Stat 245 – Test 2

Sharon Velpula

December 06, 2022

Reading in Data

```
myDataSet <- read.csv("https://osf.io/jtkn9/download")

nrow(myDataSet)
```

```
## [1] 935
```

Research Question

Response variable: indegree which is the size of the researcher's collaborative network Predictor variables: no.of.papers, Gender, n.years

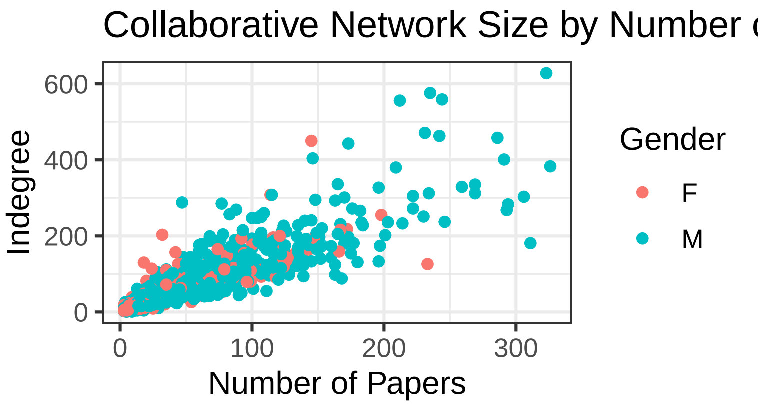
The question then is, is there an association between the size of the researcher's collaborative network (indegree) and number of papers published. Additionally, I would also like to answer if there is evidence of an interaction between gender and number of papers published.

Model Plan

I want to see how the number of papers published, gender, and years of academic career influences the size of the researcher's collaborative network. I have chosen these specific predictors because (1) according to the paper, there were gender differences observed in their study. Therefore, I have also added an interaction between gender and number of papers published because I would like to see how (if) gender influences the number of papers published. (2) I wanted to observe whether the years of a researcher's career progression affects the size of their collaborative network in a way that usually more time spent could lead to well built professional relationships. The number of predictors I have are 3 which is well within the rules of choosing predictors for the size of the given data set, where the number of observations (935) is more than 15 times larger than the number of predictors (3).

Data Exploration

```
gf_point(indegree ~ no.of.papers, color = ~Gender, data = myDataSet,
  xlab = 'Number of Papers',
  ylab = 'Indegree',
  title = "Collaborative Network Size by Number of Paper Published")
```



From this graph, we see that the size of the researcher's collaborative network is bigger as the number of papers published increases which affirms our research question by showing an association between the response and predictor variables. Additionally, men seem to have the highest number of publications with the biggest indegree value. This also ties in with what the paper said about gender differences in a lot of the variables such as the indegree, number of publications, number of co-authors, etc. We will now attempt to fit a model to further validate the claims made by this graph.

Fit & Summary of Model

```
require(glmmTMB)

mC1 <- glmmTMB(indegree ~ no.of.papers*Gender + n.years, data = myDataSet,
family = nbinom1(link = 'log'), na.action = 'na.fail')

mC2 <- glmmTMB(indegree ~ no.of.papers*Gender + n.years, data = myDataSet,
family = nbinom2(link = 'log'), na.action = 'na.fail')

BIC(mC1, mC2)
```

```
##     df      BIC
## mC1  6 9102.850
## mC2  6 9052.789
```

```
summary(mC2)
```

```
##  Family: nbinom2  ( log )
## Formula:          indegree ~ no.of.papers * Gender + n.years
## Data: myDataSet
##
##      AIC      BIC   logLik deviance df.resid
##   9023.7   9052.8  -4505.9   9011.7      929
##
##
## Dispersion parameter for nbinom2 family (): 2.96
##
## Conditional model:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.140781   0.076932  27.827  < 2e-16 ***
## no.of.papers             0.017324   0.001163  14.896  < 2e-16 ***
## GenderM                  0.330309   0.069489   4.753 2.00e-06 ***
## n.years                  0.038154   0.003080  12.386  < 2e-16 ***
## no.of.papers:GenderM -0.005710   0.001203  -4.745 2.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To better explain the coefficient estimate 0.017324 of the no.of.papers, this coefficient indicates a 0.017324 increase in indegree for every 1 count of no.of.papers, given that the other predictors n.years and Gender are kept constant.

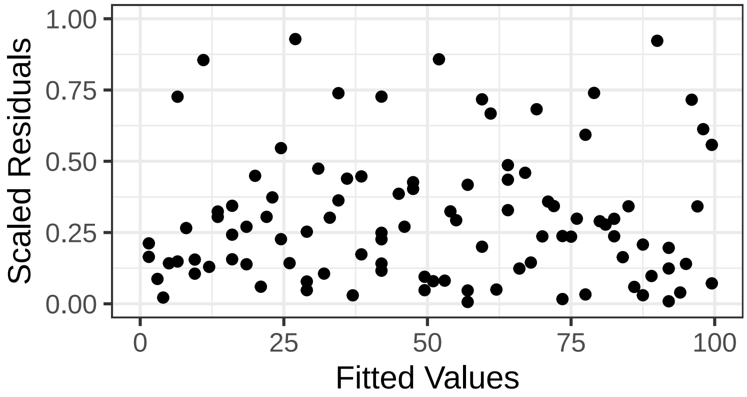
Choice of Model Family

I am using Negative Binomial Regression because I have to model the number of papers published which is basically count data. In my model, the response variable was the number of papers published by each researcher, and each researcher in the same experimental setting for the same amount of time, so there was no need to account for effort or time spent in each case. Hence, I did not find it necessary to include an offset in my model. The log link function is used in the negative binomial model here, which exponentiates the predictors.

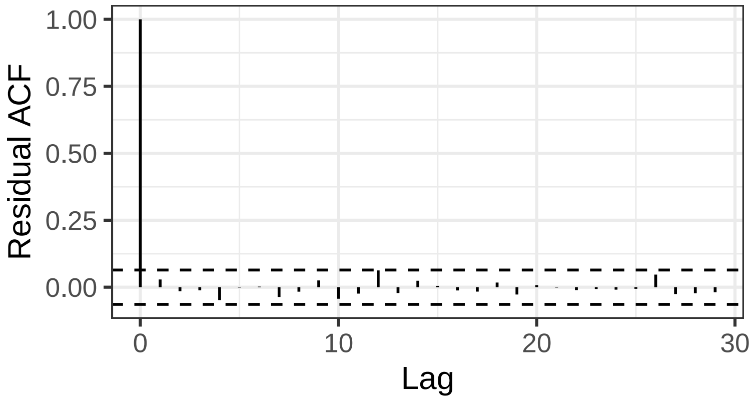
I also used BIC to select the best model, out of mC1 and mC2. It turns out that mC2 had the smaller BIC value so I chose it as my model.

Assessment

```
require(DHARMa)
sim_mC2 <- simulateResiduals(mC2)
gf_point(sim_mC2$scaledResiduals ~ rank(fitted(mC2))) %>%
gf_lims(x = c(0,100)) %>%
gf_labs(x = 'Fitted Values', y = 'Scaled Residuals')
```



```
s245::gf_acf(~mC2)
```



From what we can see, there is no evidence of any real trend and it looks like there is pretty uniform spread except for the top area of the plot because it seems sparse. I think this is because of some missing values or less observations. Overall, we can say that this model passes the condition of linearity but we must proceed with caution.

Model Selection

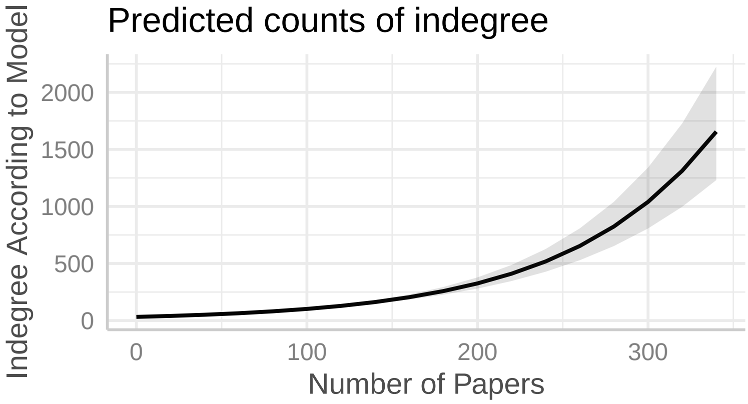
```
library(MuMIn)
dredge(mC2, rank = 'BIC')
```

```
## Global model call: glmmTMB(formula = indegree ~ no.of.papers * Gender + n.years,
##     data = myDataSet, family = nbinom2(link = "log"), na.action = "na.fail",
##     ziformula = ~0, dispformula = ~1)
## ---
## Model selection table
##    cnd((Int)) dsp((Int)) cnd(Gnd) cnd(n.yrs) cnd(no.of.ppr) cnd(Gnd:no.of.ppr)
## 16      2.141          +        +    0.03815        0.01732                  +
## 7       2.321          +             0.04067        0.01264
## 8       2.289          +        +    0.04044        0.01252
## 14      2.825          +        +                   0.02262                  +
## 5       3.150          +                            0.01637
## 6       3.097          +        +                   0.01617
## 4       1.806          +        +    0.08790
## 3       1.930          +             0.09105
## 2       4.019          +        +
## 1       4.397          +
##    df    logLik     BIC   delta weight
## 16  6 -4505.873  9052.8    0.00  0.997
## 7   4 -4518.687  9064.7   11.95  0.003
## 8   5 -4517.638  9069.5   16.69  0.000
## 14  5 -4577.249  9188.7  135.91  0.000
## 5   3 -4598.343  9217.2  164.42  0.000
## 6   4 -4596.395  9220.2  167.36  0.000
## 4   4 -4786.855  9601.1  548.28  0.000
## 3   3 -4798.869  9618.3  565.47  0.000
## 2   3 -5025.971 10072.5 1019.67  0.000
## 1   2 -5051.987 10117.7 1064.87  0.000
## Models ranked by BIC(x)
```

As we can see, model 16 seems to be the best model with number of years, number of papers, gender and the interaction between Gender and number of papers. It has the lowest BIC value.

Prediction Plot

```
require(ggeffects)
pred_plot_data_1 <- ggpredict(mC2, terms = 'no.of.papers')
plot(pred_plot_data_1) %>%
gf_labs(x = 'Number of Papers', y = 'Indegree According to Model')
```

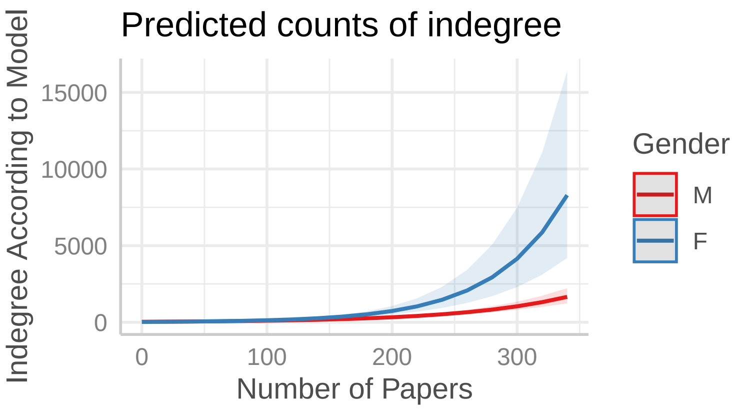


```
pred_plot_data_1
```

```
## # Predicted counts of indegree
##
## no.of.papers | Predicted |             95% CI
## ----------------------------------------------
##            0 |     31.92 | [  29.33,   34.73]
##           40 |     50.79 | [  48.11,   53.61]
##           80 |     80.82 | [  76.98,   84.85]
##          120 |    128.61 | [ 119.46,  138.45]
##          180 |    258.16 | [ 226.70,  294.00]
##          220 |    410.82 | [ 346.40,  487.21]
##          260 |    653.74 | [ 528.88,  808.08]
##          340 |   1655.45 | [1231.47, 2225.39]
##
## Adjusted for:
## *  Gender =     M
## * n.years = 26.00
```

From this prediction plot, we can see that that there is an increase in the size of a researcher's collaborative network with increasing number of papers. The predictors held constant were Gender: M and n.years: 26.

```
require(ggeffects)
pred_plot_data_2 <- ggpredict(mC2, terms = c('no.of.papers', 'Gender'))
plot(pred_plot_data_2) %>%
gf_labs(x = 'Number of Papers', y = 'Indegree According to Model')
```



```
pred_plot_data_2
```

```
## # Predicted counts of indegree
##
## # Gender = M
##
## no.of.papers | Predicted |              95% CI
## --------------------------------------------
##            0 |     31.92 | [  29.33,   34.73]
##           60 |     64.07 | [  61.14,   67.14]
##          120 |    128.61 | [ 119.46,  138.45]
##          180 |    258.16 | [ 226.70,  294.00]
##          220 |    410.82 | [ 346.40,  487.21]
##          340 |   1655.45 | [1231.47, 2225.39]
##
## # Gender = F
##
## no.of.papers | Predicted |               95% CI
## ---------------------------------------------
##            0 |     22.94 | [  20.33,    25.88]
##           60 |     64.86 | [  59.79,    70.36]
##          120 |    183.40 | [ 151.67,   221.78]
##          180 |    518.60 | [ 376.25,   714.81]
##          220 |   1037.03 | [ 688.04,  1563.04]
##          340 |   8291.81 | [4195.33, 16388.27]
##
## Adjusted for:
## * n.years = 26.00
```

From this prediction plot, we can see that that there is an increase in the size of a researcher's collaborative network with increasing number of papers, but more so for men than for women. The predictor n.years was held constant here with the value 26.

Interpretation and Conclusion

According to everything we have done, it seems that our model seems to answer our research question well enough. Our prediction plots showed an increase in indegree for increase in number of papers. Furthermore, the interaction between Gender and Number of papers also seems to exist since our prediction plot showed an even greater increase in the size of male researchers' collaborative network than female researchers' collaborative network. We had a preliminary evidence of this interaction through our data exploration graph. We are able to decide our predictions are valid because our model passed the condition of linearity as confirmed by the scaled residuals vs fitted plot, which showed no definite trend and had a uniform spread for the most part. Additionally, the choice of this model as being the best one was confirmed by our model selection, BIC where mC2 had the lowest BIC value than that of mC1. Within the model itself, the best model as suggested by dredge was model 16 with number of years, number of papers, Gender as predictors with an interaction between Gender and number of papers. This makes sense since there is evidence that Gender influences the number of papers published from the original paper. Therefore, the model we began with is the best one. If there is one thing that concerns me, it is that there is a greater uncertainty as the number of papers increase according to our prediction plots. Perhaps if more conditions were assessed, the model could be better validated. But overall, I think we are still able to infer from this model the answer for our research question, yes there is an association between the size of a researcher's collaborative network and the number of papers published.