Introduction to R Markdown

Ikuma Ogura (Georgetown)

Last Update: August 11, 2018

R Markdown

R Markdown is a tool to make documents including R scripts and their associated outputs. This document is meant to introduce you to R Markdown and to be used as a template for future use.

Markdown

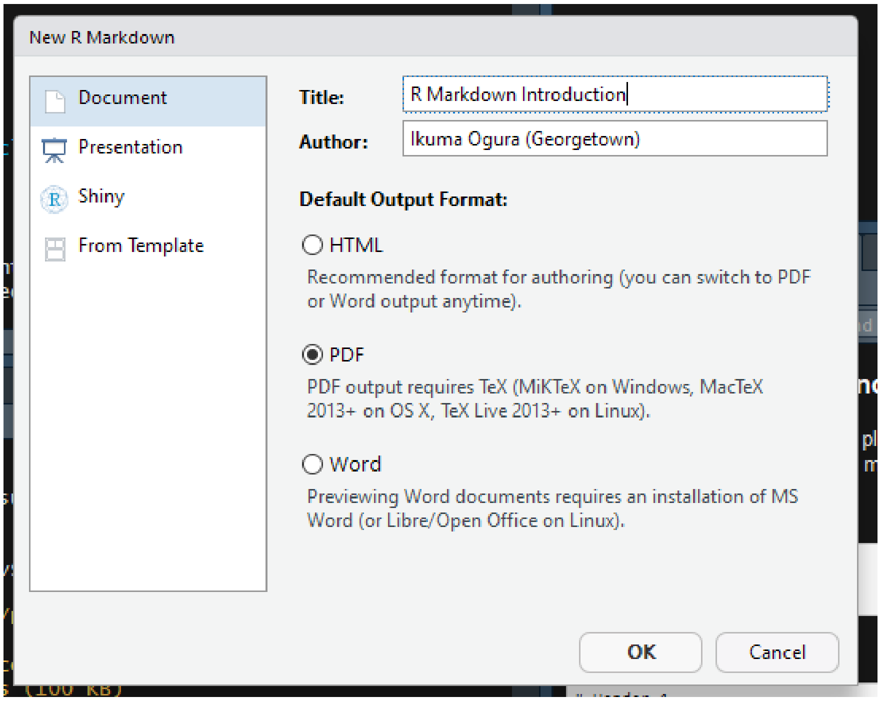
In R Markdown, we write the documents folowing the Markdown syntax rules. Markdown is a type of lightweight markup language to create documents. To see how the Markdown syntax looks like, click [Help] → [Markdown Quick Reference] in RStudio. I also recommend you download official R Markdown Cheet Sheeet from [Help] → [Cheetsheets] → [R Markdown Cheeet Sheet].

Benefits of Using R Markdown

1. You can make the R scripts and outputs on your document consistent. (If you copy & paste the codes and outputs on Microsoft Word, you often forget updating one of them…)
2. You can also include nicely formatted mathematical expressions. (Best for writing your problem set answers for stats classes!)
3. Markdown is very easy to use, so it's a good way to get used to syntax of other markup languages such as includes HTML and LaTeX (which I'm sure you want to master in the future!).
    - In fact, we can use HTML and LaTeX syntax within R Markdown to better format the document.

Workflow

1. Start your R Markdown project.
    - To create a new R Markdown document, click [File] → [New File] → [R Markdown…]
2. Type in the document title and your name, and select the document type you want to create (see Figure 1).
    - NB: To create a PDF file, you need to have LaTeX in your environment.
    - To install LaTeX, go to LaTeX website and follow the instructions. **It takes some time to install LaTeX.**
3. Write your texts and R scripts and save .Rmd file.
    - R scripts shoud be written between `` ```{r} `` and `` ``` ``.
    - You can add options within the curly brackets like `{r, message = FALSE, tidy = FALSE}` to control how the R scripts are displayed.
    - If you want to run the R codes you've written while creating your document, click the green triangle buttom on top of the code chunk.



Title: R Markdown Introduction

Author: Ikuma Ogura (Georgetown)

Default Output Format:

○ HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

● PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

Figure 1: Example R Markdown Wizard

- To include short mathetical expressions into the text, write them within two dollar signs (`$`), and to insert an independent line of mathetical expressions, write them between `\[` and `\]`. You need to follow the LaTeX syntax rule to write mathematical expressions. To write maths of multiple lines, use `aligned` environment within `\[` and `\]`.
- Helpful references on LaTeX math expressions.
  - https://www.codecogs.com/eqnedit.php
  - http://quicklatex.com/
  - https://en.wikibooks.org/wiki/LaTeX/Mathematics

4. Convert (or Knit) .Rmd file.
   - You just need to click the **Knit** button on top of the script window, and RStudio will automatically execute the R scripts and generate the document.

Example

As there are many good freely-available descriptions on (R) Markdown syntax online, I don't explain detailed rules of Markdown grammers here. Instead, I show you an example of brief R Markdown project below. Please refer to the `rmarkdown_intro.Rmd` and learn how each Markdown syntax translates to document features. You can use this .Rmd file as the template for your future project.

1. Purpose

Here I use data on publication activities of recent biochemistry PhDs, originally used in Long (1990),[1] and see (i) which of Pisson or Negative binomial regression models better fit this dataset and (ii) whether gender influences research activities beyond graduate schools.

2. Model

In a Poisson regression model, the dependent variable y_i is assumed to be generated as follows;

$$\begin{aligned} \mu_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \\ \Pr(y_i|\boldsymbol{x}_i) &= \mathrm{Pois}(\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \end{aligned}$$

where $\boldsymbol{x}_i$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of associated coefficients. In contrast, in a Negative binomial regression model, we assume that the y_i is generated as

$$\begin{aligned} \mu_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \\ \Pr(y_i|\boldsymbol{x}_i) &= \mathrm{Negbin}(\mu_i, \alpha) = \frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)}\Big(\frac{\alpha}{\alpha+\mu_i}\Big)^{\alpha}\Big(\frac{\mu_i}{\alpha+\mu_i}\Big)^{y_i}. \end{aligned}$$

Here, a new parameter α is introduced, which represents the degree of over- or underdispersion. This is because the mean and variance of y_i is represend as

$$\begin{aligned} \mathrm{E}(y_i) &= \mu_i \\ \mathrm{V}(y_i) &= \mu_i\Big(1+\frac{\mu_i}{\alpha}\Big) \end{aligned}$$

You might feel that the pmf of the negative binomial distribution above is different from the familiar form, but in fact they are the same. Let y be the total number of failures until α successes, and p be the success probability. If we define that $p = \alpha/(\mu+\alpha)$,

$$\begin{aligned} &{}_{y+\alpha-1}\mathrm{C}_y p^{\alpha}(1-p)^y \\ =&\frac{(y+\alpha-1)!}{y!(\alpha-1)!}p^{\alpha}(1-p)^y \\ =&\frac{\Gamma(y+\alpha)}{y_i!\Gamma(\alpha)}\Big(\frac{\alpha}{\alpha+\mu}\Big)^{\alpha}\Big(\frac{\mu}{\alpha+\mu}\Big)^{y}, \end{aligned}$$

where ${}_{y+\alpha-1}\mathrm{C}_y = \binom{y+\alpha-1}{y}$.

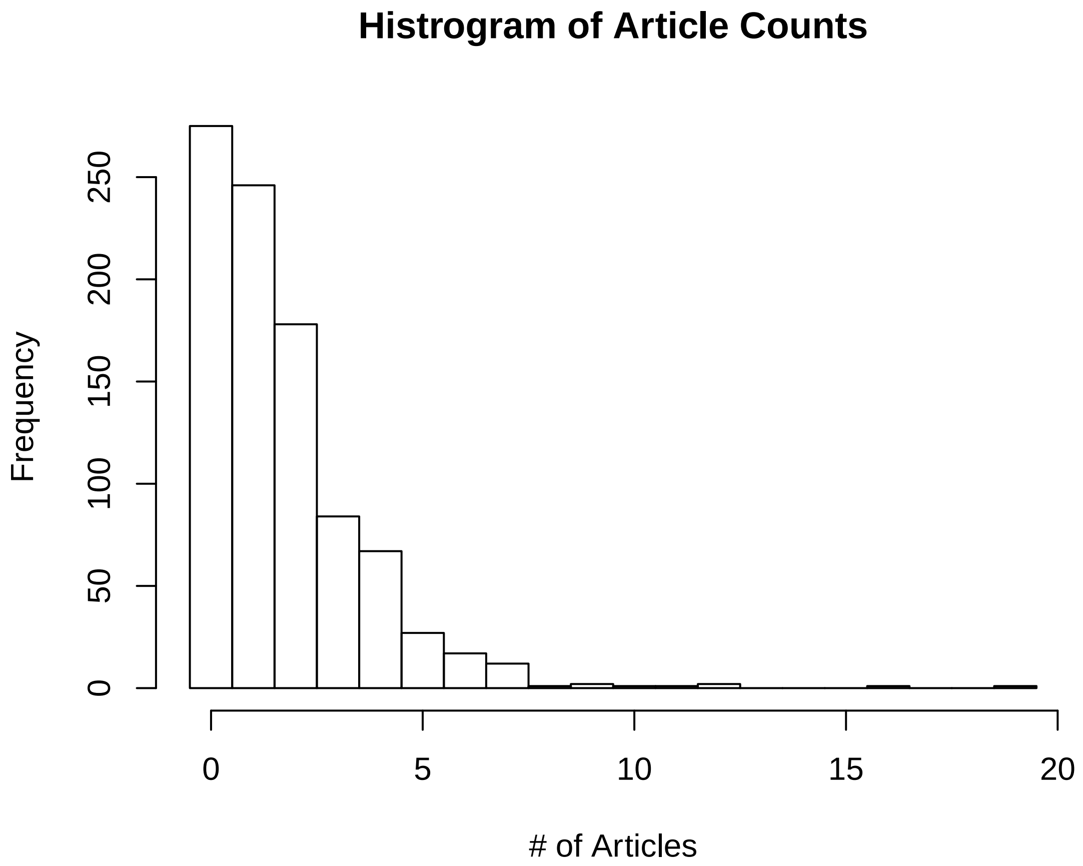
3. Preparation

```r
# Load packages
require(MASS)
require(pscl)
require(stargazer)

# setwd()
```

[1] Long, Scott. 1990. "The origins of sex differences in science." *Social Forces.* 68(3): 1297-1316.

```
# Load data
data(bioChemists)

# Create necessary variables
bioChemists$women <- ifelse(bioChemists$fem == "Women", 1, 0)
bioChemists$married <- ifelse(bioChemists$mar == "Married", 1, 0)
```

4. Data summarization/Visualization



The above figure is the histogram of our outcome of interest (`art`), the number of published articles of recent biochemistry PhDs. The histogram shows that (i) there are many 0 counts and (ii) the distribution of `art` is long-tailed, suggesting that the variance of the variable is larger than its mean. Computing the two summary statistics,

```
mean(bioChemists$art) # mean
```

```
## [1] 1.69
```

```
var(bioChemists$art) # variance
```

```
## [1] 3.71
```

Thus we can expect that the Negative binomial regression model fits the data better than the Poisson model.

5. Estimating regression models

```
out.pois <- glm(art ~ women + married + kid5 + phd + ment, data = bioChemists,
                family = poisson, x = TRUE)
out.negbin <- glm.nb(art ~ women + married + kid5 + phd + ment,
                     data = bioChemists, x = TRUE)
stargazer(list(out.pois, out.negbin), type = "text",
          covariate.labels = c("Women", "Married", "# Kids below 5",
                               "PhD Prestige", "Mentor Article Counts",
                               "Intercept"),
          omit.stat = c("ll"))
```

```
## 
## ==================================================
##                           Dependent variable:     
##                       ----------------------------
##                                   art             
##                         Poisson       negative    
##                                       binomial    
##                           (1)           (2)       
## --------------------------------------------------
## Women                  -0.225***     -0.216***    
##                         (0.055)       (0.073)     
##                                                   
## Married                 0.155**       0.150*      
##                         (0.061)       (0.082)     
##                                                   
## # Kids below 5         -0.185***     -0.176***    
##                         (0.040)       (0.053)     
##                                                   
## PhD Prestige             0.013         0.015      
##                         (0.026)       (0.036)     
##                                                   
## Mentor Article Counts  0.026***      0.029***     
##                         (0.002)       (0.003)     
##                                                   
## Intercept              0.305***       0.256*      
##                         (0.103)       (0.137)     
##                                                   
## --------------------------------------------------
## Observations              915           915       
## theta                             2.260*** (0.271)
## Akaike Inf. Crit.      3,314.000     3,136.000    
## ==================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

The above table summarizes the regression estimates of article counts by biochemistry PhDs on various explanatory factors. The left column lists the estimation results of the Poisson regression model and the right column summarizes the results of Negative binomial model.

Looking at the table, the theta parameter, which stand for the α introduced in section 2, is statistically distinguishable from 0. This means that the Negative binomial model is better for this dataset. Looking at the coefficient estimate on women dummy of the Negative binomial model, it is negative and statistically significant, indicating that women PhDs tend to publish smaller numbers of articles than male counterparts.

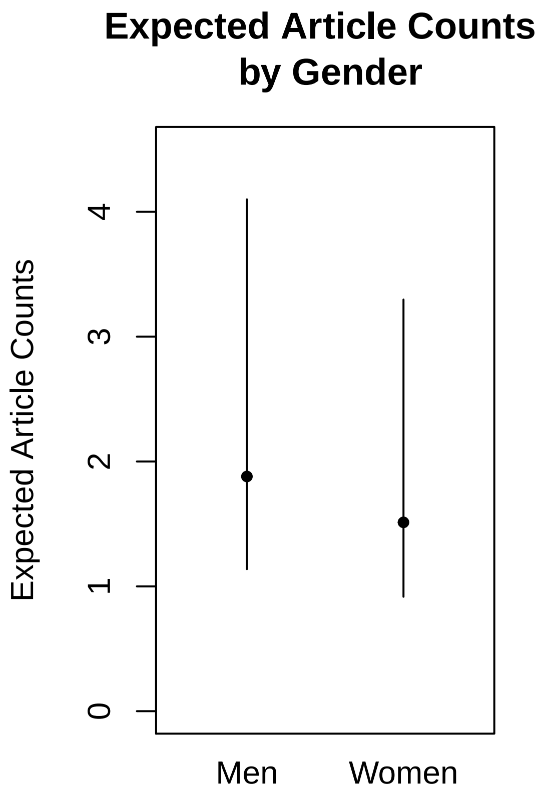
6. How Large is the Effect?

To see the substantive effect size of the `women` variable on article counts, I calculate the expected article counts of men and women using the estimation results of the Negative binomial model. I simulate the estimation uncertainty using the method proposed by King, Tomz, and Wittenberg (2000).[2] In computing the expected counts, I marginal out the independent variables other than `women` using the "observed value" approach, which is recommended by Hanmer and Ozan Kalkan (2013).[3]

```r
# Prepare
x <- out.negbin$x
x.men <- x
x.men[, 2] <- 0 # design matrix for men
x.women <- x
x.women[, 2] <- 1 # design matrix for women
# Clarify for expected article counts
set.seed(1234) # seed for random number generator
sim.coef <- mvrnorm(1000, mu = coef(out.negbin), Sigma = vcov(out.negbin))
sim.mu.men <- exp(x.men %*% t(sim.coef))
sim.mu.women <- exp(x.women %*% t(sim.coef))
count.men.mean <- mean(apply(sim.mu.men, 1, mean))
count.men.ci <- quantile(apply(sim.mu.men, 1, mean), probs = c(0.025, 0.975))
count.women.mean <- mean(apply(sim.mu.women, 1, mean))
count.women.ci <- quantile(apply(sim.mu.women, 1, mean), probs = c(0.025, 0.975))
first.diff.mean <- mean(apply(sim.mu.men, 1, mean) - apply(sim.mu.women, 1, mean))
first.diff.ci <- quantile(apply(sim.mu.men, 1, mean) - apply(sim.mu.women, 1, mean),
                          probs = c(0.025, 0.975))
# Visualize
plot(c(1, 2), c(count.men.mean, count.women.mean), pch = 20,
     xlim = c(0.5, 2.5), ylim = c(0, 4.5),
     xlab = "", ylab = "Expected Article Counts",
     main = "Expected Article Counts \n by Gender",
     xaxt = "n")
mtext(c("Men", "Women"), side = 1, line = 0.5, at = c(1, 2))
segments(1, count.men.ci[1], 1, count.men.ci[2])
segments(2, count.women.ci[1], 2, count.women.ci[2])
```

[2] King, Gary, Michael Toomz, and Jason Wittenberg. 2000. "Making Most of the Statistical Analysis: Improving Interpretation and Presentation." *American Journal of Political Science.* 44(2): 341-355.

[3] Hanmer, Michael J. and Karem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science.* 57(1): 263-277.



The results of the analysis is visually shown in the figure above. I also summarize the outcome in the table below.

	Expected Counts (Men)	Expected Counts (Women)	First Difference
Mean	1.88	1.513	0.367
95% CI	[1.138, 4.1]	[0.917, 3.297]	[0.221, 0.802]

The figure and table show that, while the differences in expected counts between men and women PhDs are statistically distinguishable from 0, it is not substantially large. Considering that the standard deviation of the dependent variable is about 1.93 ($\approx \sqrt{3.71}$), we can conclude that the calculated first difference is (although not small) not substantively large.