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 → 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, \( z_i^2 = (x_i - \bar{x})^2/s^2 \sim \chi^2_1 \).
  - 2 variables: If uncorrelated, squared distance of \((x_{i1}, x_{i2})\) from the mean is \( D_i^2 = z_{i1}^2 + z_{i2}^2 \sim \chi^2_2 \).

\[ D_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}) \sim \chi^2_p \]

where \( S \) is the \( p \times p \) sample covariance matrix.
Multivariate normality: Chi-square QQ plot

- QQ plot of ordered distances, $D_i^2$, against corresponding $\chi^2(p)$ quantiles should give a straight line through the origin for multivariate normal data.

Chisq QQ plot: Uncorrelated

Chisq QQ plot: Correlated

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

```
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 $\chi^2$ distribution.
- But: outliers can make the graphical test less sensitive.

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

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

<table>
<thead>
<tr>
<th>Test</th>
<th>Var</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>-0.6993</td>
<td>-0.8885</td>
<td>0.790</td>
<td>0.00001</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>-0.3040</td>
<td>-1.0806</td>
<td>0.829</td>
<td>0.00008</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>-1.0216</td>
<td>-1.0246</td>
<td>0.560</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>-0.5421</td>
<td>-1.8244</td>
<td>0.608</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>-0.8124</td>
<td>0.2587</td>
<td>0.863</td>
<td>0.00006</td>
</tr>
<tr>
<td></td>
<td>V6</td>
<td>-0.5955</td>
<td>-0.2693</td>
<td>0.883</td>
<td>0.00026</td>
</tr>
<tr>
<td></td>
<td>V7</td>
<td>-0.4687</td>
<td>-1.7688</td>
<td>0.671</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>V8</td>
<td>-0.9541</td>
<td>-0.5410</td>
<td>0.702</td>
<td>0.00000</td>
</tr>
<tr>
<td>All Mardia Skew</td>
<td>40.7550</td>
<td>0.242</td>
<td>264.040</td>
<td>0.00000</td>
<td></td>
</tr>
<tr>
<td>All Mardia Kurt</td>
<td>81.1770</td>
<td>0.263</td>
<td>81.040</td>
<td>0.79241</td>
<td></td>
</tr>
</tbody>
</table>

- 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;
```

Ex 1: Scatterplot shows bivariate outliers

- 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, PRIN2 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

- 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

- 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?

- The \( \chi^2 \) plot for multivariate normality is not resistant to the effects of outliers.
- A few discrepant observations affect the mean vector, \( \bar{\mathbf{x}} \), and—worse—the variance-covariance matrix, \( \mathbf{S} \).
- Inflating \( \mathbf{S} \) 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\(_i\) = \( P(\chi^2_p > D^2_{(i)}) \)
  3. Set weight\(_i\) = 0 for any observation with prob\(_i\) < \( \alpha \).
  4. Repeat steps 1–3.
The **outlier** macro
- performs 1 or more passes of multivariate trimming.
- produces a $\chi^2$ QQ plot.

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

```r
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**

\[
\begin{array}{cccc}
  \text{X1} & -3 & -2 & -1 \\
  \text{X2} & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4
\end{array}
\]

**Squared Distance**

\[
\begin{array}{cccc}
  0 & 10 & 20 & 30 & 40 & 50 & 60 & 70 & 80 & 90 & 100
\end{array}
\]

**Chi-square quantile**

\[
\begin{array}{cccc}
  0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8
\end{array}
\]

Observations trimmed in calculating Mahalanobis distance

<table>
<thead>
<tr>
<th><em>PASS</em></th>
<th><em>CASE</em></th>
<th>DSQ</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>9.6729</td>
<td>.0079353</td>
</tr>
<tr>
<td>51</td>
<td>25.2015</td>
<td>.0000034</td>
<td>*</td>
</tr>
<tr>
<td>52</td>
<td>25.1222</td>
<td>.0000035</td>
<td>*</td>
</tr>
</tbody>
</table>

See: [www.math.yorku.ca/SCS/sssg/outlier.html](http://www.math.yorku.ca/SCS/sssg/outlier.html)

*Multivariate outliers: Mammals teeth*

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

  ![Multivariate normality QQ plot](image)

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

  ![Effect of multivariate trimming](image)

<table>
<thead>
<tr>
<th><em>PASS</em></th>
<th>MAMMAL</th>
<th><em>CASE</em></th>
<th>DSQ</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mole</td>
<td>2</td>
<td>18.7217</td>
<td>0.016421</td>
</tr>
<tr>
<td>51</td>
<td>Mole</td>
<td>2</td>
<td>60.3655</td>
<td>0.000000</td>
</tr>
<tr>
<td>52</td>
<td>Elephant seal</td>
<td>28</td>
<td>48.6327</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

*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 substantive knowledge.
- Which obs. are trimmed depends on on the $p$-value used (e.g., Mammals teeth: Racoon trimmed at $p$-value=0.07).
- The **outlier** macro uses $p$-value=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 → case deleted (listwise)
  - missing on response → case not used, but a fitted value is generated.
- Multiple responses (MANOVA, Multivar regression): software differs
- 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 → results (predicted values, coefficients) are biased

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 → 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 → 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 $X$’s as predictors, i.e., $E(X | \text{others})$.
    - Sub-group means, i.e., $E(X | \text{group})$.
    - Cluster-based imputation: group into clusters; fill-in cluster mean
  - Stochastic conditional imputation
    - Regression: Fill-in $\hat{x} = x + r$, where $r \sim N(0, \sigma^2)$.
    - “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)

```r
%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=.);
```

Michael Friendly
What's wrong with mean substitution?

- Problems:
  - Corrupts marginal distribution of each imputed variable: \( \hat{\theta} \) too small (\( \bar{\theta} \) OK)
  - Corrupts covariances and correlations with other variables: \( \hat{\rho} \) too small

### 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

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

#### Cluster-based imputation

- PROC FASTCLUS

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

### Single Imputation: Examples

- Inflates covariances and correlations with other variables
**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, ...)

<table>
<thead>
<tr>
<th>Nominal coverage (%)</th>
<th>90</th>
<th>95</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual coverage (%)</td>
<td>77</td>
<td>85</td>
<td>94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nominal $\alpha$</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual $\alpha$</td>
<td>0.57</td>
<td>0.45</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Multiple imputation**

- Missing values replaced by $m > 1$ simulated versions, ($3 \leq m \leq 10$).
- 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 $\gamma$ = rate of missing info (about a parameter)

<table>
<thead>
<tr>
<th>$m$</th>
<th>.1</th>
<th>.3</th>
<th>.5</th>
<th>.7</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97</td>
<td>91</td>
<td>86</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>94</td>
<td>91</td>
<td>88</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>93</td>
<td>96</td>
</tr>
</tbody>
</table>


**Multiple imputation:** Combining estimates

Rubin (1987) method for MI inference, scalar quantities $(\hat{\theta})$—

- $m$ imputations $\rightarrow m$ estimates, $\hat{\theta}_i$, each with an estimated sampling variance $\text{var}(\hat{\theta}_i)$.
- MI point estimates: average the $m$ values of $\hat{\theta}_i$, $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$
- Proper tests and CI for imputed data must take into account:
  - Within-imputation variance: average sampling variance of the $m$ estimates.
    $$W = \frac{1}{m}\sum_{i=1}^{m}\text{var}(\hat{\theta}_i)$$
  - Between-imputation variance: variability of the estimates across $m$ imputations.
    $$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})^2$$
- These are combined to give the Total-imputation variance of $\hat{\theta}$,
  $$T = W + \frac{1}{m}B$$
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:
  $$df = (m-1) \left(1 + \frac{mW}{(m+1)B}\right)^2$$
- Fraction of missing info ($\gamma$), relative increase in variance due to nonresponse ($r$):
  $$\gamma = r + 2/(df + 3) \quad r = \frac{T - W}{W} = \frac{(1 + m^{-1})B}{W}$$

Multiple imputation: Software

- SAS:
  - 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: $Y_j = \ldots \Rightarrow Y_k = \ldots, \forall k > j$.
  - 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 → monotone pattern. Then, use monotone methods.
    - Generates $m$ imputed data sets, with index variable _Imputation_.

- 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, à 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: \( \log(\text{salary}) \sim \min(\text{years}, 7) + \text{trpc} + \text{batavgc} \)

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

```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

```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

<table>
<thead>
<tr>
<th>Group</th>
<th>years7</th>
<th>trpc</th>
<th>batavgc</th>
<th>logsal</th>
<th>Freq</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>263</td>
<td>81.68</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td></td>
<td>.</td>
<td>59</td>
<td>18.32</td>
</tr>
</tbody>
</table>

Missing Data Patterns

---
Group Means---

<table>
<thead>
<tr>
<th>Group</th>
<th>years7</th>
<th>trpc</th>
<th>batavgc</th>
<th>logsal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.224335</td>
<td>94.246679</td>
<td>262.794677</td>
<td>5.927417</td>
</tr>
<tr>
<td>2</td>
<td>5.237288</td>
<td>73.385852</td>
<td>255.474576</td>
<td>.</td>
</tr>
</tbody>
</table>

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

2. Analyze \( m \) complete data sets

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

Print parameter estimates:

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

Output:

```
  _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

```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
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Between</th>
<th>Within</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.095088</td>
<td>0.095850</td>
<td>0.209955</td>
</tr>
<tr>
<td>years7</td>
<td>0.000010826</td>
<td>0.000219</td>
<td>0.000232</td>
</tr>
<tr>
<td>trpc</td>
<td>0.000000399</td>
<td>0.000000537</td>
<td>0.000001015</td>
</tr>
<tr>
<td>batavgc</td>
<td>0.000001878</td>
<td>0.000001779</td>
<td>0.000004032</td>
</tr>
</tbody>
</table>

Parameter estimates, standard errors and CI:

```
```

```
The MIANALYZE Procedure
Multiple Imputation Parameter Estimates
```

```
Parameter Estimate Std Error 95% Confidence Limits DF
Intercept  2.741474 0.458209 1.74651 3.736435 12.38
years7    0.256181 0.015215 0.22546 0.286901 241.56
trpc      0.000000399 0.000000537 0.0000001015 16.254
batavgc   0.000001878 0.000001779 0.000004032 11.73
```

```
Parameter estimates, standard errors and CI:
```

```
Parameter Estimate Std Error 95% Confidence Limits DF
Intercept  2.741474 0.458209 1.74651 3.736435 12.38
years7    0.256181 0.015215 0.22546 0.286901 241.56
trpc      0.000000399 0.000000537 0.0000001015 16.254
batavgc   0.000001878 0.000001779 0.000004032 11.73
```

```
Parameter estimates, standard errors and CI:
```

```
Parameter estimate, standard errors and CI:
```

```
Parameter estimate, standard errors and CI:
```

```
Parameter estimate, standard errors and CI:
```
3. Combine results with PROC MIANALYZE

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

| Parameter | Theta0 | t for H0 | Parameter=Theta0 | Pr > |t| |
|-----------|--------|----------|-------------------|-------|---|
| Intercept | 0      | 5.98     | 0                 | <.0001|
| years?    | 0      | 16.71    | 0                 | <.0001|
| trpc      | 0      | 7.49     | 0                 | <.0001|
| batavgc   | 0      | 2.16     | 0                 | 0.0519|

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Theta0</th>
<th>F for H0</th>
<th>Parameter=Theta0</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>7499.13</td>
<td>0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>years?</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>trpc</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>batavgc</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

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

Marginal plots for missing data
- 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, ~10% of observations with missing salary also had batavgc set to missing.

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, misses as typical value (mean, median), with error bars for variability, over the m imputed data sets.
- miplot macro: takes input data, imputed data → plot for $(x, y)$; std. error bars for missing on one, diamonds for missing on both.

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