Streamline Your Workflow: Integrating SAS, LaTeX, and R into a Single Reproducible Document *A 538 Star Wars Example*

Lucy D'Agostino McGowan

February 28, 2017

1 Introduction

I obtained data from a 538 Star Wars Survey (found here: `https://github.com/fivethirtyeight/data/tree/master/star-wars-survey`) and will read it into SAS in order to analyze whether age, gender, or education level are associated with incorrectly believing that Greedo shot first.

```
libname data "/folders/myshortcuts/report-example/data";

filename reffile
  '/folders/myshortcuts/report-example/data/star-wars-survey-538.csv';

proc import datafile=reffile
dbms=csv
replace
out=data.starwars;
getnames=yes;
run;

data data;
 set data.starwars;
  if Jar_Jar_Binks in (" ", "Unfamiliar (N/A)")
    then wrong_jar_jar = " ";
  else if Jar_Jar_Binks in ("Very favorably","Somewhat favorably")
    then wrong_jar_jar = 1;
    else wrong_jar_jar = 0;
  if shot_first = "Han"
    then wrong_han = 0;
  if shot_first = "Greedo"
    then wrong_han = 1;
```

```
     if education in ("Bachelor degree", "Graduate degree")
       then college = "College degree";
     if education in ("Some college or Associate degree",
       "High school degree", "Less than high school degree")
       then college = "No college degree";
   run;


   ods graphics on;

   proc freq data = data;
   table wrong_han wrong_jar_jar;
   run;

   ods graphics off;
```

Figure 1: Wrong about who shot first

```
                       The FREQ Procedure

                                    Cumulative    Cumulative
   wrong_han    Frequency    Percent    Frequency      Percent
   -----------------------------------------------------------
          0          325      62.26          325        62.26
          1          197      37.74          522       100.00

                  Frequency Missing = 664
```

Figure 2: Wrong aboud Jar Jar Binks

```
   wrong_
   jar_                                Cumulative    Cumulative
   jar        Frequency    Percent    Frequency      Percent
   -----------------------------------------------------------
   0               470      66.01          470        66.01
   1               242      33.99          712       100.00

                  Frequency Missing = 474
```

```
   ods graphics on;

   proc logistic data = data plots = oddsratio;
    class age (ref = FIRST) gender college Star_Trek_fan;
    model wrong_han(event = "1") = age gender college Star_Trek_fan;
   run;
```

```
ods graphics off;

ods graphics on;

proc logistic data = data plots = oddsratio;
 class age (ref = FIRST) gender college Star_Trek_fan;
 model wrong_jar_jar(event = "1") = age gender college Star_Trek_fan;
run;

ods graphics off;
```

Figure 3: Wrong about who shot first Odds Ratio

```
                    The LOGISTIC Procedure


                    Odds Ratio Estimates


                                         Point        95% Wald
Effect                                  Estimate   Confidence Limits


Age          30-44 vs 18-29              1.566      0.941      2.605
Age          45-60 vs 18-29              1.232      0.730      2.079
Age          > 60  vs 18-29              2.190      1.218      3.936
Gender       Fema vs Male                1.402      0.970      2.026
college      College degree vs No college deg  1.027   0.705   1.497
Star_Trek_fan No vs Yes                  1.194      0.818      1.742
```
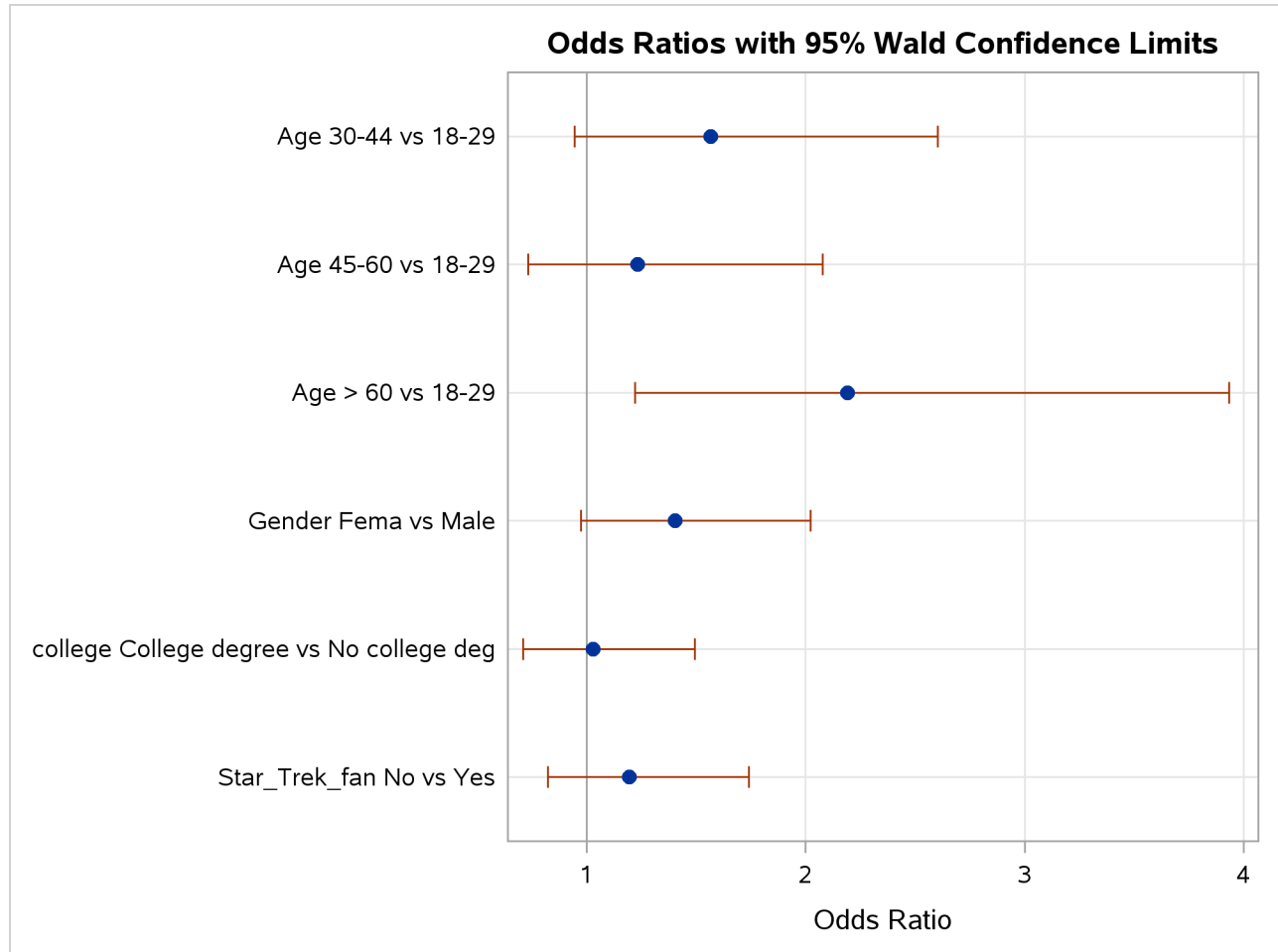
Figure 4: Wrong about Jar Jar Binks

Odds Ratios with 95% Wald Confidence Limits

Figure 5: Wrong about who shot first Odds Ratio

```
                        The LOGISTIC Procedure

                        Odds Ratio Estimates

                                         Point        95% Wald
Effect                                 Estimate    Confidence Limits

Age          30-44 vs 18-29              0.939     0.587      1.501
Age          45-60 vs 18-29              1.608     1.016      2.546
Age          > 60  vs 18-29              1.630     1.001      2.654
Gender       Fema vs Male                1.694     1.224      2.343
college      College degree vs No college deg  0.566     0.407      0.787
Star_Trek_fan No vs Yes                  1.095     0.785      1.528
```

Figure 6: Wrong about Jar Jar Binks

Odds Ratios with 95% Wald Confidence Limits
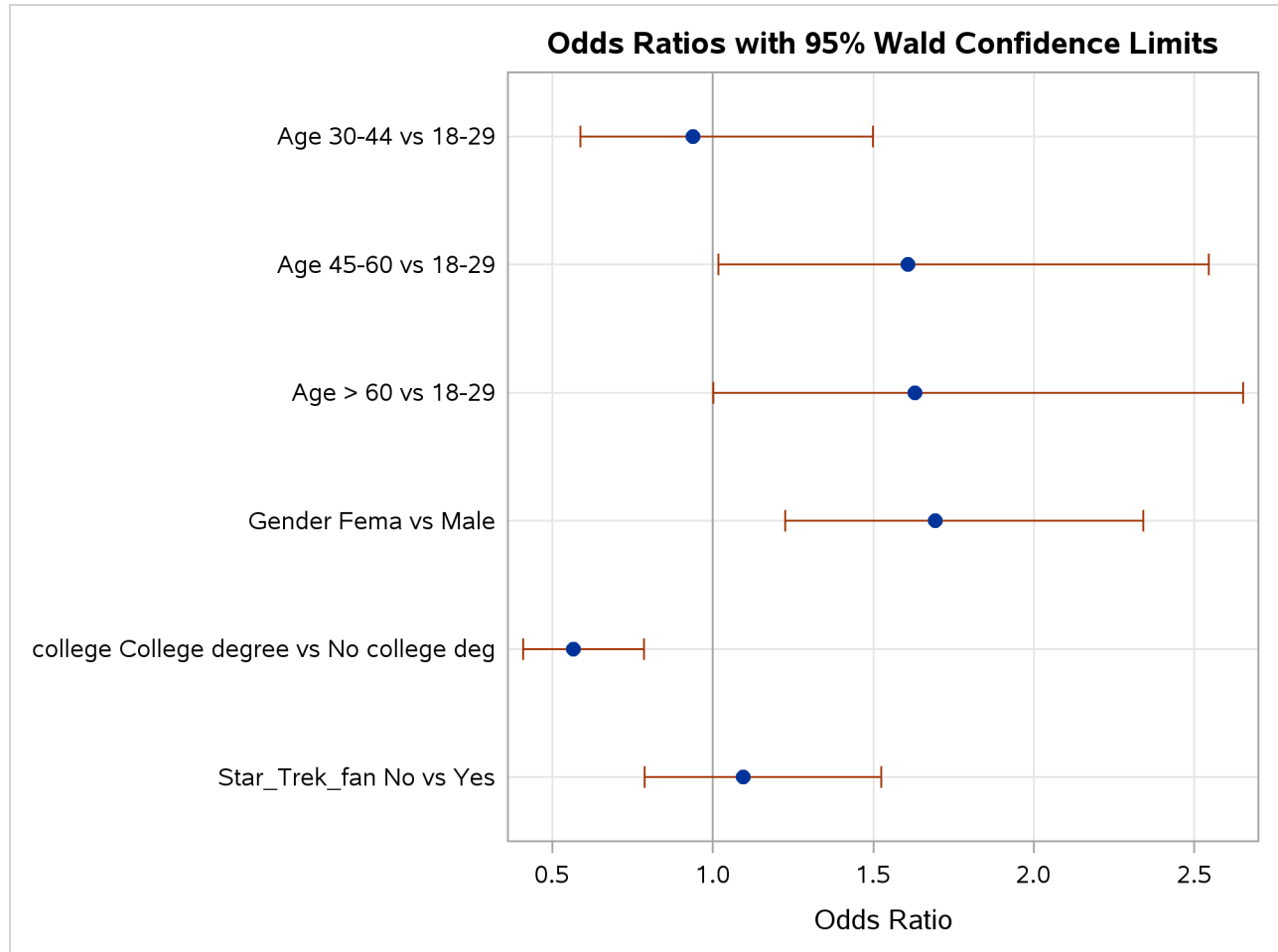
```
proc export data=data.starwars
   outfile =
     '/folders/myshortcuts/report-example/data/starwars_sasedit.csv'
   replace
   dbms = dlm;
   delimiter = ',';
run;
```

2 Test

```r
filename = "../data/starwars_sasedit.csv"
if (file.exists(filename)){
  #load libraries
  library('dplyr')
  library('rphylopic')
  library('png')
  library('ggplot2')

  #read in data
  starwars <- read.csv(filename)

  #load some cute pics
  chewie <- readPNG("../data/img/chewie.png")
  stormtrooper <- readPNG("../data/img/storm_trooper.png")


  starwars %>%
    filter(shot_first %in% c("Han","Greedo")) %>%
    select(shot_first,The_Phantom_Menace,
           Attack_of_the_Clones,Revenge_of_the_Sith,
           A_New_Hope,The_Empire_Strikes_Back,
           Return_of_the_Jedi) %>%
    group_by(shot_first) %>%
    summarise_each(funs(mean(., na.rm = TRUE)))  %>%
    tidyr::gather("film","rank",2:7) %>%
    mutate(film = gsub("_", " ", film)) %>%
    data.frame() -> plot_data
 plot_data$film <- reorder(plot_data$film, rep(1:6, each = 2))
 p <-  ggplot(plot_data,aes(film, rank, group = shot_first))  +
   geom_line(aes(color = shot_first)) +
   scale_colour_manual(values = c("black","brown")) +
   ylim(0,6)
   for (i in 1:6) {
   p <- p + add_phylopic(chewie, 1, i,
                         plot_data[plot_data$shot_first=="Han",
                                   "rank"][i],
                         ysize = 1)
  }
    for (i in 1:6) {
   p <- p + add_phylopic(stormtrooper, 1, i,
                         plot_data[plot_data$shot_first=="Greedo",
                                   "rank"][i],
                         ysize = 1)
 }
 p +
```

```
    ggtitle("Star Wars Film Rankings by Who Shot First") +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
}
```

Star Wars Film Rankings by Who Shot First