Statistics for the Sciences

Poisson Regression and Negative Binomial Regression

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Poisson Regression
- Negative Binomial Distribution
- Negative Binomial Regression
- Lab

Poisson Regression

- Often,the outcome of a variable is numerical in the form of counts.
- Conditions:
  - (a) The probability of at least one occurrence of an event in a given time interval is proportional to the length of the interval.
  - (b) The probability of two or more occurrences of an event within an extremely small interval is negligible. -(c) The number of occurrences of an event in disjoint time intervals are mutually independent.
- If the above conditions are met, then Poisson distribution can be used to model the response

$$p(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots, \lambda > 0,$$

where λ is the average number of successes (the average count)in a time or space interval.

Poisson Regression

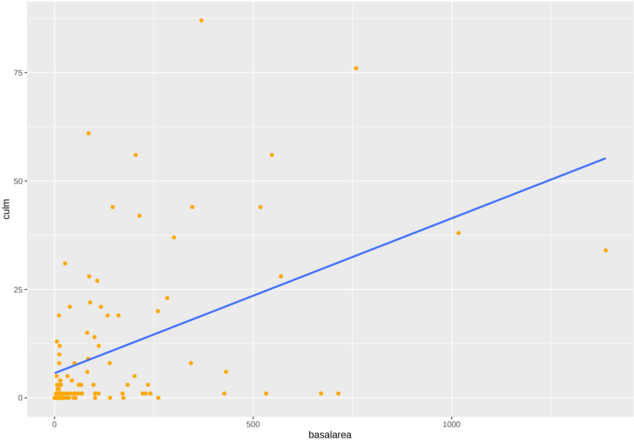
- Example (fill.csv): Fill et al (2021) studied the effect of duff (leaf litter) on the post-fire ecology of wiregrass (Aristida beyrichiana) in a section of pine savanna. They sampled 99 plants in an area of 0.1 km^2, recorded plant basal area and allocated each plant to one of three treatments: high duff, low duff, low duff with added pine cones. They then burnt the area and five months later, counted the number of culms on each plant. We will model numbers of culms per plant against basal area and duff treatment using each plant as the unit of analysis.
  - Response variable: `culm` Number of Culms per Plant
  - covariates `basalarea` Basal Area
  - Factor `treatment` with three levels: `High Duff`, `Low Duff` and 'Low Duff + Pinecones

Poisson Regression

```
##    culm reprod basalarea            treatment
## 1     8      1  11.83780            High Duff
## 2    19      1 160.92500            High Duff
## 3     3      1  15.93550            High Duff
## 4     9      1  84.78000            High Duff
## 5    44      1 346.97000            High Duff
## 6    15      1  82.15025            High Duff
## 7     4      1  43.52040            High Duff
## 8     0      0   8.24250            High Duff
## 9     1      1  41.01625            High Duff
## 10    1      1 229.69100            High Duff
## 11    1      1 102.05000            High Duff
## 12    1      1 714.50700            High Duff
## 13    0      0  52.75200            High Duff
## 14    1      1 671.17500            High Duff
## 15   21      1 116.80800            High Duff
## 16    1      1  51.02500            High Duff
## 17    1      1 221.95875            High Duff
## 18   22      1  89.49000            High Duff
## 19    2      1  10.99000            High Duff
## 20    5      1   5.27520            High Duff
## 21   19      1  11.05280            High Duff
## 22   10      1  11.97125            High Duff
## 23    1      1  32.49900            High Duff
## 24    3      1  61.04160            High Duff
## 25    3      1  67.11750            High Duff
## 26    5      1  32.49900            High Duff
## 27    8      1 343.24125            High Duff
## 28    1      1 532.70100            High Duff
## 29   31      1  26.49375             Low Duff
```

Poisson Regression

- For simplicity, let's ignore the treatment. Does SLR model work?



Poisson Regression

- Distribution of Count data
  - non-negative integers
  - not normally distributed
- The relationship between the predictors and the count outcome is often non-linear.
- So GLM should be used with

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

and the coefficients in Poisson regression can be interpreted as the effect of the predictors on the log of the expected count.

Poisson Regression

- Model fit

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         2102.      98 -1004. 2012. 2017.    1729.          97    99
```

```
## # A tibble: 2 x 7
##   term         estimate std.error statistic   p.value conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)  2.03    0.0384         52.9 0            1.96      2.11
## 2 basalarea    0.00172 0.0000750      22.9 5.50e-116  0.00157   0.00186
```

Poisson Regression

- Likelihood ratio test

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: culm
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         98     2101.6
## basalarea  1   372.99        97     1728.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Negative Binomial Distribution

- But if we check the response again ...

```
##   mean_culm var_culm
## 1  11.16162 314.2389
```

Negative Binomial Distribution

- It often occurs that the variance of the count data exceeds that of Poisson. This phenomenon is called **over-dispersion** which is quite common in practice.
  - Poisson model is not adequate to modelling the over-dispersed count data.
- Let Y be a negative binomial random variable. Then its probability mass function is given by

$$Pr(Y=y) = \binom{k+y-1}{y} p^k (1-p)^y,\ k>0, 0 \leq p \leq 1, y = 0, 1, 2, \ldots$$

where y is the number of failures before the kth success.

Negative Binomial Distribution

- A better parameterization (Bliss and Owen, 1958) for a negative binomial distribution with mean μ and coefficient c, write $Y \sim NB(\mu, c)$, is

$$Pr(Y = y|\mu, c) = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})} \left(\frac{c\mu}{1 + c\mu}\right)^y \left(\frac{1}{1 + c\mu}\right)^{c^{-1}}, \ 0 < \mu, c < \infty, y = 0, 1, 2, .$$

  - It can be shown that $E(Y) = \mu$ and $var(Y) = \mu + c\mu^2$.
  - The positiveness $Var(Y) = \mu(1 + \mu c)$ implies that $c > -1/\mu$
  - $NB(\mu, c)$ distribution becomes the Poisson distribution when $c \to 0$.
  - Dispersion parameter c can take a positive as well as a negative value.

Negative Binomial Regression

- Response $Y \sim NB(\mu, c)$
- Log link (same as Poisson regression)

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

Negative Binomial Regression

- Model fit

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik      AIC   BIC deviance df.residual  nobs
##           <dbl>   <int> <logLik>  <dbl> <dbl>    <dbl>       <int> <int>
## 1          128.      98 -308.7716  624.  631.     110.          97    99

## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  1.73     0.186         9.31 1.22e-20  1.32      2.16
## 2 basalarea    0.00315  0.000662      4.75 2.00e- 6  0.00149   0.00519
```

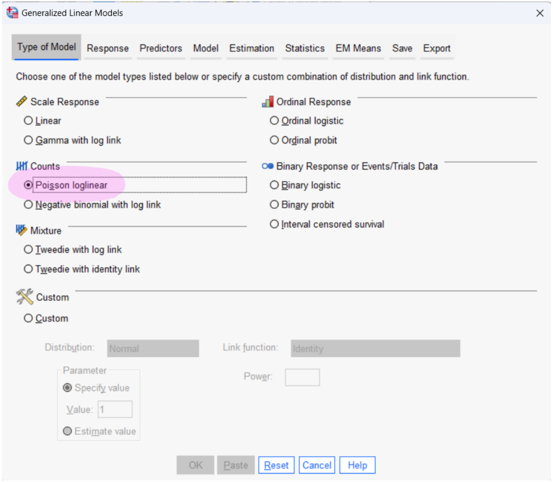
Negative Binomial Regression

- Likelihood ratio test

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.4539), link: log
##
## Response: culm
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         98     128.03
## basalarea  1   17.546        97     110.49 2.805e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

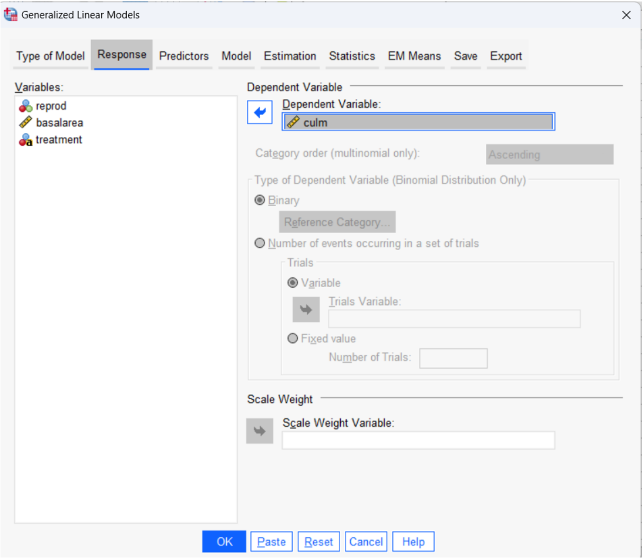
Lab

- After importing data `fill.csv`, click on `Analyze` $\rightarrow$ `Generalized Linear Models` $\rightarrow$ `Generalized Linear Models ...`



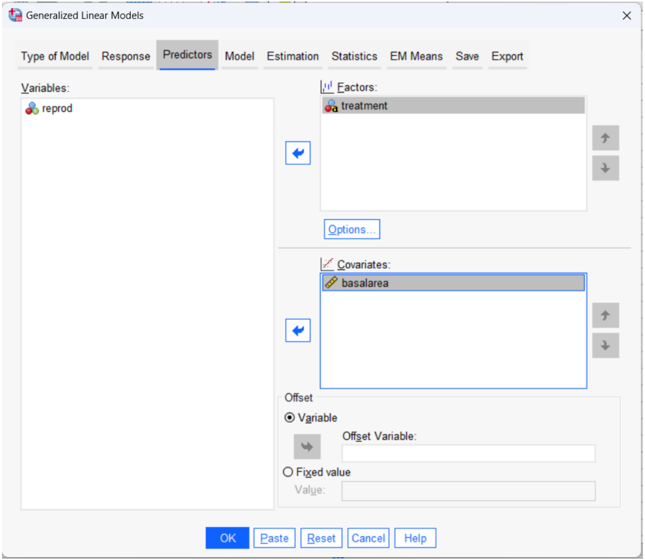
Lab

- Define the Response Variable



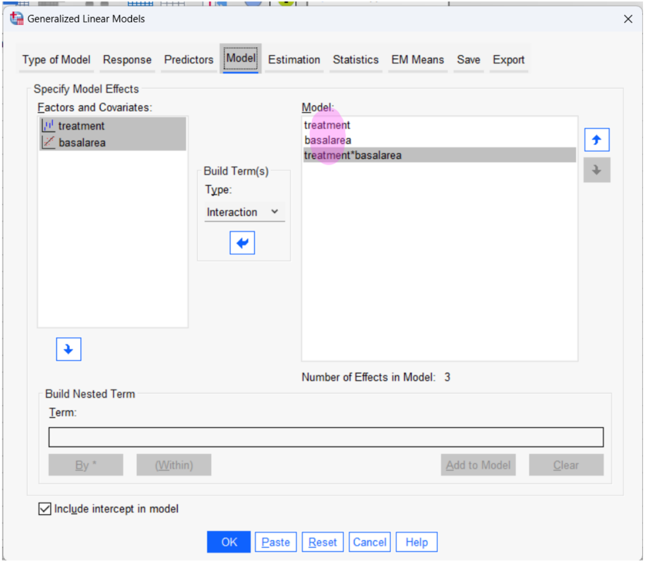
Lab

- Define Predictors. Let's add the factor as well



Lab

- Specify the Model



Lab

- Run the analysis

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	1646.819	1	<.001
treatment	33.740	2	<.001
basalarea	147.356	1	<.001
treatment * basalarea	89.023	2	<.001

Dependent Variable: culm
Model: (Intercept), treatment, basalarea, treatment * basalarea

Lab

- Re-fit the model using Negative Bimomial Regression

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	115.088	1	<.001
treatment	3.227	2	.199
basalarea	13.751	1	<.001
treatment * basalarea	6.602	2	.037

Dependent Variable: culm
Model: (Intercept), treatment, basalarea, treatment * basalarea

Double-click to activate

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	Upper	Hypothesis Test Wald Chi-Square	df	Sig.
(Intercept)	1.274	.3181	.651	1.897	16.041	1	<.001
[treatment=High Duff]	.725	.4282	-.114	1.565	2.870	1	.090
[treatment=Low Duff]	.580	.3824	-.170	1.329	2.300	1	.129
[treatment=Low Duff + Pinecones]	0[a]	.	.	.	.	.	.
basalarea	.005	.0017	.002	.009	10.765	1	.001
[treatment=High Duff] * basalarea	-.005	.0022	-.010	-.001	6.260	1	.012
[treatment=Low Duff] * basalarea	-.002	.0019	-.006	.001	1.707	1	.191
[treatment=Low Duff + Pinecones] * basalarea	0[a]	.	.	.	.	.	.
(Scale)	1[b]						
(Negative binomial)	1[b]						

Dependent Variable: culm
Model: (Intercept), treatment, basalarea, treatment * basalarea

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.