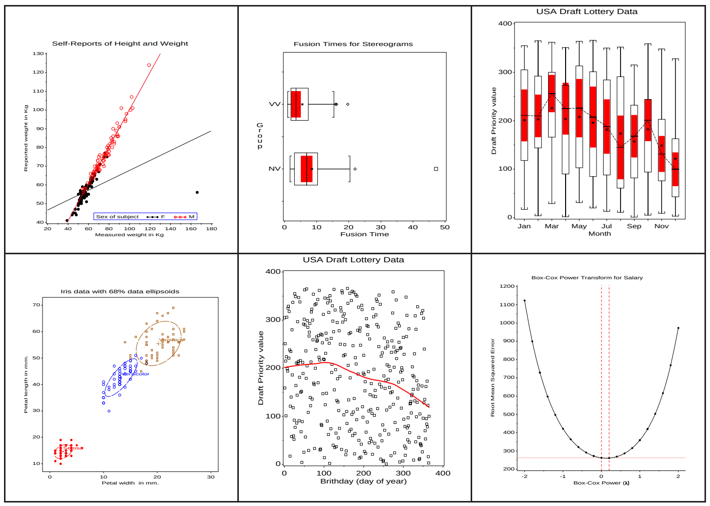
Data Screening: Part 2



Michael Friendly

York University

SCS Short Course

October, 2004

Part 2: Assessing bivariate problems

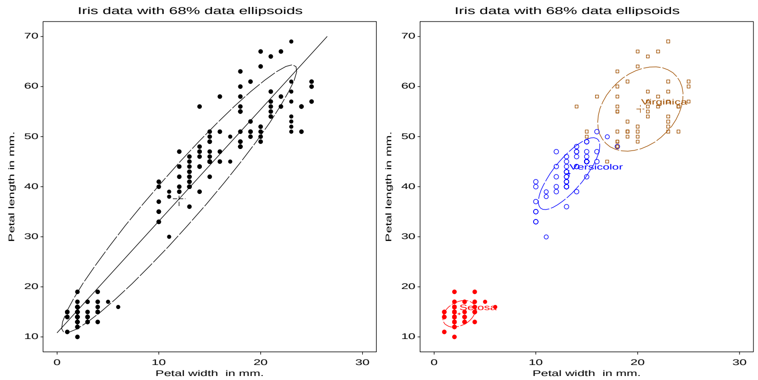
- Enhanced scatterplots
  - Data ellipse: Visualizing bivariate spread and correlation
  - Group differences: total sample vs. within sample analysis
  - Smoothing relations (`LOWESS` macro and `PROC LOESS`)
  - Plotting discrete data
- Transformations to linearity
  - Resistant lines and the double 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)

Data Ellipse

Scatterplots can be enhanced by adding an elliptical confidence region (the ***data ellipse***) around the mean

- Shows the variability of each variable, correlation, and regression line (assuming bivariate normality).
- A 50% data ellipse is analogous to the central box in a boxplot.
- A 67.7% data ellipse is analogous $\bar{X} \pm 1$std. dev.
- The `CONTOUR` macro produces plots with a data ellipse

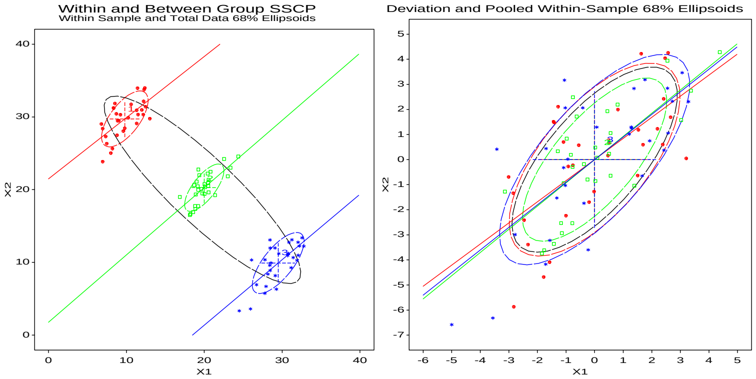
```
%include data(iris);
%contour(data=iris, x=petalwid, y=petallen);  /*total*/
%contour(data=iris, x=petalwid, y=petallen,   /*within*/
   group=species);
```



See http://www.math.yorku.ca/SCS/sssg/contour.html

Total Sample vs. Within Sample Analysis

- In any correlational analysis (e.g., regression, factor analysis) with multiple groups, differences among group *means* can affect correlations between variables
- If group means differ substantially, better to (a) include group as a factor, or (b) subtract group means before calculating correlations ('within-group correlations')



- Correlations: Total and Within group

```
Total sample    Group 1    Group 2    Group 3
  -0.88591      0.65414    0.71811    0.71464
```

Total Sample vs. Within Sample Analysis

- Total-sample correlations

```
proc corr data=groups;
   var x1 x2;
-->
Total sample:    -0.88591
```

- Within-sample correlations

```
proc corr data=groups;
   var x1 x2;
   by group;
-->
Group 1    Group 2     Group 3
0.65414    0.71811     0.71464
```

- Pooled within-sample correlations (from $X_{ij} \longrightarrow X_{ij} - \bar{X}_{.j}$)

```
proc standard data=groups out=dev m=0;
   var x1 x2;
   by group;

proc corr data=dev;
   var x1 x2;
-->
Within sample:    0.69480
```


Total Sample vs. Within Sample Analysis

What to do?

- Do the means differ substantially over grouping variables (ANOVA or MANOVA)?

```
proc glm data=groups;
   class group;
   model x1 x2 = group/ nouni;
   manova h=group;
```

- Regression: Yes $\longrightarrow$ include GROUP as a factor.

```
proc glm data=groups;
   class group;
   model y = group x1 x2;
```

- Do the variance-covariance (corelation) matrices differ over groups?

```
proc discrim data=groups pool=test;
   class group;
   var x1 x2;
-->
Test Chi-Square Value =     4.506765
with      6 DF      Prob > Chi-Sq = 0.6084

The chi-square value is not significant at the 0.05 level,
a pooled covariance matrix will be used...
```

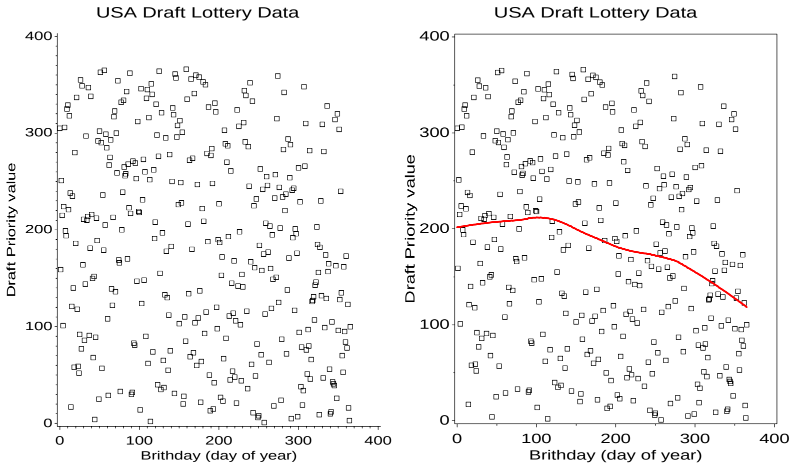
- Factor analysis: No $\longrightarrow$ use pooled within-group correlation matrix.

- Factor analysis: Yes $\longrightarrow$ separate analyses by group (or include dummy vars for group).

Smoothing

Example: 1970 USA Draft Lottery

- Our eyes can often see patterns not easily captured in numbers.
- Sometimes relationships may be too weak to see the trend in a scatterplot.
- Drawing a smoothed curve helps show the trend.
- A general and useful technique is **LO**cally **WE**ighted **S**catterplot **S**moothing ('LOWESS'), a form of non-parametric regression.




```
%lowess(data=draftusa, x=birthday, y=priority,
   symbol=square, step=5);
```

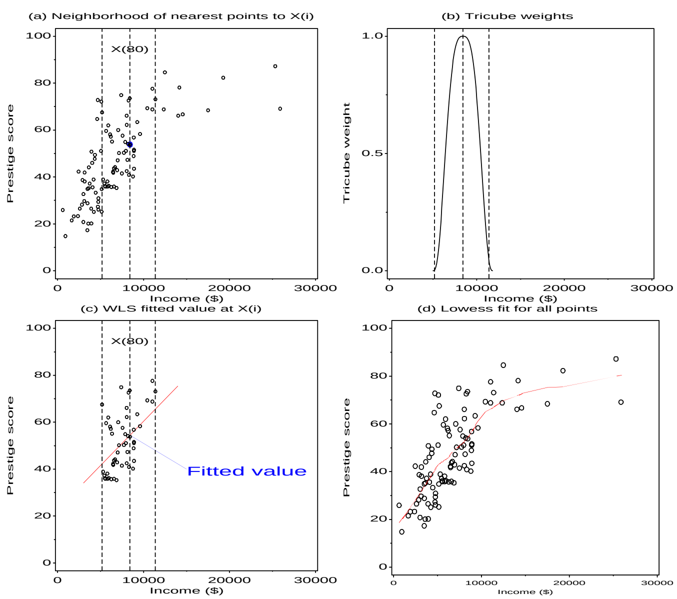

Lowess Smoothing

- Finds a smoothed fitted value, $\hat{y}_i$, for *each* x_i by fitting a weighted regression (linear, quadratic) to points in the 'neighborhood' of x_i.
- Neighborhood depends on a ***smoothing parameter***, f, $0 < f \leq 1$, the fraction of the data points to be considered in the calculation of $\hat{y}_i$.
- Points closest to x_i receive the greatest weight. Only the $r = [fn]$ points closest to x_i have non-zero weights.
- Increasing f makes the fitted curve smoother; decreasing f lets the curve follow the data more closely.
- A "robustness step" calculates residuals, $(y_i - \hat{y}_i)$, down-weights observations $\sim$ squared residual, and re-computes the smoothed values using adjusted weights.
- SAS:
  - `LOWESS` macro calculates the lowess smooth, and plots y vs. x with the smoothed curve.
  - SAS V7/V8: `LOWESS` macro uses `PROC LOESS` for the calculations (fast). SAS 6.12: uses `PROC IML`.
  - See `http://www.math.yorku.ca/SCS/vcd/lowess.html`

Lowess Smoothing

Canadian occupational prestige vs. Income

- $r = [fn]$ nearest neighbors to x_i
- tricube function gives weights for WLS



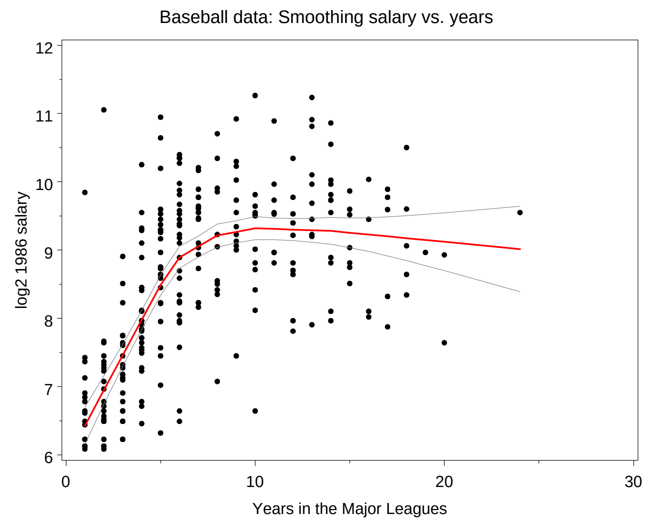
- WLS gives smoothed (fitted) y_i at x_i
- Repeat for all points

Lowess Smoothing

Example: Baseball data – Salary vs. Years

- All other analyses suggested transforming Salary to log(Salary)
- Smoothing the relation between log(Salary) and Years suggested "years up to 7" (players become free agents)

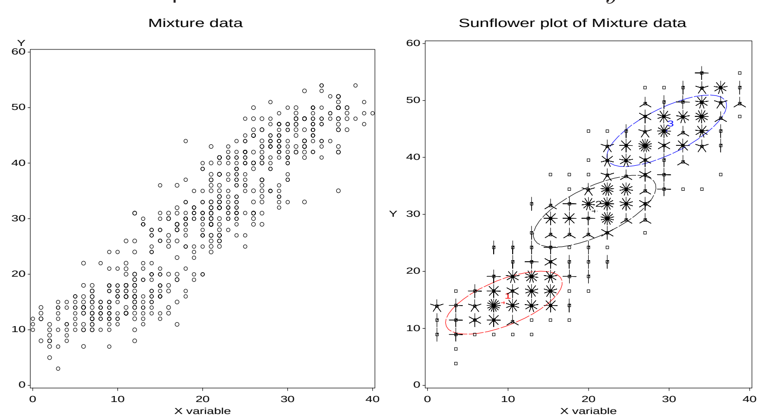
$$\log(\text{Salary}) \sim \text{years}^\star = \min(\text{years}, 7)$$




```
%lowess(data=baseball, y=logsal, x=years, id=name,
   clm=0.05, f=0.67);
```

Plotting discrete data

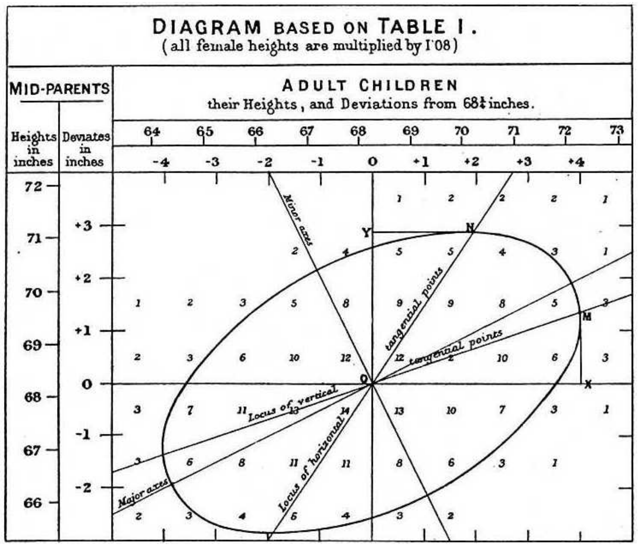
- When the x or y variables are discrete (or for large data sets), it may be difficult to see patterns because of ***overplotting***.
- Two solutions:
  - Sunflower plots: bin data into (x, y) cells, plot using a 'sunflower symbol' showing cell frequency.
  - Jittering: add a small random quantity to each point to reduce overplotting.
- Sunflower example: Mixture data: 3 bivariate normals with $\bar{y} \sim \bar{x}$



See: www.math.yorku.ca/SCS/sasmac/sunplot.html

Example: Galton's data

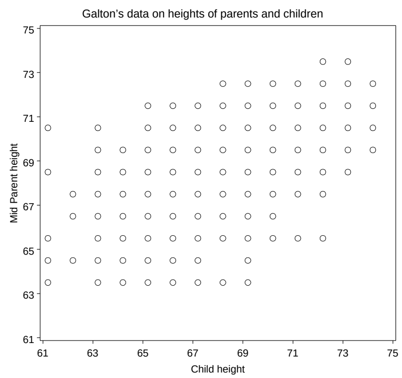
- Francis Galton, in his studies on heredity, collected data on characteristics of parents and their children
- By "inspection" of his graphs, he derived the theory of correlation and regression.



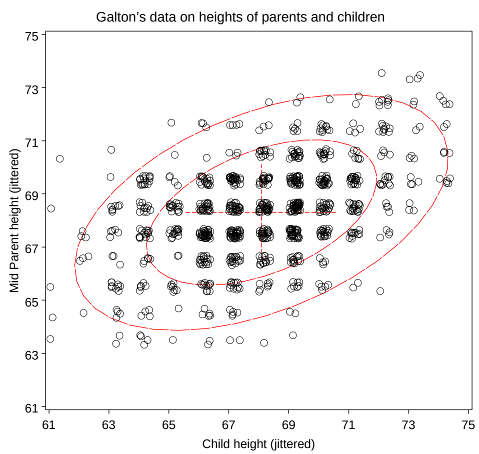
- How did he arrive at this insight?

Example: Galton's data

A plot of his data (grouped into class intervals) is unrevealing:

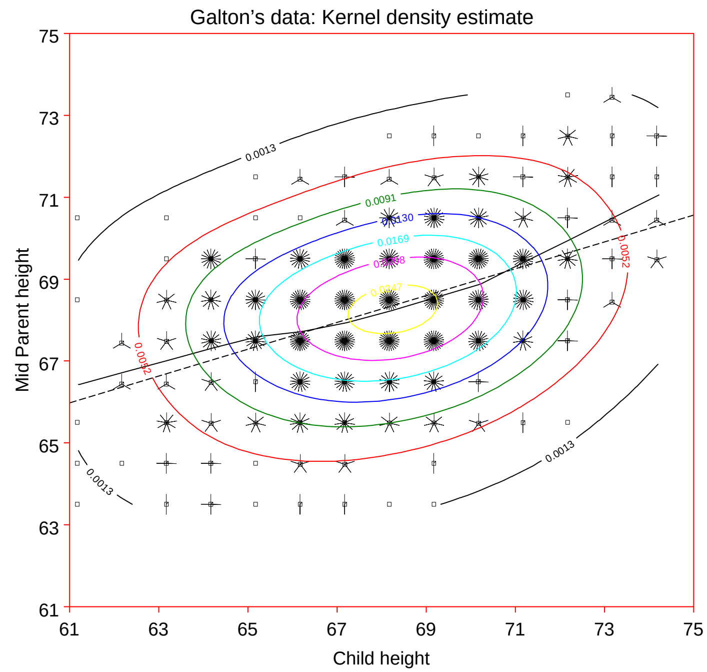


A jittered plot of the observations shows the concentration of observations more clearly:



Example: Galton's data

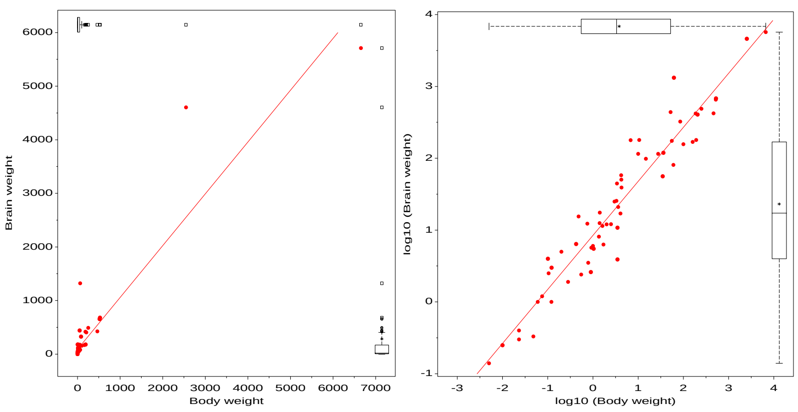
Sunflower symbols, loess smoothed curve, regression line, and non-parametric bivariate density contours (`PROC KDE`):



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,

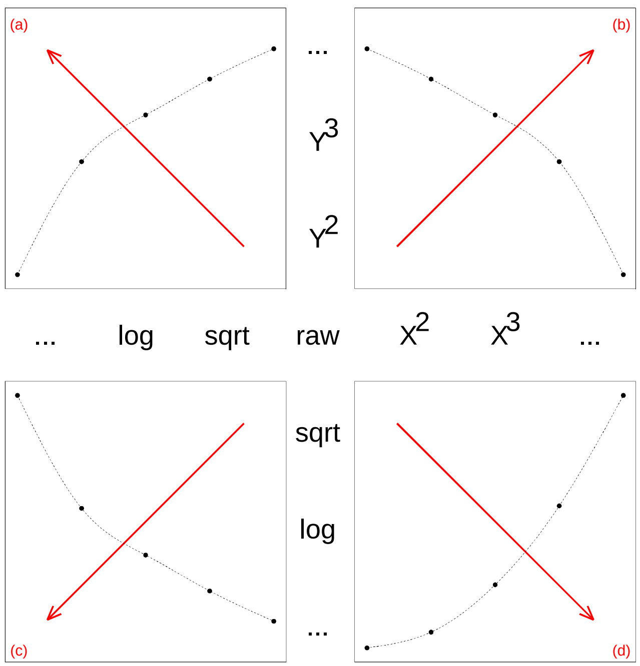
$$y' = a + b\,x' + \text{residual}$$

- Ladder of powers and Tukey's "arrow rule" indicate which direction to go.
- A "ratio of slopes" table pinpoints good power transformations.

Transformations to linearity

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

Resistant lines and the ratio of slopes table (Tukey, 1977):

- Least squares regression can give misleading results with highly skewed data or with outliers
- A resistant line often does better with ill-behaved data
- Summary values – medians of thirds, dividing by X-values (but neither end-third can cover more than 1/2 the range)

```
                 Summary Values

                     X          Y       n

Low          0.122      2.500      21
Mid         10.000     80.996      39
High      4600.627   5157.515       2 R
```

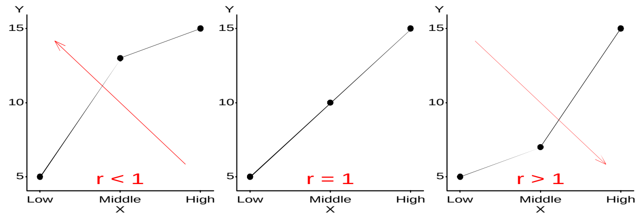
- Ratio of slopes
  - The curvature of the data can be measured by the ratio of slopes

$$r = \frac{\text{upper slope}}{\text{lower slope}} = \frac{(y_H - y_M)/(x_H - x_M)}{(y_M - y_L)/(x_M - x_L)}$$

```
           X          Y       half-slope    ratio
High   4600.627   5157.515
                                 1.1058
Mid      10.000     80.996                  0.1391
                                 7.9465
Low       0.122      2.500
```

  - A linear relation $\Rightarrow r \approx 1$ (or $\log r \approx 0$)
  - The effect of any transformation, $x \to x^p$, $y \to y^q$, can be judged by the effect it has on the ratio of slopes,

$$r^{(p,q)} = \frac{(y_H^q - y_M^q)/(x_H^p - x_M^p)}{(y_M^q - y_L^q)/(x_M^p - x_L^p)}$$



- The `resline` macro calculates the ratio of slopes for a set of powers of x and of y

```
%resline(data=brains,
   x = bodywt, y = brainwt, id=mammal);
```

  - For this data, values of $r \approx 1$ tend to run along the diagonal
  - log-log is the best combination

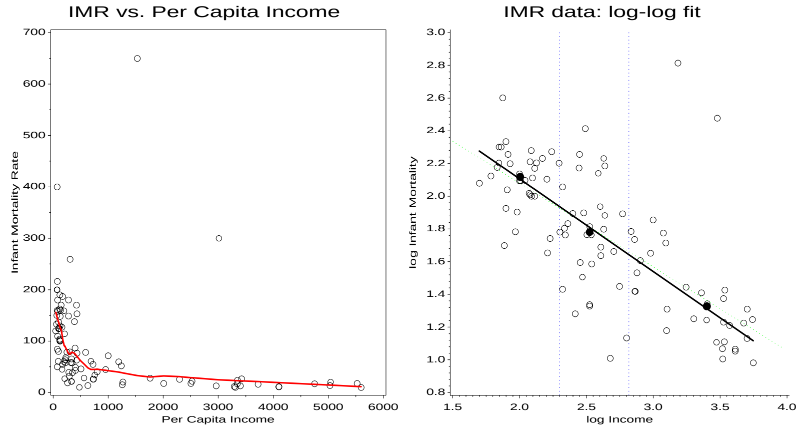
```
      ----- Ratio of Slopes table ------
      Rows are powers of X, columns are powers of Y

           -1.0     -0.5      log     sqrt       raw       2.0

-1.0      2.544   15.127   96.908  687.070  5247.745  329241.7
-0.5      0.265    1.575   10.089   71.527   546.314  34275.54
log       0.023    0.134    0.858    6.085    46.477  2915.947
sqrt      0.001    0.008    0.052    0.368     2.813   176.504
raw       0.000    0.000    0.003    0.018     0.139     8.731
 2.0      0.000    0.000    0.000    0.000     0.000     0.019

         ------- 5 Best powers -------
      Power of X Power of Y   Slope Ratio   log Ratio

      log        log               0.858      -0.066
      -0.5       -0.5              1.575       0.197
      -1.0       -1.0              2.544       0.405
      sqrt       sqrt              0.368      -0.434
      sqrt       raw               2.813       0.449
```

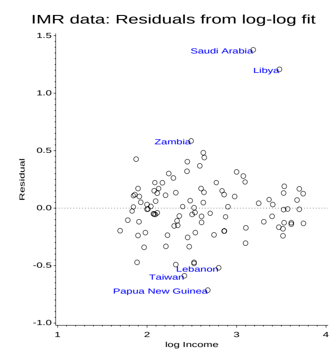
- See http://www.math.yorku.ca/SCS/sasmac/resline.html

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,

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

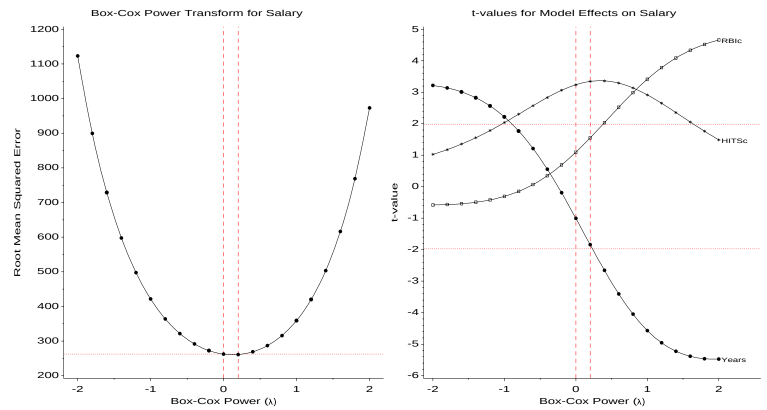
  where λ is another parameter, the power in (the 'ladder')

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

- 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);
```



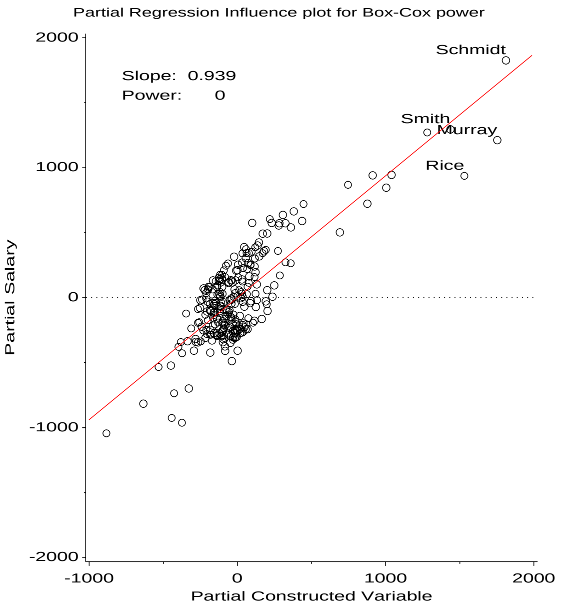
Box-Cox: Score test and influence plot

- A score test is based on the slope of the $\log L$ function at $\lambda = 1$ (slope $\approx 0 \leftrightarrow$ at maximum)
- For Box-Cox, this can be formulated as the t statistic for a *constructed variable*, $\boldsymbol{g}$,

$$g_i = y_i(\log \frac{y_i}{\tilde{y}} - 1)$$

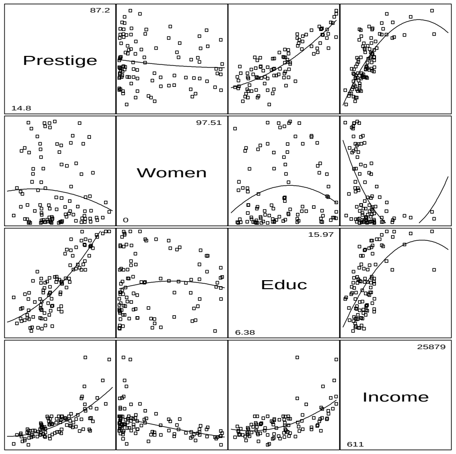
  where $\tilde{y}$ is the geometric mean of the y_i.
- Fit the model $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \phi \boldsymbol{g}$.
- Test $H_0 : \phi = 0 \quad (\leftrightarrow \lambda = 1)$. Another estimate of λ is $1 - \phi$.
- A partial regression plot for $\boldsymbol{g}$ shows the influence of individual observations on the choice of the transformation.

- Baseball data: predicting Salary from Years, RBIc, HITSc.
  - The influence plot shows that a few players are strongly determining the choice of power, but they are not out of line with the rest.
  - The slope (ϕ) again leads to the choice $\lambda = 0 \Rightarrow \log y$
- Plot produced by the `boxcox` macro (with GPLOT=INFL):



Transformations of predictors

- 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.)
- e.g., Canadian occupational prestige: %women, income, education



  - $\longrightarrow$ Prestige non-linear w.r.t. Educ and Income

Box-Tidwell Transformations

- Box and Tidwell (1962) suggested a model to determine transformations of the Xs,

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1^{\gamma_1} + \cdots \beta_k \boldsymbol{x}_k^{\gamma_k} + \boldsymbol{\epsilon}$$

- Parameters of this model—$\beta_0, \beta_1 \ldots \beta_k, \gamma_1 \ldots \gamma_k$ can be estimated by:
  1. Regress y on $x_1, \ldots, x_k \rightarrow b_0, b_1, \ldots b_k$.
  2. Create constructed variables, $x_1 \log x_1, \ldots x_k \log x_k$.
  3. Regress y on $x_1, \ldots, x_k,\ x_1 \log x_1, \ldots x_k \log x_k$ $\rightarrow b'_0, b'_1, \ldots b'_k, g_1, \ldots g_k$
  4. Estimate of the power γ_i is given by $\hat{\gamma} = 1 + g_i / b_i$
  5. Repeat steps 3, 4 until $\hat{\gamma}$ converge (gives MLE).
- The constructed variables, $x_i \log x_i$, can be used to test the need for a transformation of x_i: Test $H_0 :\ \gamma_i = 1$ from test of coefficient of $x_i \log x_i = 0$.
- Partial regression plots for the constructed variables help to assess the leverage and influence on the decision to transform an x variable.

Box-Tidwell transformations: Example

Canadian Occupational Prestige – find powers for Educ and Income

- The BOXTID macro carries out this procedure:

```
%boxtid(data=prestige,
   yvar=Prestige,  id=job,
   xvar=Women Educ Income,  /* vars in model   */
   xtrans=Educ Income,      /* vars to xform   */
   round=.5,                /* round powers    */
   out=boxtid);             /* output data set */
```

- Printed results show the iteration history...

```
Iteration History: Transformation Powers
Iteration    EDUC   INCOME  Criterion

     1    2.2551  -0.9132     1.9132
     2    2.3790   0.8273     1.9059
     3    2.3593  -0.6834     1.8261
     4    2.3221   0.4444     1.6503
    ...
    13    2.2109  -0.0426     0.0005
```

- ... and (score) tests for power transformations

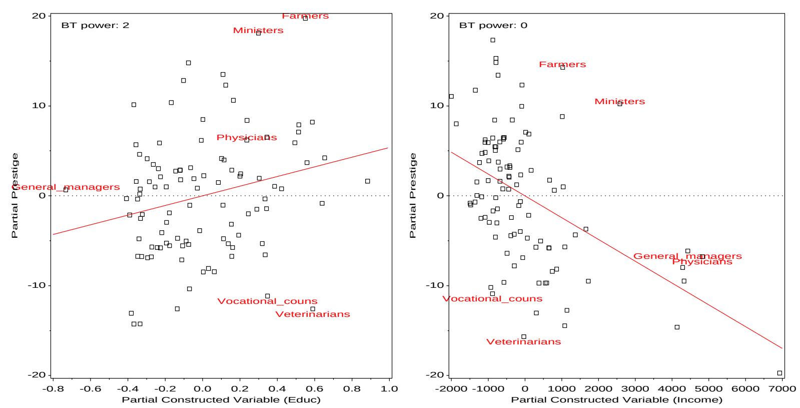
```
   Score tests for power transformations
            Power   StdErr   Score Z Prob>|Z|

EDUC       2.2109   4.9114   2.4097   0.0160
INCOME    -0.0426   0.0000  -5.2625   0.0000
```

- Powers are rounded to the nearest 0.5:

  Educ $\longrightarrow$ Educ^2, Income $\longrightarrow \log$ Income.

Box-Tidwell transformations: Example

- Partial regression plots for the transformed variables show that several observations are influential for the choice of power for Income.

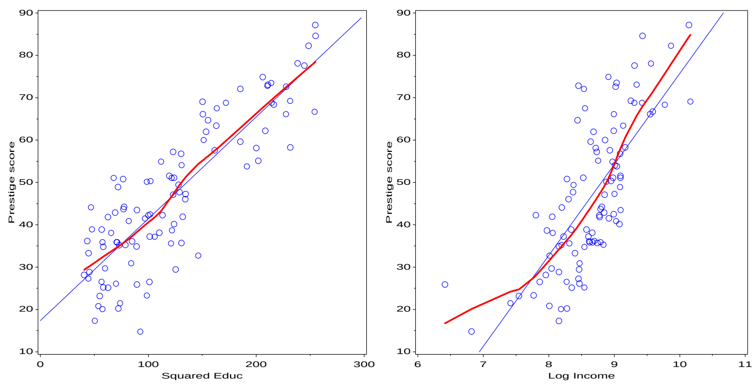


Box-Tidwell transformations: Example

- The BOXTID macro creates the transformed variables for you (e.g., t_income).
- Plot with LOWESS macro, adding linear regression lines:

```
%lowess(data=boxtid, x=t_educ, y=prestige, id=job,
   f=.667, interp=rl);
%lowess(data=boxtid, x=t_income, y=prestige, id=job,
   f=.667, interp=rl);
```

- Plots of Prestige vs. Educ^2 and $\log(\text{Income})$ show that both variables are now approx. linearly related to Prestige.



- The lowest two occupations on $\log(\text{Income})$ should be looked at more closely.

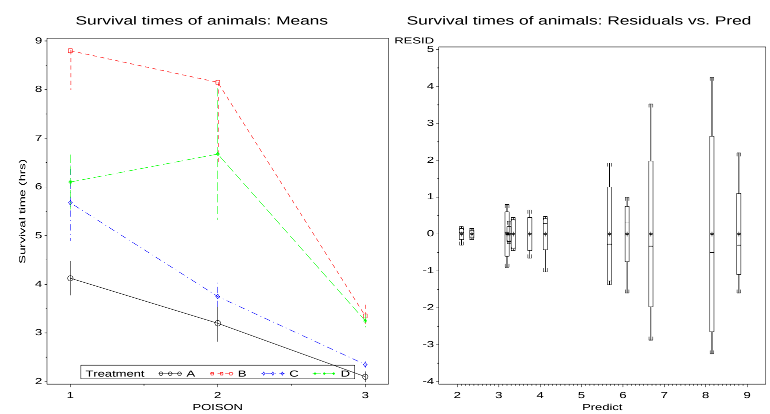
Dealing with heteroscedasticity

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

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

- 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);
```



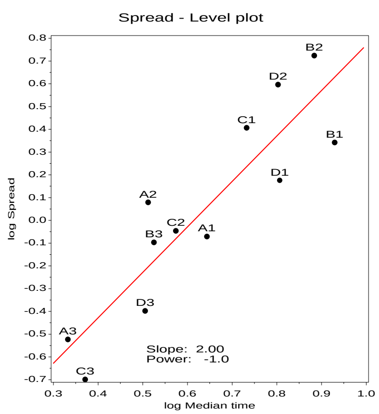
- Both plots show greater variance associated with longer survival time.

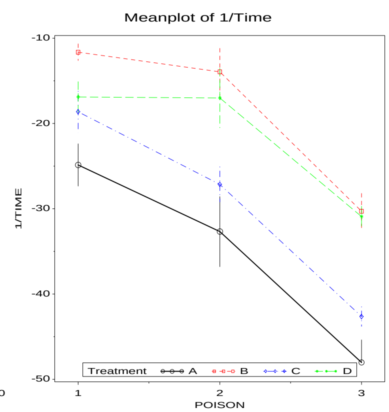
Dealing with heteroscedasticity

Spread 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 $y \longrightarrow y^p$, with $p = 1 - b$.

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





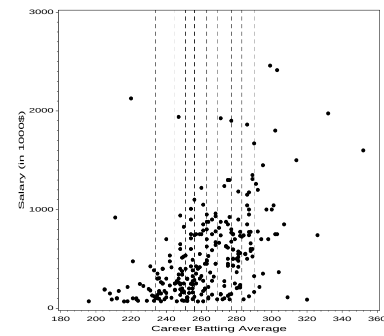
- The plot suggests transforming Time $\longrightarrow$ 1/Time.
- 1/Time also reduces apparent interaction of Poison * Treatment

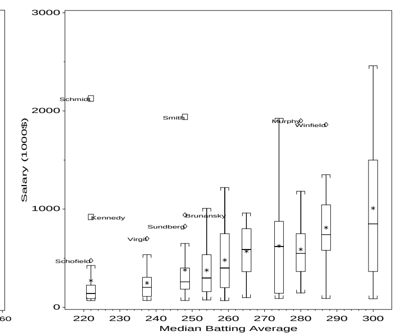
Dealing with heteroscedasticity

Regression data

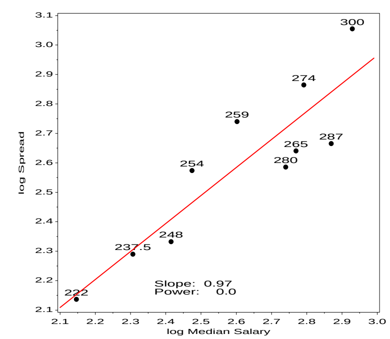
- Divide an x variable into ordered groups (e.g., deciles)

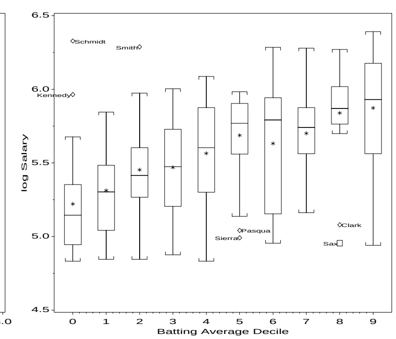
```
proc rank data=baseball out=grouped groups=10;
   var batavgc;
   ranks decile;
```





- Use Spread vs. level plot on grouped x





- log Salary is again indicated

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. 57

Box, G. E. P. and Tidwell, P. W. Transformation of the independent variables. *Technometrics*, 4:531–550, 1962. 62

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.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

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

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, 1987.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977. 54