LOGISTIC REGRESSION WITH GROUPED DATA

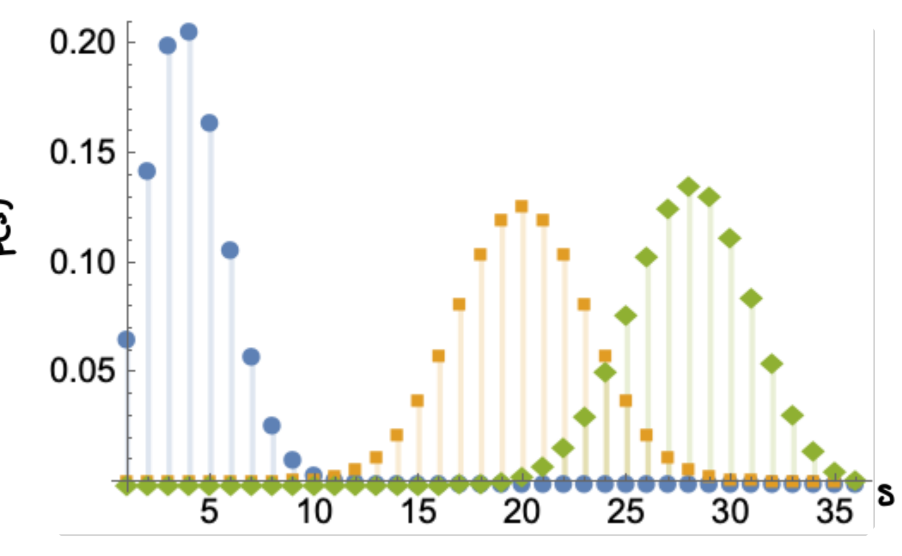
Let's consider again the beetle data.
The experiment has been run on several beetles for every dose: I can count how many beetles are dead or alive for each level. I obtain the grouped data

# killed (1)	6	13	…	60
# alive (0)	53	47	…	0
log(dose)	1.69	1.724	…	1.88

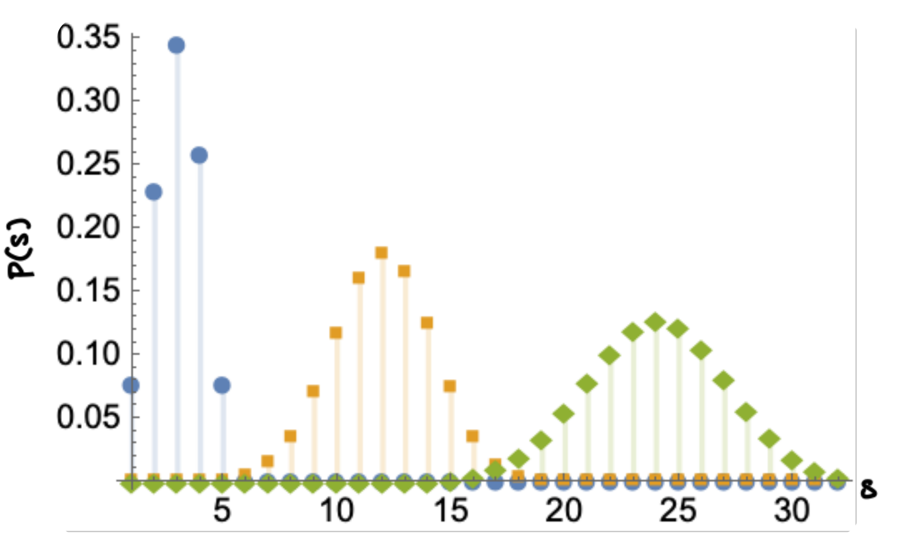
For the grouped data, an adequate distribution is the BINOMIAL distribution.
Recall that

$S \sim Bi(m, \pi)$

- parameter space: $m \in \{0, 1, 2, \dots\}$ number of trials
  $\pi \in [0, 1]$ success probability
- support: $\mathcal{S} = \{0, 1, \dots, m\}$ number of successes on m trials
- probability mass function: $p_S(s; m, \pi) = \mathbb{P}(S = s) = \binom{m}{s} \pi^s (1-\pi)^{m-s}$ with $\binom{m}{s} = \frac{m!}{s!(m-s)!}$
- moments: $\mathbb{E}[S] = m\pi$
  $var(S) = m\pi(1-\pi)$
- relationship with the Bernoulli distribution: consider a sequence of m independent Bernoulli random variables $T_1, \dots, T_m$ with common success probability π: $T_k \sim Bern(\pi)$ $k = 1, \dots, m$ independent.
  Then $S = \sum_{k=1}^{m} T_k \sim Bi(m, \pi)$.



$S \sim Bi(40, \pi)$ $\pi = \{0.1, 0.5, 0.7\}$



$S \sim Bi(m, 0.6)$ $m = \{5, 20, 40\}$

How do we define a model for grouped data, e.g., in the beetle example?
Assume that in the ungrouped data we observed $T_k \sim Bern(\pi(x_k))$ $k = 1, \dots, N$, with N the total number of beetles that were used in the experiment, and $\pi(x_k)$ the probability of "success" using a dose equal to x_k.
However, the experiment was repeated several times for each poison level.
Let's denote with n the number of different levels of poison used in the experiment.
For each dose level x_i ($i = 1, \dots, n$), m_i beetles were observed: we can group together the outcome of the experiment for each experimental condition.
Indeed, beetles with a dose $= x_i$ all have the same probability $= \pi(x_i)$

The log(dose) is x_i $i = 1, \dots, n$.
For each level x_i, we count the number of dead and alive beetles.
$S_i = \sum_{k=1}^{N} \mathbb{1}(T_k = 1 \mid x_k = x_i)$ number of successes at a dose $= x_i$

m_i is the total number of beetles observed at a dose x_i
$m_i = \sum_{k=1}^{N} \mathbb{1}(x_k = x_i)$

Since the T_k $k = 1, \dots, N$ were independent, with distribution $Bern(\pi(x_k))$ → the success probability is common for r.v. with the same x_k. Hence the distribution of S_i is
$S_i \sim Bi(m_i, \pi_i = \pi(x_i))$ $i = 1, \dots, n$ independent
with support $\{0, 1, \dots, m_i\}$.

Indeed the grouped data can be expressed as

S	m	x
s_1	m_1	x_1
⋮	⋮	⋮
s_i	m_i	x_i
⋮	⋮	⋮
s_n	m_n	x_n
# successes	# trials	

or

S	m−S	x
s_1	$m_1 - s_1$	x_1
⋮	⋮	⋮
s_i	$m_i - s_i$	x_i
⋮	⋮	⋮
s_n	$m_n - s_n$	x_n
# successes	# failures	

LOGISTIC REGRESSION (general case: grouped data)

With GLMs we model the MEAN of the random variables.
In this case we have $S_1, \dots, S_n$ independent, $S_i \sim Bi(m_i, \pi_i)$
$\mathbb{E}[S_i] = m_i \pi_i$
If we model directly the S_i, we study $(m_i \pi_i)$. But in a study the quantity of interest is actually $\pi(x_i)$: the success probability at a level x_i (not $m_i \cdot \pi_i$, also notice that m_i changes with i).
How do we define a model for π_i?

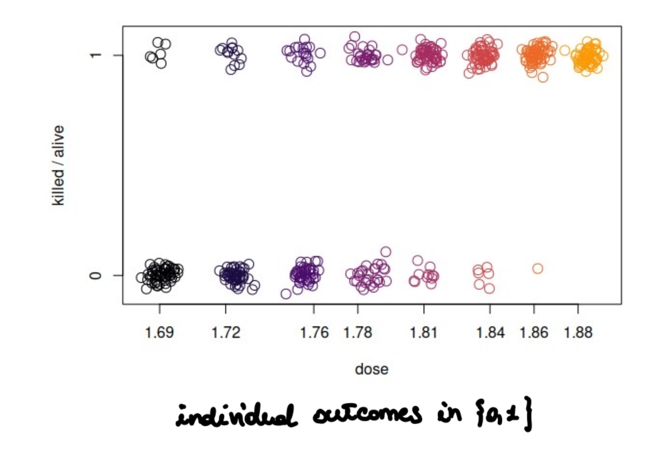
Consider a transformation of the random variables
$Y_i = \frac{S_i}{m_i}$ $i = 1, \dots, n$
The expected value is $\mathbb{E}[Y_i] = \mathbb{E}\left[\frac{S_i}{m_i}\right] = \frac{1}{m_i}\mathbb{E}[S_i] = \frac{m_i \pi_i}{m_i} = \pi_i$
The mean of Y_i is our parameter of interest π_i.
Support $\mathcal{Y} = \left\{0, \frac{1}{m_i}, \frac{2}{m_i}, \dots, \frac{m_i - 1}{m_i}, 1\right\}$
what is the distribution of these new r.v.?

$$\mathbb{P}(Y_i = y_i) = \mathbb{P}\left(\frac{S_i}{m_i} = y_i\right) = \mathbb{P}(S_i = y_i m_i) \underset{S_i \sim Bi(m_i, \pi_i)}{=} \binom{m_i}{m_i y_i} \pi_i^{y_i m_i} (1-\pi_i)^{m_i - y_i m_i} = p_S(m_i y_i; m_i, \pi_i)$$

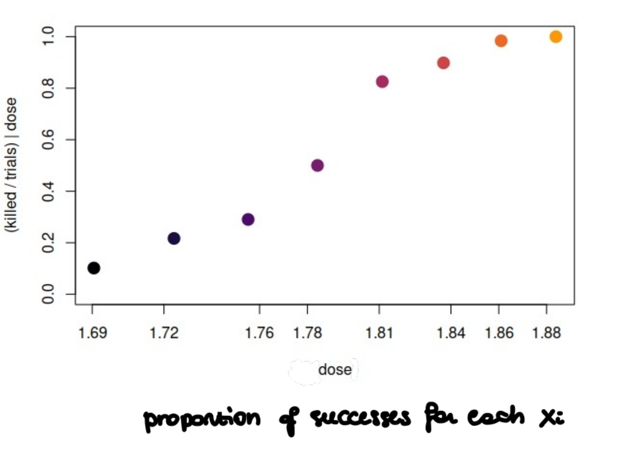
i.e. $m_i Y_i \sim Bi(m_i, \pi_i)$ $i = 1, \dots, n$ independent.
It is possible to show that the distribution of Y_i is in the exponential family.
$Var(Y_i) = Var\left(\frac{S_i}{m_i}\right) = \frac{1}{m_i^2} var(S_i) = \frac{1}{m_i^{\cancel{2}}} \cancel{m_i} \pi_i (1-\pi_i) = \frac{\pi_i(1-\pi_i)}{m_i}$ heteroscedastic

We can fit a glm on these new random variables.
The model is basically the same we have seen for $\{0, 1\}$ data.
The canonical link function is again $g(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \eta_i = \tilde{\underline{x}}_i^T \underline{\beta}$
The interpretation of the parameters is the same.

Data visualisation with ungrouped and (transformed) grouped data



individual outcomes in $\{0, 1\}$



proportion of successes for each x_i

INFERENCE:

$p_{Y_i}(y_i; m_i, \pi_i) = p_{S_i}(m_i y_i; m_i, \pi_i) = \binom{m_i}{m_i y_i} \pi_i^{y_i m_i} (1-\pi_i)^{m_i - y_i m_i}$ with $\pi_i = \frac{e^{\tilde{x}_i^T \beta}}{1 + e^{\tilde{x}_i^T \beta}}$

$$L(\underline{\beta}) = \prod_{i=1}^{n} p_{Y_i}(y_i; m_i, \pi_i) = \prod_{i=1}^{n} \binom{m_i}{m_i y_i} \pi_i^{y_i m_i} (1-\pi_i)^{m_i - y_i m_i}$$
$$\propto \prod_{i=1}^{n} \pi_i^{y_i m_i} (1-\pi_i)^{m_i - y_i m_i}$$

$$\ell(\underline{\beta}) = \sum_{i=1}^{n} \{ y_i m_i \log \pi_i + m_i (1-y_i) \log(1-\pi_i) \} = \sum_{i=1}^{n} \{ m_i \underbrace{[y_i \log \pi_i + (1-y_i) \log(1-\pi_i)]}_{\text{this is what we had in the Bernoulli case}} \}$$
$$= \sum_{i=1}^{n} \{ m_i [y_i \tilde{x}_i^T \beta - \log(1 + e^{\tilde{x}_i^T \beta})] \}$$

$$\ell_*(\underline{\beta}) = \sum_{i=1}^{n} \left\{ m_i \tilde{x}_i \left(y_i - \frac{e^{\tilde{x}_i^T \beta}}{1 + e^{\tilde{x}_i^T \beta}} \right) \right\} = \sum_{i=1}^{n} \tilde{x}_i \left(m_i y_i - m_i \frac{e^{\tilde{x}_i^T \beta}}{1 + e^{\tilde{x}_i^T \beta}} \right) = \sum_{i=1}^{n} \tilde{x}_i (m_i y_i - m_i \pi_i)$$

the likelihood equations are $\sum_{i=1}^{n} \tilde{x}_i m_i y_i = \sum_{i=1}^{n} \tilde{x}_i m_i \pi_i$

$\sum_{i=1}^{n} \tilde{x}_i m_i y_i = \sum_{i=1}^{n} \tilde{x}_i m_i \frac{e^{\tilde{x}_i^T \beta}}{1 + e^{\tilde{x}_i^T \beta}}$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_r \partial \beta_s} = - \sum_{i=1}^{n} m_i x_{ir} x_{is} \pi_i (1-\pi_i)$$

$j(\underline{\beta}) = -\ell_{**}(\underline{\beta}) = X^T U X$ $U = diag\{ m_1 \pi_1 (1-\pi_1), \dots, m_n \pi_n (1-\pi_n) \} = U(\underline{\beta})$
$j(\hat{\underline{\beta}}) = X^T U(\hat{\underline{\beta}}) X$

- INFERENCE about the ESTIMATOR of the REGRESSION COEFFICIENT
  $\hat{\underline{\beta}} \dot\sim N_p(\underline{\beta}, j(\hat{\underline{\beta}})^{-1})$

- TEST about NESTED MODELS (about subsets of $\underline{\beta}$)

$$\underline{\beta} = \begin{bmatrix} \underline{\beta}^{(0)} \\ \underline{\beta}^{(1)} \end{bmatrix} \begin{matrix} \in \mathbb{R}^{p_0} \\ \in \mathbb{R}^{p-p_0} \end{matrix}$$

$$\begin{cases} H_0: \underline{\beta}^{(1)} = \underline{0} \\ H_1: \underline{\beta}^{(1)} \neq \underline{0} \end{cases}$$

likelihood ratio test:
$W = 2\{ \hat{\ell}(\text{model}) - \tilde{\ell}(\text{restricted}) \} \dot\sim \chi^2_{p-p_0}$ under H_0

$$w^{obs} = 2 \left\{ \sum_{i=1}^{n} \{ m_i [y_i \log \hat{\pi}_i + (1-y_i) \log(1-\hat{\pi}_i)] - m_i [y_i \log \tilde{\pi}_i + (1-y_i) \log(1-\tilde{\pi}_i)] \} \right\}$$
$$= 2 \left\{ \sum_{i=1}^{n} \left\{ m_i \left[y_i \log \frac{\hat{\pi}_i}{\tilde{\pi}_i} + (1-y_i) \log \frac{(1-\hat{\pi}_i)}{(1-\tilde{\pi}_i)} \right] \right\} \right\}$$

- TEST about the OVERALL SIGNIFICANCE

$$\begin{cases} H_0: \beta_2 = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \end{cases}$$

as usual, $W \dot\sim \chi^2_{p-1}$ under H_0

under H_0: we have a common $\pi = e^{\beta_1}$ for all $i = 1, \dots, n$
What is the estimate in this case?
$\ell(\underline{\beta}) = \sum_{i=1}^{n} \{ y_i m_i \log \pi + m_i (1-y_i) \log(1-\pi) \}$

$\ell_*(\beta) = \sum_{i=1}^{n} \frac{y_i m_i}{\pi} - \frac{m_i (1-y_i)}{1-\pi}$

$\ell_*(\beta) = 0 \Rightarrow \sum y_i m_i - \pi \cancel{\sum y_i m_i} - \pi \sum m_i + \pi \cancel{\sum y_i m_i} = 0$

$$\tilde{\pi} = \frac{\sum_{i=1}^{n} y_i m_i}{\sum_{i=1}^{n} m_i} = \frac{\#\text{ successes}}{\#\text{ trials}} = \frac{\sum_{i=1}^{N} t_i}{N} = \bar{t}$$

- DEVIANCE
  I set a separate parameter π_i $\forall i$: saturated model
  $\ell(\pi_i) = m_i y_i \log \pi_i + m_i (1-y_i) \log(1-\pi_i)$
  $\ell_*(\pi_i) = \frac{m_i y_i}{\pi_i} - \frac{m_i (1-y_i)}{1-\pi_i}$

  $\ell_*(\pi_i) = 0 \Rightarrow m_i y_i - \cancel{\pi_i m_i y_i} - \pi_i m_i + m_i \cancel{\pi_i} y_i = 0$
  $\tilde{\pi}_i = y_i$
  $\ell(\tilde{\pi}_i) = m_i y_i \log y_i + m_i (1-y_i) \log(1-y_i)$ ← now it is not always zero!
  $y_i \in \{0, \frac{1}{m_i}, \frac{2}{m_i}, \dots, 1\}$

So we have now a proper deviance
$$D = 2\{ \tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model}) \}$$
$$= 2 \left\{ \sum_{i=1}^{n} [m_i y_i \log y_i + m_i (1-y_i) \log(1-y_i) - m_i y_i \log \hat{\pi}_i - m_i (1-y_i) \log(1-\hat{\pi}_i)] \right\}$$
$$= 2 \left\{ \sum_{i=1}^{n} m_i \left[y_i \log \frac{y_i}{\hat{\pi}_i} + (1-y_i) \log \frac{(1-y_i)}{(1-\hat{\pi}_i)} \right] \right\}$$

With GROUPED data, <u>if all m_i are large</u>, we have an approximate distribution for D.
In this case, the deviance can be used as a measure of the goodness of fit of the model: small values indicate a good fit.
$D < n - p$ is generally ok (where n is the number of different configurations of the covariates, e.g. the number of levels of $x_i = \log(dose)$)

- RESIDUALS
  Similarly to the Poisson regression case, we can carry out the model checking through graphical analysis of the residuals.
  Pearson's residuals: $e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{V(\hat{\pi}_i)}}$ $i = 1, \dots, n$ with $V(\hat{\pi}_i) = \frac{1}{m_i} \hat{\pi}_i (1-\hat{\pi}_i)$