Lab Seven

Shenyang Guo

3/1/2018

Example 1: Doctoral publications

- This example (Long, 1990) studies the number of articles published by biochemists in the 3 years prior to receiving their doctorate.
- Below we run two models, the Poisson regression and the negbin regression. We than compare the two models.

Load the libraries and get the data

```
#install.packages("margins")
library(haven) #read dta file
```

```
## Warning: package 'haven' was built under R version 3.4.3
```

```
library(MASS) #Negative binomial regression
```

```
## Warning: package 'MASS' was built under R version 3.4.3
```

```
library(lmtest) #model comparison
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(stargazer) #models presentation
```

```
## Warning: package 'stargazer' was built under R version 3.4.3
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```
library(sandwich)  #robust
```

```
## Warning: package 'sandwich' was built under R version 3.4.3
```

```
library(margins) #Marginal effects
```

```
## Warning: package 'margins' was built under R version 3.4.3
```

```
library(ggplot2) #Graphs
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
couart4 <- read_dta("https://github.com/kijohnson/Advanced-Data-Analysis/blob/master/Class%207%20Poisson%20and%20NegBin/Run%20Poisson%20&%20negbin%20with%20Stata/couart4.dta?raw=true")
```

Poisson regression

```
modP<- glm(art ~ factor(female) + factor(married) + kid5 + phd + mentor, family="poisson", data=couart4)
summary(modP)
```

```
##
## Call:
## glm(formula = art ~ factor(female) + factor(married) + kid5 +
##     phd + mentor, family = "poisson", data = couart4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.304617   0.102981   2.958   0.0031 **
## factor(female)1  -0.224594   0.054613  -4.112 3.92e-05 ***
## factor(married)1  0.155243   0.061374   2.529   0.0114 *
## kid5             -0.184883   0.040127  -4.607 4.08e-06 ***
## phd               0.012823   0.026397   0.486   0.6271
## mentor            0.025543   0.002006  12.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: 3314.1
##
## Number of Fisher Scoring iterations: 5
```

Negative Binomial Regression

```
modN<- glm.nb(art ~ factor(female) + factor(married) + kid5 + phd + mentor, data=couart4)
summary(modN)
```

```
##
## Call:
## glm.nb(formula = art ~ factor(female) + factor(married) + kid5 +
##     phd + mentor, data = couart4, init.theta = 2.264387695, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.256144   0.137348   1.865 0.062191 .
## factor(female)1  -0.216418   0.072636  -2.979 0.002887 **
## factor(married)1  0.150489   0.082097   1.833 0.066791 .
## kid5             -0.176415   0.052813  -3.340 0.000837 ***
## phd               0.015271   0.035873   0.426 0.670326
## mentor            0.029082   0.003214   9.048  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2644) family taken to be 1)
##
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
## AIC: 3135.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.264
##           Std. Err.:  0.271
##
##  2 x log-likelihood:  -3121.917
```

Compare models using likelihood ratio test

```
lrtest(modP, modN)
```

```
## Likelihood ratio test
##
## Model 1: art ~ factor(female) + factor(married) + kid5 + phd + mentor
## Model 2: art ~ factor(female) + factor(married) + kid5 + phd + mentor
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   6 -1651.1
## 2   7 -1561.0  1 180.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The highly significant p-value means that the negbin regression is better than the Poisson regression, we should accept the negbin results other than the Poisson regression.

Further compare two models

Below we further compare the estimates between the two models. As the results show, the Poisson regression estimates SEs that are always smaller (as shown by the narrower CIs) than those from the negbin. This implies that the Poisson regression leads to biased significance tests, and tends to make non-significant predictors significant.

```
stargazer(modP, modN, title="Model Comparison",
          type="text",align=TRUE,single.row=TRUE)
```

```
##
## Model Comparison
## ================================================
##                          Dependent variable:
##                    ---------------------------
##                                art
##                         Poisson      negative
##                                      binomial
##                           (1)           (2)
## ------------------------------------------------
## factor(female)1    -0.225*** (0.055) -0.216*** (0.073)
## factor(married)1    0.155** (0.061)   0.150* (0.082)
## kid5               -0.185*** (0.040) -0.176*** (0.053)
## phd                  0.013 (0.026)     0.015 (0.036)
## mentor             0.026*** (0.002)  0.029*** (0.003)
## Constant           0.305*** (0.103)   0.256* (0.137)
## ------------------------------------------------
## Observations             915               915
## Log Likelihood       -1,651.056        -1,561.958
## theta                                2.264*** (0.271)
## Akaike Inf. Crit.     3,314.113         3,135.917
## ================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Example 2: Use robust SE and obtain incidence rate ratios (IRRs)

To follow the convention in running negbin, we always use robust estimator of standard errors. This should also be done when you run the Poisson regression.

Always use robust SE for the final model

```
robust<-coeftest(modN, vcov = sandwich)
robust
```

```
##
## z test of coefficients:
##
##                    Estimate Std. Error z value  Pr(>|z|)
## (Intercept)       0.2561440  0.1397356  1.8331 0.0667934 .
## factor(female)1  -0.2164184  0.0707922 -3.0571 0.0022349 **
## factor(married)1  0.1504895  0.0806362  1.8663 0.0620028 .
## kid5             -0.1764152  0.0527560 -3.3440 0.0008258 ***
## phd               0.0152712  0.0374419  0.4079 0.6833744
## mentor            0.0290823  0.0033511  8.6783 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below are the syntax and output for obtaining incidence rate ratios and 95%CIs

```
est <- cbind(IRR = coef(modN), "2.5%"=robust[,1]-1.96*robust[,2], "97.5%"=robust[,1]+1.96*robust[,2])
exp(est)
```

```
##                        IRR      2.5%     97.5%
## (Intercept)      1.2919388 0.9824186 1.6989762
## factor(female)1  0.8053982 0.7010535 0.9252737
## factor(married)1 1.1624030 0.9924709 1.3614312
## kid5             0.8382698 0.7559220 0.9295883
## phd              1.0153884 0.9435415 1.0927061
## mentor           1.0295094 1.0227695 1.0362937
```

Get incidence rate ratios for continuous variables per standard deviation increase (instead of per unit change)

Use the z-transformed values (z = (x-mean(x))/sd(x)) of the predictor to fit the model. The coefficients will then be log odds ratios for one SD change of the predictor.

```
couart4$mentor_sd<-(couart4$mentor-mean(couart4$mentor))/sd(couart4$mentor)
couart4$kid5_sd<-(couart4$kid5-mean(couart4$kid5))/sd(couart4$kid5)
couart4$phd_sd<-(couart4$phd-mean(couart4$phd))/sd(couart4$phd)
modN2<- glm.nb(art ~ factor(female) + factor(married) + kid5_sd + phd_sd + mentor_sd, data=couart4)
#Compare two models
est1 <- cbind(UnitChange = coef(modN), SDChange=coef(modN2))
exp(est1)
```

```
##                 UnitChange  SDChange
## (Intercept)       1.2919388 1.6018563
## factor(female)1   0.8053982 0.8053982
## factor(married)1  1.1624030 1.1624030
## kid5              0.8382698 0.8737714
## phd               1.0153884 1.0151441
## mentor            1.0295094 1.3176034
```

Interpretations

- For a categorical variable ("female"):
- - *Being a female scientist decreases the expected number of articles by a factor of 0.805, holding other variables constant.*
- - *Being a female scientist decreases the expected number of articles by 19.5%, holding other variables constant.*
- For a continuous variable ("mentor"):
- - *For a standard deviation increase in the mentor's productivity, roughly 10 articles, a scientist's expected productivity increases by a factor of 1.318, holding other variables constant.*
- - *For every additional article by the mentor, a scientist's expected productivity increase by 3.0%, holding other variables constant.*
- - *For a standard deviation increase in the mentor's productivity, roughly 10 articles, a scientist's expected productivity increases by 31.8%, holding other variables constant.*

Example 3: Predicted probabilities (MEMs, MERs, & AMEs)

- Below I show the commands and output of obtaining predicted probabilities. There are three types of them: MEMs, MERs, and AMEs.

MEMs

```
#Get the marginal effects at mean
margins(modN, couart4, atmean=TRUE) ##Explain
```

```
## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!
```

```
## Average marginal effects
```

```
## glm.nb(formula = art ~ factor(female) + factor(married) + kid5 +     phd + mentor, data = couart4, init.theta = 2.264387695, link = log)
```

```
##     kid5     phd  mentor female1 married1
##  -0.3008 0.02604 0.04958 -0.3637   0.2505
```

The number of papers published decreases by 0.3 with one more children under 5, holding other variable at mean levels.

```
#Define our outcomes from 0 to 5
#Obtain means to create predicted probabilities
modN1<- glm.nb(art ~ female + married + kid5 + phd + mentor, data=couart4)
summary(modN1)
```

```
##
## Call:
## glm.nb(formula = art ~ female + married + kid5 + phd + mentor,
##     data = couart4, init.theta = 2.264387695, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.256144   0.137348   1.865 0.062191 .
## female      -0.216418   0.072636  -2.979 0.002887 **
## married      0.150489   0.082097   1.833 0.066791 .
## kid5        -0.176415   0.052813  -3.340 0.000837 ***
## phd          0.015271   0.035873   0.426 0.670326
## mentor       0.029082   0.003214   9.048  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2644) family taken to be 1)
##
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
```

```
## AIC: 3135.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.264
##           Std. Err.:  0.271
##
##  2 x log-likelihood:  -3121.917
```

```
newdata1 <- data.frame(female = mean(couart4$female), married=mean(couart4$married), kid5=mean(couart4$kid5), phd=mean(couart4$phd), mentor= mean(couart4$mentor))
newdata1
```

```
##      female   married     kid5      phd   mentor
## 1 0.4601093 0.6622951 0.495082 3.103109 8.767213
```

```
phat <- predict(modN1, newdata1, type = "response")
cbind(No_of_Ppaers=0:5, MEM=dnbinom(0:5, mu=phat, size=modN1$theta))#events (no of papers) = 0 to 5
```

```
##      No_of_Ppaers        MEM
## [1,]            0 0.29775427
## [2,]            1 0.27936151
## [3,]            2 0.18892789
## [4,]            3 0.11127287
## [5,]            4 0.06067851
## [6,]            5 0.03149933
```

The probability of publishing 0 to 5 papers is 29.8%, 27.9%, 18.9%, 11.1%, 6.1% and 3.1% respectively, holding all variables at mean levels.

MERs (marginal effect at representative values) to see the effect of mentor's publications

```
margins(modN, at = list(mentor = 1:6))
```

```
## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!
```

```
## Average marginal effects at specified values
```

```
## glm.nb(formula = art ~ factor(female) + factor(married) + kid5 +     phd + mentor, data = couart4, init.theta = 2.264387695, link = log)
```

```
##  at(mentor)    kid5     phd  mentor female1 married1
##           1 -0.2279 0.01973 0.03757 -0.2767   0.1899
##           2 -0.2346 0.02031 0.03868 -0.2849   0.1955
##           3 -0.2415 0.02091 0.03982 -0.2933   0.2012
##           4 -0.2487 0.02153 0.04099 -0.3019   0.2072
##           5 -0.2560 0.02216 0.04220 -0.3108   0.2133
##           6 -0.2636 0.02282 0.04345 -0.3200   0.2196
```

When the mentor publishes 1 paper, the number of papers published by PhD students decreases by 0.2 with one more children under 5, holding all other variable at mean levels.

```
#Define our outcomes from 0 to 5
##Obtain means to create predicted probabilities
options(digits=3)
newdata2 <- data.frame(female = mean(couart4$female), married=mean(couart4$married), kid5=mean(couart4$kid5), phd=mean(couart4$phd), mentor= 1:6)
newdata2
```

```
##   female married  kid5 phd mentor
## 1   0.46   0.662 0.495 3.1      1
## 2   0.46   0.662 0.495 3.1      2
## 3   0.46   0.662 0.495 3.1      3
## 4   0.46   0.662 0.495 3.1      4
## 5   0.46   0.662 0.495 3.1      5
## 6   0.46   0.662 0.495 3.1      6
```

```
phat <- predict(modN1, newdata2, type = "response")
mentor1=dnbinom(0:5, mu=phat[1], size=modN1$theta) #mentor=1; events (no of papers) = 0 to 5
mentor2=dnbinom(0:5, mu=phat[2], size=modN1$theta) #mentor=2; events (no of papers) = 0 to 5
mentor3=dnbinom(0:5, mu=phat[3], size=modN1$theta) #mentor=3; events (no of papers) = 0 to 5
mentor4=dnbinom(0:5, mu=phat[4], size=modN1$theta) #mentor=4; events (no of papers) = 0 to 5
mentor5=dnbinom(0:5, mu=phat[5], size=modN1$theta) #mentor=5; events (no of papers) = 0 to 5
```

```
mentor6=dnbinom(0:5, mu=phat[6], size=modN1$theta) #mentor=6; events (no of papers) = 0 to 5
event<-c(0:5) #events (no of papers) = 0 to 5
rbind(event, mentor1, mentor2,mentor3,mentor4,mentor5, mentor6)
```

```
##          [,1]  [,2]  [,3]   [,4]   [,5]   [,6]
## event   0.000 1.000 2.000 3.0000 4.0000 5.0000
## mentor1 0.363 0.297 0.175 0.0896 0.0425 0.0192
## mentor2 0.354 0.295 0.177 0.0924 0.0447 0.0206
## mentor3 0.346 0.293 0.179 0.0953 0.0469 0.0220
## mentor4 0.337 0.291 0.181 0.0981 0.0492 0.0235
## mentor5 0.329 0.289 0.183 0.1010 0.0516 0.0251
## mentor6 0.321 0.287 0.185 0.1037 0.0539 0.0267
```

When the mentor publishes 1 paper, the probability of publishing 0 to 5 papers is 36.3%, 29.7%, 17.5%, 9.0%, 4.3% and 1.9% respectively, holding all variables at mean levels.

Dr.Guo developed an Excel file (in the class repository) to obtain model predicted probabilities, primarily for MEMs and MERs. Results confirm that the two programs provide exactly the same results.

Obtain predicted probabilities based on a negative binomial regression
Developed by Dr. Shenyang Guo December 29, 1999 (All Rights Reserved)
Predicted Probability based on Estimated Negative Bionomial Regression Model
Example taken from Long, S. (1997), p.228 & formula on p. 237.

Formula for computation of predicted probabilities:

$$\hat{\Pr}(y \mid x) = \frac{\Gamma(y+\hat{\alpha}^{-1})}{y!\Gamma(\hat{\alpha}^{-1})}\left(\frac{\hat{\alpha}^{-1}}{\hat{\alpha}^{-1}+\hat{\mu}}\right)^{\hat{\alpha}^{-1}}\left(\frac{\hat{\mu}}{\hat{\alpha}^{-1}+\hat{\mu}}\right)^{y}$$

$$where \quad \hat{\mu} = \exp(x\hat{\beta})$$

1	2.264387636	=POWER(D20,-1)		
2	2.264387636	=D21+G10		
3	1.142440587	=EXP(GAMMALN(G11))		
4	1.142440587	=EXP(GAMMALN(G10))		
5	1.142440587	=FACT(D21)*G13		
6	0.297754333	=POWER((G10/(G10+D19)),G10)		
7	1	=POWER((D19/(G10+D19)),D21)		
8	1	=G12/G14		
		P=8*6*7		

Variable	*B*	*Mean*	*B*Mean*	*Mean*
Constant	0.256144	1	0.256144	1
female	-0.2164184	0.4601093	-0.099576119	0.4601093
married	0.1504895	0.6622951	0.099668458	0.6622951
kid5	-0.1764152	0.495082	-0.08733999	0.495082
phd	0.0152712	3.103109	0.047388198	3.103109
mentor	0.0290823	8.767213	0.254970719	8.767213
		sum or xB	0.471255267	
		u=exp(xB)	1.602003874	
		Dispersion or a	0.4416205	
		y	0	

Pr(Y)	0.297754333

MEMs (all variables are at means)

Pr(y=0)	Pr(y=1)	Pr(y=2)	Pr(y=3)	Pr(y=4)	Pr(y=5)
0.298	0.279	0.189	0.111	0.061	0.031

MERs (The impact of mentor's publications, all other variables at means)

mentor's # of pub	Pr(y=0)	Pr(y=1)	Pr(y=2)	Pr(y=3)	Pr(y=4)	Pr(y=5)
1	0.363	0.297	0.175	0.090	0.043	0.019
2	0.354	0.295	0.177	0.092	0.045	0.021
3	0.346	0.293	0.179	0.095	0.047	0.022
4	0.337	0.291	0.181	0.098	0.049	0.024
5	0.329	0.289	0.183	0.101	0.052	0.025
6	0.321	0.287	0.185	0.104	0.054	0.027

Example

AMEs

The commands below create predicted probabilities for the count equal to 1, 2, …m for each case. Averaging the sample all cases gives the AME, as:

```
margins(modN, couart4)
```

```
## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!
```

```
## Average marginal effects
```

```
## glm.nb(formula = art ~ factor(female) + factor(married) + kid5 +     phd + mentor, data = couart4, init.theta = 2.264387695, link = log)
```

```
##     kid5     phd  mentor female1 married1
##  -0.3008 0.02604 0.04958 -0.3637   0.2505
```

On average, with a one-unit increase in the number of papers published by mentor, the student is predicted to have 0.05 more papers published, other things equal.

Example 4: Graphic representation of the findings

Below I show how to present a line chart depicting the impact of a continuous variable on the expected number of articles by group. I use the number of mentor's publications as the continuous variable, and gender as the group. In order to determine the scale of x-axis in the chart, I first take a look at the distribution of the continuous variable.

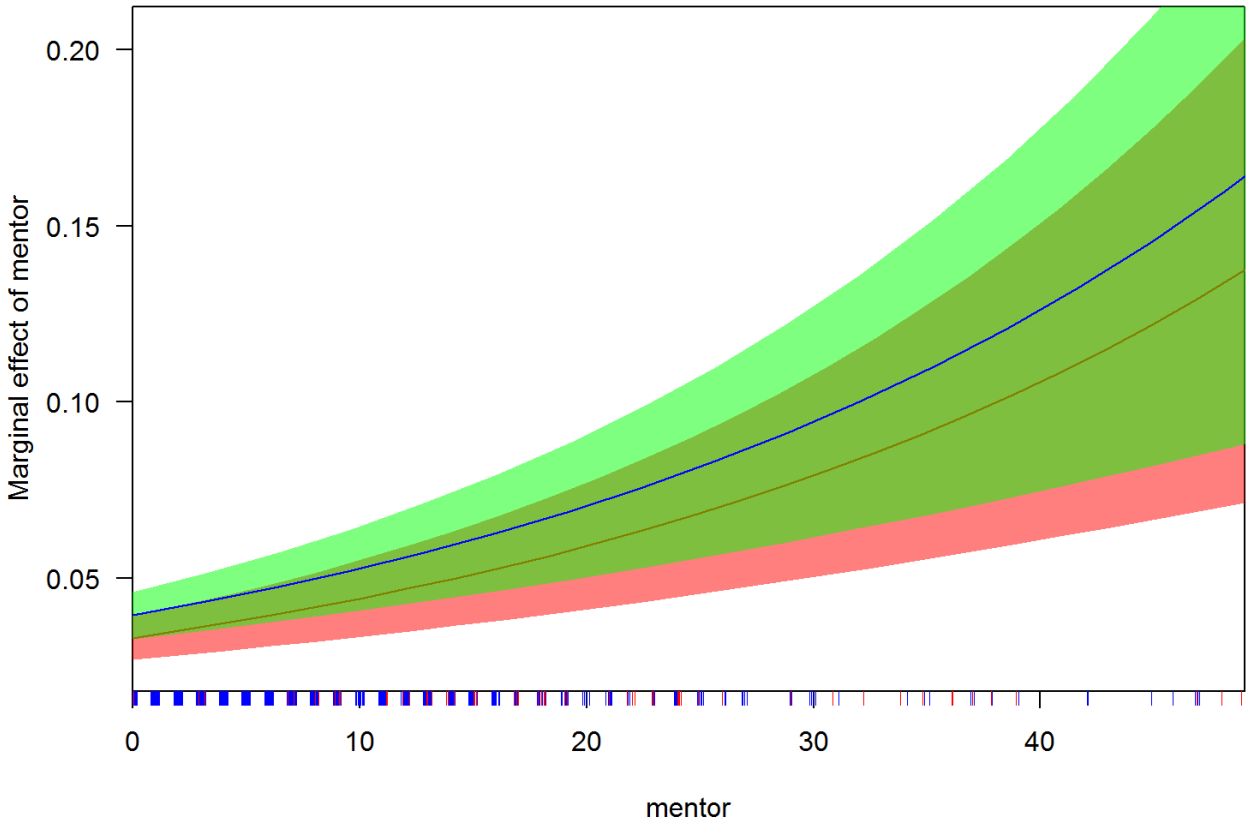
```
table(couart4$mentor)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
## 90 52 79 70 72 77 52 44 47 32 32 34 27 21 23 19 16  9 14 14  4 12  6  5  8
## 25 26 27 29 30 31 32 34 35 36 37 38 39 42 45 46 47 48 49 53 55 57 66 77
##  7  4  4  3  5  2  1  2  3  2  3  2  2  2  1  1  4  1  1  2  1  1  1  1
```

```
local({
  cplot(modN1, x="mentor", what="effect", data=couart4[couart4$female==1,],col="red",se.fill = rgb(1,0,0,.5))
  cplot(modN1, x="mentor", what="effect", data=couart4[couart4$female==0,],draw="add",col = "blue",se.fill = rgb(0,1,0,.5))
})
```

```
## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!

## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!
```

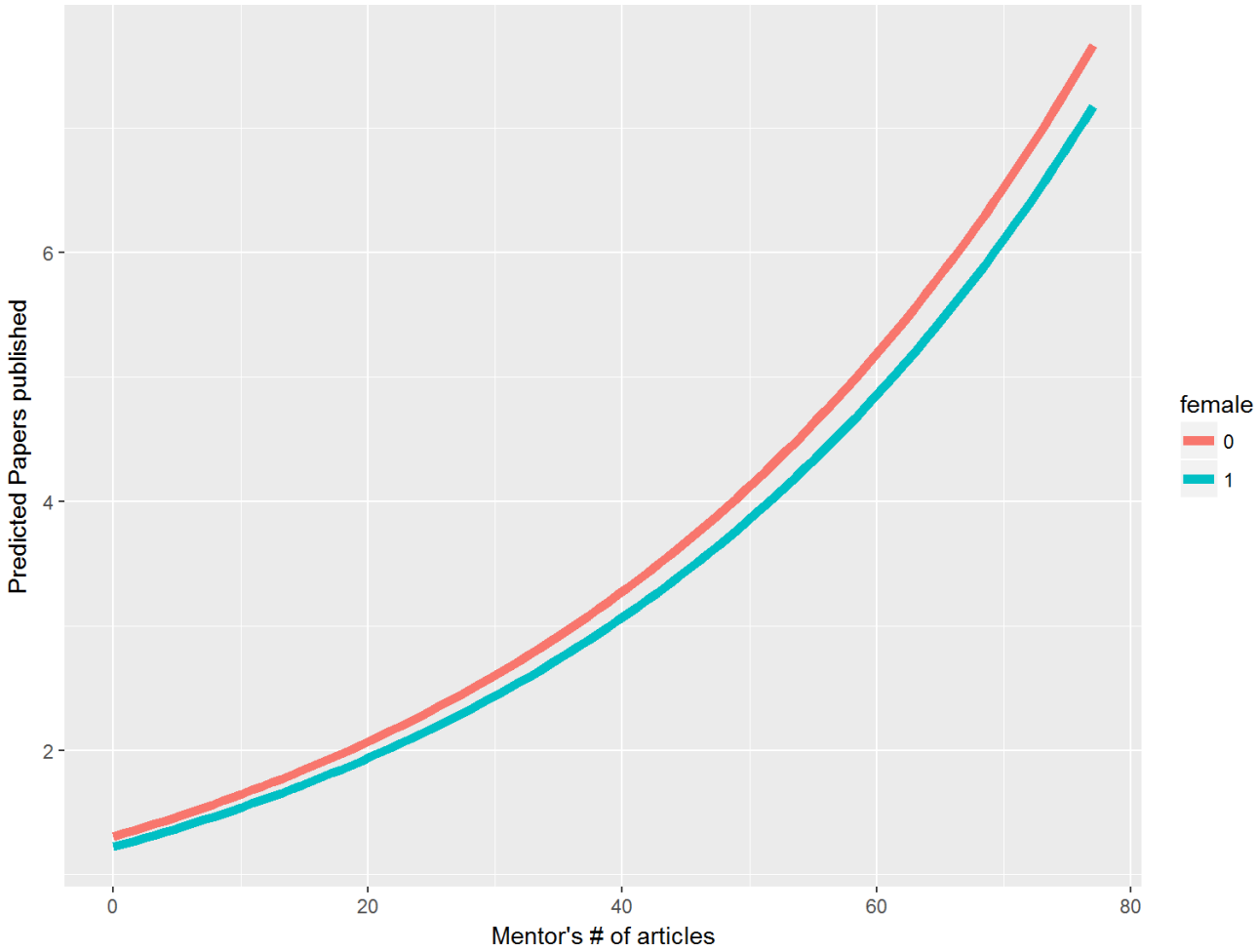



```
#OR if we want y axis to be the counts of papers
newdata2 <- data.frame(
  mentor = rep(seq(from = min(couart4$mentor), to = max(couart4$mentor), length.out = 100), 2),
  female = factor(rep(0:1, each = 100)),
  married = factor(rep(0:1, each = 100)),
  phd = rep(seq(from = min(couart4$phd), to = max(couart4$phd), length.out = 100), 2),
  kid5 = rep(seq(from = min(couart4$kid5), to = max(couart4$kid5), length.out = 100), 2))

newdata2 <- cbind(newdata2, predict(modN, newdata2, type = "link", se.fit=TRUE))
newdata2 <- within(newdata2, {
  art <- exp(fit)
})

ggplot(newdata2, aes(mentor, art)) +
```

```
  geom_line(aes(colour = female), size = 2) +
  labs(x = "Mentor's # of articles", y = "Predicted Papers published")
```



Example 5: Test interaction

Testing interaction is an important procedure in statistical modeling. This often determines by a study's research questions or conceptual model. Researchers also sometimes test significant interaction through a data-driven procedure. In this example, I am interested in the joint effect of PhD program's prestige and mentor's number of publications on the study student's number of publications. Since both variables are continuous, we need to categorize one of the two variables first so that the interaction effects show the impact of a continuous variable on the DV by the level of the categorical variable. We first tabulate the two continuous variables. From the distribution, I choose mentor's publications as a variable on which I dichotomous it. I use 20 articles as a cutoff: those who published more than 20 articles as high. So the study investigates how the impact of PhD program's prestige on students' publication varies by mentor's high- versus low-publication status. Results of this interaction is very interesting, and important. Note that the interaction is statistically significant.

```
couart4$mentor_cat<-ifelse(couart4$mentor>20,1,0)
modN3<- glm.nb(art ~ female + married + kid5 + phd*mentor_cat, data=couart4)
summary(modN3)
```

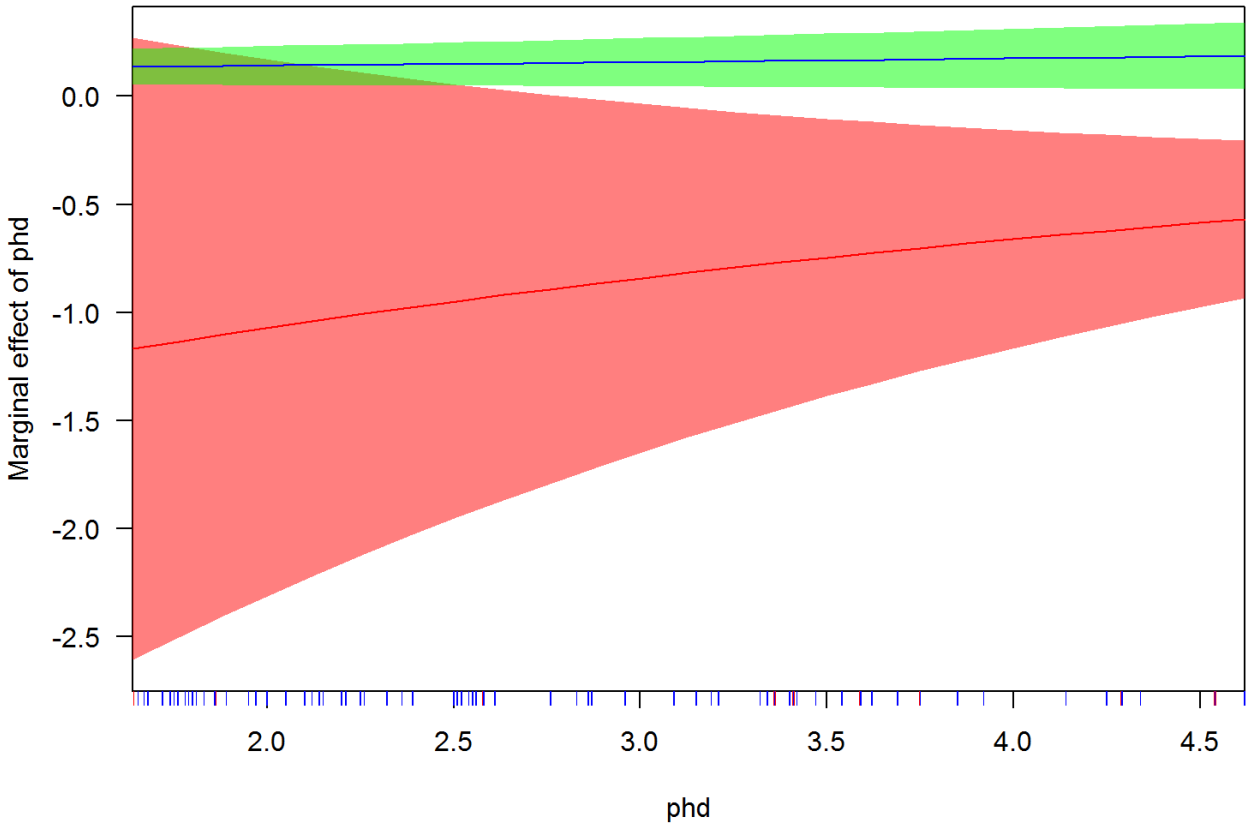
```
##
## Call:
## glm.nb(formula = art ~ female + married + kid5 + phd * mentor_cat,
##     data = couart4, init.theta = 2.178114933, link = log)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.280  -1.406  -0.275   0.403   3.484
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.2161     0.1440    1.50   0.1333
## female          -0.2388     0.0730   -3.27   0.0011 **
## married          0.1118     0.0828    1.35   0.1767
## kid5            -0.1536     0.0528   -2.91   0.0036 **
## phd              0.1033     0.0377    2.74   0.0062 **
## mentor_cat       1.8413     0.3935    4.68  2.9e-06 ***
## phd:mentor_cat  -0.3445     0.1098   -3.14   0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.18) family taken to be 1)
##
##     Null deviance: 1094.4  on 914  degrees of freedom
## Residual deviance: 1012.0  on 908  degrees of freedom
## AIC: 3159
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.178
##           Std. Err.:  0.260
##
##  2 x log-likelihood:  -3142.526
```

```
modN3<- glm.nb(art ~ female + married + kid5 + phd*mentor_cat, data=couart4)
local({
  cplot(modN3, x="phd", what="effect", data=couart4[couart4$mentor_cat==1,],col="red",se.fill = rgb(1,0,0,.5))
  cplot(modN3, x="phd", what="effect", data=couart4[couart4$mentor_cat==0,],draw="add",col = "blue",se.fill = rgb(0,1,0,.5))
})
```

```
## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!

## Warning in warn_for_weights(model): 'weights' used in model estimation are
## currently ignored!
```

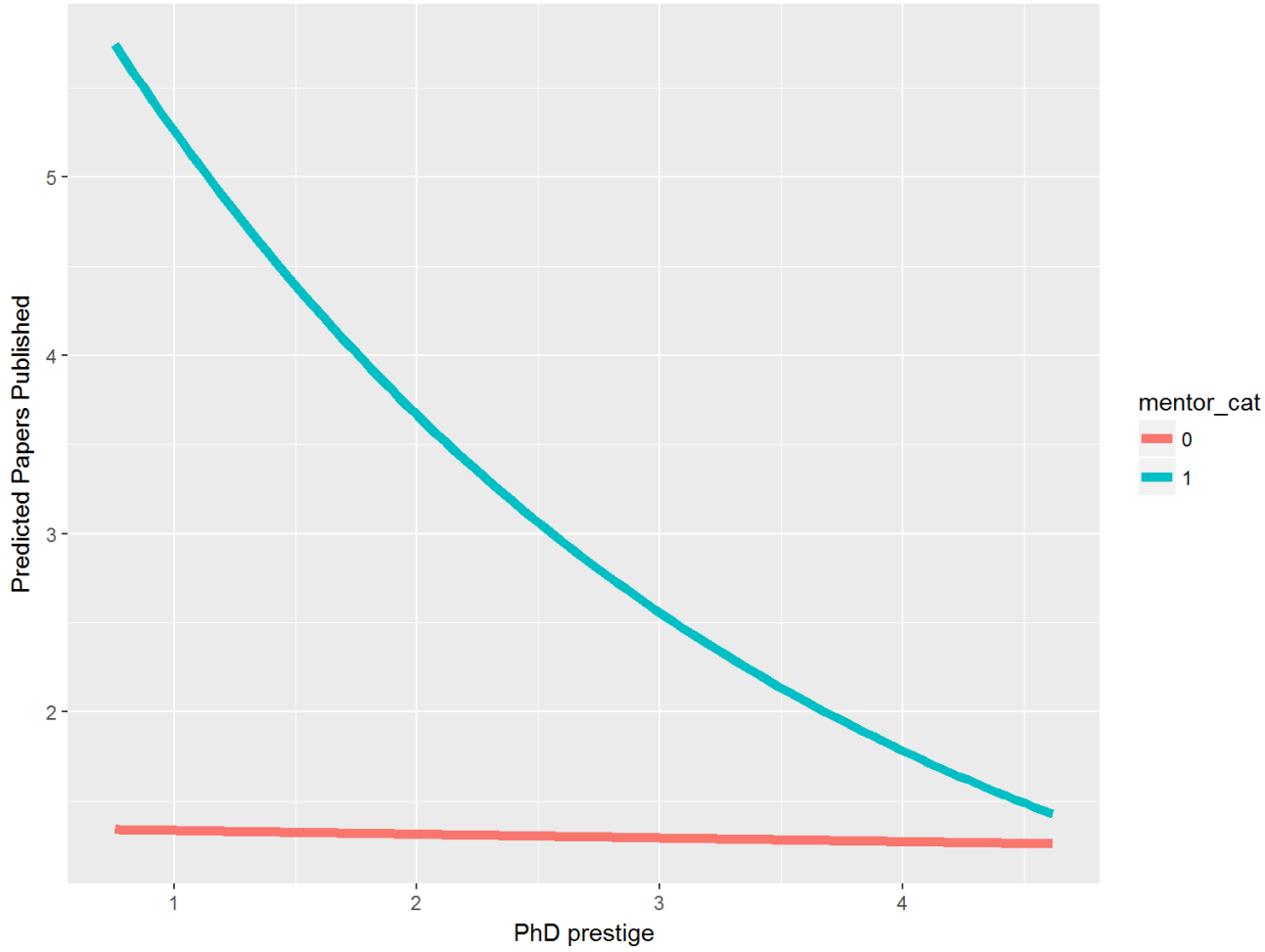



```
#OR if we want y axis to be the counts of papers
modN4<- glm.nb(art ~ factor(female) + factor(married) + kid5 + phd*factor(mentor_cat), data=couart4)

newdata4 <- data.frame(
  mentor_cat = factor(rep(0:1, each = 100)),
  female = factor(rep(0:1, each = 100)),
  married = factor(rep(0:1, each = 100)),
  phd = rep(seq(from = min(couart4$phd), to = max(couart4$phd), length.out = 100), 2),
  kid5 = rep(seq(from = min(couart4$kid5), to = max(couart4$kid5), length.out = 100), 2))

newdata4 <- cbind(newdata4, predict(modN4, newdata4, type = "link", se.fit=TRUE))
newdata4 <- within(newdata4, {
  art <- exp(fit)
})
```

```
ggplot(newdata4, aes(phd, art)) +
  geom_line(aes(colour = mentor_cat), size = 2) +
  labs(x = "PhD prestige", y = "Predicted Papers Published")
```



The results show that PhD program's impact on students' publications varies by the mentor's productivity rate. For a mentor who is not productive, the impact of PhD program's prestige has a positive impact on students' expected number of publications: the higher the prestige, the more articles the student produced. However, if the mentor is very productive, the impact of PhD program's prestige on students' productivity is negative: the higher the prestige, the lower the students' productivity. This finding is very interesting! Am I right that working with a very productive professor in a highly prestigious program is not necessarily productive? The chart shows that working with a professor who is very

productive in a program with the highest prestige is the same as working with a not-productive professor in the same program. Don't generalize the findings without caution. Note that this study is about biochemist PhD students.