MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 5: Count responses

Outline

Poisson distribution

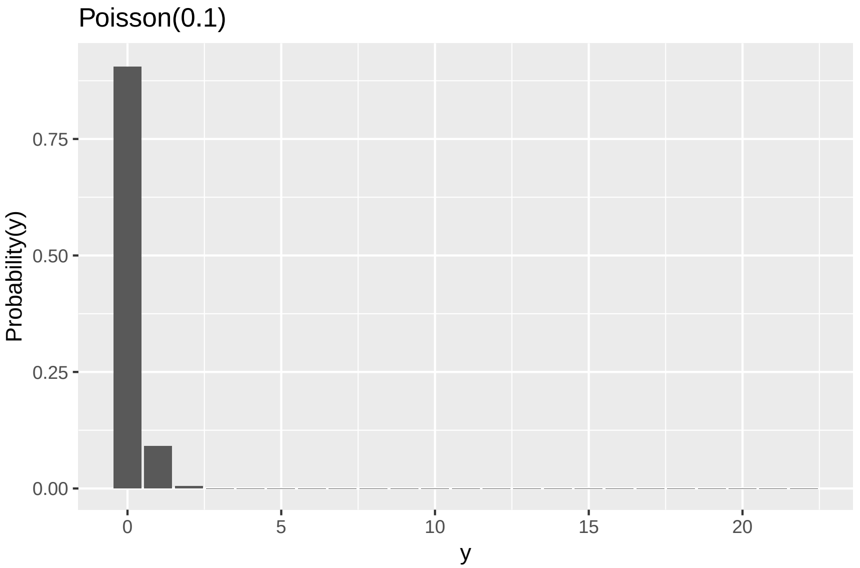
Let Y = number of events in given time interval. If events independent, and prob of event proportional to length of interval, then Y is Poisson distributed.

Poisson(μ) distribution

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

- $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$
- 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.

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.
- Likelihood: $L = \prod_{i=1}^{n} \frac{\exp^{-\mu_i} \mu_i^{y_i}}{y_i!}$

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.
- Likelihood: $L = \prod_{i=1}^{n} \frac{\exp^{-\mu_i} \mu_i^{y_i}}{y_i!}$

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

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.
- Likelihood: $L = \prod_{i=1}^{n} \frac{\exp^{-\mu_i} \mu_i^{y_i}}{y_i!}$

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

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

Deviance

$$
\begin{aligned}
D = -2\log L &= -2\sum_{i=1}^{n}\left[-\hat{\mu}_i + y_i \log(\hat{\mu}_i) - \log(y_i!)\right] \\
&= 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right)
\end{aligned}
$$

Deviance

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

- 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".

Dispersed Poisson model

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

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

Outline

Quasi-Poisson models

Assumes:

$$Y \in \{0, 1, 2, \dots\} \qquad E(Y) = \mu, \qquad Var(Y) = \phi\mu$$

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

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

Outline

Negative binomial distribution

In series of independent trials, each with prob p of success, let Z be number of trials until kth success.

$$P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}, \qquad z = k, k+1, \ldots$$

- $k = 1$ gives the *geometric* distribution.
- The NegBin distribution also arises when $Y \sim \text{Poisson}(\lambda)$ and $\lambda\theta \sim \gamma$ for some constant θ.
- For negative binomial regression, we model $Y = Z - k$.
- $\text{E}(Y) = \mu = k(1-p)/p$ and $\text{Var}(Y) = \mu + \mu^2/k$.
- $p = k/(k + \mu)$

Negative binomial regression

$Y_i \sim \text{NegBin} - k$ with mean μ_i and variance $\mu_i + \mu_i^2/k$.

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

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

$$L = \prod_{i=1}^{n} \binom{y_i + k - 1}{k - 1} p_i^k (1 - p_i)^{y_i}$$

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

Negative binomial regression

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

# Estimated k
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).

$$P(Y = 0) = p$$

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

- p is probability of zero; f 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).

$$P(Y = 0) = p + (1 - p)f(0)$$

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

- p 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.