

# Regression Methods

Myrto Linnios

Autumn 2025 - Week 8

## General Models - Count Data

# Types of count data

- ▶  $y \in \{0, 1, 2, \dots\}$ , perhaps with upper bound  $m$ , depending on sampling scheme:
  - ▶ counts, with no fixed total;
  - ▶  $m$  individuals, subdivided into various categories:
    - ▶ **nominal response**—unordered categories (nationality, ...)
    - ▶ **ordinal response**—ordered categories (pain level, income, ...)
- ▶ Simplest models:
  - ▶ single unbounded response, or Poisson approximation to binomial, takes  $Y \sim \text{Pois}(\mu)$ ;
  - ▶ group of responses  $(Y_1, \dots, Y_d)$  with fixed total  $\sum Y_j = m$  has multinomial distribution, probabilities  $(\pi_1, \dots, \pi_d)$  and denominator  $m$ .

↑  
probability  
of picking an individual  
from 1st group.

# Poisson and multinomial distributions

- ▶  $Y \sim \text{Pois}(\mu)$  implies that

$$f(y; \mu) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0, 1, 2, \dots, \quad \mu > 0.$$

- ▶ Exponential family with natural parameter  $\theta = \log \mu$ , GLM with canonical logarithmic link,  $x^T \beta = \eta = \log \mu$ .
- ▶ If  $Y$  is number of events in Poisson process of rate  $\lambda$  observed for period of length  $T$ , then  $\mu = \lambda T$  and we set  $\eta = x^T \beta + \log T$ 
  - ▶ **offset**  $\log T$  is fixed part of linear predictor  $\eta$
- ▶ If  $Y_r \stackrel{\text{ind}}{\sim} \text{Pois}(\mu_r)$ ,  $r = 1, \dots, d$ , then the joint distribution of  $Y_1, \dots, Y_d$  given  $Y_1 + \dots + Y_d = m$  is **multinomial**, with denominator  $m$ , and probabilities

$$\pi_1 = \frac{\mu_1}{\sum_{r=1}^d \mu_r}, \quad \dots, \quad \pi_d = \frac{\mu_d}{\sum_{r=1}^d \mu_r}.$$

- ▶ If  $(Y_1, \dots, Y_d) \sim \text{Mult}(m; \pi_1, \dots, \pi_d)$ , then marginal and conditional distributions, e.g., of

$$(Y_1 + Y_2, Y_3 + Y_4 + Y_5, Y_6, \dots, Y_d), \quad (Y_1, Y_2, Y_4) \mid (Y_3, Y_5, \dots, Y_d),$$

are also multinomial.

# Log-linear and logistic regressions

$$(Y_1, x_1) \sim \mathcal{P}(\mu_1), \quad (Y_2, x_2) \sim \mathcal{P}(\mu_2)$$

where  $\mu_2 = e^{\alpha_2^T \beta + \gamma}$

- Special case: if  $d = 2$ , then

$$Y_2 \mid Y_1 + Y_2 = m \sim B\left(m, \pi = \frac{\mu_2}{\mu_1 + \mu_2}\right)$$

- If  $\mu_1 = \exp(\gamma + x_1^T \beta)$ ,  $\mu_2 = \exp(\gamma + x_2^T \beta)$ , then

$$\pi = \frac{\exp(\gamma + x_2^T \beta)}{\exp(\gamma + x_1^T \beta) + \exp(\gamma + x_2^T \beta)} = \frac{\mu_2}{\mu_1 + \mu_2} = \frac{\exp\{(x_2 - x_1)^T \beta\}}{1 + \exp\{(x_2 - x_1)^T \beta\}}$$

*Handwritten notes:*  $\mu_2$  (circled),  $\mu_1 + \mu_2$  (circled),  $\beta$  (circled),  $\mu_1$  (circled),  $\mu_2$  (circled). A red bracket groups the denominator terms. A red arrow points to the  $\beta$  term with the note "Probab of success".

which corresponds to a logistic regression model for  $Y_2$  with denominator  $m$  and probability  $\pi$ .

- Can estimate  $\beta$  using log linear model or logistic model—but can't estimate  $\gamma$  from logistic model.

## General Models - Poisson Regression

# Premier League data

```
> soccer
  month day year   team1   team2 score1 score2
1   Aug  19 2000  Charlton ManchesterC    4     0
2   Aug  19 2000   Chelsea   WestHam    4     2
3   Aug  19 2000  Coventry  Middlesbr    1     3
4   Aug  19 2000    Derby  Southampton    2     2
5   Aug  19 2000    Leeds   Everton    2     0
6   Aug  19 2000  Leicester  AstonVilla    0     0
7   Aug  19 2000  Liverpool   Bradford    1     0
8   Aug  19 2000  Sunderland   Arsenal    1     0
9   Aug  19 2000  Tottenham   Ipswich    3     1
10  Aug  20 2000 ManchesterU  Newcastle    2     0
11  Aug  21 2000   Arsenal   Liverpool    2     0
12  Aug  22 2000   Bradford   Chelsea    2     0
13  Aug  22 2000   Ipswich  ManchesterU    1     1
14  Aug  22 2000  Middlesbr  Tottenham    1     1
15  Aug  23 2000   Everton   Charlton    3     0
16  Aug  23 2000 ManchesterC  Sunderland    4     2
17  Aug  23 2000  Newcastle   Derby    3     2
18  Aug  23 2000  Southampton  Coventry    1     2
19  Aug  23 2000   WestHam   Leicester    0     1
20  Aug  26 2000   Arsenal   Charlton    5     3
...

```

# Premier League data

- ▶ 380 soccer matches in English Premier League in 2000–2001 season.
- ▶ Data: home score  $y_{ij}^h$  and away score  $y_{ij}^a$  when team  $i$  is at home to team  $j$ , for  $i, j, = 1, \dots, 20, i \neq j$ .
- ▶ Treat these as Poisson counts with means

$$\mu_{ij}^h = \exp(\Delta + \alpha_i - \beta_j), \quad \mu_{ij}^a = \exp(\alpha_j - \beta_i)$$

where

- ▶  $\Delta$  represents the home advantage;
  - ▶  $\alpha_i$  and  $\beta_i$  represent the offensive and defensive strengths of team  $i$ .
- ▶ Two possibilities for fitting:
    - ▶ Poisson GLM, with 39 parameters;
    - ▶ binomial GLM, with 20 parameters.

# Premier League data: Analysis of deviance

Poisson model			Binomial model		
Terms	df	Deviance reduction	Terms	df	Deviance reduction
Home	1	33.58	Home	1	33.58
Defence	19	39.21	Team	19	79.63
Offence	19	58.85			
Residual	720	801.08	Residual	332	410.65

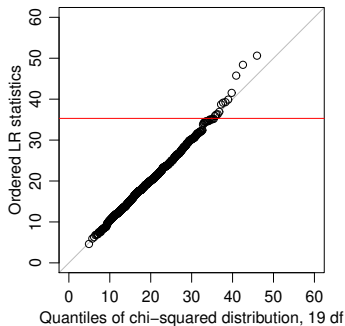
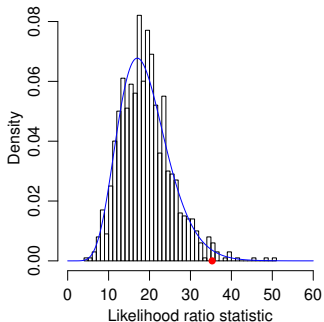
$\Delta(\text{Home})$   
-  $\Delta(\text{Home} + \text{Defence})$

- ▶ There's a strong effect of playing at home, and lots of evidence of differences among the teams—more in offence than defence.
- ▶ Both residual deviances are a little large, but since the counts are small, we don't expect the large-sample  $\chi^2$  distribution to apply well to the residual deviance (36 of the individual scores exceed three goals).
- ▶ Simulations from the fitted model suggest that the residual deviances are not unusually large, so there's no evidence of a lack of fit.

# Premier League data: Null deviance for defence effect

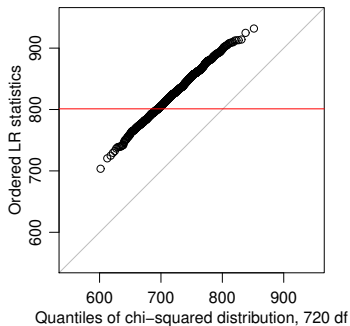
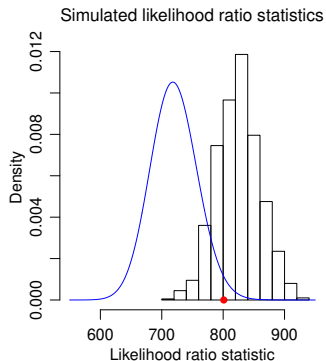
Defence effect deviance (in red) for the Poisson model is large(ish) relative to  $\chi^2_{19}$  distribution, but the asymptotics seem OK, based on simulations from a model without this effect (i.e., **Home + Offence**). It seems we can trust asymptotic distributions for differences of deviances, even though the counts are small.

Simulated likelihood ratio statistics



# Premier League data: Residual deviance

Residual deviance of 801 (in red) for the Poisson model seems large(ish) relative to  $\chi^2_{720}$  distribution, but the asymptotics are suspect because most of the counts are small. Comparison of observed deviance with  $\chi^2_{720}$  distribution shows that 801 is in fact somewhat smaller than average for datasets simulated from the fitted model.



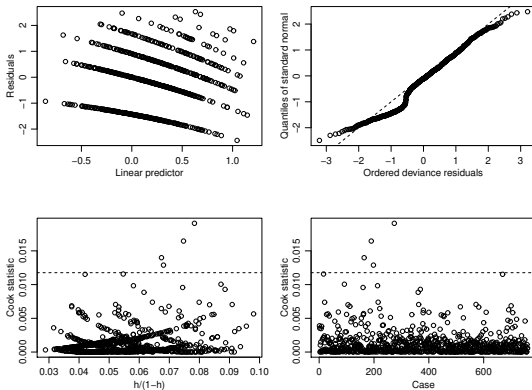
## Premier League data: Estimates

	Overall ( $\delta$ )	Offensive ( $\alpha$ )	Defensive ( $\beta$ )
Manchester United	0.39	0.22	0.15
Liverpool	0.13	0.12	-0.08
Arsenal	—	0.04	—
Chelsea	-0.09	0.08	-0.22
Leeds	-0.10	0.02	-0.17
Ipswich	-0.16	-0.10	-0.13
Sunderland	-0.33	-0.31	-0.10
Aston Villa	-0.48	-0.31	-0.15
West Ham	-0.53	-0.33	-0.30
Middlesborough	-0.53	-0.35	-0.17
Charlton	-0.55	-0.21	-0.43
Tottenham	-0.58	-0.28	-0.38
Newcastle	-0.59	-0.35	-0.30
Southampton	-0.60	-0.45	-0.25
Everton	-0.75	-0.32	-0.46
Leicester	-0.77	-0.47	-0.31
Manchester City	-0.90	-0.40	-0.56
Coventry	-0.93	-0.53	-0.52
Derby	-0.93	-0.51	-0.45
Bradford	-1.29	-0.71	-0.62
SEs	0.29	0.20	0.20

Home advantage:  $\hat{\Delta} = 0.37$  (0.07),  $\exp(\hat{\Delta}) = 1.45$  representing the incidence ratio, i.e., the increase in mean score when playing at home.

# Premier League data: Assessment of fit

Diagnostic plots for fitted model: residuals against  $\hat{\eta}$  (top left); normal QQ-plot of residuals (top right); Cook statistic  $C_j$  against leverage ratio  $h_j/(1 - h_j)$  (lower left); Cook statistic  $C_j$  against case number (lower right).



## General Models - Contingency Tables

# Sampling schemes

$A \setminus B$	0	1
0	$y_{00}$	$y_{01}$
1	$y_{10}$	$y_{11}$

- ▶ A **contingency table** contains individuals (sampling units) cross-classified by various categorical variables.
- ▶ The sampling scheme underlying a table may fix certain totals. Suppose a pollster wants to find out how people will vote. She might
  - ▶ wait in the street for a morning, and get opinions from those people willing to talk to her;
  - ▶ wait until she has the views of a fixed number, say  $m$ , of people;
  - ▶ wait until she has the views of fixed numbers of men and women.

## Example 18

Find the likelihoods for each of these sampling schemes, under (unrealistic!) assumptions of independence of voters.

## Note to Example 18

- ▶ An  $R \times C$  table arises by randomly sampling a population over a fixed period and then classifying the resulting individuals.
- ▶ In the first scheme there are no constraints on the row and column totals, and a simple model is that the count in the  $(r, c)$  cell,  $y_{rc}$ , has a Poisson distribution with mean  $\mu_{rc}$ . The resulting likelihood is

$$\prod_{r,c} \left\{ \frac{\mu_{rc}^{y_{rc}}}{y_{rc}!} e^{-\mu_{rc}} \right\};$$

this is simply the Poisson likelihood for the counts in the  $RC$  groups.

- ▶ The pollster may set out with the intention of interviewing a fixed number  $m$  of individuals, stopping only when  $\sum_{rc} y_{rc} = m$ . In this case the data are multinomially distributed, with likelihood

$$\frac{m!}{\prod_{r,c} y_{rc}!} \prod_{r,c} \pi_{rc}^{y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1,$$

with  $\pi_{rc} = \mu_{rc} / \sum_{s,t} \mu_{st}$  the probability of falling into the  $(r, c)$  cell.

- ▶ A third scheme is to interview fixed numbers of men and of women, thus fixing the row totals  $m_r = \sum_c y_{rc}$  in advance. In effect this treats the row categories as subpopulations, and the column categories as the response. This yields independent multinomial distributions for each row, and product multinomial likelihood

$$\prod_r \left\{ \frac{m_r!}{\prod_c y_{rc}!} \prod_c \pi_{rc}^{y_{rc}} \right\}, \quad \sum_c \pi_{1c} = \dots = \sum_c \pi_{Rc} = 1,$$

in which  $\pi_{rc} = \mu_{rc} / \sum_t \mu_{rt}$ .

# Contingency tables and Poisson response models

- ▶ Multinomial models can be fitted using Poisson errors, provided the appropriate baseline terms are always included in the linear predictor.
- ▶ Write the data as two-way layout, with  $C$  columns and  $R$  rows with fixed totals (e.g.,  $6 \times 8 = 48$  rows each with 3 columns for the jacamar data).
- ▶ Consider Poisson model with means  $\mu_{rc} = \exp(\gamma_r + x_{rc}^T \beta)$ :
  - ▶ the row parameters  $\gamma_1, \dots, \gamma_R$  are **nuisance parameters**, not of interest;
  - ▶ we want inference for the **parameter of interest**,  $\beta$ .
- ▶ Corresponding multinomial model has fixed row totals  $m_r$  and probabilities

$$\pi_{rc} = \frac{\mu_{rc}}{\sum_{d=1}^C \mu_{rd}} = \frac{\exp(\gamma_r + x_{rc}^T \beta)}{\sum_{d=1}^C \exp(\gamma_r + x_{rd}^T \beta)} = \frac{\exp(x_{rc}^T \beta)}{\sum_{d=1}^C \exp(x_{rd}^T \beta)}, \quad \eta_{rc} = x_{rc}^T \beta.$$

*Handwritten notes: "subpop" with arrow to  $\pi_{rc}$ , "cat." with arrow to denominator, and  $\eta_{rc} = x_{rc}^T \beta$  to the right.*

for  $r = 1, \dots, R$ ,  $c = 1, \dots, C$ ; i.e., one multinomial variable for each row.

- ▶ The resulting multinomial log likelihood is *one*

$$\begin{aligned} \ell_{\text{Mult}}(\beta; y \mid m) &\equiv \sum_{r=1}^R \sum_{c=1}^C y_{rc} \log \pi_{rc} \\ &= \sum_{r=1}^R \left\{ \sum_{c=1}^C y_{rc} x_{rc}^T \beta - m_r \log \left( \sum_{d=1}^C e^{x_{rd}^T \beta} \right) \right\}. \end{aligned}$$

*Handwritten notes: "one" with arrow to the log likelihood expression, and  $\log \frac{e^{x_{rc}^T \beta}}{\sum_{d=1}^C e^{x_{rd}^T \beta}} = x_{rc}^T \beta - \log \sum_{d=1}^C e^{x_{rd}^T \beta}$  to the right.*

$$m_r = \sum_{d=1}^C y_{rd} \uparrow$$

# Contingency tables and Poisson response models, II

$$\tau_r = \sum_{d=1}^c \mu_{rd}$$

## Lemma 19

If parameters  $\tau_r$  for the row margins are included in the above setup, then we can write

$$\ell_{\text{Poiss}}(\beta, \tau) = \ell_{\text{Poiss}}(\tau; m) + \ell_{\text{Mult}}(\beta; y | m).$$

### ► Implications:

- the MLEs of  $\beta$  and  $\tau$  based on the LHS are the same as those from separate maximisations of the terms on the right:
    - $\hat{\beta}$  equals the MLE for the multinomial model,
    - $\hat{\tau}_r = m_r$
  - the observed and expected information matrices for  $\beta, \tau$  are block diagonal.
  - SEs based on the multinomial and Poisson models are equal (exercise).
- General conclusion: inferences on  $\beta$  are the same for multinomial and Poisson models,
- provided the parameters associated to the margins fixed under the multinomial model, i.e., the  $\gamma_r$ , are included in the Poisson fit.*

# Note to Lemma 19

- The Poisson model has no conditioning, so with  $\log \mu_{rc} = \gamma_r + x_{rc}^T \beta$  the log likelihood is

$$\ell_{\text{Pois}}(\beta, \gamma) \equiv \sum_{r,c} (y_{rc} \log \mu_{rc} - \mu_{rc}) = \sum_{r=1}^R \left( m_r \gamma_r + \sum_{c=1}^C y_{rc} x_{rc}^T \beta - e^{\gamma_r} \sum_{c=1}^C e^{x_{rc}^T \beta} \right).$$

*Handwritten notes:*  $m_{rc} = e^{\gamma_r + x_{rc}^T \beta}$ ,  $m_r = \sum_{c=1}^C y_{rc}$

- Now we reparametrise in terms of the row totals  $\tau_r = \sum_c \mu_{rc}$ , noting that

$$\tau_r = e^{\gamma_r} \left( \sum_{c=1}^C e^{x_{rc}^T \beta} \right), \quad \gamma_r = \log \tau_r - \log \left\{ \sum_{c=1}^C \exp(x_{rc}^T \beta) \right\},$$

*Handwritten notes:*  $\tau_r = \sum_c m_{rc}$ ,  $\tau_r = \sum_c e^{\gamma_r + x_{rc}^T \beta} = e^{\gamma_r} \sum_c e^{x_{rc}^T \beta}$

$$\begin{aligned} \ell_{\text{Pois}}(\beta, \tau) &\equiv \sum_{r=1}^R (m_r \log \tau_r - \tau_r) + \sum_{r=1}^R \left\{ \sum_{c=1}^C y_{rc} x_{rc}^T \beta - m_r \log \left( \sum_{c=1}^C e^{x_{rc}^T \beta} \right) \right\}, \\ &= \ell_{\text{Pois}}(\tau; \mathbf{m}) + \ell_{\text{Mult}}(\beta; \mathbf{y} | \mathbf{m}), \end{aligned}$$

which is the log likelihood corresponding to

- independent Poisson row totals  $m_r$  with means  $\tau_r$ , and, independent of this,
- the multinomial log likelihood for the contingency table.