北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

Generalized Linear Models

Lecture 5: Count responses

Outline

Poisson distribution

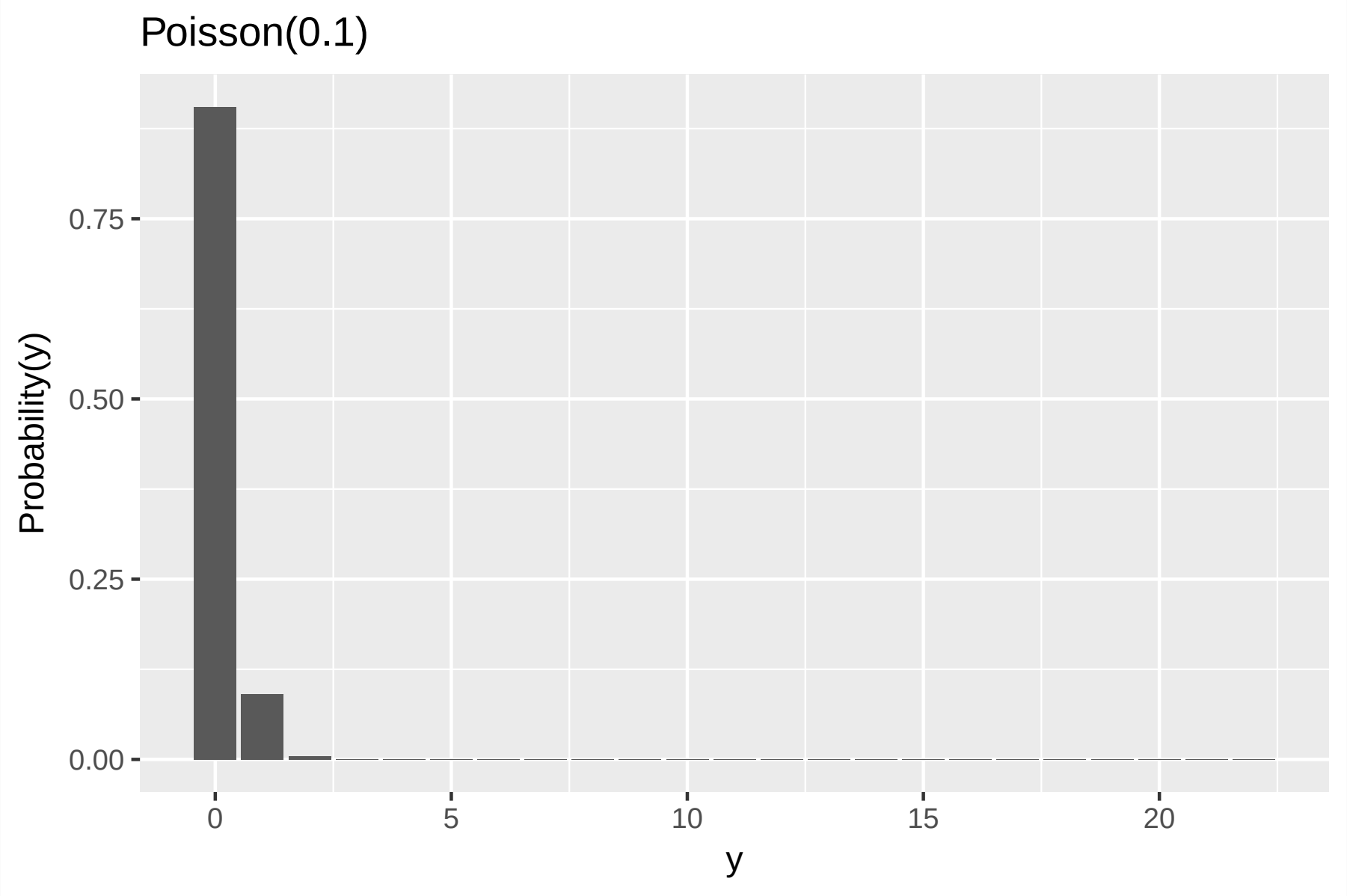
Let Y = number of events in given time interval. If events independent, and prob of event proportional to length of interval, then Y is Poisson distributed.

Poisson(μ) distribution

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- $E(Y) = V(Y) = \mu$
- If $Y \sim B(n, p)$, then $Y \approx \text{Poisson}(np)$ for small p/n.
- If $Y \sim \text{Poisson}(\mu)$, then $Y \approx N(\mu, \mu)$ for large μ.
- $\text{Poisson}(\mu_1) + \text{Poisson}(\mu_1) \sim \text{Poisson}(\mu_1 + \mu_2)$.

Poisson distribution



Regression with count data

Suppose response Y is a count $(0,1,2,\ldots)$.

- If count is bounded and bound is small, use binomial regression.
- If min count is large, use normal approximation.
- Otherwise, use Poisson or negative binomial.

Poisson regression

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

- Log link function forces positive mean.
- Log-likelihood:

$$\log L = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \beta - \exp(\mathbf{x}_i^T \beta) - \log(y_i!) \right]$$

```
fit <- glm(y ~ x1 + x2,
  family='poisson', data)
```

Deviance

$$D = 2\sum_{i=1}^{n}(y_i \log(y_i/\hat{\mu}) - (y_i - \hat{\mu}_i))$$

- Check distributional assumptions by comparing D against χ^2
- Compare changes in deviance using a χ^2 test as for binomial regression.
- Use profile likelihood to find confidence intervals for parameters.
- Common for a model to be “over-dispersed”.

Outline

Dispersed Poisson model

When model is over-dispersed (variance too large):

- estimates of β consistent, but standard errors incorrect.
- could correct model using negative binomial, or quasi-Poisson model.

Quasi-Poisson models

```
fit <- glm(y ~ x1 + x2,
  family='quasipoisson', data)
```

- Use *F*-tests not χ^2 tests when using quasi-Poisson models
- Overdispersion parameter represents the variance inflation

Negative binomial distribution

In a series of independent trials, each with probability of success p, let Y be the number of fails until the kth success.

$$P(Y = y) = \binom{y + k - 1}{k - 1} p^k (1 - p)^y, \qquad y = 0, 1, \ldots$$

- $\mathrm{E}(Y) = \mu = k(1 - p)/p$ and $\mathrm{V}(Y) = \mu + \mu^2/k$.

Negative binomial regression

We can link μ_i to a linear combination of X:

$$\eta_i = \log\left(\frac{\mu_i}{\mu_i + k}\right) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

- k (the "dispersion" parameter) is usually estimated along with the coefficients by MLE:

$$\log L = \sum_{i=1}^{n} \left(y_i \log\left(\frac{\mu_i}{\mu_i + k}\right) - k\log(1 + \mu_i/k) + \sum_{j=0}^{y_i - 1} \log(j + k) - \log(y_i!) \right)$$

Negative binomial regression

```
fit <- glm(y ~ x1 + x2,
  family=negative.binomial(k), data)

fit <- MASS::glm.nb(y ~ x1 + x2, data)
```

Outline

Zero inflated count models

Examples of zero-inflated data:

- Number of insurance claims for each account
- Number of arrests for criminal offences for each individual
- Number of articles written by PhD students

Over-dispersed models do not deal adequately with this type of data.

Zero-inflated count models

Solution 1: Hurdle model

- Model for probability of zero (logistic).
- Model for non-zero counts (truncated Poisson).

$$P(Y = 0) = f_1(0)$$

$$P(Y = j) = (1 - f_1(0))\frac{f_2(j)}{1 - f_2(0)}, \qquad j > 0$$

- f_1 is binomial probability; f_2 is Poisson probability.
- Two sets of coefficients for the two parts of the model.

```
fit <- pscl::hurdle(y ~ x1 + x2, data)
```

Zero-inflated count models

Solution 2: Mixture model

- Model for probability of always zero (logistic).
- Model for counts (regular Poisson).

$$P(Y = 0) = \phi + (1 - \phi)f(0)$$

$$P(Y = j) = (1 - \phi)f(j), \qquad j > 0$$

- ϕ is probability of zero; f is Poisson probability.
- Two sets of coefficients for the two parts of the model.

```
fit <- pscl::zeroinfl(y ~ x1 + x2, data)
```

Zero-inflated count models

- Often difficult to select between these – what makes most sense for the application?
- Can have different predictors for the two sub-models:

```
fit <- zeroinfl(y ~ x1 + x2 | x3, data)
```

count model before the `|` and zero model after.

Outline

Conclusion

- The basic GLM for count data is the Poisson model with log link.
- Frequently, however, when the response variable is a count, its conditional variance increases more rapidly than its mean, producing a condition termed overdispersion, and invalidating the use of the Poisson distribution. The quasi-Poisson GLM adds a dispersion parameter to handle overdispersed count data; this model can be estimated by the method of quasi-likelihood.
- A similar model is based on the negative-binomial distribution, which is not an exponential family. Negative-binomial GLMs can nevertheless be estimated by maximum likelihood.
- The zero-inflated Poisson regression model may be appropriate when there are more zeroes in the data than is consistent with a Poisson distribution.

Outline

Assignment 3

Exercise 5 on Page 100 of our text book.