Poisson Regression

Adam J Sullivan, PhD

03/07/2018

- Another situation where we can use a GLM is with Poisson data.
- The poisson model fits data where
  - Response is a count that follows a Poisson distribution.
  - If the events are recurrent than the probability of a 2^{nd} event must not have an increase over the probability of the 1^{st} event. (This would fail for say blood clots.)
  - Incidence rates remain constant over time.
  - Incidence rate multiplied by exposure gives the expected number of events.
  - Over a very small exposure time t the probability of more than one event happening is 0.

Poisson Regression

Poisson Regression Strengths

- Poisson regression generalizes crude and stratified incidence rates.
- Does not require that subjects are followed for the same amount of time
  - This can be a weakness of logistic regression when we assume that subjects are followed for the same amount of time given that most studies never achieve this.

Link Function

- With Poisson Regression it can be shown that we are interested in

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- This would show us we would be using the $\log$ link.

Example

- Unlike Logistic Regression where we took more time to go through the math the benefit of using a GLM is that we no longer need to discuss how to estimate our coefficients.
- We once again will by using maximum likelihood theory.
- We will instead discuss Poisson Regression through example.

Recall the Colorectal Cancer component of the Physicians Health Study

NAME	DESCRIPTION
age	Age in years at time of Randomization
asa	0 - placebo, 1 - aspirin
bmi	Body Mass Index (kg/m^2)
hypert	1 - Hypertensive at baseline, 0 - Not
alcohol	0 - less than monthly, 1 - monthly to less than daily, 2 - daily consumption

NAME	DESCRIPTION
dm	0 = No diabetes Mellitus, 1 - diabetes Mellitus
sbp	Systolic BP (mmHg)
exer	0 - No regular, 1 - Sweat at least once per week
csmoke	0 - Not currently, 1 - $<$ 1 pack per day, 2 - $\geq$ 1 pack per day
psmoke	0 - never smoked, 1 - former < 1 pack per day, 2 - former $\geq$ 1 pack per day
pkyrs	Total lifetime packs of cigarettes smoked
crc	0 - No colorectal Cancer, 1 - Colorectal cancer
cayrs	Years to colorectal cancer, or death, or end of follow-up.

What Each Subject Contributed

1. Information on whether of not they had a Colorectal Cancer(CRC) during follow-up
2. Follow-up time in years, specified as time from randomization until first of
   - end of Study
   - death
   - Colorectal Cancer
   - Loss to follow-up

Loading Data

```
library(tidyverse)
library(haven)
phscrc <- read_dta("phscrc.dta")
phscrc <- phscrc %>% mutate(age.cat = cut(age, c(40, 50, 60,
    70, 90), right = FALSE)) %>% mutate(alcohol.use = factor(phscrc$alcohol >
    0, labels = c("no", "yes")))
```

Alcohol Use by Age

We then can consider the following table of information

	ALCOHOL USERS	NON-ALCOHOL USERS
Ages	$\frac{\text{Events(MI)}}{\text{Person-Years}}$	$\frac{\text{Events(MI)}}{\text{Person-Years}}$
40-49	$\frac{8}{69.723} = 0.1147$	$\frac{31}{208.093} = 0.1490$
50-59	$\frac{21}{172.485} = 0.1217$	$\frac{59}{426.540} = 0.1383$
60-69	$\frac{32}{233.063} = 0.1373$	$\frac{62}{410.415} = 0.1511$
70+	$\frac{20}{121.789} = 0.1642$	$\frac{21}{129.177} = 0.1626$
Total	$\frac{81}{597.060} = 0.13566$	$\frac{173}{1174.225} = 0.1473$

Reasoning for Poisson Regression

- Before we continue we will discuss why we may use Poisson regression here rather than logistic.
  - Poisson regression is used to model expected number of events given covariates.
  - We can use either categorical or continuous covariates.
  - The number of events for each subject is independent from subject to subject and each subject has a distribution:

$$Y_i \approx Poisson(\mu_i = \lambda_i t_i)$$

    - This incidence rate (λ_i) of CRC is constant over time but may vary individually based on covariates for subject i.

What do we have?

- For i^{th} subject we have a follow up of t_i years $i = 1, \ldots, 22071$.

$$Y_i = \begin{cases} 1 & \text{if patient develops CRC} \\ 0 & \text{Otherwise} \end{cases}$$

- Y_i is not binomial since t_i is different for every subject.
- We can then assume that Y_i is Poisson

$$\begin{aligned} E(Y_i) &= \mu_i = \lambda_i t_i \\ \log(\mu_i) &= \log(\lambda_i) + \log(t_i) \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \log(t_i) \end{aligned}$$

The Offset

- $\log(t_i)$ is called an offset.
- We fit a regression model and fix the offset coefficient so that it is 1.

Rationale for $\log()$ transform

- Many times we call Poisson regression, log-linear regression. The rationale behind using the log transform is:
  - $\log(\lambda)$ has a range of $-\infty$ to ∞ even though $\lambda > 0$. This means there are no restrictions to a specific range.
  - Maximum likelihood estimation works extremely well with the $\log()$ relationship. This is due to the fact that the $\log()$ link is something called the canonical link between outcome and covariates.

Analysis of Grouped or Individual Data

We can actually enter data in different ways

1. Individual Data: one line per patient.
2. Grouped Data: grouping data by a covariate pattern.
    - This happens when all covariates are categorical
    - There are no differences with inferences in either case.

Interpreting Coefficients

- Given our data we consider the following model:

$$\log(\lambda_{x_1,x_2}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- where
  - X_1 is Alcohol Use
    - 1 for Daily
    - 0 for Less than Daily
  - X_2 is mean-centered age at baseline

What is β_0?

- β_0 can be interpreted as:

$$\beta_0 = \log(\lambda_{x_1=0,x_2=0})$$

- This then represents the log CRC rate for less than daily drinkers who are at the mean age.
- The CRC rate for less than daily drinkers who are at the mean age is $\exp(\beta_0)$.

What is β_1?

- β_1 can be interpreted as:
- Consider 2 subjects who are the same age but differ in drinking status:

$$\begin{aligned}
\log(\lambda_{x_1=0,x_2}) &= \beta_0 + \beta_2 x_2 \\
\log(\lambda_{x_1=1,x_2}) &= \beta_0 + \beta_1 + \beta_2 x_2 \\
\beta_1 &= \log(\lambda_{x_1=0,x_2}) - \log(\lambda_{x_1=1,x_2}) \\
&= \log\left(\frac{\lambda_{x_1=1,x_2}}{\lambda_{x_1=0,x_2}}\right)
\end{aligned}$$

How do you interpret β_1?

- This is the log CRC rate ratio comparing daily drinking to less than daily drinking in subjects who are the same age.
- The CRC rate ratio comparing daily drinking to less than daily drinking in subjects who are the same age is $\exp(\beta_1)$.

What is β_2?

1. What is β_2?:

- Consider 2 subjects who differ in age by one year but have the same drinking status:

$$\begin{aligned}
\log(\lambda_{x_1,x_2}) &= \beta_0 + \beta_2 x_2 \\
\log(\lambda_{x_1,x_2+1}) &= \beta_0 + \beta_1 + \beta_2(x_2+1) \\
\beta_2 &= \log(\lambda_{x_1,x_2+1}) - \log(\lambda_{x_1,x_2}) \\
&= \log\left(\frac{\lambda_{x_1,x_2+1}}{\lambda_{x_1=0,x_2}}\right)
\end{aligned}$$

How do you interpret β_2?

- This is the log CRC rate ratio comparing a one year increase over mean age for patients who have the same drinking status.
- The CRC rate ratio comparing a one year increase over mean age for subjects who have the same drinking status is $\exp(\beta_2)$.

Model in R

```
phscrc$mean.cent.age <- phscrc$age - mean(phscrc$age, na.rm = TRUE)
fit5 <- glm(crc ~ alcohol.use + mean.cent.age + offset(log(cayrs)),
    data = phscrc, family = poisson(link = "log"))
```

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	0.001	0.000	0.001	0.001
alcohol.useyes	1.413	0.026	1.051	1.936
mean.cent.age	1.080	0.000	1.067	1.093

Interpretation of Coefficients

- :
  - The CRC rate for less than daily drinkers who are 53 years old is 0.001.
- :
  - The CRC rate ratio comparing daily drinking to less than daily drinking in subjects who are the same age is 1.1976 although it is insignificant.
  - The CRC rate for daily drinkers is 19.76% greater than the CRC rate of less than daily drinkers although it is insignificant.

Interpretation of Coefficients

- :
  - The CRC rate ratio comparing a one year increase over mean age for subjects who have the same drinking status is 1.0781.
  - The CRC rate for one year increase in mean age is 7.81% larger than the CRC rate for subjects at the mean age and who have the same drinking status.

Model fit for Poisson Regression

Deviance Goodness-of-Fit Test

- **Deviance** is a a measure of how close our model predicts the actual observed outcomes.
    -
- We can use this as a basic test for goodness of fit since we hope our predictions are close to actual outcomes.

Distribution

- We first must understand what the distribution of this would be
  - If our model is correctly specified we must determine how much variation we expect in the observed outcomes around the predicted means under the assumption that our data is Poisson.
  - It can be shown that deviance follows a χ^2 distribution equal to the difference in parameters between the model fit at the saturated model, $n - p$.

Chi-Square Test

- This means we can use a χ^2 test for this with the hypothesis of:

$$H_0: \text{ The Model is Correctly Specified}$$

$$\text{vs.}$$

$$H_1: \text{ The Model is Not Correctly Specified}$$

```
pchisq(fit5$deviance, df = fit5$df.residual, lower.tail = FALSE)
```

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

Overdispersion

-When we deal with Poisson data we are saying that

$$E(X) = Var(X)$$

- In other words we are saying that the mean is equal to the variance. If this is not true:
  - We still have valid estimates of relevant event rates
  - We tend to underestimate variance and then have p-values that are too small and confidence intervals that are too narrow.
  - We correct for this using Negative-Binomial Regression.

Dispersion Test

We can test for this in R:

```
library(AER)
dispersiontest(fit5)
```

```
##
##  Overdispersion test
##
## data:  fit5
## z = -8, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##      0.984
```

Results of Dispersion Test

- From this test we see that our dispersion is 1 and so we have correctly scaled the variance compared to the mean.
- If it was not we could make a change that would correct it without having to learn a new regression model:

```
summary(fit5)
```

```
##
## Call:
## glm(formula = crc ~ alcohol.use + mean.cent.age + offset(log(cayrs)),
##     family = poisson(link = "log"), data = phscrc)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.531  -0.199  -0.148  -0.116   4.026
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -7.13694    0.14604  -48.87   <2e-16 ***
## alcohol.useyes  0.34583    0.15557    2.22    0.026 *
## mean.cent.age   0.07665    0.00621   12.35   <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: 2495.0  on 16017  degrees of freedom
## Residual deviance: 2342.6  on 16015  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 2857
```

```
summary(fit5, dispersion = 0.9841428)
```

```
##
## Call:
## glm(formula = crc ~ alcohol.use + mean.cent.age + offset(log(cayrs)),
##     family = poisson(link = "log"), data = phscrc)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.531  -0.199  -0.148  -0.116   4.026
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.13694    0.14488  -49.26   <2e-16 ***
## alcohol.useyes  0.34583    0.15433    2.24    0.025 *
## mean.cent.age   0.07665    0.00616   12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 0.984)
##
##     Null deviance: 2495.0  on 16017  degrees of freedom
## Residual deviance: 2342.6  on 16015  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 2857
```