Reproducible Research

Dave Kent

10/21/2016

Remember our core values?

- Constant improvement
- Reproducible research
- Team work
- Excellence in execution

Remember our core values?

- Constant improvement
- **Reproducible research**
- Team work
- Excellence in execution

What is reproducible research?

- Reproducibility is the degree to which an experiment can be duplicated
- A complete and truthful description of methods
- Ambiguity → less reproducible

What is reproducible research?

- Not reproducible:
  - “Each BT raw milk samples was tested for mesophilic spores.”
- Reproducible:
  - “Aliquots of each BT raw milk sample were spore pasteurized at 80°C for 12 min, followed by rapid cooling to ≤6°C. Spore pasteurized samples were spiral plated onto brain heart infusion agar in duplicate and incubated at 32°C for 48 h for a mesophilic spore count.”

What is reproducible research?

- Not reproducible:
  - “A linear model was fit to the data to investigate the effect of temperature on bacteria count.”
- Reproducible:
  - “Using the lm function in R, a linear model was fit to the data with log-transformed bacteria count as the response, temperature as a main effect, and test day as a blocking effect. Effect of temperature on log bacteria count was assessed with the lsmeans package in R.”

Why is reproducible research important?

- …that's kind of the essence of the scientific method…
  - If I want to test a claim I find in a paper, I need to know *how* to test it.
  - If I want to test a hypothesis, I need to know the conditions where the hypothesis is supposed to hold
- Practically, it makes our lives easier
  - When Martin thinks your results are weird, and you know exactly what happened, you can blame the weirdness on the method or the samples! (ok yeah sure, now you have to justify why you chose a weird method)

A (mostly fictional) anecdote

- You're in a meeting with Martin, and in your manuscript he reads:
- "The geometric mean MS level for samples with spores detected by direct plating (≥10 spores/mL) was **3400 spores/mL**. The geometric mean TS level for samples with spores detected by direct plating (≥10 spores/mL) was **14 spores/mL**."

A (mostly fictional) anecdote

- “The geometric mean MS level for samples with spores detected by direct plating (≥10 spores/mL) was **3400 spores/mL**. The geometric mean TS level for samples with spores detected by direct plating (≥10 spores/mL) was **14 spores/mL**.”
- Martin says “That 590 spores/mL seems too high. I need you to double check these numbers.”
- Seems like a reasonable request – he’s probably right, as usual. Maybe I made a transcription error.

A (mostly fictional) anecdote

- “The geometric mean MS level for samples with spores detected by direct plating (≥10 spores/mL) was **3400 spores/mL**. The geometric mean TS level for samples with spores detected by direct plating (≥10 spores/mL) was **14 spores/mL**.”
- I go back to the spreadsheet and check.
- A-ha! The geometric mean MS level is **33.6 spores/mL**. I probably rounded and missed a decimal place or something.
- I’ll just go ahead and check the TS level for kicks. Oh. Oh, no. **16.3 spores/mL**?

A (mostly fictional) anecdote

- At this point, we’ve already produced the *data*
- We are having trouble reproducing the *analysis*

A (mostly fictional) anecdote

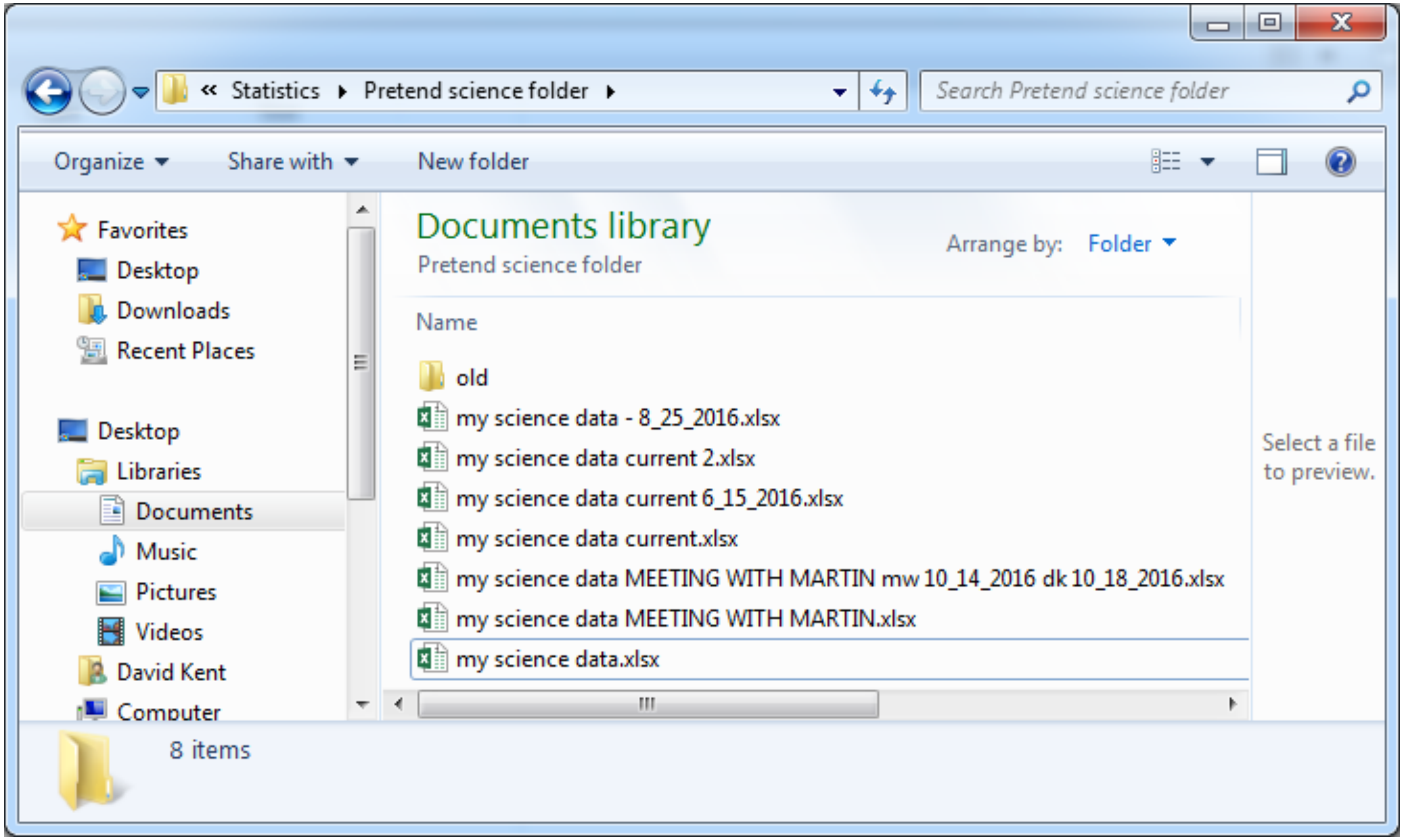
- At this point, we've already produced the *data*
- We are having trouble reproducing the *analysis*
- **Even though we are the ones that did it in the first place!**

Ugh, what’s the problem?

- We adjust the raw data by correcting, transforming, normalizing
- We don’t always keep good track of the age or succession of our data files.

Ugh, what’s the problem?

- We often don’t know what we did:



Documents library

Pretend science folder

Name

old

my science data - 8_25_2016.xlsx

my science data current 2.xlsx

my science data current 6_15_2016.xlsx

my science data current.xlsx

my science data MEETING WITH MARTIN mw 10_14_2016 dk 10_18_2016.xlsx

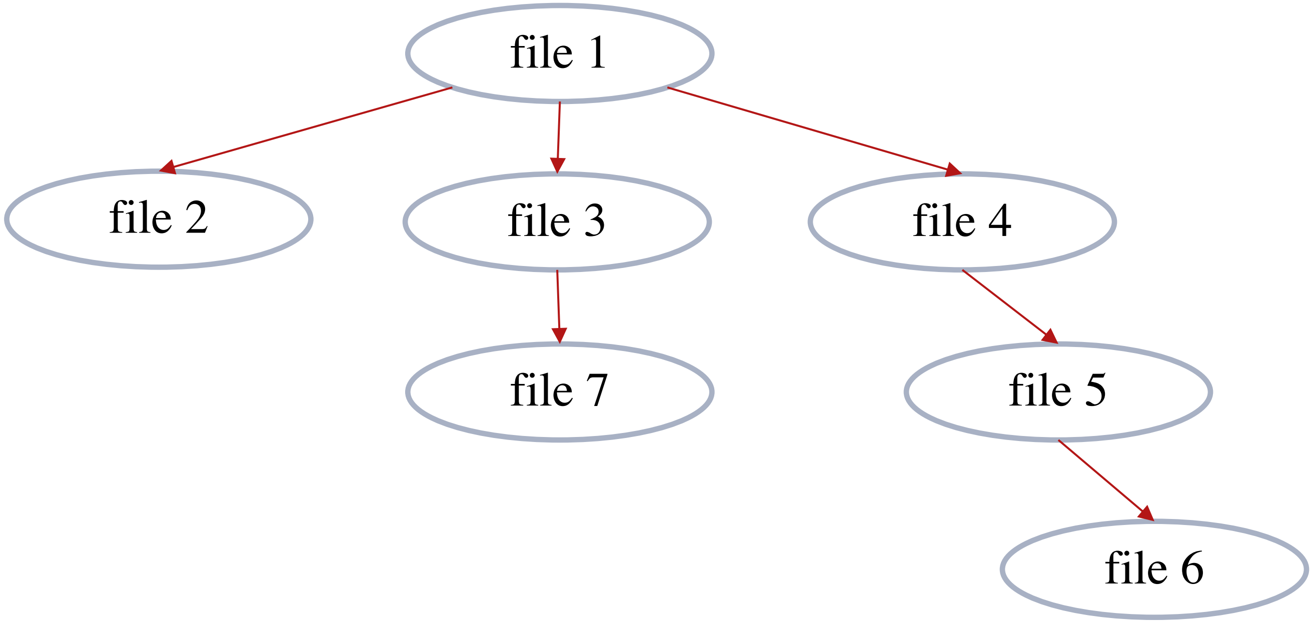
my science data MEETING WITH MARTIN.xlsx

my science data.xlsx

8 items

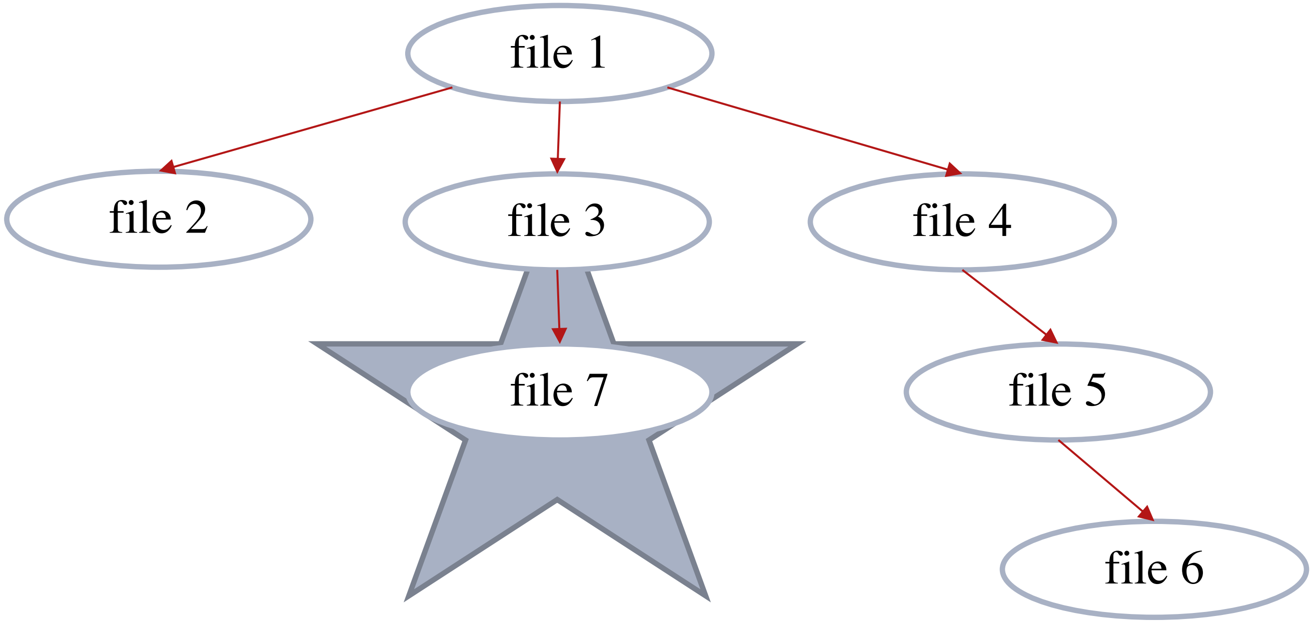
Ugh, what's the problem?

- We often don't know what we did:



Ugh, what’s the problem?

- We often don’t know what we did:

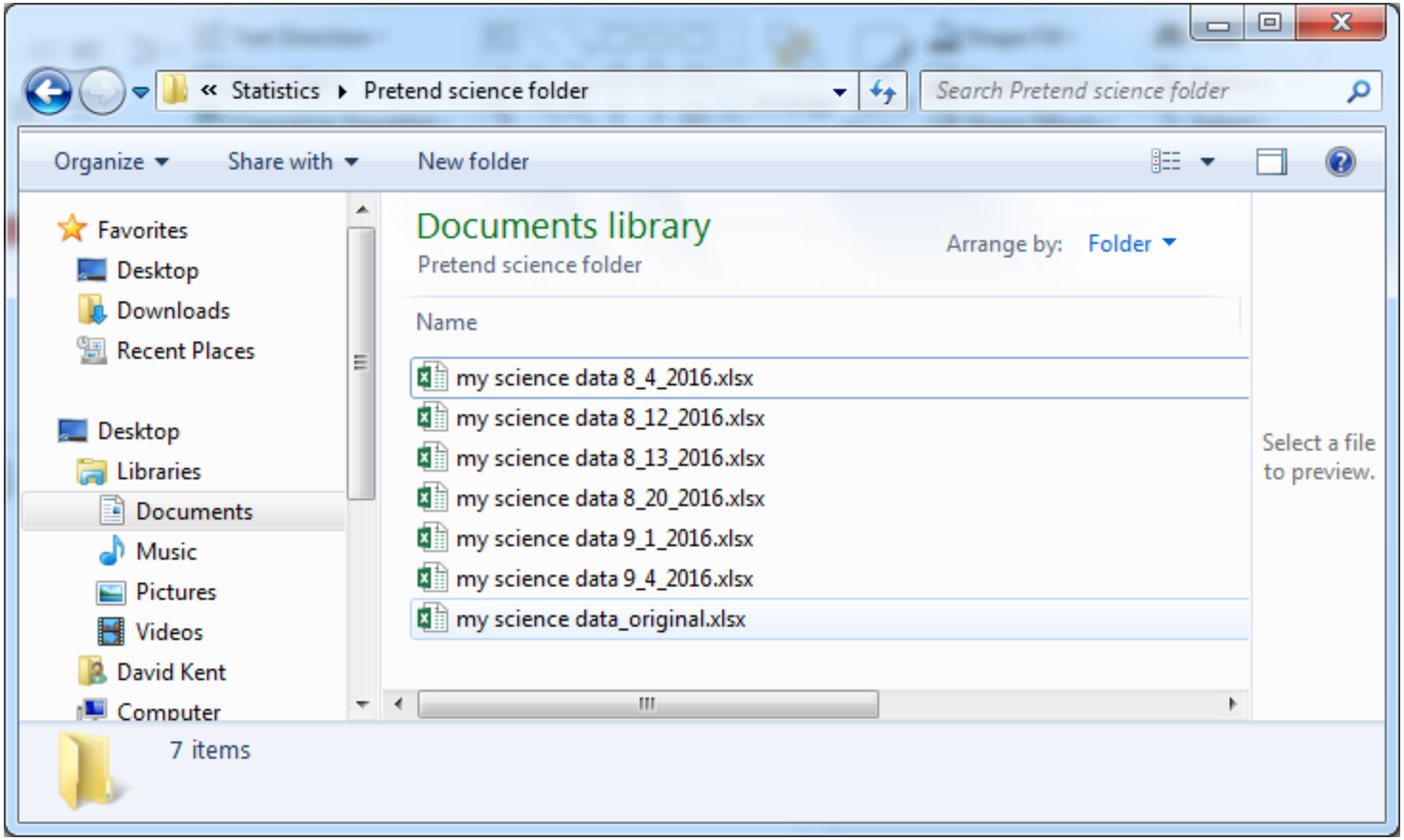


Constant improvement

- Let’s do better!
- There are solutions for someone of any technical proficiency

Low-tech

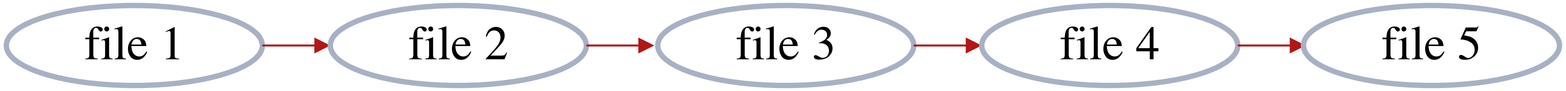
- Take control of the succession of your files:



my science data 8_4_2016.xlsx
my science data 8_12_2016.xlsx
my science data 8_13_2016.xlsx
my science data 8_20_2016.xlsx
my science data 9_1_2016.xlsx
my science data 9_4_2016.xlsx
my science data_original.xlsx

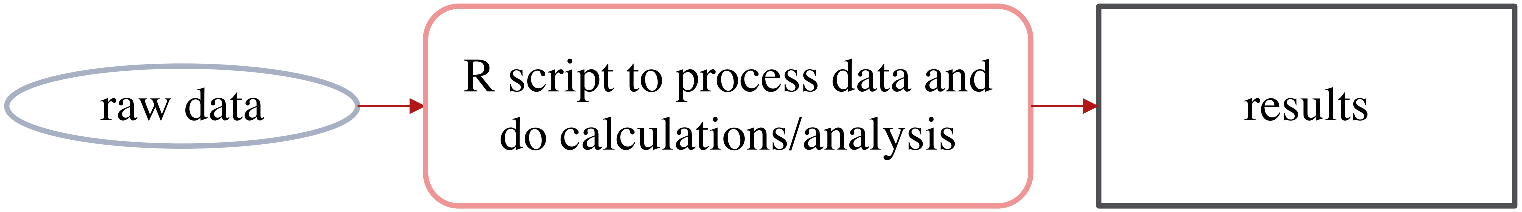
Low-tech

- Take control of the succession of your files:
  - Always save-as before you start editing
  - Always date the files
- Now the most recent file ought to be the “most correct,” and we have a snapshot of the history



Medium-tech (this is what I usually do)

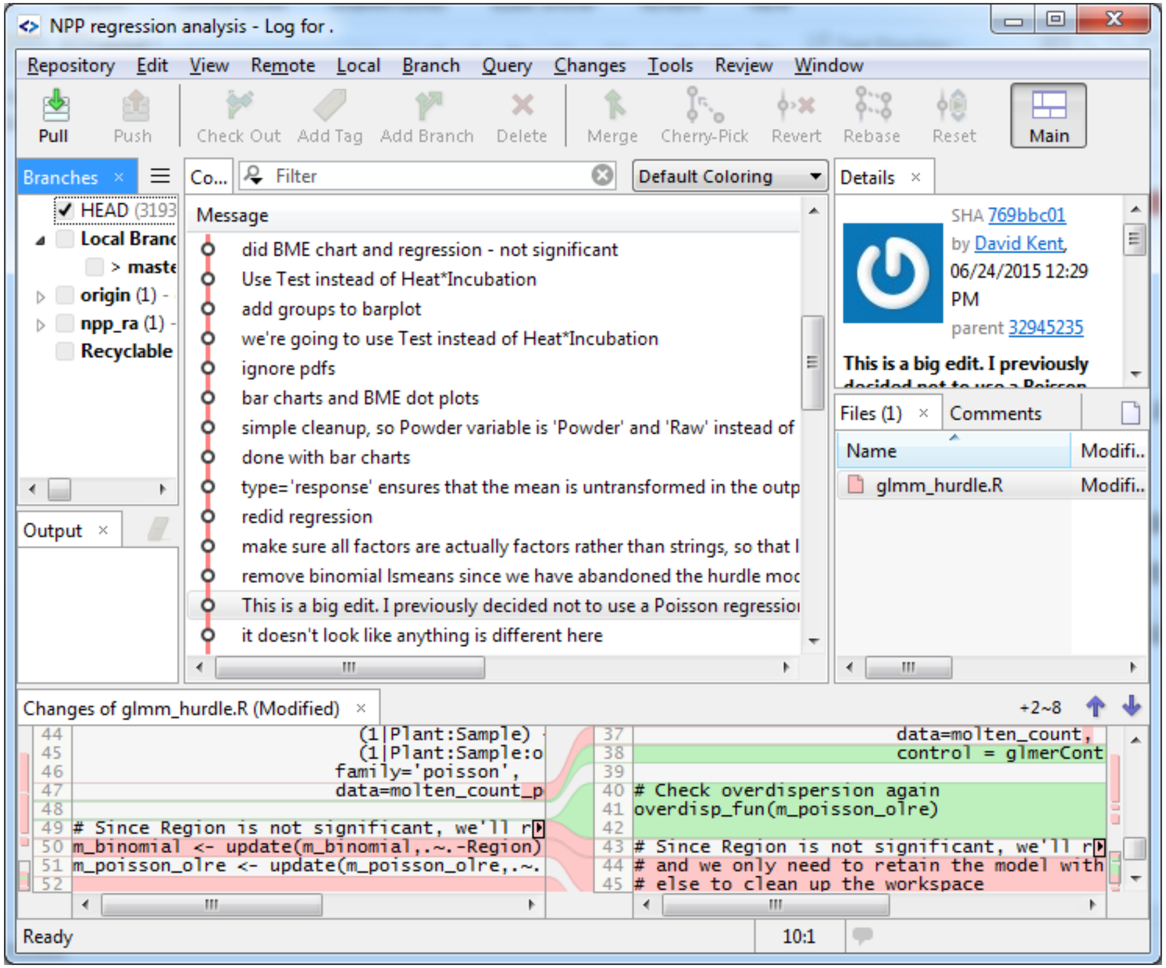
- Retain exactly ONE copy of your raw data
- Write an R script which does all of the corrections, normalization, and transformation
- The data itself is never changed, and corrections are made to the script



High-tech

- Just like medium-tech, but keep the data and script in “version control”
- This tracks *all* changes that have been made to the files, along with (mandatory) comments
- Most useful for more complicated analysis, multiple scripts

High-tech



Message

- did BME chart and regression - not significant
- Use Test instead of Heat*Incubation
- add groups to barplot
- we're going to use Test instead of Heat*Incubation
- ignore pdfs
- bar charts and BME dot plots
- simple cleanup, so Powder variable is 'Powder' and 'Raw' instead of
- done with bar charts
- type='response' ensures that the mean is untransformed in the outp
- redid regression
- make sure all factors are actually factors rather than strings, so that I
- remove binomial lsmeans since we have abandoned the hurdle moc
- This is a big edit. I previously decided not to use a Poisson regressio
- it doesn't look like anything is different here

SHA 769bbc01
by David Kent,
06/24/2015 12:29 PM
parent 32945235

This is a big edit. I previously decided not to use a Poisson

Name	Modifi..
glmm_hurdle.R	Modifi..

Changes of glmm_hurdle.R (Modified)

+2~8

```
                    (1|Plant:Sample)
                    (1|Plant:Sample:o
                  family='poisson',
                  data=molten_count_p

# Since Region is not significant, we'll r
m_binomial <- update(m_binomial,.~.-Region)
m_poisson_olre <- update(m_poisson_olre,.~.
```

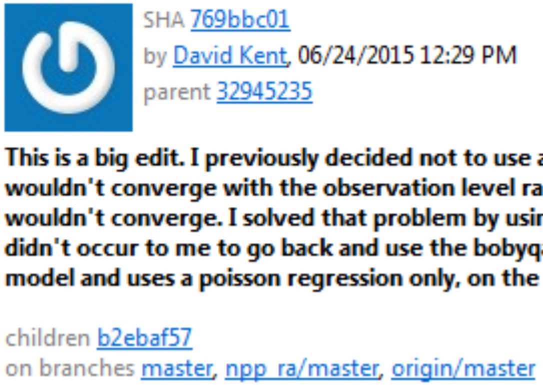
```
                          data=molten_count,
                          control = glmerCont

# Check overdispersion again
overdisp_fun(m_poisson_olre)

# Since Region is not significant, we'll r
# and we only need to retain the model with
# else to clean up the workspace
```

High-tech

- Real life example
- When I look back, I don’t have to wonder why I made certain changes, large *or* small.



SHA 769bbc01
by David Kent, 06/24/2015 12:29 PM
parent 32945235

This is a big edit. I previously decided not to use a Poisson regression on the raw counts (including zeroes) because the model wouldn't converge with the observation level random effects. I moved to a hurdle model, and then the binomial regression wouldn't converge. I solved that problem by using the 'bobyqa' optimizer based on a suggestion by the lme4 author, but it didn't occur to me to go back and use the bobyqa optimizer with the poisson regression. This revision abandons the hurdle model and uses a poisson regression only, on the data including zero counts.

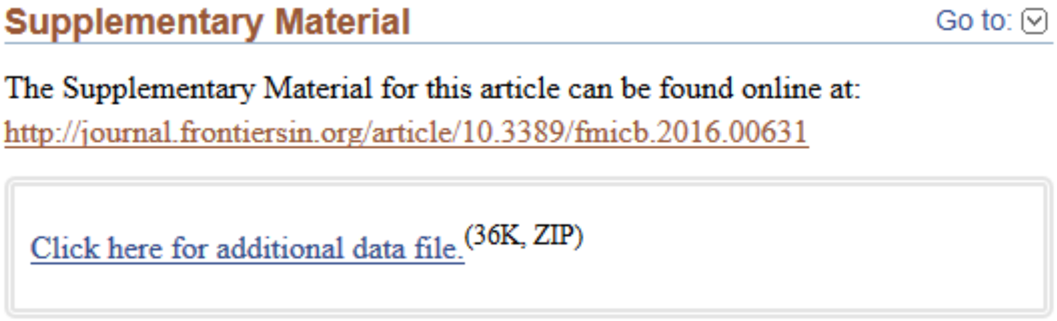
children b2ebaf57
on branches master, npp_ra/master, origin/master

Publishing the process

- Medium- and high-tech solutions yield an explicit description of the analysis process
- This can be published along with the text to remove any ambiguity about the analysis

Publishing the process

- Tom recently published a paper with the analysis script in the supplemental
- The analysis is *fully* reproducible



Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00631

Click here for additional data file. (36K, ZIP)

On another note

- Statistics only provides the right answer when we ask the right question
- We need to make explicit three things
  - What can we take for granted?
  - What information did the experiment collect?
  - What do we want to quantify?
- Sometimes I only know the answer in my gut, rather than my brain
- If we get counterintuitive results from the statistics, it’s important to go back and make sure we’re asking the right question.

Questions?