Count data and the Poisson distribution

Frank Edwards

Count data

- Counts are cumulative totals of the number of incidences of some event, generally across time or place

Count data

- Counts are cumulative totals of the number of incidences of some event, generally across time or place
- Counts are positive integers $\in [0, \infty]$

Count data

- Counts are cumulative totals of the number of incidences of some event, generally across time or place
- Counts are positive integers $\in [0, \infty]$

Counts as extensions of binary data

- Counts can be thought of as repeated binary trials
- $\sum y_i$ where y is equal to 1 or 0 provides a count
- Generally, we could treat `sum(y==1) + sum(y==0)` or `nrow(y)` as the exposure, or denominator for a rate. Why?

An example of count data

```
load("./data/fieldplayer_overall_season_stats.rda")
load("./data/player.rda")

nwsl_stats<-fieldplayer_overall_season_stats
nwsl_players<-player

head(nwsl_players)
```

```
## # A tibble: 6 x 5
##   person_id player          nation pos   name_other
##       <dbl> <chr>           <chr>  <chr> <chr>
## 1       342 Marisa Abegg    USA    DF    <NA>
## 2       117 Danesha Adams   USA    FW,MF <NA>
## 3         6 Adriana         ESP    FW    <NA>
## 4       300 Leigh Ann Brown USA    DF,MF <NA>
## 5       202 Jazmyne Avant   USA    DF    <NA>
## 6        28 Amy Barczuk     USA    DF    <NA>
```

make a joined table with players names

```r
### attaching names
dat<-nwsl_stats %>%
  left_join(nwsl_players %>%
              select(person_id, player))

### check to ensure that the dimensions are what we want
nrow(dat) == nrow(nwsl_stats)
```

```
## [1] TRUE
```

```r
### if we have multiple positions for players
dat<-nwsl_stats %>%
  left_join(nwsl_players)
```

Approaches to modeling count data

The Poisson model

Where y is a non-negative integer (count)

$$y \sim Poisson(\lambda)$$

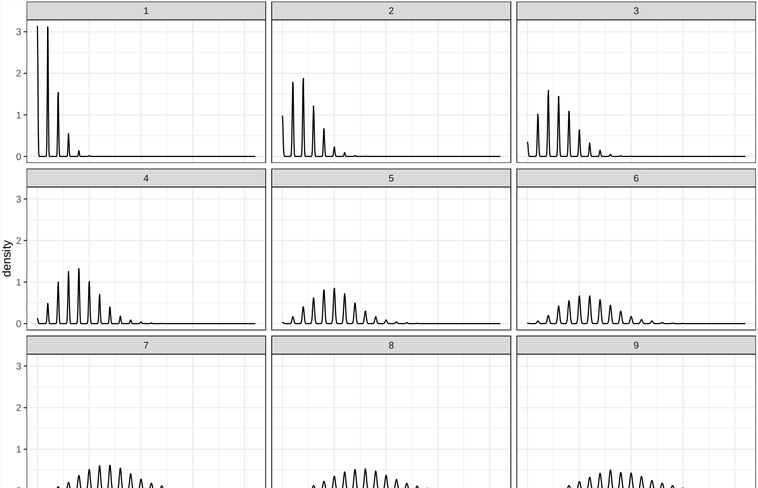
$$E(y) = \bar{y} = \lambda$$

$$Var(y) = \lambda$$

$$Pr(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Shape of the Poisson distribution

```
ggplot(pois_demo, aes(x=count)) +
  geom_density(adjust = 1/4) +
  facet_wrap(~lambda)
```



Let’s look at each Poisson variable

```
pois_demo%>%group_by(lambda)%>%

  summarise(mean = mean(count),
            variance = var(count))
```

```
## # A tibble: 9 x 3
##   lambda  mean variance
## *  <int> <dbl>    <dbl>
## 1      1  1.01     1.01
## 2      2  2.00     1.99
## 3      3  2.97     2.94
## 4      4  4.00     3.99
## 5      5  5.02     5.04
## 6      6  6.02     6.27
## 7      7  7.00     6.93
## 8      8  7.98     8.11
## 9      9  8.98     9.20
```

Poisson models as a GLM

For a count variable y, we can specify a Poisson GLM with a log link function

$$y \sim Poisson(\lambda)$$

$$\lambda = e^{\beta X} = e^{\beta_0 + \beta_1 x_1 \cdots \beta_n x_n}$$

$$E(y|x) = e^{\lambda}$$

$$log(E(y|x)) = \lambda = \beta X$$

Model NWSL data as a Poisson GLM

```r
m0_mp<-glm(mp ~
             1,
           data = dat,
           family = "poisson")

### check to see if the data looks similar to the poisson
data_sim<-data.frame(
  y_pred = rpois(nrow(dat), 13.54815),
  mp = dat$mp
)

## evaluate the distributions
ggplot(data_sim,
       aes(x = y_pred)) +
  geom_histogram(fill = "red", alpha = 0.5) +
  geom_histogram(aes(x = mp), fill = "blue", alpha = 0.5)
```

Modeling assists

```
assist_0<-glm(ast ~ 1,
              data = dat,
              family = "poisson")

sim_dat<-data.frame(
  y_pred = rpois(nrow(dat), 0.9782119),
  ast = dat$ast
)

ggplot(sim_dat,
       aes(x = y_pred)) +
  geom_histogram(fill = "red", alpha =0.5) +
  geom_histogram(aes(x = ast), fill = "blue", alpha = 0.5)
```

let’s generate predictions for players

```
### take our two predictors
### define reasonable ranges to predict over
mp<-1:24
gls<-0:5
### generate expected values
fake_data<- expand_grid(mp, gls)

fake_data<-fake_data%>%
  mutate(expected_assists =
           predict(assist_3, fake_data,
                   type = "response"))

ggplot(fake_data,
       aes(x = mp, y = expected_assists,
           color = factor(gls))) +
  geom_line()
```

Advantages of the Poisson distribution for regression

1. Constrained to non-negative integers
2. Variance scales with the expectation of y
3. Relatively simple to interpret

However:

$$\lambda = E(y|x) = var(y)$$

Homework

1. Visualize the distribution of goals across players for the 2019 season (your choice on geom)
2. Define a linear predictor for goals made during a season, where the players' position is the only predictor.
3. Estimate this model with a Normal likelihood (OLS)
4. Estimate this model with a Poisson likelihood (family = "poisson")
5. Generate predictions for each position for both models
6. Compare the predictions. Which model do you prefer? Why?