NRES 776 Lecture 17

GLM - Poisson regression

Sunny Tseng

Our schedule today

- Announcement (0.5 min) - recording
- Review of GLM (10 min)
- Poisson regression (30 min)
- Wrap up (5 min)

Generalized linear model (GLM)

GLM is a statistical modelling technique formulated by John Nelder and Robert Wedderburn. It allows the response variable y to have an error distribution other than a normal distribution.The models include Linear Regression, Logistic Regression, and Poisson Regression.

- Generalized: GLM can accommodates other error structures (e.g., Poisson, Binomial) in addition to Normal
- Linear: The parameters, coefficients (i.e., β) are linearly combined

https://www.youtube.com/watch?app=desktop&v=ddCO2714W-o

Assumptions of GLM

- Independent observation
- The variance function (i.e., distribution type) is correctly specified
- The link function is correctly specified
- The dispersion parameter, or scale parameter (ϕ) equals 1
  - Over-dispersed ($\phi > 1$) or under-dispersed ($\phi < 1$)
  - Can change the variance function to account for this (use `quasipoisson` or `quasibinomial`)

Model goodness of fit

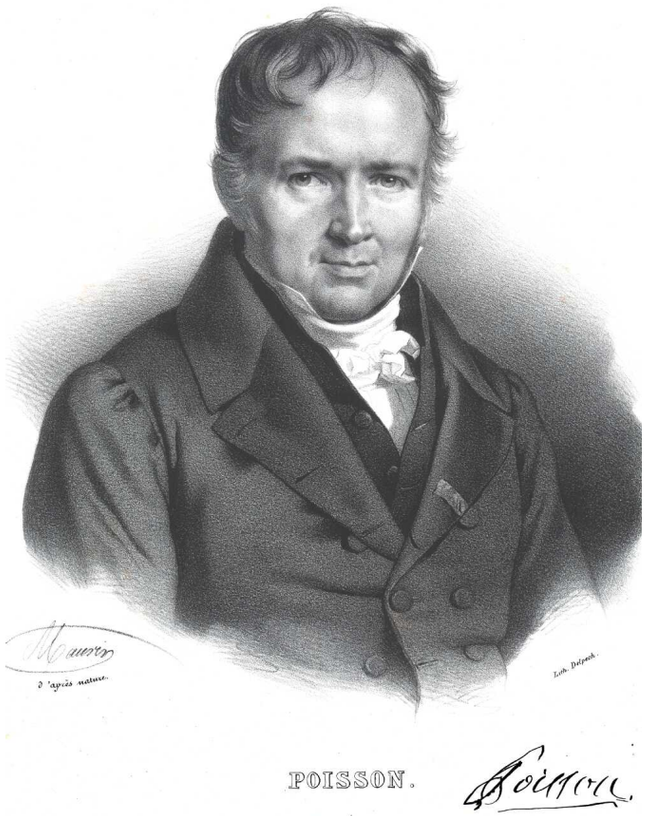
- Better models have higher likelihood
- Better models have higher Pseudo R-squared (interpretation of this is similar to the R-squared in linear models)
- Better models have lower AIC value
- Pearson residuals versus fitted values, or predictors. They should have no patterns

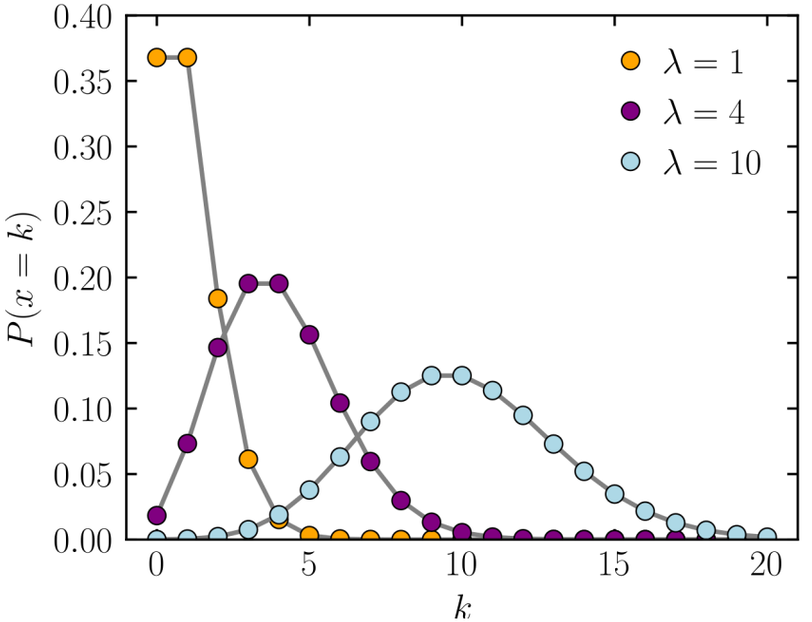
Poisson regression

Count data follows Poisson distribution

- Scores, number of vehicles, number of individuals within certain area and time
- Count data follows Poisson distribution $(0, \infty)$, which only has one parameter λ
- The mean is equal to variance, both are λ

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$


POISSON.


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$P(x = k)$
0
5
10
15
20
k
$\lambda = 1$
$\lambda = 4$
$\lambda = 10$

Overview of Poisson

1. Systematic component

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

2. Link function $g()$ –> most often **log,** makes $(0, \infty)$ to $(-\infty, \infty)$

$$\eta_i = g(\lambda_i) = log(\lambda_i)$$

3. Random component

$$var(y_i) = \lambda = \mu$$

Overview of Poisson (con'd)

$$ln(\mu_i) = ln(\hat{y}_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

Family: Poisson; Link function: log (the most commonly used)

When to use

- y is count (no natural denominator, else use y as a proportion)
- y must be a variable that is counted within defined area and time

Not to use

- y is not count or non-positive
- Non-constant sample area or time ($trees/km$ vs $trees/m$)
- Mean count $\geq$ 30 -> consider using normal distribution
- Over-dispersed -> consider quasi-Poisson
- Too many zeros -> consider zero-inflated Poisson

Overview of Poisson (con'd)

- By default `family = poisson(link = "log")`
- You can change the link to `"identity"` or `"sqrt"`
- The `variance = "mu"` for this distribution and you cannot change it from the default

```
glm_poission <- glm(formula = score ~ player,
                    data = scores,
                    family = "poisson")
glm_poission %>% summary
```

```
Call:
glm(formula = score ~ player, family = "poisson", data = scores)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6420  -3.1986  -0.2182   0.7354   6.5921

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.82666    0.08111  34.849  < 2e-16 ***
playerCindy  -0.71982    0.14175  -5.078 3.81e-07 ***
playerGilliam -0.15635    0.11946  -1.309    0.191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 379.23  on 26  degrees of freedom
Residual deviance: 350.15  on 24  degrees of freedom
```

```
AIC: 453.58

Number of Fisher Scoring iterations: 6
```

Research question (1 categorical x)

There is a investigation on how the tension (Low, Median, or High) on the number of warp breaks per loom. The “breaks” is the response variable which is a count of number of breaks. And the tension (L, M, H) is taken as the predictor variable.

```
# A tibble: 18 × 4
   round     L     M     H
   <int> <dbl> <dbl> <dbl>
 1     1    26    42    36
 2     2    30    26    21
 3     3    54    19    24
 4     4    25    16    18
 5     5    70    39    10
 6     6    52    28    43
 7     7    51    21    28
 8     8    26    39    15
 9     9    67    29    26
10    10    27    18    20
11    11    14    21    21
12    12    29    29    24
13    13    19    17    17
14    14    29    12    13
15    15    31    18    15
16    16    41    35    15
17    17    20    30    16
18    18    44    36    28
```

Take a look at the group means

```
1 warpbreaks %>%
2   summarize(mean_breaks = mean(breaks),
3             .by = tension)
```

```
  tension mean_breaks
1       L    36.38889
2       M    26.38889
3       H    21.66667
```

Model formulation

$$ln(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

```r
breaks_poisson <- glm(formula = breaks ~ tension, data = warpbreaks, family = "poisson")

breaks_poisson %>% summary
```

```
Call:
glm(formula = breaks ~ tension, family = "poisson", data = warpbreaks)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2464  -1.6031  -0.5872   1.2813   4.9366

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.59426    0.03907  91.988  < 2e-16 ***
tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 297.37  on 53  degrees of freedom
Residual deviance: 226.43  on 51  degrees of freedom
AIC: 507.09

Number of Fisher Scoring iterations: 4
```

Coef. interpretation

- For Low tension:

$$ln(\mu_i) = \beta_0$$

$$\mu_i = exp(\beta_0) = exp(3.59) = 36.38$$

- For Median tension:

$$ln(\mu_i) = \beta_0 + \beta_1$$

$$\mu_i = exp(\beta_0 + \beta_1) = exp(3.59) * exp(-0.32) = 26.38$$

- For High tension:

$$ln(\mu_i) = \beta_0 + \beta_2$$

$$\mu_i = exp(\beta_0 + \beta_2) = exp(3.59) * exp(-0.51) = 21.66$$

Output interpretation

Group mean prediction

- The predicted group means are the same as the ones we calculated based on data

R output

- **Intercept (3.59)**: The number of breaks for reference level (low tension) is $exp(3.59)$
- **tensionM (-0.32)**: The number of breaks for median tension is $exp(-0.32)$ **times** less than reference level (low tension)
- **tensionH (-0.51)**: The number of breaks for high tension is $exp(-0.51)$ **times** less than reference level (low tension)
- **Dispersion parameter (1)**: wonderful. Need to consider other methods if dispersion larger than 1 (over-dispersion) or smaller than 1 (under-dispersion)
- **AIC (507.09)**: Can be used to compare the goodness of fit between models

Model goodness of fit: Likelihood ratio test

```r
breaks_poisson_null <- glm(formula = breaks ~ 1,
                           data = warpbreaks,
                           family = "poisson")
```

H_0: The model performance is the same as a null model (making predictions by chance)

H_1: The model performance is significantly different comparing to a null model

- Use `lrtest()` function in the `lmtest` package

```r
lrtest(breaks_poisson, breaks_poisson_null)
```

```
Likelihood ratio test

Model 1: breaks ~ tension
Model 2: breaks ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -250.55
2   1 -286.02 -2 70.942  3.938e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model goodness of fit: Likelihood ratio test

```r
breaks_poisson_null <- glm(formula = breaks ~ 1,
                           data = warpbreaks,
                           family = "poisson")
```

H_0: The model performance is the same as a null model (making predictions by chance)

H_1: The model performance is significantly different comparing to a null model

- Or, Use `anova()` and specify `test = "Chisq"`

```r
anova(breaks_poisson, breaks_poisson_null, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: breaks ~ tension
Model 2: breaks ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        51     226.43
2        53     297.37 -2  -70.942 3.938e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictor significance: Likelihood ratio test

- Test whether adding one more variable `wool` (type of wool) could increase the model performance

H_0: The full model performance is the same as a reduced model (whichever model have fewer predictors)

H_1: The full model performance is significantly different comparing to a reduced model

```
breaks_poisson_add <- glm(formula = breaks ~ tension + wool,
                          data = warpbreaks,
                          family = "poisson")

lrtest(breaks_poisson, breaks_poisson_add)
```

```
Likelihood ratio test

Model 1: breaks ~ tension
Model 2: breaks ~ tension + wool
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -250.55
2   4 -242.53  1 16.039  6.206e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model prediction

- Use `predict()` and specify `type = "response"` to get back transformed y_i

```
1 warpbreaks_p <- warpbreaks %>%
2   mutate(breaks_p = predict(breaks_poisson_add, type = "response"))
3
4 warpbreaks_p
```

```
   breaks wool tension breaks_p
1      26    A       L 40.12354
2      30    A       L 40.12354
3      54    A       L 40.12354
4      25    A       L 40.12354
5      70    A       L 40.12354
6      52    A       L 40.12354
7      51    A       L 40.12354
8      26    A       L 40.12354
9      67    A       L 40.12354
10     18    A       M 29.09722
11     21    A       M 29.09722
12     29    A       M 29.09722
13     17    A       M 29.09722
14     12    A       M 29.09722
15     18    A       M 29.09722
16     35    A       M 29.09722
17     30    A       M 29.09722
18     36    A       M 29.09722
19     36    A       H 23.89035
20     21    A       H 23.89035
21     24    A       H 23.89035
22     18    A       H 23.89035
23     10    A       H 23.89035
```

```
24     43    A       H 23.89035
25     28    A       H 23.89035
26     15    A       H 23.89035
27     26    A       H 23.89035
28     27    B       L 32.65424
```

Model prediction

- Or, use `fitted()`, which provides back transformed y_i by default

```
1 warpbreaks_p <- warpbreaks %>%
2   mutate(breaks_p = fitted(breaks_poisson_add))
3
4 warpbreaks_p
```

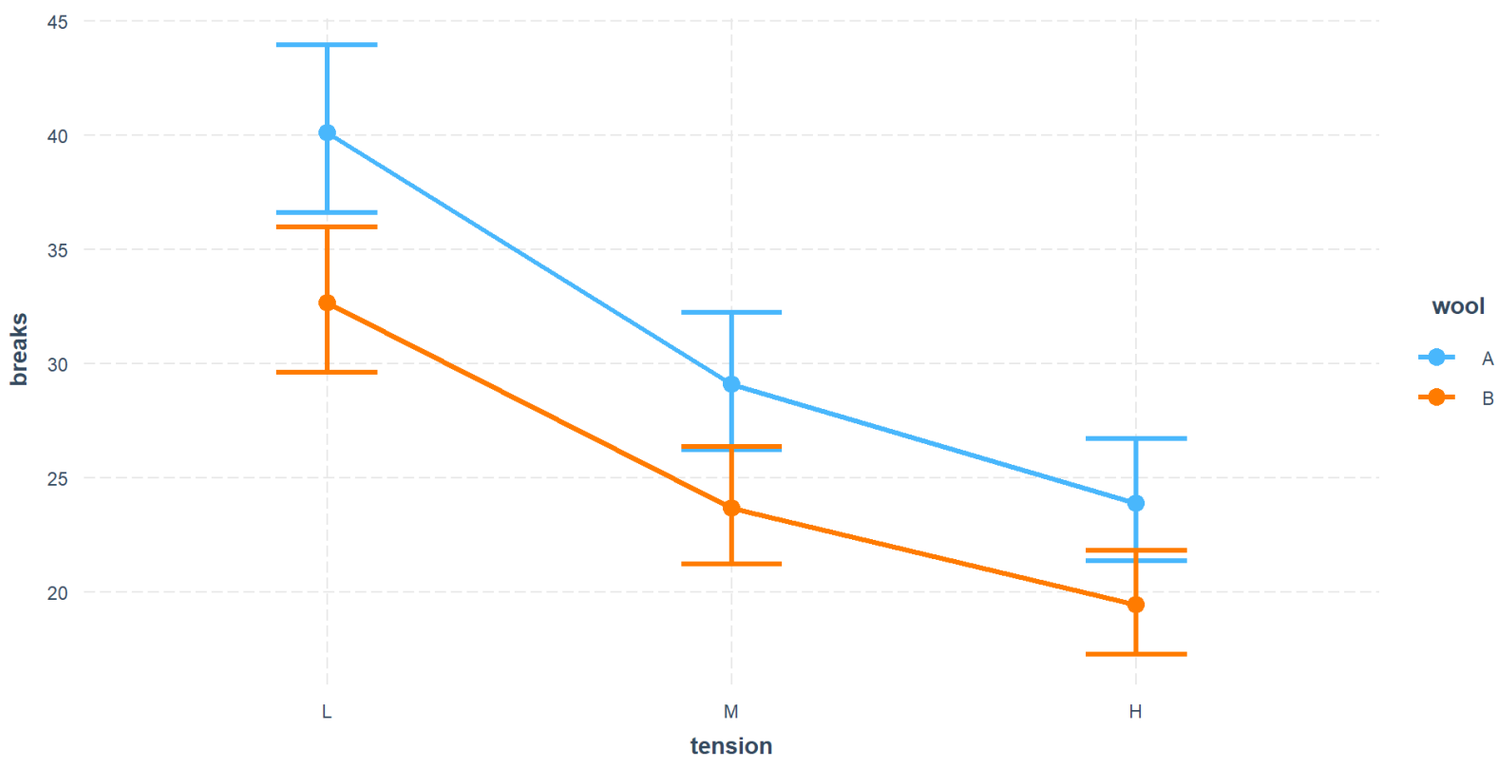
```
   breaks wool tension breaks_p
1      26    A       L 40.12354
2      30    A       L 40.12354
3      54    A       L 40.12354
4      25    A       L 40.12354
5      70    A       L 40.12354
6      52    A       L 40.12354
7      51    A       L 40.12354
8      26    A       L 40.12354
9      67    A       L 40.12354
10     18    A       M 29.09722
11     21    A       M 29.09722
12     29    A       M 29.09722
13     17    A       M 29.09722
14     12    A       M 29.09722
15     18    A       M 29.09722
16     35    A       M 29.09722
17     30    A       M 29.09722
18     36    A       M 29.09722
19     36    A       H 23.89035
20     21    A       H 23.89035
21     24    A       H 23.89035
22     18    A       H 23.89035
23     10    A       H 23.89035
```

```
24     43    A       H 23.89035
25     28    A       H 23.89035
26     15    A       H 23.89035
27     26    A       H 23.89035
28     27    B       L 32.65424
```

Model visualization

- Box plot: categorical predictor (y: count; x = tension)

```
cat_plot(breaks_poisson_add,
         pred = tension,
         modx = wool,
         geom = "line")
```



Research question (1 continuous x)

A survey was done for 915 Candian PhD students to investigate the relationship between number of article published during the PhD and the number of mentors they have.

```
    articles mentor
1          0      7
2          0      6
3          0      6
4          0      3
5          0     26
6          0      2
7          0      3
8          0      4
9          0      6
10         0      0
11         0     14
12         0     13
13         0      3
14         0      4
15         0      0
16         0      1
17         0      7
18         0     13
19         0      7
20         0      9
21         0      6
22         0      3
23         0      5
24         0      4
25         0      1
26         0      3
27         0      8
28         0      3
```

Take a look at the group means

```
PhDPublications %>%
  select(articles, mentor) %>%
  summarise(articles_mean = mean(articles), .by = me
  arrange(mentor) %>%
  head()
```

```
  mentor articles_mean
1      0     0.9777778
2      1     0.8846154
3      2     1.0000000
4      3     1.3857143
5      4     1.6250000
6      5     1.6363636
```

Model formulation

$$ln(\mu_i) = \beta_0 + \beta_1 x_{1i}$$

```r
articles_poisson <- glm(formula = articles ~ mentor,
                        data = PhDPublications,
                        family = "poisson")

articles_poisson %>% summary()
```

```
Call:
glm(formula = articles ~ mentor, family = "poisson", data = PhDPublications)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5700  -1.6316  -0.3598   0.5068   5.9483

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.259906   0.034361   7.564 3.91e-14 ***
mentor      0.026050   0.001917  13.586  < 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: 1669.5  on 913  degrees of freedom
AIC: 3341.3
```

```
Number of Fisher Scoring iterations: 5
```

Coef. interpretation

- For PhD students with no mentor:

$$ln(\mu_i) = \beta_0$$

$$\mu_i = exp(\beta_0) = exp(0.25) = 1.29$$

- For PhD students having one mentor:

$$ln(\mu_i) = \beta_0 + \beta_1$$

$$\mu_i = exp(\beta_0 + \beta_1) = exp(0.25) * exp(0.02) = 1.33$$

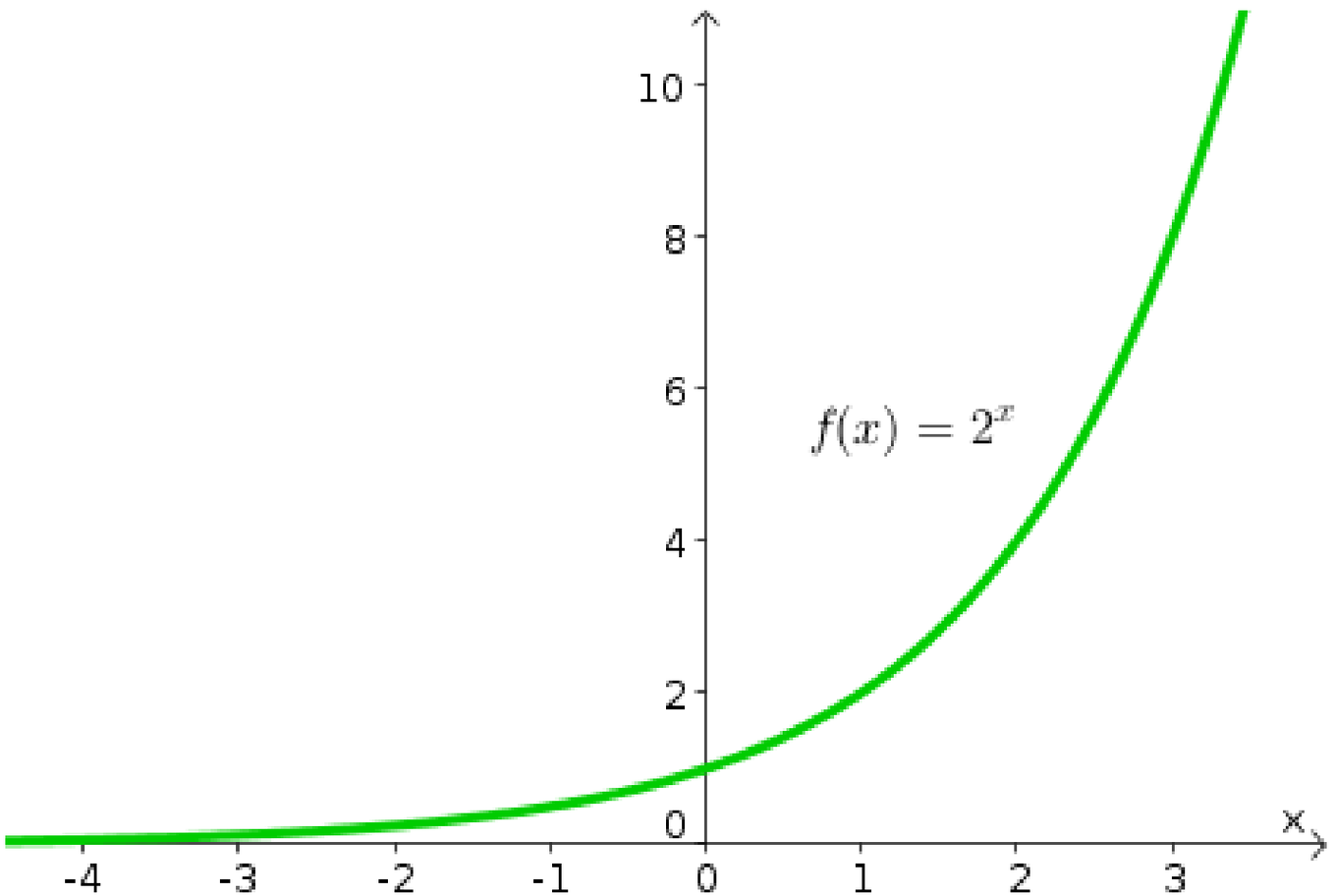
- For PhD students having more than one mentor(s):

$$ln(\mu_i) = \beta_0 + \beta_1 x_{1i}$$

$$\mu_i = exp(\beta_0 + \beta_1 x_{1i}) = exp(0.25) * exp(0.02 x_{1i})$$

Coef. interpretation

- If x is positive, then $exp(x)$ is larger than 1
- If x is positive, then $exp(x)$ is smaller than 1



Output interpretation

Mean prediction

- If the variable coefficient is greater than 0 -> the counts gets higher as the variable increases, vice versa

R output

- **Intercept (0.25)**: The number of articles as baseline (no mentor) is $exp(0.25)$
- **mentor (0.02)**: The number of articles with one mentor is $exp(0.02)$ **times** more than baseline
- **Dispersion parameter (1)**: wonderful. Need to consider other methods if dispersion larger than 1 (over-dispersion) or smaller than 1 (under-dispersion)
- **AIC (3341.3)**: Can be used to compare the goodness of fit between models

Model goodness of fit: Likelihood ratio test

```r
articles_poisson_null <- glm(formula = articles ~ 1,
                             data = PhDPublications,
                             family = "poisson")
```

H_0: The model performance is the same as a null model (making predictions by chance)

H_1: The model performance is significantly different comparing to a null model

- Use `lrtest()` function in the `lmtest` package

```r
lrtest(articles_poisson, articles_poisson_null)
```

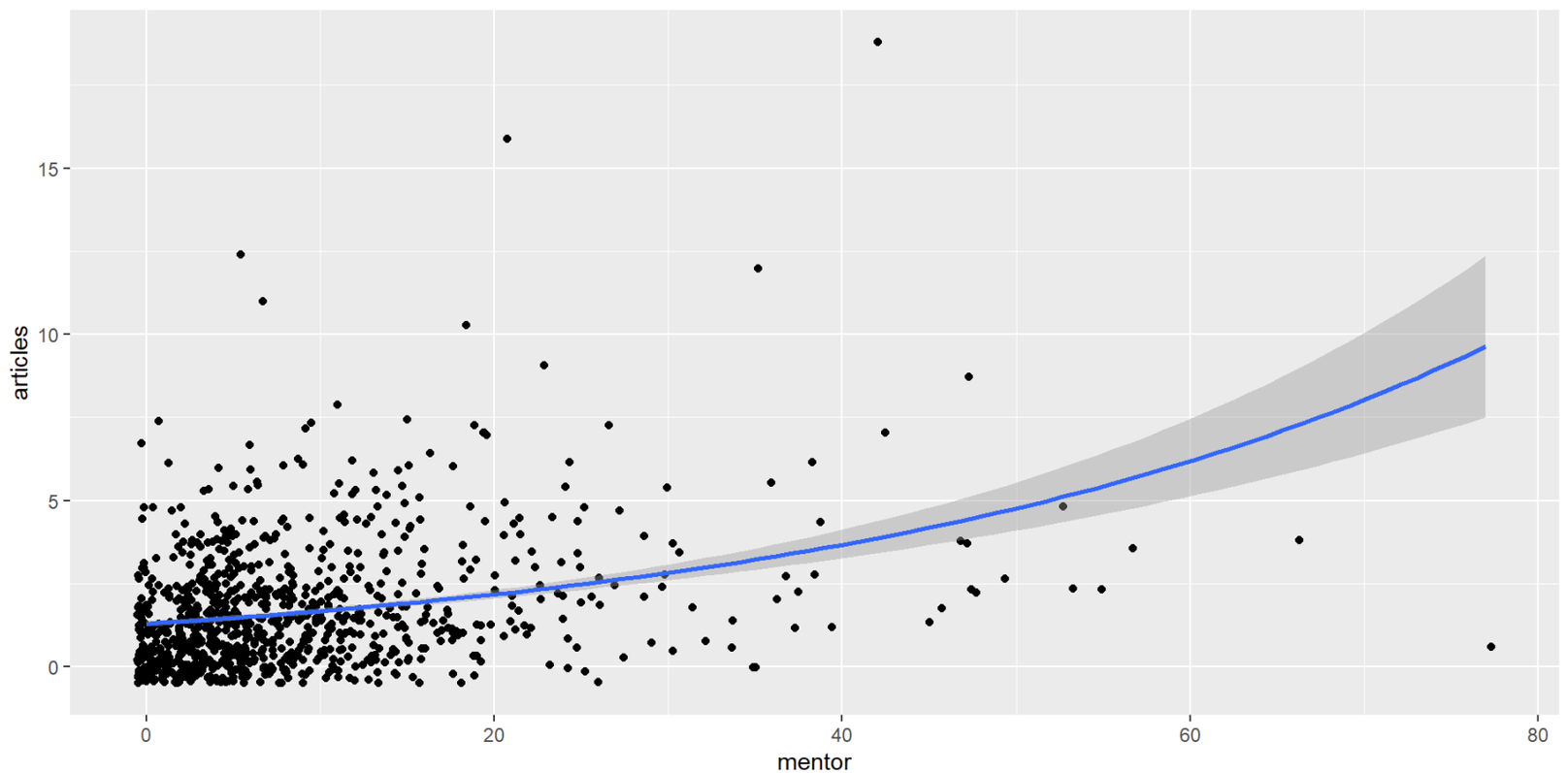
```
Likelihood ratio test

Model 1: articles ~ mentor
Model 2: articles ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -1668.6
2   1 -1742.6 -1 147.86  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model visualization

- Smooth: continuous (y: count; x = continuous)

```
ggplot(aes(x = mentor, y = articles), data = PhDPublications) +
  geom_jitter(width = 0.5, height = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "poisson"))
```



What we learned today

- Count data follows Poisson distribution
- Poisson uses log link as the most common link function
- Coefficient interpretation on Poisson regression when there is 1 variable
- Test goodness of fit using Likelihood Ratio Test
- Poisson regression model prediction
- Poisson regression model visualization

Wrap up

Before we meet again

- Review intro to GLM and Poisson regression
- Enjoy weekend!

Next time

- Next Tuesday 12:30 virtual lecture