Organizing a Data Analysis

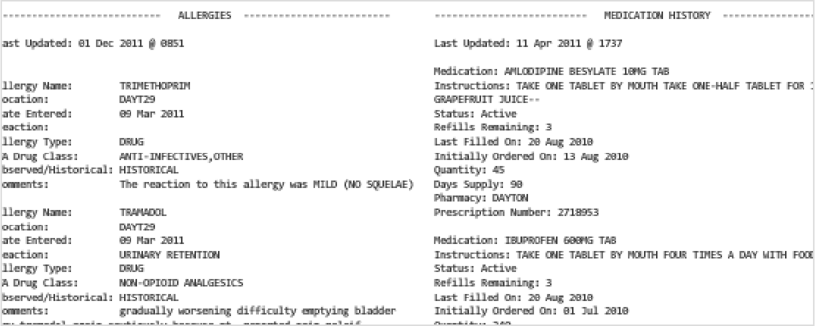
Roger D. Peng, Associate Professor of Biostatistics

May 18, 2016

Data analysis files

- Data
- Raw data
- Processed data
- Figures
- Exploratory figures
- Final figures
- R code
- Raw / unused scripts
- Final scripts
- R Markdown files
- Text
- README files
- Text of analysis / report

Raw Data

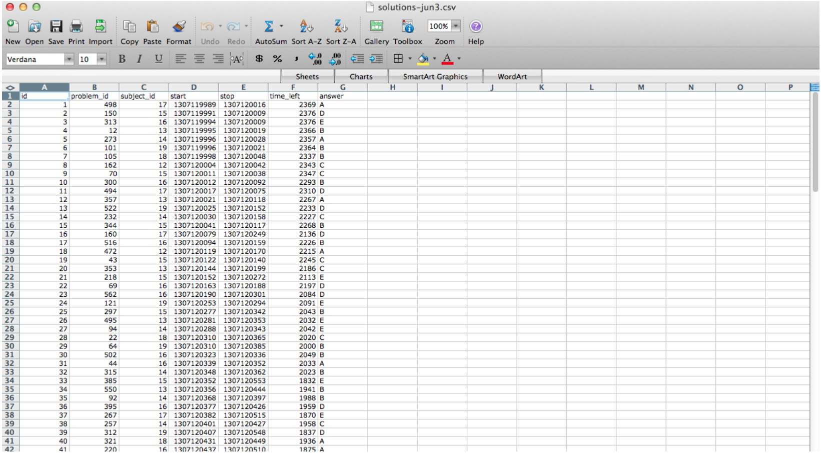


```
---------------------------   ALLERGIES   -------------------------

ast Updated: 01 Dec 2011 @ 0851


llergy Name:         TRIMETHOPRIM
ocation:             DAYT29
ate Entered:         09 Mar 2011
eaction:
llergy Type:         DRUG
A Drug Class:        ANTI-INFECTIVES,OTHER
bserved/Historical:  HISTORICAL
omments:             The reaction to this allergy was MILD (NO SQUELAE)

llergy Name:         TRAMADOL
ocation:             DAYT29
ate Entered:         09 Mar 2011
eaction:             URINARY RETENTION
llergy Type:         DRUG
A Drug Class:        NON-OPIOID ANALGESICS
bserved/Historical:  HISTORICAL
omments:             gradually worsening difficulty emptying bladder
ny tramadol again particularly because pt. reported pain relief
```

```
--------------------------   MEDICATION HISTORY   ---------------

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPINE BESYLATE 10MG TAB
Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR 
GRAPEFRUIT JUICE--
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 13 Aug 2010
Quantity: 45
Days Supply: 90
Pharmacy: DAYTON
Prescription Number: 2718953

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010
Quantity: 240
```

- Should be stored in your analysis folder
- If accessed from the web, include url, description, and date accessed in README

Processed data

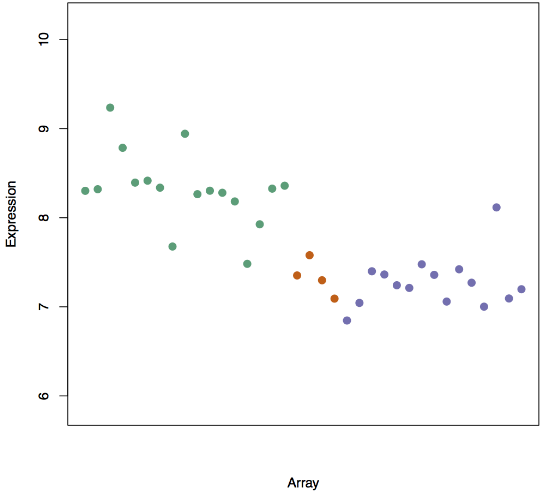


id	problem_id	subject_id	start	stop	time_left	answer
1	498	17	1307119989	1307120016	2369	A
2	150	15	1307119991	1307120009	2376	D
3	313	16	1307119994	1307120009	2376	E
4	12	13	1307119995	1307120019	2366	B
5	273	14	1307119996	1307120028	2357	A
6	101	19	1307119996	1307120021	2364	B
7	105	18	1307119998	1307120048	2337	B
8	162	12	1307120004	1307120042	2343	C
9	70	15	1307120011	1307120038	2347	C
10	300	16	1307120012	1307120092	2293	B
11	494	17	1307120017	1307120075	2310	D
12	357	13	1307120021	1307120118	2267	A
13	522	19	1307120025	1307120152	2233	D
14	232	14	1307120030	1307120158	2227	C
15	344	15	1307120041	1307120117	2268	B
16	160	17	1307120079	1307120249	2136	D
17	516	16	1307120094	1307120159	2226	B
18	472	12	1307120119	1307120170	2215	A
19	43	15	1307120122	1307120140	2245	C
20	353	13	1307120144	1307120199	2186	C
21	218	15	1307120152	1307120272	2113	E
22	69	16	1307120163	1307120188	2197	D
23	562	16	1307120190	1307120301	2084	D
24	121	19	1307120253	1307120294	2091	E
25	297	15	1307120277	1307120342	2043	B
26	495	13	1307120281	1307120353	2032	E
27	94	14	1307120288	1307120343	2042	E
28	22	18	1307120310	1307120365	2020	C
29	64	19	1307120310	1307120385	2000	B
30	502	16	1307120323	1307120336	2049	B
31	44	16	1307120339	1307120352	2033	A
32	315	14	1307120348	1307120362	2023	B
33	385	15	1307120352	1307120553	1832	E
34	550	13	1307120356	1307120444	1941	B
35	92	14	1307120368	1307120397	1988	B
36	395	16	1307120377	1307120426	1959	D
37	267	17	1307120382	1307120515	1870	E
38	257	14	1307120401	1307120427	1958	C
39	312	19	1307120407	1307120548	1837	D
40	321	18	1307120431	1307120449	1936	A
41	220	16	1307120437	1307120510	1875	A

*

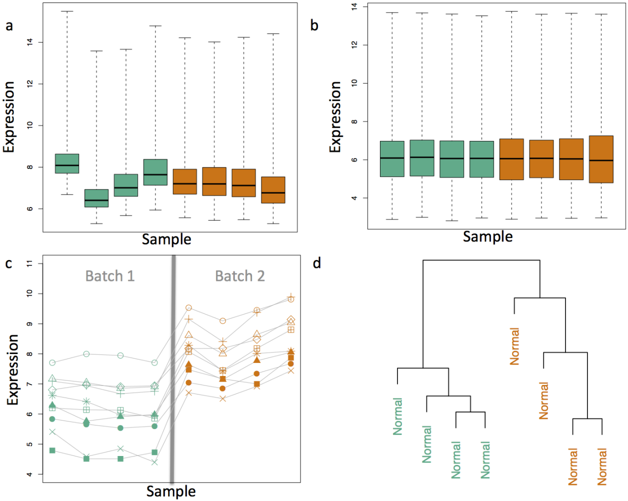
Processed data should be named so it is easy to see which script generated the data. * The processing script - processed data mapping should occur in the README * Processed data should be tidy

Exploratory figures



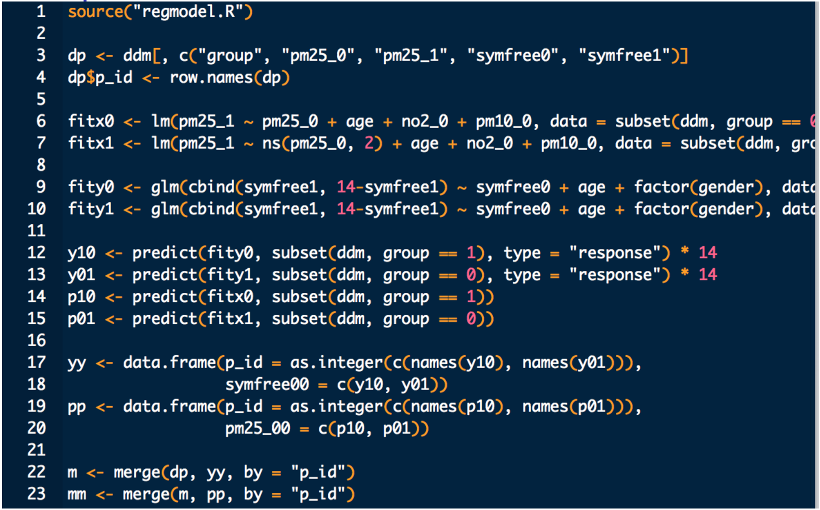
- Figures made during the course of your analysis, not necessarily part of your final report.

Final Figures



- Usually a small subset of the original figures
- Axes/colors set to make the figure clear

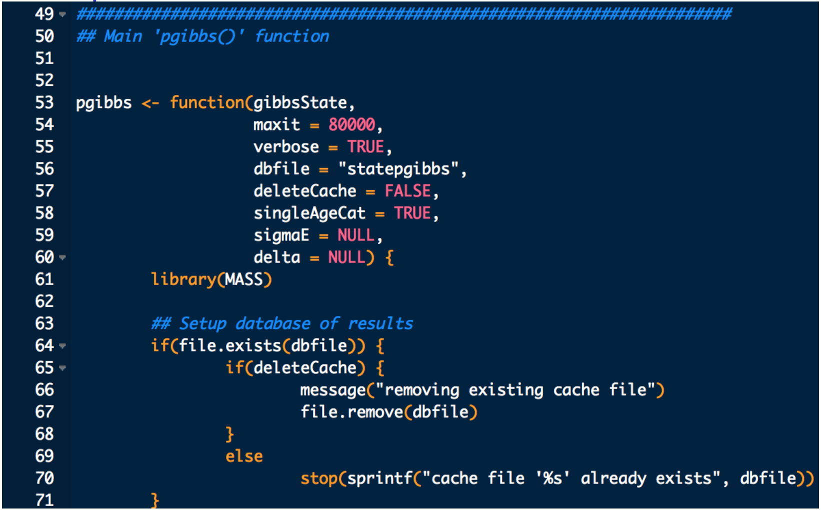
Raw scripts



```r
source("regmodel.R")

dp <- ddm[, c("group", "pm25_0", "pm25_1", "symfree0", "symfree1")]
dp$p_id <- row.names(dp)

fitx0 <- lm(pm25_1 ~ pm25_0 + age + no2_0 + pm10_0, data = subset(ddm, group == 0
fitx1 <- lm(pm25_1 ~ ns(pm25_0, 2) + age + no2_0 + pm10_0, data = subset(ddm, gro

fity0 <- glm(cbind(symfree1, 14-symfree1) ~ symfree0 + age + factor(gender), data
fity1 <- glm(cbind(symfree1, 14-symfree1) ~ symfree0 + age + factor(gender), data

y10 <- predict(fity0, subset(ddm, group == 1), type = "response") * 14
y01 <- predict(fity1, subset(ddm, group == 0), type = "response") * 14
p10 <- predict(fitx0, subset(ddm, group == 1))
p01 <- predict(fitx1, subset(ddm, group == 0))

yy <- data.frame(p_id = as.integer(c(names(y10), names(y01))),
                 symfree00 = c(y10, y01))
pp <- data.frame(p_id = as.integer(c(names(p10), names(p01))),
                 pm25_00 = c(p10, p01))

m <- merge(dp, yy, by = "p_id")
mm <- merge(m, pp, by = "p_id")
```

- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded

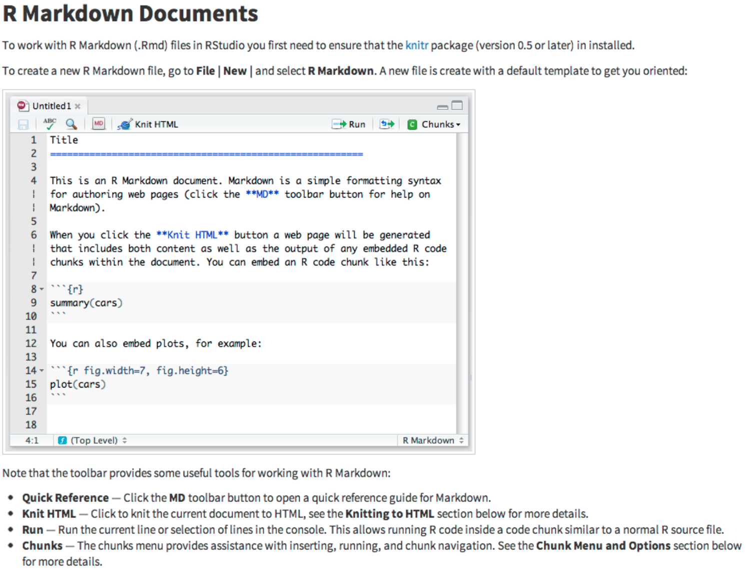
Final scripts



```r
##################################################################
## Main 'pgibbs()' function


pgibbs <- function(gibbsState,
                   maxit = 80000,
                   verbose = TRUE,
                   dbfile = "statepgibbs",
                   deleteCache = FALSE,
                   singleAgeCat = TRUE,
                   sigmaE = NULL,
                   delta = NULL) {
        library(MASS)

        ## Setup database of results
        if(file.exists(dbfile)) {
                if(deleteCache) {
                        message("removing existing cache file")
                        file.remove(dbfile)
                }
                else
                        stop(sprintf("cache file '%s' already exists", dbfile))
        }
```

- Clearly commented
- Small comments liberally - what, when, why, how
- Bigger commented blocks for whole sections

R markdown files



R Markdown Documents

To work with R Markdown (.Rmd) files in RStudio you first need to ensure that the knitr package (version 0.5 or later) in installed.

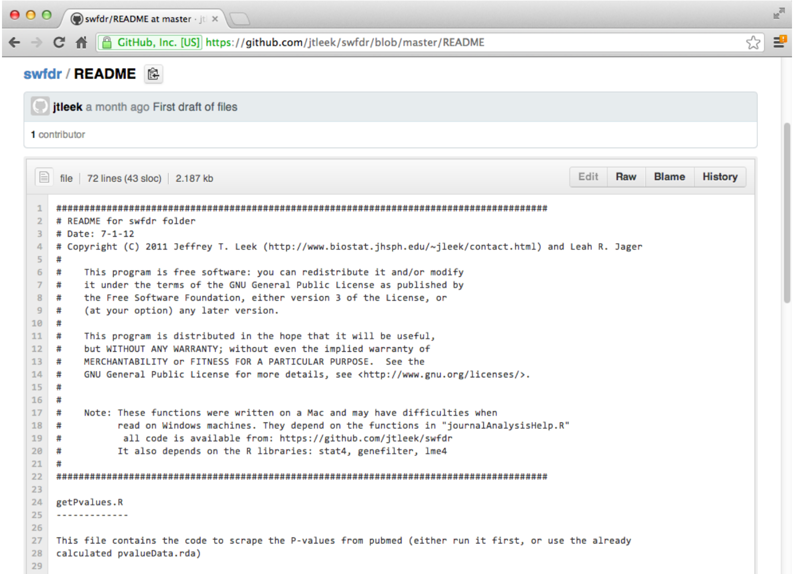
To create a new R Markdown file, go to **File | New |** and select **R Markdown**. A new file is create with a default template to get you oriented:

~~~
Title
========================================================


This is an R Markdown document. Markdown is a simple formatting syntax
for authoring web pages (click the **MD** toolbar button for help on
Markdown).

When you click the **Knit HTML** button a web page will be generated
that includes both content as well as the output of any embedded R code
chunks within the document. You can embed an R code chunk like this:

```{r}
summary(cars)
~~~

You can also embed plots, for example:

```{r fig.width=7, fig.height=6}
plot(cars)
```
```

Note that the toolbar provides some useful tools for working with R Markdown:

- **Quick Reference** — Click the **MD** toolbar button to open a quick reference guide for Markdown.
- **Knit HTML** — Click to knit the current document to HTML, see the **Knitting to HTML** section below for more details.
- **Run** — Run the current line or selection of lines in the console. This allows running R code inside a code chunk similar to a normal R source file.
- **Chunks** — The chunks menu provides assistance with inserting, running, and chunk navigation. See the **Chunk Menu and Options** section below for more details.

- R markdown files can be used to generate reproducible reports
- Text and R code are integrated
```

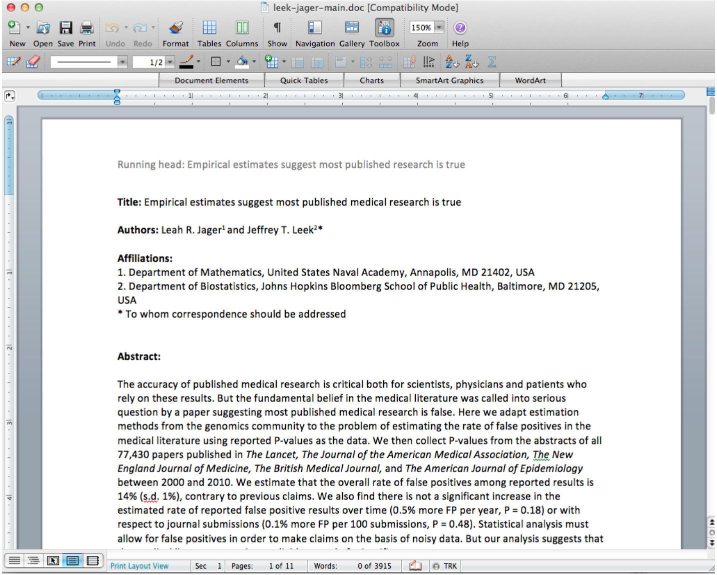
# Readme files


swfdr/README at master · jt
GitHub, Inc. [US] https://github.com/jtleek/swfdr/blob/master/README
swfdr / README
jtleek a month ago First draft of files
1 contributor
file | 72 lines (43 sloc) | 2.187 kb
Edit
Raw
Blame
History

```
##
README for swfdr folder
Date: 7-1-12
Copyright (C) 2011 Jeffrey T. Leek (http://www.biostat.jhsph.edu/~jleek/contact.html) and Leah R. Jager
#
This program is free software: you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation, either version 3 of the License, or
(at your option) any later version.
#
This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
GNU General Public License for more details, see <http://www.gnu.org/licenses/>.
#
#
Note: These functions were written on a Mac and may have difficulties when
read on Windows machines. They depend on the functions in "journalAnalysisHelp.R"
all code is available from: https://github.com/jtleek/swfdr
It also depends on the R libraries: stat4, genefilter, lme4
#
##

getPvalues.R

This file contains the code to scrape the P-values from pubmed (either run it first, or use the already
calculated pvalueData.rda)

```

- Not necessary if you use R markdown
- Should contain step-by-step instructions for analysis

# Text of the document



Running head: Empirical estimates suggest most published research is true

**Title:** Empirical estimates suggest most published medical research is true

**Authors:** Leah R. Jager[1] and Jeffrey T. Leek[2]*

**Affiliations:**
1. Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA
2. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
* To whom correspondence should be addressed

**Abstract:**

The accuracy of published medical research is critical both for scientists, physicians and patients who rely on these results. But the fundamental belief in the medical literature was called into serious question by a paper suggesting most published medical research is false. Here we adapt estimation methods from the genomics community to the problem of estimating the rate of false positives in the medical literature using reported P-values as the data. We then collect P-values from the abstracts of all 77,430 papers published in *The Lancet, The Journal of the American Medical Association, The New England Journal of Medicine, The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010. We estimate that the overall rate of false positives among reported results is 14% (s.d. 1%), contrary to previous claims. We also find there is not a significant increase in the estimated rate of reported false positive results over time (0.5% more FP per year, P = 0.18) or with respect to journal submissions (0.1% more FP per 100 submissions, P = 0.48). Statistical analysis must allow for false positives in order to make claims on the basis of noisy data. But our analysis suggests that

- It should include a title, introduction (motivation), methods (statistics you used), results (including measures of

# Further resources

- Information about a non-reproducible study that led to cancer patients being mistreated: The Duke Saga Starter Set
- Reproducible research and Biostatistics
- Managing a statistical analysis project guidelines and best practices
- Project template - a pre-organized set of files for data analysis