Chapter 14. Poisson/loglinear regression

1. Poisson regression or loglinear regression is useful for describing count data, rate data and contingency table.

2. **Poisson regression** refers to the situation where we want to relate the events (e.g., number of car accidents, number of dead beetles, or the number of cured patients) to various explanatory variables that can be continuous or categorical.

3. **Loglinear model** refers to contingency table analysis where the response variable is the frequency or count in each cell of the table. The variables used to define the table are treated as explanatory variables.

4. Example: model the number of species (or the abundance of *Ocotea whitei*) in a quadrat, denoted by *y*, in terms of *slope*, *meanelev*, *convex* and *habcat*.

   The reasonable model for *y* is the Poisson distribution:

$$f(y;\mu) = \frac{\lambda^y e^{-\lambda}}{y!}, \qquad y = 0, 1, 2, 3, \ldots.$$

   In the standard format of the exponential family:

$$f(y;\mu) = \exp\left(y \log \lambda - \lambda - \log y!\right)$$

   Recall that the general form of the exponential family can be written as:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

   If $a(y) = y$, the pdf is said to be in the *canonical form*, and $b(\theta)$ is called the *natural parameter*.

5. Systematic component – Specify the slope *x* (explanatory variable) as

$$\beta_0 + \beta_1 x$$

6. Link function – As in the logistic regression, it is most typical to use the natural parameter as a link function. In our case, the natural parameter is log(λ):

$$\log(\lambda) = \beta_0 + \beta_1 x$$

Other links include:

Identity link: $\lambda = \beta_0 + \beta_1 x$

Square root link: $\sqrt{\lambda} = \beta_0 + \beta_1 x$

7. R-code:

bcisp.log=glm(rich~slope,family=poisson(link=log),data=bcisp.dat)

Let's look at the outputs:

> summary(bcisp.log)

Call:

glm(formula = rich ~ slope, family = poisson(link = log), data = bcisp.dat)

Deviance Residuals:

Min	**1Q**	**Median**	**3Q**	**Max**
-3.77294	**-0.84163**	**-0.00519**	**0.76878**	**3.92256**

Coefficients:

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	**3.882422**	**0.006615**	**586.91**	**<2e-16 *****	(Wald test: $\frac{\hat{\beta}_0}{se(\hat{\beta}_0)} \sim N(0,1)$)
slope	**0.012194**	**0.001034**	**11.79**	**<2e-16 *****	($\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim N(0,1)$)

Signif. codes: 0 `***' 0.001 `' 0.01 `*' 0.05 `.' 0.1 ` ' 1**

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1904.7 on 1249 degrees of freedom

Residual deviance: 1768.3 on 1248 degrees of freedom

AIC: 8984.8

Number of Fisher Scoring iterations: 4

8. Plot the observed data and prediction

   > plot(bcisp.dat$slope,bcisp.dat$rich)

   > lines(lowess(bcisp.dat$slope,bcisp.dat$rich,f=0.1),col="red")

   > id=order(bcisp.dat$slope)

   > lines(bcisp.dat$slope[id],fitted.values(bcisp.log)[id],col="blue")

9. Use high order polynomial terms

   Try this model: log(y)=x+x^2+x^3.

   Let's use orthogonal polynomials here:

   >bcisp.log=glm(rich~poly(slope,3),family=poisson(link=log),data=bcisp.dat)

   Open a new graphical window: **>X11()**

10. Hypothesis test – Wald test, H_0: $\beta_1 = 0$

$$z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim N(0,1)$$

   If |z| is large, H_0 will be more likely to be rejected. The glm outputs are based on this test.

11. Check model adequacy using residuals

(1) The Pearson residuals: $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$. The denominator uses $\sqrt{e_i}$ because for the Poisson distribution, $\text{var}(Y_i) = E(Y_i)$, the standard error of Y_i is estimated by $\sqrt{e_i}$.

(2) Standardized residuals: $r_{pi} = \frac{o_i - e_i}{\sqrt{e_i}\sqrt{1-h_i}}$, where the leverage, h_i, is the i^{th} element on the diagonal of the **hat matrix**: $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (note: $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$).

(3) Chi-squared goodness-of-fit statistic is related to the Pearson residuals by:

$$\chi^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}.$$

(4) Deviance: $D = 2\sum\left[o_i \log\frac{o_i}{e_i} - (o_i - e_i)\right]$. However, for most models, $\sum o_i = \sum e_i$, so the deviance is simplified to: $D = 2\sum\left[o_i \log\frac{o_i}{e_i}\right]$

(5) The deviance residuals are defined as:

$d_i = sign(o_i - e_i)\sqrt{2\left[o_i \log\frac{o_i}{e_i} - (o_i - e_i)\right]}$. **R-code: >resid(bcisp.log)**

The deviance defined in (3) is simply: $D = \sum d_i^2$.

Note: It is quite easy to show that $D \approx \chi^2$. Use the Taylor expansion:

$$o\log\frac{o}{e} = (o-e) + \frac{1}{2}\frac{(o-e)^2}{e} + \ldots$$

(6) Pseudo R^2 provides an overall test of the hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_p = 0$:

$R^2 = \frac{2[l(b_{\min}) - l(b)]}{l(b_{\min})}$, where $l(b_{\min})$ is the maximum value of the log-likelihood function for a minimal model with no covariates (only including intercept β_0), and the maximum of the log-likelihood function for the model with p parameters.

12. There are three types of link functions for Poisson regression in **R**. Try

    bcisp.identity=glm(rich~slope,family=poisson(link=identity),data=bcisp.dat)

    bcisp.sqrt=glm(rich~slope,family=poisson(link=sqrt),data=bcisp.dat)

13. Build the "best" model

    What is the best model is up to anyone's imagination. Compare these:

    > bcisp.slope3=glm(rich~poly(slope, 3), family=poisson, data=bcisp.dat)

    > bcisp.logslope=glm(rich~log(slope), family=poisson, data=bcisp.dat)

    Compare the two predictions against observations. We will choose the second model as it is simpler.

    Now to consider including other explanatory variables:

    > bcisp.pois=glm(rich~log(slope)+meanelev+convex+habcat,family=poisson(link=sqrt),data=bcisp.dat)

14. Variable selection:

    (1). Start from model with constant term only:

    > bcisp.1=glm(rich~1,family=poisson(link=sqrt),data=bcisp.dat)

    (2). Use **step**

    > **bcisp.step=step(bcisp.1,~log(slope)+meanelev+convex+habcat)**

15. Loglinear models

    Loglinear models are used to model cell frequency in a contingency table. In the nonparametric section, we have learned how to test the independence of the classification schemes (factors). The loglinear model is a parametric approach.

Before specifying loglinear models for frequency data summarized in contingency table, it is important to consider how the design of a study may constrain on the data. The study design also affects the choice of probability models to describe the data. There are two basic designs (see Dobson's book, page 157-160). One has fixed total, i.e., the total number of samples (patients) are prefixed before survey/experiment is conducted. The other is the row or column totals are fixed before conducting the experiment. In R, both design share the same model formula.

16. Example: Cross-sectional study of malignant melanoma – These data are from a cross-sectional study of patients with a form of skin cancer called malignant melanoma. For a sample of $n = 400$ patients, the site of the tumor and its histological type were recorded. The data, numbers of patients with each combination of tumor type and site, are given below.

Tumor type	Site			
	Head & Neck	Trunk	Extremities	Total
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Question of interest: Is there any association between tumor type and site?

To answer this question, we can use contingency table analysis. However, loglinear regression can give you more detailed information about the relationship. For example, the results will tell you whether the occurrence of Indeterminate tumor is different across sites.

17. Model

Let Y_{ij} denote the frequency for the $(i, j)^{th}$ cell with i = 1, 2, 3, 4 rows, and j = 1, 2, 3 columns. In this study the total number of patients is $n = \sum_{i=1}^{4}\sum_{j=1}^{3} Y_{ij} = 400$, where 400 is fixed by the design of the study – it was determined before survey was conducted. In this case, the Poisson model has to be constrained by this condition (n = 400).

If the Y_{ij}'s are independent random variables with Poisson distribution with parameters $E(Y_{ij}) = \mu_{ij}$, then their sum has the Poisson distribution with parameter $E(n) = \mu = \sum_{i=1}^{4}\sum_{j=1}^{3} \mu_{ij}$. Hence, the joint probability distribution of the Y_{ij}'s, conditional on their sum n, is the **multinomial distribution**:

$$f(\mathbf{y} \mid n) = n! \prod_{i=1}^{4}\prod_{j=1}^{3}\left[\frac{\theta_{ij}^{Y_{ij}}}{Y_{ij}!}\right], \quad \text{where } \theta_{ij} = \frac{\mu_{ij}}{\mu}, \text{ and } \sum\sum \theta_{ij} = 1.$$

θ_{ij} in this model can be interpreted as the probability of an observation in the $(i, j)^{th}$ cell of the table. The expected value of Y_{ij} is: $E(Y_{ij}) = \mu_{ij} = n\theta_{ij}$. The link function is then:

$$\log(\mu_{ij}) = \log(n) + \log(\theta_{ij}).$$

18. For a two contingency table (like the melanoma data), the most commonly considered hypothesis is that the row and column variables are independent so that:

$$\theta_{ij} = \theta_{i.}\theta_{.j},$$

where $\theta_{i.}$ and $\theta_{.j}$ are the marginal probabilities with $\sum\theta_{i.} = 1$ and $\sum\theta_{.j} = 1$.

This hypothesis can be tested by comparing the fit of the following two linear models:

Full model: $\log(\mu_{ij}) = \log(n) + \log(\theta_{ij})$ (# of parameters = # of cells)

Reduced model: $\log(\mu_{ij}) = \log(n) + \log(\theta_{i}.) + \log(\theta_{.j})$

Note: this is analogous to the analysis of variance for two factor experiment without replication.

Full model: $\log(\mu_{ij}) = \log(n) + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

Reduced model: $\log(\mu_{ij}) = \log(n) + \alpha_i + \beta_j$

So the key question is whether $(\alpha\beta)_{ij} = 0$.

19. Inference for loglinear models

The adequacy of the loglinear model can be assessed by using the same goodness-of-fit statistics given in (11) in the above.

20. Data format in R:

> melanoma.dat

	type	site	freq
1	**freckle**	**headneck**	**22**
2	**freckle**	**trunk**	**2**
3	**freckle**	**extremities**	**10**
4	**spread**	**headneck**	**16**
5	**spread**	**trunk**	**54**
6	**spread**	**extremities**	**115**
7	**nodular**	**headneck**	**19**
8	**nodular**	**trunk**	**33**
9	**nodular**	**extremities**	**73**
10	**indeterminate**	**headneck**	**11**
11	**indeterminate**	**trunk**	**17**
12	**indeterminate**	**extremities**	**28**

21. R-code:

Full model:

> melanoma.glm1=glm(freq~type*site,family=poisson,data=melanoma.dat)

Reduced model:

> melanoma.glm2=glm(freq~type+site,family=poisson,data=melanoma.dat)

Minimum model:

> melanoma.glm3=glm(freq~1,family=poisson,data=melanoma.dat)

22. Interpretation of R outputs

> melanoma.glm2

Call: glm(formula = freq ~ type + site, family = poisson, data = melanoma.dat)

Coefficients:

(Intercept)	typeindeterminate	typenodular	typespread	sitetrunk	siteheadneck
2.9554	0.4990	1.3020	1.6940	-0.7571	-1.2010

Degrees of Freedom: 11 Total (i.e. Null); 6 Residual

Null Deviance: 295.2

Residual Deviance: 51.8 AIC: 122.9

Model setting in R is slightly different from the model in section 17. The program treats the category of "freckle" and "extremities" as the reference category – it picks up the cell combination according to the alphabetic order of the category names. The model is:

$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$, where i = "head & neck", and "trunk"

j = "spread", "nodular", and "indeterminate"

Therefore, the expected cell frequencies can be estimated as follow.

The expected frequency for "freckle" on "extremities": $e^{2.9554} = 19.21$

The expected frequency for "freckle" on the "head & neck" is: $e^{2.9554-1.201} = 5.78$.

The expected frequency for "spreading" on "trunk" is: $e^{2.9554+1.694-0.7571} = 49.02$.

These estimated frequencies can be confirmed by: > **fitted.values(melanoma.glm2)**. Or it can be confirmed using: **row total × column total/cell total**.

23. The test for the association between two classification factors can be easily solved using the contingency table analysis. The advantages of loglinear model are twofold. (1) It can give more detailed information about the relationship. For example, the results will tell you whether the occurrence of Indeterminate tumor is different across sites. (2) Loglinear model can easily handle more complicated problems with several

classification factors. It is easier to analyze multiple cross-tabulated data using loglinear models.

24. Note: contingency tables may include cells which cannot have any observations (e.g., male hysterectomy cases). This phenomenon, termed structural zeros, may not be easily incorporated in Poisson regression unless the parameters can be specified to accommodate the situation. See Agresti (1990) Categorical data analysis. Wiley, NY.