Academic Productivity between Ph.D Students and Their Mentors

Cheng Peng

A Sample Case Study Report for STA321

Contents

1 Introduction

In this case study, we use data from Long (1990) on the number of publications produced by Ph.D. biochemists to illustrate the application of Poisson models. The variables in the data set are listed below.

1.1 Variable Description

- articles: integer. articles in the last three years of Ph.D.
- gender: factor. coded one for females.
- married: factor. coded one if married.
- kids: integer. the number of children under age six.
- prestige: numeric.the prestige of Ph.D. program
- mentor: integer. articles by the mentor in last three years

1.2 Research Question

We want to assess how factors affect the number of articles published in the last three years in the Ph.D. programs.

2 EDA and Feature Engineering

Variable **kids** is a discrete variable. We created a frequency table of **kids** and found that 16 of 915 Ph.D. students had 3 kids. After additional exploratory analysis. We decide to dichotomize **kids** and redefine

a new variable under the name **newkids**.

```
phd=read.table("w10-ph-data.txt",skip=10, header=TRUE )[,-1] # drop the ID variable
id.3 = which(phd$kids > 0)
newkids = phd$kids
newkids[id.3] = 1
phd$newkids = newkids
```

3 Poison Regression Modeling

We build both the regular Poisson and Quasi-Poisson regression models and extract the dispersion parameter to decide which model should be used as a working model.

3.1 Extracting Dispersion Index

Recall that the theoretical dispersion index ϕ of a distribution is defined to be

$$\phi = \frac{\text{variance}}{\text{mean}}.$$

Under the assumption of Poisson regression, i.e., $\text{var(Y)} = \text{E}(Y)$, the variance of Y can be estimated by either Pearson residuals or deviance residuals which are detailed in the lecture note. This means that $\phi = 1$. On the other hand, using asymptotic tools in *mathematical statistics*, we can show that the sum of squared Pearson (or deviance) residuals is approximately distributed as χ^2_{df} with $\text{df} = n - p$, n is the sample size and p is the number of parameters in the model. Furthermore, we know that the $\text{E}(\chi^2_{\text{df}}) = \text{df}$, therefore, the dispersion index is estimated by Pearson residuals in the following.

$$\hat{\phi} = \sum_{i=1}^{n} \left[\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}} \right]^2 /(n-p).$$

We can also estimate the dispersion index with deviance residual in the following.

$$\hat{\phi}_{\text{deviance}} = \frac{2\left[y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)\right]}{n-p}.$$

Next, we extract the approximated dispersion index using both Pearson and deviance residuals. **Note that the dispersion index in the R base library uses Pearson residuals to approximate the distribution index**.

```
## phd=read.table("w10-ph-data.txt",skip=10, header=TRUE )[,-1] # drop the ID variable
## Regular Poison Model
pois.model = glm(article ~ gender + married + factor(newkids) + prestige + mentor,
                 family = poisson(link="log"), data =phd)
## predicted y: yhat
yhat = pois.model$fitted.values
pearson.resid = (phd$article - yhat)/sqrt(yhat)
Pearson.disp = sum(pearson.resid^2)/pois.model$df.residual
##
Deviance.disp = (pois.model$deviance)/pois.model$df.residual
##
disp = cbind(Pearson.disp = Pearson.disp, Deviance.disp = Deviance.disp)
kable(disp, caption="Dispersion parameter", align = 'c')
```

Table 1: Dispersion parameter

Pearson.disp	Deviance.disp
1.841542	1.805213

The dispersion index under deviance is about 1.80 and 1.84 based on Pearson residual indicating that the Poison assumption is not seriously violated. Therefore, the Poisson model is appropriate. For illustrative purposes, we still adjust the standard error by fitting the quasi-Poisson model in the following (using the estimated dispersion index based on the default deviance).

Next, we summarize the inferential statistics about the regression coefficients in the following table.

```
quasi.model = glm(article ~ gender + married + factor(newkids) + prestige + mentor,
                  family = quasipoisson, data =phd)
summary(quasi.model )
```

```
##
## Call:
## glm(formula = article ~ gender + married + factor(newkids) +
##     prestige + mentor, family = quasipoisson, data = phd)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.45794    0.12904   3.549 0.000407 ***
## genderWomen       -0.21793    0.07425  -2.935 0.003421 **
## marriedSingle     -0.15170    0.08553  -1.774 0.076467 .
## factor(newkids)1 -0.24956    0.08596  -2.903 0.003781 **
## prestige           0.01027    0.03591   0.286 0.774815
## mentor             0.02582    0.00274   9.423  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.841565)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1640.9  on 909  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
SE.quasi.pois = summary(quasi.model)$coef
kable(SE.quasi.pois, caption = "Summary statistics of quasi-poisson regression model")
```

Table 2: Summary statistics of quasi-poisson regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4579404	0.1290407	3.5488055	0.0004068
genderWomen	-0.2179247	0.0742536	-2.9348721	0.0034208
marriedSingle	-0.1516973	0.0855315	-1.7735846	0.0764666
factor(newkids)1	-0.2495633	0.0859576	-2.9033304	0.0037815
prestige	0.0102754	0.0359069	0.2861675	0.7748150
mentor	0.0258173	0.0027397	9.4233384	0.0000000

The estimated dispersion index is 1.84 based on the Pearson residuals.

In the above quasi-Poisson regression, variable prestige is insignificant (p-value = 0.77). The p-value for testing the significance of the variable married is 0.079. We refit the quasi-Poisson model by dropping **prestige** and **married**.

```
quasi.model.02 = glm(article ~ gender + factor(newkids) + mentor,
                     family = quasipoisson, data =phd)
kable(summary(quasi.model.02)$coef, caption = "Inferential statistics of
the Poisson regression coefficients  in the final working model.")
```

Table 3: Inferential statistics of the Poisson regression coefficients in the final working model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4238445	0.0645972	6.561345	0.0000000
genderWomen	-0.2332786	0.0738848	-3.157329	0.0016446
factor(newkids)1	-0.1796153	0.0767849	-2.339200	0.0195404
mentor	0.0257762	0.0026586	9.695432	0.0000000

The above model will be used as the final model. The interpretation of the regression coefficient of the Poisson model is not as straightforward as that in the linear regression models since the response variable in the model is at a log scale.

For example, the coefficient associated with gender is -0.233. This is the estimated Poisson regression coefficient comparing females to males, given the other variables are held constant in the model. The difference in the logs of expected publications is expected to be 0.2332786 units lower for females compared to males while holding the other variables constant in the model. This is still not easy to understand for the general audience.

3.2 Some Visual Comparisons

Next, we make a visualization to show how the explanatory variables in the final working model affect the **actual** number of publications of doctoral students.

To this end, we classify all Ph.D. students into the following four groups defined by **gender** and **status** of having at least one child:

phd.m0 = male and had no child

phd.m1 = male and had at least one child

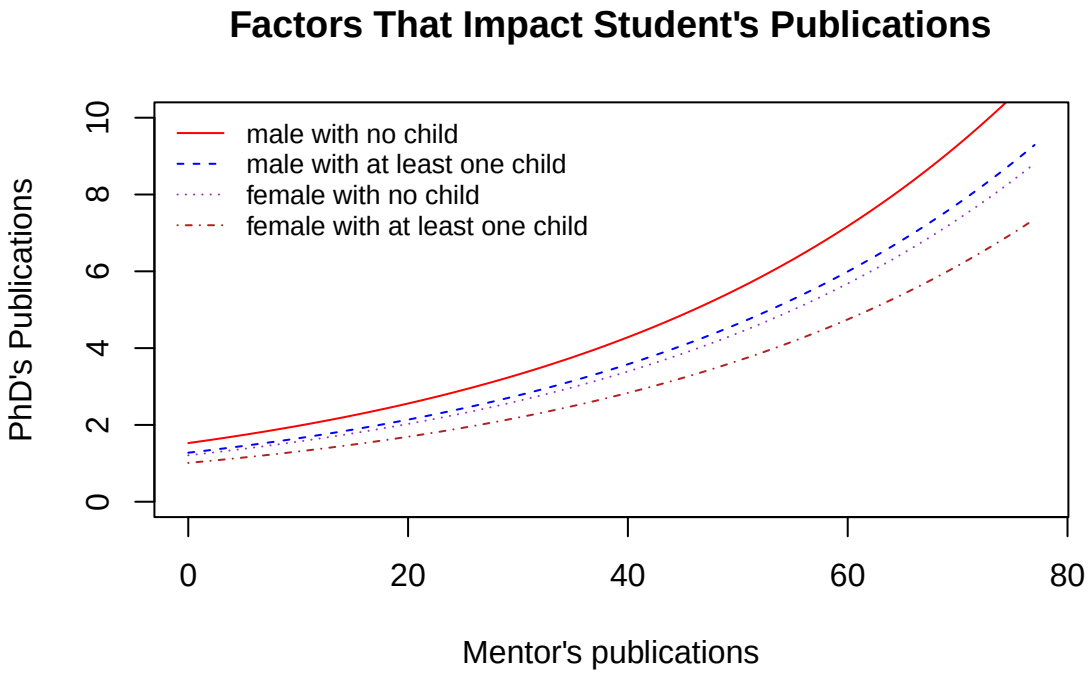
phd.f0 = female and had no child

phd.f1 = female and had at least one child

Next, We exponentiate the log count of publications of Ph.D. students to the actual number of publications and then make graphs to show the relationship between doctoral students and their mentors in terms of the number of publications in each of the groups defined above.

```
mentors =  range(phd$mentor)[1]:range(phd$mentor)[2]
phd.m0 = 0.42384447 + 0.02577624*mentors
phd.m1 = 0.42384447 - 0.17961531 + 0.02577624*mentors
phd.f0 = 0.42384447 -0.23327860 + 0.02577624*mentors
phd.f1 = 0.42384447 -0.23327860 -0.17961531 + 0.02577624*mentors
##
plot(mentors, exp(phd.m0), ylim=c(0,10),
     type = "l",
     col = "red",
```

```
     lty = 1,
     ylab = "PhD's Publications",
     xlab = "Mentor's publications",
     main = "Factors That Impact Student's Publications")
lines(mentors, exp(phd.m1), col = "blue", lty = 2)
lines(mentors, exp(phd.f0), col = "darkorchid", lty = 3)
lines(mentors, exp(phd.f1), col = "firebrick", lty = 4)
legend("topleft", c("male with no child", "male with at least one child",
                    "female with no child", "female with at least one child"),
       col=c("red", "blue", "darkorchid", "firebrick"),  lty=1:4, bty="n", cex=0.8)
```



We can see the relationship between the number of publications of doctoral students and other factors.

1. the number of publications of doctoral students is positively associated with their mentor publication.
2. Male doctoral students with no kids published more articles than those who had at least one kid. Female doctoral students also have the same pattern.
3. Overall, male doctoral students published more than female students.

4 Conclusions

The Poisson regression model is used for modeling counts/rates-based data sets. If the model is appropriate, its results are explainable and comparable and backed by statistical theory.

If Poisson regression is not appropriate, we can consider other models depending on the situation. The complex alternatives to the Poisson regression model that can be considered are negative binomial regression, zero-inflated regression models, random-forest-based regression models, and neural-network-based regression models. The last two models are "black-box" models because of the lack of interpretability.