Course > Unit 7 Generalized Linear Models > Homework 11 > 1. Poisson regression

1. Poisson regression

*Instructions: For this problem, whenever a formula box requires you to enter a factorial, enter **fact** to indicate the factorial function. For instance, **fact(10)** denotes $10!$.*

(a)

5/5 points (graded)

We want to model the rate of infection with an infectious disease depending on the day after outbreak t. Denote the recorded number of outbreaks at day t by k_t.

We are going to model the distribution of k_t as a Poisson distribution with a time-varying parameter λ_t, which is a common assumption when handling count data.

First, recall the likelihood of a Poisson distributed random variable Y in terms of the parameter λ,

$$\mathbf{P}\left(Y=k\right)=\mathbf{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

Rewrite this in terms of an exponential family. In other words, write it in the form

$$\mathbf{P}\left(Y=k\right)=h\left(k\right)\exp\left[\eta\left(\lambda\right)T\left(k\right)-B\left(\lambda\right)\right].$$

Since this representation is only unique up to re-scaling by constants, take the convention that $T\left(k\right)=k$.

$\eta\left(\lambda\right)=$ ln(lambda) ✔ **Answer:** ln(lambda)

$B\left(\lambda\right)=$ lambda ✔ **Answer:** lambda

$h\left(k\right)=$ 1/(fact(k)) ✔ **Answer:** 1/(fact(k))

We can write this in canonical form, e.g. as

$$\mathbf{P}\left(Y=k\right)=h\left(k\right)\exp\left[k\eta-b\left(\eta\right)\right].$$

What is $b\left(\eta\right)$?

$b\left(\eta\right)=$ exp(eta) ✔ **Answer:** exp(eta)

Recall that the mean of a Poisson(λ) distribution is λ. What is the canonical link function $g\left(\mu\right)$ associated with this exponential family, where $\mu=\mathbb{E}\left[Y\right]$? Write your answer in terms of λ.

$g\left(\mu\right)=$ ln(lambda) ✔ **Answer:** ln(lambda)

Solution:

We can rewrite the likelihood as

$$\mathbf{P}\left(Y=k\right)=\quad\mathbf{e}^{-\lambda}\frac{\lambda^k}{k!}$$

$$= \frac{1}{k!}\exp\left[-\lambda + k\ln(\lambda)\right].$$

Hence, given the convention $T(k) = k$ for this specific case, we set

$$\begin{aligned} h(k) &= \frac{1}{k!} \\ B(\lambda) &= \lambda \\ \eta(\lambda) &= \ln(\lambda). \end{aligned}$$

In order to rewrite this in canonical form, solve

$$\ln(\lambda) = \eta \iff \lambda = \mathbf{e}^{\eta},$$

so

$$b(\eta) = \mathbf{e}^{\eta}.$$

The canonical link function is b'^{-1}, which is

$$b'^{-1}(\mu) = \ln(\mu)$$

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

(b)

2/2 points (graded)

What range will the values in Y belong to?

- $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- $\mathbb{Z}_{+} = \{1, 2, 3, \ldots\}$
- $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$ ✔
- $\mathbb{R}$
- $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$
- $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$

According to the canonical Generalized Linear Model (your answer from (a)), what is the range of possible predictions for λ?

- $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- $\mathbb{Z}_{+} = \{1, 2, 3, \ldots\}$
- $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$

- $\mathbb{R}$
- $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$
- $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$ ✔

Solution:

Y has the Poisson distribution, so it lives in $\{0, 1, 2, \ldots\}$.

Since the canonical model states $\lambda = e^{\eta}$, the range of λ_t is the full range of parameters for a Poisson distribution: $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$.

Submit   You have used 3 of 3 attempts

Answers are displayed within the problem

(c)

2/3 points (graded)

Return to the original model. We now introduce a Poisson intensity parameter λ_t for every time point and denote the parameter (η) that gives the canonical exponential family representation as above by θ_t . We choose to employ a linear model connecting the time points t with the canonical parameter θ of the Poisson distribution above, i.e.,

$$\theta_t = a + bt.$$

In other words, we choose a generalized linear model with Poisson distribution and its canonical link function. That also means that conditioned on t, we assume the Y_t to be independent.

Imagine we observe the following data:

$t_1 = 1$ 1 outbreaks

$t_2 = 2$ 3 outbreaks

$t_3 = 4$ 10 outbreaks

We want to produce a maximum likelihood estimator for (a, b). To this end, write down the log likelihood $\ell(a, b)$ of the model for the provided three observations at t_1, t_2, and t_3 (plug in their values).

$\ell(a, b) =$

14*a+47*b-exp(a+b)-exp(a+2*b)-exp(a+4*b) + ln(1/6) + ln(1/362880) ✖

Answer: -ln(6)-ln(fact(10))-exp(a+b)-exp(a+2*b)-exp(a+4*b)+(14*a)+(47*b)

$$14 \cdot a + 47 \cdot b - \exp(a+b) - \exp(a + 2 \cdot b) - \exp(a + 4 \cdot b) + \ln\left(\frac{1}{6}\right) + \ln\left(\frac{1}{362880}\right)$$

What is its gradient? Enter your answer as a pair of derivatives.

$\partial_a \ell(a, b) =$

14 - exp(a + b) - exp(a + 2*b) - exp(a + 4*b) ✔

Answer: -exp(a+b)-exp(a+2*b)-exp(a+4*b)+14

$$14 - \exp(a+b) - \exp(a + 2 \cdot b) - \exp(a + 4 \cdot b)$$

$\partial_b \ell(a,b) =$

47 - exp(a + b) - 2*exp(a + 2*b) - 4*exp(a + 4*b)

Answer: -exp(a+b)-2*exp(a+2*b)-4*exp(a+4*b)+47

$47 - \exp(a+b) - 2 \cdot \exp(a + 2 \cdot b) - 4 \cdot \exp(a + 4 \cdot b)$

Solution:

The likelihood for one observation is given by

$$\mathbf{P}(Y_t = k_t) = \frac{1}{k_t!} \exp\left[-\exp(a+bt) + k_t(a+bt)\right].$$

That means the log likelihood for the model for n observations is

$$\ell(a,b) = \sum_{i=1}^{n} \left[-\ln(k_t!) - \exp(a+bt_i) + k_{t_i}(a+bt_i)\right].$$

Plugging in the provided values, we get

$$\begin{aligned}\ell(a,b) = \; & -\ln(1!) - \ln(3!) - \ln(10!) \\ & -\exp(a+b) - \exp(a+2b) - \exp(a+4b) \\ & +14a + 47b.\end{aligned}$$

Its derivative with respect to a is

$$\partial_a \ell(a,b) = \; -\exp(a+b) - \exp(a+2b) - \exp(a+4b) + 14.$$

Its derivative with respect to b is

$$\partial_b \ell(a,b) = \; -\exp(a+b) - 2\exp(a+2b) - 4\exp(a+4b) + 47.$$

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

(d)

1/1 point (graded)

In order to find the maximum likelihood estimator, we have to solve the nonlinear equation

$$\nabla \ell(a,b) = 0,$$

which in general does not have a closed solution.

Assume that we can reasonably estimate the likelihood estimator using numerical methods, and we obtain

$$\widehat{a} \approx -0.43, \quad \hat{b} \approx 0.69.$$

Given these results, what would be the predicted expected number of outbreaks for $t = 3$? Round your answer to the nearest 0.001.

5.1552

Answer: 5.1551695

Solution:

We obtain the expected number of outbreaks as

$$\mathbb{E}\left[Y_t|t\right] = \lambda_t,$$

since the expectation of a Poisson random variable is equal to its rate parameter. With this and the relation $\lambda_t = \exp\left(a + bt\right)$, we obtain the prediction

$$\lambda_3 = \exp\left(\widehat{a} + \hat{b}t\right) \approx 5.1551695.$$

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 7 Generalized Linear Models:Homework 11 / 1. Poisson regression