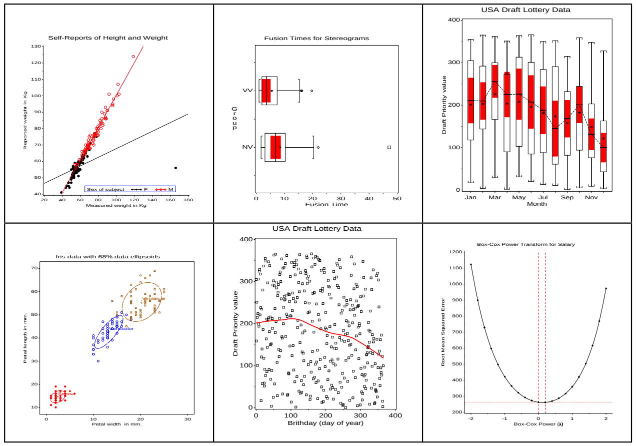
Data Screening



Michael Friendly
Psychology 6140

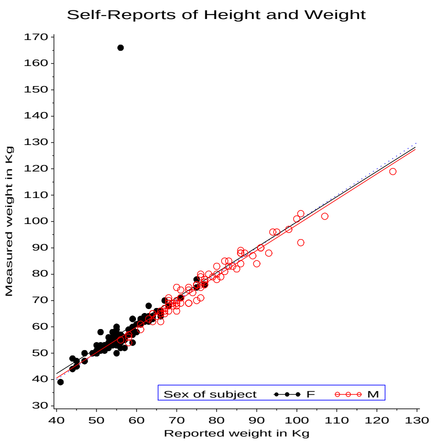
Outline

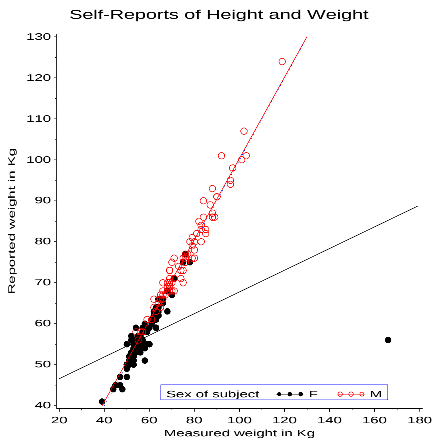
- Part 1: Getting started
  - Failures to screen data
  - Entering and checking raw data
  - Assessing univariate problems
    - Boxplots and outliers
    - Transformations to symmetry
    - Normal probability plots
- Part 2: Assessing bivariate problems
  - Transformations to linearity
  - Dealing with non-constant variance
- Part 3: Multivariate problems
  - Assessing multivariate problems
    - Multivariate normality
    - Multivariate outliers
- SAS macro programs:
  - `http://datavis.ca/sasmac/`

Failures to Screen Data

Data on Self-Reports of height and weight among men and women active in exercise

- Regression of reported weight on measured weight gave very different regressions for men and women
- Plotting the data suggested an answer





Checking variables

- Descriptive statistics checks: verify correct ranges, amount of missing, etc.
  - R - `summary()`
  - SPSS - Frequencies
  - SAS - `PROC UNIVARIATE`
    - Min, Max, # missing
    - Mean, median, std. dev, skewness, etc.
  - Use `plot` option for stem-leaf/boxplot and normal probability plot
  - Use `ID` statement to identify highest/lowest obs.

```
proc univariate plot data=baseball;
   var atbat -- salary ;
   id name;
```

- Consistency checks (e.g., unmarried teen-aged widows?)
  - SPSS - Crosstabs
  - SAS - `PROC FREQ`

```
proc freq;
   tables age * marital;
```

- But: these can generate too much output!

Checking numeric variables - the `DATACHK` macro

- Uses `PROC UNIVARIATE` to extract descriptive stats, high/low obs.
- Formats output to 5 variables/page
- Boxplot of standardized scores to show distribution shape, outliers
- Lists observations with more than `nout` (default: 3) extreme z scores, $|z| >$ `zout` (default: 2)
- Example:

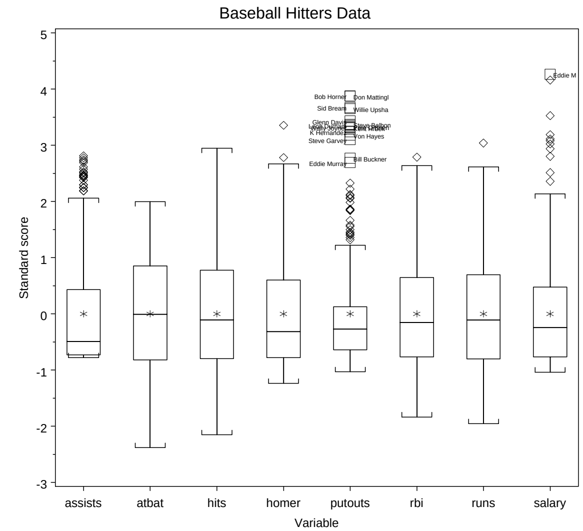
```
  %include data(baseball);
  %datachk(data=baseball, id=name,
var=salary runs hits rbi atbat homer assists putouts);
```

Documentation: `http://datavis.ca/sasmac/datachk.html`

R:

```
data(baseball, package="corrgram")
bb <- scale(baseball)
boxplot(bb)
```


Boxplots of standard scores show the ‘shape’ of each variable, with labels for ’far-out’ observations.




```
Variable              Stat    Value        Extremes Id
 ...
RBI                   N            322          0 Doug Baker
                      Miss           0          0 Mike Schmidt
Runs Batted In        Mean    48.02795          0 Tony Armas
                      Std     26.16689          2 Bob Boone
                      Skew    0.608377
                                              113 Don Mattingly
                                              116 Dave Parker
                                              117 Jose Canseco
                                              121 Joe Carter
                      -------------------------------------------
RUNS                  N            322          0 Mike Schmidt
                      Miss           0          1 Cliff Johnson
Runs                  Mean    50.90994          1 Doug Baker
                      Std      26.0241          1 Tony Armas
                      Skew    0.415779
                                              108 Joe Carter
                                              117 Don Mattingly
                                              119 Kirby Puckett
                                              130 R Henderson
                      -------------------------------------------
SALARY                N            263         68 B Robidoux
                      Miss          59         68 Mike Kingery
Salary (in 1000$)     Mean    535.9658         70 Al Newman
                      Std      451.104         70 Curt Ford
                      Skew    1.589077 *
                                             1975 Don Mattingly
                                             2127 Mike Schmidt
                                             2413 Jim Rice
                                             2460 Eddie Murray
                      -------------------------------------------
```


Sidebar: Using SAS macros

- SAS macros are high-level, general programs consisting of a series of `DATA` steps and `PROC` steps.
- Keyword arguments substitute your data names, variable names, and options for the named macro parameters.
- Use as:

  `%macname(data=dataset, var=variables, ...);`

  e.g.,

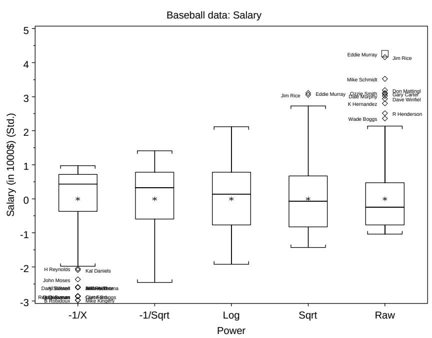
  `%boxplot(data=nations, var=imr, class=region, id=nation);`
- Most arguments have default values (e.g., `data=_last_`)
- All *SSSG* and *VCD* macros have internal and/or online documentation, `http://datavis.ca/sasmac/`
- Macros can be installed in directories *automatically* searched by SAS. Put the following `options` statement in your `AUTOEXEC.SAS` file:

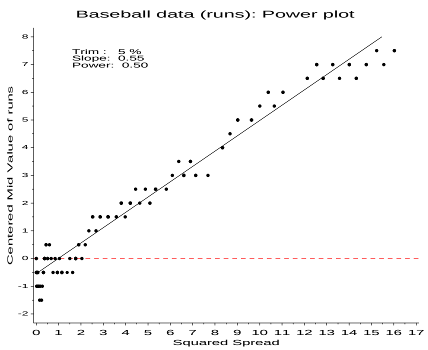
  `options sasautos=('c:\sasuser\macros' sasautos);`

Assessing univariate problems

- Boxplots
- Transformations to symmetry
- Outliers
- Normal probability plots

Note: Normality is **not** required for all variables (e.g., predictors in regression). However, extremely skewed distributions can cause both univariate and bivariate problems.

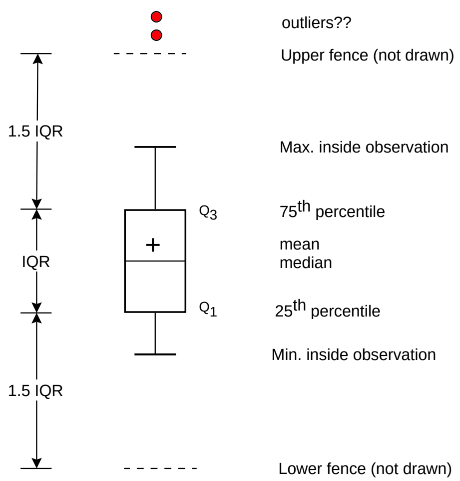




Boxplots

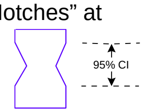
Boxplots provide a *schematic* graphical summary of important features of a distribution, including:

- the center (mean, median)
- the spread of the middle of the data (IQR)
- shape: symmetric? skewed?
- the behavior of the tails
- outliers (plotted individually)



- Notched boxplots for multiple groups: "Notches" at

$$\text{Median} \pm 1.58 \frac{\text{IQR}}{\sqrt{n}}$$



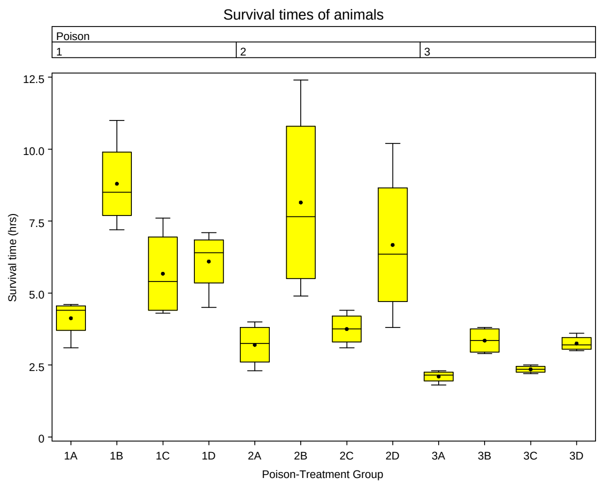
show approximate 95% confidence intervals around the medians. Medians differ if the notches do not overlap (McGill et al., 1978).

Boxplots - ANOVA data

- Boxplots are particularly useful for comparing groups
- ANOVA: Do means differ?
- ANOVA: Assumes equal within-group variance!

Example: Survival times of animals (Box and Cox, 1964)

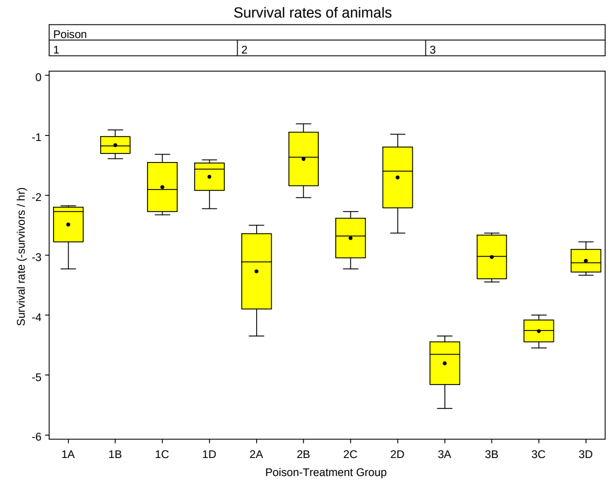
- Animals exposed to one of 3 types of poison
- Given one of 4 treatments
- $\rightarrow 3\times$ 4 design, $n = 4$ per group



- Boxplot shows that variance increases with mean (why?)

Boxplots - ANOVA data

- Methods we will learn today suggest that power transformations, $y \rightarrow y^p$ are often useful.
- These suggest: rate = 1 / time to reduce heterogeneity of variance



Transformations to symmetry

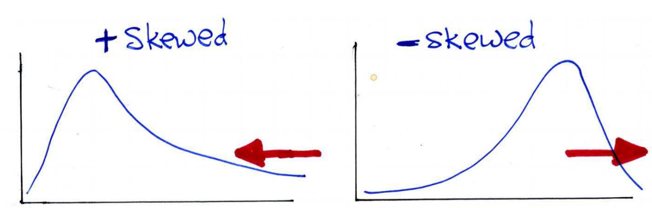
- Transformations have several uses in data analysis, including:
  - making a distribution more symmetric.
  - equalizing variability (spreads) across groups.
  - making the relationship between two variables linear.
- These goals often coincide: a transformation that achieves one goal will *often* help for another (but not *always*).
- Some tools (Friendly, 1991):
  - Understanding the *ladder of powers*.
  - `SYMBOX` macro - boxplots of data transformed to various powers.
  - `SYMPLOT` macro - various plots designed to assess symmetry. POWER plot: line with slope $b \Rightarrow y \rightarrow y^p$, where $p = 1 - b$ (rounded to 0.5).
  - `BOXCOX` macro - for regression model, transform $y \rightarrow y^p$ to minimize MSE (or maximum likelihood); influence plot shows impact of observations on choice of power (Box and Cox, 1964).
  - `BOXGLM` macro - for GLM (anova/regression), transform $y \rightarrow y^p$ to minimize MSE (or max. likelihood)
  - `BOXTID` macro - for regression, transform $x_i \rightarrow x_i^p$ (Box and Tidwell, 1962).

Transformations – Ladder of Powers

- Power transformations are of the form $x \rightarrow x^p$.
- A useful family of transformations is *ladder of powers* (Tukey, 1977), defined as $x \rightarrow t_p(x)$,

$$t_p(x) = \begin{cases} \frac{x^p - 1}{p} & p \neq 0 \\ \log_{10} x & p = 0 \end{cases} \quad (1)$$

- Key ideas:
  - $\log(x)$ plays the role of x^0 in the family.
  - $1/p \rightarrow$ keeps order of x the same for $p < 0$, e.g., $1/x = x^{-1}$.
  - Thinking rule: which direction to go, to compress (←) or expand (→) the upper tail?



- For simplicity, usually use only simple integer and half-integer powers (sometimes, $p = 1/3 \rightarrow \sqrt[3]{x}$)
- You are free to scale the values to keep results simple.

Power	Transformation	Re-expression
3	Cube	x^3 /100
2	Square	x^2 /10
1	NONE (Raw)	x
1/2	Square root	$\sqrt{x}$
0	Log	$\log_{10} x$
-1/2	Reciprocal root	$-10/\sqrt{x}$
-1	Reciprocal	$-100/x$

Ladder of Powers – Properties

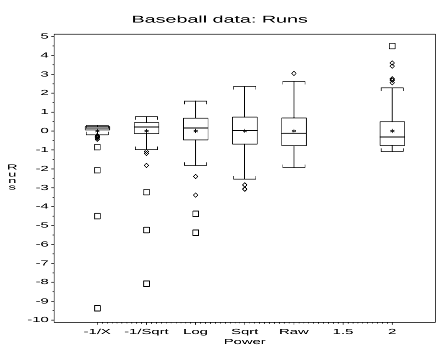
- ***Preserve the order of data values***. Larger data values on the original scale will be larger on the transformed scale. (That's why negative powers have their sign reversed.)
- ***They change the spacing of the data values***. Powers $p < 1$, such as $\sqrt{x}$ and $\log x$ compress values in the upper tail of the distribution relative to low values; powers $p > 1$, such as x^2, have the opposite effect, expanding the spacing of values in the upper end relative to the lower end.
- ***Shape of the distribution changes systematically with*** p. If $\sqrt{x}$ pulls in the upper tail, $\log x$ will do so more strongly, and negative powers will be stronger still.
- ***Requires all*** $x > 0$. If some values are negative, add a constant first, i.e., $x \rightarrow t_p(x + c)$
- Has an effect only if the ***range of*** x ***values is moderately large***.

Ladder of Powers – Example

Baseball data - runs

- SYMBOX macro - transforms a variable to a list of powers, show standardized scores using the BOXPLOT macro

```
%include data(baseball);
title 'Baseball data: Runs';
%symbox(data=baseball, var=Runs, powers =-1 -.5 0 .5 1 2);
```



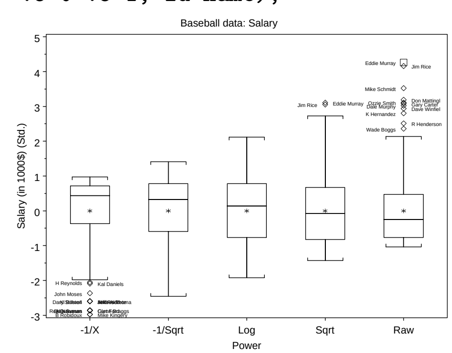
- runs $\rightarrow \sqrt{\text{runs}}$ looks best.

Ladder of Powers – Example

Baseball data - salary

- SYMBOX macro - transforms a variable to a list of powers, show standardized scores using the BOXPLOT macro

```
title 'Baseball data: Salary';
%symbox(data=baseball, var=Salary,
    powers =-1 -.5 0 .5 1, id=name);
```



- salary $\rightarrow \log(\text{salary})$ looks best.

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

Plots for assessing symmetry

Power plot: Mid vs. z^2 plots

- Emerson and Stoto (1982) suggest a variation of the Mid vs. Spread plot, scaled so that a slope, b indicates the power $p = 1 - b$ for a transformation to approximate symmetry.
- In this display, we plot the centered mid value,

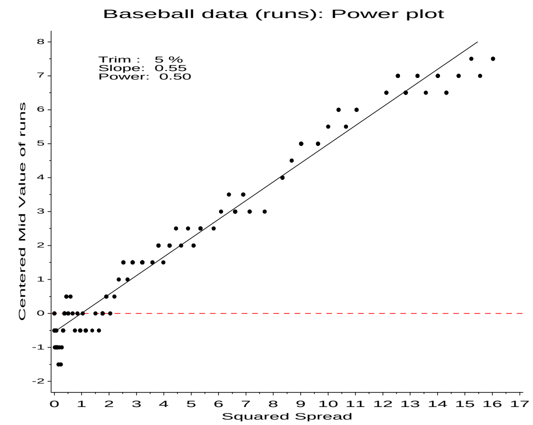
$$\frac{x_{(i)} + x_{(n+1-i)}}{2} - \text{M}$$

against a squared measure of spread,

$$z^2 \equiv \frac{\text{Lower}^2 + \text{Upper}^2}{4M} = \frac{[M - x_{(i)}]^2 + [x_{(n+1-i)} - M]^2}{4M}$$

- SYMPLOT macro - Power plots (`plot=power`). Points should plot as a horizontal line with slope = 0 in a symmetric distribution.

```
title 'Baseball data (runs):  Power plot';
%symplot(data=baseball, var=runs, plot=power);
```

- Symmetry is indicated by a line with slope=0 and intercept=0.
- The SYMPLOT macro rounds $p = 1 - b$ to the nearest half-integer.
- It is often useful to exclude (trim) the highests/lowest 5–10% of observations for automatic diagnosis.

See `http://datavis.ca/sasmac/symplot.html`

Normal probability plots

- Compare observed distribution to some theoretical distribution (e.g., the normal or Gaussian distribution)
- Ordinary histograms not particularly useful for this, because
  - they use arbitrary bins (class intervals)
  - they lose resolution in the tails (where differences are likely)
  - the standard for comparison is a curve
- **Quantile-comparison plots** (Q-Q plots) plot the quantiles of the data against corresponding quantiles in the theoretical distribution, i.e.,

$$x_{(i)} \quad \text{vs.} \quad z_i = \Phi^{-1}(p_i)$$

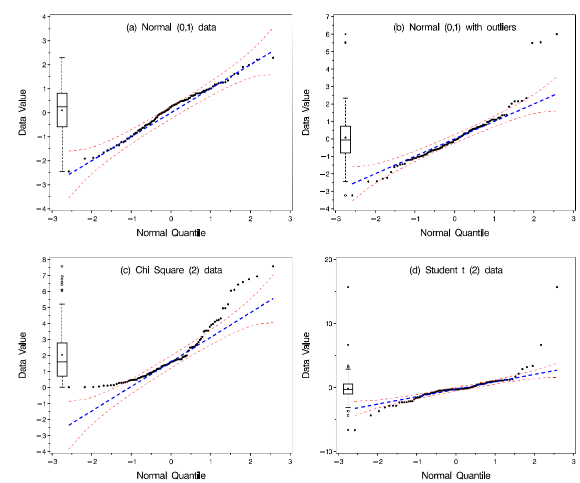
where $x_{(i)}$ is the i-th *sorted* data value, having a proportion, $p_i = \frac{i - 1/2}{n}$ of the observations below it, and $z_i = \Phi^{-1}(p_i)$ is the corresponding quantile in the normal distribution.

- When the data follows the normal distribution, the points in such a plot will follow a straight line with slope = 1.
- Departures from the line shows *how* the data differ from the assumed distribution.

Normal probability plots

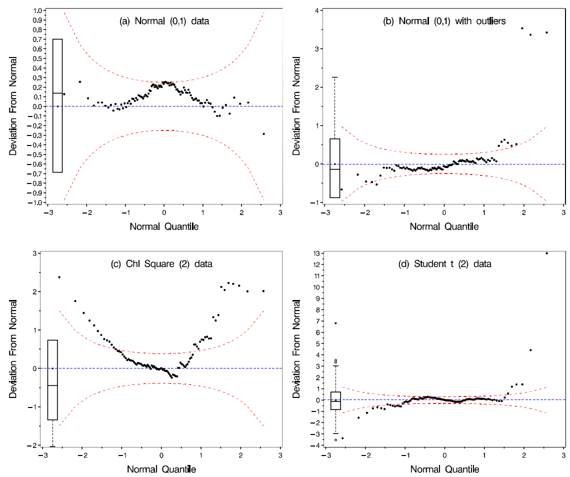
Patterns of deviation for Normal Q-Q plots:

- ***Postive (negative) skewed***: Both tails above (below) the comparison line
- ***Heavy tailed***: Lower tail below, upper tail above the comparison line



Normal probability plots: detrended

- De-trended plots show the deviations more clearly
- Plot $x_{(i)} - z_i$ vs. z_i.

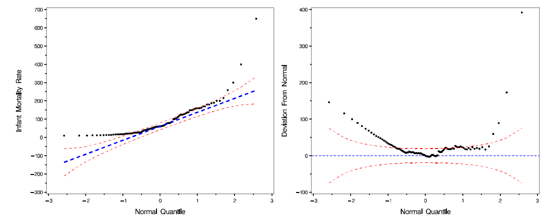


Normal probability plots: confidence bands

- Points in a Q-Q plot are not equally variable—observations in the tails vary most for normal.
- Calculate estimated standard error, $\hat{s}(z_i)$, of the ordinate z_i and plot curves showing the interval $z_i \pm 2\,\hat{s}(z_i)$ to give approximate 95% confidence intervals. (Chambers et al. (1983) provide formulas.)

$$\hat{s}(z_i) = \frac{\hat{\sigma}}{f(z_i)} \sqrt{\frac{p_i\,(1-p_i)}{n}}$$

- Confidence bands help to judge how well the data follow the assumed distribution



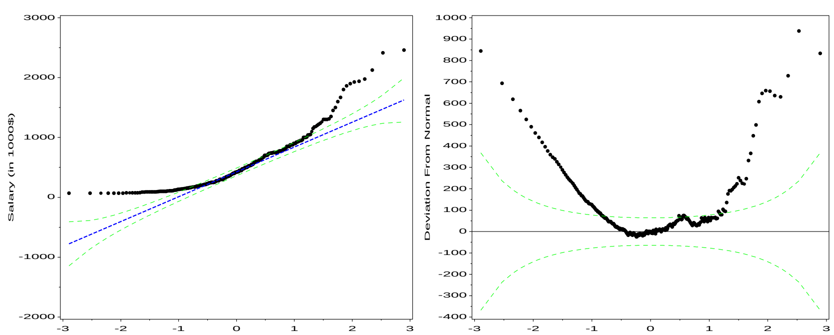
See `http://datavis.ca/sasmac/nqplot.html`

Normal probability plots

Baseball data - salary

- Raw data

```
%nqplot(data=baseball, var=salary);
```

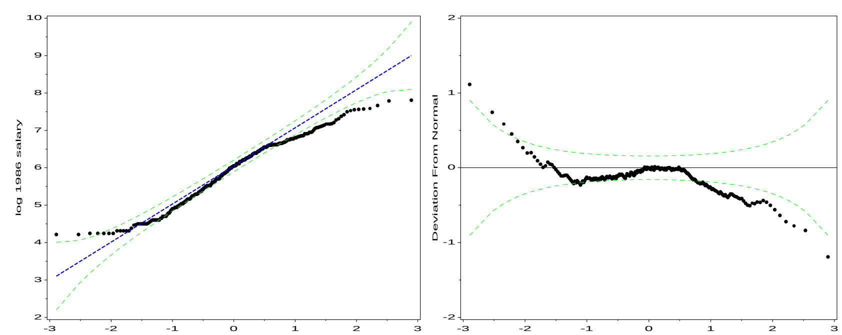


R: use `qqPlot()` from the `car` package

```
data(baseball, package="corrgram")
car::qqPlot(Baseball$Salary)
```


- Try $\log$ salary — better, but not perfect (who is?)

```
data baseball;
   set baseball;
   label logsal = 'log 1986 salary';
   logsal = log(salary);
%nqplot(data=baseball, var=logsal);
```



R:

```
car::qqPlot(baseball$logSal)
```

Part 2: Assessing bivariate problems

- Transformations to linearity
  - The “Arrow Rule” and the double ladder of powers
  - Box-Cox transformation for y (`BOXCOX` macro, `BOXGLM` macro)
  - Box-Tidwell transformation for Xs (`BOXTID` macro)
- Dealing with heteroscedasticity (non-constant error variance)
  - Spread vs. level plots (`SPRDPLOT` macro)

Transformations to linearity

Brain weight and body weight of mammals:

- Marginal boxplots show that both variables are highly skewed
- Most points bunched up at origin
- Relation is strongly non-linear
- Log transform removes both problems

Transformations to linearity

- If y is a ***response*** (“dependent”) and x is a predictor, we often want to fit

$$y = f(x) + \text{residual}$$

- Generally we prefer a “simple” $f(x)$, like a linear function, $y = a + b\,x + \text{residual}$.
- If the relation between y and x is substantially non-linear, we have two choices:

Bend the model: Try fitting a quadratic, cubic, or other polynomial (easy: linear in parameters), or else a non-linear model, e.g., $y = a \exp(bx)$ (harder).

Unbend the data: Transform either $y \rightarrow y'$, or $x \rightarrow x'$ (or both), so that relation is linear,

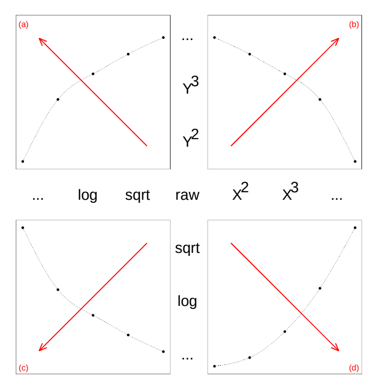
$$y' = a + b\,x' + \text{residual}$$

- Ladder of powers and Tukey’s “arrow rule” indicate which direction to go.

Transformations to linearity: Arrow Rule

Tukey’s arrow rule and the double ladder of powers:

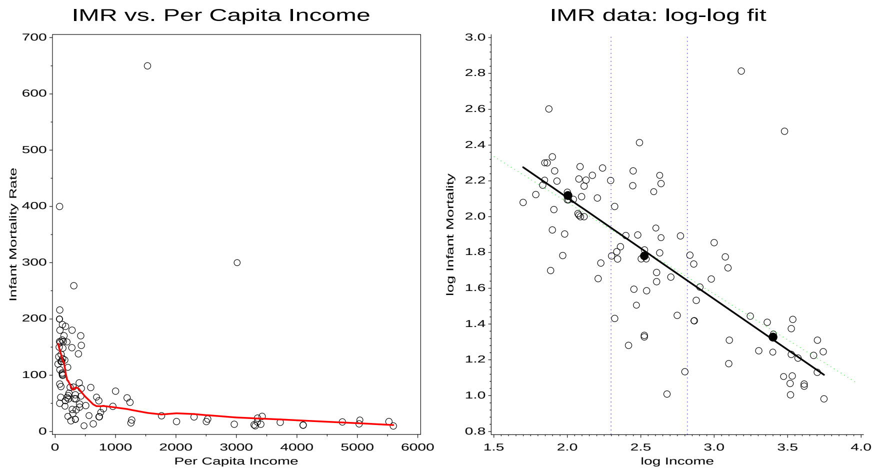
- Draw an arrow in the direction of the “bulge”.
- The arrow points in the direction to move along the ladder of powers for x or y (or both).



Transformations to linearity

Infant mortality rate and per-capita income

- Arrow points toward lower powers of x and/or y
- Ratio of slopes suggest $\log x, \log y$



Box-Cox Transformations

- Another way to select an “optimal” transformation of y in regression is to add a parameter for the power to the model,

$$y^{(\lambda)} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

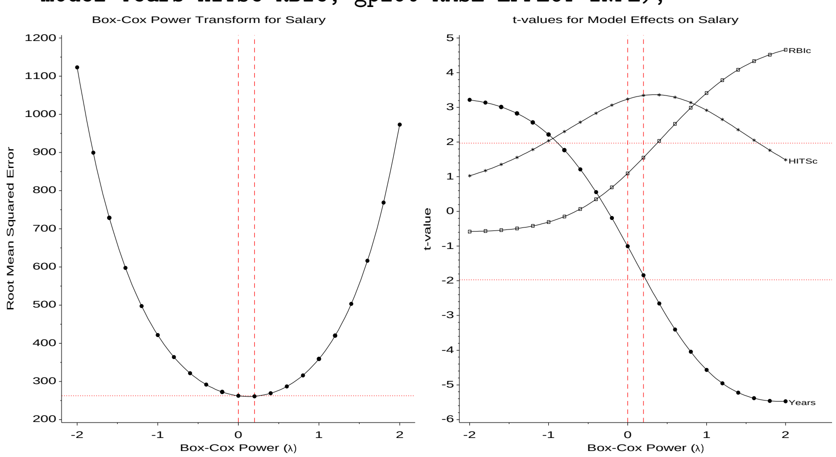
where λ is another parameter, the power in (the ‘ladder’)

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log\ y, & \lambda = 0 \end{cases}$$

- Box and Cox (1964) proposed a maximum likelihood procedure to estimate the power (λ) along with the regression coefficients ($\boldsymbol{\beta}$).
- This is equivalent to minimizing $\sqrt{MSE}$ over choices of λ. $\Rightarrow$ fit the model for a range of λ (-2 to +2, say)
- The maximum likelihood method also provides a 95% confidence interval for λ.
- Can also plot the partial t or F statistic for each regressor vs. λ.

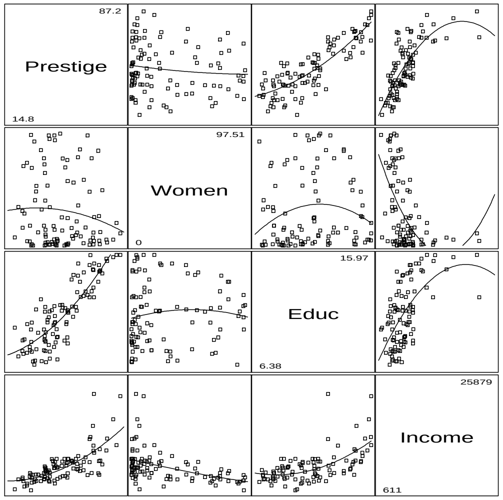
- Baseball data: predicting Salary from Years, RBIc, HITSc.
  - CI (λ) includes $\lambda = 0 \rightarrow \log$(Salary)
  - Effects plot shows t statistic for each regressor
- The boxcox macro provides the RMSE, EFFECTS, and INFL plots:

```
title 'Box-Cox transformation for Baseball salary';
%include data(baseball);
%boxcox(data=baseball, id=name, resp=Salary,
   model=Years HITSc RBIc, gplot=RMSE EFFECT INFL);
```



Transformations of predictors

- Another statistical method: Box-Tidwell transformation– like Box-Cox, but for predictors in regression models
- In any correlational analysis (e.g., regression, factor analysis) we can get a simple overview of the relations by
  - Plotting all pairs of variables together (`scatmat` macro)
  - Drawing a *quadratic* regression curve for each pair `%scatmat(...,interp=rq)`.
  - “curves” will be straight when the relations are linear.
  - (lowess fits are better, but more computationally intensive.)
- Simple method: Canadian occupational prestige: %women, income, education



- → Prestige non-linear w.r.t. Educ and Income
- smoothed loess curves are more useful (but computationally harder)

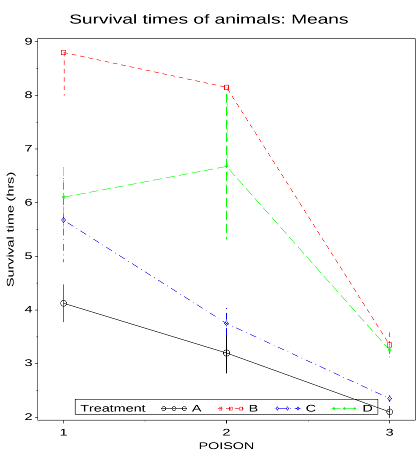
Dealing with heteroscedasticity

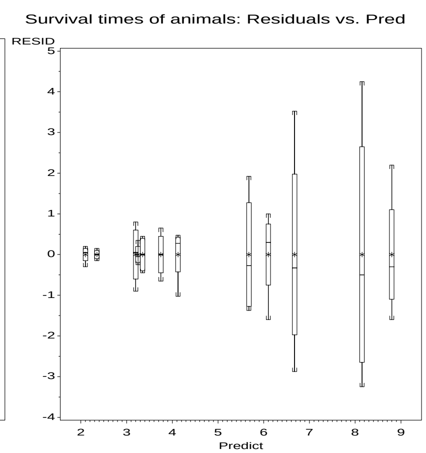
- Classical linear models (ANOVA, regression) assume constant (residual) variance

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \quad \text{Var}(\epsilon) = \sigma^2$$

- ANOVA: examine std. dev. of residuals by groups
  - Plot means $\pm$ 1 std. error (`meanplot` macro)
  - Boxplots of residuals vs. predicted (`boxplot` macro)

```
%meanplot(data=animals, class=poison treatmt,
   response=time);

proc glm data=animals;
   class poison treatmt;
   model time = poison | treatmt;
   output out=results p=predict r=resid;
%boxplot(data=results, class=Predict, var=resid);
```



- Both plots show greater variance associated with longer survival time.

Dealing with heteroscedasticity: Spread-Level plots

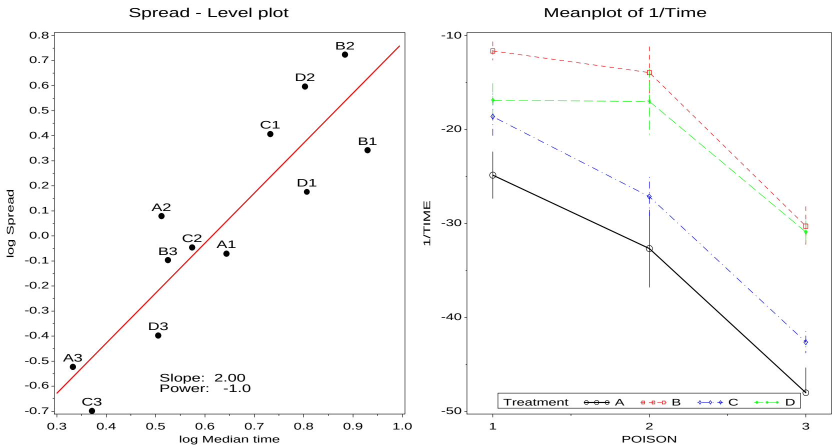
Spread vs. level plots (the `sprdplot` macro)

- Plot log(spread) vs. log(level) e.g., log(IQR) vs. log(Median)
- If a linear relation exists, with slope b, transform $y \rightarrow y^p$, with $p = 1 - b$.

```
%sprdplot(data=animals, class=poison treatmt, var=time);
%meanplot(data=animals, class=poison treatmt,
   response=t_time);
```

- In R: use `car::spreadLevelPlot()`

```
spreadLevelPlot(time ~ poison + treatment, data=animals)
```



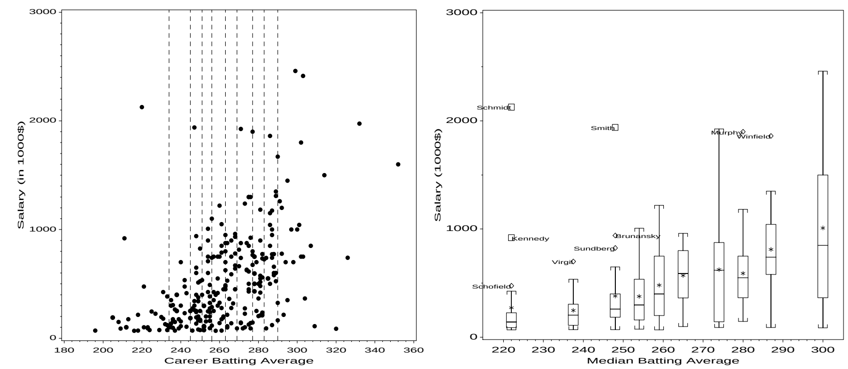
- The plot suggests transforming Time $\longrightarrow$ 1/Time.
- 1/Time also reduces apparent interaction of Poison * Treatment

Dealing with heteroscedasticity

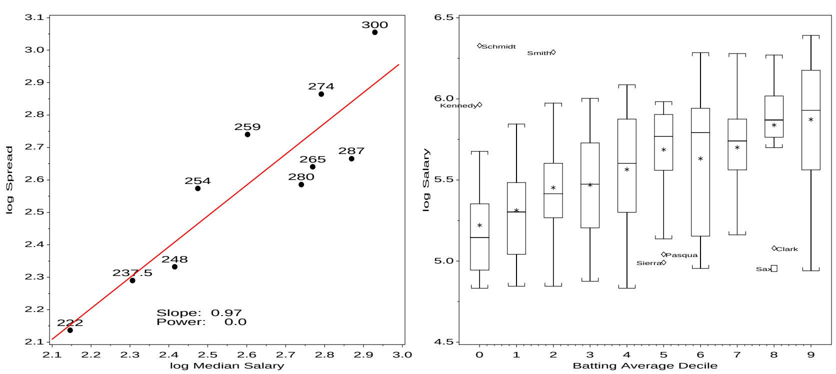
Regression data

- Divide an x variable into ordered groups (e.g., deciles)

```
proc rank data=baseball out=grouped groups=10;
   var batavgc;
   ranks decile;
```



- Use Spread vs. level plot on grouped x



- log Salary is again indicated

Part 3: Multivariate problems

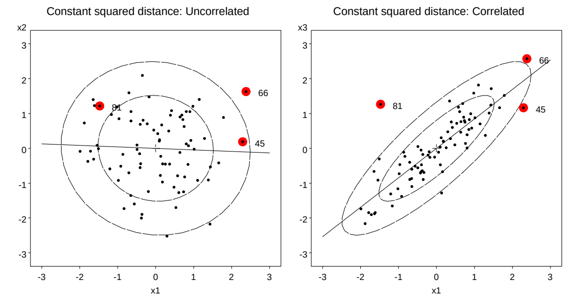
- Assessing multivariate problems
  - Multivariate normality
  - Outliers: univariate, bivariate, multivariate
  - Robust outlier detection

Multivariate normality

- Some multivariate statistical methods assume that all measures are jointly multivariate normal.
  - e.g., Factor analysis, discriminant analysis, MANOVA (for Y variables)
  - Regression:
    - Usually *not* required for predictors
    - *Is* required for multivariate MRA (Y variables)
  - Better to check for (multivariate) normality of *residuals*
- Statistical measures
  - Univariate: Skewness, kurtosis $\longrightarrow$ Shapiro-Wilk test
  - Multivariate: Mardia's multivariate skewness, kurtosis
  - But: these are sensitive to small deviations from strict (multi-) normality.
  - Don't worry about small to moderate departures

Multivariate normality: Chi-square QQ plot

- Graphical method: Chi-square QQ plot
  - 1 variable: $z_i = (x_i - \bar{x})/s \sim \mathcal{N}(0,1)$, or, $z_i^2 = \frac{(x_i - \bar{x})^2}{s^2} \sim \chi^2_{(1)}$.
  - 2 variables: If uncorrelated, squared distance of (x_{i1}, x_{i2}) from the mean is $D_i^2 = z_{i1}^2 + z_{i2}^2 \sim \chi^2_{(2)}$.



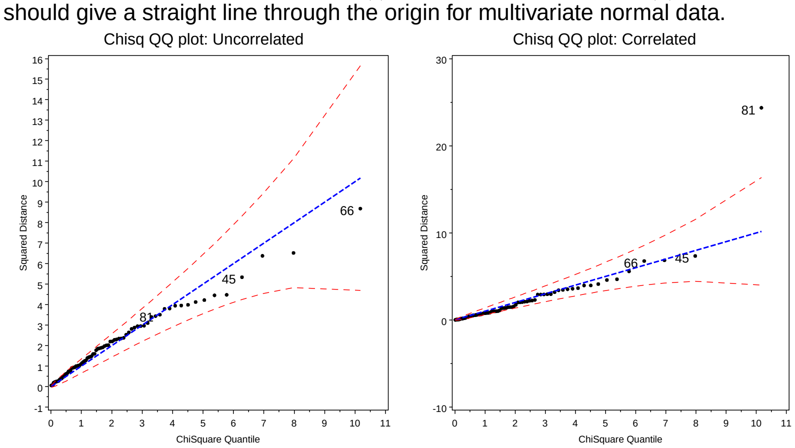
  - p variables: Calculate generalized (Mahalanobis) squared distance, D_i^2 of each observation $\boldsymbol{x}_i$ from the mean vector,

$$D_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\mathsf{T} \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \sim \chi^2_{(p)}$$

    where $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix.

Multivariate normality: Chi-square QQ plot

- $\Rightarrow$ QQ plot of *ordered* distances, $D^2_{(i)}$, against corresponding $\chi^2_{(p)}$ quantiles should give a straight line through the origin for multivariate normal data.



Multivariate normality: Chi-square QQ plot

Computation:

- The D_i^2 can be easily calculated by transforming the data to *standardized* principal component scores, i.e., $D_i^2 = \sum_j^p z_{ij}^2$:

```
proc princomp STD out=PC;
   var X1-X10;
data pc;
   set pc;
   Dsq = USS(of PRIN1-PRIN10);
```

- The `multnorm` macro calculates univariate and multivariate normality tests, and produces the Chi-square QQ plot.
  - Confidence bands for the distribution help to judge how close the D_i^2 are to a χ^2 distribution.
  - But: outliers can make the graphical test lest sensitive.
- R: `mahalanobis()` for D^2; `heplots::cqplot()` for plots

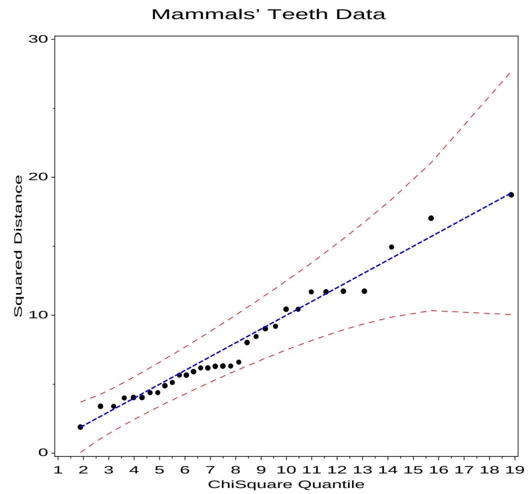
Example: Mammals teeth: number of incisors, canines, molars, etc. in 32 species

```
%include data(teeth);
%multnorm(data=teeth, var=v1-v8, id=mammal);
```

Var	Test	Skewness	Kurtosis	Test Statistic	p-value
V1	Shapiro-Wilk	-0.6993	-0.8885	0.790	0.00001
V2	Shapiro-Wilk	-0.3040	-1.0806	0.829	0.00008
V3	Shapiro-Wilk	-1.0216	-1.0246	0.560	0.00000
V4	Shapiro-Wilk	-0.5421	-1.8244	0.608	0.00000
V5	Shapiro-Wilk	-0.8124	0.2587	0.863	0.00060
V6	Shapiro-Wilk	-0.5955	-0.2693	0.883	0.00206
V7	Shapiro-Wilk	-0.4687	-1.7688	0.671	0.00000
V8	Shapiro-Wilk	-0.9541	-0.5410	0.702	0.00000
All	Mardia Skew	40.7550	.	242.640	0.00000
All	Mardia Kurt	.	81.1770	0.263	0.79241

- All test statistics indicate substantial deviation from univariate and multivariate normality
- QQ plot does not reveal anything strange. Why?

In R: `mardiaTest()` and others in the MVN package

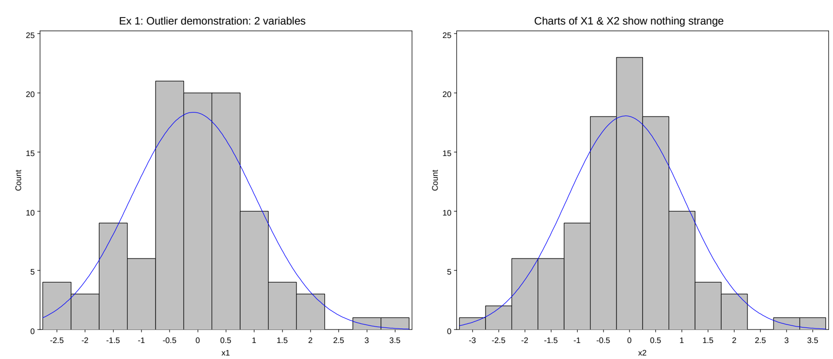


Outliers

- Different kinds of outliers: univariate, bivariate, multivariate, or just observations which don't fit your model (large residuals)
- Univariate outliers:
  - Typical analysis: Examine standardized scores $z_i = (x_i - \bar{x})/s$, for $|z_i| > \pm 2$ (1.96: $p < 0.05$)
  - But: outliers will shift the mean, inflate the std. dev., making obs. look less outlying!
  - Better: Boxplot uses inner fences– quartiles $\pm 1.5 IQR$, $(p < 0.05)$, outer fences– quartiles $\pm 3 IQR$, $(p < 0.001)$.
  - `datachk` macro gives a brief summary for a collection of variables

Outliers

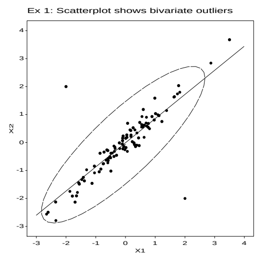
- Univariate checks are useful, but not always sufficient: Can you spot the outliers?



Bivariate outliers

- Bivariate plots can reveal— bivariate outliers!

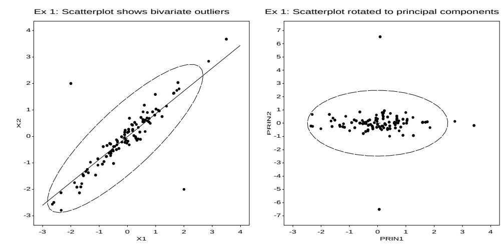
```
data outlier1;
   do i = 1 to 100;
      x1 = normal(33445);          * Correlated;
      x2 = x1 + normal(22345)/4;   * bivariate normal;
      output;
      end;
   *-- Generate two additional obs: outliers;
   x1 = 2; x2 = -2; output;
   x1 =-2; x2 =  2; output;
```



- But, *only* bivariate outliers
- Bivariate plot suggests rotation to principal components

Multivariate outliers

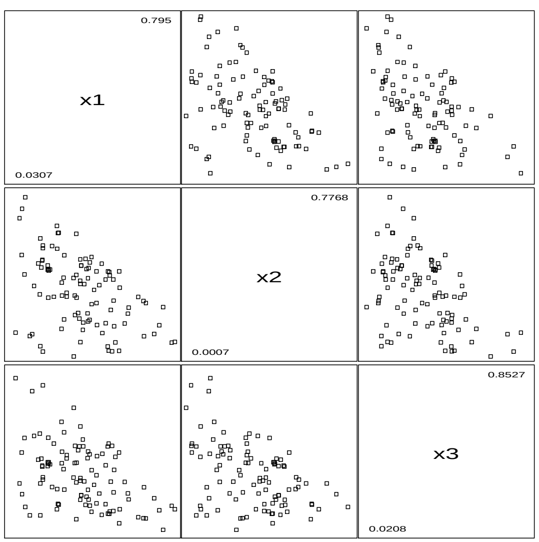
- Transforming variables to principal components:
  - Principal components rotate the cloud of points to new (orthogonal) axes.
  - PRIN1 has greatest variance, PRINp smallest variance
  - Outliers will usually appear as extreme values on the *last* principal component.




```
proc princomp std noprint data=outlier1 out=prin;
   var x1-x2;
title 'Ex 1: Scatterplot rotated to principal components';
%contour( data=prin, y=prin2, x=prin1, pvalue=.95);
```


Multivariate outliers

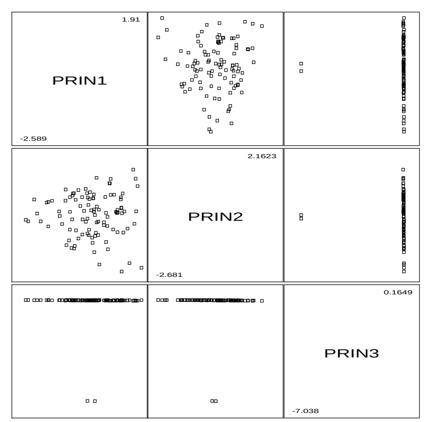
- With 3 or more variables, bivariate plots may show nothing strange.



Multivariate outliers

- Again, outliers show up clearly on the last PC

```
proc princomp std noprint data=outlier2 out=prin;
   var x1-x3;
%scatmat(data=prin, var=prin1-prin3, symbols=square);
```



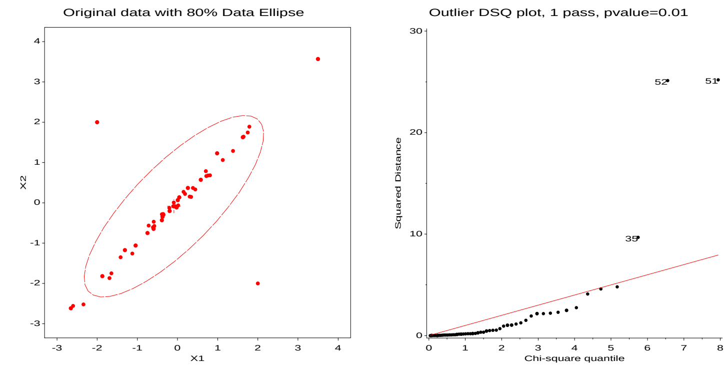
Robust Outlier Detection

- The χ^2 plot for multivariate normality is not resistant to the effects of outliers.
- A few discrepant observations affect the mean vector, $\bar{\boldsymbol{x}}$, and—worse—the variance-covariance matrix, $\boldsymbol{S}$.
- Inflating $\boldsymbol{S} \rightarrow$ decreases D^2: extreme obs. look less discrepant!
- One simple solution is to use ***multivariate trimming*** (Gnanadesikan and Kettenring, 1972) to calculate D^2 values not affected by potential outliers:
  1. Calculate $D^2_{(i)}$ values
  2. Find prob$_i$ $= \Pr(\chi^2_p > D^2_{(i)})$
  3. Set weight$_i$ = 0 for any observation with prob$_i$ $< \alpha$.
  4. Repeat steps 1–3.
- State-of-art ("high breakdown bounds") methods now available in R:
  - `cqplot()` in heplots package
  - robust package; mvoutlier package, ...
  - robust linear and generalized linear models

`outlier` macro

- The `outlier` macro
  - performs 1 or more passes of multivariate trimming,
  - produces a χ^2 QQ plot.

```
title 'Original data with 80% Data Ellipse';
%contour(data=outlier1, y=x2, x=x1, pvalue=.80);

title 'Outlier DSQ plot, 1 pass, pvalue=0.01';
%outlier(data=outlier1, var=x1-x2, id=sub, out=chiplot,
   passes=1, pvalue=.01);
```

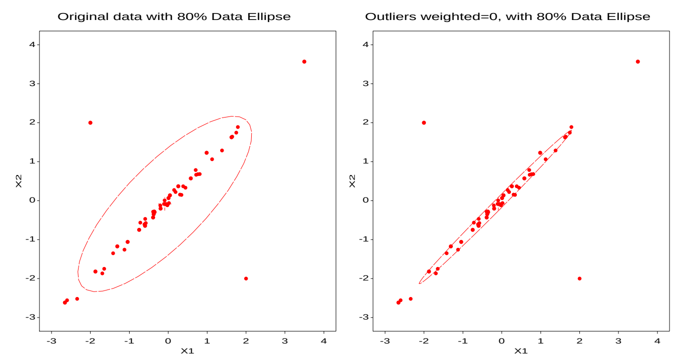



```
          Outlier DSQ plot, 1 pass, pvalue=0.01
Observations trimmed in calculating Mahalanobis distance

      _PASS_    _CASE_      DSQ        PROB

        1         35       9.6729    .0079353
                  51      25.2015    .0000034   *
                  52      25.1222    .0000035   *
```

See: `datavis.ca/sasmac/outlier.html`

`outlier` macro

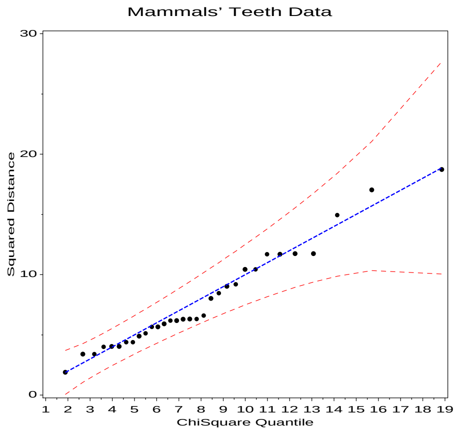
- Comparing data ellipse for original data and weighted data shows the effect of multivariate trimming



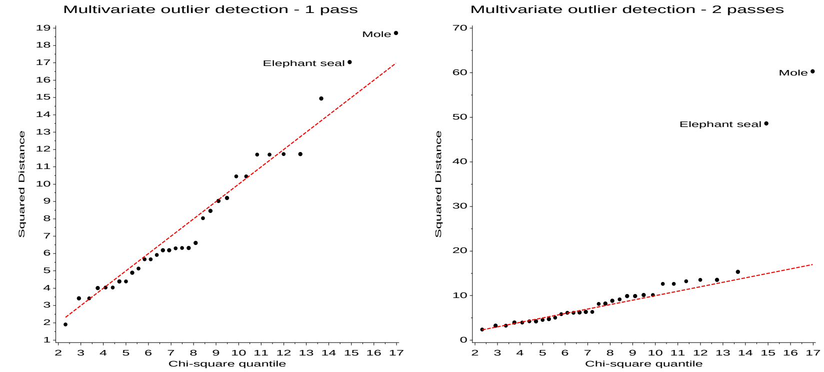

```
title 'Original data with 80% Data Ellipse';
%contour(data=outlier1, y=x2, x=x1, pvalue=.80);

title 'Outliers weighted=0, with 80% Data Ellipse';
%contour(data=chiplot,  y=x2, x=x1, weight=_weight_,
   pvalue=.80);
```

Multivariate outliers: Mammals teeth

- Multivariate normality QQ plot (no trimming) looked OK:



- Effect of multivariate trimming: D^2 increases for outliers



PASS	MAMMAL	_CASE_	DSQ	PROB
1	Mole	2	18.7217	0.016421
	Elephant seal	28	17.0421	0.029674
2	Mole	2	60.3055	0.000000
	Elephant seal	28	48.6327	0.000000

Multivariate outliers: Practical issues

- 2 passes usually sufficient; more obs. may be trimmed in later passes.
- An effective, but *ad hoc* procedure: No hypothesis tests.
- Results of any automatic procedure must be tempered by substantitive knowledge.
- Which obs. are trimmed depends on on the p-value used (e.g., Mammals teeth: Racoon trimmed at `pvalue=0.07`).
- The `outlier` macro uses `pvalue=0.05` by default. A more conservative p-value (e.g., $p < 0.001$) may be more appropriate.
- "OK, I've got outliers." What to do?
  - Answer depends on the context and the analysis.
  - Generally, prefer to remove only probable errors or truly extreme outliers.
  - Sensitivity test: Do analysis with and without. Do the conclusions or main results change?
  - Consider a more robust model fitting method (retain, but down-weight outliers), e.g., `robust` macro, `robmlm()` in heplotspackage.

References

Box, G. E. P. and Cox, D. R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26:211–252, 1964.

Box, G. E. P. and Tidwell, P. W. Transformation of the independent variables. *Technometrics*, 4: 531–550, 1962.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983.

Emerson, J. D. and Stoto, M. A. Exploratory methods for choosing power transformations. *Journal of the American Statistical Association*, 77:103–108, 1982.

Friendly, M. *SAS System for Statistical Graphics*. SAS Institute, Cary, NC, 1st edition, 1991.

Gnanadesikan, R. and Kettenring, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.

McGill, R., Tukey, J. W., and Larsen, W. Variations of box plots. *The American Statistician*, 32: 12–16, 1978.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.