Taking a step towards reproducibility in research

A Michelle Edwards, Ph.D., MLIS

September 9, 2019

Table of Contents


```
## SAS found at C:/Program Files/SASHome/SASFoundation/9.4/sas.exe
## sas, saslog, sashtml, sashtml5, and sashtmllog & sashtml5log engines
##    are now ready to use.
```

What is Reproducible Research?

Definitions

Reproducible:

- Able to be shown, done, or made again (https://dictionary.cambridge.org/dictionary/english/reproducible)
- Able to be reproduced or copied (https://www.lexico.com/en/definition/reproducible)
- Capable of being reproduced (https://www.vocabulary.com/dictionary/reproducible)

Using these definitions - how would you define "Reproducible Research"?

"Reproducible Crisis"

A study has claimed that up to 2/3 of researchers have tried and failed to reproduce another scientist's experiments (https://www.bbc.com/news/science-environment-39054778).

Experiments are meant to be replicable or reproducible. Materials and Methods section should be similar to a recipe - with step-by-step type of guidelines to follow. Data collection and statistical analysis methods form part of the Materials and Methods section of papers and theses.

We want to aim to make our research **transparent**, **open**, and **reproducible**.

Exercise No. 1

Read the Materials and Methods section of the paper that was handed out. Working with a partner, can you identify the steps that were used for data collection and statistical analysis? Is anything missing? Can you identify all the data (variables) that were collected? Did the author(s) create any new variables for their analysis? Did they delete any observations? Were there any transformations? What analysis did they conduct? Did they check the assumptions of the analysis they used?

What can YOU do to help make your research REPRODUCIBLE?

1) _

2) _

3) _

4) _

5) _

Let's discuss and work through a couple of ways you can incorporate into your data collection and analysis to increase the transparency and reproducibility of your work.

Documenting your R or SAS syntax

R syntax

Here is a sample of R syntax:

```
#  Using the base aov() function in base R this will conduct an ordinary leas
t squares ANOVA            #


model <- aov(nitrogen ~ block + trmt, data=rcbd_data)
model

summary(model)


#  Let's fix our block and trmt effects
#


rcbd_data$block_fac = as.factor(rcbd_data$block)
rcbd_data$trmt_fac = as.factor(rcbd_data$trmt)

#  Running the aov() with block and trmt as factors
#
model2 <- aov(nitrogen ~ block_fac + trmt_fac, data=rcbd_data)
model2

#  Run summary on your new updated model
summary(model2)
```

```
#  What if we wanted to see Tukey means comparisons?
model3 <- aov(nitrogen ~ block_fac + factor(trmt_fac), data=rcbd_data)
TukeyHSD(model3,which='factor(trmt_fac)')
summary(model3)
```

SAS syntax

Here is a sample of SAS code:

```
*  Partioning of the Variation for an RCBD trial;
Proc glimmix data=rcbd;
  class block trmt;
  model Nitrogen = trmt/s;
  random block;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid
 student(noblup)=smresid;
Run;
```

Exercise No. 2

Working with a partner, can you identify the steps that were used in the statistical analysis? Do you understand what the analysis is doing? Are these pieces of syntax helpful to have in a thesis? or accessible to a published paper?

Adding Comments to your code

For R syntax - you can add information embedded in your code by using the “#” at the beginning of the line of code. In the sample code above - you will notice that there are a number of comments included in the syntax. You can make these are long or as short as you wish. These will help YOU remember what you did and why you did it.

For SAS syntax - you have two options to add comments in your program. If you only have one line - then you would add an “*" at the beginning of the line and a “;” at the end of the line. If you have several lines to add as a comment you can add “/ *" at the beginning and a “* /” at the end of your comment.

Exercise No. 3

Take a few moments to review the R and the SAS code in the previous section to see how comments have been used. Would you add more comments in the code? Could you add more comments? Why?

Code and results ALL together

Whether you use R or SAS, your syntax or code is in one file, while your output is usually in a second file. We tend to create the code, run the code, and save our output. We may then print the output and use it to writeup our results. What happens when you look at the output, say 6 months down the road and you notice a problem? Maybe you decide that you would like to rerun the code. Will you be able to find the exact code you ran? Chances are you've moved on from that code and modified it.

Maybe the answer is to create a document where you see your code and your output together?

MARKDOWN

Check out these examples - the first one for R and the second for SAS.

1. Example using R Markdown:

```
# Analysis conducted 20190913
# Reading data from the Excel file that contains the RCBD trial sample data u
sing the READXL package
library(readxl)

rcbd <- read_excel("~/Workshops/Common_Files/RCBD_excel.xlsx", col_names=T)
rcbd

# A tibble: 24 x 5
   block  trmt Nitrogen  Weed Bin_weed
   <dbl> <dbl>    <dbl> <dbl>    <dbl>
 1     1     1     35.0    81     0.81
 2     2     1     41.2    87     0.87
 3     3     1     36.9    89     0.89
 4     4     1     40.0    79     0.79
 5     1     2     40.9    88     0.88
 6     2     2     46.7    85     0.85
 7     3     2     46.6    99     0.99
 8     4     2     41.9    86     0.86
 9     1     3     42.1    54     0.54
10     2     3     49.4    23     0.23
# ... with 14 more rows

# Running the Mixed Model ANOVA for the RCBD trial data using the LME4 packag
e
library(lme4)

Loading required package: Matrix
```

```
model1 <- lmer(Nitrogen ~ ((1|block) + trmt), data=rcbd)
model1

Linear mixed model fit by REML ['lmerMod']
Formula: Nitrogen ~ ((1 | block) + trmt)
   Data: rcbd
REML criterion at convergence: 135.3343
Random effects:
 Groups   Name        Std.Dev.
 block    (Intercept) 2.871
 Residual             4.025
Number of obs: 24, groups:  block, 4
Fixed Effects:
(Intercept)         trmt
    41.5702       0.1433
```

2. Example using SAS Markdown

```
* Analysis conducted 20190913;
* Reading data that contains the RCBD trial sample data;

Data rcbd;
  input block trmt Nitrogen Weed Bin_weed;
  datalines;
1 1 34.98 81 0.81
2 1 41.22 87 0.87
3 1 36.94 89 0.89
4 1 39.97 79 0.79
1 2 40.89 88 0.88
2 2 46.69 85 0.85
3 2 46.65 99 0.99
4 2 41.90 86 0.86
1 3 42.07 54 0.54
2 3 49.42 23 0.23
3 3 52.68 11 0.11
4 3 42.91 16 0.16
1 4 37.18 10 0.10
2 4 45.85 23 0.23
3 4 40.23 10 0.10
4 4 39.20 69 0.69
1 5 37.99 61 0.61
2 5 41.99 06 0.06
3 5 37.61 02 0.02
4 5 40.45 75 0.75
1 6 34.89 21 0.21
2 6 50.15   08 0.08
3 6 44.57   27 0.27
4 6 43.29   03 0.03
;
Run;


* Running the Mixed Model ANOVA for the RCBD trial data;

Proc glimmix data=rcbd;
  class block trmt;
  model Nitrogen = trmt/s;
  random block;
  title "Proc GLIMMIX Results";
  lsmeans trmt / pdiff adjust=tukey ilink lines;
  output out=second predicted=pred residual=resid residual(noblup)=mresid stu
dent=studentresid student(noblup)=smresid;
Run;
```

Proc GLIMMIX Results

The GLIMMIX Procedure

Model Information

Data Set	WORK.RCBD
Response Variable	Nitrogen
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
block	4	1 2 3 4
trmt	6	1 2 3 4 5 6

Number of Observations Read	22
Number of Observations Used	21

Dimensions

G-side Cov. Parameters	1
R-side Cov. Parameters	1
Columns in X	7
Columns in Z	4
Subjects (Blocks in V)	1
Max Obs per Subject	21

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled
Residual Variance	Profiled
Starting From	Data

Iteration History

Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient

0	0	4	82.130540078	.		2.66E-15

Convergence criterion (ABSGCONV=0.00001) satisfied.

Fit Statistics

-2 Res Log Likelihood	82.13
AIC (smaller is better)	86.13
AICC (smaller is better)	87.13
BIC (smaller is better)	84.90
CAIC (smaller is better)	86.90
HQIC (smaller is better)	83.44
Generalized Chi-Square	92.02
Gener. Chi-Square / DF	6.13

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
block	6.2503	6.1256
Residual	6.1344	2.5044

Solutions for Fixed Effects

Effect	trmt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		37.5808	2.9097	3	12.92	0.0010
trmt	1	0.6967	2.9047	12	0.24	0.8145
trmt	2	6.4517	2.9047	12	2.22	0.0463
trmt	3	9.1892	2.9047	12	3.16	0.0082
trmt	4	3.0342	2.9047	12	1.04	0.3168
trmt	5	1.9292	2.9047	12	0.66	0.5191
trmt	6	0	.	.	.	.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trmt	5	12	6.86	0.0031

trmt Least Squares Means

trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean

1	38.2775	1.7596	12	21.75	<.0001	38.2775	1.7596
2	44.0325	1.7596	12	25.02	<.0001	44.0325	1.7596
3	46.7700	1.7596	12	26.58	<.0001	46.7700	1.7596
4	40.6150	1.7596	12	23.08	<.0001	40.6150	1.7596
5	39.5100	1.7596	12	22.45	<.0001	39.5100	1.7596
6	37.5808	2.9097	12	12.92	<.0001	37.5808	2.9097

Differences of trmt Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

trmt	_trmt	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
1	2	-5.7550	1.7513	12	-3.29	0.0065	0.0565
1	3	-8.4925	1.7513	12	-4.85	0.0004	0.0041
1	4	-2.3375	1.7513	12	-1.33	0.2068	0.7619
1	5	-1.2325	1.7513	12	-0.70	0.4950	0.9779
1	6	0.6967	2.9047	12	0.24	0.8145	0.9999
2	3	-2.7375	1.7513	12	-1.56	0.1440	0.6347
2	4	3.4175	1.7513	12	1.95	0.0748	0.4198
2	5	4.5225	1.7513	12	2.58	0.0240	0.1753
2	6	6.4517	2.9047	12	2.22	0.0463	0.2961
3	4	6.1550	1.7513	12	3.51	0.0043	0.0385
3	5	7.2600	1.7513	12	4.15	0.0014	0.0133
3	6	9.1892	2.9047	12	3.16	0.0082	0.0692
4	5	1.1050	1.7513	12	0.63	0.5399	0.9863
4	6	3.0342	2.9047	12	1.04	0.3168	0.8936
5	6	1.9292	2.9047	12	0.66	0.5191	0.9828

Conservative Tukey-Kramer Grouping for
trmt Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

trmt	Estimate	
3	46.7700	A
		A
2	44.0325	A
		A
4	40.6150	A
		A
5	39.5100	A
		A
1	38.2775	A
		A
6	37.5808	A

```
    The LINES display does not reflect all significant comparisons. The
              following additional pairs are significantly
                     different: (3,4), (3,5), (3,1).
```

What can you do with R markdown?

1. Keep code and results together - in one file!
2. A way to archive your code
3. A way to submit your code to journals

How do you create R markdown

1. Create PDF, Word, HTML documents
2. R markdown documents are made up of R (SAS) code and notes - like you would in a Word or text editor.
3. Each R or SAS code section creates the output.

R Markdown example

One way to learn R markdown is to review an example. This document was created using R markdown. The RMD file or R markdown file used to create this document is available for download on the blog.

Concluding remarks and thoughts

- We should all try to ensure that our research is transparent and reproducible
- Including comments in your syntax or code is a great place to start. It can help you remember what you did when you've been away from your analysis for a while. It can also encourage you to be more efficient with your coding, explaining what you did every step. YES! It will take time, and it can be difficult to slow down and add all the information into your code as you work.
- Creating a document that contains both your syntax/code along with the results is a great way to keep your syntax/code together. Requires time to learn the R markdown and it needs you to spend more time documenting. But it will be worth the extra work in the end.