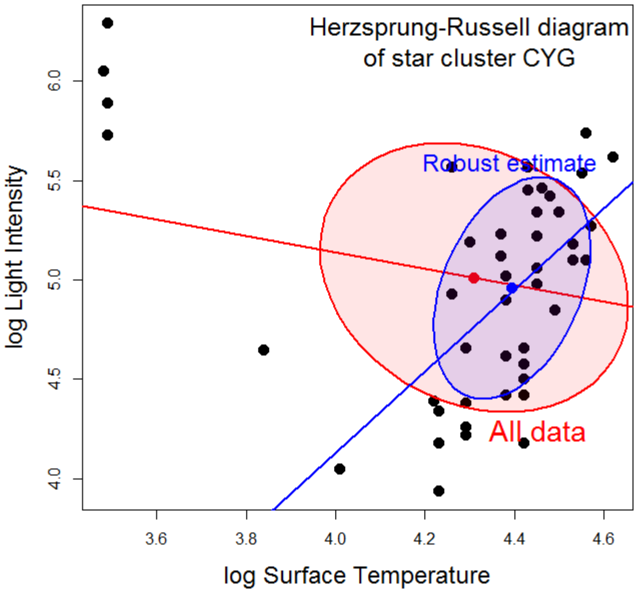
Robust methods: robust=TRUE

- Outliers and high-leverage points challenge routine use of classical, gaussian methods: so yesterday! you say…
  - Robust estimation of center (**μ**) and scatter (**Σ**) is now well established
  - High breakdown bound methods: MCD, MVE, …
  - These are increasingly available in standard software

```
dataEllipse(logtemp, loglight, ...)
```

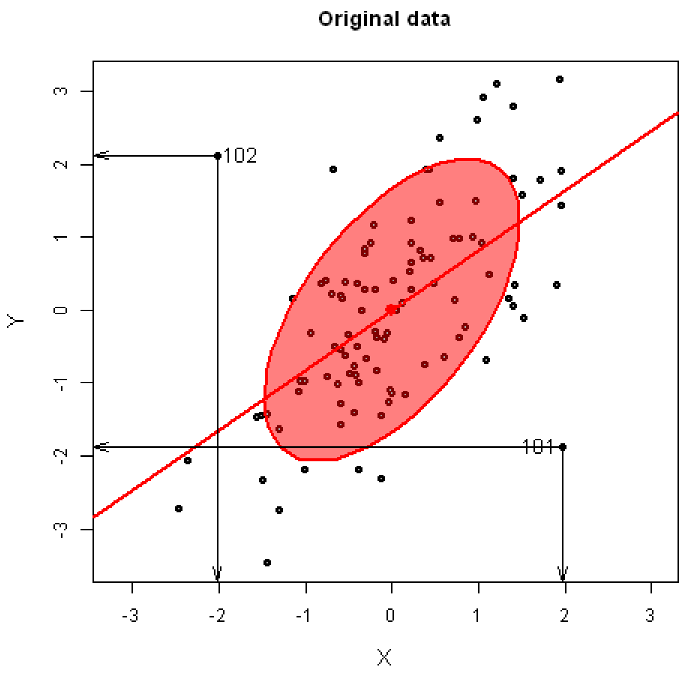


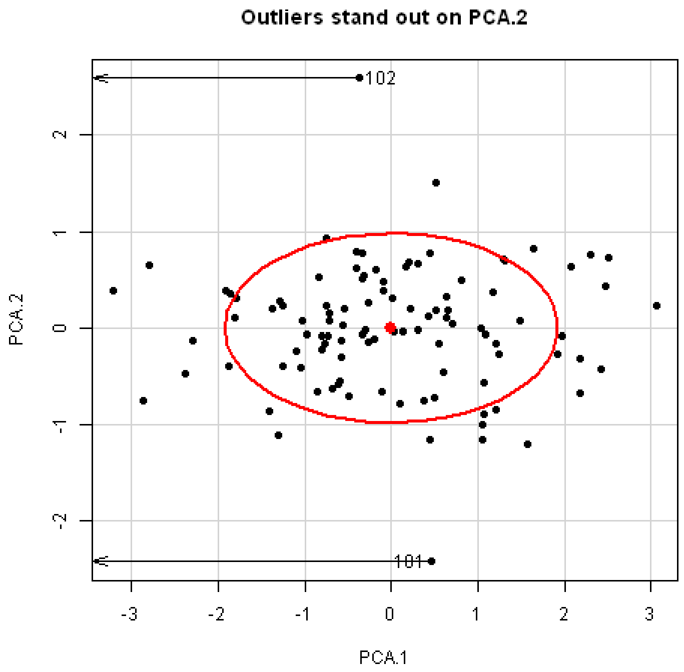
```
dataEllipse(logtemp, loglight,
robust=TRUE, ...)
```



Insight: Outlier detection

- Multivariate outliers might be invisible in univariate views, but become readily apparent on the *smallest* principal component
  - 100 observations on two correlated normal variables with two bivariate outliers near (2,2), (-2, -2)



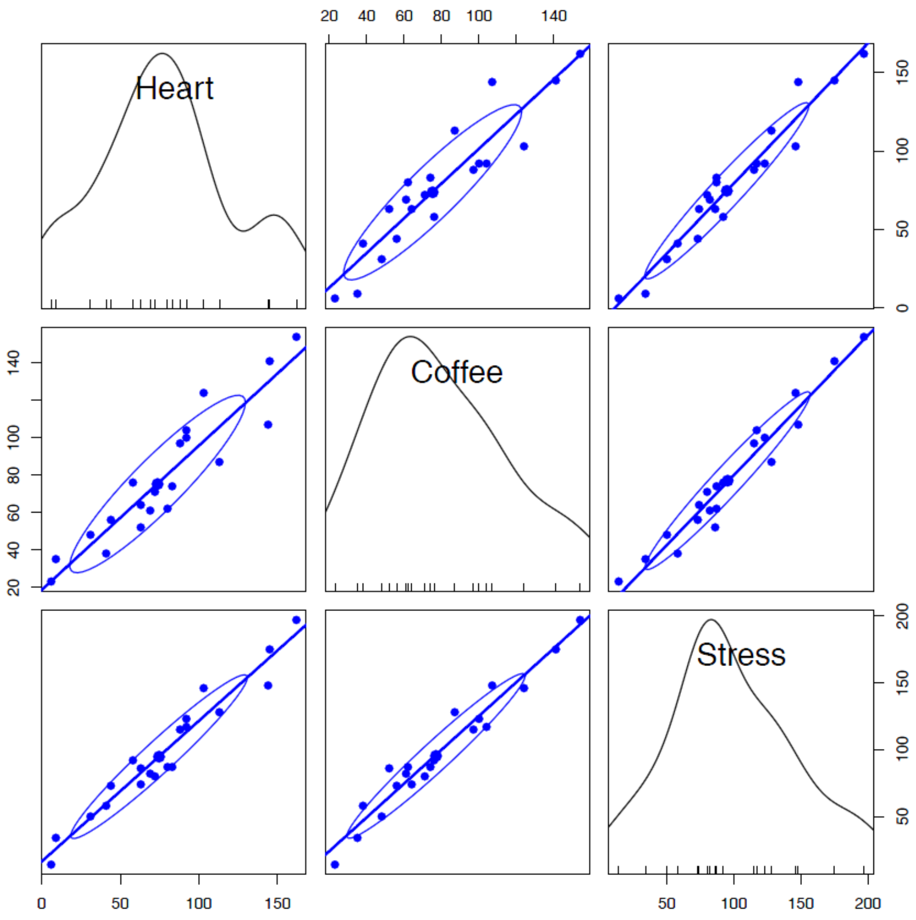


Animation: http://www.datavis.ca/gallery/animation/outlier-demo/

Insight: Measurement error

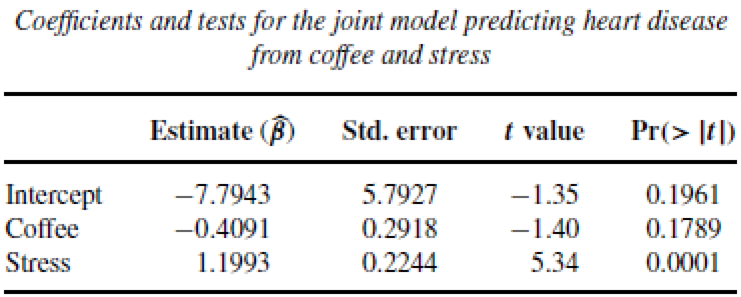
- In classical linear models, predictors (X) are usually assumed to be fixed (non-random), or measured w/o error
  - Rarely true in social science and medicine
  - Structural equation models often used to account for this
  - What effects do errors in predictors have on typical regression models?
- Main ideas:
  - Ellipses in data space show effects on bias and precision
  - The same effects can be seen in parameter (**β**) space

Coffee, stress and heart disease



Imagine a small study investigating the relation between a measure of heart disease (y), and coffee consumption (x1) and stress (x2)

Results: lm(Heart ~ Coffee + Stress)



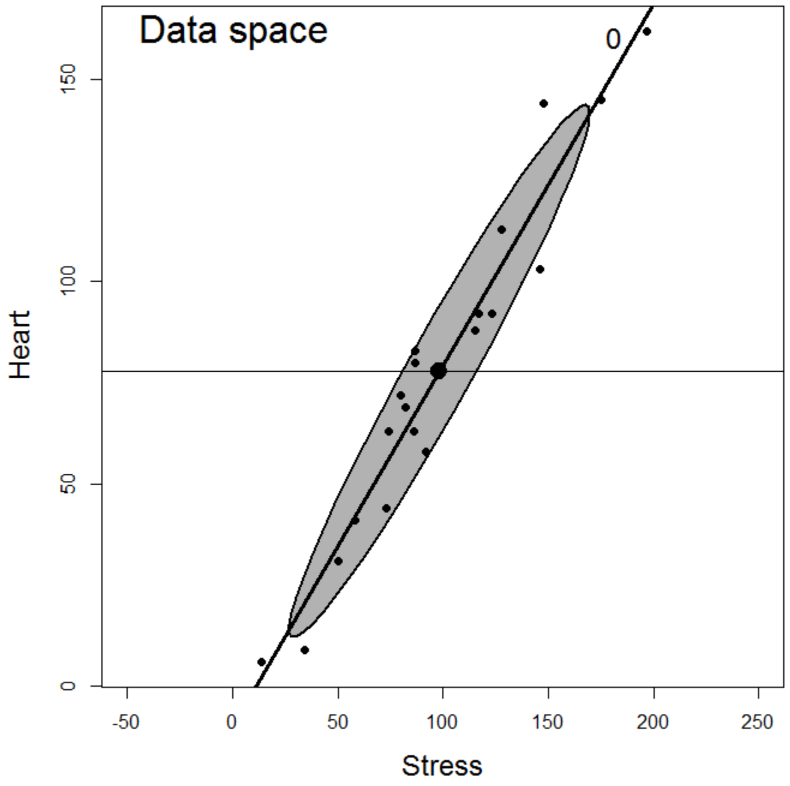
Coefficients and tests for the joint model predicting heart disease from coffee and stress

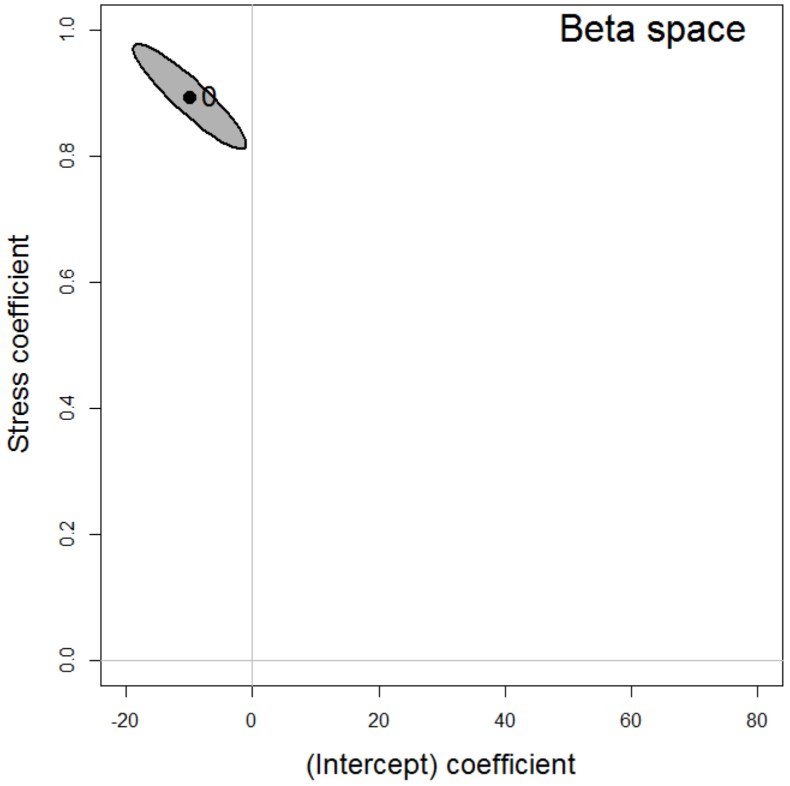
	Estimate ($\hat{\beta}$)	Std. error	t value	Pr(> $\|t\|$)	
Intercept	−7.7943	5.7927	−1.35	0.1961	
Coffee	−0.4091	0.2918	−1.40	0.1789	✗
Stress	1.1993	0.2244	5.34	0.0001	✓

Wow! That means I can drink all the coffee I want as long as I avoid stress.

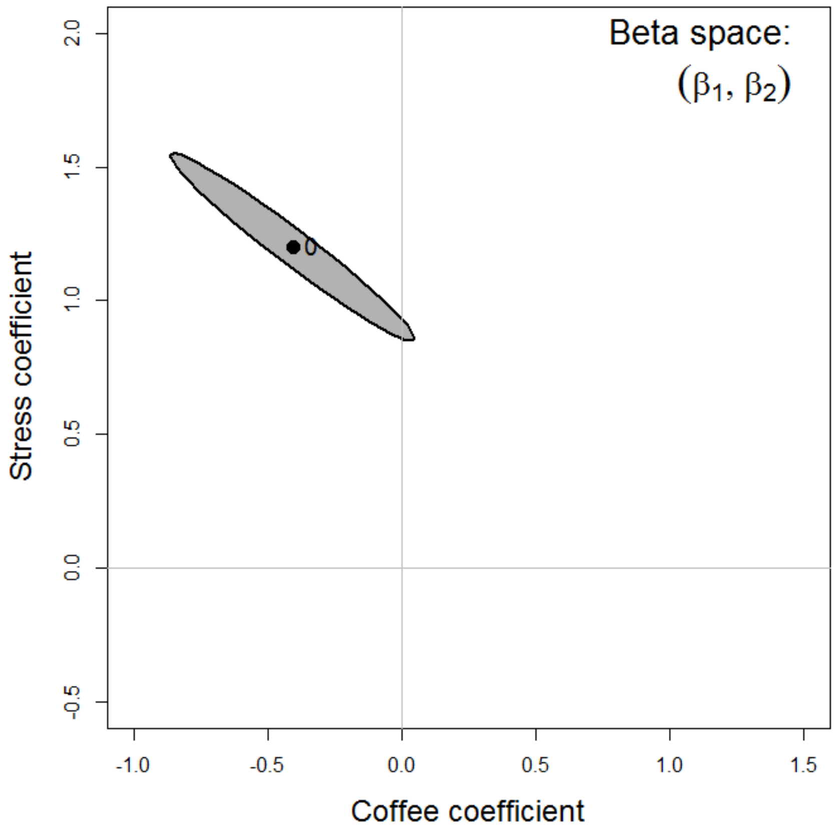
Adding measurement error

- Measurement error in Heart (y) decreases precision, but does not add bias
- Measurement error in Stress (x_2) biases its effect $\beta_{Stress} \rightarrow 0$ & decreases precision





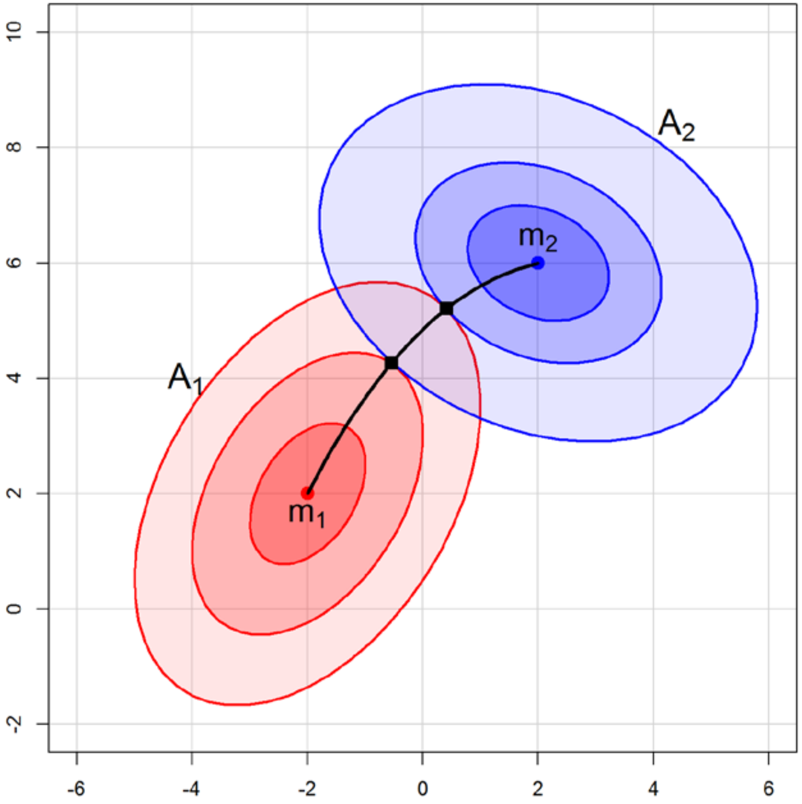
Measurement error: even worse!



- As error in Stress increases, $\beta_{Stress} \to 0$.
  - OK, understand that now.
- But, at the same time, the effect of coffee, β_{Coffee} seems to become larger!
- Elliptical insight:
  - Increasing error in x_2 drives coefficient for x_1 toward the marginal model, ignoring x_2
  - You can also see that in this case the std. errors of β_{Coffee} decrease with error in Stress!

Kissing ellipsoids

$$\mathbf{A}_1 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 1.5 & -0.3 \\ -0.3 & 1.0 \end{pmatrix}$$



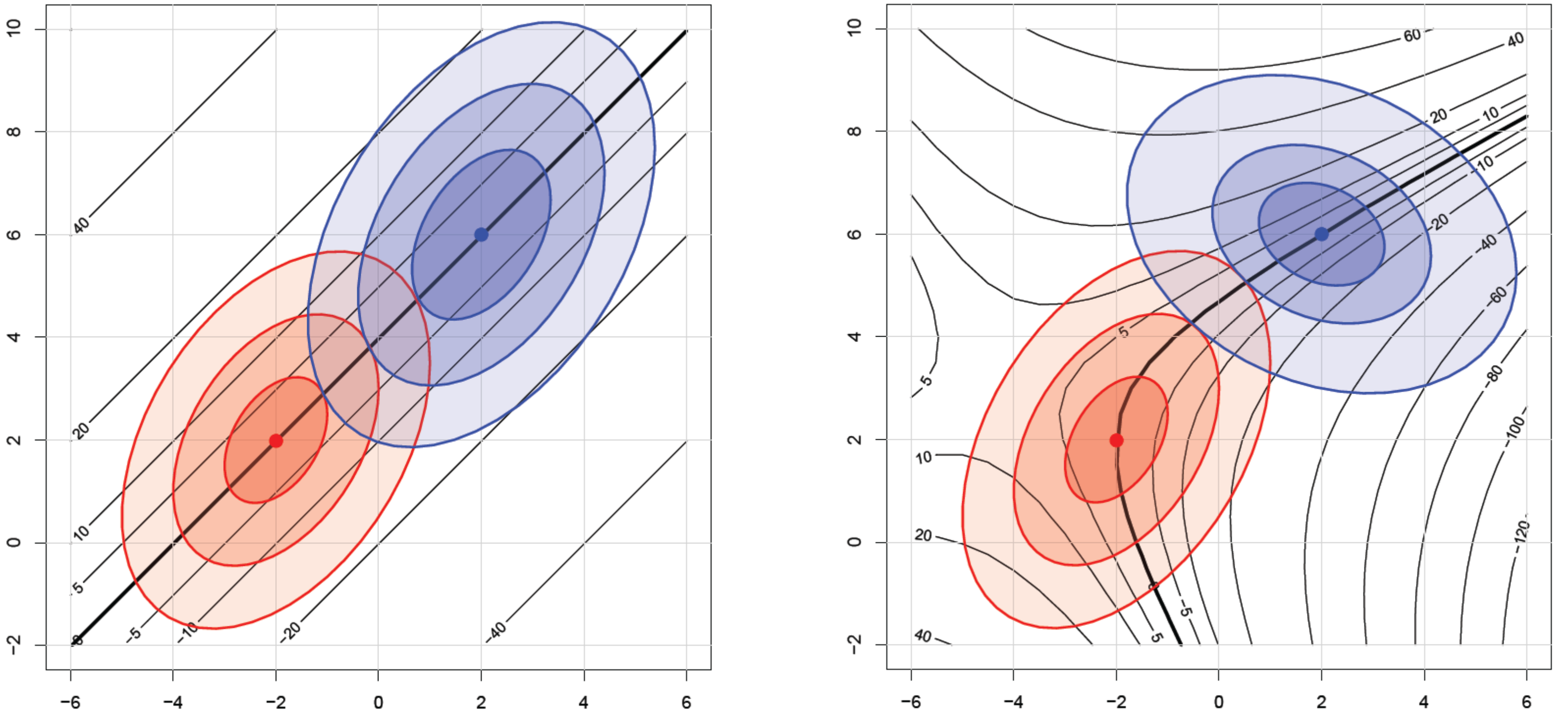
- Imagine 2 magic stones which generate elliptical waves when dropped into a pond at locations $\mathbf{m}_1$ & $\mathbf{m}_2$
- Their locus of osculation is the set of points where the tangents to the ellipses are parallel– where the ellipses kiss!
- The solution has a lovely bilinear (bisexual?) form

$$(\mathbf{x}-\mathbf{m}_2)^T \mathbf{A}_2^T \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \mathbf{A}_1^T (\mathbf{x}-\mathbf{m}_1) = 0$$

Kissing ellipsoids: Discriminant analysis

This is exactly the situation in two group discriminant analysis:

- Under the assumption of equal covariance matrices, $\Sigma_1 = \Sigma_2$, the locus of osculation is linear--- the discriminant axis, and we have LDA, with $\mathbf{b} = \mathbf{S}_{\text{pooled}}^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$
- If $\Sigma_1 \neq \Sigma_2$, the classification axis is quadratic, and we have QDA
- This is why you need to test for $\Sigma_1 = \Sigma_2$!



The HE plot framework

- Hypothesis-Error (HE) plots
  - Visualize multivariate tests in the MLM
  - Linear hypotheses--- lower-dimensional ellipsoids
  - Extension: HE plot matrices
- Canonical displays
  - low-dimensional multivariate juicers
  - shows data in the space of maximal effects
- Covariance ellipsoids
  - visualize tests of homogeneity of covariance matrices
- For all: robust methods are available or good research projects!

HE plot framework: Trivial example

Two groups of middle-school students are taught algebra by instructors using different methods, and then tested on:

- **BM**: basic math problems (7 * 23 – 2 * 9 = ?)
- **WP**: word problems (“a train travels at 23 mph for 7 hours, but for 2 hours …”)

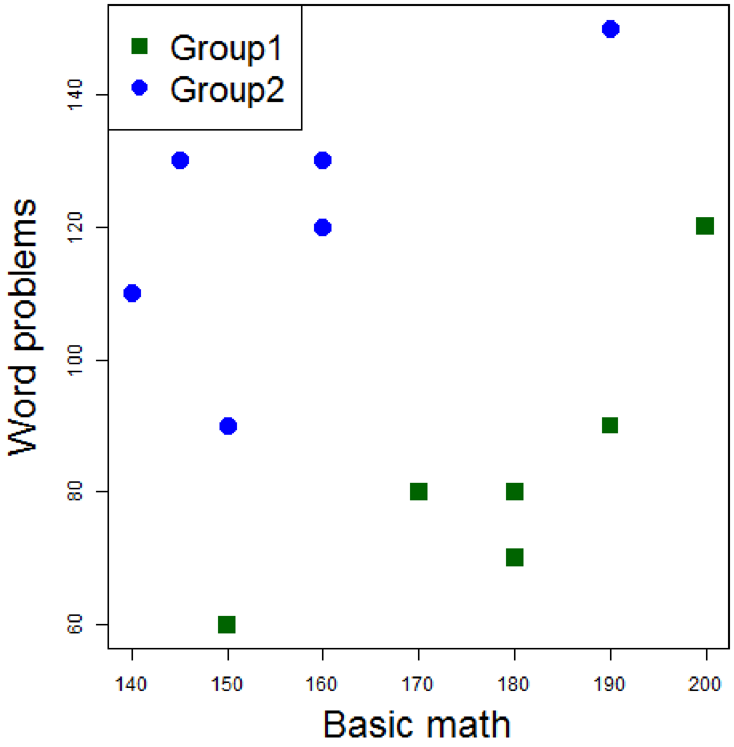
Do the groups differ on (BM, WP) by a multivariate test?
If so, how ???

```
> mod <- lm(cbind(BM, WP) ~ group, data=mathscore)
> Anova(mod)

Type II MANOVA Tests: Pillai test statistic
      Df test stat approx F num Df den Df    Pr(>F)
group  1   0.86518   28.878      2      9 0.0001213 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

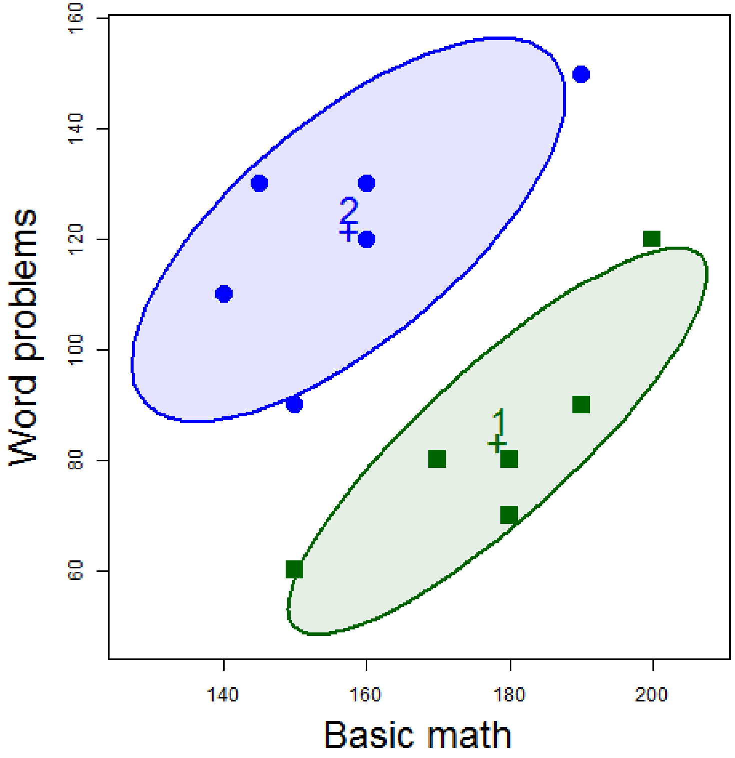
HE plot framework: Visual overview

The data ellipsoid is a sufficient visual summary for multivariate location & scatter, just as $(\bar{y}, \mathbf{S})$ are sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Data

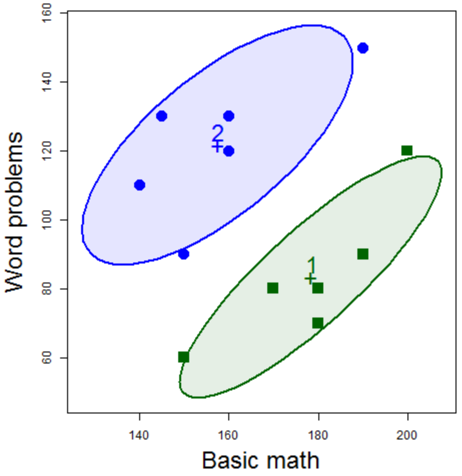


Data ellipses

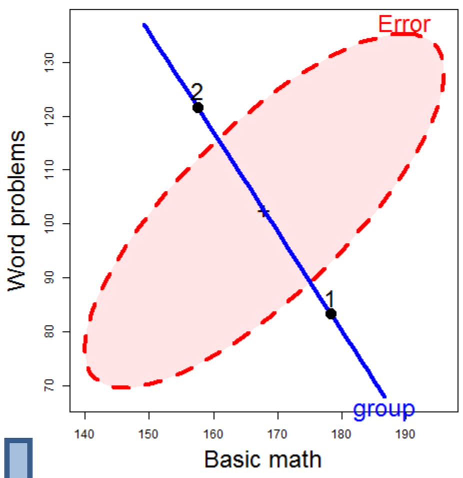


Visual overview

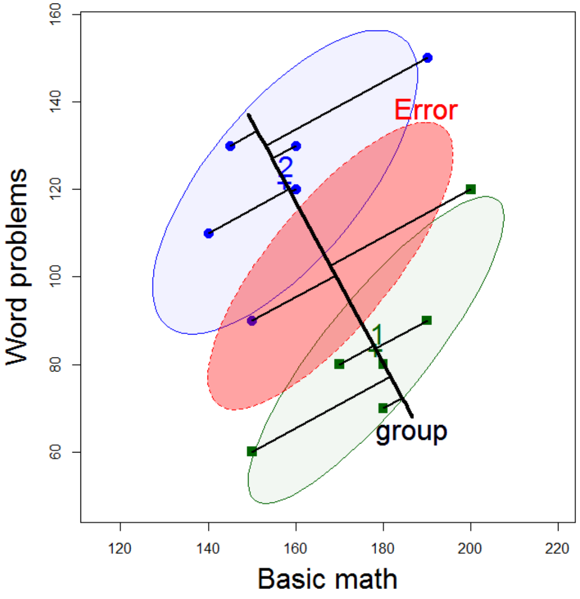
Data ellipses



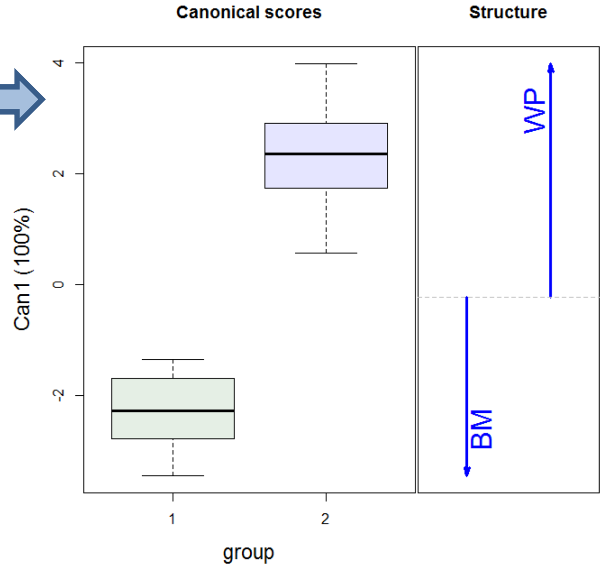
HE plot



Discriminant scores



Canonical space



HE plots: Details

- Hypothesis - Error (HE) plots provide a simple framework for visualizing MLMs
- All hypothesis tests correspond to statistics based on the eigenvalues, λ_i of $\mathbf{HE}^{-1}$ (**H** "relative to" **E**):
  - **H**: the sum of squares & products (SSP) matrix for the hypothesis
  - **E**: the SSP matrix for error
- For any term, t,

  $\mathbf{H}_t = \widehat{\mathbf{Y}}_t^T \widehat{\mathbf{Y}}_t$ SSP of fitted values

  $\mathbf{E} =$ SSP of residuals in full model

- They answer the question:
  - "How big is the **H** ellipsoid relative to the **E** ellipsoid?"
  - Equivalent Q: How big is the data ellipsoid of fitted values relative to data ellipsoid of residuals?

Ex: Neuro- & Social-Cognitive measures in psychiatric groups

- A study by Leah Hartman @York examined whether patients classified as ‘schizophrenic’ or ‘schizoaffective’ (on DSM-IV) could be distinguished from a normal, control sample on standardized tests in the following domains:
  - Neuro-Cognitive: processing speed, attention, verbal learning, visual learning, problem solving
  - Social-cognitive: managing emotions, theory of mind, externalizing bias, personalizing bias
- Research questions → MANOVA contrasts
  - Do the two psychiatric groups differ from the controls?
  - Do the two psychiatric groups differ from each other?

See: Friendly & Sigal (2017), Graphical Methods for Multivariate Linear Models in Psychological Research: An R Tutorial *The Quantitative Methods for Psychology, 13*, 20-45, http://dx.doi.org/10.20982/tqmp.13.1.p020

Neuro-cognitive measures

```
library(heplots)
library(candisc)
data(NeuroCog, package="heplots")

# fit the MANOVA model, test hypotheses
NC.mlm <- lm(cbind(Speed, Attention, Memory, Verbal, Visual,ProbSolv) ~ Dx,
             data=NeuroCog)
Anova(NC.mlm)


Type II MANOVA Tests: Pillai test statistic
   Df test stat approx F num Df den Df    Pr(>F)
Dx  2    0.2992   6.8902     12    470 1.562e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

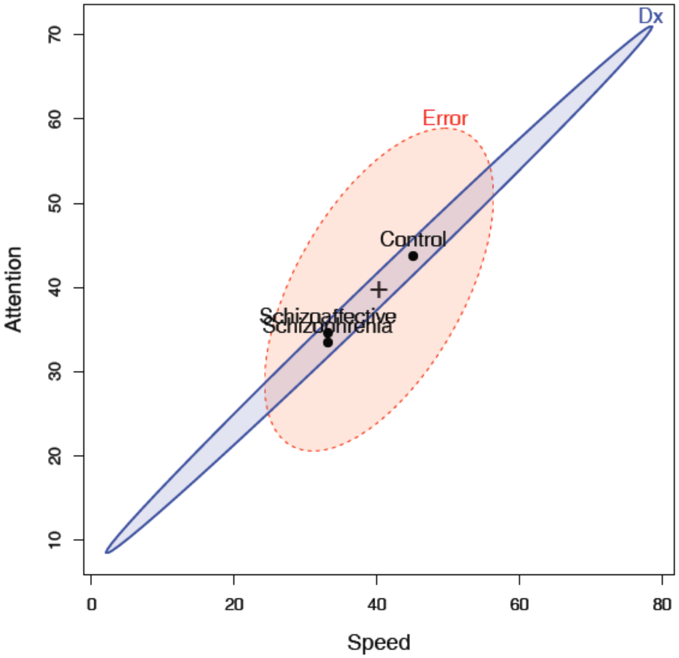
So, the groups differ. But how?
What about the research hypotheses?

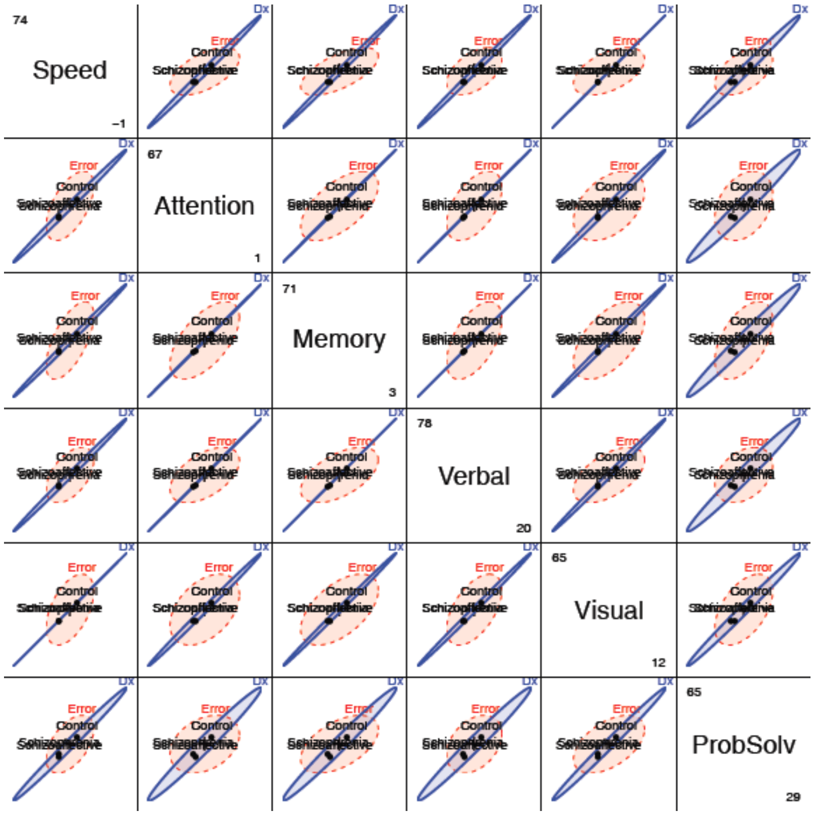
```
> contrasts(NeuroCog$Dx)
                [,1] [,2]
Schizophrenia   -0.5    1
Schizoaffective -0.5   -1
Control          1.0    0
```

Visualize me: in data space

```
# Bivariate view for any 2 responses:
heplot(NC.mlm, var=1:2, ...)
```

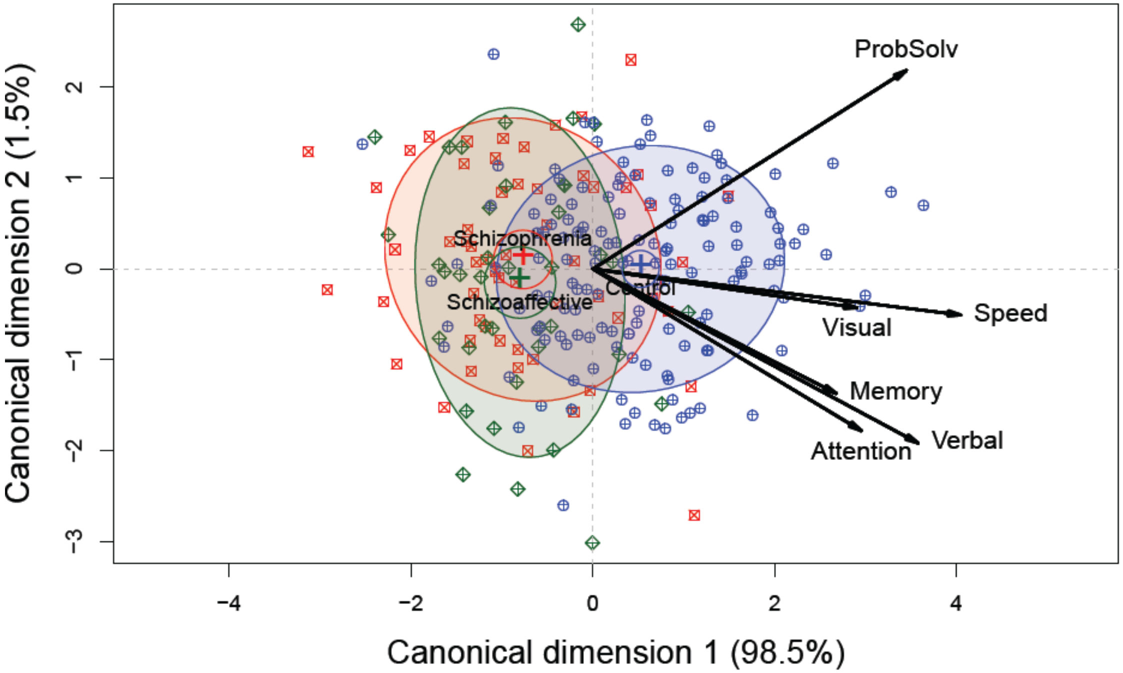
```
# HE plot matrix: for all responses
pairs(NC.mlm, ...)
```





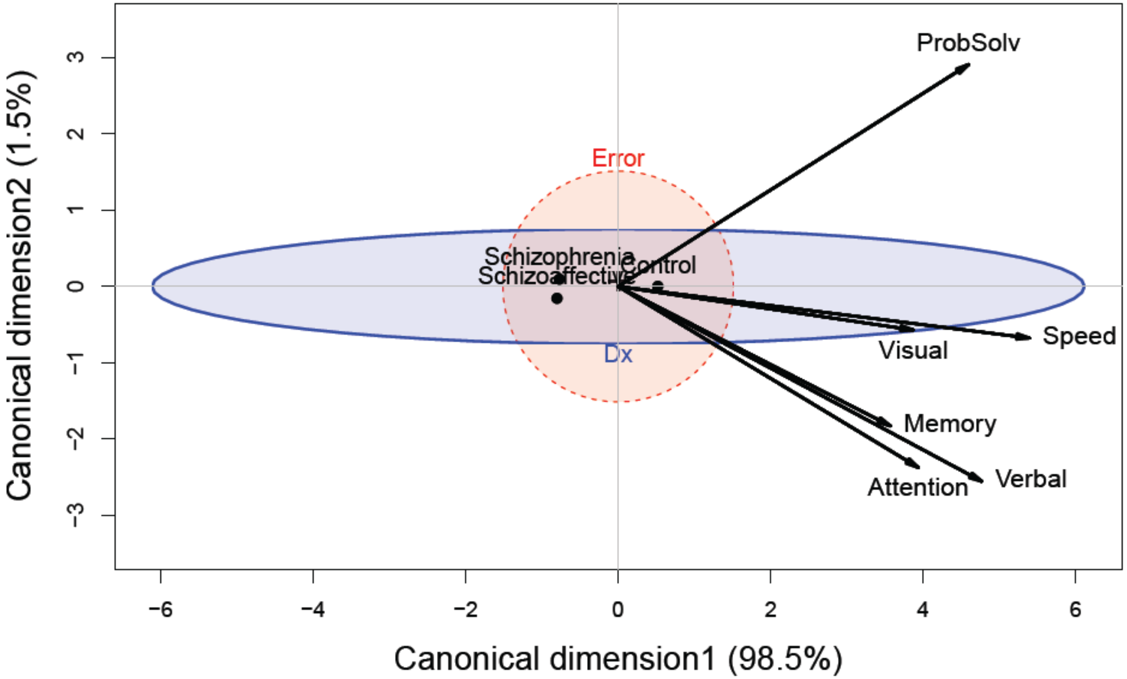
Visualize me: in canonical space

- As with biplot, we can visualize MLM hypothesis variation for *all* responses by projecting ***H*** and ***E*** into low-rank space.
- Canonical projection: $\boldsymbol{Y}_{n \times p} \mapsto \boldsymbol{Z}_{n \times s} = \boldsymbol{Y}\boldsymbol{E}^{-1/2}\boldsymbol{V}$, where ***V*** = eigenvectors of $\boldsymbol{H}\boldsymbol{E}^{-1}$.
- This is the view that maximally discriminates among groups, ie max. ***H*** wrt ***E*** !



Visualize me: canonical HE plots

- Canonical HE plot is just the HE plot of canonical scores, $(\mathbf{z}_1, \mathbf{z}_2)$ in 2D,
- or, $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, in 3D.
- As in biplot, we add vectors to show relations of the $\mathbf{y}_i$ response variables to the canonical variates.
- variable vectors here are structure coefficients = correlations of variables with canonical scores.



Social cognitive measures

```
> data(SocialCog, package="heplots")
> SC.mlm <-  lm(cbind(MgeEmotions,ToM, ExtBias, PersBias) ~ Dx,
               data=SocialCog)
> Anova(SC.mlm)

Type II MANOVA Tests: Pillai test statistic
   Df test stat approx F num Df den Df   Pr(>F)
Dx  2     0.212     3.97      8    268 0.00018 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

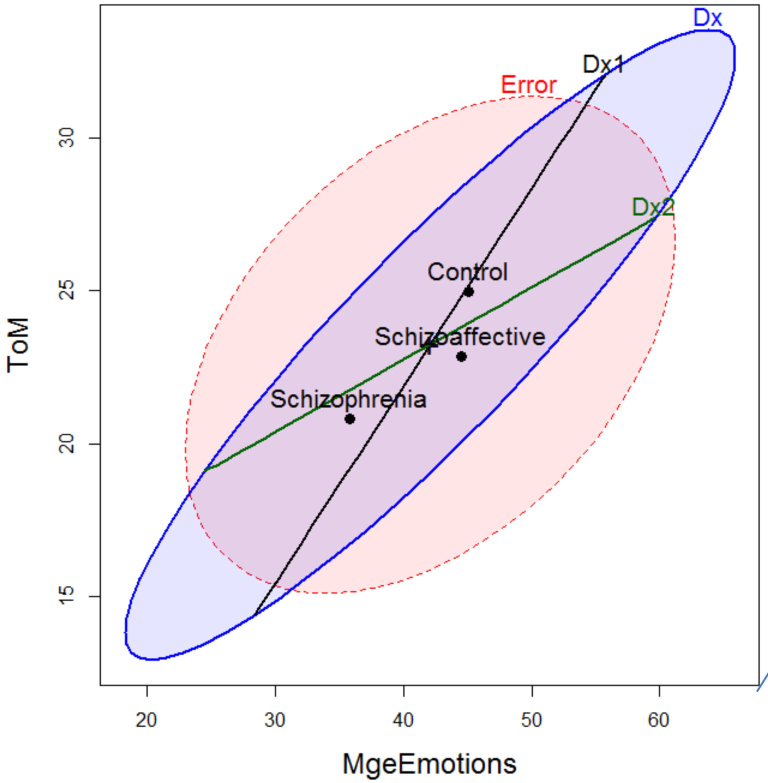
Test contrasts: Dx1 = Normal vs. Patient; Dx2 = Schizo vs. Schizoaffective

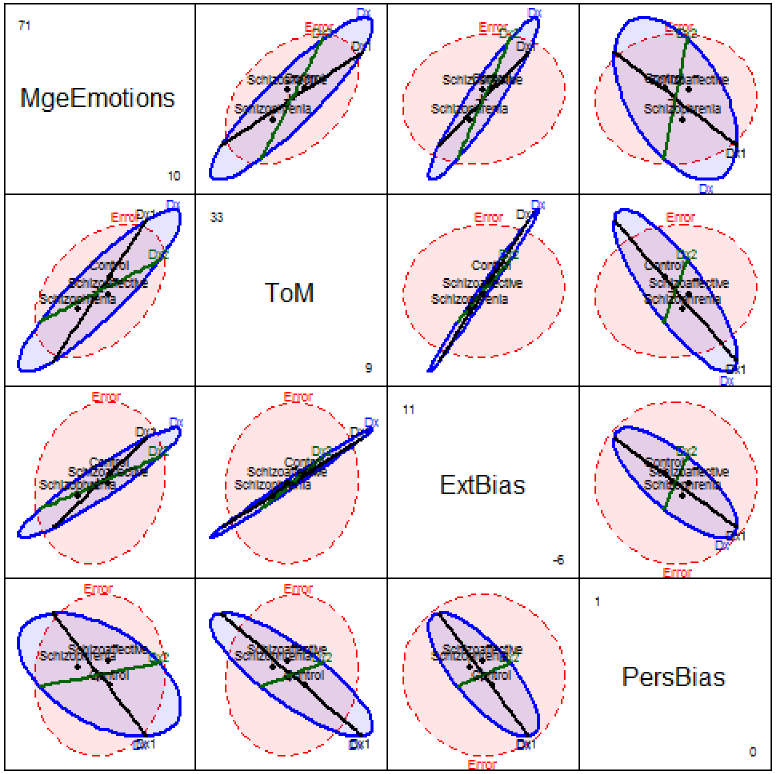
```
> print(linearHypothesis(SC.mlm, "Dx1"), SSP=FALSE)
Multivariate Tests:
                 Df test stat approx F num Df den Df   Pr(>F)
Pillai            1    0.1355    5.212      4    133 0.000624 ***

> print(linearHypothesis(SC.mlm, "Dx2"), SSP=FALSE)
Multivariate Tests:
                 Df test stat approx F num Df den Df Pr(>F)
Pillai            1    0.0697    2.493      4    133 0.0461 *
```

Visualize me: data space

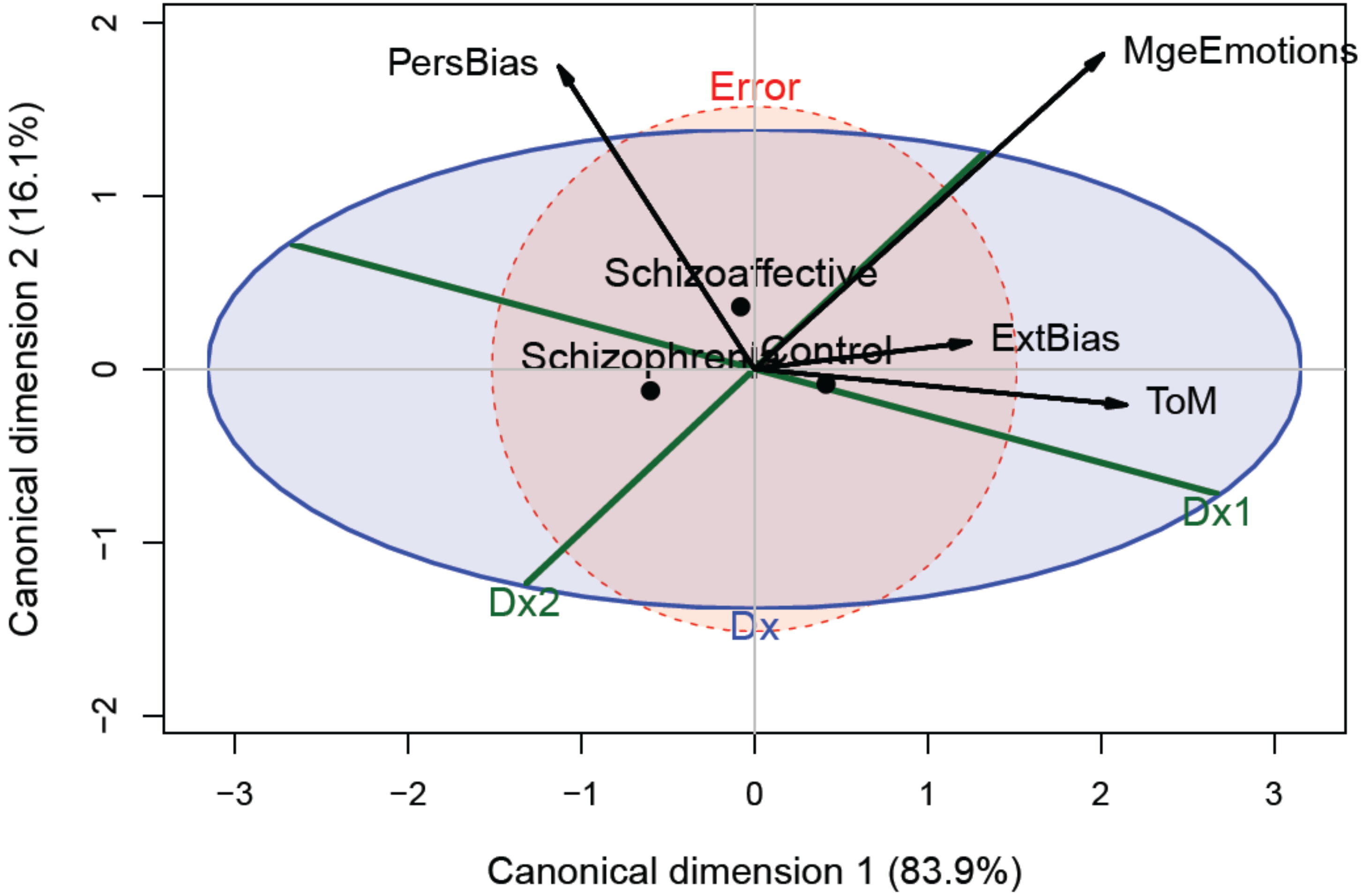
```
heplot(SC.mlm,
hypotheses=list("Dx1", "Dx2"),...)
```

```
pairs(SC.mlm,
hypotheses=list("Dx1", "Dx2"),...)
```





Visualize me: canonical space



Robust MLMs

- Robust methods for univariate LMs are now well-developed and implemented
  - → proper SEs, CIs and hypothesis tests
- Analogous methods for multivariate LMs are a current hot research topic
- The heplots package now provides `robmlm()` for the fully general MLM (MANOVA, MMReg)
  - Uses simple M-estimator via IRLS
  - Weights: calculated from Mahalanobis $D^{2,}$ a robust covariance estimator and weight function, ψ(D^2)

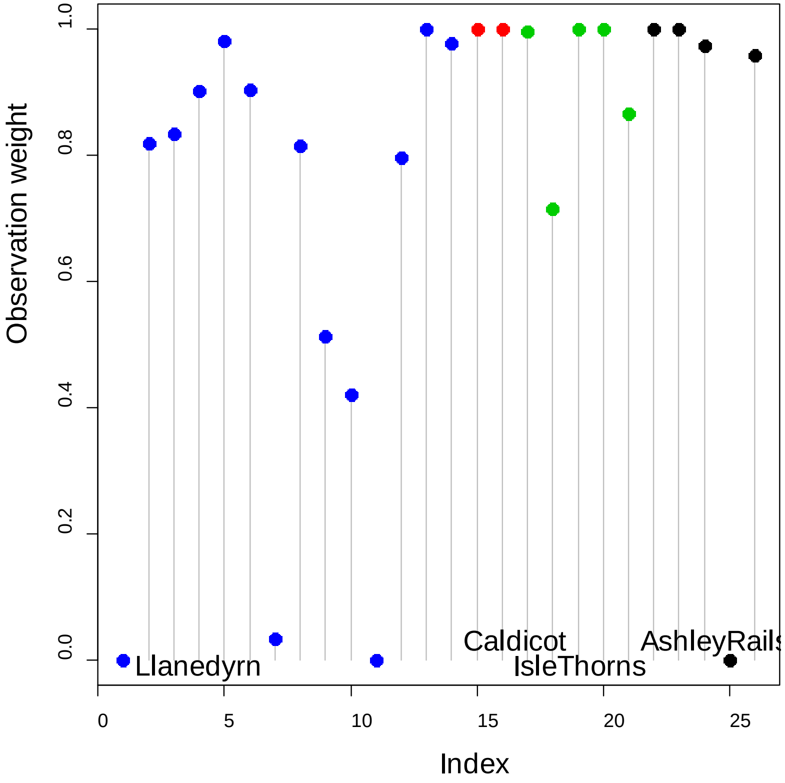
$$D^2 = (\mathbf{Y} - \bar{\mathbf{Y}})^T \mathbf{S}_{\text{robust}}^{-1} (\mathbf{Y} - \bar{\mathbf{Y}}) \sim \chi_p^2$$

  - Downside: SEs, *p*-values only approximate

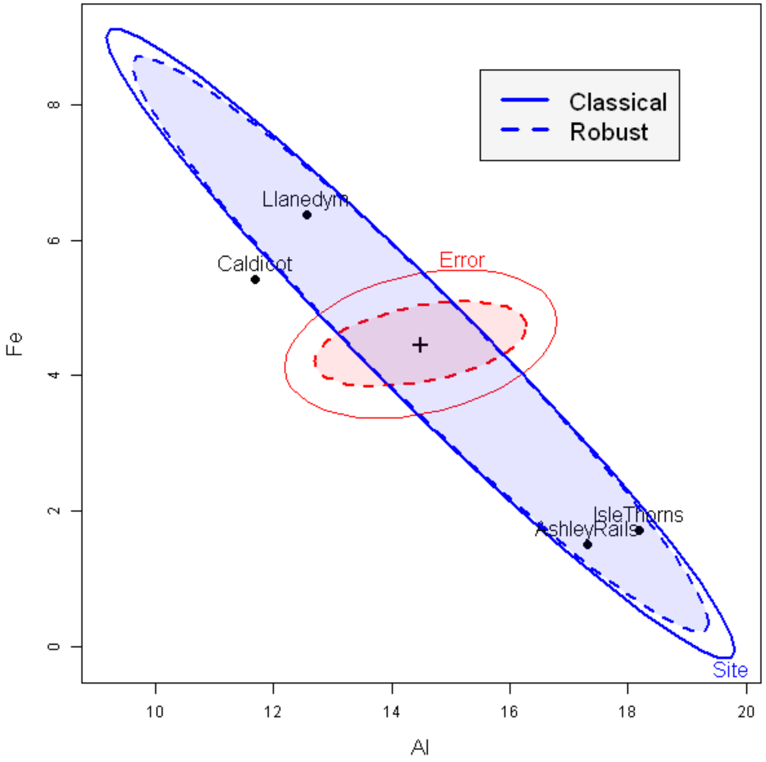
Robust MLMs: Example

```
> pottery.mod <- lm(cbind(Al,Fe,Mg,Ca,Na)~Site, data=Pottery)
> pottery.rmod <- robmlm(cbind(Al,Fe,Mg,Ca,Na)~Site, data=Pottery)
```

Observation weights



overlaid HE plots



Ridge regression: Visualizing bias & precision

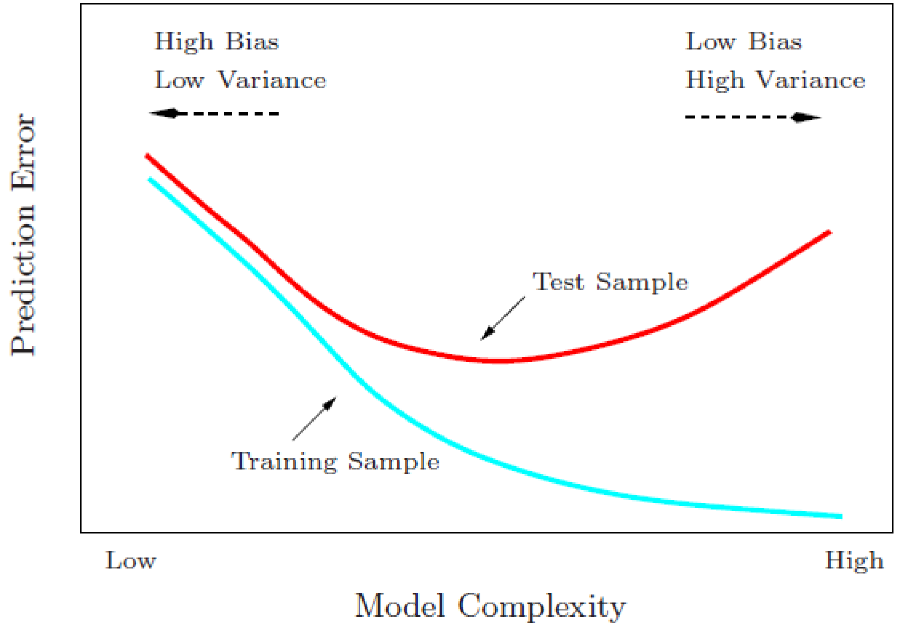
- In the classical linear model, collinearity -- high $R^2(X_i|\text{other Xs})$ -- causes problems:
  - Std errors of coefficients **β** are inflated
  - OLS estimates tend to be too large on average
- Ridge regression & shrinkage methods
  - Desire: increase precision (decrease Var(β))
  - OLS estimates are constrained, shrinking $\boldsymbol{\beta}^T\boldsymbol{\beta} \rightarrow 0$
  - All methods use some tuning parameter (k) to quantify tradeoff
  - How to choose?
    - Numerical criteria, generalized cross-validation, bootstrap, etc.

Bias *vs.* Precision tradeoff

- Particularly important when the goal is predictive accuracy
  - Complex models, many predictors, e.g., demand for medical care
  - In-sample prediction error decreases with model complexity
- But, in new samples prediction error suffers from high variance of complex models

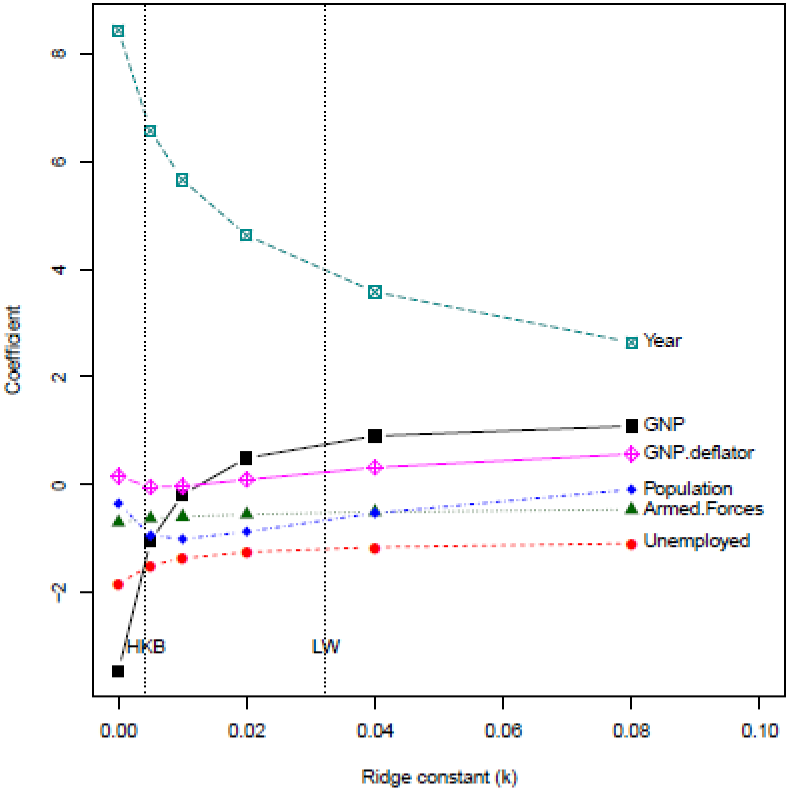
How to visualize the tradeoff?

+1 if you guessed an ellipse !



Univariate ridge trace plots

- Typical: univariate line plot of β_k vs. shrinkage, k
- What can you see here regarding bias vs. precision?
- This is the wrong graphic form, for a multivariate problem!
- Goal: visualize $\widehat{\beta}_k$ vs. $\widehat{\mathrm{Var}}(\widehat{\beta}_k)$



Example: Longley data

Longley (1965) data: economic time series ($n = 16$) of yearly data from 1947 – 1962, often used as an example of extreme collinearity.

```
> names(longley)
```

```
[1] "GNP.deflator" "GNP"          "Unemployed"   "Armed.Forces"
[5] "Population"   "Year"         "Employed"
```

We take number of people Employed as the response:

```
> lmod <- lm(Employed ~ GNP + Unemployed + Armed.Forces +
     Population + Year + GNP.deflator, data = longley)
> vif(lmod)
```

```
      GNP   Unemployed Armed.Forces   Population         Year  GNP.deflator
 1788.513       33.619        3.589      399.151      758.981       135.532
```

As suspected, almost all VIFs are very large.

Ridge regression: Properties

- OLS estimates:

$$\begin{aligned}\widehat{\beta}^{\mathrm{OLS}} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \ , \\ \widehat{\mathrm{Var}}(\widehat{\beta}^{\mathrm{OLS}}) &= \widehat{\sigma}^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}.\end{aligned}$$

- Ridge regression: replaces $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ with $\mathbf{X}^{\mathsf{T}}\mathbf{X} + k\mathbf{I}$
  - drives $|\mathbf{X}^{\mathsf{T}}\mathbf{X} + k\mathbf{I}|$ away from zero even if $|\mathbf{X}^{\mathsf{T}}\mathbf{X}| \approx 0$.
  - drives $||\beta|| = (\beta^{\mathsf{T}}\beta)^{1/2}$ toward zero— increasing “bias”
  - decreases the “size” of $\widehat{\mathrm{Var}}(\widehat{\beta})$— increasing precision— in that

$$|\widehat{\mathrm{Var}}(\widehat{\beta}^{\mathrm{OLS}})| \quad \geq \quad |\widehat{\mathrm{Var}}(\widehat{\beta}_k^{\mathrm{RR}})| \qquad \text{decreases with } k$$

Ridge regression: Properties

- Ridge estimates:

$$\begin{aligned} \widehat{\beta}_k^{\mathrm{RR}} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \qquad (1) \\ &= \mathbf{G}_k \, \widehat{\beta}^{\mathrm{OLS}} \ , \\ \widehat{\mathrm{Var}}(\widehat{\beta}_k^{\mathrm{RR}}) &= \widehat{\sigma}^2 \mathbf{G}_k (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \mathbf{G}_k^\mathsf{T} \ . \qquad (2) \end{aligned}$$

where $\mathbf{G}_k = \left[\mathbf{I} + k(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right]^{-1}$, the $(p \times p)$ "shrinkage" matrix.

- Equivalent to penalized LS criterion,

$$\mathrm{RSS}(k) = (\mathbf{y} - \mathbf{X}\beta)^\mathsf{T}(\mathbf{y} - \mathbf{X}\beta) + k\beta^\mathsf{T}\beta \qquad (k \geq 0) \ , \qquad (3)$$

- Or, to a constrained LS minimization problem,

$$\widehat{\beta}^{\mathrm{RR}} = \underset{\beta}{\mathrm{argmin}}(\mathbf{y} - \mathbf{X}\beta)^\mathsf{T}(\mathbf{y} - \mathbf{X}\beta) \quad \text{subject to} \quad \beta^\mathsf{T}\beta \leq t(k) \qquad (4)$$

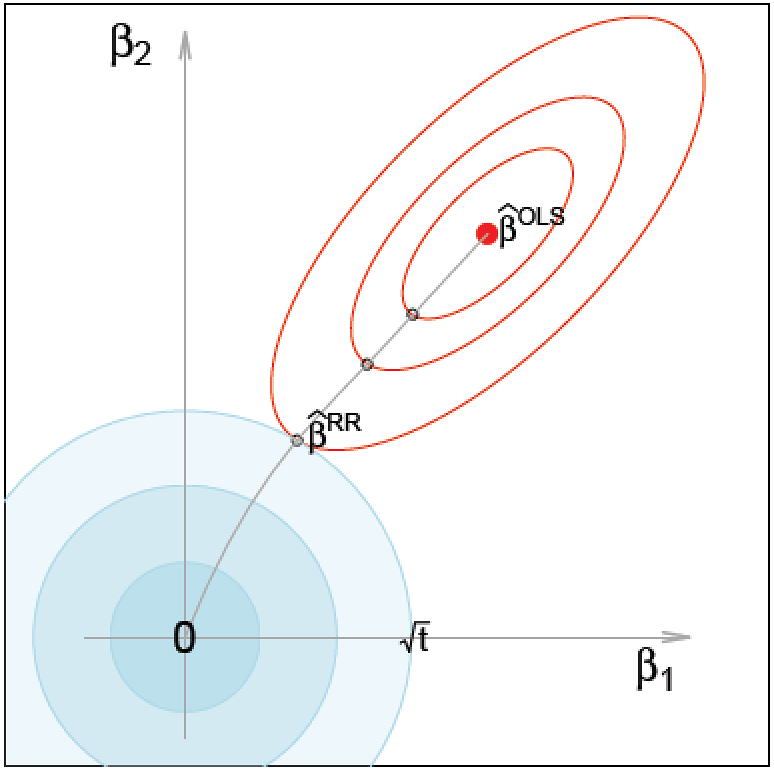
Ridge regression: Geometry

Ridge regression solution has a simple geometric interpretation based on ellipsoids of the $RSS(k)$ function,

$$\mathrm{RSS}(k) = (\mathbf{y} - \mathbf{X}\beta)^\mathsf{T}(\mathbf{y} - \mathbf{X}\beta) + k\beta^\mathsf{T}\beta$$

OLS coefficients are shrunk toward **0** along the locus of osculation of

- Covariance ellipsoid of β^{OLS}
- Unit sphere $\beta^\mathsf{T}\beta \leq t(k)$

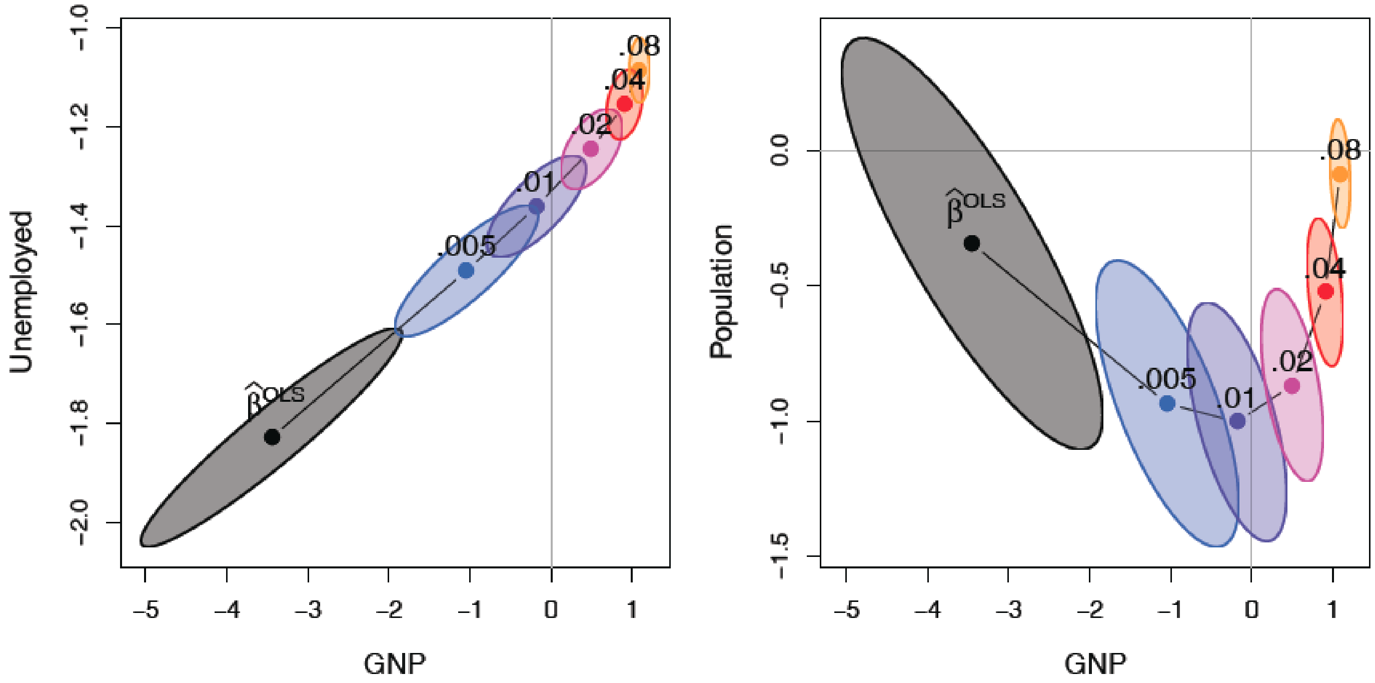


The matrix $\mathbf{G}_k = [\mathbf{I} + k(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}]^{-1}$ shrinks the covariance matrix of β_k in a similar way

Generalized ridge trace plots

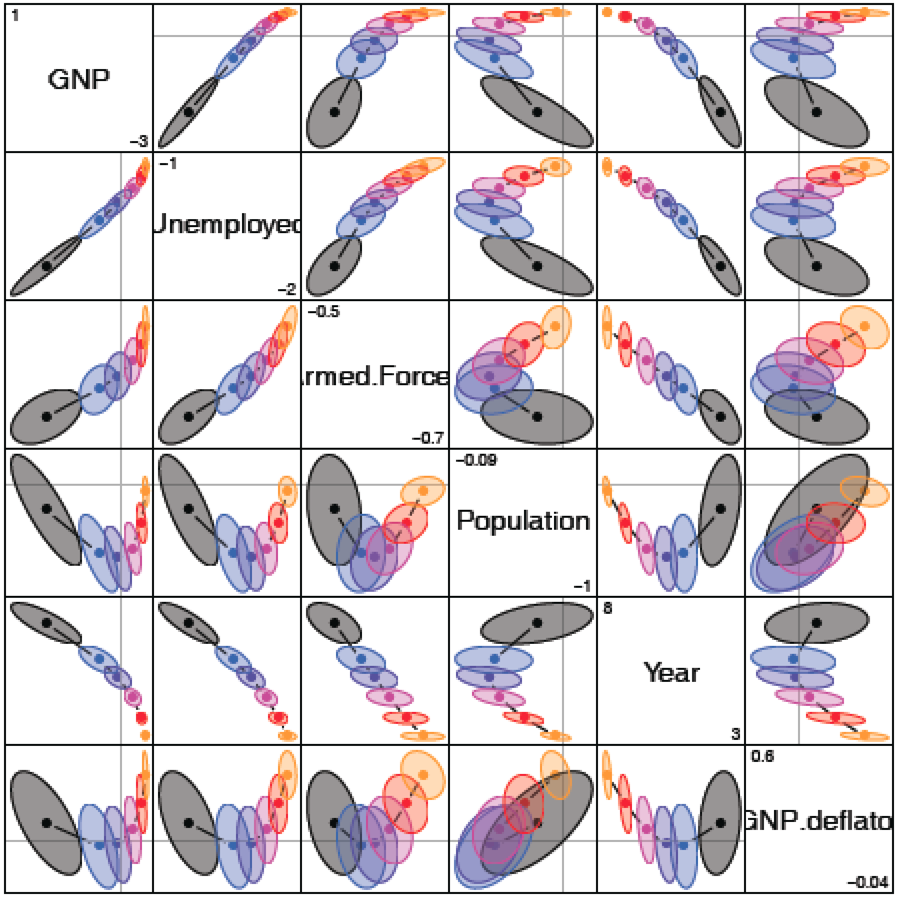
Rather than plotting just the univariate trajectories of β_k vs. k, plot the covariance ellipsoids of $\widehat{\Sigma}_k \equiv \widehat{\text{Var}}(\widehat{\beta}_k)$ over same range of k

- Centers of the ellipsoids are $\widehat{\beta}_k$ – same info as in univariate plot
- Can see how change in one coefficient is related to changes in others
- Relative size & shape of ellipsoids shows directly effect on precision



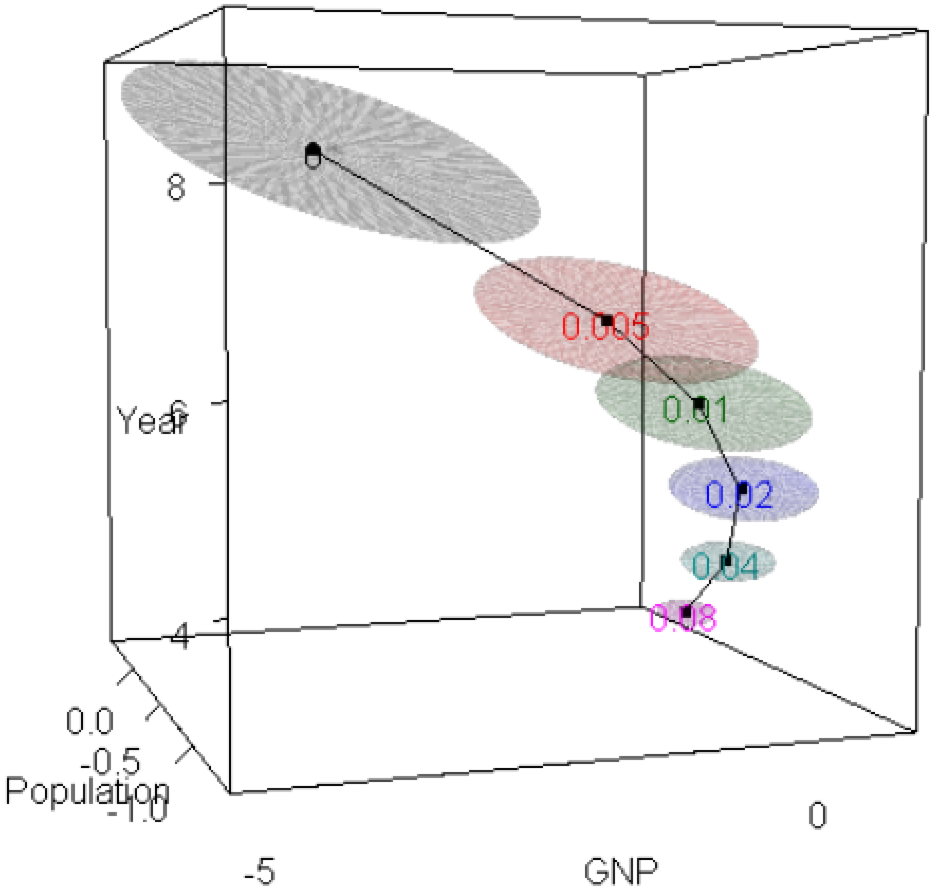
Scatterplot matrix of ridge trace plots

```
> pairs(lridge, radius=0.5, diag.cex=1.75, col=clr, fill=TRUE)
```



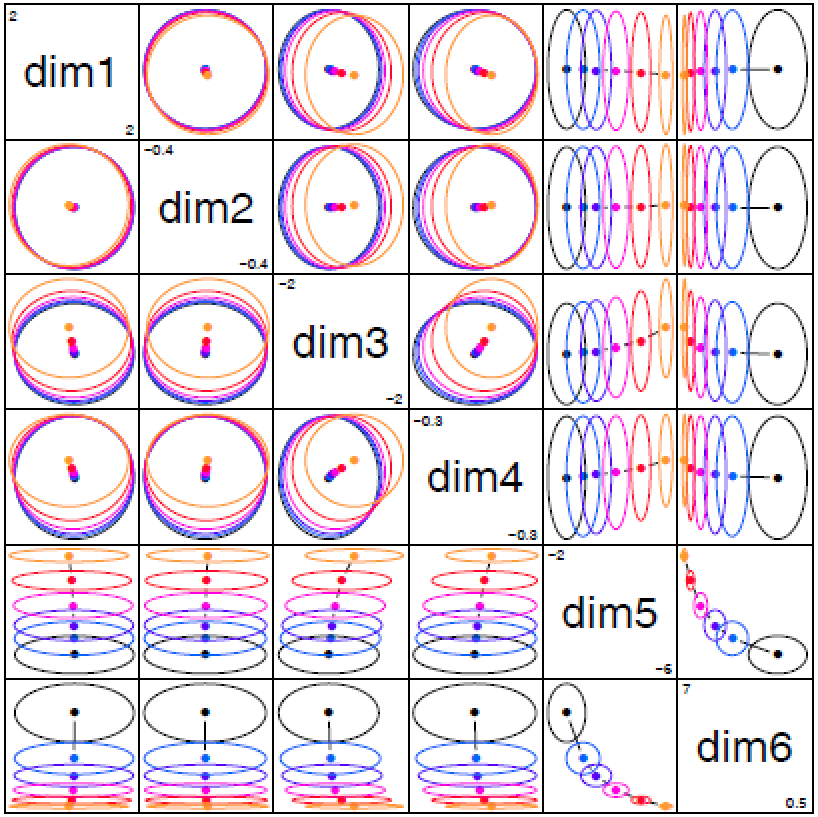
plot3d() method

```
> plot3d(lridge, radius=0.5)
```



Plots in PCA/SVD space: pca() method

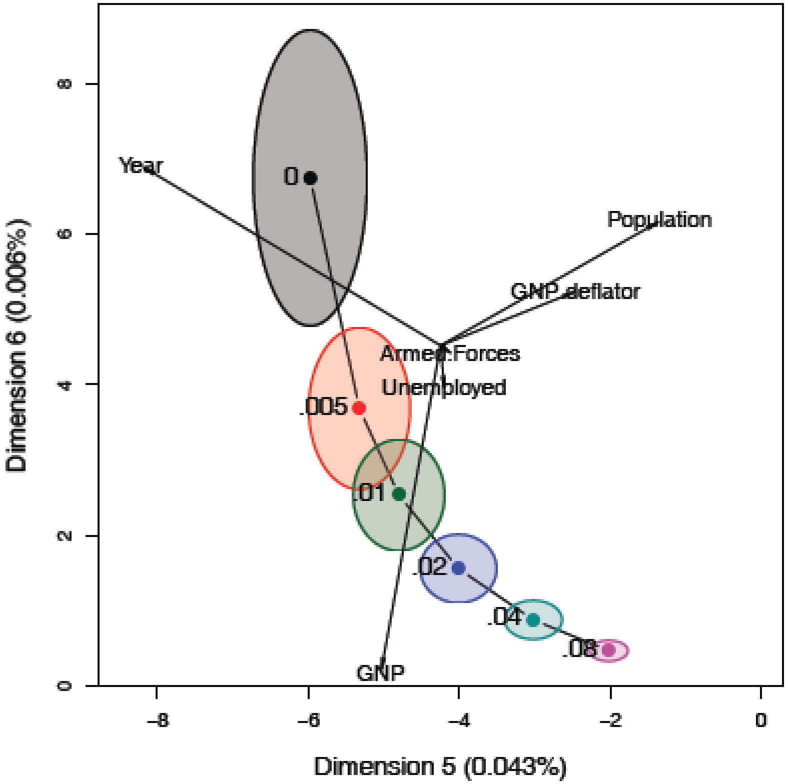
```
> plridge <- pca.ridge(lridge)
> pairs(plridge, col=clr, radius=0.5, diag.cex=3)
```



- The ellipsoids are rotated to the principal axes of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$
- SVD of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$ implies: $\mathcal{E}(\beta, \Sigma) \mapsto \mathcal{E}(\mathbf{V}\beta, \mathbf{V}^{\mathsf{T}}\Sigma\mathbf{V})$
- Transformed ellipsoids have their major/minor axes aligned with coordinate axes.
- It is easy to see that shrinkage occurs only in the space of the smallest eigenvalues

View in PCA space of smallest dimensions: biplot() method

```
> biplot(plridge, col=clr, radius=.5, cex.lab=1.25, prefix="Dimension ")
```



- View the variance ellipsoids in the space of the smallest dimensions
- This is where the greatest shrinkage takes place!
- Variable vectors show how these dimensions relate to the original variables [“biplot”]
- GNP, Year & Pop contribute most to Dim 6

Summary & conclusions

- This presentation has been brought to you by the letter $\mathcal{E}$,

$$\mathcal{E}$$

- It stands for all I have come to appreciate about the deep relationships among:
  - geometry,
  - statistics, and
  - data visualization
- The history of data vis progressed from 1D → 2D → *n*D [1→2→many]
  - The visual discovery of the data ellipse by Galton is the inception of modern statistical methods
  - It was then only a small step from 2D → *n*D for multivariate data vis methods.
- The connections among these are still tools for thought & continue to give new insights