

Generalised linear models in R

Richard B. Sherley

University of Exeter, Penryn Campus, UK

March 2020



Recap: Linear regression

Assumptions:

- 1 A **linear** mean function is relevant.
- 2 Variances are equal across all predicted values of the response (**homoscedastic**).
- 3 Errors are **normally** distributed.
- 4 Samples collected at **random**.
- 5 Errors are **independent**.

Generalised linear models (GLMs)

- 1 A linear mean (including any explanatory variables you want to)
i.e $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$
- 2 A **link function** (like an “internal” transformation).
- 3 An **error structure**.
So far we assumed normality $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Link functions

Links your **mean** function to the *scale* of the **observed data** e.g.

$$E(Y) = g^{-1}(\beta_0 + \beta_1 X)$$

- $\mathbb{E}(Y)$ is the **expected value** (i.e. mean of Y).
- The function $g(\cdot)$ is known as the **link function**, and $g^{-1}(\cdot)$ denotes the **inverse** of $g(\cdot)$.

Simple linear regression is a special case of a GLM

① A linear mean: $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

② An **error structure**: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

③ **Link function: identity**

$$\mu = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 X$$

GLMs in R

```
lm(height ~ weight, data=df)
```

Is equivalent to:

```
glm(height ~ weight, data=df, family=gaussian(link=identity))
```

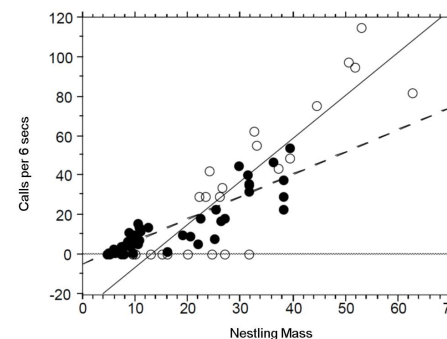
family specifies the error structure **and** link function

Default link functions

| Family | Link |
|------------------|--------------------------|
| gaussian | identity |
| binomial | logit, probit or cloglog |
| poisson | log, identity or sqrt |
| Gamma | inverse, identity or log |
| inverse.gaussian | 1/mu ² |
| quasi | user-defined |
| quasibinomial | logit |
| quasipoisson | log |

Poisson regression

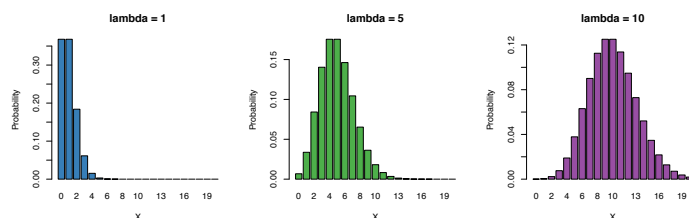
Count data is **discrete** and **non-negative**



$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad Y \sim \text{Pois}(\mu)$$
$$\mu = \beta_0 + \beta_1 X \quad \log \mu = \beta_0 + \beta_1 X$$

Poisson distribution

- **Discrete** variable, defined on the range $0, 1, \dots, \infty$.
- A single **rate** parameter λ , where $\lambda > 0$.
- **Mean** = λ
- **Variance** = λ



Poisson regression

$$Y \sim \text{Pois}(\lambda)$$
$$\log \lambda = \beta_0 + \beta_1 X$$

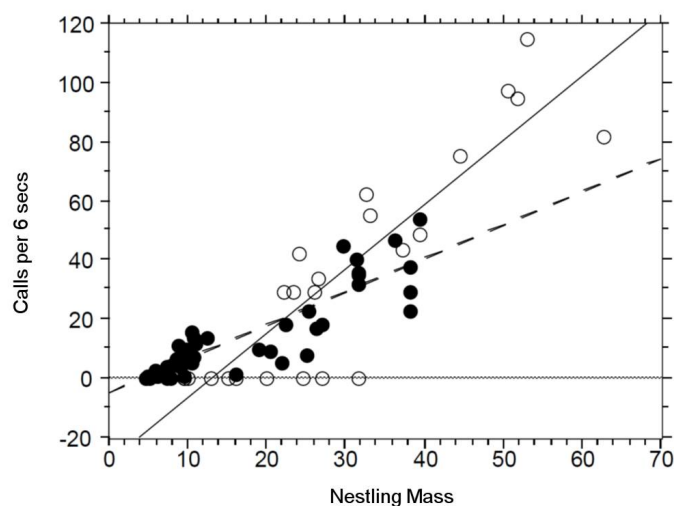
Using the rules of logarithm (i.e $\log \lambda = k$, then $\lambda = e^k$):

$$\log \lambda = \beta_0 + \beta_1 X$$
$$\lambda = e^{\beta_0 + \beta_1 X}$$

Thus we are effectively modelling the observed counts using an exponential distribution

```
glm(outcome ~ explanatory, data=df, family=poisson(link=log))
```

Cuckoo data

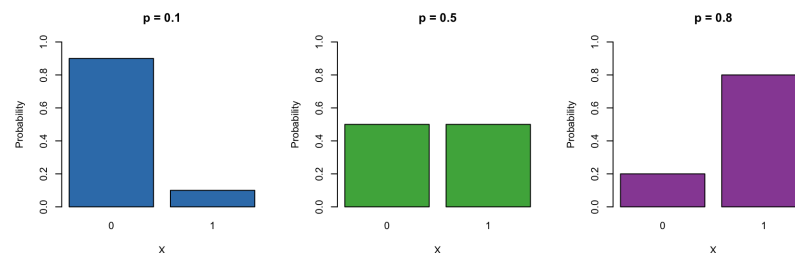


Logistic regression

Consider a **categorical** response variable with two levels (e.g pass/fail). These type of **binary** data are assumed to be **Bernoulli** distributed:

$$Y \sim \text{Bern}(p)$$

- A **probability** parameter p , where $0 < p < 1$.
- **Mean** = p
- **Variance** = $p(1 - p)$



Logistic regression

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad Y \sim \text{Pois}(\lambda) \quad Y \sim \text{Bern}(p)$$

$$\mu = \beta_0 + \beta_1 X \quad \log \lambda = \beta_0 + \beta_1 X \quad ?? = \beta_0 + \beta_1 X$$

$$Y \sim \text{Bern}(p)$$

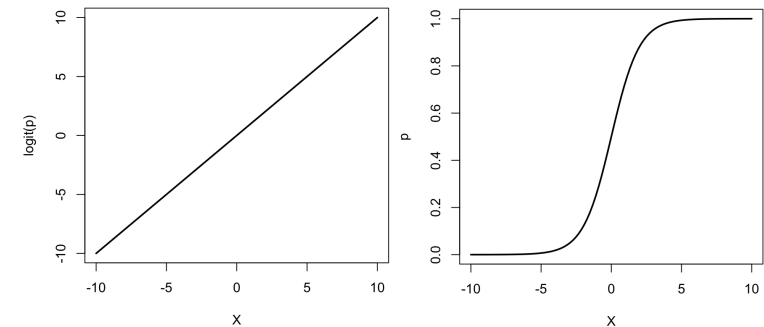
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

Logistic regression

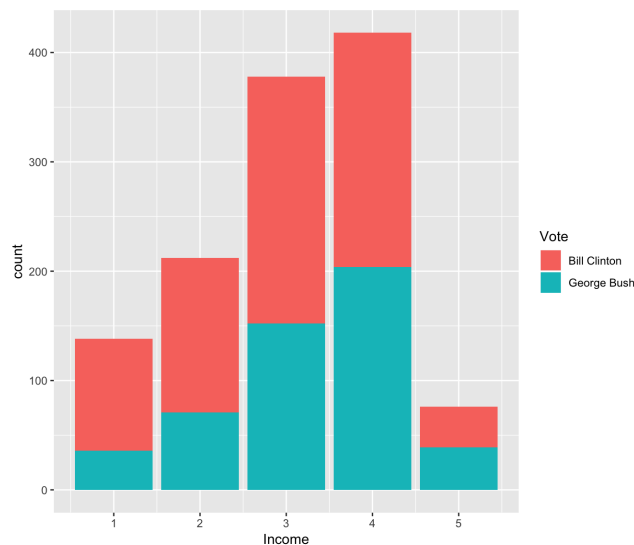
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$



```
glm(response ~ explanatory, data=df, family=binomial(link=logit))
```

1992 US election survey



1992 US election survey

```
fit <- glm(Vote ~ Income, data=USA, family=binomial(link=logit))
summary(fit)
```

```
##
## Call:
## glm(formula = Vote ~ Income, family = binomial(link = logit),
##      data = USA)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2699  -1.0162  -0.8998   1.2152   1.6199
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3017    0.1828  -7.122 1.06e-12 ***
## Income         0.3033    0.0551   5.505 3.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1655.0  on 1221  degrees of freedom
## Residual deviance: 1623.5  on 1220  degrees of freedom
## AIC: 1627.5
##
## Number of Fisher Scoring iterations: 4
```

$$Y \sim \text{Bern}(p)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

- ‘(Intercept)’ = $\beta_0 = -1.3$
- ‘Income’ = $\beta_1 = 0.303$

It is common to interpret variables according to some **central tendency**
e.g at the central income category (i.e $X = 3$)

$$\begin{aligned} P(\text{Republican vote at } X = 3) &= \text{logit}^{-1}(-1.3 + 0.3 \times 3) \\ &= \frac{e^{-1.3+0.3 \times 3}}{1 + e^{-1.3+0.3 \times 3}} \\ &= 0.48. \end{aligned}$$