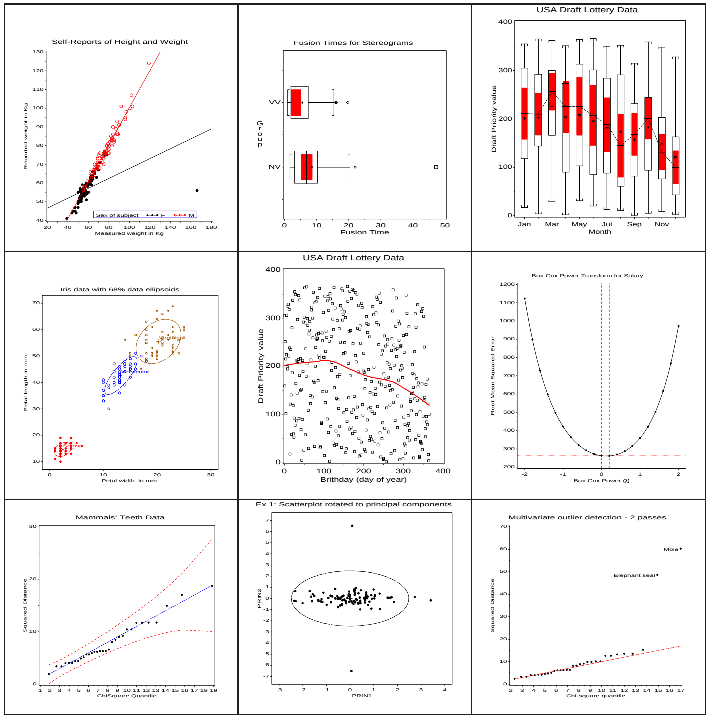
Data Screening: Part 3



Michael Friendly

York University

SCS Short Course

October, 2004

Part 3: Multivariate problems and missing data

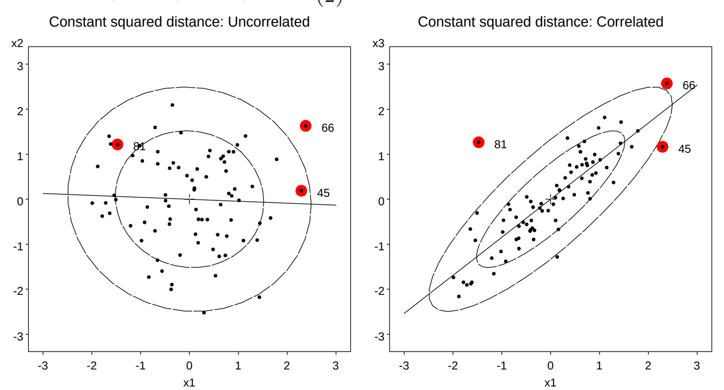
- Assessing multivariate problems
  - Multivariate normality
  - Outliers: univariate, bivariate, multivariate
  - Robust outlier detection
- Dealing with missing data
  - Estimation with missing data (EM algorithms)
  - Simple Imputation
  - Multiple imputation
  - Plots for missing data

Multivariate normality

- Some multivariate statistical methods assume that all measures are jointly multivariate normal.
  - e.g., Factor analysis, discriminant analysis, MANOVA
  - Regression: Usually *not* required for predictors
    - Usually *not* required for predictors
    - *Is* required for multivariate MRA
- 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.

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 $D_i^2 = z_{i1}^2 + z_{i2}^2 \sim \chi^2_{(2)}$.



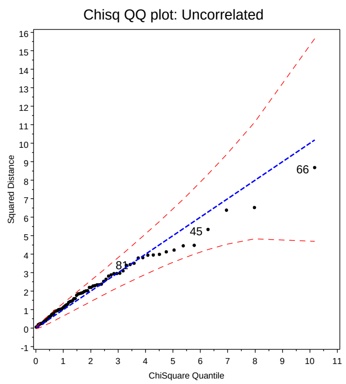
  - p variables: Calculate generalized (Mahalanobis) squared distance, D_i^2 of each observation $\boldsymbol{x}_i$ from the mean vector,

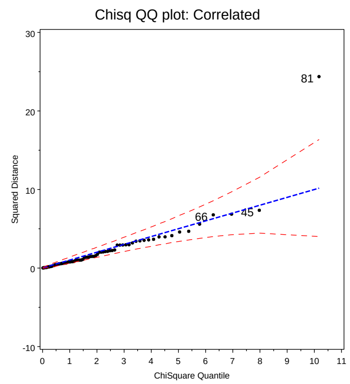
$$D_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathsf{T}} \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \sim \chi^2_{(p)}$$

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.





Computation:

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

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

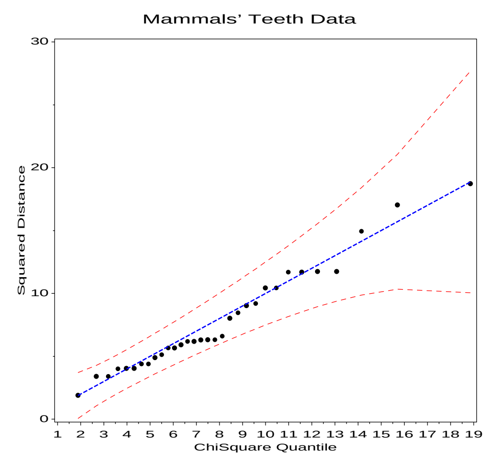
- The `multnorm` macro calculates univariate and multivariate normality tests, and produces the Chi-square QQ plot.
  - Confidence bands for the distribution help to judge how close the D^2_i are to a χ^2 distribution.
  - But: outliers can make the graphical test lest sensitive.

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

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

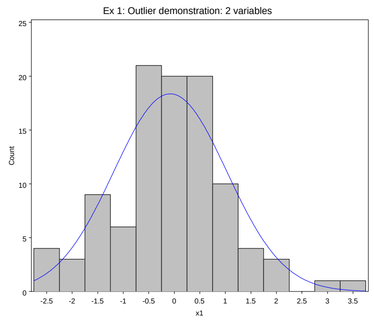
Var	Test	Skewness	Kurtosis	Test Statistic	p-value
V1	Shapiro-Wilk	-0.6993	-0.8885	0.790	0.00001
V2	Shapiro-Wilk	-0.3040	-1.0806	0.829	0.00008
V3	Shapiro-Wilk	-1.0216	-1.0246	0.560	0.00000
V4	Shapiro-Wilk	-0.5421	-1.8244	0.608	0.00000
V5	Shapiro-Wilk	-0.8124	0.2587	0.863	0.00060
V6	Shapiro-Wilk	-0.5955	-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

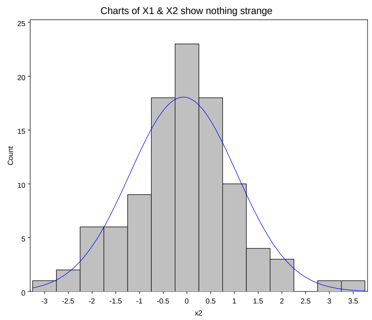
- All test statistics indicate substantial deviation from univariate and multivariate normality
- QQ plot does not reveal anything strange. Why?



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 $\pm 1.5IQR$, $(p < 0.05)$, outer fences– quartiles $\pm 3IQR$, $(p < 0.001)$.
  - `datachk` macro gives a brief summary for a collection of variables
- Univariate checks are useful, but not always sufficient: Can you spot the outliers?

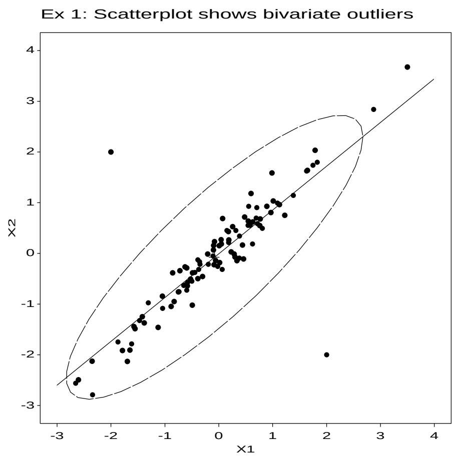




Bivariate outliers

- Bivariate plots can reveal— bivariate outliers!

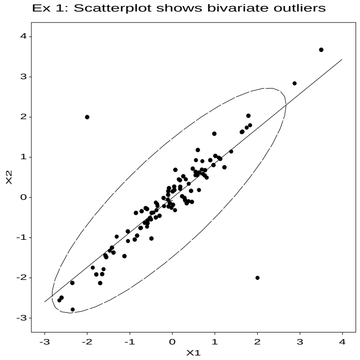
```
data outlier1;
   do i = 1 to 100;
      x1 = normal(33445);           * Correlated;
      x2 = x1 + normal(22345)/4;    * bivariate normal;
      output;
      end;
   *-- Generate two additional obs: outliers;
   x1 = 2; x2 = -2; output;
   x1 =-2; x2 =  2; output;
```

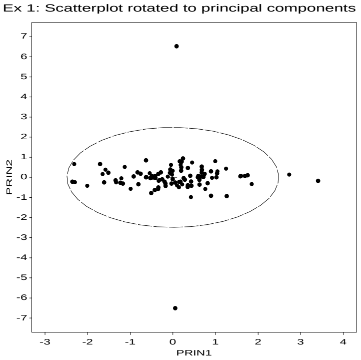


- But, *only* bivariate outliers
- Bivariate plot suggests rotation to principal components

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.





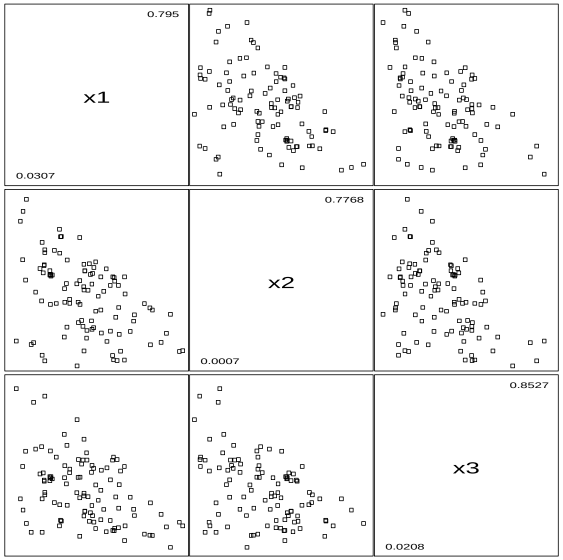

```
proc princomp std noprint data=outlier1 out=prin;
   var x1-x2;
title 'Ex 1: Scatterplot rotated to principal components';
%contour( data=prin, y=prin2, x=prin1, pvalue=.95);
```


Multivariate outliers

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

```
data outlier2;
   do i = 1 to 100;
      x1 = uniform(54321);
      x2 = uniform(54321);
      x3 = uniform(54321);
      x1 = x1 / sum(of x1-x3);
      x2 = x2 / sum(of x1-x3);
      x3 = x3 / sum(of x1-x3);
      output;
   end;
   x1 = .1;  x2 = .1; x3 = .1;  output;   /* outlier */
   x1 = .15; x2 = .05; x3 = .1; output;   /* outlier */
```

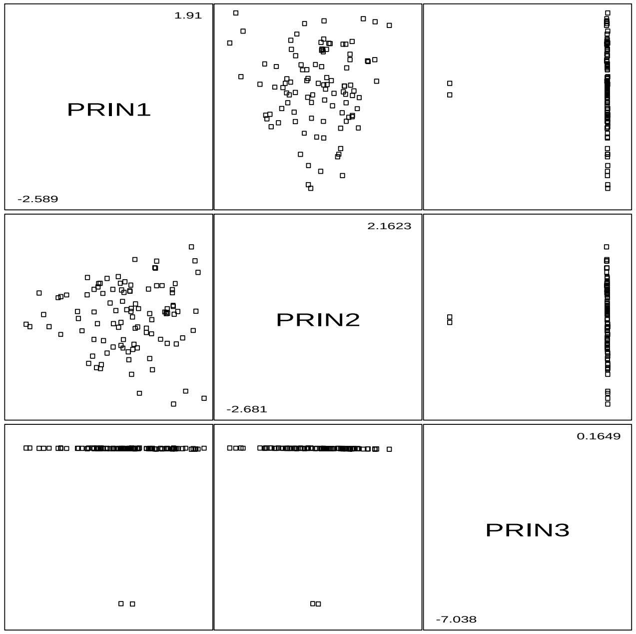
Can you spot the outliers?



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);
```



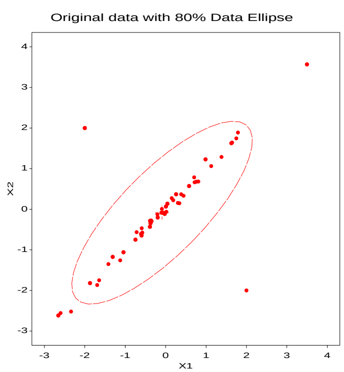
Robust Outlier Detection

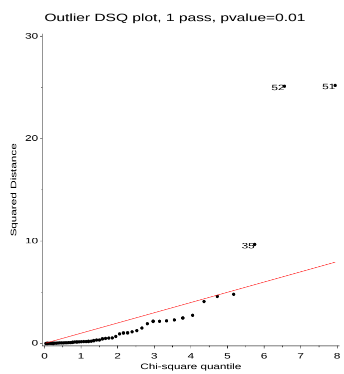
- The χ^2 plot for multivariate normality is not resistant to the effects of outliers.
- A few discrepant observations affect the mean vector, $\bar{\boldsymbol{x}}$, and—worse—the variance-covariance matrix, $\boldsymbol{S}$.
- Inflating $\boldsymbol{S} \rightarrow$ decreases D^2: extreme obs. look 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.

`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);
```





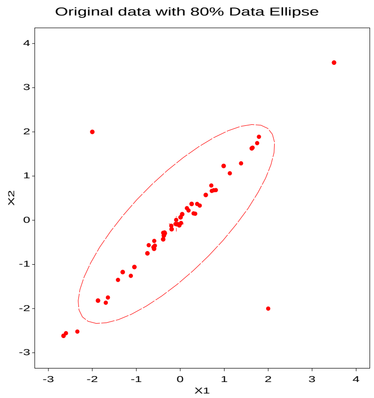
Outlier DSQ plot, 1 pass, pvalue=0.01
Observations trimmed in calculating Mahalanobis distance

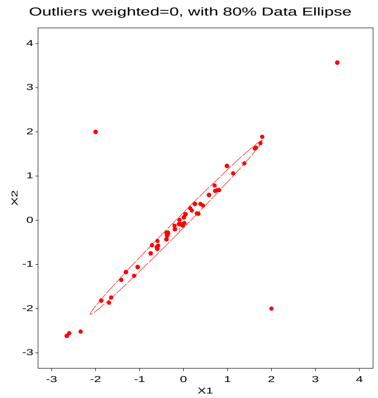
PASS	_CASE_	DSQ	PROB	
1	35	9.6729	.0079353	
	51	25.2015	.0000034	*
	52	25.1222	.0000035	*

See: www.math.yorku.ca/SCS/sssg/outlier.html

`outlier` macro

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

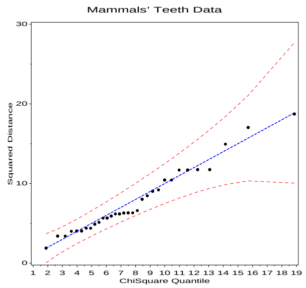




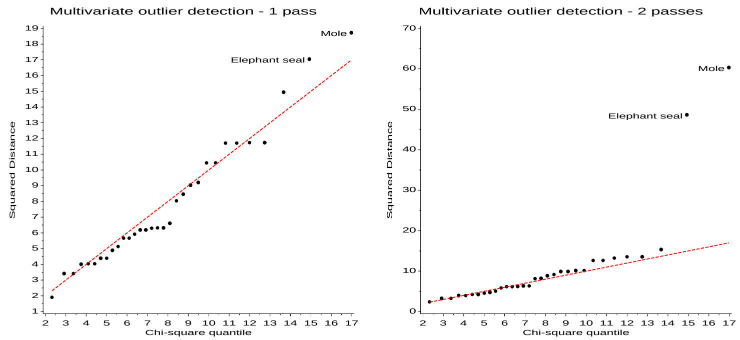

```
title 'Original data with 80% Data Ellipse';
%contour(data=outlier1, y=x2, x=x1, pvalue=.80);

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.
  - Classical methods: 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.

Dealing with missing data

- Gertrude Cox: “*The best thing to do about missing data is not to have any*.”
- Most software uses one of two procedures for missing data:
  - Complete case analysis (listwise deletion) — Discard data with *any* missing variables
  - Available case analysis (pairwise deletion)— Discard data with missing values on the analysis variables
- E.g.,, in linear models (regression, ANOVA, etc.):
  - univariate statistics based on available data
  - missing on *any* predictor $\longrightarrow$ case deleted (listwise)
  - missing on response $\longrightarrow$ case not used, but a fitted value is generated.
  - multiple responses (MANOVA, Multivar regression): software difffers
- Caveats:
  - Must assume at least *missing-at-random* (MAR): “missingness” on X is unrelated to value of X
  - e.g., *not* MAR if respondents with high income are more likely not to report income.
  - Failure of MAR $\longrightarrow$ results (predicted values, coefficients) are biased
  - Factor analysis, PCA, etc: available case analysis can $\longrightarrow$ improper correlation matrices (not PD)

Missing data: General strategies

- ***Single imputation***: replace missing by ‘suitable estimates’, use complete-case analysis.
- ***Weighting***: discard if any missing, but weight complete cases to compensate for incomplete cases.
- ***Direct analysis*** of incomplete data. Two forms:
  - Available case analysis
  - Maximum likelihood estimation over available data (e.g., `PROC MIXED`, E-M algorithm)
- ***Multiple imputation***:
  - Impute $m > 1$ from appropriate distribution for each missing observation.
  - Combine $\longrightarrow$ estimates, std. errors that incorporate missing-data uncertainty.
- See: General FAQ 25: Handling missing or incomplete data http://www.utexas.edu/cc/faqs/stat/general/gen25.html

Missing data: Imputation

- Trade-offs
  - $+$ Get complete data $\longrightarrow$ use software not handling missings
  - $+$ Makes good use of info on incomplete cases
  - $-$ Analyses overstate precision: nominal 95% CI may have only 80%–90% coverage; p-values $< .05$ may be really 0.10–0.20!
- Some imputation methods:
  - Unconditional imputation: fill in the grand mean
    - Only gives illusion of progress!
    - Estimates of variance are understated!
  - Conditional imputation
    - Regression-based imputation: Fill-in $\hat{x}$ using other Xs as predictors, i.e., $\mathcal{E}(X \mid \text{others})$.
    - Sub-group means, i.e., $\mathcal{E}(X \mid \text{group})$.
    - Cluster-based imputation: group into clusters; fill-in cluster mean
  - Stochastic conditional imputation
    - Regression: Fill-in $\hat{x}_i + r_i$, where $r_i \sim \mathcal{N}(0, \sigma^2)$.
    - "Hot-deck" imputation: cluster, then fill-in randomly chosen observation in same cluster.
  - Multiple imputation: Impute $m \geq 2$ values for each missing obs.; use variability of these imputations to correct std. errors, p-values

Single Imputation: Examples

Auto data: some missing values for repair record in 77 & 78

- Unconditional means (Not A Good Idea)

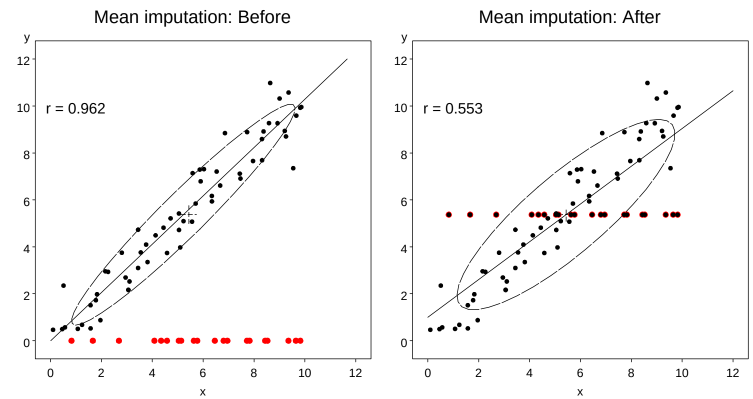
```
%include data(auto);
data auto;
   set auto;
   r77 = rep77;  *-- copy original vars;
   r78 = rep78;
proc standard REPLACE;
   var rep77 rep78;
proc print;
   where (r77=. or r78=.);
```

$\longrightarrow$

MODEL	ORIGIN	REP77	REP78	R77	R78
AMC SPIRIT	A	3.20	3.41	.	.
BUICK OPEL	A	3.20	3.41	.	.
FORD FIESTA	A	3.20	4.00	.	4
MERC. MONARCH	A	3.20	3.00	.	3
PEUGEOT 604 SL	E	3.20	3.41	.	.
PLYM. HORIZON	A	3.20	3.00	.	3
PLYM. SAPPORO	A	3.20	3.41	.	.
PONT. PHOENIX	A	3.20	3.41	.	.

What's wrong with mean substitution?

- Problems:
  - Corrupts marginal distribution of each imputed variable: s^2 too small ($\bar{x}$ OK)
  - Corrupts covariances and correlations with other variables: $|r|$ too small



- Conditional means (by region of origin)

```
proc sort data=auto;
   by origin;
proc standard REPLACE;
   by origin;
   var rep77 rep78;
```

$\longrightarrow$

MODEL	ORIGIN	REP77	REP78	R77	R78
AMC SPIRIT	A	2.98	3.02	.	.
BUICK OPEL	A	2.98	3.02	.	.
FORD FIESTA	A	2.98	4.00	.	4
MERC. MONARCH	A	2.98	3.00	.	3
PLYM. HORIZON	A	2.98	3.00	.	3
PLYM. SAPPORO	A	2.98	3.02	.	.
PONT. PHOENIX	A	2.98	3.02	.	.
PEUGEOT 604 SL	E	2.90	4.00	.	.

Single Imputation: Examples

- Regression estimates
  - Include vars with missing as dependents
  - Replace missing values with predicted values from regression
  - `PROC REG`: Output data set with predicted values

```
proc reg data=auto;
   model rep77 rep78 = price mpg hroom rseat trunk
                weight length turn displa gratio;
   output out=newauto p=p77 p78;
```

$\longrightarrow$

MODEL	ORIGIN	P77	P78	REP77	REP78
AMC SPIRIT	A	3.85	4.38	.	.
BUICK OPEL	A	3.76	3.58	.	.
FORD FIESTA	A	2.76	3.70	.	4
MERC. MONARCH	A	3.00	3.06	.	3
PLYM. HORIZON	A	3.42	4.34	.	3
PLYM. SAPPORO	A	4.07	3.87	.	.
PONT. PHOENIX	A	3.08	2.86	.	.
PEUGEOT 604 SL	E	3.31	4.20	.	.

- Problems:
  - Inflates covariances and correlations with other variables

Single Imputation: Examples

- Cluster-based imputation: `PROC FASTCLUS` $\longrightarrow$

```
proc fastclus IMPUTE data=auto out=auto2
     maxclusters=10 delete=1 summary;
   var price -- gratio;
   id model;
```

MODEL	ORIGIN	REP77	REP78	CLUSTER	DIST	_IMPUTE_
AMC SPIRIT	A	2.5	2.7	1	702.05	2
BUICK OPEL	A	3.6	4.0	8	347.71	2
FORD FIESTA	A	3.6	4.0	8	291.73	1
MERC. MONARCH	A	2.5	3.0	1	408.48	1
PEUGEOT 604 SL	E	3.0	2.7	2	941.74	2
PLYM. HORIZON	A	3.6	3.0	8	274.28	1
PLYM. SAPPORO	A	3.9	4.4	4	411.99	2
PONT. PHOENIX	A	2.5	2.7	1	410.23	2

Problems with single imputation

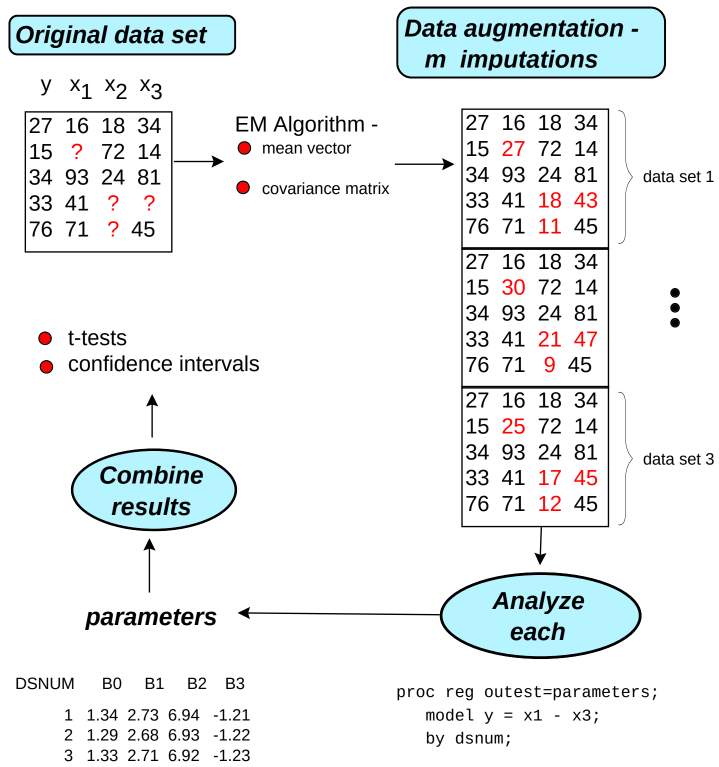
- Subsequent analyses do not reflect *missing data uncertainty*
  - sample size, N and degrees of freedom are overstated
  - confidence intervals too narrow
  - Type I error rates too high
- Problem gets worse as rate of missing and model complexity (number of parameters) increase
- Example:
  - 30% missing
  - One confidence interval (regression coeff., odds ratio, ...)

Nominal coverage (%)	90	95	99
Actual coverage (%)	77	85	94

  - Testing a 10-parameter H_0 (e.g., regression F-test)

Nominal α	0.10	0.05	0.01
Actual α	0.57	0.45	0.25

Multiple imputation



Multiple imputation

- Obtaining valid inferences from imputed data: Little and Rubin (1987), Rubin (1987), Schafer (1997)
- Missing values replaced by $m > 1$ simulated versions, $(3 \le m \le 10)$.
- Each imputed complete dataset is analyzed by standard methods,
- Results combined to produce estimates and confidence intervals that incorporate missing-data uncertainty
- High efficiency, even for small m.

$$\text{Rel. Efficiency} = \left(1 + \frac{\gamma}{m}\right)^{-1}$$

where γ = rate of missing info (about a parameter)

	γ				
m	.1	.3	.5	.7	.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	96

See The Multiple Imputation FAQ page, http://www.stat.psu.edu/~jls/mifaq.html

Multiple imputation: Combining estimates

Rubin (1987) method for MI inference, ***scalar quantities*** (θ)—

- m imputations $\longrightarrow$ m estimates, $\hat{\theta}_i$, each with an estimated sampling variance $\widehat{var(\theta_i)}$.
- MI point estimates: average the m values of $\hat{\theta}_i$

$$\bar{\theta} = \frac{1}{m} \sum_i^m \hat{\theta}_i$$

- Proper tests and CI for imputed data must take into account:
  - ***Within-imputation variance***: average sampling variance of the m estimates.

$$\bar{W} = \sum \widehat{var(\theta_i)} / m$$

  - ***Between-imputation variance***: variability of the estimates across m imputations.

$$B = \sum (\hat{\theta}_i - \bar{\theta})^2 / (m - 1)$$

- These are combined to give the ***Total-imputation variance*** of $\bar{\theta}$,

$$T \equiv var(\bar{\theta}) = \bar{W} + (1 + \frac{1}{m}) B$$

Multiple imputation: Significance tests and CI

- MI hypothesis tests: $t_{obs} = \bar{\theta}/\sqrt{T} \sim t_{df}$
- MI adjusted confidence interval: $\bar{\theta} \pm t_{df}\sqrt{T}$
- Degrees of freedom:

$$df = (m-1)\left(1 + \frac{m\bar{W}}{(m+1)B}\right)^2$$

- Fraction of missing info (γ), relative increase in variance due to nonresponse (r):

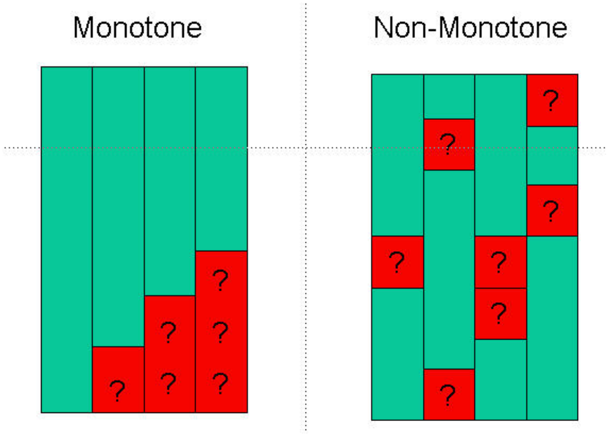
$$\gamma = \frac{r + 2/(df+3)}{r+1} \qquad r = \frac{T - \bar{W}}{\bar{W}} = \frac{(1+m^{-1})B}{\bar{W}}$$

Multiple imputation: Software

- SAS:
  - SAS V8.1+: `PROC MI`, `PROC MIANALYZE` (`www.sas.com/rnd/app/papers/multipleimputation.pdf`)
- SPSS: Missing Value Analysis (MVA) add-in module [EM only]
- Splus and Win95/NT (Joe Schafer)
  - NORM - Multivariate normal
  - CAT - Multivariate categorical
  - MIX - Continuous and categorical
  - PAN - Panel or clustered
- Other software listed at `www.utexas.edu/cc/faqs/stat/general/gen25.html`

Multiple imputation: PROC MI and PROC MIANALYZE

- PROC MI — Different methods for different missing patterns:



- Monotone missing data pattern: $Y_j = . \Rightarrow Y_k = ., \forall k > j$.
  - Parametric regression (assumes multivariate normality)
  - Non-parametric, propensity scores method
- Non-monotone missing data pattern
  - Markov chain monte carlo (MCMC) for all (comp. intensive) [default!]
  - MCMC $\longrightarrow$ monotone pattern. Then, use monotone methods.
- Generates m imputed data sets, with index variable _Imputation_.

Multiple imputation: PROC MI and PROC MIANALYZE

- *Any* analysis step BY _Imputation_, producing estimates. (PROC REG, PROC GLM, PROC MIXED, PROC GENMOD, etc.)

```
proc reg data=outmi outest=outreg covout noprint;
   model Oxygen = RunTime RunPulse;
    by _Imputation_;
```

- PROC MIANALYZE
  - Combines estimates, á la Rubin (1987), Schafer (1997)
  - Provides both univariate (t) and multivariate (F) tests.

Baseball data: PROC MI and PROC MIANALYZE

- Salary and performance data for $n = 322$ players
- Salary missing for 18% of players (monotone pattern)
- Model: $\log(\text{salary}) \sim \min(\text{years}, 7) + trpc + batavgc$
- MI assumes:
  - Variables are multivariate-normal (transform first if not)
  - Model used for imputation is the same as the analysis model

0. Screen and transform variables

basemi.sas ···

```
%include data(baseball);
*-- Screen/Transform variables;
data baseball;
    set baseball;
    if salary ^=.
        then logsal = log(salary);
    years7 = min(years,7);
    trpc = (runsc + rbic + homerc) / years;
    label logsal = 'log Salary'
        trpc='Total career runs/year'
        years7='Years, up to 7';
```


1. Generate m imputed data sets

··· basemi.sas ···

```
title 'Proc MI: Regression method (monotone)';
proc mi data=baseball seed=42424241 out=basemi;
    monotone method=regression;
    var years7 trpc batavgc logsal;
    run;
```

Printed output:

```
                         The MI Procedure

                        Model Information

     Data Set                                WORK.BASEBALL
     Method                                  Regression
     Number of Imputations                   5
     Seed for random number generator        42424241

                      Missing Data Patterns

 Group     years7     trpc     batavgc     logsal        Freq     Percent

     1     X          X        X           X              263       81.68
     2     X          X        X           .               59       18.32
                      Missing Data Patterns

           -------------------------Group Means-----------------------
Group             years7             trpc            batavgc            logsal

    1           5.224335        94.246679         262.794677          5.927417
    2           5.237288        73.385852         255.474576                 .
```

Notes:

- Are there differences in means among different missing patterns?
- Is there evidence that data is not MAR?

2. Analyze m complete data sets

- Use output dataset from `PROC MI` as input to `PROC REG`
- Obtain output dataset containing parameter estimates (and covariance matrices)
- Use `by _Imputation_;` to repeat analysis m times

··· basemi.sas ···

```
proc reg data=basemi noprint outest=outreg covout;
   model logsal = years7 trpc batavgc;
   by _Imputation_;
   run;
```

Print parameter estimates:

··· basemi.sas ···

```
proc print data=outreg;
   id _Imputation_;
   by _Imputation_;
   where (_Type_ = 'PARMS');
   var Intercept years7 trpc batavgc;
   title2 'Parameter estimates from imputed data sets';
   run;
```

Output:

Parameter estimates from imputed data sets

Imputation	Intercept	years7	trpc	batavgc
1	2.62497	0.25947	.007388496	.004761202
2	2.56478	0.25190	.007207141	.005198336
3	2.48735	0.25515	.006764982	.005615983
4	2.76897	0.25300	.008008711	.004010438
5	3.26130	0.25139	.008340637	.002135631

3. Combine results with `PROC MIANALYZE`

- Use output dataset from `PROC REG` as input to `PROC MIANALYZE`

··· basemi.sas ···

```
title 'Proc MIANALYZE to combine and test';
proc mianalyze data=outreg mult edf=318;
    var Intercept years7 trpc batavgc;
    run;
```

Printed output:

The MIANALYZE Procedure

Multiple Imputation Variance Information

Parameter	Variance: Between	Variance: Within	Variance: Total	DF
Intercept	0.095088	0.095850	0.209955	12.38
years7	0.000010826	0.000219	0.000232	241.56
trpc	0.000000399	0.000000537	0.000001015	16.254
batavgc	0.000001878	0.000001779	0.000004032	11.73

Parameter estimates, standard errors and CI:

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	2.741474	0.458209	1.74651	3.736435	12.38
years7	0.254181	0.015215	0.22421	0.284153	241.56
trpc	0.007542	0.001008	0.00541	0.009675	16.254
batavgc	0.004344	0.002008	-0.00004	0.008730	11.73

3. Combine results with `PROC MIANALYZE`

Individual hypothesis tests ($H_0 : \theta_i = 0$):

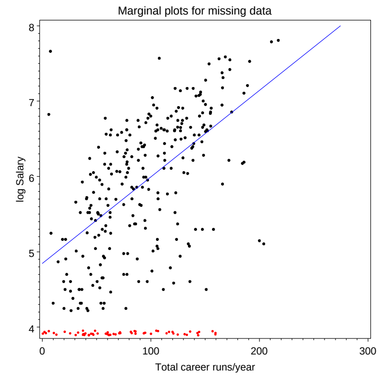
```
         Multiple Imputation Parameter Estimates

                                          t for H0:
Parameter           Theta0    Parameter=Theta0    Pr > |t|

Intercept                0                5.98      <.0001
years7                   0               16.71      <.0001
trpc                     0                7.49      <.0001
batavgc                  0                2.16      0.0519
```

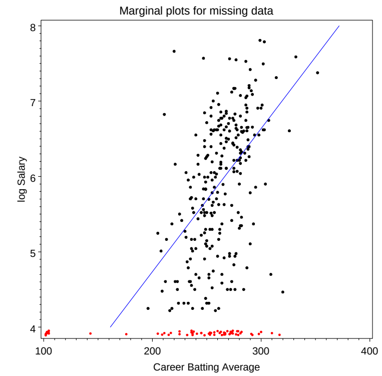
Multivariate hypothesis test ($H_0 : \theta_1 = \theta_2 = \cdots = 0$):

```
          Multiple Imputation Multivariate Inference
   Assuming Proportionality of Between/Within Covariance Matrices

  Avg Relative
      Increase                                F for H0:
   in Variance    Num DF    Den DF     Parameter=Theta0     Pr > F

      0.502386         4    94.202              7499.13     <.0001
```


Plots for missing data

- Marginal plots
  - Ordinary bivariate plots ignore all missing observations
  - Instead, show missing observations as marginal points
  - Are the missing observations consistent with the marginal distributions? (weak test of MAR)
- Example: Baseball data, marginal plots for Career runs/year (`trpc`) and Career batting average (`batavgc`)
  - Only `logsal` is missing here.
  - Missing observations shown at margins (red).
  - For illustration, $\sim 10\%$ of observations with missing salary also had `batavgc` set to missing.





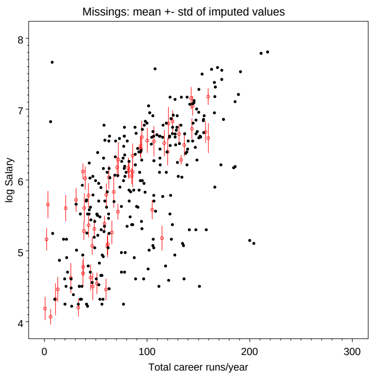
Plots for missing data

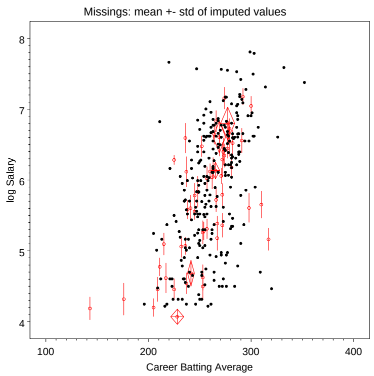
- Imputation plots
  - For missing observations, calculate typical value (mean, median) and variability (std., stderr, IQR) over the m imputed data sets.
  - Show fully observed data as points, missings as typical value (mean, median), with error bars for variability, over the m imputed data sets.
  - `miplot` macro: takes input data, imputed data $\longrightarrow$ plot for (x, y); std. error bars for missing on one, diamonds for missing on both.

··· basemiplt.sas

```
proc mi data=baseball out=basemi nimpute=10;
    monotone method=regression;
    var years7 trpc batavgc logsal;
    run;

%miplot(data=baseball, imputed=basemi,
    x=trpc,    y=logsal, id=name);
%miplot(data=baseball, imputed=basemi,
    x=batavgc, y=logsal, id=name);
```

Plots:





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