STATS 111/202

Lecture 14: Poisson Regression

Ana Maria Kenney

(Thanks again to Koko for this week's lecture help)

Statistical Modeling: GLM Count Data

- Modeling counts of the occurrence of an event over time or space is of interest in many settings.
- Examples:
  - The number of vehicles passing on a toll road per weekend.
  - Number of requests to a cloud server per day.
  - Number of cases of a disease over a year time.
  - Number of credit cards someone owns.
  - Number of smartphones a household owns.
  - Number of trees per square mile in a city.
- Natural to consider a Poisson distribution for such types of data.

Poisson distribution.

- Let Y follow a Poisson distribution with parameter μ where $\mu > 0$.
- Then $P(Y = k) = \frac{e^{-\mu}\mu^k}{k!}$, for k=0,1,2,... .
- $E(Y) = \mu$ and $var(Y) = \mu$.
- μ is the expected value of Y where $\mu > 0$.
- Note that the mean and variance are set to equal one another by the model.

Statistical Modeling: GLM Count Data

In the examples before:

- μ is the mean/expected number of vehicles that pass a toll road over a weekend.
- μ is the mean number of cloud server request in a day.
- μ is the mean number of disease cases over a year.
- μ is the mean number of credit cards someone has.

Statistical Modeling: GLM Count Data

Poisson rate parameter.

- Focus commonly relies on the estimation of the *rate parameter*, λ, when analyzing Poisson count data.
- We can consider Y to be a count of events that came about at a rate of λ per unit-time of exposure for an exposure period of A.
- A should be chosen to make scientific sense.
- So what we have is $E(Y) = \mu = A * \lambda$.

Statistical Modeling: GLM Count Data

- Example: If we have rate of $\lambda = 2$ events per minute. Let A be a 1 hour period (60 minutes). Then μ is equal to $A * \lambda = 60 * \lambda = 60 * 2 = 120$ events per hour.
- Example: If we have rate of $\lambda = 20$ events per hour. Let A be a 30-minute period (so half of the time of the lambda rate). Then $\mu = A * \lambda = \frac{1}{2} * 20 = 10$ per 30-minute period.
- Example: If we have rate of $\lambda = 0.05$ event per subject. Let A be 1000 subjects. Then $\mu = 1000 * 0.05 = 50$ events per 1000 people/subjects.

Statistical Modeling: GLM Count Data

Poisson regression model set-up.

- Let λ_i be the incidence (event) rate in the i-th observation (or population) with explanatory variables $X_{i1}, ..., X_{ip}$.
- The random component is the Poisson distribution noted earlier.
- The systematic component is $\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$.
- The link is the log link: $log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$, where log is the natural log.
- Which leads us to $\lambda_i = e^{\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}}$.

Interpretation in a Poisson regression.

- We interpret coefficients (and covariate effects on response) on the exponentiated scale.
- Note that when all covariates are set to 0 ($X = 0$), that $\lambda_i = e^{\beta_0}$.
- Now take a unit increase in X_{i1} given all other explanatory variables held constant. Take two observations
- For the unit increase, have that $log(\lambda_{i+}) = \beta_0 + \beta_1(X_{i1} + 1) + ... + \beta_p X_{ip}$, where log is the natural log.
- For the original X, have that $log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$, where log is the natural log.

Statistical Modeling: GLM Count Data

- Taking the difference of these we have:

$$log(\lambda_{i+}) - log(\lambda_i) = \beta_1$$

- Which leads to:

$$log\left(\frac{\lambda_{i+}}{\lambda_i}\right) = \beta_1 \Rightarrow \frac{\lambda_{i+}}{\lambda_i} = e^{\beta_1}$$

- A unit increase in X_1 results in a relative change in rates (numerator rate is for the increased X_1) of e^{β_1}.
- That is to say a unit increase in X_1 will lead to a rate that is e^{β_1} times the old rate.
- Moreover, a M unit change in X_1 will result in $\frac{\lambda_{i+}}{\lambda_i} = e^{\beta_1 * M}$.
- An M unit increase in X_1 results in a relative change in rates of $e^{\beta_1 * M}$.
- That is to say a M unit increase in X_1 will lead to a rate that is $e^{\beta_1 * M}$ times the old rate.

Statistical Modeling: GLM Count Data

- If $\beta_j = 0$, then Y and X_j are not related.
- If $\beta_j > 0$, then $e^{\beta_j} > 1$, and so Y and X_j are positively related (as X increases, the expected count/rate will increase).
- If $\beta_j < 0$, then $e^{\beta_j} < 1$, and so Y and X_j are negatively related (as X increases, the expected count/rate will decrease).

Statistical Modeling: GLM Count Data

Confidence intervals and testing.

- Like before, can compute a confidence interval for β_j and then exponentiate to get confidence interval for e^{β_j}.
- Can also compute confidence interval for linear combinations of the β's. This would be needed if there are interactions in the model, comparing two option groups of the covariate, or to obtain intervals for the rate λ itself.
- Can also conduct test of several coefficients using the likelihood ratio test like we did with logistic regression.
- We can use all the R functions we used for logistic regression here with our Poisson regression (glmCI, linContr.glm, lrtest and robust.se.glm).

Statistical Modeling: GLM Count Data

Poisson model with offset term.

- What we have is Y_i is assumed to be Poisson, where we have rate λ_i and exposure period A_i.
- $E(Y_i) = \mu_i = A_i * \lambda_i$ and $var(Y_i) = \mu_i$.
- And so (using natural log)
  $log(\mu_i) = log(A_i \lambda_i) = log(A_i) + log(\lambda_i)$.
- Thus $log(\mu_i) = log(A_i) + \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$.
- The $log(A_i)$ term must be added as what is called an *offset term*.
- Offset term is a term that has its coefficient set to 1.
- This offset term will have no effect when $A_i = 1$ (rate period same as exposure period) or when all A_i are equal to each other (each observation has same exposure period).

Statistical Modeling: GLM Count Data

- Will account for A_i with the offset term $log(A_i)$ when the exposure period for all observations are not the same.
- Say we have two observations, where observation 1 is the number of events across a 24 hour period and observation 2 is the number of events across a 12 hour period.
- Cannot just compare the rates, because observation 1 will result in a higher rate just because it had a longer exposure period.
- By adjusting for the exposure periods (here $A_1 = 24$ and $A_2 = 12$), we can compare rates as they would be for a 1 hour period.

Statistical Modeling: GLM Count Data

We'll start with an example where A_i is the same for all observations.

Female crabs were monitored to see how many male crabs came to nest near the female crab. The response variable is how many males came to nest near the female crab. Explanatory variables are color (C=1,2,3,4, higher the brighter), spine condition (S=1,2,3, higher the wider), weight (Wt, in pounds), and carapace width (W, in inches).

All crabs were observed for the same exposure period. The number of counts varied from 0 to 15.

Statistical Modeling: GLM Count Data

Let us fit the model $log(\mu_i) = log(\lambda_i) = \beta_0 + \beta_1 W_i$.

In R, the call is the same as logistic regression, but change the family and link option.

Call would be: glm(Y∼X, family=poisson(link="log"), data=data)

Statistical Modeling: GLM Count Data

```
glm(formula = Sa ~ W, family = poisson(link = log),
data = crab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8526  -1.9884  -0.4933   1.0970   4.9221

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
W            0.16405    0.01997   8.216  < 2e-16 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18

Number of Fisher Scoring iterations: 6
```

Statistical Modeling: GLM Count Data

Interpreting the model.

- $e^{\hat{\beta}_0} = e^{-3.304} = 0.0367$. We estimate that a female crab with width=0 will have the expected count of events to be 0.0367.
- Now say W=25 inches. $e^{\hat{\beta}_0 + \hat{\beta}_1 * 25} = e^{-3.304 + 0.164 * 25} = 2.216$. For a female crab with W=25, we estimate the expected count to be 2.216 events.
- A 1 unit increase in W will result in a relative change in expected counts by $e^{\hat{\beta}_1} = e^{0.164} = 1.178$.
- That is to say a female crab with 1 inch greater W will have expected count that is 1.178 times that of a similar crab who has the original W width.

Statistical Modeling: GLM Count Data

Interpreting the model.

- Now say we are interested in a 5 unit change in W.
- A 5 unit increase in W will result in a relative change in expected counts by $e^{\hat{\beta}_1 * 5} = e^{0.164 * 5} = 2.270$.
- That is to say a female crab with 5 inch greater W will have expected count that is 2.270 times that of a similar crab who has the original W width.
- As before, we can get confidence intervals for linear combinations of β's and then exponentiate to get intervals for the rate/count.

Statistical Modeling: GLM Count Data

Confidence intervals for the exponentiate coefficients.

	exp(Est)	ci95.lo	ci95.hi	z value	Pr(>|z|)
(Intercept)	0.0367	0.0127	0.1062	-6.0946	0
W	1.1783	1.1331	1.2253	8.2165	0

This is to say that for a 1 unit increase in W, we are 95% confident that the new rate/count will be between 1.1331 and 1.2253 times that of the original rate/count.

Or can say we are 95% confident that a 1 unit increase in W will result in a relative change in rate/count that is between 1.1331 and 1.2253.

Statistical Modeling: GLM Count Data

Now say we fit a model with explanatory variables being W (width) and Wt (weight) along with an interaction.

```
glm(formula = Sa ~ W + Wt + W * Wt, family = poisson, data = crab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1120  -1.8423  -0.5578   0.9171   4.9420

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.90638    1.90575  -3.624 0.000290 ***
W            0.23014    0.07520   3.061 0.002209 **
Wt           3.06607    0.74642   4.108    4e-05 ***
W:Wt        -0.08658    0.02478  -3.494 0.000476 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 545.27  on 169  degrees of freedom
AIC: 908.57

Number of Fisher Scoring iterations: 6
```

poisson(link=log)

Interpreting the model:
$log(\mu_i) = log(\lambda_i) = \beta_0 + \beta_1 W_i + \beta_2 Wt_i + \beta_3 W_i Wt_i$.

- Say a female crab weighs 3 pounds.
- A 1 unit increase in W will result in a relative change in expected counts by $e^{\hat{\beta}_1 + \hat{\beta}_3 3} = e^{0.230 - 0.0865 * 3} = 0.970$.
- That is to say a 3 pound female crab with 1 inch greater W will have expected count that is 0.970 times that of a similar crab who has the original W width.

Confidence intervals for model with interaction.

```
Test of H_0: exp( 1*W + 3*W:Wt ) = 1 :

  exp( Est ) se.est  zStat pVal ci95.lo ci95.hi
1      0.971  0.052 -0.568 0.57   0.877   1.075
```

This is to say that for a 1 unit increase in W for crab that is 3 pounds, we are 95% confident that the new rate/count will be between 0.877 and 1.075 times that of the original rate/count.

Or can say we are 95% confident that for a 3 pound crab a 1 unit increase in W will result in a relative change in rate/count that is between 0.877 and 1.075.

Statistical Modeling: GLM Count Data

We can compare the two models we just fit (reduced and full model) by using the lrtest() function just like we did with logistic regression.

Reduced model had just W as explanatory variable and the full model had W, Wt, and an interaction.

```
> lrtest(crab.model,crab.model2)

Assumption: Model 1 nested within Model 2

  Resid. Df Resid. Dev Df Deviance pValue
1       171    567.879
2       169    545.273  2   22.606      0
```

The null H_0 is that the reduced model fits well enough. The alternative H_a is that the the full model is fits better. P-value is approximately 0, so we reject the null and conclude evidence for the alternative. Have evidence that the full model fits better than the reduced.

Statistical Modeling: GLM Count Data

Modeling the rate.

- So far we only considered the scenario where all observations are observed for the same exposure period (observed for same amount of time).
- In some cases, we do not observe each observation for the same exposure period.
- For example say we have two subjects in my dataset. Subject 1 is observed for 3 hours and subject 2 is observed for 10 hours.
- It will not be correct to just compare the counts between these subjects, as subject 2 is observed much longer than subject 1 (and therefore likely to have higher counts just because of the longer exposure period).
- Instead of trying to find the lowest common denominator among the exposure periods, the regression will just model the rate per unit time (in this example that would be a 1 hour period).

Statistical Modeling: GLM Count Data

- At this point, we will model the *rate* λ and not the counts themselves.
- Will model the rate per hour for each observation by utilizing the offset term described earlier in the slides. To get the estimated counts, will just multiply the rate by the exposure period.
    - Here in our example on the previous slide, we will have an estimate of the rate per unit time (depending on the data, could be per hour, per minute, per year, or even per person). To get subject 1's estimated expected counts, would just multiply the rate by 3. To get subject 2's estimated expected counts, would just multiply the rate by 10.

Person years exposure period.

- Now say each observation has numerous subjects.
- That is to say that each explanatory variable setting, several subjects are followed for a period of time (and the number of events is observed).
- The exposure period is going to be defined as the total time observed across all subjects.
- This is to say at a specific level of the explanatory variables, if we observe 10 subjects and each one for 5 hours, then in total we have 50 hours of observations (or the first 5 for 1 hour each and the second 5 subject for 9 hours each, for a total of 50 hours).

Statistical Modeling: GLM Count Data

Example of such type of data. The Gail data on the class site. Interest lies in the incidence (or rate) of melanoma by latitude, adjusting for age.

Nine cities/locales were selected and given a latitude designation. Age-specific populations were determined and followed for a set period of time, yielding age- and locale-specific person-time of exposure (total number of years observed across all subjects in each age and location group).

Number of new cases of melanoma were recorded by age and locale. Age is categorized to be below 35 years old, between 35-44, between 45-54, between 55-64, between 65-74 and 75+.

Statistical Modeling: GLM Count Data

The data is structured as follows:

	locale	ageg	inccases	persyrs	latitude
1	Detroit	<35 years	18	991371	Northern
2	Detroit	35-44 years	25	199104	Northern
3	Detroit	45-54 years	33	209192	Northern
4	Detroit	55-64 years	39	146400	Northern
5	Detroit	65-74 years	18	82961	Northern
6	Detroit	>=75 years	13	44096	Northern

Look at the first observation. This is to say in Detroit, among those who are younger than 35 years old, a total of 991371 years were observed. Out of this exposure period, 18 events were observed.

Second observation says that in Detroit, among those 35-44 years old, a total of 199104 years were observed. A total of 25 events were observed in this exposure period.

Other locales include Atlanta, San Francisco, Dallas, and Pittsburgh.

Statistical Modeling: GLM Count Data

- For example if we observed 10,000 subjects for 10 years each, that will result in a total of 100,000 years observed.
- Since the amount of years observed at each explanatory setting (location, age, and latitude) is not the same, we cannot just compare the count of events.
- We will model the rate λ, which will be the rate of event per year.
- Will then multiply this per year rate by the years observed to obtain an estimate of the expected counts.
- To fit a model with this data, must specify the offset term to be log(persyrs) (where persyrs is our A_i exposure period).

Statistical Modeling: GLM Count Data

```
glm(formula = inccases ~ factor(ageg) + latitude.ord, family = poisson(link = l
    data = gail, offset = log(persyrs))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9430  -0.7380  -0.2058   0.6037   2.9128

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -10.91171    0.09722 -112.23   <2e-16 ***
factor(ageg)>=75  years    2.71086    0.10982   24.68   <2e-16 ***
factor(ageg)35-44 years    1.66212    0.09798   16.96   <2e-16 ***
factor(ageg)45-54 years    1.87783    0.09359   20.06   <2e-16 ***
factor(ageg)55-64 years    2.04197    0.09694   21.06   <2e-16 ***
factor(ageg)65-74 years    2.19106    0.10796   20.30   <2e-16 ***
latitude.ord               0.41011    0.03467   11.83   <2e-16 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1196.489  on 53  degrees of freedom
Residual deviance:   71.135  on 47  degrees of freedom
AIC: 342.91
```

Statistical Modeling: GLM Count Data

- The model is:
  $log(\mu_i) = \beta_0 + \beta_1 \mathtt{I}(agegrp \geq 75) + \beta_2 \mathtt{I}(agegrp35 - 44)$
  $+\beta_3 \mathtt{I}(agegrp \geq 45 - 54) + \beta_4 \mathtt{I}(agegrp55 - 64)$
  $+\beta_5 \mathtt{I}(agegrp65 - 74) + \beta_6 Latitude_i + log(persyrs_i)$
- Latitude is an ordinal variable with levels 1,2, and 3. North is 1, middle is 2, and south is 3.
- This is to signify that an increase in latitude means closer to the equator (more sunlight).
- The base age group is age<35.
- The offset term is $log(persyrs_i)$.

- The model is now estimating the rate of event per year observed.
- Say we are in age group ≥ 75 and are at latitude=1.
- This results in $log(\hat{\lambda}_i) = -10.911 + 2.710 + 0.410 = -7.791$ and so $\hat{\lambda}_i = e^{-10.911+2.710+0.410} = e^{-7.791} = 0.0004134392$.
- This is to say that this explanatory setting (age≥ 75 and latitude=1) will result in an estimated rate of 0.0004134392 events per year observed.

- We would then multiply this by a scientifically interesting exposure period to get the estimated expected counts.
- Say we observe 10,000 subjects for 10 years each, resulting in 100,000 years observed.
- $\hat{\lambda}_i A_i = 0.0004134392 * 100000 = 41.34$.
- According to the model, we have an estimated rate of 0.0004134392 events per year observed. If we observe 100,000 years in total, then we would have an estimated expected counts of 41.34.

Statistical Modeling: GLM Count Data

- In models with offsets, interpretation is done with respect to the rates per unit exposure time.
- Say we go from latitude=1 to latitude=2 (a 1 unit increase in latitude, go from North to Middle), holding age group constant.
- This will result in an estimated relative change in rates of $e^{0.4101} = 1.5069$.
- This is to say that the Middle latitude (latitude=2) will result in an estimated rate (per year observed) that is 1.5069 times that of the North (latitude=1) latitude (holding all other covariates constant, i.e. age group remains the same).

Statistical Modeling: GLM Count Data

- The relative change in rates $\frac{\lambda_+}{\lambda}$ is the systematic component exponentiated.
- The glmCI() function and linContr.glm() function can be used just like with logistic regression.
- Can obtain confidence intervals for the relative change and for the rate based upon a certain set of explanatory variables.

Statistical Modeling: GLM Count Data

Example using the Gail data and model from a few slides ago:

```
> glmCI(model.gail)
                          exp( Est ) ci95.lo ci95.hi   z value Pr(>|z|)
(Intercept)                   0.0000  0.0000  0.0000 -112.2344        0
factor(ageg)>=75  years      15.0423 12.1293 18.6549   24.6847        0
factor(ageg)35-44 years       5.2705  4.3496  6.3863   16.9646        0
factor(ageg)45-54 years       6.5393  5.4434  7.8559   20.0642        0
factor(ageg)55-64 years       7.7057  6.3723  9.3182   21.0637        0
factor(ageg)65-74 years       8.9447  7.2389 11.0526   20.2950        0
latitude.ord                  1.5070  1.4080  1.6130   11.8280        0
```

Statistical Modeling: GLM Count Data

- We are 95% confident that going from base age group (<35 years) to the oldest group ($\geq$ 75 years) will result in a relative rate change between 12.129 to 18.654.
- That is to say we are 95% confident the oldest group will have rate that is between 12.129 to 18.654 times that of the base (youngest) group.
- We are 95% confident that a 1 unit increase in latitude (closer to equator) will result in a relative rate change between 1.408 to 1.613.
- That is to say we are 95% confident that increasing latitude by 1 (say going from middle to south) will result in a rate that is 1.408 to 1.613 times that of the rate of the original latitude.

Statistical Modeling: GLM Count Data

Use linContr.glm() function to obtain a confidence interval for the rate of events for the 45-54 age group and latitude=3 (south).

```
> linContr.glm(c("(Intercept)","factor(ageg)45-54 years","latitude.ord") ,
c(1,1,3), model.gail)

Test of H_0: exp( 1*(Intercept) + 1*factor(ageg)45-54 years +
3*latitude.ord ) = 1 :

  exp( Est )      se.est     zStat pVal    ci95.lo    ci95.hi
1 0.00040828 0.07267891 -107.3702    0 0.00035408 0.00047079
```

We are 95% confident the rate (per year) for the event (melanoma) for age group 45-54 and latitude=3 (south) is between 0.000345408 and 0.00047079.