First thoughts on the gender citation gap

Mio Hienstorfer-Heitmann

10/12/2021

Contents

Simulation of Dion, Sumner, and Mitchell (2018)

To give you a small idea of my skills, but also to ingitiate interest in the topic of the project, I replicated the logistic regression model of the paper by Dion, Sumner, and Mitchell (2018) (Table 3). The authors used various logistic models to predict the probability that the authorship of a citation is all female given by the gender of the authors of the cited article.

Further, I will show a first scheme for a different approach that aims to rather the model individual citation counts as opposed to who cites whom.

The following lines will be heavy with code, as I want to be as transparent as possible with my approaches.

Defining the model

First, we define the logistic regression model function, whose likelihood we are going to maximise using a 'hill climbing' algorithm that traces the coefficients by iteratively maximising the loglikelihood function of the model.

```r
logit_fun <- function(y, X, theta){

    if(!is.null(ncol(X))){

      beta <- theta[1:ncol(X)]

      mu <- X %*% beta


    } else {

      beta <- theta[1]

      mu <- X * beta


    }

    p <- 1 / ( 1 + exp(-mu) )

    logll <-  sum( y * log(p) + (1 - y) * log ( 1- p) )

    return(logll)

  }
```

Estimating the coefficients

The following code defines a function that will:

1. Filter the data used in Dion, Sumner, and Mitchell (2018) to only use data of a specific journal.
2. Estimate the model parameters by applying a BFGS-hillclimbing algorithm.
3. Calculate robust standard errors using the `rms`-package.
4. Create a table that visualises the results and shows regression diagnostics.

```r
logistic_per_journal <- function(journal, vcovcoef = FALSE){

  require(tidyverse)
  require(MASS)
  require(rms)
  require(optimx)


  if(journal != "Pooled"){
    df_analysis <- df %>%
      filter(newjnlid %in% journal & refauthcomplete == 1) %>%
      dplyr::select(newjnlid, authorteam, reffemonly, newartid) %>%
      na.omit() %>%
      mutate(Female = ifelse(authorteam == "Female", 1, 0),
             Mixed = ifelse(authorteam == "Mixed", 1, 0)) %>%
      dplyr::select(-authorteam)


      y <- df_analysis$reffemonly

  X <- cbind(1,
             df_analysis$Female,
             df_analysis$Mixed)



    # start values
  startvals <- rep(0, ncol(X))

  # optimize
  res <- optim(
    par = startvals,
    fn = logit_fun,
    y = y,
    X = X,
    control = list(fnscale = -1),
```

```
    hessian = TRUE,
    method = "BFGS"
  )

startvals2 <- c(0, 0) # Why three this time?

restricted <- optim(
  startvals2,
  logit_fun,
  y = y,
  X = X[, 1],
  # restricted model
  control = list(fnscale = -1),
  method = "BFGS"
)

  coef <- res$par
  # vcov <- solve(-res$hessian)
  # se <- sqrt(diag(vcov))

  # Unfortunately, I am not yet able to compute robust standard errors
  # clustered at article level by hand. But I am on it.
  fit=lrm(data = df_analysis, reffemonly ~ Female + Mixed, x=T, y=T)
  vcov <- vcov(robcov(lrm(data = df_analysis, reffemonly ~ Female + Mixed,
                          x=T, y=T),
               cluster = df_analysis$newartid)
               )
  se <- sqrt(diag(vcov))
  ##robust standard error

  } else {

    df_analysis <- df %>%
      filter(refauthcomplete == 1) %>%
      dplyr::select(newjnlid, authorteam, reffemonly, newartid) %>%
      na.omit() %>%
```

```
    mutate(Female = ifelse(authorteam == "Female", 1, 0),
           Mixed = ifelse(authorteam == "Mixed", 1, 0),
           APSR = ifelse(newjnlid == "APSR", 1, 0),
           PG = ifelse(newjnlid == "Politics & Gender", 1, 0),
           PA = ifelse(newjnlid == "Political Analysis", 1, 0),
           Econ. = ifelse(newjnlid == "Econometrica", 1, 0),
           SMR = ifelse(newjnlid == "Soc. Methods & Res.", 1, 0)) %>%
    dplyr::select(-authorteam)

    y <- df_analysis$reffemonly

X <- cbind(1,
           df_analysis$Female,
           df_analysis$Mixed,
           df_analysis$PG,
           df_analysis$PA,
           df_analysis$Econ.,
           df_analysis$SMR)



  # start values
startvals <- rep(0, ncol(X))

# optimize
res <- optim(
  par = startvals,
  fn = logit_fun,
  y = y,
  X = X,
  control = list(fnscale = -1),
  hessian = TRUE,
  method = "BFGS"
)

startvals2 <- c(0, 0) # Why three this time?
```

```
restricted <- optim(
  startvals2,
  logit_fun,
  y = y,
  X = X[, 1],
  # restricted model
  control = list(fnscale = -1),
  method = "BFGS"
)

  coef <- res$par
  # vcov <- solve(-res$hessian)
  # se <- sqrt(diag(vcov))

  # Unfortunately, I am not yet able to compute robust standard errors
  # clustered at article level by hand. But I am on it.
  vcov <- vcov(robcov(lrm(data = df_analysis, reffemonly ~ Female + Mixed +
                            PG + PA + Econ. + SMR, x=T, y=T),
                cluster = df_analysis$newartid))
  #               )
  se <- sqrt(diag(vcov))
  ##robust standard error
  }

  Names = c("Intercept", "Female", "Mixed", "P&G", "PA", "Econ", "SMR",
            "Pseudo R2", "NullLL",  "LL", "Clusters", "Observations")
  if(journal == "Pooled"){

    ModelTable <- data.frame(Name = c(paste0(round(coef, 2), " (", round(se, 2), ")"),
                                                  round( 1- (restricted$value / res$value),
                                                  round(restricted$value, 0),
                                                  round(res$value,0),
                                                  length(unique(df_analysis$newartid)),
                                                   nrow(df_analysis))
                                )
```

```
  } else {

    ModelTable <- data.frame(Name = c(paste0(round(coef, 2), " (", round(se, 2), ")"),
                                                rep("", 4),
                                                round( 1- (restricted$value / res$value),
                                                round(restricted$value, 0),
                                                round(res$value,0),
                                                length(unique(df_analysis$newartid)),
                                                nrow(df_analysis)))
  }

  rownames(ModelTable) <- Names
  colnames(ModelTable) <- journal


  if(vcovcoef == FALSE){

     return(ModelTable)

  } else {

    return(list(coef = coef,
                vcov = vcov))
  }

}
```

The regression model

```
models <- do.call("cbind", lapply(unique(df$newjnlid), logistic_per_journal) )
pooled <- logistic_per_journal("Pooled")

flextable(cbind.data.frame(models, pooled) %>% rownames_to_column(" "))
```

	APSR	Politics & Gender	Political Analysis	Econometrics	Soc. Methods & Res.	Pooled
Intercept	-2.07 (0.05)	-0.01 (0.11)	-2.84 (0.09)	-3.18 (0.06)	-2.46 (0.1)	-2.02 (0.05)
Female	0.99 (0.16)	0.53 (0.12)	0.42 (0.38)	1.14 (0.22)	0.76 (0.28)	0.86 (0.1)
Mixed	0.21 (0.13)	-0.15 (0.16)	-0.08 (0.16)	0.07 (0.14)	0.06 (0.18)	0.11 (0.08)
P&G						1.73 (0.1)
PA						-0.89 (0.09)
Econ						-1.14 (0.07)
SMR						-0.47 (0.1)
Pseudo R2	-0.026	-0.0165	-7e-04	-0.0106	-0.0078	-0.2796
NullLL	-6359	-4007	-1249	-1951	-1185	-18566
LL	-6198	-3942	-1248	-1931	-1175	-14509
Clusters	464	332	295	604	232	1927
Observations	15648	5883	5891	10869	4053	42344

The table depicts the results of the regression analysis. Dion, Sumner, and Mitchell (2018) could show that the probability of an all-female citating authorship is associated with the gender of the cited authorship: 'female' citations are much more likely for female authors compared to male authors showing that women are more likely to cite each other.

What I dislike about this approach is that it may uncover gendered citation networks, but that it does not give any new insights on how to adress the gap in citations.

In the following section I want to approach the citation gap from a different angle.

Citation counts as a poisson process

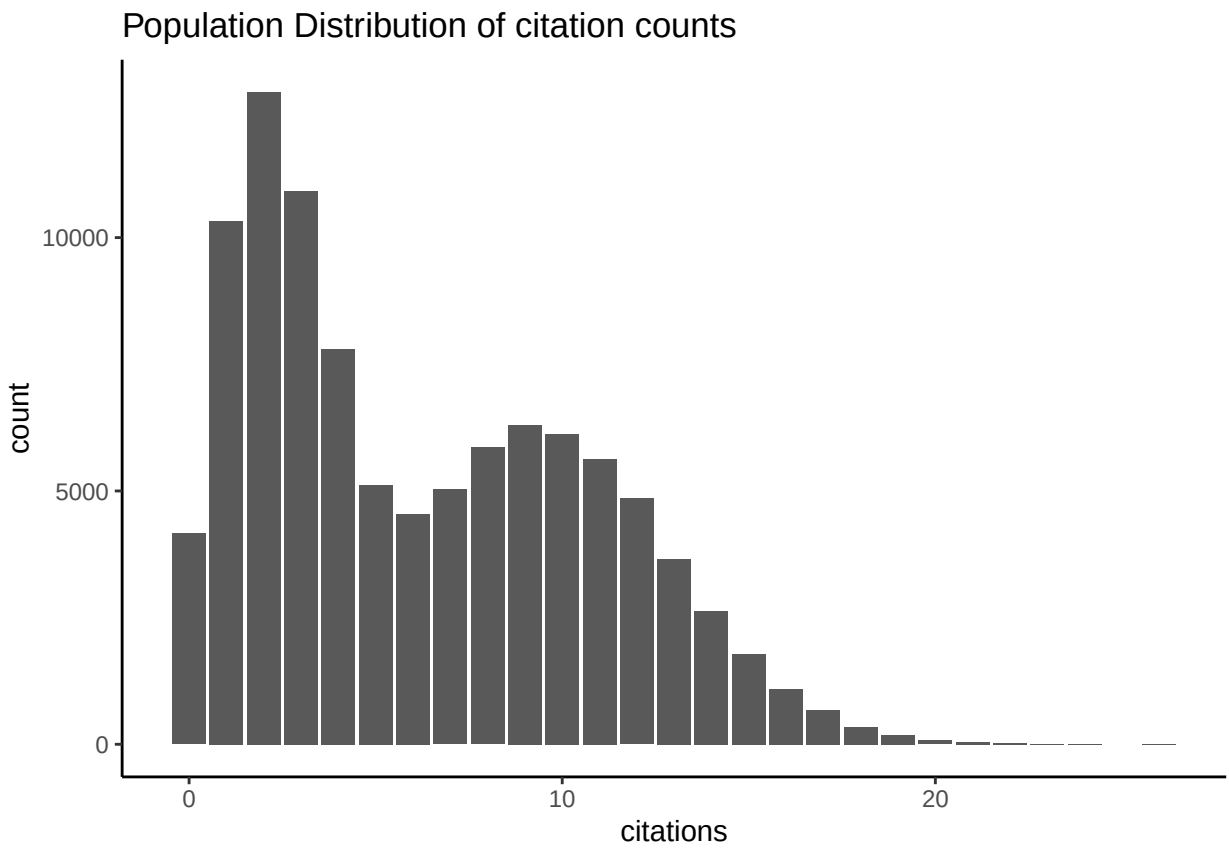
If the citation gap is the difference between men and women in their citation count, then the citation gap can be described by a count process.

For the sake of simplicity, I will assume a poisson distribution that creates different citation counts (which might be true for scientists higher up the career ladder, whereas for

the entire population including the academic *Mittelbau* a negative binomial process might be more appropriate.). And this process is (perhaps indirectly driven) by a variable gender.

So let's assume a population of 100,000 international political scientists (an increadibly rough estimate) , equally split between males and females (probably not entirely true) with a mean in citation counts around 10 ($beta_0$). In that population, we have a coefficient $beta_1$ that decreases the citation count of women by 4 on average. The figure depicts the distribution of the data in the population.

```
n <- 100000

female <- sample(size = n, x = c(0,1), replace = TRUE)

beta0 <- log(10)
beta1 <- -log(4)

lambda <- exp(beta0 + beta1 * female)

citations <- rpois(n, lambda)

population <- data.frame(citations = citations,
                 female = female)


ggplot(data = population,
       aes(x = citations,
           group = as.factor(female)))+
  geom_bar(stat = "count")+
  theme_classic()+
  labs(title = "Population Distribution of citation counts")
```

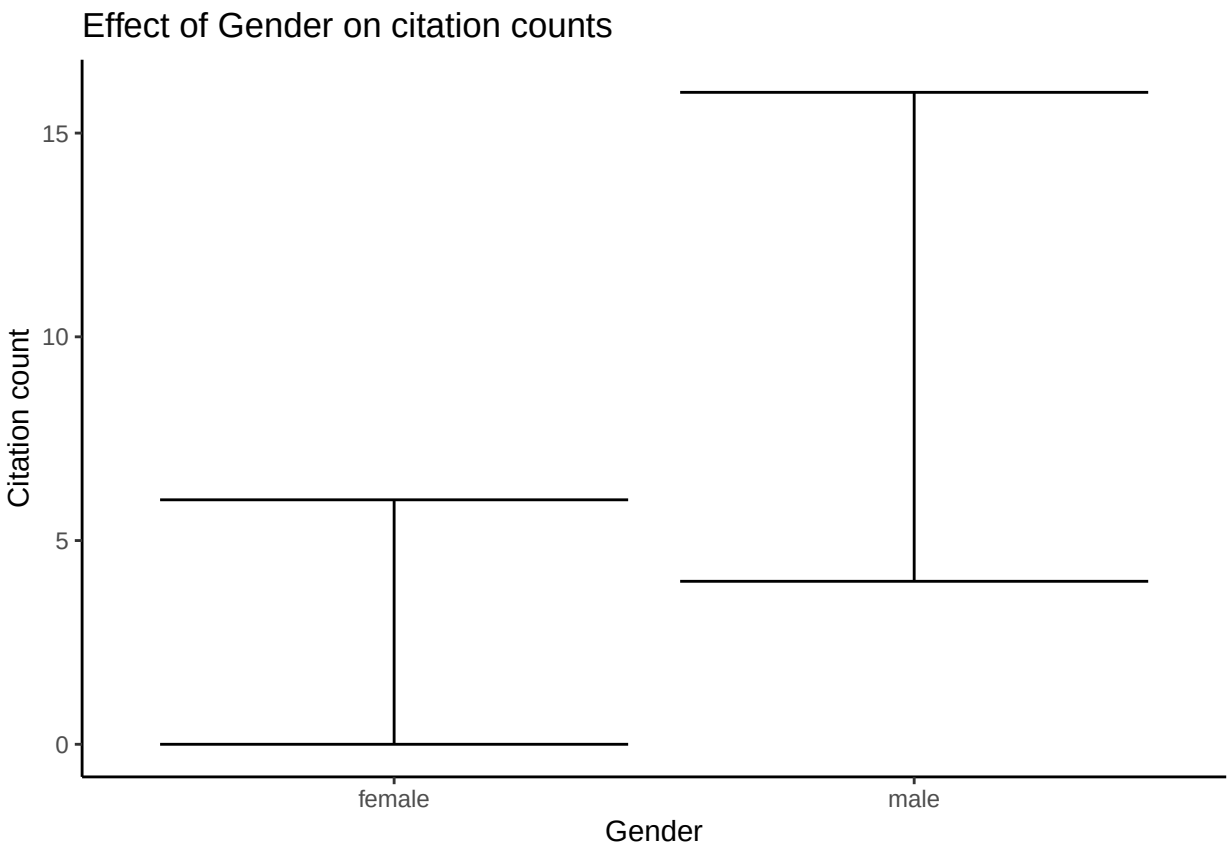


We can then model citation counts and simulate the data using MLE and a simulation to differentiate citation counts between men and women.

```
pois_ll <- function(y, X, theta){

  beta <- theta[1:ncol(X)]


  logll <- -sum(exp(beta %*% t(X))) +
    sum(y * (beta%*% t(X))) - sum(log(factorial(y)))


  return(logll)

}


sample <- population %>%
```

```r
  sample_n(500)

stval <- c(0, 0)
res <-
  optim(
    stval,
    pois_ll,
    y = sample$citations,
    X = cbind(rep(1, nrow(sample)), sample$female),
    control = list(fnscale = -1), # this is important
    hessian = TRUE,
    method = "BFGS"
    # and tell optimx to maximize rather than to minimize.
  )
coef <- res$par
varcov <- solve(-res$hessian)


## simulate the data


scenario <- c(0,1)
nsim <- 1000
S <- mvrnorm(nsim, coef, solve(-res$hessian))


sim_pois <- function(scenario, nsim, S){

  scenarios <- c(1, scenario)

  mu <- S %*% scenarios

  # response function
  lambda <- exp(mu)

  pois <- rpois(nsim, lambda)
```

```r
  #pois <- lambda

  mean <- mean(pois)
  lower <- quantile(pois, 0.025)
  upper <- quantile(pois, 0.975)

  return(data.frame(mean = mean,
                    lower = lower,
                    upper = upper,
                    scenario = scenario))

}

sim_df <- do.call("rbind", lapply(scenario, sim_pois, S = S, nsim = nsim))

sim_df$scenario <- ifelse(sim_df$scenario == 1, "female", "male")

ggplot(data = sim_df, aes(x = scenario, ymin = lower, max = upper))+
  labs(title = "Effect of Gender on citation counts",
       x = "Gender",
      y = "Citation count")+
  geom_errorbar()+
  theme_classic()
```



We can easily see that there is a difference in the population in the citation counts for male and female researchers.

In my view, the core aim of the CITATIONGAP project is then, to find all the factors that lie between gender and the citation count which may help to moderate the effect of gender on the frequency with which a scientist is cited. Because those moderating factors may help reduce that gap.

References

Dion, Michelle L, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. "Gendered Citation Patterns Across Political Science and Social Science Methodology Fields." *Political Analysis* 26 (3): 312–27.