HW3

Jinhong Du - 12243476

2020/01/03

Contents

6.4. With three outcome categories and a single explanatory variable, suppose

$$\pi_{ij} = \frac{\exp(\beta_{j0} + \beta_j x_i)}{1 + \exp(\beta_{10} + \beta_1 x_i) + \exp(\beta_{20} + \beta_2 x_i)},$$

$j = 1, 2$. **Show that π_{i3} is (a) decreasing in x_i if $\beta_1 > 0$ and $\beta_2 > 0$, (b) increasing in x_i if $\beta_1 < 0$ and $\beta_2 < 0$, and (c) nonmonotone when β_1 and β_2 have opposite signs.**

Proof. Since

$$\pi_{i3} = 1 - \sum_{j=1}^{2} \pi_{ij} = \frac{1}{1 + \exp(\beta_{10} + \beta_1 x_i) + \exp(\beta_{20} + \beta_2 x_i)},$$

we have

$$\frac{\mathrm{d}\pi_{i3}}{\mathrm{d}x_i} = -\frac{\beta_1 \exp(\beta_{10} + \beta_1 x_i) + \beta_2 \exp(\beta_{20} + \beta_2 x_i)}{[1 + \exp(\beta_{10} + \beta_1 x_i) + \exp(\beta_{20} + \beta_2 x_i)]^2}$$

(a) If $\beta_1 > 0$ and $\beta_2 > 0$, then $\frac{\mathrm{d}\pi_{i3}}{\mathrm{d}x_i} < 0$. So π_{i3} is decreasing in x_i.

(b) If $\beta_1 < 0$ and $\beta_2 < 0$, then $\frac{\mathrm{d}\pi_{i3}}{\mathrm{d}x_i} > 0$. So π_{i3} is increasing in x_i.

(c) If $\beta_1\beta_2 < 0$, without loss of generality, suppose that $\beta_1 < 0$ and $\beta_2 > 0$. Consider the function $g(x_i) = \beta_1 \exp(\beta_{10} + \beta_1 x_i) + \beta_2 \exp(\beta_{20} + \beta_2 x_i)$. Since $\lim_{x \to +\infty} g(x) = +\infty$ and $\lim_{x \to -\infty} g(x) = -\infty$, $g(x_i)$ cannot be always positive or negative. So π_{i3} is nonmonotone. □

6.10. Does it make sense to use the cumulative logit model of proportional odds form with a nominal-scale response variable? Why or why not? Is the model a special case of a baseline-category logit model? Explain.

It does not make sense to use the cumulative logit model form with a nominal-scale response variable. Since in the cumulative logit model, we assume that the responses are ordinal data, the values of the categories matters in the sense that there exists some ordinal characteristics such as a monotone trend. If we fit the nominal data, we add extra assumption that the values of the responses and the number of parameters are increased.

In the baseline-category logit model,

$$\mathrm{logit}[\mathbb{P}(y_i \le k)] = \log\left(\frac{p_{i_1} + \cdots + p_{i_k}}{p_{i_{k+1}} + \cdots + p_{i_c}}\right) = \log\left(\frac{\frac{p_{i_1}}{p_{i_c}} + \cdots + \frac{p_{i_k}}{p_{i_c}}}{\frac{p_{i_{k+1}}}{p_{i_c}} + \cdots + 1}\right) = \log\left(\frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}_1} + \cdots + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}_k}}{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}_{k+1}} + \cdots + 1}\right)$$

which is not necessarily equal to $\alpha_k + \boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}$ for some α_k and $\tilde{\boldsymbol{\beta}}$. Therefore, the cumulative logit model is not a special case of a baseline-category logit model.

6.13. Consider the cumulative logit model, $\mathbf{logit}[\mathbb{P}(y_i \le j)] = \alpha_j + \beta_j x_i$.

(a) With continuous x_i taking values over the real line, show that the model is improper, in that cumulative probabilities are misordered for a range of x_i values.

For $j < k$, since $\mathbb{P}(y_i \le j) \le \mathbb{P}(y_i \le k)$, we have

$$\mathrm{logit}[\mathbb{P}(y_i \le j)] - \mathrm{logit}[\mathbb{P}(y_i \le k)] = (\alpha_j - \alpha_k) + (\beta_j - \beta_k)x_i \le 0.$$

However, if $\beta_j - \beta_k > 0$, then as $x_i \to +\infty$, $\mathrm{logit}[\mathbb{P}(y_i \le j)] - \mathrm{logit}[\mathbb{P}(y_i \le k)] \to +\infty$; if $\beta_j - \beta_k < 0$, then as $x_i \to -\infty$, $\mathrm{logit}[\mathbb{P}(y_i \le j)] - \mathrm{logit}[\mathbb{P}(y_i \le k)] \to +\infty$. So the cumulative probabilities are misordered for a range of x_i values.

(b) When x_i is a binary indicator, explain why the model is proper but requires constraints on $(\alpha_j + \beta_j)$ (as well as the usual ordering constraint on $\{\alpha_j\}$) and is then equivalent to the saturated model.

For binary data, we want $(\alpha_j - \alpha_k) + (\beta_j - \beta_k) \leq 0$, i.e., $(\alpha_j + \beta_j) - (\alpha_k + \beta_k) \leq 0$. So we need $\alpha_j + \beta_j$ to be monotone increasing in j.

We can organize the data as grouped form, that is, c columns $(y = 1, \ldots, c)$ and 2 rows $x = 0, 1$, then there are $2c$ many $\tilde{y}_{ij}$ in total to be estimated. The number of parameters in the cumulative logit model is also $2c$. So $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{\pi}}$, which is the saturate model.

6.17. A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, do not know). A two-part model uses a logistic regression model for the probability of a don't know response and a separate ordinal model for the ordered categories conditional on response in one of those categories. Explain how to construct a likelihood function to fit the two parts simultaneously.

Denote the order of categories `strongly agree`, `mildly agree`, `mildly disagree`, `strongly disagree` and `do not know` by $y_i = 4, 3, 2, 1, 0$ respectively. Then the model is given by

$$\begin{aligned}\text{logit}[\mathbb{P}(y_i \neq 0)] &= \beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta} \\ \text{logit}[\mathbb{P}(y_i \leq k | y_i \neq 0)] &= \alpha_k + \boldsymbol{X}_i^\top \boldsymbol{\eta}\end{aligned}$$

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta}, \boldsymbol{\eta} \in \mathbb{R}^p$, $\boldsymbol{X}_i \in \mathbb{R}^p$ does not include the intercept term. For the first part of the model, we only want to see if y_i equals to 0 or not. For the second part of the model, conditioning on $y_i \neq 0$, it becomes a cumulative logit model with categories $1, 2, 3, 4$. Let $y_{i_k} = \mathbb{1}_{y_i = k}$ for $k = 0, 1, \ldots, 4$, then the log-likelihood function is given by

$$\begin{aligned}L(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \log\left[\prod_{i=1}^{N}\prod_{k=0}^{4} \mathbb{P}(y_i = k)^{y_{i_k}}\right] \\ &= \sum_{i=1}^{N}\left\{y_{i_0}\log[\mathbb{P}(y_i = 0)] + \sum_{k=1}^{4} y_{i_k}\log[\mathbb{P}(Y_i = k | Y_i \neq 0)\mathbb{P}(y_i \neq 0)]\right\} \\ &= \sum_{i=1}^{N}\left\{y_{i_0}\log[\mathbb{P}(y_i = 0)] + \sum_{k=1}^{4} y_{i_k}\log[\mathbb{P}(Y_i = k | Y_i \neq 0)] + \sum_{k=1}^{4} y_{i_k}\log[\mathbb{P}(y_i \neq 0)]\right\} \\ &= \sum_{i=1}^{N}\Bigg\{\underbrace{y_{i_0}\log[\mathbb{P}(y_i = 0)] + \left(\sum_{k=1}^{4} y_{i_k}\right)\log[\mathbb{P}(y_i \neq 0)]}_{\text{log-likelihood of the logistic regression}} + \underbrace{\sum_{k=1}^{4} y_{i_k}\log[\mathbb{P}(Y_i = k | Y_i \neq 0)]}_{\substack{\text{log-likelihood of the} \\ \text{cumulative logit model}}}\Bigg\}.\end{aligned}$$

We can see that the log-likelihood function is just the mixture of log-likelihoods of the logistic regression and the cumulative logit model. So by taking the derivative with respect to the parameters, it will be the sum of two derivative of two parts. Then Newton's Method, Fisher Scoring Method or IRLS can be use to sovled for MLEs.

6.23. A 1976 article by M. Madsen (Scand. J. Stat. 3: 97–106) showed a $4 \times 2 \times 3 \times 3$ contingency table (the file `Satisfaction.dat` at the text website) that cross classifies a sample of residents of Copenhagen on the type of housing, degree of contact with other residents, feeling of influence on apartment management, and satisfaction with housing conditions. Treating satisfaction as the response variable, analyze these data.

First we relabel the categorical covariates and responses. Let random variables $y = 1, 2, 3$ denote low, medium and high satisfaction respectively, $X_1 = 1, 2$ denote low and high contact respectively, $X_2 = 1, 2, 3$ denote the low, medium and high influence and $X_3 = 1, 2, 3, 4$ denote `Tower blocks`, `Apartments`, `Atrium houses` and `Terraced`

`houses` respectively. Since the responses are ordinal, we should use the cumulative logit model,

$$\text{logit}[\mathbb{P}(y_i \le k)] = \alpha_k + \boldsymbol{X}_i^\top \boldsymbol{\beta}, \qquad k = 1, 2,$$

where $\boldsymbol{X}_i \in \mathbb{R}^3$ is the ith sample of $(X_1, X_2, X_3)^\top$ without the intercept.

Now we fit the corresponding model. The results below show very strong significance of the estimated coefficients.

```
N <- 3*3*2*4
df <- data.frame(Satisfaction=rep(1:3, N),
                 Contact = rep(rep(1:2, each=3), N/6),
                 Influence = rep(rep(1:3, each=6), N/18),
                 Housing = rep(1:4, each=N/4),
                 n=c(21,21,28,14,19,37,
                     34,22,36,17,23,40,
                     10,11,36,3,5,23,
                     61,23,17,78,46,43,
                     43,35,40,48,45,86,
                     26,18,54,15,25,62,
                     13,9,10,20,23,20,
                     8,8,12,10,22,24,
                     6,7,9,7,10,21,
                     18,6,7,57,23,13,
                     15,13,13,31,21,13,
                     7,5,11,5,6,13))
df2 <- aggregate(.~Contact+Influence+Housing, df, sum)[,1:3]
Contact = as.factor(df2$Contact)
Influence = as.factor(df2$Influence)
Housing = as.factor(df2$Housing)
for(i in 1:3){
    df2[paste0("y", i)] = aggregate(.~Contact+Influence+Housing,
                                    df[df['Satisfaction']==i,], sum)['n']
}

library(VGAM)

fit <- vglm(as.matrix(df2[,4:6]) ~ Contact + Influence + Housing,
            family = cumulative(link = "logitlink", parallel = T))
summary(fit)
```

```
##
## Call:
## vglm(formula = as.matrix(df2[, 4:6]) ~ Contact + Influence +
##     Housing, family = cumulative(link = "logitlink", parallel = T))
##
## Pearson residuals:
##                       Min      1Q   Median    3Q   Max
## logitlink(P[Y<=1]) -3.012 -0.9969 -0.38340 1.109 4.059
## logitlink(P[Y<=2]) -2.888 -1.5650 -0.07981 1.635 2.879
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.49614    0.07190  -6.900 5.20e-12 ***
## (Intercept):2  0.69071    0.07229   9.555  < 2e-16 ***
## Contact2      -0.36028    0.05505  -6.544 5.98e-11 ***
## Influence2    -0.56639    0.06060  -9.346  < 2e-16 ***
```

```
## Influence3    -1.28882    0.07315 -17.618  < 2e-16 ***
## Housing2       0.57235    0.06856   8.348  < 2e-16 ***
## Housing3       0.36619    0.09051   4.046 5.21e-05 ***
## Housing4       1.09101    0.08748  12.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 143.1829 on 40 degrees of freedom
##
## Log-likelihood: -197.2611 on 40 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##   Contact2 Influence2 Influence3   Housing2   Housing3   Housing4
##  0.6974781  0.5675685  0.2755961  1.7724273  1.4422234  2.9772934
```

7.7. For the one-way layout for Poisson counts, derive a test of $H_0 : \mu_1 = \cdots = \mu_c$ by applying a Pearson chi-squared goodness-of-fit test (with df $= c - 1$) for a multinomial distribution that compares sample proportions in c categories against H_0 values of multinomial probabilities, (a) when $n_1 = \cdots = n_c$, (b) for arbitrary $\{n_i\}$, with $n = \sum_i n_i$.

Let $Y_i \sim \text{Poisson}(\mu_i)$ be the random variable of counts for group i, $i = 1, \ldots, c$. For the one-way layout for a count response, let y_{ij} be observation j of Y_i in group i, $i = 1, \ldots, c$, $j = 1, \ldots, n_i$, with $n = \sum_{i=1}^{c} n_i$. Suppose that $\{y_{ij}\}$ are independent Poisson with $\mathbb{E}(y_{ij}) = \mu_i$. Let $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$.

Conditional on y_{++}, we have $(n_1 Y_1, \ldots, n_c Y_c) | \sum_{i=1}^{c} n_i Y_i = y_{++} \sim \text{Multinomial}(y_{++}, \boldsymbol{p})$, where $\boldsymbol{p} = \left(\frac{n_1 \mu_1}{\sum_{i=1}^{c} n_i \mu_i}, \ldots, \frac{n_c \mu_c}{\sum_{i=1}^{c} n_i \mu_i} \right)^\top \overset{H_0}{=\!=} \left(\frac{n_1}{n}, \ldots, \frac{n_c}{n} \right)^\top$.

(a) Under H_0, $\mu_1 = \cdots = \mu_c$ and $\frac{n_i}{n} = \frac{1}{c}$, the Pearson chi-squared goodness-of-fit test is given by

$$X^2 = \sum_{i=1}^{c} \frac{\left(y_{i+} - \frac{y_{++}}{c} \right)^2}{\frac{y_{++}}{c}}$$

which is approximately χ^2_{c-1} distributed under H_0.

(b) Under H_0, $\mu_1 = \cdots = \mu_c$, the Pearson chi-squared goodness-of-fit test is given by

$$X^2 = \sum_{i=1}^{c} \frac{y_{++} \left(\frac{y_{i+}}{y_{++}} - \frac{n_i}{n} \right)^2}{\frac{n_i}{n}} = \sum_{i=1}^{c} \frac{\left(y_{i+} - \frac{n_i y_{++}}{n} \right)^2}{\frac{n_i y_{++}}{n}}$$

which is approximately χ^2_{c-1} distributed under H_0.

7.15. For a $2 \times c \times l$ table, consider the loglinear model by which A is jointly independent of B and C. Treat A as a response variable and B and C as explanatory, conditioning on $\{n_{+jk}\}$. Construct the logit for the conditional distribution of A, and identify the corresponding logistic model.

Since A is jointly independent of B and C, we have

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}.$$

We are then modeling cl binomial distributions on A. For $j = 1, \ldots, c$ and $k = 1, \ldots, l$,

$$\begin{aligned}\text{logit}(\mathbb{P}(A = 1|B = j, C = k)) &= \log \frac{\mathbb{P}(A = 1|B = j, C = k)}{\mathbb{P}(A = 2|B = j, C = k)} \\ &= \log \frac{\mu_{1jk}}{\mu_{2jk}} = \log \mu_{1jk} - \log \mu_{2jk} \\ &= (\beta_0 + \beta_1^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}) \\ &\quad - (\beta_0 + \beta_2^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}) \\ &= \beta_1^A - \beta_2^A.\end{aligned}$$

Therefore, the logit model has the form as

$$\text{logit}(\mathbb{P}(A = 1|B = j, C = k)) = \lambda.$$

7.16. For the homogeneous association loglinear model (7.7) for a $r \times c \times l$ contingency table, treating A as a response variable, find the equivalent baseline category logit model.

For $i = 1, \ldots, c - 1$,

$$\begin{aligned}\log \frac{\mathbb{P}(A = i|B = j, C = k)}{\mathbb{P}(A = r|B = j, C = k)} &= \log \frac{\mu_{ijk}}{\mu_{cjk}} = \log \mu_{ijk} - \log \mu_{cjk} \\ &= (\beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ij}^{AB}) \\ &\quad - (\beta_0 + \beta_c^A + \beta_j^B + \beta_k^C + \gamma_{ck}^{AC} + \gamma_{jk}^{BC} + \gamma_{cj}^{AB}) \\ &= (\beta_i^A - \beta_c^A) + (\gamma_{ik}^{AC} - \gamma_{ck}^{AC}) + (\gamma_{ij}^{AB} - \gamma_{cj}^{AB}).\end{aligned}$$

Let $\pi_{i|jk} = \mathbb{P}(A = i|B = j, C = k)$ $(i = 1, \ldots, c)$, $\beta_{0i} = \beta_i^A - \beta_c^A$, $\beta_{1jk} = \gamma_{ik}^{AC} - \gamma_{ck}^{AC}$ and $\beta_{2jk} = \gamma_{ij}^{AB} - \gamma_{cj}^{AB}$. Therefore, the equivalent baseline category is given by

$$\frac{\pi_{i|jk}}{\pi_{i|jk} + \pi_{r|jk}} = \frac{e^{\beta_{0i} + \beta_{1jk} + \beta_{2jk}}}{1 + e^{\beta_{0i} + \beta_{1jk} + \beta_{2jk}}}.$$

7.20. A county's highway department keeps records of the number of automobile accidents reported each working day on a superhighway that runs through the county. Describe factors that are likely to cause the distribution of this count over time to show overdispersion relative to the Poisson distribution.

(1) Types of automobiles (car, truck, etc). The ratios of different types of automobiles are different and so does the rates of accidents.

(2) Weather situations (rainy day, sunny day, etc). In rainy days, accidents are easier to happen.

(3) Time. Accidents are more likely to happen at night when drivers cannot see the paths well.

These factors may contribute to a mixture of several Poisson populations with different means for the response. This heterogeneity results in an overall response distribution at that weight having greater variation than the Poisson, i.e., overdispersion.

7.28. Other than a formal goodness-of-fit test, one analysis that provides a sense of whether a particular GLM is plausible is the following: Suppose the ML fitted equation were the true equation. At the observed x values for the n observations, randomly generate n variates with distributions specified by the fitted GLM. Construct scatterplots. Do they look like the scatterplots that were actually observed? Do this for a Poisson loglinear model for the horseshoe crab data, with $y =$

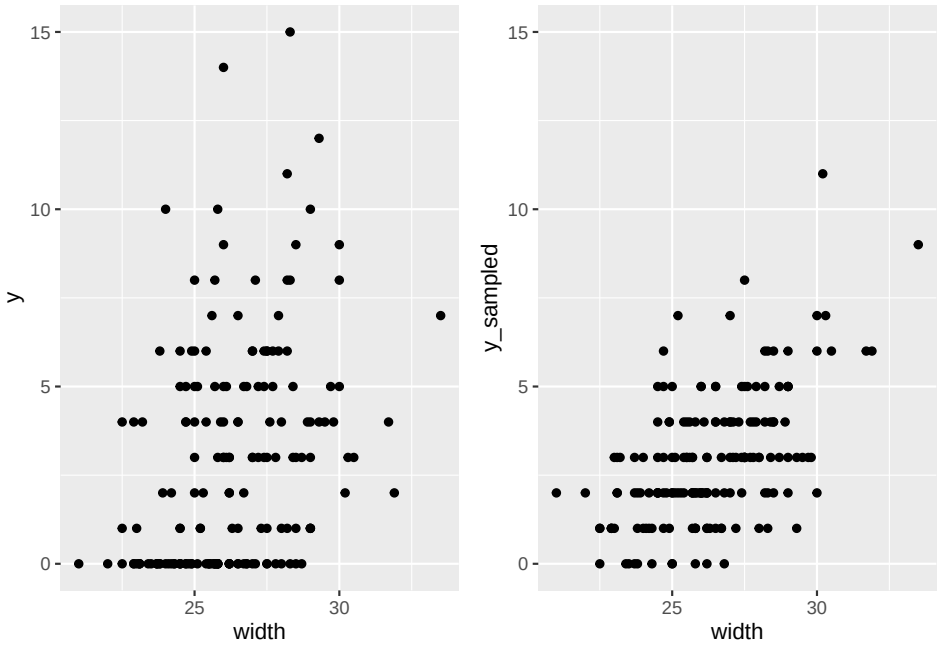
number of satellites and x = width. Does the variability about the fit resemble that in the actual data, including a similar number of 0's and large values? Repeat this a few times to get a better sense of how the scatterplot observed differs from what you would observe if the Poisson GLM truly held.

The scatterplots look like the below figure. As we can see, the variability varies in the two data. The sampled data has more zeros and large numbers. So this model may be unsuitable.

```
library(ggplot2)
library(gridExtra)

set.seed(0)
for(i in (1:10)){
    Crabs <- read.table("Crabs.dat", header=T)
    fit <- glm(y ~ width, family=poisson, data=Crabs)
    mu <- predict(fit, type='response')
    Crabs$y_sampled <- rpois(length(mu), mu)


    p1 <- ggplot(Crabs) + geom_point(aes(x=width, y=y)) + ylim(0,15)
    p2 <- ggplot(Crabs) + geom_point(aes(x=width, y=y_sampled)) + ylim(0,15)
    grid.arrange(p1, p2, ncol = 2)
}
```



7.30. A headline in *The Gainesville Sun* (February 17, 2014) proclaimed a worrisome spike in shark attacks in the previous 2 years. The reported total number of shark attacks in Florida per year from 2001 to 2013 were 33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23. Are these counts consistent with a null Poisson model or a null negative binomial model? Test the Poisson model against the negative binomial alternative. Analyze the evidence of a positive linear trend over time.

We fit the two models and run a likelihood ratio test. The p-value of the test is about $0.01 < 0.05$. Therefore, we cannot reject the null hypothesis that these counts are consistent with a null Poisson model.

```
library(MASS)
df <- data.frame(counts=c(33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23),
                 year=2001:2013)
fit_poisson <- glm(counts ~ 1, family=poisson, data=df)
fit_nb <- glm.nb(counts ~ 1, data = df)
pchisq(as.numeric(2 * (logLik(fit_nb) - logLik(fit_poisson))), df = 1, lower.tail = FALSE)
```

```
## [1] 0.01066321
```

Next we add the covariate x_i (year) and fit the Poisson GLM. From the summary below, we can see that β_0 and β_1 are both significant. Since $\hat{\beta}_1 < 0$ and is near 0, there is no positive linear trend over time.

```
fit_poisson <- glm(counts ~ year, family=poisson, data=df)
summary(fit_poisson)
```

```
##
## Call:
## glm(formula = counts ~ year, family = poisson, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8268  -1.4791   0.5308   1.0992   1.7207
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 66.89554   31.46400   2.126   0.0335 *
## year        -0.03178    0.01568  -2.027   0.0427 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 31.392  on 12  degrees of freedom
## Residual deviance: 27.267  on 11  degrees of freedom
## AIC: 95.003
##
## Number of Fisher Scoring iterations: 4
```