Introducing Generalized Linear Models

Mark Andrews

Logistic regression as a generalization of linear regression

- If y is continuous, it may be appropriate to model y as a noisy linear function of x, i.e.

$$y = \alpha + \beta x + \epsilon.$$

- If $y \in \{0, 1\}$, then we can model the *probability* that y takes the value 1 as an *inverse logit* function of x

$$P(y = 1) = \phi(\alpha + \beta x) = \frac{1}{1 + e^{-\alpha - \beta x}}.$$

Logistic regression as generalized linear regression

- In linear-Gaussian regression, we assume the following:

$$y_i \sim N(\mu_i, \sigma^2), \quad \text{where } \mu_i = \alpha + \beta x_i.$$

- By contrast, in binary logistic regression we assume the following:

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{where } p_i = \phi(\alpha + \beta x) = \frac{1}{1 + e^{-\alpha - \beta x_i}}.$$

Alternative description of logistic regression

- Rather than viewing $P(y = 1)$ as a transformed linear function of x, we can view
$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right),$$
as a linear function of x.
- In other words,
$$y_i \sim \text{Bernoulli}(p_i), \quad \text{where } \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i.$$

Logit and inverse-logit function

- The two regression models, i.e.

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{where } p_i = \frac{1}{1 + e^{-\alpha - \beta x_i}},$$

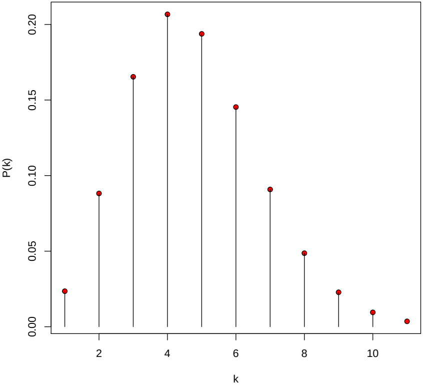
and

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{where } \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i.$$

are identical.

The Poisson Distribution

- The Poisson distribution is a discrete probability distribution over the non-negative integers $0, 1, 2 \ldots$.



The Poisson Distribution

- The Poisson distribution is used to model the probability of a given number of events occurring in a fixed interval of time, e.g. the number of trams that pass Chaucer during this class.
- It has a single parameter λ, known as the *rate*.
- If x is a Poisson random variable whose, its probability mass function is
$$P(x = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Poisson Regression

- In any regression problem, our data are $(y_1, x_1), (y_2, x_2) \ldots (y_n, x_n)$, where each y_i is modelled as a stochastic function of x_i.
- In Poisson regression, we assume that each y_i is a Poisson random variable rate λ_i and

$$\log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki},$$

or equivalently

$$\lambda_i = e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}}.$$

Poisson Regression

- As an example of Poisson regression, we can look at the number visits to a doctor in a fixed period as a function of predictors such as gender.
- Using a data-set of over 5000 people, we estimate (using mle) that

$$\log(\lambda_i) = 1.65 + 0.43 \times x_i$$

where $x_i = 1$ for a female, and $x_i = 0$ for a male.

Poisson Regression

- Using this example, we see that for a female

$$\lambda_{\text{Female}} = e^{1.65+0.43} = 8.004$$

and for males

$$\lambda_{\text{Male}} = e^{1.65} = 5.2$$

- In other words, the expected value for females is 8.2 and for males it is 5.2.