Poisson Regression with offset

9/14/2021

Learning Goals: - Identify when an offset is needed in a Poisson regression model and identify the appropriate offset variable - Fit a Poisson regression model with offset in R - Interpret intercept and slope coefficients for a Poisson regression model with offset - Describe how coefficient estimates will or will not change if different units are used for the offset

Complete # 1 - 7 before class.

Be sure you have read Chapter 4.5 - 4.7, including the Campus Crime case study. In the book example, we use an offset to model the number of crimes per 1000 students, rather than the number of crimes.

Today's dataset is Bald Eagle count data collected from the year 1981 to 2017, in late December, by birdwatchers in the Ontario, Canada area. The data was made available by the Bird Studies Canada website and distributed through the R for Data Science TidyTuesday project.

year - year of data collection

count - number of birds observed

hours - total person-hours of observation period

count_per_hour - count divided by hours

count_per_week - count_per_hour multiplied by 168 hours per week

Source https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-06-18/bird_counts.csv

1. We are interested in how the Bald Eagle population has changed over time. Formulate this into a more specific question that can be addressed with these data. *Is the number of eagles observed (count) related to year?*

2. Read in the data from `Class > Data > bald_eagles.csv`. How many rows? How many columns?

```
library(tidyverse)
bald_eagles <- read_csv("~/Stats_316_F24/Class/Data/bald_eagles.csv")
```

3. Look at the top and bottom of the data using the `head` and `tail` functions. Identify the observational units, response variable, and explanatory variable to answer the question you formulated in #1.

```
head(bald_eagles)
```

```
## # A tibble: 6 x 5
##    year count hours count_per_hour count_per_week
##   <dbl> <dbl> <dbl>          <dbl>          <dbl>
## 1  1981     0   167        0                  0
## 2  1982     0   164        0                  0
## 3  1983     0   168        0                  0
## 4  1984     1   178        0.00562            0.944
## 5  1985     2   143        0.0140             2.35
## 6  1986     1   182        0.00549            0.923
```

```
tail(bald_eagles)
```

```
## # A tibble: 6 x 5
##    year count hours count_per_hour count_per_week
##   <dbl> <dbl> <dbl>          <dbl>          <dbl>
## 1  2012     3  195.         0.0154           2.59
## 2  2013    11  182.         0.0604          10.1
## 3  2014     6  179.         0.0335           5.62
## 4  2015     7  221          0.0317           5.32
## 5  2016     3  217.         0.0138           2.33
## 6  2017    12  199.         0.0604          10.1
```

- Observational units: Each year
- Response: Count (number of eagles)
- Explanatory: year

4. Calculate summary statistics and make plots that show the distribution for `count`. Describe the shape of this distribution.

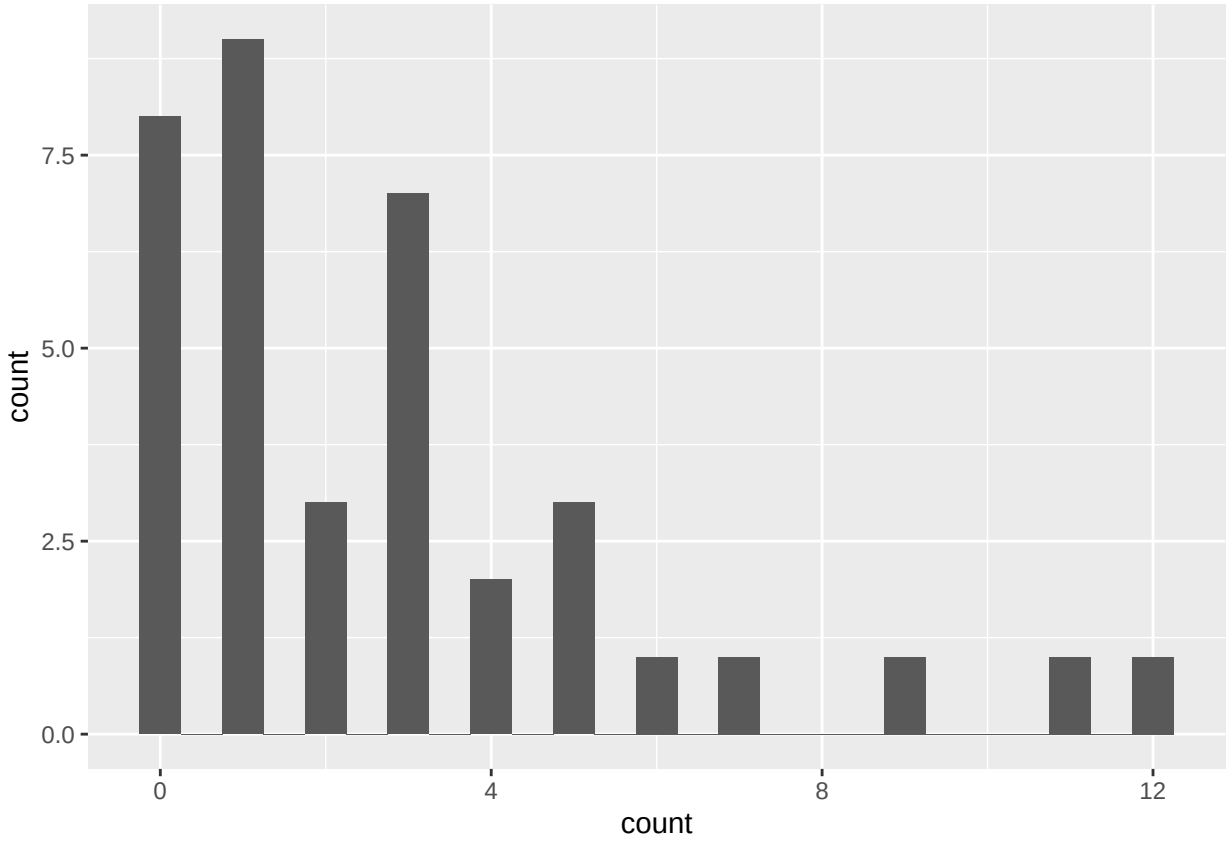
```
summary(bald_eagles$count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.811   4.000  12.000
```

```
sd(bald_eagles$count)
```

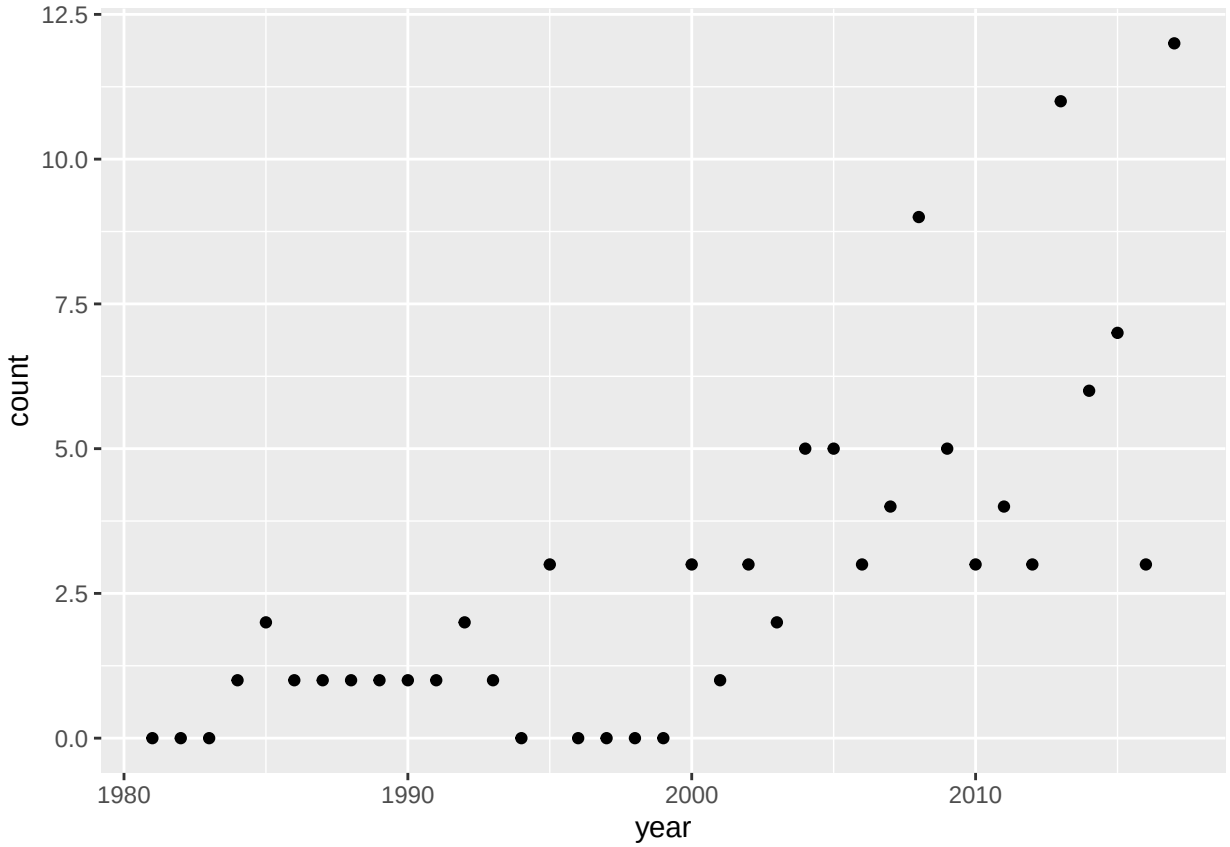
```
## [1] 3.026162
```

```
ggplot(bald_eagles, aes(x = count)) +
  geom_histogram(binwidth = 0.5)
```



5. Make a plot that show the relationship between `count` and `year`. What does this tell you in relation to your main question?

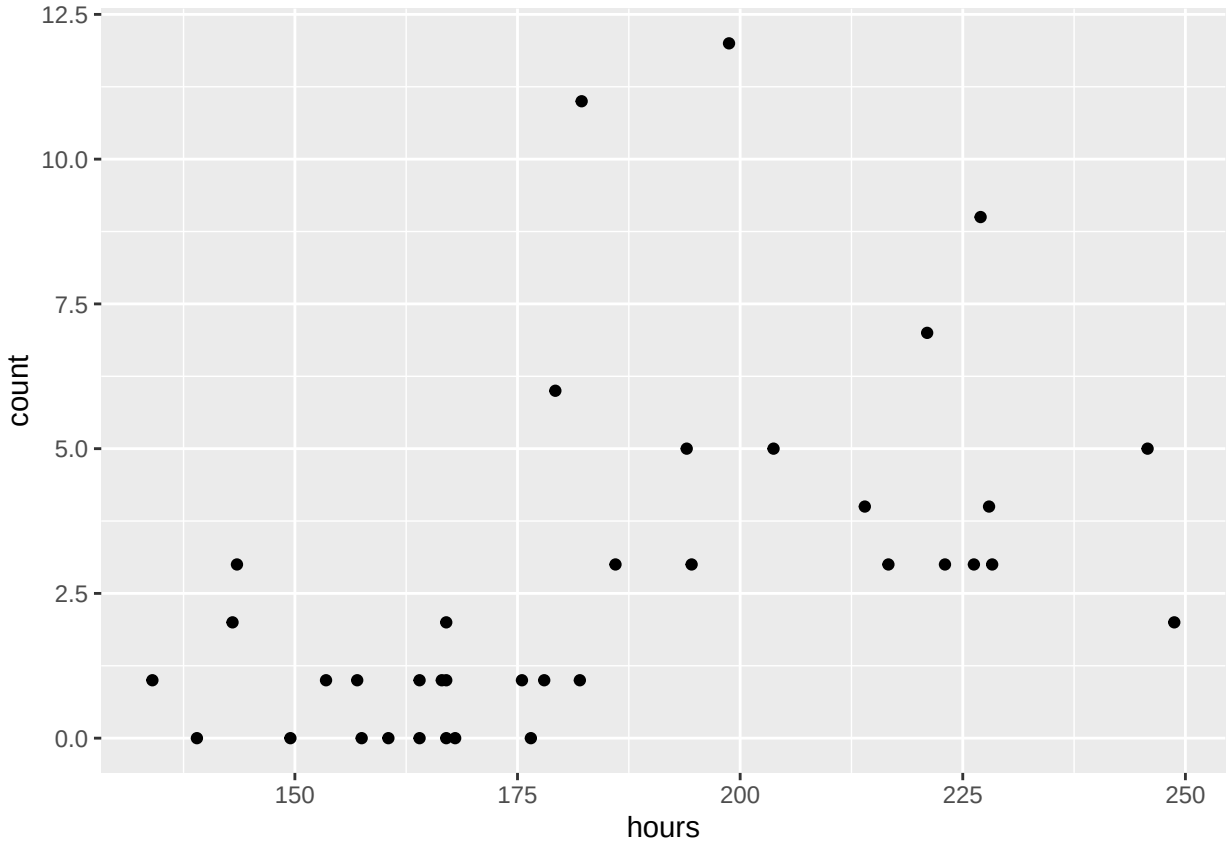
```
ggplot(bald_eagles, aes(x = year, y = count)) +
  geom_point()
```



Count seems to be increasing over time

6. How would you expect the `hours` variable to be related to the count? Make a plot of `count` and `hours` to confirm.

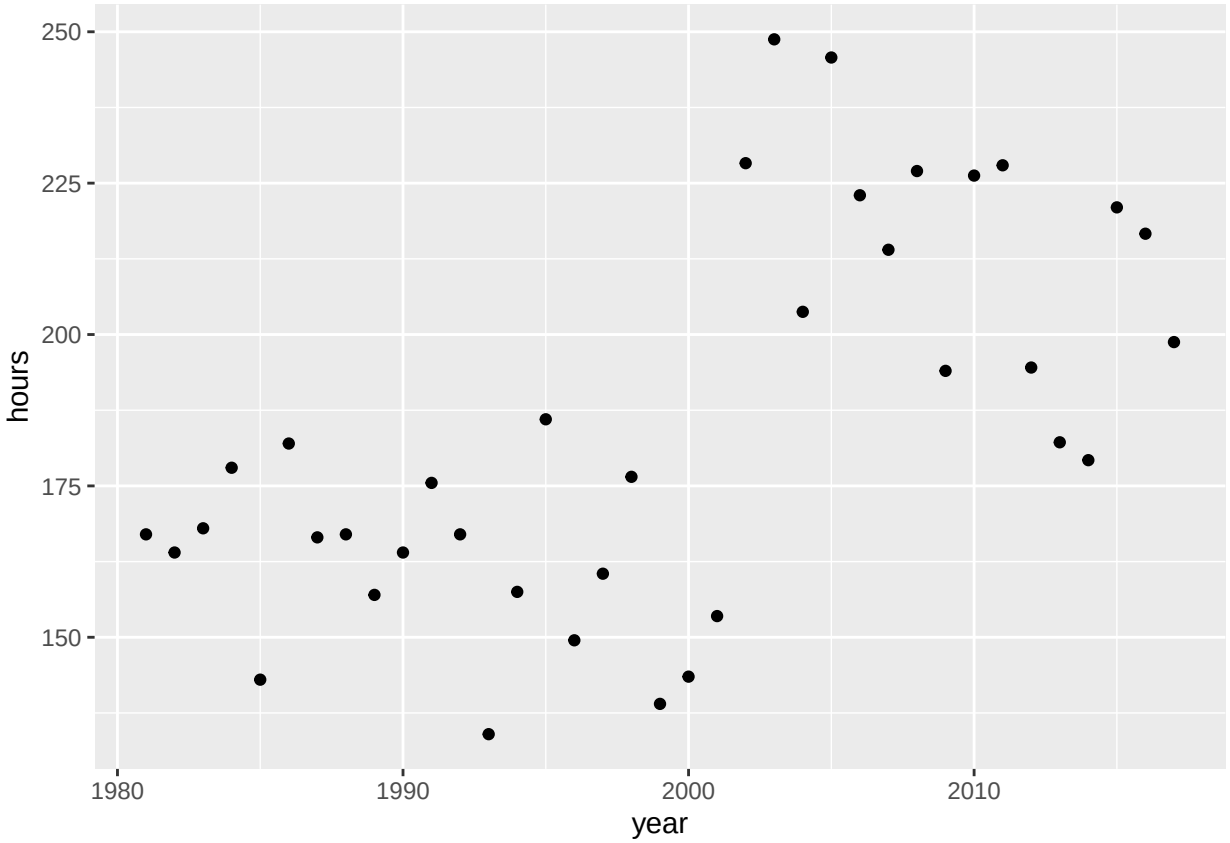
```
ggplot(bald_eagles, aes(x = hours, y = count)) +
  geom_point()
```



More eagles are counted when more time is spent observing.

7. Make a plot of `year` and `hours` to determine whether there is also a trend in observation hours over time. What problesm might this lead to in your analysis?

```
ggplot(bald_eagles, aes(x = year, y = hours)) +
  geom_point()
```



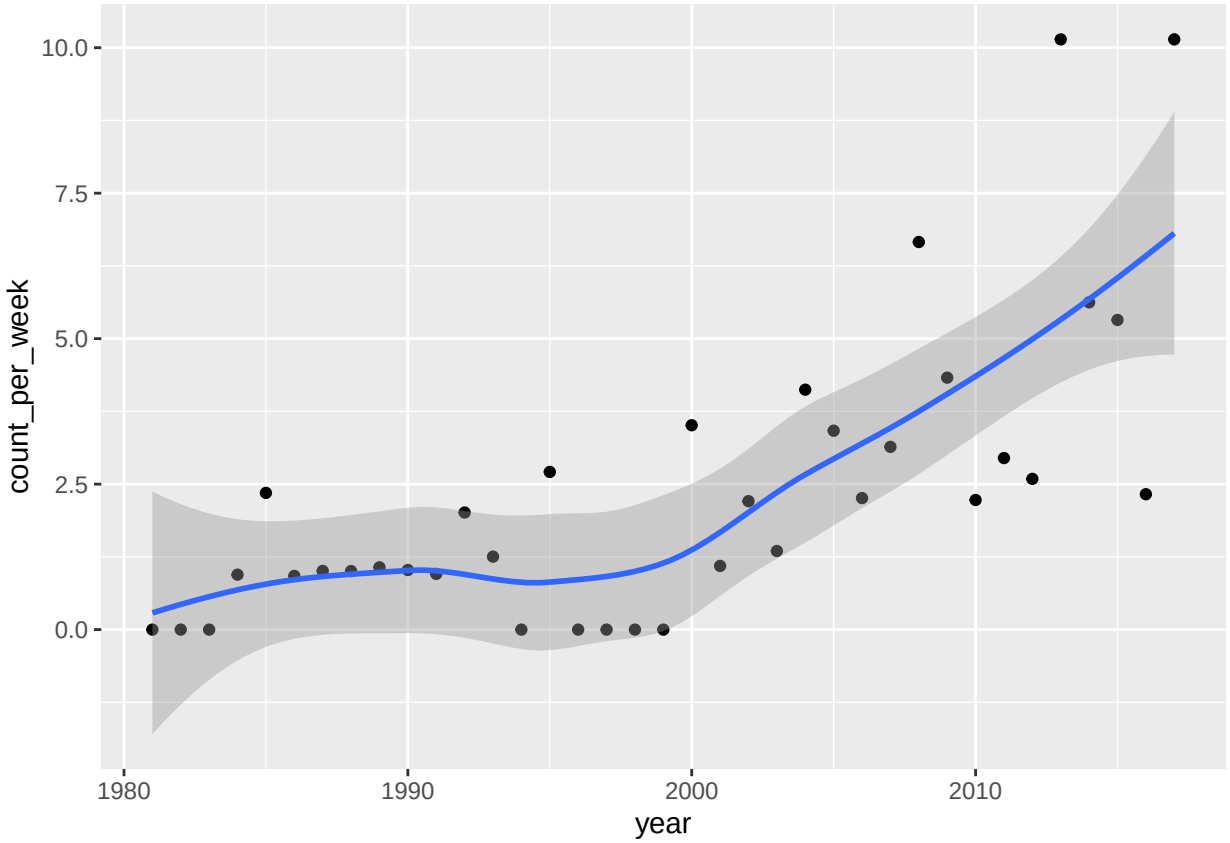
This tells that we need to account for differences in hours spend observing, especially since this trends upward in time

We'll start here in class.

8. The plots above suggest that it will be important to account for how many person-hours were spent in observing birds. What do the plots below tell you about the suitability of Poisson regression?

```
ggplot(bald_eagles, aes(x = year, y = count_per_week)) +
  geom_point() +
  geom_smooth()
```

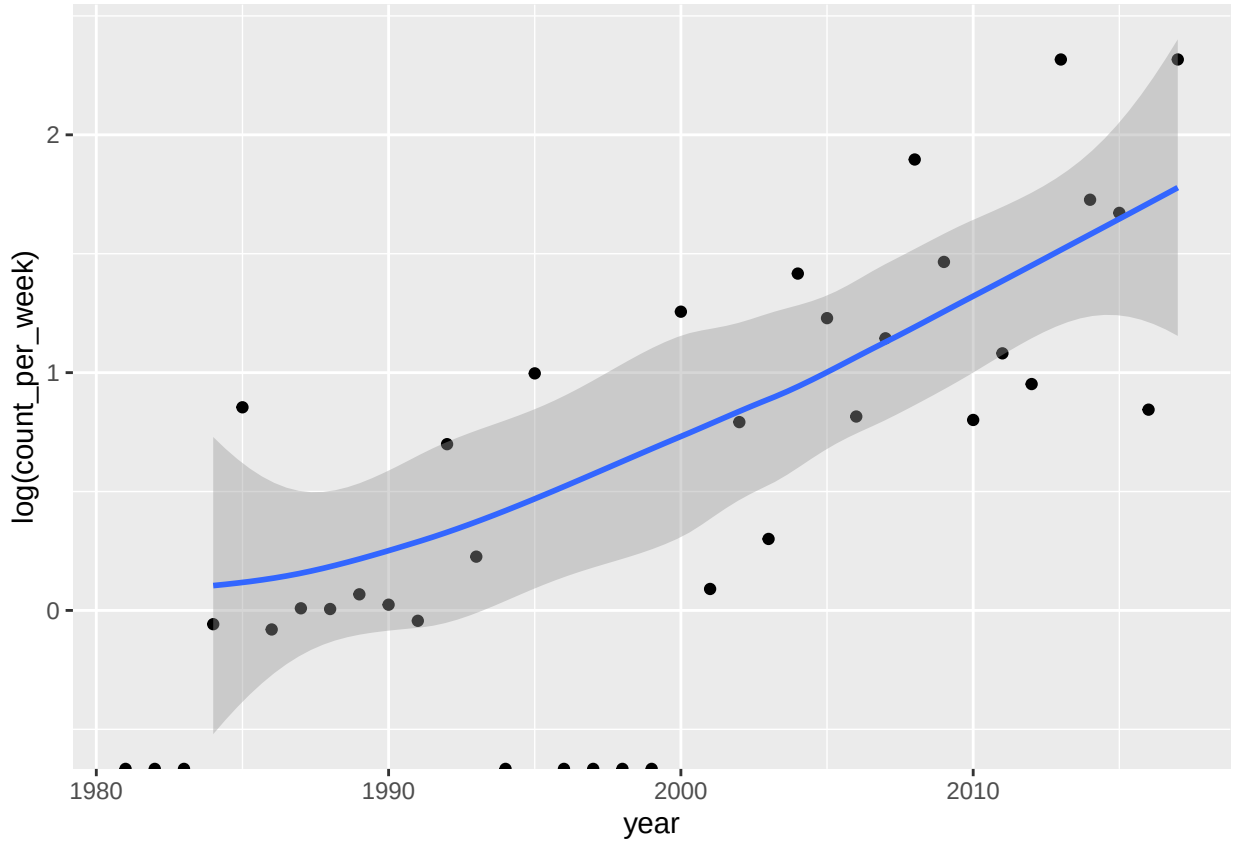
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
ggplot(bald_eagles, aes(x = year, y = log(count_per_week))) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



The relationship is log-linear, so that is a good sign!

Introducing the Poisson model with offset

The Poisson distribution is for discrete (whole number) values, therefore we can't just use `count_per_week` directly as the response in a Poisson regression model because this variable is not discrete (whole number) values. Instead, we consider a model with an *offset*.

To review, here is the standard Poisson regression model, in which Y is the observed count (number of eagles) λ represents the expected count. We assume the log expected count is related to our explanatory variable, X.

$$Y \sim Poisson(\lambda)$$
$$log(\lambda) = \beta_0 + \beta_1 X$$

In a Poisson regression model, we assume that the expected number of eagles, λ is related to the explanatory variables. When we have an offset term, we assume that the *rate* of spotting eagles per week is related to the explanatory variables. To the expected rate of eagles per week is λ divided by the number of weeks of observation. So for this model,

$$log(\frac{\lambda}{weeks}) = \beta_0 + \beta_1 X$$

IMPORTANT!!!! This means our interpretation of β_0 and β_1 should be in *eagles per week* rather than *number of eagles*.

Note that we can rewrite this equation and move weeks to the right hand side:

$$log(\lambda) = \beta_0 + \beta_1 X + log(weeks)$$

Notice then that the expected number of eagles λ depends on the number of weeks, but this does not add any parameters to the model (no extra βs).

Fit Poisson model with offset

The bald_eagles dataset does not have a column for "weeks". Let's see what happens if we use log(hours) as the offset instead:

```
bald_eagles
```

```
## # A tibble: 37 x 5
##     year count hours count_per_hour count_per_week
##    <dbl> <dbl> <dbl>          <dbl>          <dbl>
##  1  1981     0  167         0                0
##  2  1982     0  164         0                0
##  3  1983     0  168         0                0
##  4  1984     1  178         0.00562          0.944
##  5  1985     2  143         0.0140           2.35
##  6  1986     1  182         0.00549          0.923
##  7  1987     1  166.        0.00601          1.01
##  8  1988     1  167         0.00599          1.01
##  9  1989     1  157         0.00637          1.07
## 10  1990     1  164         0.00610          1.02
## # i 27 more rows
```

```
eagle.glm <- glm(count ~ year, data = bald_eagles,
                 family = poisson, offset = log(hours))
summary(eagle.glm)
```

```
##
## Call:
## glm(formula = count ~ year, family = poisson, data = bald_eagles,
##     offset = log(hours))
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -155.81571   23.20053  -6.716 1.87e-11 ***
## year           0.07566    0.01156   6.547 5.88e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 95.026  on 36  degrees of freedom
## Residual deviance: 42.390  on 35  degrees of freedom
## AIC: 130.25
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(eagle.glm))
```

```
##  (Intercept)         year
## 2.138438e-68 1.078601e+00
```

9. Interpret the coefficients from the Poisson regression model with offset above.

Intercept is not very interpretable, but would be the number of eagles per hour in the year 0.

The number of eagles observed per hour increases by 7.8% each year.

10. Notice that the intercept value is really tiny... "Eagles per hour" might be a difficult scale on which to interpret the intercept. There are 24*7 = 168 hours per week. Create a new variable that is "weeks" and use this as the offset. Interpret your regression coefficients. Which changed? Which stayed the same?

```
bald_eagles <- bald_eagles %>%
  mutate(weeks = hours/168)

eagle.glm.weeks <- glm(count ~ year, data = bald_eagles,
                   family = poisson, offset = log(weeks))

coef(eagle.glm.weeks) %>% exp()
```

```
##  (Intercept)         year
## 3.592576e-66 1.078601e+00
```

Intercept changed, because now it is the number of eagles per WEEK, but still in year 0, so still silly.

The slope is the same value... it's interpertaion is that the number of eagles observed per week increases by 7.8% each year. Since it is a multiplicative (or ratio, percent, etc) change it doesn't matter if we consider per hour or per week, the value of the increase is the same

11. The intercept is still not very interpretable here because year starts at 1981. How might you make the intercept a more interpretable value? Use mutate, fit the model and interpret the coefficients. Which changed? Which stayed the same?

```
bald_eagles <- bald_eagles %>%
  mutate(year1980 = year-1980)

eagle.glm.weeks1980 <- glm(count ~ year1980, data = bald_eagles,
                  family = poisson, offset = log(hours))

coef(eagle.glm.weeks1980) %>% exp()
```

```
## (Intercept)    year1980
## 0.002480389 1.078601003
```

The average number of bald eagles observed in 1980 is 0.4 per week.

The number observed per week increases by 7.8% each year.

The slope stayed the same again, but the intercept changed.

Conceptual questions

12. Consider each of the following examples. Imagine what data you would collect to answer the question (e.g. what is your observational unit, response variable, and explanatory variable). Also describe if an offset is needed.

a. Are the number of motorcycle deaths in a given year related to a state's helmet laws?

observational unit: State [possibly also across multiple years]. Repsponse: Number of deaths. Explanatory: Law status (e.g. required to wear helmet or not). Need an offset for state pouplation size

b. Does the daily number of asthma-related visits to an Emergency Room differ depending on air pollution indices?

THere are several ways to think about this: *observational unit: Day; Response: Asthma ER visits, Explanatory: Air pollution; No offset needed* OR, if you are at different emergency rooms: *observational unit: ER (and possibly day), Response: Asthma ER visits, Explanatory: Air pollution; we would need an offset for the size of the population served by the particular ER department.*

c. In a drug treatment program, does the number of relapses within five years of initial treatment depend upon a patient's mental health screening score at the beginning of their program?

Obs unit: patient; Response: Number of relapses; Explanatory: Initial MH sccore; Offset not needed because all patients are observed for 5 years&

d. Has the number of deformed fish in randomly selected Minnesota lakes been affected by changes in trace minerals in the water over the last decade?

Again, this is vague... *Obs unit: Lake and year, Response: number of deformed fish; Explan: mineral content. Would need an offset for either number of fish caught, or estimated population of fish... or perhaps size of lake as a proxy for population size?*

More time, More practice.

Check out the NYCairbnb data described in Open Ended Exercises. https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html#open-ended-exercises-2

```
NYCairbnb <- read_csv("~/Stats_316_F24/Class/Data/NYCairbnb.csv")
```

```
## Rows: 40628 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): last_scraped, host_since, room_type
## dbl (9): id, days, bathrooms, bedrooms, price, number_of_reviews, review_sco...
## lgl (1): instant_bookable
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
NYCairbnb
```

```
## # A tibble: 40,628 x 13
##       id  days last_scraped host_since room_type       bathrooms bedrooms price
##    <dbl> <dbl> <chr>        <chr>      <chr>                <dbl>    <dbl> <dbl>
##  1  2515  3130 4/2/2017     9/6/2008   Private room             1        1    59
##  2  2595  3127 4/2/2017     9/9/2008   Entire home/apt          1        0   230
##  3  3647  3050 4/2/2017     11/25/2008 Private room             1        1   150
##  4  3831  3038 4/2/2017     12/7/2008  Entire home/apt          1        1    89
##  5  4611  3012 4/2/2017     1/2/2009   Private room            NA        1    39
##  6  5099  2981 4/2/2017     2/2/2009   Entire home/apt          1        1   212
##  7  5107  2981 4/2/2017     2/2/2009   Entire home/apt          1        2   250
##  8  5121  2980 4/2/2017     2/3/2009   Private room            NA        1    60
##  9  5172  2980 4/2/2017     2/3/2009   Entire home/apt          1        1   129
## 10  5178  2952 4/2/2017     3/3/2009   Private room             1        1    79
## # i 40,618 more rows
## # i 5 more variables: number_of_reviews <dbl>, review_scores_cleanliness <dbl>,
```

```
## #   review_scores_location <dbl>, review_scores_value <dbl>,
## #   instant_bookable <lgl>
```

Create a model with Number of reviews as the response. What is an appropriate offset? (You might have to create this variable!) Do some EDA first to see which variables might be related to our response! Some particular variables of interest might be the price (or find price/bedroom), review_scores_value, and instant_bookable.

For this one you might consider creating a variable that is months the room has been on airbnb and use this as an offset.