Solution to missing data exercises

Statistical analysis of correlated and repeated measurements

Exercise 1

A randomized controlled was performed to evaluate the effect of topical zink sulfate on epidermal wound healing. Thirty healthy volunteers (half male, half female) had suction-blister wounds induced on each buttock and were randomized to receive two out of the following three treatments:

A. Active treatment with a zink shower gel.
B. Placebo treatment with a gel with contents identical to A except from containing zink sulfate.
C. Control treatment with demineralized water.

Several clinical outcomes were evaluated in the trial, but in this exercise we will only be concerned with the pain sensation experienced for each wound just after application of the treatments. This was assessed by VAS-scores (0-100mm visual analogue scale) after each treatment application and summarized by area under the curve for the whole treatment course.


1. Read in the data from the file vasscores.txt.

data vaswide;
infile "C:\Documents\mydata\vasscores.txt" FIRSTOBS=2;
input id group $ vasA vasB vasC;
run;

- How many subjects were allocated to each of the groups AB, AC, and BC?

proc freq data=vaswide;
table group;
run;

The FREQ Procedure

<table>
<thead>
<tr>
<th>group</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>10</td>
<td>33.33</td>
<td>10</td>
<td>33.33</td>
</tr>
<tr>
<td>AC</td>
<td>10</td>
<td>33.33</td>
<td>20</td>
<td>66.67</td>
</tr>
<tr>
<td>BC</td>
<td>10</td>
<td>33.33</td>
<td>30</td>
<td>100.00</td>
</tr>
</tbody>
</table>
We see that subjects were distributed evenly on the three allocation groups.

- Make summary statistics for the variables `vasA`, `vasB`, and `vasC`. How many observations of each variable are missing?

We run `proc corr` to produce relevant summary statistics. To assess whether data is approximately normally distributed we use the option `plots=all` to make pairwise scatterplots.

```plaintext
ods graphics on;

proc corr data=vaswide plots=all;
var vasA vasB vasC;
run;
```

Further output looks as follows:

```
The CORR Procedure

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>vasA</td>
<td>20</td>
<td>64.62500</td>
<td>44.68424</td>
<td>1293</td>
<td>3.50000</td>
<td>177.0000</td>
</tr>
<tr>
<td>vasB</td>
<td>20</td>
<td>78.12500</td>
<td>46.52217</td>
<td>1563</td>
<td>13.00000</td>
<td>166.5000</td>
</tr>
<tr>
<td>vasC</td>
<td>20</td>
<td>56.95000</td>
<td>46.61062</td>
<td>1139</td>
<td>3.50000</td>
<td>168.5000</td>
</tr>
</tbody>
</table>
```

As appears, the highest average vasscores were found for treatment B and the lowest for treatment C. Standard deviations are fairly high compared to the sample means,
which implies a notable inter-individual variation in pain sensation. Since vas-scores cannot be negative, the large standard deviations could be an indication of skewness. However, save from a few outliers, the deviations from the normal distribution are not that obvious, so we proceed with the untransformed data. There are 20 observations of response to treatment A implying that the ten remaining subjects have missing values of $\text{vasA}$. These must be the ten from allocation group BC who never received treatment A in the first place. Likewise for treatments B and C. We finally note that the correlations between the paired observation are overall high, but the highest correlation is between the responses to treatments A and B which were similarly composed save from the zink sulfate:

\[
\begin{array}{c|c|c|c}
& \text{vasA} & \text{vasB} & \text{vasC} \\
\hline
\text{vasA} & 1.00000 & 0.92145 & 0.83667 \\
& 0.0002 & 0.0025 & \\
20 & 10 & 10 & \\
\hline
\text{vasB} & 0.92145 & 1.00000 & 0.65760 \\
& 0.0002 & 0.0388 & \\
10 & 20 & 10 & \\
\hline
\text{vasC} & 0.83667 & 0.65760 & 1.00000 \\
& 0.0025 & 0.0388 & \\
10 & 10 & 20 & \\
\end{array}
\]

- **What type of missing data are we dealing with?**
  We are dealing with data that are *missing completely at random* (MCAR). E.g. the ten persons that are missing $\text{vasA}$ were randomly allocated to the BC-group and this is the reason why their response to treatment A is missing. Similarly for the other missing treatment responses.

- **Would it make sense to perform a complete case analysis on these data?**
  Obviously not, as there are no complete cases in this data.

2. **Make two simple analyses to compare the response to treatment A and C:**
   I. A paired t-test based on the data from allocation group AB.
   II. A two-sample t-test based on the data from allocation groups AC and BC.

   **Estimate the mean difference in response between treatment A and C and its standard error.**
   We pick out the data from allocation group AC and apply a paired t-test to compare the mean pain sensation experienced with these two treatments:

   \[
   \begin{array}{c|c|c|c}
   & \text{vasA} & \text{vasB} & \text{vasC} \\
   \hline
   \text{vasA} & 1.00000 & 0.92145 & 0.83667 \\
   & 0.0002 & 0.0025 & \\
   20 & 10 & 10 & \\
   \hline
   \text{vasB} & 0.92145 & 1.00000 & 0.65760 \\
   & 0.0002 & 0.0388 & \\
   10 & 20 & 10 & \\
   \hline
   \text{vasC} & 0.83667 & 0.65760 & 1.00000 \\
   & 0.0025 & 0.0388 & \\
   10 & 10 & 20 & \\
   \end{array}
   \]

   ```
   data paired; set vaswide; if group eq 'AC'; run;
   proc ttest data=paired;
   paired vasA*vasC;
   run;
   ```

3
The TTEST Procedure

Difference:  vasA - vasC

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>16.5000</td>
<td>24.5028</td>
<td>7.7485</td>
<td>26.0000</td>
<td>51.5000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.5000</td>
<td>-1.0283</td>
<td>34.0283</td>
<td>24.5028</td>
<td>16.8539</td>
</tr>
</tbody>
</table>

|       | DF | t Value | Pr > |t| |
|-------|----|---------|------|
|       | 9  | 2.13    | 0.0621 |

The estimated difference is 16.5 (SE 7.7), which result in a p-value just above 5%.

Next we combine the relevant data from allocation groups AB and BC and carry out an unpaired t-test with:

data unpaired; set vaswide; if group ne 'AC';
if group eq 'AB' then vas=vasA;
if group eq 'BC' then vas=vasaucC;
run;

proc ttest data=unpaired;
class group; var vas;
run;

The TTEST Procedure

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>10</td>
<td>60.9500</td>
<td>47.4883</td>
<td>15.0171</td>
<td>9.5000</td>
<td>177.0</td>
</tr>
<tr>
<td>BC</td>
<td>10</td>
<td>62.1000</td>
<td>53.0570</td>
<td>16.7781</td>
<td>7.5000</td>
<td>168.5</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>1.1500</td>
<td>50.3497</td>
<td>22.5171</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>group</th>
<th>Method</th>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>60.9500</td>
<td>26.9789</td>
<td>94.9211</td>
<td>47.4883</td>
<td>32.6641</td>
<td>86.6951</td>
</tr>
<tr>
<td>BC</td>
<td>62.1000</td>
<td>24.1453</td>
<td>100.1</td>
<td>53.0570</td>
<td>36.4945</td>
<td>96.8614</td>
</tr>
<tr>
<td>Diff (1-2) Pooled</td>
<td>-1.1500</td>
<td>-48.4566</td>
<td>46.1566</td>
<td>50.3497</td>
<td>38.0449</td>
<td>74.4583</td>
</tr>
<tr>
<td>Diff (1-2) Satterth</td>
<td>-1.1500</td>
<td>-48.4980</td>
<td>46.1980</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|       | Method | Variances | DF  | t Value | Pr > |t| |
|-------|--------|-----------|-----|---------|------|
|       | Pooled | Equal     | 18  | -0.05   | 0.9598 |
|       | Satterthwaite | Unequal | 17.738 | -0.05  | 0.9598 |

This yields an estimated difference of -1.2 (SE 22.5) which is insignificant (P=0.96).

Why can’t we just analyse all of the data at the same time using a suitable t-test?

Preferably, the conclusion about the treatment difference should be based on a single estimate using all available data. However, there is no such thing as a half-way paired half-way unpaired t-test, so we need a mixed model to do the optimal analysis.
3. Transform data to the long format and make spaghettiplots to illustrate the data from the three allocation groups.

``` Sas 
data vaslong; set vaswide;
treat='A'; vas=vasA; output;
treat='B'; vas=vasB; output;
treat='C'; vas=vasC; output;
keep id group treat vas;
run;
```

``` Sas 
proc sgpanel data=vaslong noautolegend;
panelby group / columns=3;
series x=treat y=vas / group=id lineattrs=(pattern=solid);
run;
```

4. Make an analysis of the data based on an appropriate mixed model.

We apply a mixed model with fixed effect of treatment and an unstructured covariance for the repeated measurements on each subject.

``` Sas 
proc mixed data=vaslong plots=all;
class id treat;
model vas=treat / ddfm=kr vciry;
lsmeans treat / diff cl;
repeated treat / subject=id type=un;
run;
```

``` Question 2 ```

![Image of Question 2 graph showing spaghettiplots for different groups A, B, and C with repeated measurements over time.]
The `lsmeans`-statement is convenient for estimating all of the treatment means and the differences between them and `plots=all` produces residual diagnostics. Due to the small sample size, the high correlation between the outcomes, and the missing data, it is recommended to use the `scaled residuals` invoked by the `vciry`-option following the `model`-statement.

The residual plots indicate a decent model fit.

Select output from proc mixed is shown below:

- **Estimate the mean responses for treatments A, B, and C and their standard errors.**

  **Least Squares Means**

  | Effect | treat | Estimate | Error | DF | t Value | Pr > |t| |
  |--------|--------|----------|-------|----|---------|-------|---|
  | treat  | A      | 70.1231  | 8.5888| 27 | 8.16    | <.0001|
  | treat  | B      | 77.3554  | 9.0145| 27 | 8.58    | <.0001|
  | treat  | C      | 53.7009  | 9.3979| 24 | 5.71    | <.0001|

- **Estimate the difference between treatments A and C and its standard error. How does this compare with the results of the two t-tests from question 2?**

  **Differences of Least Squares Means**

  | Effect | treat _treat | Estimate | Error | DF | t Value | Pr > |t| |
  |--------|--------------|----------|-------|----|---------|-------|---|
  | treat  | A B          | -7.2323  | 5.9799| 13 | -1.21   | 0.2477|
  | treat  | A C          | 16.4222  | 7.2863| 15 | 2.25    | 0.0396|
  | treat  | B C          | 23.6545  | 8.3451| 26 | 2.83    | 0.0086|
Contrary to the two t-test, we now find that the pain sensation experienced with treatment A is significantly higher than with treatment C. The estimated difference is similar to the one we found with the paired t-test but the standard error has decreased since we are making use of all the data.

- **Find the estimated standard deviations of the responses to treatment A, B, and C and the correlations between them.**

We find the estimated variances and covariances in the output from proc mixed:

\[
\text{Covariance Parameter Estimates}
\]

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>id</td>
<td>2038.72</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>id</td>
<td>1892.17</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>id</td>
<td>2111.42</td>
</tr>
<tr>
<td>UN(3,1)</td>
<td>id</td>
<td>1797.46</td>
</tr>
<tr>
<td>UN(3,2)</td>
<td>id</td>
<td>1369.79</td>
</tr>
<tr>
<td>UN(3,3)</td>
<td>id</td>
<td>2157.88</td>
</tr>
</tbody>
</table>

From this we compute the standard deviations and correlations:

\[
\begin{align*}
\sigma_A &= \sqrt{2038.72} = 45.15 \\
\rho_{AB} &= \frac{1892.17}{\sqrt{2038.72 \cdot 2111.42}} = 0.91 \\
\sigma_B &= \sqrt{2111.42} = 45.95 \\
\rho_{BC} &= \frac{1369.79}{\sqrt{2111.42 \cdot 2157.88}} = 0.64 \\
\sigma_C &= \sqrt{2157.88} = 46.45 \\
\rho_{AC} &= \frac{1797.46}{\sqrt{2111.42 \cdot 2157.88}} = 0.86
\end{align*}
\]

which are quite similar to the summary statistics we found in question 1.

5. **Make an analysis based on a dataset where missing data have been replaced by mean value predictions.**

Using data in the wide format, we insert the treatment means from question 1 in place of the missing data in each allocation group.

```plaintext
data widemv;
set vaswide;
if group eq 'AB' then vasC=56.950;
if group eq 'AC' then vasB=78.125;
if group eq 'BC' then vasA=64.625;
run;
```
• Make spaghetti plots to illustrate the resulting data.
We transform data to the long format so that we can draw the spaghetti plots.

data longmv; set widemv;
treat='A'; vas=vasA; output;
treat='B'; vas=vasB; output;
treat='C'; vas=vasC; output;
keep id group treat vas;
run;

proc sgpanel data=longmv noautolegend;
panelby group / columns=3;
series x=treat y=vas / group=id lineattrs=(pattern=solid);
run;

The only good think that can be said about mean value imputation is that is it easy do to. The resulting data looks very unnatural since there is no variation in the imputed values. It is difficult to imagine any situation where valid statistical results would arise when using this very naive imputation strategy!

• Apply proc means to obtain estimates of the mean response for treatments A and B and C and their standard errors. How does these results compare with the mixed model results?

proc means data=widemv n mean stderr;
var vasA vasB vasC;
run;
The MEANS Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>vasA</td>
<td>30</td>
<td>64.625</td>
<td>6.6034593</td>
</tr>
<tr>
<td>vasB</td>
<td>30</td>
<td>78.125</td>
<td>6.8750705</td>
</tr>
<tr>
<td>vasC</td>
<td>30</td>
<td>56.95</td>
<td>6.8881416</td>
</tr>
</tbody>
</table>

There are no missing values in the imputed dataset so the means of the three treatment responses and their standard errors can be computed straight-forwardly as explained in any basic statistics course. We note that the estimated means are exactly the same as for the observed data in question 1. These estimates are similar to the estimates we obtained the mixed model and both are unbiased since data is MCAR. However, the standard errors are notably smaller than the estimates we got from the mixed model. This is due to the reduced variation in the mean value imputations. A highly unwarranted side-effect of naive imputations is that they tend to make themselves into self-fulfilling prophecies. Not only have we replaced the missing data with unrealistic guesses; we also act as if we are much more certain about our statistical results due to these!

- **Compute the variances of and correlations between the treatment responses. How do they compare to the estimates obtained from the mixed model?**

  We run `proc corr` on the imputed dataset:

```r
proc corr data=widemv;
var vasA vasB vasC;
run;
```

The CORR Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>vasA</td>
<td>30</td>
<td>64.625</td>
<td>36.168</td>
<td>1939</td>
<td>3.500</td>
<td>177.000</td>
</tr>
<tr>
<td>vasB</td>
<td>30</td>
<td>78.125</td>
<td>37.656</td>
<td>2344</td>
<td>13.000</td>
<td>166.500</td>
</tr>
<tr>
<td>vasC</td>
<td>30</td>
<td>56.95</td>
<td>37.728</td>
<td>1709</td>
<td>3.500</td>
<td>168.500</td>
</tr>
</tbody>
</table>

**Pearson Correlation Coefficients, N = 30**

<table>
<thead>
<tr>
<th></th>
<th>vasA</th>
<th>vasB</th>
<th>vasC</th>
</tr>
</thead>
<tbody>
<tr>
<td>vasA</td>
<td>1.00000</td>
<td>0.50615</td>
<td>0.34116</td>
</tr>
<tr>
<td></td>
<td>0.0043</td>
<td>0.0650</td>
<td></td>
</tr>
<tr>
<td>vasB</td>
<td>0.50615</td>
<td>1.00000</td>
<td>0.33908</td>
</tr>
<tr>
<td></td>
<td>0.0668</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vasC</td>
<td>0.34116</td>
<td>0.33908</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0668</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard deviations and the correlations in the mean value imputed data are substantially smaller than in the observed data (question 1) and than estimated by the mixed model (question 4). This could bias the comparison of the treatment in either direction.
Use a paired t-test to estimate the difference between treatments A and C and its standard error. Compare the result with question 4.

```
proc ttest data=widemv;
paired vasA*vasC;
run;
```

The TTEST Procedure

Difference: vasA - vasC

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>7.6750</td>
<td>42.4322</td>
<td>7.7470</td>
<td>-103.9</td>
<td>120.1</td>
</tr>
</tbody>
</table>

Mean 95% CL Mean Std Dev 95% CL Std Dev
7.6750 -8.1694 23.5194 42.4322 33.7933 57.0422

DF t Value Pr > |t|
29 0.99 0.3300

From this we get an estimated difference of 7.7 (SE 7.7, P=0.33) between treatments A and C. By coincidence the standard error is similar to that from the mixed model but otherwise the two approaches yield different results and only the mixed model results are to be trusted.

**Exercise 2**

Run `calcium_demo2.sas` to reproduce the multipel imputations from the lecture.

Follow the instructions within the demo that will take you through the following steps:

1. Read in the data from `calcium.txt` (it’s in the wide format).
2. Make a long version of the data so that you can compare the data from completers and non-completers in the two treatment groups in spaghettiplots.
3. For benchmarking run the standard cLMM analysis to estimate treatment effect at final follow-up. Check whether the multivariate normal distribution fits the data well.
4. Make a multiply imputed dataset by applying `proc mi` to data in the wide format.
5. Transform to the long format and make spaghettiplots to visualize some of the multiple imputations.
6. Run the cLMM on the imputed data by _Imputation_ and save the resulting multiple estimates of the treatment effect in an output dataset.
7. Use `proc mianalyze` to combine the results into a final estimate and standard error.