Lecture 2: Poisson and logistic regression

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton, UK

S^3RI, 11 - 12 December 2014

introduction to Poisson regression

application to the BELCAP study

introduction to logistic regression

confounding and effect modification

comparing of different generalized regression models

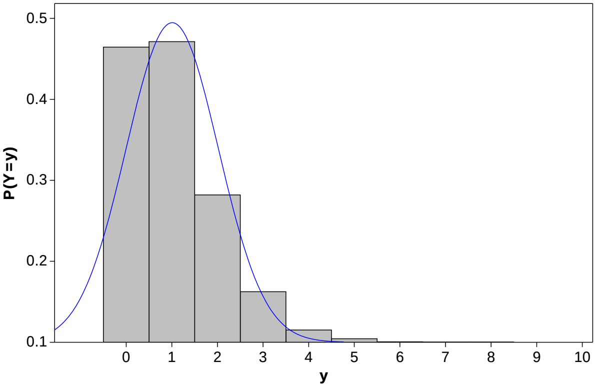
meta-analysis of BCG vaccine against tuberculosis

the Poisson distribution

- count data may follow such a distribution, at least approximately
- Examples: number of deaths, of diseased cases, of hospital admissions and so on ...
- $Y \sim Po(\mu)$:

$$P(Y = y) = \mu^y \exp(-\mu)/y!$$

where $\mu > 0$


0.5
0.4
0.3
0.2
0.1
P(Y=y)
0
1
2
3
4
5
6
7
8
9
10
y

but why not use a linear regression model?

- for a Poisson distribution we have $E(Y) = Var(Y)$. This violates the constancy of variance assumption (for the conventional regression model)
- a conventional regression model assumes we are dealing with a normal distribution for the response Y, but the Poisson distribution may not look very normal
- the conventional regression model may give negative predicted means (negative counts are impossible!)

the Poisson regression model

$$\log E(Y_i) = \log \mu_i = \alpha + \beta x_i$$

- the RHS of the above is called the **linear predictor**
- $Y_i \sim Po(\mu_i)$
- this model is the **log-linear model**

the Poisson regression model

$$\log E(Y_i) = \log \mu_i = \alpha + \beta x_i$$

can be written equivalently as

$$\mu_i = \exp[\alpha + \beta x_i]$$

Hence it is clear that any fitted log-linear model will always give non-negative fitted values!

an interesting interpretation in the Poisson regression model

suppose x represents a binary variable (yes/no, treatment present/not present)

$$x = \begin{cases} 1 & \text{if person \ is in intervention group} \\ 0 & \text{otherwise} \end{cases}$$

$$\log E(Y) = \log \mu = \alpha + \beta x$$

- $x = 0$: $\log \mu_{\text{no intervention}} = \alpha + \beta x = \alpha$
- $x = 1$: $\log \mu_{\text{intervention}} = \alpha + \beta x = \alpha + \beta$
- hence

$$\log \mu_{\text{intervention}} - \log \mu_{\text{no intervention}} = \beta$$

an interesting interpretation in the Poisson regression model

- hence

$$\log \mu_{\text{intervention}} - \log \mu_{\text{no intervention}} = \beta$$

- or

$$\frac{\mu_{\text{intervention}}}{\mu_{\text{no intervention}}} = \exp(\beta)$$

- the coefficient $\exp(\beta)$ corresponds to the **risk ratio** comparing the mean risk in the treatment group to the mean risk in the control group

Poisson regression model for several covariates

$$\log E(Y_i) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

- where $x_{1i}, \cdots, x_{pi}$ are the **covariates of interest**
- testing the effect of covariate x_j is done by the size of the estimate $\hat{\beta}_j$ of β_j

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

- if $|t_j| > 1.96$ covariate effect is **significant**

estimation of model parameters

consider the likelihood (the probability for the observed data)

$$L = \prod_{i=1}^{n} \mu_i^{y_i} \exp(-\mu_i)/y_i!$$

for model with p covariates:

$$\log \mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

- finding parameter estimates by maximizing the likelihood L (or equivalently the log-likelihood $\log L$)
- guiding principle: choosing the parameters that make the observed data the most likely

The simple regression model for BELCAP

with $Y = DMFSe$:

$$\log E(DMFSe_i) =$$

$$\alpha + \beta_1 OHE_i + \beta_2 ALL_{2i} + \beta_4 ESD_i + \beta_5 MW_i + \beta_6 OHY_i + \beta_7 DMFSb_i$$

- $OHE_i = \begin{cases} 1 & \text{if child } i \text{ is in intervention OHE} \\ 0 & \text{otherwise} \end{cases}$
- $ALL_i = \begin{cases} 1 & \text{if child } i \text{ is in intervention ALL} \\ 0 & \text{otherwise} \end{cases}$
- ...

analysis of BELCAP study using the Poisson regression model including the DMFS at baseline

covariate	$\hat{\beta}_j$	$s.e.(\hat{\beta}_j)$	t_j	P-value
OHE	-0.7043014	0.0366375	-6.74	0.000
ALL	-0.5729402	0.0355591	-8.97	0.000
ESD	-0.8227017	0.0418510	-3.84	0.000
MW	-0.6617572	0.0334654	-8.16	0.000
OHY	-0.7351562	0.0402084	-5.63	0.000
DMFSb	1.082113	0.0027412	31.15	0.000

Introduction to logistic regression

Binary Outcome Y

$$Y = \begin{cases} 1, & \text{Person diseased} \\ 0, & \text{Person healthy} \end{cases}$$

Probability that Outcome $Y = 1$

$Pr(Y = 1) = p$ is probability for $Y = 1$

Odds

$$odds = \frac{p}{1-p} \Leftrightarrow p = \frac{odds}{odds+1}$$

Examples

- $p = 1/2 \Rightarrow odds = 1$
- $p = 1/4 \Rightarrow odds = 1/3$
- $p = 3/4 \Rightarrow odds = 3/1 = 3$

Odds Ratio

$$
\begin{aligned}
OR &= \frac{odds(\text{ in exposure })}{odds(\text{ in non-exposure })} \\
&= \frac{p_1/(1-p_1)}{p_0/(1-p_0)}
\end{aligned}
$$

Properties of odds ratio

- $0 < OR < \infty$
- $OR = 1(p_1 = p_0)$ is reference value

Examples

$$\text{risk} = \begin{cases} p_1 = 1/4 \\ p_0 = 1/8 \end{cases} \quad \text{effect measure} = \begin{cases} OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{1/3}{1/7} = 2.33 \\ RR = \frac{p_1}{p_0} = 2 \end{cases}$$

$$\text{risk} = \begin{cases} p_1 = 1/100 \\ p_0 = 1/1000 \end{cases} \quad \text{eff. meas.} = \begin{cases} OR = \frac{1/99}{1/999} = 10.09 \\ RR = \frac{p_1}{p_0} = 10 \end{cases}$$

Fundamental Theorem of Epidemiology

$$p_0 \text{ small} \Rightarrow OR \approx RR$$

benefit: OR is interpretable as RR which is easier to deal with

A simple example: Radiation Exposure and Tumor Development

	cases	non-cases	
E	52	2820	2872
NE	6	5043	5049

odds and *OR*

odds for disease given exposure (in detail):

$$\frac{52/2872}{2820/2872} = 52/2820$$

odds for disease given non-exposure (in detail):

$$\frac{6/5049}{5043/5049} = 6/5043$$

A simple example: Radiation Exposure and Tumor Development

	cases	non-cases	
E	52	2820	2872
NE	6	5043	5049

OR

odds ratio for disease (in detail):

$$OR = \frac{52/2820}{6/5043} = \frac{52 \times 5043}{6 \times 2820} = 15.49$$

or, $\log OR = \log 15.49 = 2.74$

for comparison

$$RR = \frac{52/2872}{6/5049} = 15.24$$

Logistic regression model for this simple situation

$$\log \frac{p_x}{1-p_x} = \alpha + \beta x$$

where

- $p_x = Pr(Y = 1|x)$
- $x = \begin{cases} 1, & \text{if exposure present} \\ 0, & \text{if exposure not present} \end{cases}$
- $\log \frac{p_x}{1-p_x}$ is called the **logit link** that connects p_x with the linear predictor

benefits of the logistic regression model

$$\log \frac{p_x}{1-p_x} = \alpha + \beta x$$

is **feasible**

- since

$$p_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \in (0, 1)$$

  whereas

$$p_x = \alpha + \beta x$$

  is **not feasible**

Interpretation of parameters α and β

$$\log \frac{p_x}{1-p_x} = \alpha + \beta x$$

$$x = 0 : \log \frac{p_0}{1-p_0} = \alpha \quad (1)$$

$$x = 1 : \log \frac{p_1}{1-p_1} = \alpha + \beta \quad (2)$$

now

$$(2) - (1) = \underbrace{\log \frac{p_1}{1-p_1} - \log \frac{p_0}{1-p_0}}_{\log \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \log OR} = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

A simple illustration example

	cases	non-cases	
E	60	1100	1160
NE	1501	3100	4601

OR

odds ratio:

$$OR = \frac{60 \times 3100}{1501 \times 1100} = 0.1126$$

stratified:

Stratum 1:

	cases	non-cases	
E	50	100	150
NE	1500	3000	4500

$$OR = \frac{50 \times 3000}{100 \times 1500} = 1$$

Stratum 2:

	cases	non-cases	
E	10	1000	1010
NE	1	100	101

$$OR = \frac{10 \times 100}{1000 \times 1} = 1$$

	Y	E	S	freq
1.	1	1	0	50
2.	0	1	0	100
3.	1	0	0	1500
4.	0	0	0	3000
5.	1	1	1	10
6.	0	1	1	1000
7.	1	0	1	1
8.	0	0	1	100

The logistic regression model for simple confounding

$$\log \frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}} = \alpha + \beta E + \gamma S$$

where

$$\mathbf{x} = (E, S)$$

is the covariate combination of exposure E and stratum S

in detail for stratum 1

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S$$

$$E = 0, S = 0 : \log \frac{p_{0,0}}{1-p_{0,0}} = \alpha \quad (3)$$

$$E = 1, S = 0 : \log \frac{p_{1,0}}{1-p_{1,0}} = \alpha + \beta \quad (4)$$

now

$$(4) - (3) = \log OR_1 = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

the log-odds ratio in the first stratum is β

in detail for stratum 2:

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S$$

$$E=0, S=1 : \log \frac{p_{0,1}}{1-p_{0,1}} = \alpha + \gamma \quad (5)$$

$$E=1, S=1 : \log \frac{p_{1,1}}{1-p_{1,1}} = \alpha + \beta + \gamma \quad (6)$$

now:

$$(6) - (5) = \log OR_2 = \alpha + \beta + \gamma - \alpha - \gamma = \beta$$

the log-odds ratio in the second stratum is β

important property of the confounding model:

assumes the identical exposure effect in each stratum!

```
(crude analysis) Logistic regression
Log likelihood = -3141.5658
------------------------------------------------------------
  Y | Odds Ratio   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------
  E |   .1126522   .0153479       .0862522    .1471326
------------------------------------------------------------


(adjusted for confounder) Logistic regression
Log likelihood = -3021.5026
------------------------------------------------------------
 Y | Odds Ratio   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------
 E |          1   .1736619       .7115062    1.405469
 S |        .02   .0068109       .0102603    .0389853
------------------------------------------------------------
```

A simple illustration example: passive smoking and lung cancer

	cases	non-cases	
E	52	121	173
NE	54	150	204

OR
odds ratio:

$$OR = \frac{52 \times 150}{54 \times 121} = 1.19$$

stratified:

Stratum 1 (females):

	cases	non-cases	
E	41	102	143
NE	26	71	97

$$OR = \frac{41 \times 71}{26 \times 102} = 1.10$$

Stratum 2 (males):

	cases	non-cases	
E	11	19	30
NE	28	79	107

$$OR = \frac{11 \times 79}{19 \times 28} = 1.63$$

interpretation:

effect changes from one stratum to the next stratum!

The logistic regression model for effect modification

$$\log \frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}} = \alpha + \beta E + \gamma S + \underbrace{(\beta\gamma)}_{\text{effect modif. par.}} E \times S$$

where

$$\mathbf{x} = (E, S)$$

is the covariate combination of exposure E and stratum S

in detail for stratum 1

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S + (\beta\gamma) E \times S$$

$$E = 0, S = 0 : \log \frac{p_{0,0}}{1-p_{0,0}} = \alpha \quad (7)$$

$$E = 1, S = 0 : \log \frac{p_{1,0}}{1-p_{1,0}} = \alpha + \beta \quad (8)$$

now

$$(8) - (7) = \log OR_1 = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

the log-odds ratio in the first stratum is β

in detail for stratum 2:

$$\log \frac{p_x}{1-p_x} = \alpha + \beta E + \gamma S + (\beta\gamma) E \times S$$

$$E = 0, S = 1 : \log \frac{p_{0,1}}{1-p_{0,1}} = \alpha + \gamma \quad (9)$$

$$E = 1, S = 1 : \log \frac{p_{1,1}}{1-p_{1,1}} = \alpha + \beta + \gamma + (\beta\gamma) \quad (10)$$

now:

$$(10) - (9) = \log OR_2 = \alpha + \beta + \gamma + (\beta\gamma) - \alpha - \gamma = \beta + (\beta\gamma)$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta + (\beta\gamma)}$$

the log-odds ratio in the second stratum is $\beta + (\beta\gamma)$

important property of the effect modification model:

effect modification model allows for different effects in the strata!

Data from passive smoking and LC example are as follows:

	Y	E	S	ES	freq
1.	1	1	0	0	41
2.	0	1	0	0	102
3.	1	0	0	0	26
4.	0	0	0	0	71
5.	1	1	1	1	11
6.	0	1	1	1	19
7.	1	0	1	0	28
8.	0	0	1	0	79

```
CRUDE EFFECT MODEL

Logistic regression

Log likelihood = -223.66016

------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
           E |   .1771044   .2295221     0.77   0.440
       _cons |  -1.021651   .1586984    -6.44   0.000
------------------------------------------------------
```

CONFOUNDING MODEL

Logistic regression

Log likelihood = -223.56934

Y	Coef.	Std. Err.	z	P>|z|
E	.2158667	.2472221	0.87	0.383
S	.1093603	.2563249	0.43	0.670
_cons	-1.079714	.2101705	-5.14	0.000

EFFECT MODIFICATION MODEL

Logistic regression

Log likelihood = -223.2886

Y	Coef.	Std. Err.	z	P>|z|
E	.0931826	.2945169	0.32	0.752
S	-.03266	.3176768	-0.10	0.918
ES	.397517	.5278763	0.75	0.451
_cons	-1.004583	.2292292	-4.38	0.000

interpretation of crude effects model:

$$\log OR = 0.1771 \Leftrightarrow OR = e^{0.1771} = 1.19$$

interpretation of confounding model:

$$\log OR = 0.2159 \Leftrightarrow OR = e^{0.2159} = 1.24$$

interpretation of effect modification model:

stratum 1:

$$\log OR_1 = 0.0932 \Leftrightarrow OR_1 = e^{0.0932} = 1.10$$

stratum 2:

$$\log OR_2 = 0.0932 + 0.3975 \Leftrightarrow OR_2 = e^{0.0932+0.3975} = 1.63$$

Model evaluation in logistic regression:

the likelihood approach:

$$L = \prod_{i=1}^{n} p_{x_i}^{y_i}(1 - p_{x_i})^{1-y_i}$$

is called the **likelihood** for models

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \begin{cases} \alpha + \beta E_i + \gamma S_i + (\beta\gamma) E_i \times S_i, \ (M_1) \\ \alpha + \beta E_i + \gamma S_i, \ (M_0) \end{cases}$$

where M_1 is the effect modification model and M_0 is the confounding model

Model evaluation in logistic regression using the likelihood ratio:

let

$$L(M_1) \text{ and } L(M_0)$$

be the **likelihood** for models M_1 and M_0
then

$$LRT = 2\log L(M_1) - 2\log L(M_0) = 2\log \frac{L(M_1)}{L(M_0)}$$

is called the **likelihood ratio** for models M_1 and M_0 and has a **chi-square distribution with 1** *df* under M_0

illustration for passive smoking and LC example:

model	log-likelihood	LRT
crude	-223.66016	-
homogeneity	-223.56934	0.1816
effect modification	-223.2886	0.5615

note:

for valid comparison on chi-square scale: models must be **nested**

Model evaluation in more general:

consider the likelihood

$$L = \prod_{i=1}^{n} p_{x_i}^{y_i}(1 - p_{x_i})^{1-y_i}$$

for a general model with p **covariates**:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \; (M_0)$$

example:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 \mathrm{AGE}_i + \beta_2 \mathrm{SEX}_i + \beta_3 \mathrm{ETS}_i$$

Model evaluation in more general:

example:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 \mathrm{AGE}_i + \beta_2 \mathrm{SEX}_i + \beta_3 \mathrm{ETS}_i$$

where these covariates can be mixed:

- quantitative, continuous such as AGE
- categorical binary (use 1/0 coding) such as SEX
- non-binary ordered or unordered categorical such as ETS (none, moderate, large)

Model evaluation in more general:

consider the likelihood

$$L = \prod_{i=1}^{n} p_{x_i}^{y_i}(1 - p_{x_i})^{1-y_i}$$

for model with **additional k covariates**:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

$$+\beta_{p+1} x_{i,p+1} + ... + \beta_{k+p} x_{i,k+p} \ (M_1)$$

Model evaluation in more general for our example:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{SEX}_i + \beta_3 \text{ETS}_i$$
$$+ \beta_4 \text{RADON}_i + \beta_5 \text{AGE-HOUSE}_i$$

Model evaluation using the likelihood ratio:

again let

$$L(M_1) \text{ and } L(M_0)$$

be the **likelihood** for models M_1 and M_0
then the **likelihood ratio**

$$LRT = 2 \log L(M_1) - 2 \log L(M_0) = 2 \log \frac{L(M_1)}{L(M_0)}$$

has a **chi-square distribution with k** *df* under M_0

Model evaluation for our example:

$$\begin{cases} M_0: & \alpha + \beta_1 \text{AGE}_i + \beta_2 \text{SEX}_i + \beta_3 \text{ETS}_i \\ M_1: & ...M_0... + \beta_4 \text{RADON}_i + \beta_5 \text{AGE-HOUSE}_i \end{cases}$$

then

$$LRT = 2 \log \frac{L(M_1)}{L(M_0)}$$

has under model M_0 a chi-square distribution with 2 df

model evaluation

- for model assessment we will use criteria that compromise between **model fit** and **model complexity**
- Akaike information criterion

$$AIC = -2 \log L + 2k$$

- Bayesian Information criterion

$$BIC = -2 \log L + k \log n$$

- where k is the number of parameters in the model
- and n is the number of clustered observations
- we seek a model for which AIC and/or BIC are small

Meta-Analysis

Meta-Analysis is a methodology for investigating the study results from several, independent studies with the purpose of an integrative analysis

Meta-Analysis on BCG vaccine against tuberculosis

Colditz *et al. 1974, JAMA* provide a meta-analysis to examine the efficacy of BCG vaccine against tuberculosis

Data on the meta-analysis of BCG and TB

the data contain the following details

- 13 studies
- each study contains:
  - TB cases for BCG intervention
  - number at risk for BCG intervention
  - TB cases for control
  - number at risk for control
- also two covariates are given: *year of study* and *latitude expressed in degrees from equator*
- latitude represents the variation in rainfall, humidity and environmental mycobacteria suspected of producing immunity against TB

			intervention		control	
study	year	latitude	TB cases	total	TB cases	total
1	1933	55	6	306	29	303
2	1935	52	4	123	11	139
3	1935	52	180	1541	372	1451
4	1937	42	17	1716	65	1665
5	1941	42	3	231	11	220
6	1947	33	5	2498	3	2341
7	1949	18	186	50634	141	27338
8	1950	53	62	13598	248	12867
9	1950	13	33	5069	47	5808
10	1950	33	27	16913	29	17854
11	1965	18	8	2545	10	629
12	1965	27	29	7499	45	7277
13	1968	13	505	88391	499	88391

Data analysis on the meta-analysis of BCG and TB

these kind of data can be analyzed by taking

- *TB case* as disease occurrence response
- *intervention* as exposure
- *study* as confounder

```
Log likelihood = -15191.497                       Pseudo R2       =     0.0050
```

TB_Case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Intervention	.6116562	.024562	-12.24	0.000	.5653613	.6617421
_cons	.0091641	.0002369	-181.51	0.000	.0087114	.0096404

```
. estat ic, n(13)
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	13	-15267.81	-15191.5	2	30386.99	30388.12

Note: N=13 used in calculating BIC

```
Logistic regression                               Number of obs   =     357347
                                                  LR chi2(2)      =    1239.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -14648.082                       Pseudo R2       =     0.0406
```

TB_Case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Latitude	1.043716	.00126	35.44	0.000	1.04125	1.046189
Intervention	.6253014	.0251677	-11.67	0.000	.577869	.6766271
_cons	.0031643	.0001403	-129.85	0.000	.002901	.0034515

```
. estat ic, n(13)
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	13	-15267.81	-14648.08	3	29302.16	29303.86

Note: N=13 used in calculating BIC

```
Logistic regression                               Number of obs   =     357347
                                                  LR chi2(3)      =    1402.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -14566.659                       Pseudo R2       =     0.0459
```

TB_Case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Latitude	1.029997	.0016409	18.55	0.000	1.026786	1.033219
Intervention	.6041037	.0243883	-12.48	0.000	.5581456	.6538459
Year	.9666536	.0025419	-12.90	0.000	.9616844	.9716485
_cons	.0300164	.0053119	-19.81	0.000	.021219	.0424611

```
. estat ic, n(13)
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	13	-15267.81	-14566.66	4	29141.32	29143.58

Note: N=13 used in calculating BIC

.

model evaluation

model	log L	AIC	BIC
intervention	-15191.50	30386.99	30388.12
+ latitude	-14648.08	29302.16	29303.86
+ year	-14566.66	29141.32	29143.58