Week 8 GLM Practical Notes

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.0.5

##
## Attaching package: 'patchwork'

## The following object is masked from 'package:MASS':
##
##     area
```

```
library(rsq)
```

```
## Warning: package 'rsq' was built under R version 4.0.5
```

```
pub <- read.csv("Data/pub.csv")
# Setting female and married as factors
pub <- pub %>%
  mutate(across(2:3, as.factor))
```

Exploratory analysis

Summary

First, a basic summary of the data. We look for any possible indication of data errors.

```
summary(pub)
```

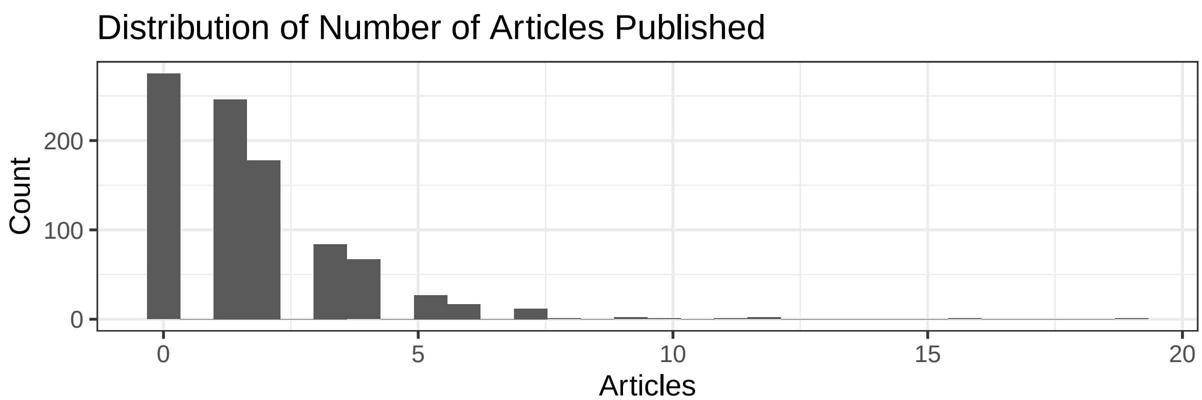
```
##     articles      female  married      kids          prestige
##  Min.   : 0.000   0:494   0:309   Min.   :0.0000   Min.   :0.755
##  1st Qu.: 0.000   1:421   1:606   1st Qu.:0.0000   1st Qu.:2.260
##  Median : 1.000                   Median :0.0000   Median :3.150
##  Mean   : 1.693                   Mean   :0.4951   Mean   :3.103
##  3rd Qu.: 2.000                   3rd Qu.:1.0000   3rd Qu.:3.920
##  Max.   :19.000                   Max.   :3.0000   Max.   :4.620
##      mentor
##  Min.   : 0.000
##  1st Qu.: 3.000
##  Median : 6.000
##  Mean   : 8.767
##  3rd Qu.:12.000
##  Max.   :77.000
```

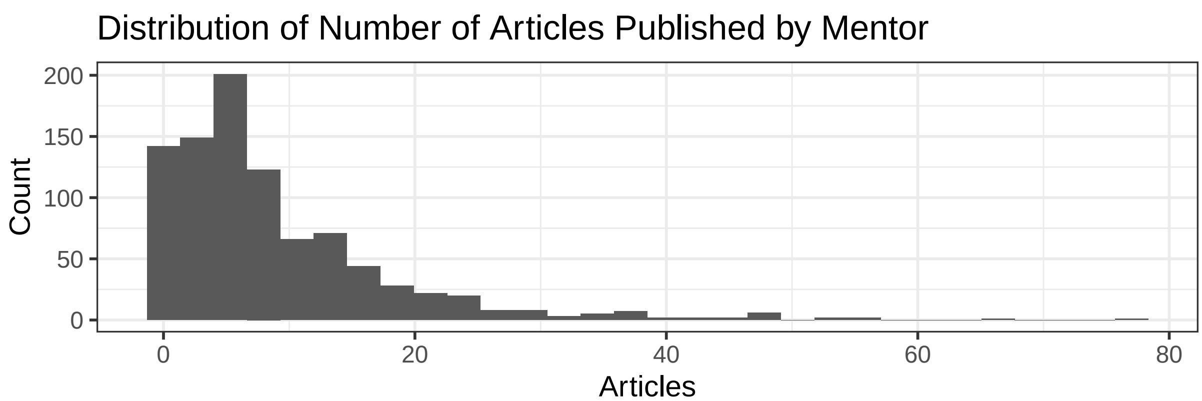
There are some values that appear to be quite high for mentor and articles, but when we look at the overall distribution we see that it is very skewed and so these values are probably not a cause for concern at the moment.

```
articles_histogram <- pub %>%
  ggplot(aes(x = articles)) +
  geom_histogram() +
  theme_bw() +
  labs(title = "Distribution of Number of Articles Published",
       x = "Articles",
       y = "Count")

mentor_histogram<- pub %>%
  ggplot(aes(x = mentor)) +
  geom_histogram() +
  theme_bw() +
  labs(title = "Distribution of Number of Articles Published by Mentor",
       x = "Articles",
```

```r
       y = "Count")

articles_histogram / mentor_histogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

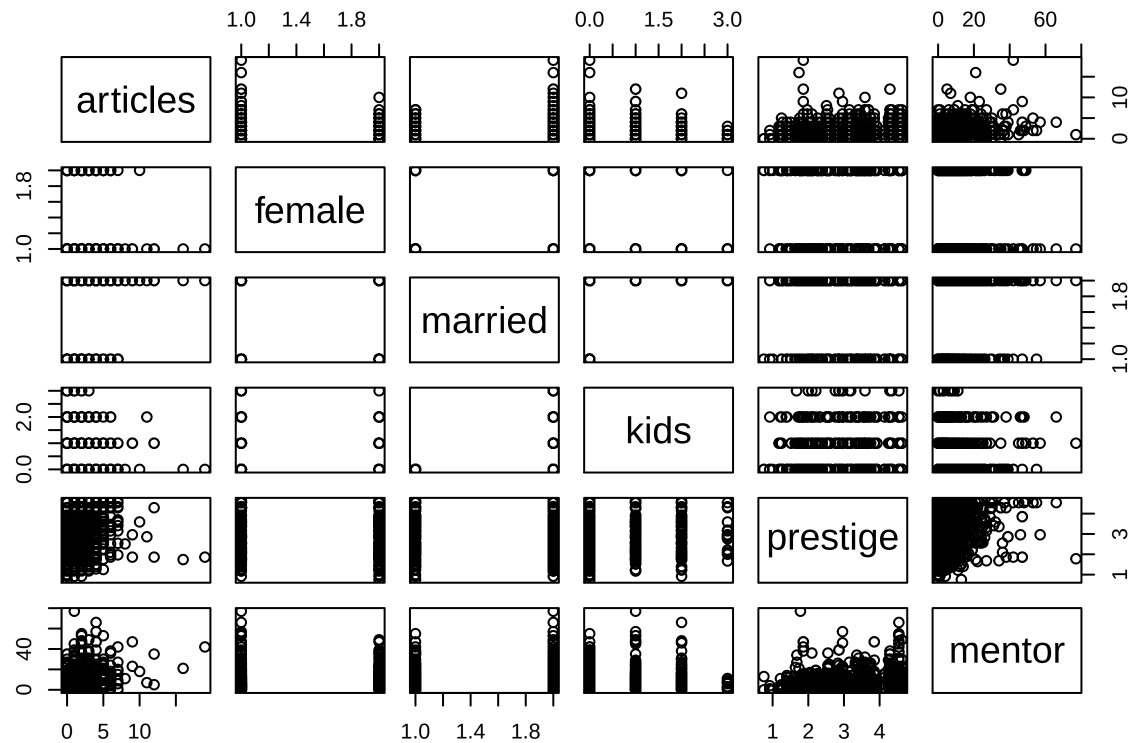




Pairs plot

A quick look at all the possible associations between the variables:

```r
pairs(pub)
```

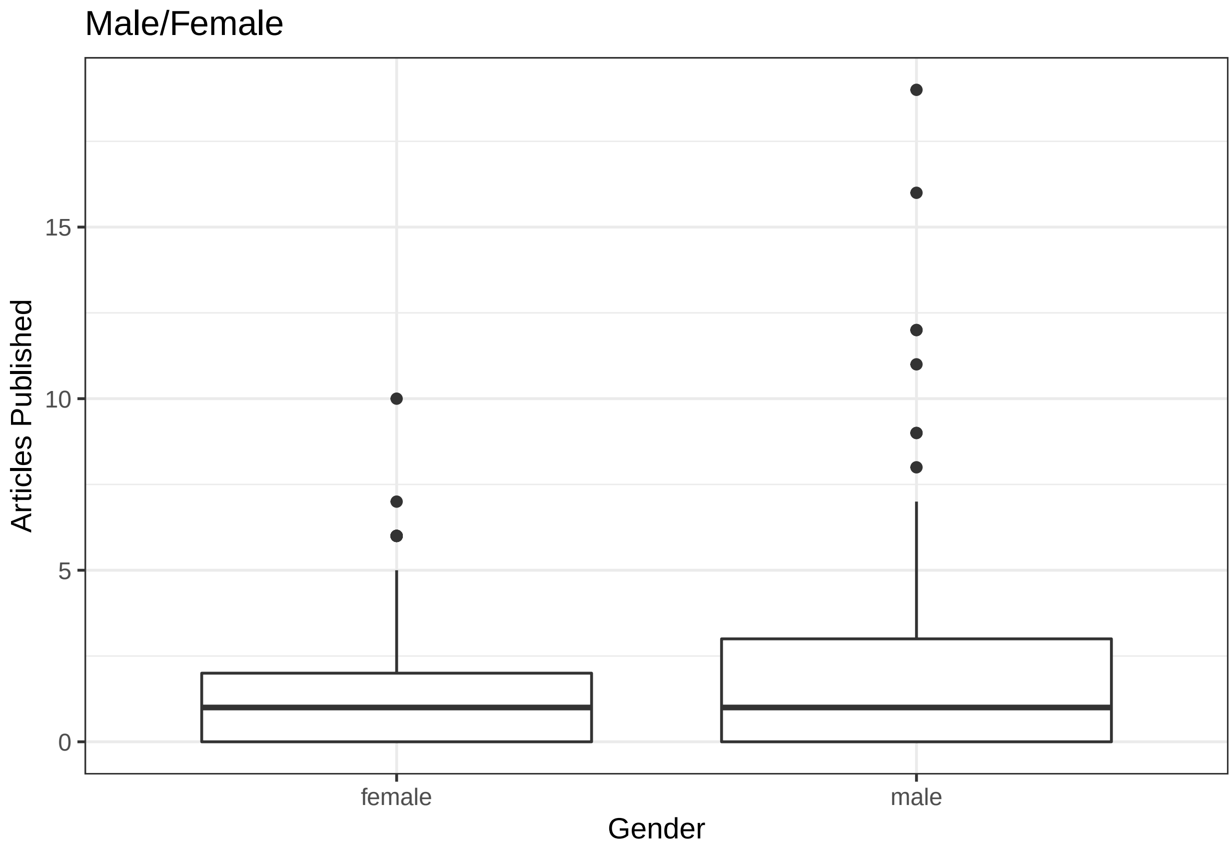


There seems to be more articles/wider spread for males than females. More articles for married than unmarried. Fewer kids more articles.There seems to be some positive relationship between prestige and articles but not extremely strong. The relationship between mentor and articles is somewhat unclear.

There are obvious relationships between some of the explanatory variables. Marriage means higher numbers of kids.The prestige of the program is positively correlated with the output of the mentor.

Boxplots

```
female_boxplot <- pub %>%
  # Changing the labelling of the factors for the plot
  mutate(female = ifelse(female == 0, "male", "female")) %>%
  ggplot() +
  geom_boxplot(aes(x = female, y = articles)) +
  theme_bw() +
  labs(title = "Male/Female",
       x = "Gender",
       y = "Articles Published")

female_boxplot
```



We see that the median number of publications is similar for men and women but that there is a larger dispersion for the number of articles produced by men.

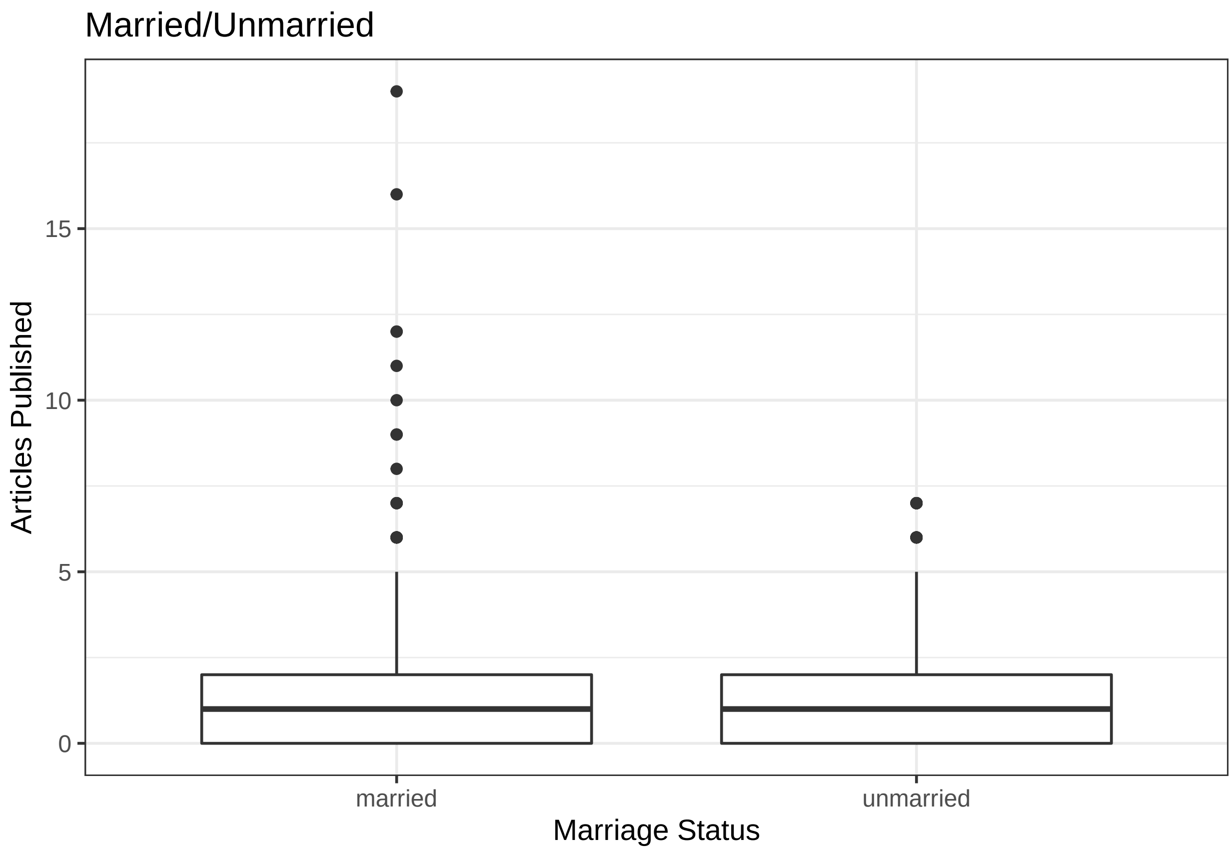
```
mean_articles_f <- pub %>%
  group_by(female) %>%
  summarise(mean = mean(articles)) %>%
  mutate(mean = signif(mean,3))
mean_articles_f
```

```
## # A tibble: 2 x 2
##   female  mean
##   <fct>  <dbl>
## 1 0       1.88
## 2 1       1.47
```

The mean number of articles is lower for women however.

```
married_boxplot <- pub %>%
  # Changing the labelling of the factors for the plot:
  mutate(married = ifelse(married == 1,"married","unmarried")) %>%
  ggplot() +
  geom_boxplot(aes(x = married, y = articles)) +
  theme_bw() +
  labs(title = "Married/Unmarried",
       x = "Marriage Status",
```

```
       y = "Articles Published")

married_boxplot
```



Larger spread for people who are married but again similar median.

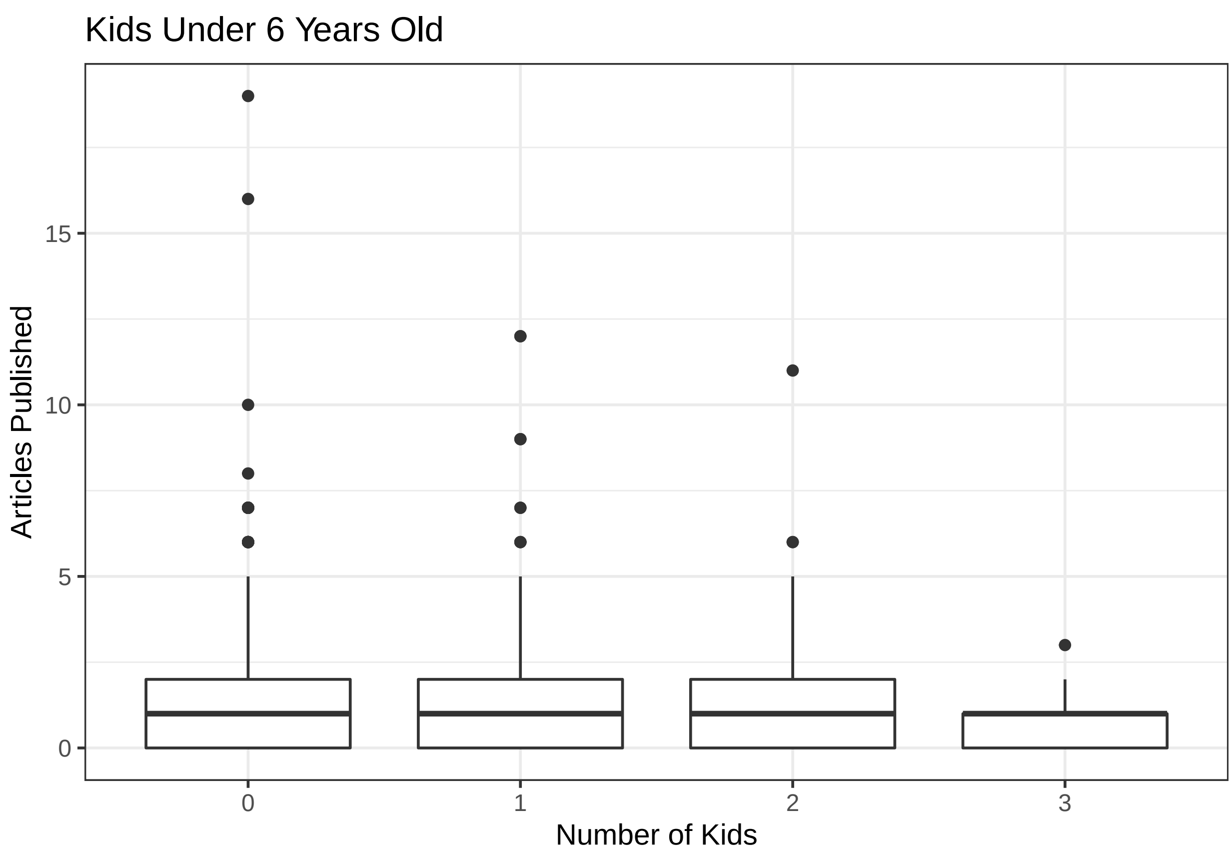
```
mean_articles_married <- pub %>%
  group_by(married) %>%
  summarise(mean = mean(articles)) %>%
  mutate(mean = signif(mean,3))

mean_articles_married
```

```
## # A tibble: 2 x 2
##   married  mean
##   <fct>   <dbl>
## 1 0        1.59
## 2 1        1.74
```

Married people have a higher mean output, but the difference is smaller.

```
kids_boxplot <- pub %>%
  ggplot() +
  geom_boxplot(aes(x = as.factor(kids), y = articles)) +
```

```r
  theme_bw() +
  labs(title = "Kids Under 6 Years Old",
       x = "Number of Kids",
       y = "Articles Published")

kids_boxplot
```

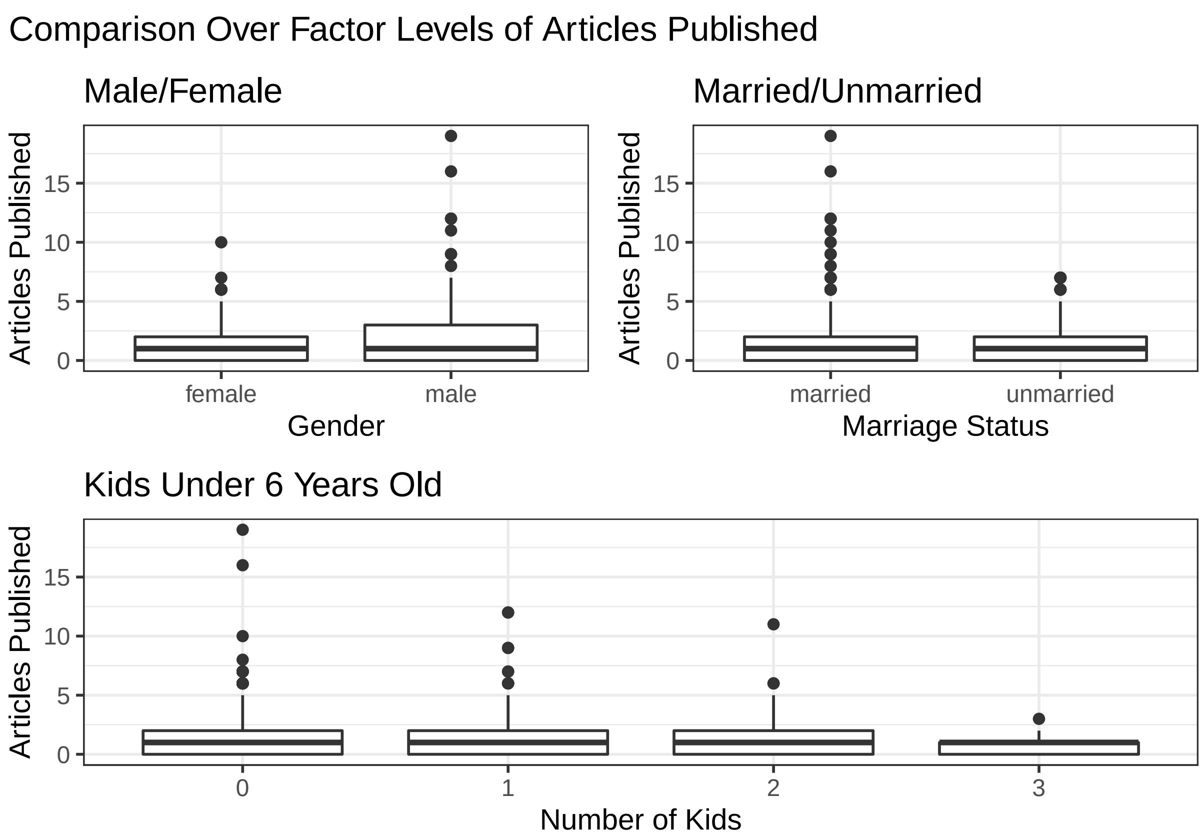


Larger spread of articles for individuals with fewer children.

```r
mean_articles_kids <- pub %>%
  group_by(kids) %>%
  summarise(mean = mean(articles)) %>%
  mutate(mean = signif(mean,3))

mean_articles_kids
```

```
## # A tibble: 4 x 2
##    kids  mean
##   <int> <dbl>
## 1     0 1.72
## 2     1 1.76
## 3     2 1.54
## 4     3 0.812
```

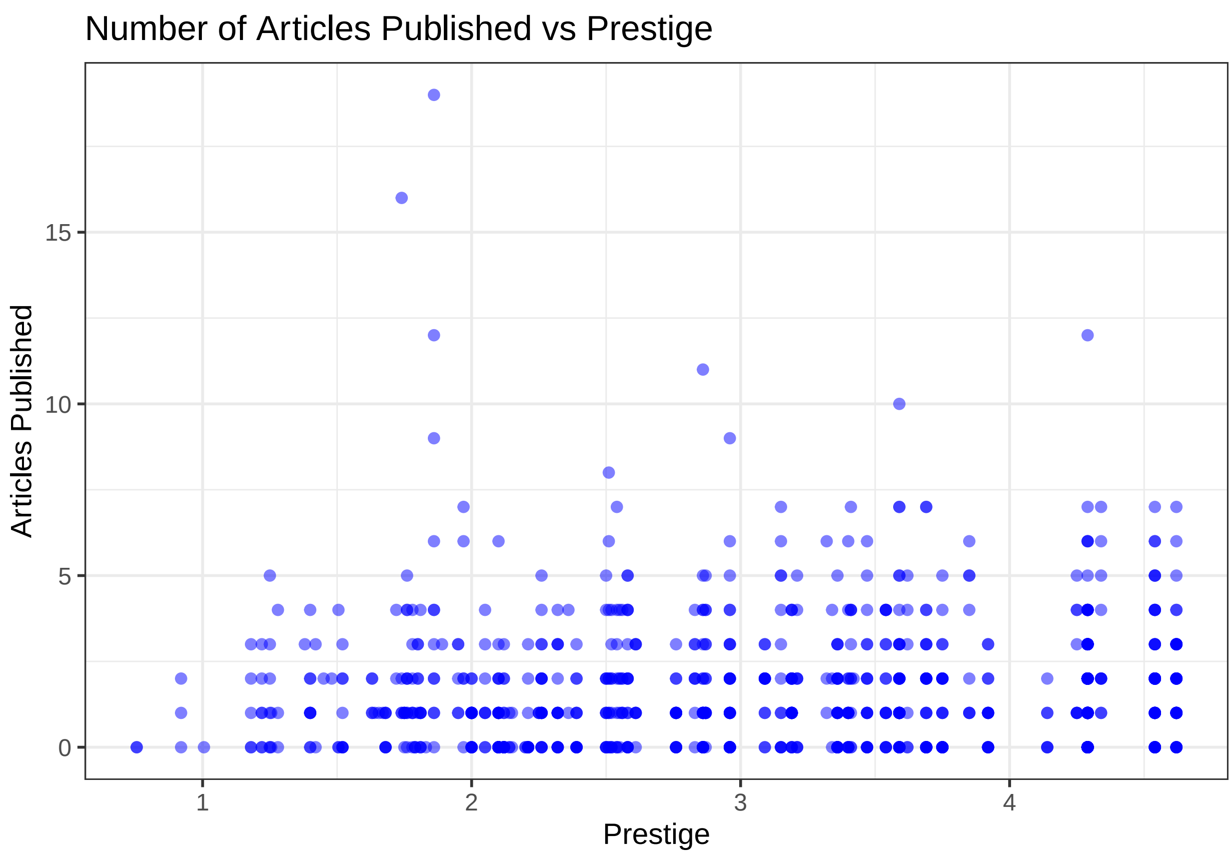
```
(female_boxplot + married_boxplot) / kids_boxplot +
  plot_annotation(
    title = "Comparison Over Factor Levels of Articles Published"
  )
```



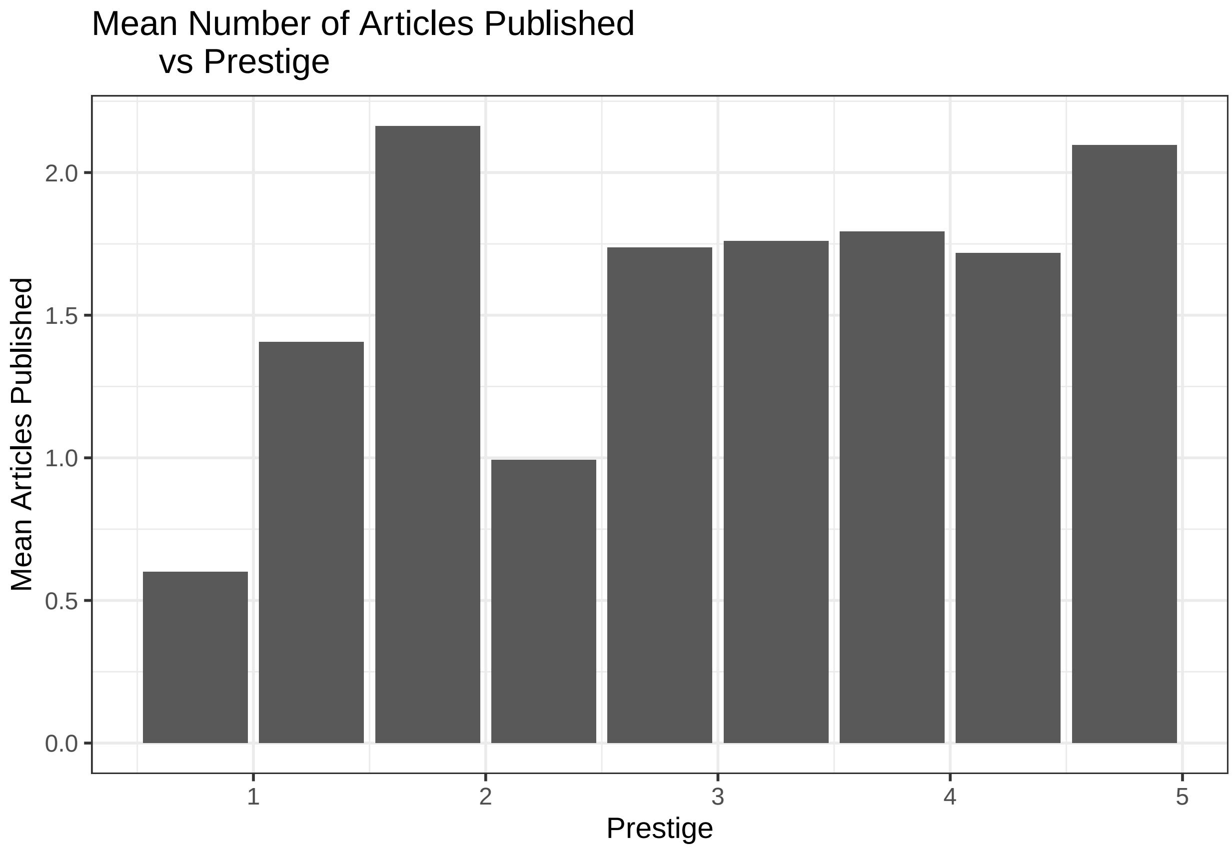
Highest output from individuals with one child, then the number of articles drops off. Since the relationship between number of kids and articles is not monotonic, it is probably best to treat the number of children as a factor.

```
pub$kids <- as.factor(pub$kids)
```

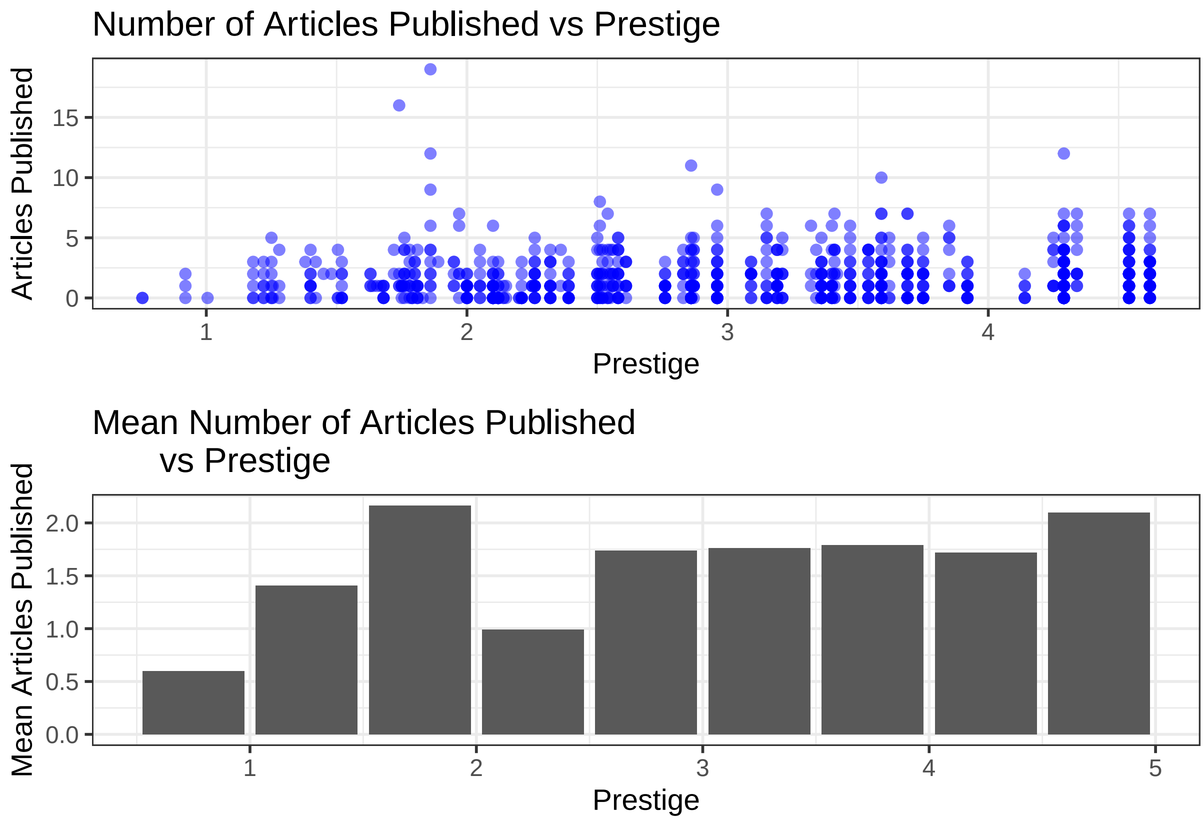
```
prestige_points <- pub %>%
  ggplot(aes(x = prestige, y = articles)) +
  geom_point(alpha = 0.5, color = "blue") +
  theme_bw() +
  labs(title = "Number of Articles Published vs Prestige",
       x = "Prestige", y = "Articles Published")

prestige_points
```




```
prestige_colplot <- pub %>%
  # Binning prestige into increments of 0.5 points
  mutate(prestige = floor(prestige*2)/2 + 0.25) %>%
  group_by(prestige) %>%
  summarise(mean = mean(articles)) %>%
  ggplot(aes(x = prestige, y = mean)) +
  geom_col() +
  theme_bw() +
  labs(title = "Mean Number of Articles Published
       vs Prestige",
       x = "Prestige", y = "Mean Articles Published")

prestige_colplot
```



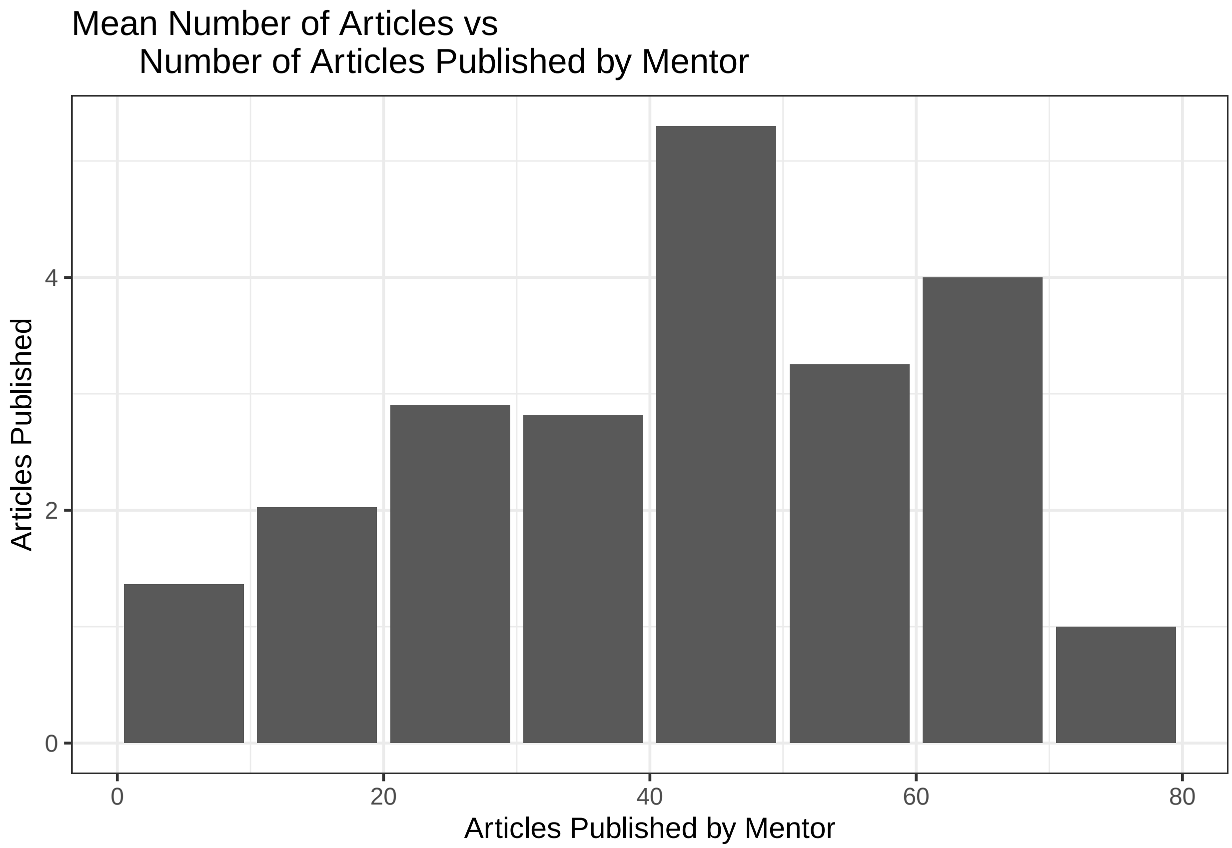

```
prestige_points / prestige_colplot
```



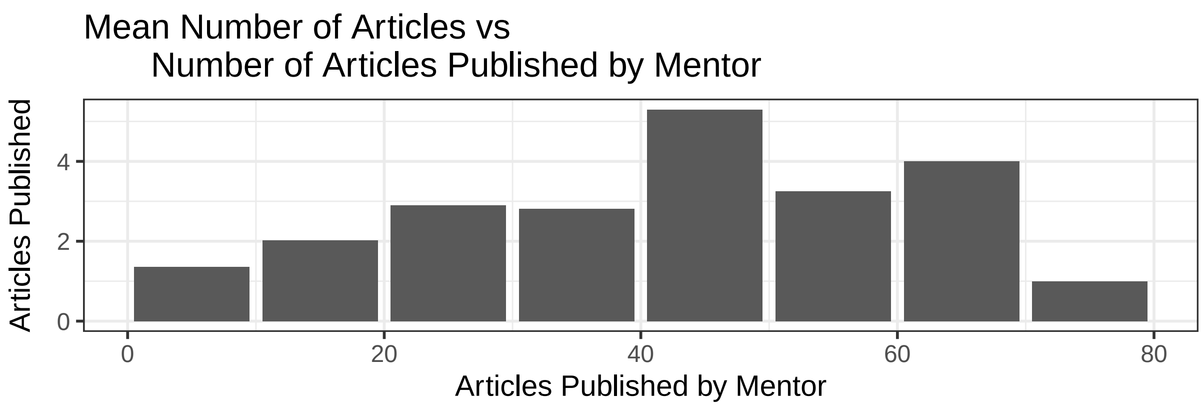
Grouping prestige into 0.5 unit increments and taking the mean we see that the mean number of publications seems to increase slightly with prestige.

```
mentor_points <- pub %>%
  ggplot(aes(x = mentor, y = articles)) +
  geom_point(alpha = 0.5, color = "blue") +
  theme_bw() +
  labs(title = "Number of Articles Published
       vs Number of Articles Published by Mentor",
       x = "Articles Published by Mentor", y = "Articles Published")
mentor_points
```


```
mentor_colplot <- pub %>%
  mutate(mentor = floor(mentor/10)*10 + 5) %>%
  group_by(mentor) %>%
  summarise(mean = mean(articles)) %>%
  ggplot(aes(x = mentor, y = mean)) +
  geom_col() +
  theme_bw() +
  labs(title = "Mean Number of Articles vs
       Number of Articles Published by Mentor",
       x = "Articles Published by Mentor", y = "Articles Published")
mentor_colplot
```



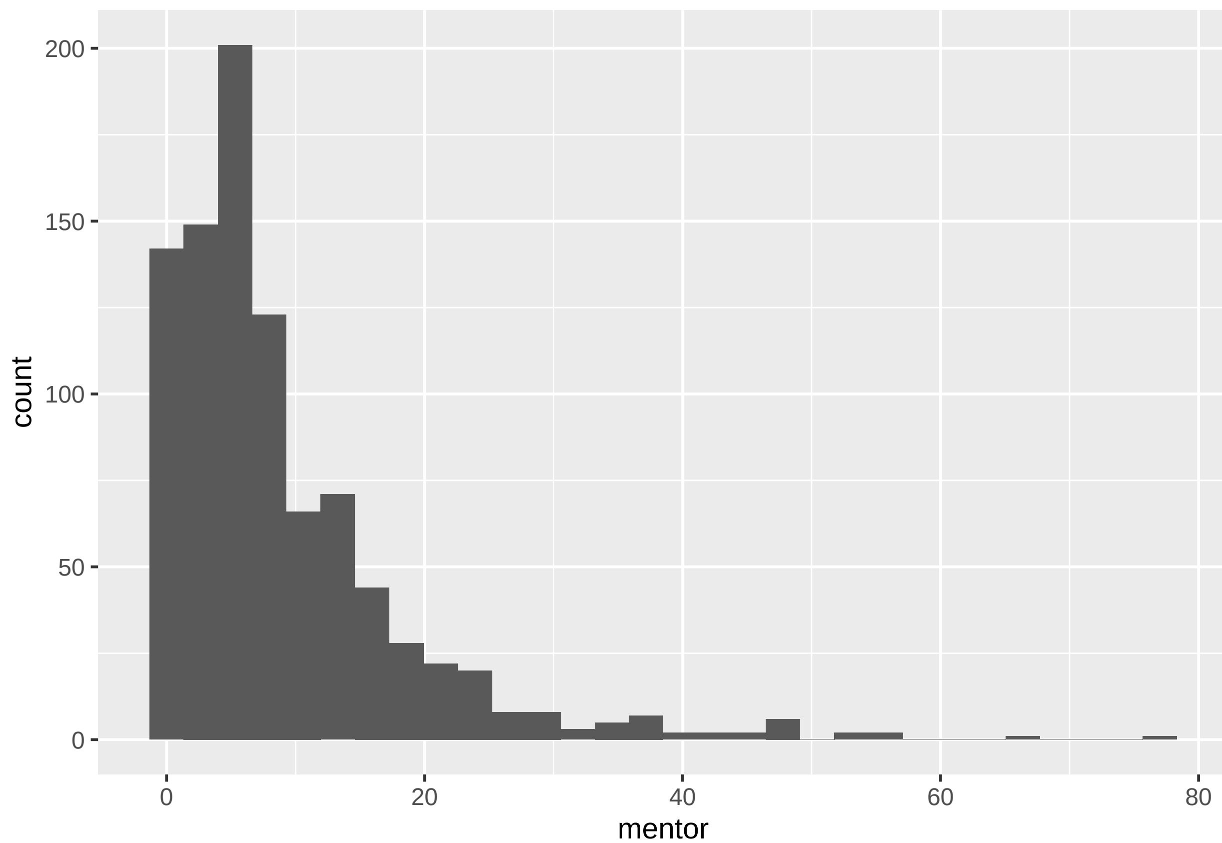

```
mentor_points / mentor_colplot
```



Grouping the number of articles published by mentors into intervals of 10 articles, there appears to be some positive relationship between the mean number of articles produced by PhD candidates and output of their mentors but the relationship is not particularly clear. There is also very sparse data at the end of the series.

```
pub %>%
  ggplot(aes(x = mentor)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Model Fitting

```
pub.glm <- glm(articles ~ 1 + female + married + kids + prestige + mentor +
                 female*(married + kids + prestige + mentor),
               data = pub,
               family = poisson)

summary(pub.glm)
```

```
##
## Call:
## glm(formula = articles ~ 1 + female + married + kids + prestige +
##     mentor + female * (married + kids + prestige + mentor), family = poisson,
##     data = pub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7415  -1.5570  -0.3722   0.5641   5.3047
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.513401   0.136150   3.771 0.000163 ***
## female1            -0.687085   0.200357  -3.429 0.000605 ***
```

```
## married1          0.093216   0.090495   1.030 0.302981
## kids1            -0.187308   0.085881  -2.181 0.029183 *
## kids2            -0.272641   0.103798  -2.627 0.008623 **
## kids3            -0.829794   0.294683  -2.816 0.004864 **
## prestige         -0.041905   0.034074  -1.230 0.218774
## mentor            0.025854   0.002289  11.297  < 2e-16 ***
## female1:married1  0.088610   0.126630   0.700 0.484077
## female1:kids1     0.028791   0.155220   0.185 0.852851
## female1:kids2    -0.231699   0.233793  -0.991 0.321663
## female1:kids3     0.308019   1.044513   0.295 0.768076
## female1:prestige  0.133147   0.055437   2.402 0.016317 *
## female1:mentor   -0.001124   0.004936  -0.228 0.819832
## ---
## 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: 1625.6  on 901  degrees of freedom
## AIC: 3321.4
##
## Number of Fisher Scoring iterations: 5
```

Being female significant negative effect, as expected. Being married has a slight positive effect but it is not statistically significant; although it looked on the exploratory plots that there might be a difference, the means were quite similar so this seems to make sense. Having children has a statistically significant negative effect, as we would expect given the plots. Somewhat surprisingly, prestige does not have a statistically significant effect in the model, but the output of the mentor has a slight positive effect.

None of the interactions appear significant apart from female1:prestige which has a slight positive effect.

AIC model selection

```
# Null model
pub.glm.null <- glm(articles ~ 1,
               data = pub,
               family = poisson)

stepAIC(pub.glm.null,
        scope = list(lower = pub.glm.null, upper = pub.glm),
        data = pub,
        direction = "forward")
```

```
## Start:  AIC=3487.15
## articles ~ 1
##
##            Df Deviance    AIC
## + mentor    1   1669.5 3341.3
## + female    1   1794.4 3466.1
## + prestige  1   1806.6 3478.3
## + kids      3   1806.1 3481.8
## + married   1   1814.6 3486.3
```

```
## <none>          1817.4 3487.1
##
## Step:  AIC=3341.29
## articles ~ mentor
##
##            Df Deviance    AIC
## + female    1   1657.0 3330.7
## + kids      3   1658.6 3336.4
## <none>          1669.5 3341.3
## + married   1   1667.6 3341.4
## + prestige  1   1669.3 3343.1
##
## Step:  AIC=3330.74
## articles ~ mentor + female
##
##                 Df Deviance    AIC
## + kids           3   1638.9 3318.6
## <none>               1657.0 3330.7
## + married        1   1656.7 3332.4
## + prestige       1   1656.8 3332.6
## + female:mentor  1   1657.0 3332.7
##
## Step:  AIC=3318.64
## articles ~ mentor + female + kids
##
##                 Df Deviance    AIC
## + married        1   1633.3 3315.1
## <none>               1638.9 3318.6
## + female:mentor  1   1638.8 3320.5
## + prestige       1   1638.8 3320.6
## + female:kids    3   1637.6 3323.3
##
## Step:  AIC=3315.07
## articles ~ mentor + female + kids + married
##
##                  Df Deviance    AIC
## <none>                1633.3 3315.1
## + female:married  1   1633.1 3316.8
## + prestige        1   1633.1 3316.8
## + female:mentor   1   1633.2 3317.0
## + female:kids     3   1632.0 3319.8
```

```
##
## Call:  glm(formula = articles ~ mentor + female + kids + married, family = poisson,
##     data = pub)
##
## Coefficients:
## (Intercept)       mentor      female1        kids1        kids2        kids3
##     0.34765      0.02557     -0.22597     -0.18031     -0.32775     -0.82152
##    married1
##     0.14812
##
## Degrees of Freedom: 914 Total (i.e. Null);  908 Residual
## Null Deviance:       1817
```

```
## Residual Deviance: 1633  AIC: 3315
```

Using forward selection we get the formula: articles ~ mentor + female + kids + married.

```r
stepAIC(pub.glm,
        scope = list(lower = pub.glm.null, upper = pub.glm),
        data = pub,
        direction = "backward")
```

```
## Start:  AIC=3321.36
## articles ~ 1 + female + married + kids + prestige + mentor +
##     female * (married + kids + prestige + mentor)
##
##                   Df Deviance    AIC
## - female:kids      3   1626.9 3316.6
## - female:mentor    1   1625.7 3319.4
## - female:married   1   1626.1 3319.9
## <none>                 1625.6 3321.4
## - female:prestige  1   1631.4 3325.2
##
## Step:  AIC=3316.64
## articles ~ female + married + kids + prestige + mentor + female:married +
##     female:prestige + female:mentor
##
##                   Df Deviance    AIC
## - female:mentor    1   1627.0 3314.8
## - female:married   1   1627.3 3315.0
## <none>                 1626.9 3316.6
## - female:prestige  1   1632.8 3320.6
## - kids             3   1649.0 3332.7
##
## Step:  AIC=3314.79
## articles ~ female + married + kids + prestige + mentor + female:married +
##     female:prestige
##
##                   Df Deviance    AIC
## - female:married   1   1627.4 3313.2
## <none>                 1627.0 3314.8
## - female:prestige  1   1632.9 3318.6
## - kids             3   1649.3 3331.0
## - mentor           1   1757.2 3442.9
##
## Step:  AIC=3313.18
## articles ~ female + married + kids + prestige + mentor + female:prestige
##
##                   Df Deviance    AIC
## <none>                 1627.4 3313.2
## - married          1   1632.2 3316.0
## - female:prestige  1   1633.1 3316.8
## - kids             3   1651.0 3330.8
## - mentor           1   1757.5 3441.2
```

```
##
```

```
## Call:  glm(formula = articles ~ female + married + kids + prestige +
##     mentor + female:prestige, family = poisson, data = pub)
##
## Coefficients:
##      (Intercept)          female1          married1             kids1
##          0.47391         -0.63170           0.13862          -0.18735
##            kids2            kids3          prestige            mentor
##         -0.32483         -0.82756          -0.03619           0.02542
## female1:prestige
##          0.12655
##
## Degrees of Freedom: 914 Total (i.e. Null);  906 Residual
## Null Deviance:       1817
## Residual Deviance: 1627  AIC: 3313
```

Using backward selection, the formula selected is:

articles ~ female + married + kids + prestige + mentor + female:prestige.

```
stepAIC(pub.glm.null,
        scope = list(lower = pub.glm.null, upper = pub.glm),
        data = pub,
        direction = "both")
```

```
## Start:  AIC=3487.15
## articles ~ 1
##
##            Df Deviance    AIC
## + mentor    1   1669.5 3341.3
## + female    1   1794.4 3466.1
## + prestige  1   1806.6 3478.3
## + kids      3   1806.1 3481.8
## + married   1   1814.6 3486.3
## <none>          1817.4 3487.1
##
## Step:  AIC=3341.29
## articles ~ mentor
##
##            Df Deviance    AIC
## + female    1   1657.0 3330.7
## + kids      3   1658.6 3336.4
## <none>          1669.5 3341.3
## + married   1   1667.6 3341.4
## + prestige  1   1669.3 3343.1
## - mentor    1   1817.4 3487.1
##
## Step:  AIC=3330.74
## articles ~ mentor + female
##
##                 Df Deviance    AIC
## + kids           3   1638.9 3318.6
## <none>               1657.0 3330.7
## + married        1   1656.7 3332.4
## + prestige       1   1656.8 3332.6
```

```
## + female:mentor  1   1657.0 3332.7
## - female         1   1669.5 3341.3
## - mentor         1   1794.4 3466.1
##
## Step:  AIC=3318.64
## articles ~ mentor + female + kids
##
##                 Df Deviance    AIC
## + married        1   1633.3 3315.1
## <none>               1638.9 3318.6
## + female:mentor  1   1638.8 3320.5
## + prestige       1   1638.8 3320.6
## + female:kids    3   1637.6 3323.3
## - kids           3   1657.0 3330.7
## - female         1   1658.6 3336.4
## - mentor         1   1775.9 3453.6
##
## Step:  AIC=3315.07
## articles ~ mentor + female + kids + married
##
##                  Df Deviance    AIC
## <none>                1633.3 3315.1
## + female:married  1   1633.1 3316.8
## + prestige        1   1633.1 3316.8
## + female:mentor   1   1633.2 3317.0
## - married         1   1638.9 3318.6
## + female:kids     3   1632.0 3319.8
## - female          1   1650.5 3330.3
## - kids            3   1656.7 3332.4
## - mentor          1   1772.5 3452.2
```

```
##
## Call:  glm(formula = articles ~ mentor + female + kids + married, family = poisson,
##     data = pub)
##
## Coefficients:
## (Intercept)       mentor      female1        kids1        kids2        kids3
##     0.34765      0.02557     -0.22597     -0.18031     -0.32775     -0.82152
##    married1
##     0.14812
##
## Degrees of Freedom: 914 Total (i.e. Null);  908 Residual
## Null Deviance:       1817
## Residual Deviance: 1633  AIC: 3315
```

So the question, how do we select between these models?

Test for inclusion of prestige

```
pub.glm.noprestige <- glm(articles ~ 1 + female + married + kids + mentor,
               data = pub,
               family = poisson)
pub.glm.prestige <- glm(articles ~ 1 + female + married + kids + mentor + prestige,
               data = pub,
```

```
               family = poisson)
dof <- pub.glm.prestige$rank - pub.glm.noprestige$rank
lrt <- deviance (pub.glm.noprestige) - deviance(pub.glm.prestige)
pval <- 1 - pchisq(lrt,dof)
cbind(lrt,dof,pval)
```

```
##            lrt dof      pval
## [1,] 0.2325111   1 0.6296681
```

Just including prestige is not significant.

```
pub.glm.noprestige <- glm(articles ~ 1 + female + married + kids + mentor,
               data = pub,
               family = poisson)
pub.glm.prestige <- glm(articles ~ 1 + female + married + kids + mentor + prestige
                        + female*prestige,
               data = pub,
               family = poisson)
dof <- pub.glm.prestige$rank - pub.glm.noprestige$rank
lrt <- deviance (pub.glm.noprestige) - deviance(pub.glm.prestige)
pval <- 1 - pchisq(lrt,dof)
cbind(lrt,dof,pval)
```

```
##           lrt dof       pval
## [1,] 5.893019   2 0.05252273
```

The hypothesis test suggests that there is a significant effect for including both prestige and female at the 10% significance level.

Let us include the prestige and female*prestige term since it has a borderline significant effect which we may want to examine in our analysis.

Marriage may not be significant, so let us test for it:

```
pub.glm.nomarriage <- glm(articles ~ 1 + female + kids + mentor,
               data = pub,
               family = poisson)
pub.glm.marriage <- glm(articles ~ 1 + female + married + kids + mentor,
               data = pub,
               family = poisson)
dof <- pub.glm.marriage$rank - pub.glm.nomarriage$rank
lrt <- deviance (pub.glm.nomarriage) - deviance(pub.glm.marriage)
pval <- 1 - pchisq(lrt,dof)
cbind(lrt,dof,pval)
```

```
##           lrt dof       pval
## [1,] 5.571698   1 0.01825305
```

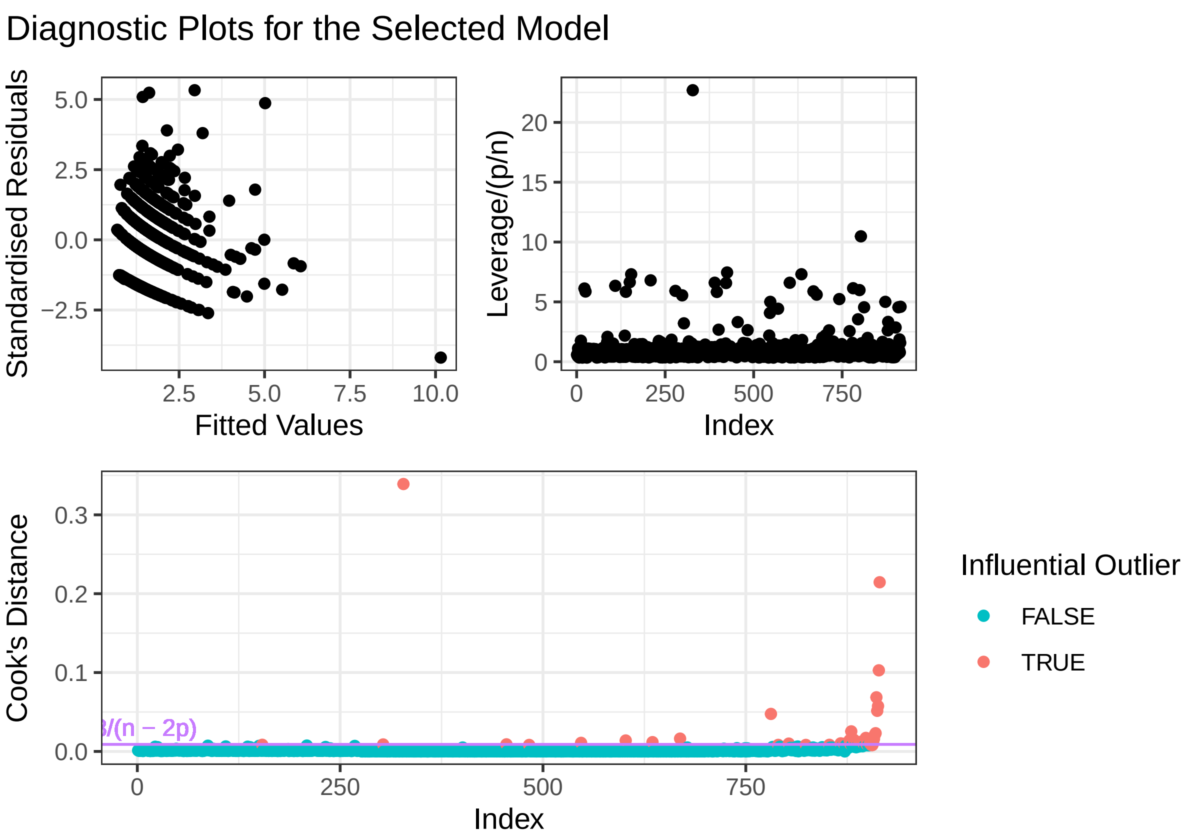
Final model:

```
pub.glm.selected <- glm(articles ~ 1 + female + married + kids + mentor + prestige
                        + female*prestige,
               data = pub,
               family = poisson)
```

Diagnostics

Here are the diagnostic plots for the model: standardised residuals, leverage and Cook's distances.

```
# Number of coefficients in the model
p <- pub.glm.selected$rank
# Number of observations
n <- nrow(model.frame(pub.glm.selected))

r_standard_plot <- data.frame(fitted_values = fitted(pub.glm.selected),
           r_standard = rstandard(pub.glm.selected)) %>%
  ggplot(aes(x = fitted_values, y = r_standard)) +
  geom_point() +
  theme_bw() +
  labs(x = "Fitted Values", y = "Standardised Residuals")

influence_plot <- data.frame(influence = influence(pub.glm.selected)$hat/(p/n),
           index = 1:length(influence(pub.glm.selected)$hat))  %>%
  ggplot(aes(x = index, y = influence)) +
  geom_point() +
  theme_bw() +
  labs(x = "Index", y = "Leverage/(p/n)")

cooks_bound <-  8 / (n - 2 * p)

cooks_distance_plot <- data.frame(cooks_distance = cooks.distance(pub.glm.selected),
                                  index = 1:n) %>%
  # Select slightly below the bound to be influential outliers in the plot too
  mutate(influential_outlier = cooks_distance >= cooks_bound - 0.001) %>%
  ggplot() +
  geom_point(aes(x = index, y = cooks_distance, colour = influential_outlier)) +
  geom_hline(aes(yintercept = cooks_bound), colour = "#C77CFF") +
  geom_text(aes( x = 10, y = cooks_bound, label = "8/(n - 2p)", vjust = -0.5),
            size = 3, colour = "#C77CFF") +
  theme_bw() +
  scale_colour_manual(name = "Influential Outlier", values = c("#00BFC4","#F8766D")) +
  labs(x = "Index", y = "Cook's Distance")

(r_standard_plot + influence_plot) / cooks_distance_plot +
  plot_annotation(
    title = "Diagnostic Plots for the Selected Model"
  )
```



If we remove the "influential outliers", this causes some changes to the analysis:

```
# These are the indices of the points which are above the bound for Cook's distance
which(cooks.distance(pub.glm.selected) > 8/(n-2*p))
```

```
## 328 455 547 602 635 669 781 803 867 871 878 879 880 883 887 890 892 893 895 896
## 328 455 547 602 635 669 781 803 867 871 878 879 880 883 887 890 892 893 895 896
## 898 899 900 903 904 907 908 909 910 911 912 913 914 915
## 898 899 900 903 904 907 908 909 910 911 912 913 914 915
```

```
# Removing influential outliers and performing analysis again

pub2 <- pub %>%
  slice(which(cooks.distance(pub.glm.selected) < 8/(n-2*p)))

# New analysis of the data
pub.glm2 <- glm(articles ~ 1 + female + married + kids + prestige + mentor +
                  female*(married + kids + prestige + mentor),
                  data = pub2,
                  family = poisson)
pub.glm2
```

```
##
## Call:  glm(formula = articles ~ 1 + female + married + kids + prestige +
##     mentor + female * (married + kids + prestige + mentor), family = poisson,
```

```
##     data = pub2)
##
## Coefficients:
##      (Intercept)           female1          married1             kids1
##         0.172387         -0.297644          0.084918         -0.156762
##            kids2             kids3          prestige            mentor
##        -0.233044         -1.315832          0.017901          0.027858
## female1:married1     female1:kids1     female1:kids2     female1:kids3
##         0.065173          0.039737         -0.355084          0.878977
## female1:prestige    female1:mentor
##         0.049186         -0.005856
##
## Degrees of Freedom: 880 Total (i.e. Null);  867 Residual
## Null Deviance:       1429
## Residual Deviance: 1298  AIC: 2872
```

Looking to see if there is any difference in automatic model selection:

```
# Null model for this data
pub.glm2.null <- glm(articles ~ 1,
               data = pub2,
               family = poisson)

stepAIC(pub.glm2.null,
        scope = list(lower = pub.glm2.null, upper = pub.glm2),
        data = pub2,
        direction = "forward")
```

```
## Start:  AIC=2976.58
## articles ~ 1
##
##            Df Deviance    AIC
## + mentor    1   1332.5 2882.1
## + prestige  1   1411.8 2961.5
## + kids      3   1413.5 2967.1
## + female    1   1419.8 2969.4
## <none>          1429.0 2976.6
## + married   1   1428.5 2978.1
##
## Step:  AIC=2882.11
## articles ~ mentor
##
##            Df Deviance    AIC
## + kids      3   1318.0 2873.6
## + female    1   1326.9 2878.6
## <none>          1332.5 2882.1
## + prestige  1   1330.7 2882.3
## + married   1   1331.9 2883.5
##
## Step:  AIC=2873.57
## articles ~ mentor + kids
##
##            Df Deviance    AIC
## + female    1   1307.5 2865.2
```

```
## + married   1   1313.6 2871.2
## <none>          1318.0 2873.6
## + prestige  1   1316.2 2873.8
##
## Step:  AIC=2865.16
## articles ~ mentor + kids + female
##
##                  Df Deviance    AIC
## + married        1   1304.6 2864.2
## <none>               1307.5 2865.2
## + prestige       1   1306.0 2865.7
## + female:mentor  1   1306.8 2866.4
## + female:kids    3   1304.4 2868.1
##
## Step:  AIC=2864.21
## articles ~ mentor + kids + female + married
##
##                   Df Deviance    AIC
## <none>                1304.6 2864.2
## + prestige         1   1302.8 2864.4
## + female:mentor    1   1303.8 2865.4
## + female:married   1   1304.4 2866.0
## + female:kids      3   1301.4 2867.1
```

```
##
## Call:  glm(formula = articles ~ mentor + kids + female + married, family = poisson,
##     data = pub2)
##
## Coefficients:
## (Intercept)       mentor        kids1        kids2        kids3      female1
##     0.22852      0.02677     -0.15225     -0.30705     -1.23641     -0.17462
##    married1
##     0.11554
##
## Degrees of Freedom: 880 Total (i.e. Null);  874 Residual
## Null Deviance:       1429
## Residual Deviance: 1305  AIC: 2864
```

```
stepAIC(pub.glm2,
        scope = list(lower = pub.glm2.null, upper = pub.glm2),
        data = pub2,
        direction = "backward")
```

```
## Start:  AIC=2871.91
## articles ~ 1 + female + married + kids + prestige + mentor +
##     female * (married + kids + prestige + mentor)
##
##                    Df Deviance    AIC
## - female:kids       3   1301.1 2868.7
## - female:married    1   1298.5 2870.1
## - female:prestige   1   1299.0 2870.6
## - female:mentor     1   1299.3 2870.9
## <none>                  1298.3 2871.9
```

```
##
## Step:  AIC=2868.74
## articles ~ female + married + kids + prestige + mentor + female:married +
##     female:prestige + female:mentor
##
##                   Df Deviance    AIC
## - female:married   1   1301.2 2866.9
## - female:prestige  1   1301.8 2867.4
## - female:mentor    1   1302.5 2868.1
## <none>                 1301.1 2868.7
## - kids             3   1322.2 2883.8
##
## Step:  AIC=2866.86
## articles ~ female + married + kids + prestige + mentor + female:prestige +
##     female:mentor
##
##                   Df Deviance    AIC
## - female:prestige  1   1301.9 2865.5
## - female:mentor    1   1302.6 2866.2
## <none>                 1301.2 2866.9
## - married          1   1304.3 2867.9
## - kids             3   1323.1 2882.7
##
## Step:  AIC=2865.5
## articles ~ female + married + kids + prestige + mentor + female:mentor
##
##                 Df Deviance    AIC
## - female:mentor  1   1302.8 2864.4
## - prestige       1   1303.8 2865.4
## <none>               1301.9 2865.5
## - married        1   1305.2 2866.8
## - kids           3   1323.7 2881.3
##
## Step:  AIC=2864.41
## articles ~ female + married + kids + prestige + mentor
##
##            Df Deviance    AIC
## - prestige  1   1304.6 2864.2
## <none>          1302.8 2864.4
## - married   1   1306.0 2865.7
## - female    1   1311.5 2871.1
## - kids      3   1325.0 2880.7
## - mentor    1   1380.7 2940.3
##
## Step:  AIC=2864.21
## articles ~ female + married + kids + mentor
##
##           Df Deviance    AIC
## <none>         1304.6 2864.2
## - married  1   1307.5 2865.2
## - female   1   1313.6 2871.2
## - kids     3   1326.9 2880.5
## - mentor   1   1397.0 2954.6
```

```
##
## Call:  glm(formula = articles ~ female + married + kids + mentor, family = poisson,
##     data = pub2)
##
## Coefficients:
## (Intercept)      female1     married1        kids1        kids2        kids3
##     0.22852     -0.17462      0.11554     -0.15225     -0.30705     -1.23641
##      mentor
##     0.02677
##
## Degrees of Freedom: 880 Total (i.e. Null);  874 Residual
## Null Deviance:       1429
## Residual Deviance: 1305  AIC: 2864
```

I do not think that there is a good justification for removing influential outliers from the analysis. I cannot think of a good reason that the information would be incorrectly recorded other than a few minor counting errors.

Model quality

This is a Poisson model so a goodness of fit test can be appropriate, however the counts are quite small. The deviance is given by $D(y) \sim \chi^2(n-p)$ under $\mathcal{H}_0$. Here \$n = \$ 915 and \$p = \$ 9

```
# Goodness of fit for the model without outliers excluded
deviance(pub.glm.prestige)
```

```
## [1] 1627.437
```

```
dim(pub)[1] - pub.glm.prestige$rank
```

```
## [1] 906
```

```
qchisq(0.95,dim(pub)[1] - pub.glm.prestige$rank)
```

```
## [1] 977.1359
```

```
# Goodness of fit for the model with outliers excluded
deviance(pub.glm2)
```

```
## [1] 1298.294
```

```
dim(pub2)[1] - pub.glm2$rank
```

```
## [1] 867
```

```
qchisq(0.95,dim(pub2)[1] - pub.glm2$rank)
```

```
## [1] 936.6119
```

This is not a good fit. With outliers removed it is still not a good fit.

Here we are looking at the R^2_{KL} value:

```
rsq(pub.glm.selected,type = "kl")
```

```
## [1] 0.104527
```

```
rsq(pub.glm2,type = "kl")
```

```
## [1] 0.09144423
```

Without removing the data the R^2_{KL} is better for the first model. It is quite close to 0, so the model really does not capture a lot of the variation in the data.

Interpretation

Confidence intervals at 95% level:

```
coefs <- summary(pub.glm.selected)$coef[2:pub.glm.selected$rank,1]
SEs <- summary(pub.glm.selected)$coef[2:pub.glm.selected$rank,2]
cval <- qnorm(0.975)

CI <- data.frame(
  estimate = coefs,
  lower = coefs - cval * SEs,
  upper = coefs + cval * SEs
)
# Transforming the confidence intervals to find the multiplicative factors
CI <- CI %>%
  mutate(exp_estimate = exp(estimate),
         exp_lower = exp(lower),
         exp_upper = exp(upper))
CI <- signif(CI,3)
CI$rownames <- rownames(CI)

CI
```

```
##                  estimate   lower   upper exp_estimate exp_lower exp_upper
## female1           -0.6320 -0.9850 -0.2780        0.532     0.373     0.757
## married1           0.1390  0.0146  0.2630        1.150     1.010     1.300
## kids1             -0.1870 -0.3260 -0.0486        0.829     0.722     0.953
## kids2             -0.3250 -0.5030 -0.1470        0.723     0.605     0.864
## kids3             -0.8280 -1.3800 -0.2750        0.437     0.252     0.759
## mentor             0.0254  0.0215  0.0294        1.030     1.020     1.030
## prestige          -0.0362 -0.1020  0.0294        0.964     0.903     1.030
## female1:prestige   0.1270  0.0221  0.2310        1.130     1.020     1.260
##                          rownames
## female1                   female1
## married1                 married1
## kids1                       kids1
```

```
## kids2                         kids2
## kids3                         kids3
## mentor                       mentor
## prestige                   prestige
## female1:prestige female1:prestige
```

Estimating ϕ

Estimate for $\hat{phi}$:

```
phi_hat <- 1/(dim(pub)[1] - pub.glm.prestige$rank) *
  sum((pub$articles - pub.glm.selected$fitted.values)^2/pub.glm.selected$fitted.values)
```

The data is actually overdispersed. This implies that the standard errors were too small and that the CI's were too narrow. We can adjust the estimated variances by multiplying by a factor of $\hat{\phi}$, giving the following confidence intervals:

```
coefs <- summary(pub.glm.selected)$coef[2:pub.glm.selected$rank,1]
SEs <- summary(pub.glm.selected)$coef[2:pub.glm.selected$rank,2]*phi_hat^(1/2)
cval <- qnorm(0.975)

CI <- data.frame(
  estimate = coefs,
  lower = coefs - cval * SEs,
  upper = coefs + cval * SEs
)
CI <- CI %>%
  mutate(exp_estimate = exp(estimate),
         exp_lower = exp(lower),
         exp_upper = exp(upper))
CI <- signif(CI,3)
CI$rownames <- rownames(CI)

CI
```

```
##                  estimate   lower     upper exp_estimate exp_lower exp_upper
## female1           -0.6320 -1.1100 -0.156000        0.532     0.330     0.856
## married1           0.1390 -0.0283  0.306000        1.150     0.972     1.360
## kids1             -0.1870 -0.3740 -0.000719        0.829     0.688     0.999
## kids2             -0.3250 -0.5650 -0.085000        0.723     0.569     0.918
## kids3             -0.8280 -1.5700 -0.084900        0.437     0.208     0.919
## mentor             0.0254  0.0201  0.030700        1.030     1.020     1.030
## prestige          -0.0362 -0.1240  0.052000        0.964     0.883     1.050
## female1:prestige   0.1270 -0.0140  0.267000        1.130     0.986     1.310
##                          rownames
## female1                   female1
## married1                 married1
## kids1                       kids1
## kids2                       kids2
## kids3                       kids3
## mentor                     mentor
## prestige                 prestige
```

```
## female1:prestige female1:prestige
```

This has some implications for the results of our analysis.