MSc Generalised Linear Models, Assessed Practical

Practical ID: P453

December 7, 2021

Contents

1 Introduction

1.1 Data Description

The data relate to the number of publications by PhD students in biochemistry during the last three years of their PhD. For each student, the available variables describe the number of articles published during the last three years of their PhD ('articles'), whether they were female or not ('female'), whether they were married or not ('married'), the number of children less than six years old that they have ('kids'), the prestige score of the graduate program ('prestige') and the number of articles published by the student's supervisor ('mentor'). Note that articles and mentor take integer values, while female and married are binary indicator variables. Kids is treated as a categorical variable and prestige is treated as a continuous variable. Interest lies in determining how the number of publications by students depends on the other variables.

1.2 Exploratory Analysis

The summary statistics of the articles, prestige and mentor variables are given in Table 1. Both the number of articles published by the student and by their supervisor has positive skew, with most of the data concentrated in low values and a small number in very large values. As expected, the number of articles published by supervisors was on average much higher than the number of articles published by students. The prestige score ranges from 0.76 to 4.62, with a mean of 3.10.

	Articles	**Prestige**	**Mentor**
Minimum	0.00	0.76	0.00
1st Quartile	0.00	2.26	3.00
Median	1.00	3.15	6.00
Mean	1.69	3.10	8.77
3rd Quartile	2.00	3.92	12.00
Maximum	19.00	4.62	77.00

Table 1: Summary statistics of articles, prestige and mentor

The split between female and non-female students is relatively even; 54.0% and 46.0% respectively as shown in Table 2. Approximately two-thirds of the students were married.

	Female	**Married**
No	494	309
Yes	421	606

Table 2: Summary statistics of female and married

The number of children that students had varied between zero and three as shown in Table 3, with 65.5% having none, 21.3% having one, 11.5% having two and the remaining 1.7% having three.

Kids	**Frequency**
0	599
1	195
2	105
3	16

Table 3: Summary statistics of kids

The distribution of the number of articles is shown in Figure 1. As noted in the numerical summaries, the distribution is positively skewed, with 76.4% of students publishing less than three articles and just 4.1% publishing more than five.

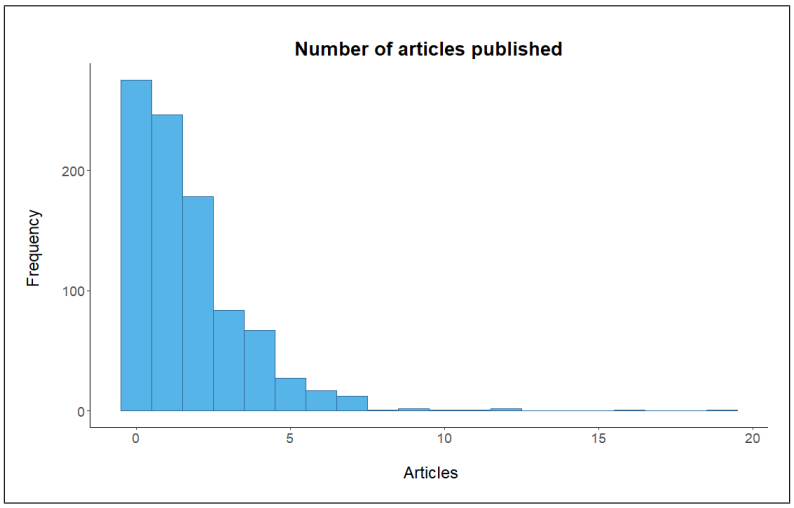

Figure 1: Histogram of articles

The relationships between the number of articles published by students and the variables mentor and prestige are shown in Figure 2. The relationships are coloured by the sex of the student, which highlights that the students with very large amounts of articles published are generally male. The number of articles published by a student and the number of articles published by their supervisor are moderately positively related, with correlation coefficient equal to 0.31. Most of the students are concentrated around low values of number of articles published and number of supervisor articles published, however there are some extreme values on the far right and in the upper section of the plot. The relationship between prestige and the number of articles published is less clear, with a correlation coefficient equal to 0.07.

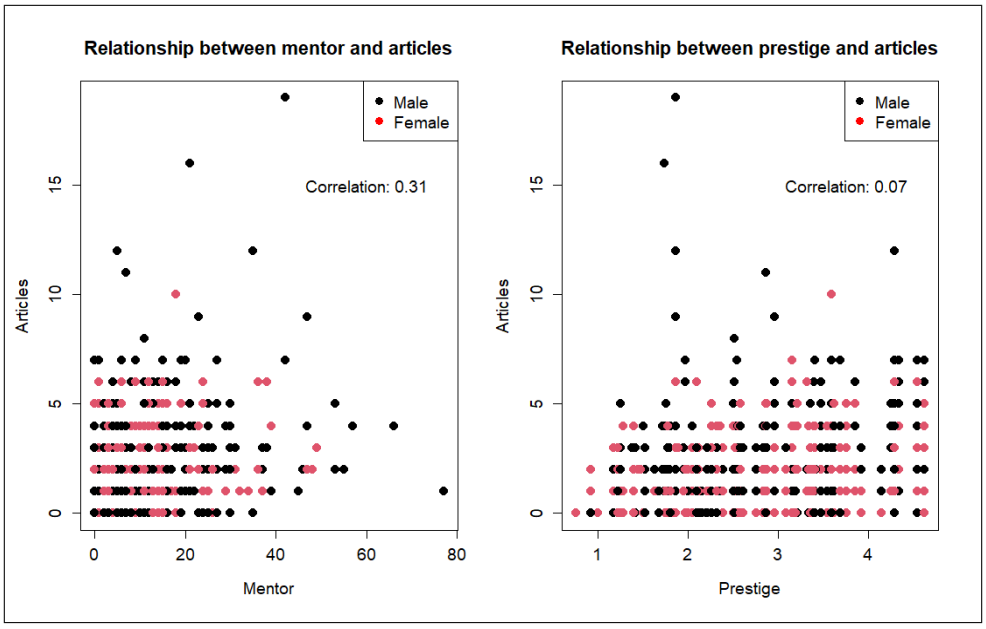

Figure 2: Scatterplots of prestige and mentor

The relationship between the number of articles published and each of the remaining variables, female, kids and married, is shown in Figure 3. Male students have both a larger mean and variance than female students in terms of number of articles published, 1.88 versus 1.47 and 4.75 versus 2.41 respectively. Although the distribution of students with zero, one, and two children are similar with regards to the number of articles published, there are fewer large values of articles published in the groups with more children. The distribution of students with three children overlaps with that of the other students, but with lower third quartiles and maximum values. Married students have both a larger mean and variance of number of articles published compared to non-married students, 1.74 versus 1.59 and 4.08 versus 2.99 respectively. There were also more large values of articles published for married students than for non-married students.

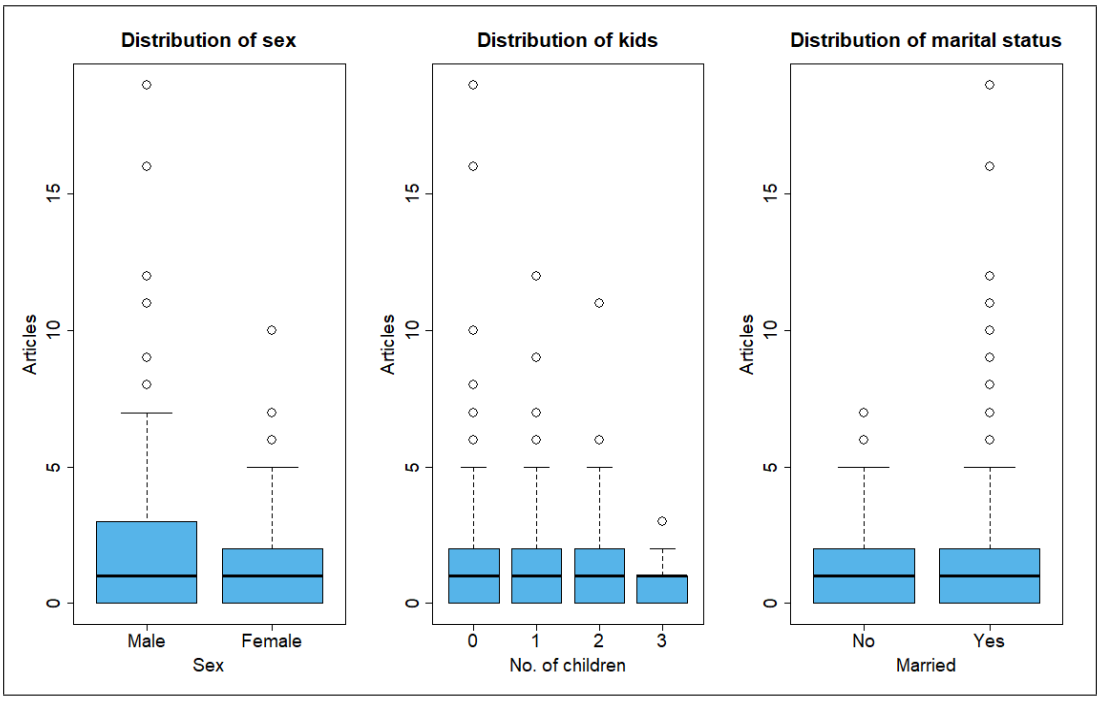

Figure 3: Boxplots of female, kids and married

2 Methods and Results

2.1 GLM Estimation

The relation between the number of publications and the other variables was modelled using the Poisson GLM with canonical link function:

$$\lambda = \exp(x\prime\beta)$$

The first model considered was of the form:

glm(articles ~ female + married + kids + prestige + mentor + female*married + female*kids + female*prestige + female*mentor, family = poisson) (1)

Model selection was performed on this model using both forward and backward stepwise regression based on the Akaike Information Criterion ('AIC'). The model chosen by this selection criteria was of the form:

glm(articles ~ female + married + kids + prestige + mentor + female*prestige, family = poisson) (2)

A likelihood ratio test was carried out to test whether the additional parameters in model (1) compared to model (2) were significantly different from zero. The degrees of freedom were equal to the difference in the number of coefficients present in each model, here equal to five. Formally,

$$H_0 : \hat{\beta}_{female*married} = \hat{\beta}_{female*kids} = \hat{\beta}_{female*mentor} = 0$$

A p-value of 0.874 was obtained and, therefore, the null hypothesis was not rejected at a significance level of 0.05. Accordingly, it was concluded that model (2) was preferable to model (1).

It is important to note that the prestige term in the summary of model (2) was not significant according to the Wald test at a significance level of 0.05, but the interaction term female*prestige was significant. Due to the principal of marginality, it was not proposed to exclude the main effect of prestige while keeping the female*prestige term in the model; either both variables should remain in the model or they both should be removed. This then motivates a comparison of model (2) with a simpler model of the form:

glm(articles ~ female + married + kids + mentor, family = poisson) (3)

A likelihood ratio test was carried out to test whether the additional parameters in model (2) compared to model (3) were significantly different from zero, with degrees of freedom equal to two. Formally,

$$H_0 : \hat{\beta}_{prestige} = \hat{\beta}_{female*prestige} = 0$$

A p-value of 0.053 was obtained for this test. At a significance level of 0.05, the null hypothesis is not rejected, and the test implies model (3) is preferable to model (2). However, the p-value is very close to the significance level, and as a result, it is not obvious which model is best.

In terms of assessing model fit, model (2) had a marginally better AIC value compared to model (3), 3313 versus 3315. Model (2) had a slightly larger R^2_{KL} measure based on the Kullback-Leibler

divergence than model (3), meaning that model (2) had a greater fraction of uncertainty explained in its fit than model (3). However, it is worth noting that neither model had a particularly large value of this measure, indicating that there is a lot of heterogeneity present in both models. Model (3) is more parsimonious than model (2) and arguably more interpretable due to the lack of an interaction effect.

In the leverage and cook's distance plots for both models, there were many points beyond the $\frac{8}{n-2p}$ cut-off, which is typically used to identify observations with large influence on the model fit. As a test of robustness of the models, different numbers of influential points were removed, starting from the most influential, and the models were refitted to check whether coefficients changed considerably. It took just four students to be removed from the data for model (2) to suggest that the terms prestige and female*prestige were not significant. Model (3) was more stable when influential observations were removed. Therefore, it was concluded that model (3) was preferred to model (2) based on the result of the Likelihood Ratio Test and since it was more robust to the exclusion of highly influential observations. Table 4 summarises the comparison of the two models.

Model	**Likelihood Ratio Test**	**AIC**	$\boldsymbol{R^2_{KL}}$	**Robustness**
2	✗	3313	0.104	✗
3	✓	3315	0.101	✓

Table 4: Comparison of model (2) and model (3)

The diagnostic plots of model (3) are shown in Figure 4. Note that the increase on the right of the cook's distance plot arises due to students being ordered according to the number of articles they published.

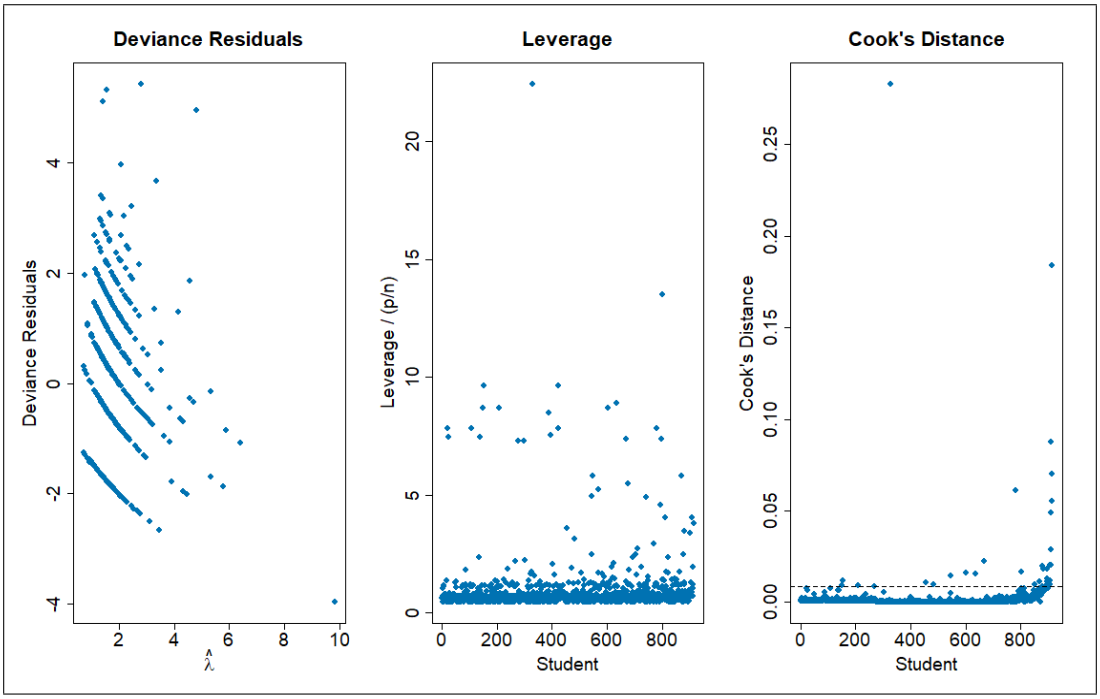

Figure 4: Diagnostic plots of model (3)

2.2 Interpretation

A summary of the final model (3) is given in Table 5. Note that the highly influential students discussed in the previous section were not removed as there is no obvious reason for their exclusion.

Coefficient	Estimate	Standard Error	p-value	exp(estimate)	Confidence Interval
Intercept	0.3477	0.0601	≈0	1.4157	1.2583 - 1.5928
Female (Yes)	-0.2260	0.0547	≈0	0.7977	0.7166 - 0.8880
Married (Yes)	0.1481	0.0628	0.0183	1.1597	1.0254 - 1.3115
Kids (One)	-0.1803	0.0706	0.0107	0.8350	0.7271 - 0.9590
Kids (Two)	-0.3278	0.0910	0.0003	0.7205	0.6029 - 0.8611
Kids (Three)	-0.8215	0.2817	0.0035	0.4398	0.2532 - 0.7638
Mentor	0.0256	0.0020	≈0	1.0259	1.0220 - 1.0299

Table 5: Summary of model (3)

The exponential of the estimate of the intercept term corresponds to the expected number of articles published by a male, non-married student, with no children, whose supervisor published no articles, which is equal to 1.42 articles. This value has 95% confidence interval (1.26 – 1.59).

For the binary indicator female, the estimated coefficient is -0.23 with standard error equal to 0.05. The exponential of this estimate is 0.80; the expected number of articles published by students increases by a multiplicative factor of 0.80 when the student is female compared to male. As this value is less than one, the fitted model suggests that the expected number of articles published decreases when students are female as opposed to male. The 95% confidence interval for this value is (0.72, 0.89). The interpretation of the binary indicator married is similar. The further the exponential of the estimate is from one, the larger the estimated effect size in either an increasing or decreasing direction, depending on whether the value is greater than or less than one.

For the variable kids, the effects are relative to having no children. The exponentiated estimate of the number of articles published by a supervisor is 1.03, which means that the expected number of articles published by a student increases by a multiplicative factor of this value for a one article increase in the number of publications by a supervisor. As an example, the estimated number of articles published for a student who is female and married with two kids and whose supervisor published eight articles is:

$$= e^{\hat{\beta}_{intercept}+\hat{\beta}_{female}+\hat{\beta}_{married}+\hat{\beta}_{kids,2}+8\times\hat{\beta}_{mentor}}$$

$$= e^{0.3477-0.2260+0.1481-0.3278+8\times0.0256} = 1.16$$

2.3 Dispersion Parameter

The dispersion parameter can be calculated using:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{V(\hat{\lambda}_i)}$$

The calculated dispersion parameter for model (3) is 1.82. Since this is greater than one, the model is overdispersed and, as a result, the standard errors reported in Table 5 are too small. Table 6 shows a summary of model (3) with dispersion parameter equal to 1.82, which highlights the increase in standard errors, the decrease in the significance of the coefficients and the wider confidence intervals. Although Married (Yes) and Kids (One) are not significant, their base levels of Married (No) and Kids (None) are absorbed in the intercept which is still highly significant. It is worth noting that model (2) was similarly overdispersed, and when the dispersion parameter was allowed to be equal to its calculated result, the terms prestige and female*prestige were no longer significant. To adjust for overdispersion, the dispersion parameter could be estimated within the model using the quasi-families in R.

Coefficient	**Estimate**	**Standard Error**	**p-value**	**exp(estimate)**	**Confidence Interval**
Intercept	0.3477	0.0812	≈0	1.4157	1.2074 - 1.6600
Female (Yes)	-0.2260	0.0739	0.0022	0.7977	0.6902 - 0.9220
Married (Yes)	0.1481	0.0848	0.0807	1.1597	0.9821 - 1.3694
Kids (One)	-0.1803	0.0954	0.0587	0.8350	0.6927 - 1.0066
Kids (Two)	-0.3278	0.1228	0.0076	0.7205	0.5664 - 0.9167
Kids (Three)	-0.8215	0.3804	0.0308	0.4398	0.2086 - 0.9269
Mentor	0.0256	0.0027	≈0	1.0259	1.0206 - 1.0313

Table 6: Summary of model (3) with $\hat{\phi} = 1.82$

3 Conclusions

The final model (3) had many observations with large values of influence and of cook's distance. For instance, student 328 had a particularly large influence on the fit due to them publishing one article while their supervisor published 77 articles. Observations like this may have been data entry errors. The model had a relatively low value of R^2_{KL} and it was overdispersed, which indicated that for similar values of the predictor variables, there were quite different numbers of articles published by students.

The model was still useful in identifying the relationship between the number of articles published and the various predictor variables, and the strength of these relationships. For instance, it is informative that the prestige of the graduate program was not very important in predicting the number of articles published, and that being a male and married student with less kids and a supervisor who publishes more is generally associated with a higher number of articles published by students. Model (2) could also have been chosen based on the AIC or the R^2_{KL}, especially given that the Likelihood Ratio Test had a p-value very close to the significance level of 0.05.

The model might be a better fit with other interaction terms included, by allowing the dispersion parameter to be estimated or by using the square root link function instead of the canonical log link function. It could also be improved by fitting a different type of model or by gathering more predictor variables that relate to the number of articles published. These might include the average GPA of the student at undergraduate or master's level, their age, or whether they are self-funded or on a scholarship, though these variables may be correlated with those already used.

Appendix

```
# MSc, Generalised Linear Models, Assessed Practical

# Section 2: Assessed Exercise

setwd('C:/Users/james/OneDrive/Documents/Oxford/Michaelmas/Applied Statistics')

#
### Q1.
#

set.seed(453)
pub <- read.csv("pub.csv")
dim(pub)
names(pub)
head(pub)

# Define factors:
pub$kids <- factor(pub$kids, levels = c("0", "1", "2", "3"))
pub$female <- factor(pub$female, levels = c("0", "1"))
levels(pub$female) <- c(levels(pub$female), "Female")
levels(pub$female) <- c(levels(pub$female), "Male")
pub$female[pub$female == "1"] <- "Female"
pub$female[pub$female == "0"] <- "Male"
pub$female <- factor(pub$female, levels = c("Male", "Female"))

pub$married <- factor(pub$married, levels = c("0", "1"))
levels(pub$married) <- c(levels(pub$married), "Yes")
levels(pub$married) <- c(levels(pub$married), "No")
pub$married[pub$married == "1"] <- "Yes"
pub$married[pub$married == "0"] <- "No"
pub$married <- factor(pub$married, levels = c("No", "Yes"))

# Numerical summaries:
summary(pub)

# Graphical summaries:
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
               "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
library(ggplot2)
library(grid)
library(GGally)

# Histogram of articles:
ggplot(pub, aes(x = articles, color = 0)) +
  geom_histogram(fill = cbPalette[3], position = "dodge", binwidth = 1) +
  ggtitle("Number of articles published") +
```

```
  xlab("\n Articles") + ylab("Frequency \n") +
  theme_classic() +
  theme(text = element_text(size = 18),
        plot.title = element_text(hjust = 0.5, size = 20, face = "bold")) +
  theme(legend.position = 'none',
        plot.margin = unit(c(2, 1, 1, 1), "lines"))

# Scatterplots:
par(mfrow=c(1,2))
plot(pub$mentor, (pub$articles), col = pub$female, pch = 16,
     cex = 1.4,
     cex.main = 1.4,
     cex.axis = 1.2,
     cex.lab = 1.2,
     main = "Relationship between mentor and articles",
     xlab = "Mentor",
     ylab = "Articles")
text(paste("Correlation:", round(cor(pub$mentor, pub$articles), 2)), x = 60, y = 15,
     cex = 1.2)
legend("topright", legend = c("Male", "Female"), col = c("black", "red"), cex = 1.2, pch = 16)
plot(pub$prestige, pub$articles, col = pub$female, pch = 16,
     cex = 1.4,
     cex.main = 1.4,
     cex.axis = 1.2,
     cex.lab = 1.2,
     main = "Relationship between prestige and articles",
     xlab = "Prestige",
     ylab = "Articles")
text(paste("Correlation:", round(cor(pub$prestige, pub$articles), 2)), x = 3.75, y = 15,
     cex = 1.2)
legend("topright", legend = c("Male", "Female"), col = c("black", "red"), cex = 1.2, pch = 16)

# Histograms of mentor and prestige:
hist(pub$mentor, breaks = 25)
hist(pub$prestige, breaks = 25)

# Boxplots of categorical variables:
par(mfrow = c(1,3))
par(mar = c(5,5,4,2)+0.1)
boxplot(articles ~ female, main = "Distribution of sex", ylab = "Articles",
        xlab = "Sex",col = cbPalette[3],
        cex = 2,
        cex.main = 2,
        cex.axis = 1.8,
        cex.lab = 1.8)
boxplot(articles ~ kids, main = "Distribution of kids", ylab = "Articles",
        xlab = "No. of children", col = cbPalette[3],
        cex = 2,
        cex.main = 2,
```

```
        cex.axis = 1.8,
        cex.lab = 1.8)
boxplot(articles ~ married, main = "Distribution of marital status", ylab = "Articles",
        xlab = "Married", col = cbPalette[3],
        cex = 2,
        cex.main = 2,
        cex.axis = 1.8,
        cex.lab = 1.8)
tapply(articles, female, var) # tapply to get mean, var, etc.
tapply(articles, kids, var)
tapply(articles, married, var)
```

```
#
### Q2.
#

# Defining a model with all possible variables and interactions:
pub.glm <- glm(formula = articles ~ female + married + kids + prestige + mentor +
                female:married + female:kids + female:prestige + female:mentor,
               data = pub, family = poisson)
summary(pub.glm)
step(pub.glm, direction = "both") # Stepwise Regression

# Model chosen by AIC:
pub.glm2 <- glm(formula = articles ~ female + married + kids + prestige + mentor +
                   female:prestige, data = pub, family = poisson)
summary(pub.glm2) # suggests prestige is not significant

# Model without prestige and female:prestige:
pub.glm3 <- glm(formula = articles ~ female + married + kids + mentor, data = pub, family = poisson)
summary(pub.glm3)

# Likelihood ratio tests. pub.glm vs pub.glm2:
dof <- pub.glm$rank - pub.glm2$rank
lrt <- deviance(pub.glm2) - deviance(pub.glm)
pval <- 1 - pchisq(lrt, dof)
cbind(lrt, dof, pval) # fail to reject

# pub.glm2 vs pub.glm3:
dof2 <- pub.glm2$rank - pub.glm3$rank
lrt2 <- deviance(pub.glm3) - deviance(pub.glm2)
pval2 <- 1 - pchisq(lrt2, dof2)
cbind(lrt2, dof2, pval2) # fail to reject but very close
```

```
#
### Q3.
#

# Note that pub.glm2 was also analysed in Q3, Q4 and Q5 using the same code.
# Cameron and Windmeijer (1997) KL/Deviance based R^2:
library(rsq)
rsq.kl(pub.glm3)

# Diagnostics:
par(mfrow=c(1,3))
plot(fitted(pub.glm3), rstandard(pub.glm3),
     xlab = expression(hat(lambda)), ylab = "Deviance Residuals",
     main = "Deviance Residuals",
     pch = 16, col = cbPalette[6],
     cex = 1.2,
     cex.main = 2,
     cex.axis = 1.8,
     cex.lab = 1.8)
p <- pub.glm3$rank
n <- nrow(model.frame(pub.glm3))
plot(influence(pub.glm3)$hat/(p/n), ylab='Leverage / (p/n)', xlab = "Student",
     pch = 16, col = cbPalette[6], main = "Leverage",
     cex = 1.2,
     cex.main = 2,
     cex.axis = 1.8,
     cex.lab = 1.8)
#text(influence(pub.glm3)$hat/(p/n), labels = rownames(pub), ylab='Leverage / (p/n)')
plot(cooks.distance(pub.glm3), ylab = "Cook's Distance", xlab = "Student",
     pch = 16, col = cbPalette[6], main = "Cook's Distance",
     cex = 1.2,
     cex.main = 2,
     cex.axis = 1.8,
     cex.lab = 1.8)
abline(h = 8/(n-2*p),lty = 2)
#text(cooks.distance(pub.glm3), labels = rownames(pub))

# Removing influential points sequentially and refitting:
pub.out <- pub[-c(328, 915, 901, 914),]
pub.glm.out <- glm(formula = articles ~ female + married + kids + prestige + mentor +
                female:married + female:kids + female:prestige + female:mentor,
                data = pub.out, family = poisson)
step(pub.glm.out, direction = "both")
pub.glm.out2 <- glm(formula = articles ~ female + married + kids + prestige + mentor +
                      female:prestige, data = pub.out, family = poisson)
summary(pub.glm.out2) # prestige and female:prestige not significant
pub.glm.out3 <- glm(formula = articles ~ female + married + kids + mentor,
                    data = pub.out, family = poisson)
summary(pub.glm.out3)
```

```
#
### Q4.
#

summary(pub.glm3)

# Confidence Intervals:
options(digits = 4)
confint.default(pub.glm3)
exp(confint.default(pub.glm3))



#
### Q5.
#

dp = sum(residuals(pub.glm3, type ="pearson") ^ 2)/pub.glm3$df.residual
dp

library(AER)
dispersiontest(pub.glm3)

summary(pub.glm3, dispersion = dp)
exp(summary(pub.glm3, dispersion = dp)$coef[1:7])

# Confidence Intervals:
dp.beta17 <- summary(pub.glm3, dispersion = dp)$coef[1:7, 1]
dp.se17 <- summary(pub.glm3, dispersion = dp)$coef[1:7, 2]
dp.cval <- qnorm(0.975)
dp.lower <- dp.beta17 - dp.cval*dp.se17
dp.upper <- dp.beta17 + dp.cval*dp.se17
dp.ci95 <- cbind(dp.lower, dp.upper)
dp.ci95
exp(dp.ci95)
```