Lab 07 Poisson Regression

Lusine Zilfimian

April 01 (Wednesday), 2020

Contents

- Libraries
- Data Preparation
- Undestanding the data
- Intercept-only model
- Poisson Regression with one explanatory variable
- Interpretation
- Prediction
- Overdispersion
- Model Selection

Needed packages

```r
library(ggplot2)
library(ggpubr)
library(dplyr)
library(AER) # for disp dispersiontest()
library(MASS) # for glm.nb()
```

Data

- Data with 915 observations on the following 6 variables.
- **art** - Articles during last 3 years of Ph.D.
- **fem** - 1 if female scientist; male - 0
- **mar** - 1 if married; else 0
- **kid5** - Number of children 5 or younger
- **phd** - Prestige of Ph.D. department
- **ment** - Articles by mentor during last 3 years

Data Preparation

```
ar <- read.csv("articles.csv")
str(ar)
```

```
## 'data.frame':    915 obs. of  6 variables:
##  $ art : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fem : int  0 1 1 0 1 1 1 0 0 1 ...
##  $ mar : int  1 0 0 1 0 1 0 1 0 1 ...
##  $ kid5: int  0 0 0 1 0 2 0 2 0 0 ...
##  $ phd : num  2.52 2.05 3.75 1.18 3.75 ...
##  $ ment: int  7 6 6 3 26 2 3 4 6 0 ...
```

```
ar$fem <- factor(ar$fem, levels = c(0,1), labels = c("Male", "Female"))
ar$mar <- factor(ar$mar, levels = c(0,1), labels = c("Else", "Married"))
unique(ar$kid5)
```

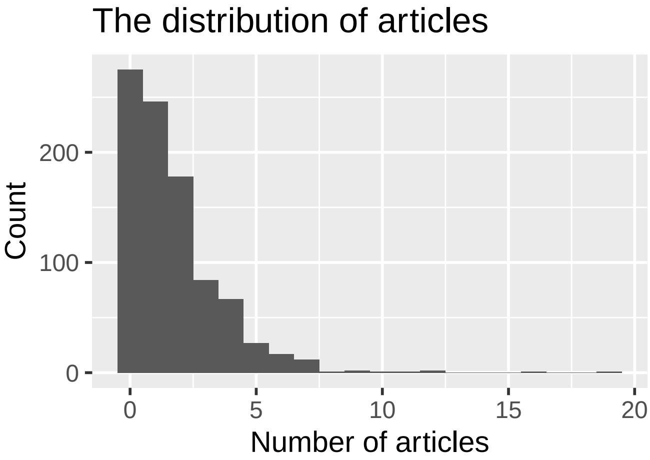
```
## [1] 0 1 2 3
```

```
unique(ar$art)
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 16 19
```

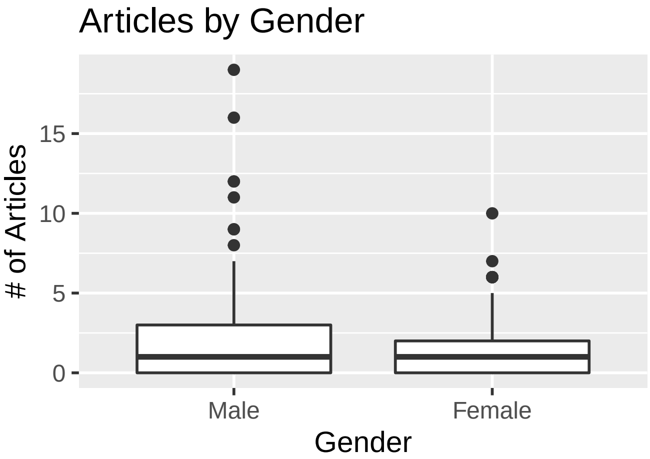
Undestanding the data

```r
ggplot(data = ar, aes(x = art)) + geom_histogram(bins = 20) +
  labs(x = "Number of articles", y = "Count") +
  ggtitle("The distribution of articles")
```



Undestanding the data

```r
ggplot(data = ar, aes(x = fem, y = art))+
  geom_boxplot()+
  labs(x = "Gender", y = "# of Articles", title = "Articles by Gender")
```



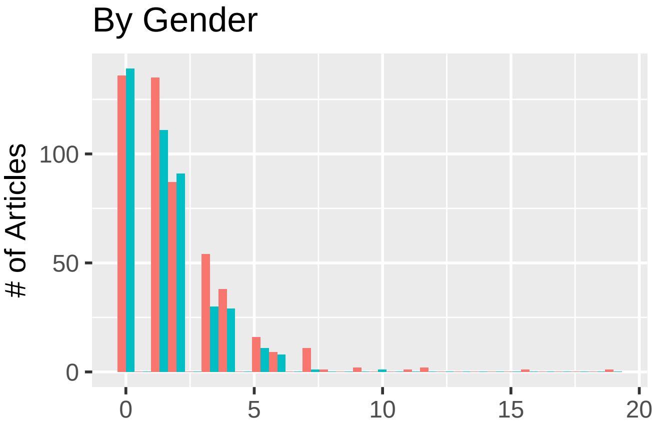
Undestanding the data

```
ar %>%
  group_by(fem) %>%
  summarise(Mean = mean(art), SD = sd(art), Min = min(art), Max = max(art))
```

```
## # A tibble: 2 x 5
##   fem     Mean    SD   Min   Max
##   <fct>  <dbl> <dbl> <int> <int>
## 1 Male    1.88  2.18     0    19
## 2 Female  1.47  1.55     0    10
```

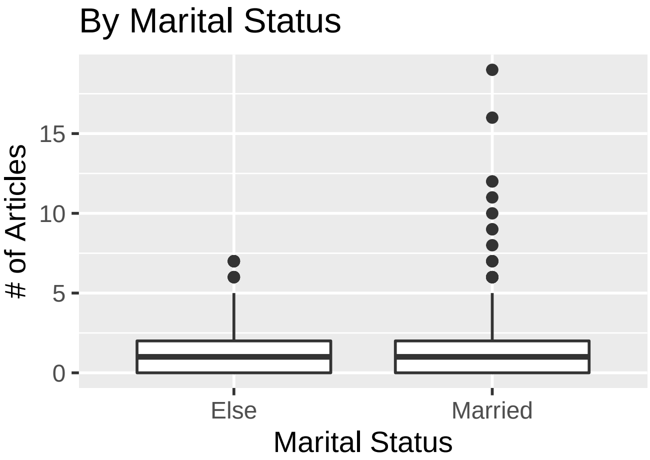
Undestanding the data

```
ggplot(ar, aes(x = art, fill = fem)) +
  geom_histogram(position = "dodge") +
  labs(x="", y="# of Articles", title="By Gender") +
  theme(legend.position = "None")
```



Undestanding the data

```
ggplot(data = ar, aes(x = mar, y = art)) +
  geom_boxplot() + labs( x = "Marital Status", y = "# of Articles",
    title="By Marital Status")
```



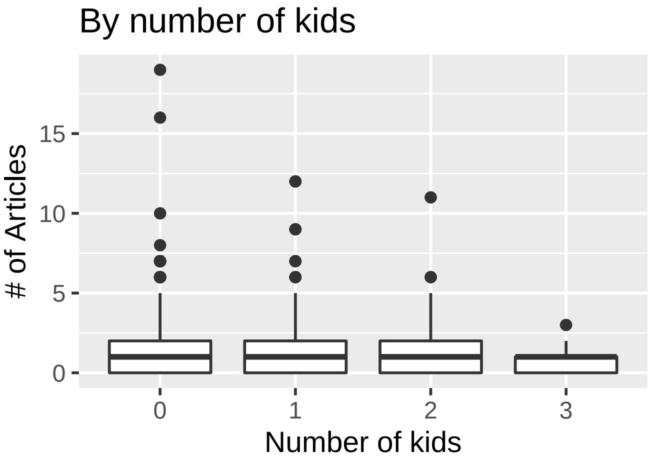
Undestanding the data

```r
ar %>%
  group_by(mar)%>%
  summarise(Mean = mean(art), SD = sd(art), Min = min(art), Max = max(art))
```

```
## # A tibble: 2 x 5
##   mar      Mean    SD   Min   Max
##   <fct>   <dbl> <dbl> <int> <int>
## 1 Else     1.59  1.73     0     7
## 2 Married  1.74  2.02     0    19
```

Undestanding the data

```
ggplot(data = ar, aes(x = factor(kid5), y = art))+
  geom_boxplot() + labs(x = "Number of kids", y = "# of Articles",
    title="By number of kids")
```



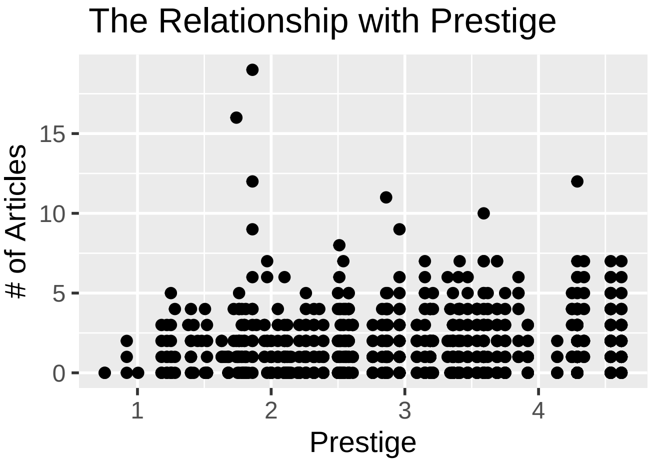
Undestanding the data

```
ar %>%
  group_by(factor(kid5))%>%
  summarise(Mean = mean(art), SD = sd(art), Min = min(art), Max = max(art))
```

```
## # A tibble: 4 x 5
##   `factor(kid5)`  Mean    SD   Min   Max
##   <fct>          <dbl> <dbl> <int> <int>
## 1 0              1.72  1.93      0    19
## 2 1              1.76  2.05      0    12
## 3 2              1.54  1.74      0    11
## 4 3              0.812 0.911     0     3
```

Undestanding the data

```
ggplot(data = ar, aes(x = phd, y = art)) +
  geom_point() + labs(x = "Prestige", y = "# of Articles",
    title=" The Relationship with Prestige")
```



Undestanding the data

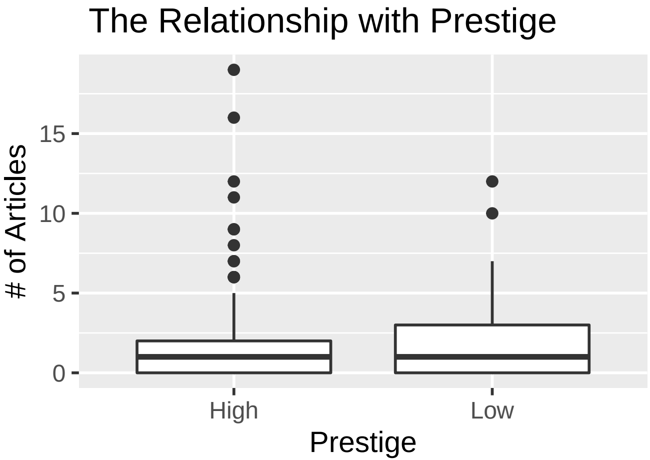
```r
summary(ar$phd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.755   2.260   3.150   3.103   3.920   4.620
```

```r
phddummy <- factor(ifelse(ar$phd > 3.1, 1, 0), levels = c(0,1), labels = c(
```

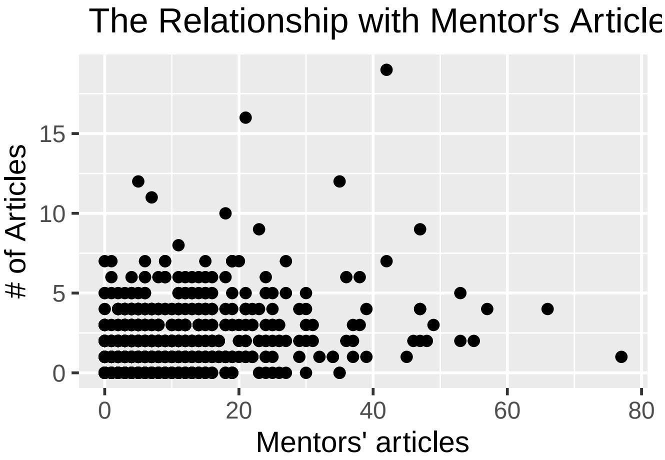
Undestanding the data

```
ggplot(data = ar, aes(x = phddummy, y = art))+
  geom_boxplot() + labs(x = "Prestige", y = "# of Articles",
    title=" The Relationship with Prestige")
```



Undestanding the data

```
ggplot(data = ar, aes(x = ment, y = art)) +
  geom_point() + labs(x = "Mentors' articles", y = "# of Articles",
    title=" The Relationship with Mentor's Articles")
```



Intercept-only model

- $\lambda = e^{\beta_0}$

```r
mod <- glm(art ~ 1, data = ar, family = poisson(link = log))
summary(mod)
```

```
##
## Call:
## glm(formula = art ~ 1, family = poisson(link = log), data = ar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8401  -1.8401  -0.5770   0.2294   7.5677
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.52644    0.02541   20.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

Intercept-only model (Outputs explanation)

```r
mean(ar$art)
```

```
## [1] 1.692896
```

```r
exp(mod$coefficients)
```

```
## (Intercept)
##    1.692896
```

```r
(z <- 0.52644/0.02541)
```

```
## [1] 20.71783
```

```r
mod$null.deviance
```

```
## [1] 1817.405
```

```r
sum(resid(mod, type = "deviance")^2)
```

```
## [1] 1817.405
```

Poisson Regression with one explanatory variable

```
mod1 <- glm(art ~ mar , data = ar, family = poisson(link = log))
summary(mod1)
```

```
##
## Call:
## glm(formula = art ~ mar, family = poisson(link = log), data = ar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8677  -1.7845  -0.5042   0.3107   7.4992
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.46514    0.04508  10.317   <2e-16 ***
## marMarried   0.09117    0.05458   1.671   0.0948 .
## ---
## 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
```

Interpretation

```r
exp(mod1$coefficients)
```

```
## (Intercept)  marMarried
##    1.592233    1.095458
```

Poisson Regression with two explanatory variables

```r
mod2 <- glm(art ~ fem + ment, data = ar, family = poisson(link = log))
summary(mod2)
```

```
##
## Call:
## glm(formula = art ~ fem + ment, family = poisson(link = log),
##     data = ar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6094  -1.5746  -0.3891   0.5651   5.7868
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.34909    0.04191   8.329  < 2e-16 ***
## femFemale   -0.18445    0.05235  -3.523 0.000426 ***
## ment         0.02510    0.00193  13.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

Interpretation

- $e^{0.0251} = 1.0254$

```
exp(mod2$coefficients)
```

```
## (Intercept)   femFemale        ment
##   1.4177815   0.8315604   1.0254201
```

Full model

```r
mod.full <- glm(art~., data = ar, family = poisson(link = log))
summary(mod.full)
```

```
##
## Call:
## glm(formula = art ~ ., family = poisson(link = log), data = ar)
##
## 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 **
## femFemale   -0.224594   0.054613  -4.112 3.92e-05 ***
## marMarried   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
## ment         0.025543   0.002006  12.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction

```r
nd <- data.frame(fem = "Female", mar = "Married", kid5 = 0, phd = 2,
  ment = 15)
as.numeric(predict(mod.full, newdata = nd))
```

```
## [1] 0.6440523
```

```r
as.numeric(lambda <- predict(mod.full, newdata = nd, type="response"))
```

```
## [1] 1.904182
```

```r
dpois(4, lambda = lambda)
```

```
## [1] 0.08159183
```

```r
ppois(4, lambda = lambda, lower.tail = T)
```

```
## [1] 0.9555782
```

```r
ppois(4, lambda = lambda, lower.tail = F)
```

```
## [1] 0.04442181
```

Goodness of Fit Test

- Chi-squared Statistics

```
qchisq(p = 0.05, 909, lower.tail = F)
```

```
## [1] 980.2518
```

```
(chi_sq <- sum(resid(mod.full, type = "pearson")^2/mod.full$fitted.values))
```

```
## [1] 1014.468
```

```
pchisq(chi_sq, df = df.residual(mod.full), lower.tail = F)
```

```
## [1] 0.008211409
```

Overdispersion

```
var(ar$art)
```

```
## [1] 3.709742
```

```
mean(ar$art)
```

```
## [1] 1.692896
```

```
ar %>% group_by(kid5) %>%
  summarise(Mean=mean(art), Var = var(art))
```

```
## # A tibble: 4 x 3
##    kid5  Mean   Var
##   <int> <dbl> <dbl>
## 1     0 1.72  3.74
## 2     1 1.76  4.19
## 3     2 1.54  3.02
## 4     3 0.812 0.829
```

```
modOver<- glm(art~kid5, data=ar, family = poisson(link = log))
```

Overdispersion

```
summary(modOver)
```

```
##
## Call:
## glm(formula = art ~ kid5, family = poisson(link = log), data = ar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8708  -1.8067  -0.5333   0.3694   7.4916
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.55960    0.02988  18.728   <2e-16 ***
## kid5        -0.06978    0.03450  -2.023   0.0431 *
## ---
## 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: 1813.2  on 913  degrees of freedom
```

Overdispertion

- $var(y) = \mu + \alpha * trafo(\mu)$
- Common specifications of the transformation function is $trafo(\mu) = \mu$
- $var(y) = \mu(1 + \alpha) = dispersion * \mu$

```
dispersiontest(modOver, trafo = NULL)
```

```
##
##  Overdispersion test
##
## data:  modOver
## z = 4.7376, p-value = 1.081e-06
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.182502
```

```
mod.qp <- glm(art ~ kid5, data = ar, family = quasipoisson(link=log))
mod.nb <- glm.nb(art ~ kid5, data = ar)
```

Overdispertion

```r
summary(mod.qp)
```

```
## 
## Call:
## glm(formula = art ~ kid5, family = quasipoisson(link = log), 
##     data = ar)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8708  -1.8067  -0.5333   0.3694   7.4916  
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.55960    0.04419  12.663   <2e-16 ***
## kid5        -0.06978    0.05102  -1.368    0.172    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for quasipoisson family taken to be 2.187224)
## 
##     Null deviance: 1817.4  on 914  degrees of freedom
```

Overdispertion

```r
summary(mod.nb)
```

```
##
## Call:
## glm.nb(formula = art ~ kid5, data = ar, init.theta = 1.715885718,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5539  -1.5139  -0.3950   0.2662   4.0655
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.56065    0.04239  13.227   <2e-16 ***
## kid5        -0.07199    0.04780  -1.506    0.132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.7159) family taken to be 1
##
##     Null deviance: 1003.7  on 914  degrees of freedom
```

Overdispertion

```
data.frame(coef(modOver), coef(mod.qp), coef(mod.nb))
```

```
##             coef.modOver. coef.mod.qp. coef.mod.nb.
## (Intercept)     0.5596001    0.5596001   0.56064809
## kid5           -0.0697818   -0.0697818  -0.07198668
```

Overdispertion

```r
summary(modOver)$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  0.5596001 0.02988037 18.728019 2.926004e-78
## kid5        -0.0697818 0.03450068 -2.022621 4.311226e-02
```

```r
summary(mod.qp)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  0.5596001 0.04419087 12.663250 5.646403e-34
## kid5        -0.0697818 0.05102398 -1.367627 1.717653e-01
```

```r
summary(mod.nb)$coefficients
```

```
##                 Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  0.56064809 0.04238579 13.227264 6.107035e-40
## kid5        -0.07198668 0.04779878 -1.506036 1.320580e-01
```

Overdispertion

```r
mod.full <- glm(art~., data = ar, family = poisson(link = log))
dispersiontest(mod.full)
```

```
##
##  Overdispersion test
##
## data:  mod.full
## z = 5.7825, p-value = 3.681e-09
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##    1.82454
```

```r
mod.full.qp <-glm(art~., data = ar, family = quasipoisson(link = log))
mod.full.nb <- glm.nb(art~., data = ar)
```

Model Selection

```
deviance(mod.full); deviance(mod.full.nb); deviance(mod.full.qp)
```

```
## [1] 1634.371

## [1] 1004.281

## [1] 1634.371
```

```
mod.full$aic
```

```
## [1] 3314.113
```

```
mod.full.nb$aic
```

```
## [1] 3135.917
```