Binomial Models

Let $i = 1, \ldots, n$ reference a particular row in our standings dataset (e.g. LA Rams in 2023)

w_i = # of wins for that team (in that seasons)

A simple model for w_i

$W_i \overset{ind}{\sim} \text{Binomial}(n_i, p_i)$

$n_i = 16$ or 17 depending on the season

p_i = probability team i wins each of its n_i games.

Problems with this model:

* don't know p_i (need to estimate it)
* Assumes $P(\text{team } i \text{ beats team } j) = p_i$ not depending on team j's probability
* Quality of team i same all season.

But it's a starting point

How do we get p_i? What should it depend on?

~~and~~ idea: p_i depends on prop. of wins last year $\underline{q_i}$

Try this: $p_i = b_0 + b_1 q_i$

problem is that p_i should be between 0 and 1

This model does not guarantee that.

Thinking ahead: to estimate any parameters, least squares is probably not appropriate because responses are not normal.

Generalized Linear Model:

1. Response distribution

   $Y_i \sim \text{Distribution}(\theta_i)$

   ~~response~~ specify a good model for the response.

2. Link function

   $\theta_i = f(a_i)$

   $a_i = f^{-1}(\theta_i)$

   ensure parameter θ_i falls within valid range. Link function should be invertible.

3. Linear form

   $a_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots$

   Relate parameter to other variables.

for the binomial model:

1. $W_i \overset{ind}{\sim} \text{Binomial}(n_i, p_i)$

2. $p_i = \dfrac{\exp(a_i)}{1+\exp(a_i)} := \text{expit}(a_i)$

$a_i = \log\left(\dfrac{p_i}{1-p_i}\right) := \text{logit}(p_i)$

3. $a_i = b_0 + b_1 q_i$ or

$a_i = b_0 + b_1 \,\text{logit}(q_i)$

No epsilons! tranforms are allowed.

Looking ahead. For hurricanes:

1. $Y_i \overset{ind}{\sim} \text{Poisson}(\mu_i)$ (hurricane count in year i)

2. $\mu_i = \exp(a_i)$

$a_i = \log(\mu_i)$

3. $a_i = b_0 + b_1 x_i$ x_i = sea surface temp.

No epsilons!

Go over the code for NFL data:

Actual wins is more variable than that predicted by the binomial GLM

Could be due to various factors:

Need to inject more variability into the model

Generalized linear mixed model:

1. $W_i \overset{ind}{\sim} \text{Binomial}(n_i, p_i)$

2. $p_i = \exp(a_i) / (1 + \exp(a_i))$

3. $a_i = b_0 + b_1 \text{logit}(q_i) + \varepsilon_i$

   $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

Basically, we are saying that the proportion of wins is that predicted by prior season, plus a random effect, which could kick it in either direction.

Beta Binomial Model:

1. $W_i \overset{ind}{\sim} \text{Binomial}(n_i, p_i)$

2. $p_i \sim \text{Beta}(\alpha_i, \beta_i)$

   α_i and β_i ~~m~~ (or a transformation thereof) may depend on other variables, such as q_i

   $$E(p_i) = \frac{\alpha_i}{\alpha_i + \beta_i} = \cancel{b_0 + b_1 \operatorname{logit}(q_i)} = \frac{\exp(a_i)}{1 + \exp(a_i)}$$

   $$\phi_i = \frac{1}{\alpha_i + \beta_i + 1} = c_0$$

3. $a_i = b_0 + b_1 \operatorname{logit}(q_i)$

Same ideas as gen. lin. mixed model:

p_i is drawn from a distribution centered on $\operatorname{expit}(b_0 + b_1 \operatorname{logit}(q_i))$

Just a different distribution.

$$E(p) = \frac{\alpha}{\alpha+\beta} \qquad \phi = \frac{1}{\alpha+\beta+1}$$

solve for α and β

$$\frac{1}{\phi} = \alpha + \beta + 1$$

$$\frac{\alpha}{E(p)} = \alpha + \beta \longrightarrow \frac{1}{\phi} - 1 = \alpha + \beta$$

$$\frac{\alpha}{E(p)} = \frac{1}{\phi} - 1$$

$$\alpha = E(p)\left(\frac{1}{\phi} - 1\right)$$

$$\left(\frac{1}{\phi} - 1\right) - E(p)\left(\frac{1}{\phi} - 1\right) = \beta$$

$$\left(1 - E(p)\right)\left(\frac{1}{\phi} - 1\right) = \beta$$

Beta binomial pmf:

$$P(X = k) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)}$$

Maximum Likelihood Estimation

likelihood is the joint density function, evaluated at the data, viewed as a fun. of parameters.

For binomial GLM:, data $w_1, \dots, w_n$, $q_1, \dots, q_n$

$$f(w) = \prod_{i=1}^{n} \binom{n_i}{w_i} p_i^{w_i} (1-p_i)^{n_i - w_i}$$

$$= \prod_{i=1}^{n} \binom{n_i}{w_i} \text{expit}(a_i)^{w_i} \left(1 - \text{expit}(a_i)\right)^{n_i - w_i}$$

$$= \prod_{i=1}^{n} \binom{n_i}{w_i} \text{expit}\left(b_0 + b_1 \text{logit}(q_i)\right)^{w_i} \left(1 - \text{expit}(b_0 + b_1 \text{logit}(q_i)\right)^{n_i - w_i}$$

complicated!

Generally, there won't be a formula for $\hat{b}_0$, $\hat{b}_1$ like in regression (i.e. $(X^TX)^{-1}X^Ty$)

But we can use numerical procedures to maximize the logarithm of the likelihood.

- Fisher scoring is the preferred method.