Statistics for Data Science: Week 14

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

We will now consider two fundamental specific GLM families:

1. Logistic Regression for Binary Data (Bernoulli GLM with natural link)
2. Loglinear Regression for Count Data (Poisson GLM with natural link)

These will give us concrete situations to keep in mind, demonstrating concepts already presented in generality.

We will conclude with remarks on the notion of a scale parameter.

$$\text{Very often have response} \to Y = \begin{cases} 1, & \text{“success”} \\ 0, & \text{“failure”} \end{cases}$$

So Y has a very simple Bernoulli structure:

$$\mathbb{P}[Y = 1] = \pi = 1 - \mathbb{P}[Y = 0], \quad \mathbb{E}\{Y\} = \pi$$

- Regression: need to connect response Y with an explanatory x.
- Use GLM. Can postulate that $g(\pi) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for some link g. Why?
- Intuition: Depending on circumstances, can imagine Y arising as

$$Y = \mathbf{1}\{Z > 0\} \implies \mathbb{P}[Y = 1] = \pi = 1 - F_Z(0)$$

  i.e. describing the level of a “hidden” variable Z:
- Now suppose $Z = \boldsymbol{x}^\top \boldsymbol{\alpha} + \sigma\varepsilon$. Then

$$\pi_i = 1 - F_Z(0) = 1 - F_\varepsilon\left(-\boldsymbol{x}^\top(\sigma^{-1}\boldsymbol{\alpha})\right) = 1 - F_\varepsilon(-\boldsymbol{x}^\top\boldsymbol{\beta})$$

  ($(\boldsymbol{\alpha}, \sigma)$ unidentifiable, but $\boldsymbol{\beta} = \sigma^{-1}\boldsymbol{\alpha}$ is ok)

So $g(x) = -F^{-1}(1-x)$ can serve as a link function

$$1 - \pi = F(-\boldsymbol{x}^\top \boldsymbol{\beta}) \implies -F^{-1}(1-\pi) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

Choice of Link $\iff$ Choice of Error Distribution F_ε

Distribution $F_\varepsilon(u)$		Link function $g(\pi)$	
Logistic	$e^u/(1+e^u)$	Logit	$\log\{\pi/(1-\pi)\}$
Normal	$\Phi(u)$	Probit	$\Phi^{-1}(\pi)$
Log Weibull	$1-\exp(-\exp(u))\}$	Log-log	$-\log\{-\log(\pi)\}$
Gumbel	$\exp\{-\exp(-u)\}$	Complementary log-log	$\log\{-\log(1-\pi)\}$

- Logit and probit symmetric, hard to distinguish in practice
- Log-log and complementary log-log are asymmetric
- Logit (canonical link) is usual choice, with nice interpretation

Assuming independence:

$$\mathbb{P}[Y_i = y] \overset{ind}{\sim} \pi_i^y (1-\pi_i)^{1-y}, \quad y \in \{0,1\}, \qquad \text{with } g(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^p$$

- Suppose $\{1, ..., N\} = M_1 \cup M_2 \cup ... \cup M_n, \quad M_k \cap M_q = \emptyset,\ k \neq q$

with $\boldsymbol{x}_i = \boldsymbol{c}_k$ for $i \in M_k$. Then we have a Binomial GLM:

$$\underbrace{R_j}_{\in[0,1]} | x_j \overset{ind}{\sim} \exp\left[m_j \left\{ \frac{r}{m_j} \log\left(\frac{\pi_j}{1-\pi_j} \right) + \log(1-\pi_j) \right\} + \log\binom{m_j}{r} \right]$$

$$\rightarrow g(\pi_j) = \boldsymbol{x}_j^\top \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad j = 1, ..., n$$

with $m_j = |M_j|$, $j = 1, ..., n \quad (\sum_j m_j = N)$.

M's are called covariate classes - se why important later.

Example (Challenger Catastrophe)

Challenger Space-shuttle exploded at launch on 28/1/1986, killing all seven astronauts on board. US Presidential Commission (incuding Richard Feynman) concluded that the cause was the leakage of gas due to behaviour of O-rings under low temp (temp at launch was an unusual $31^{\circ}F$).

Table: O-Ring Data

Mission #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Temp - °F	53	57	58	63	66	67	67	67	68	69	70
# Damaged	5	1	1	1	0	0	0	0	0	0	1
# Intact	1	5	5	5	6	6	6	6	6	6	5

Mission #	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Temp - °F	70	70	70	72	73	75	76	76	76	78	79	81
# Damaged	0	1	0	0	0	0	1	0	0	0	0	0
# Intact	6	5	6	6	6	6	5	6	6	6	6	6

Binary regression with natural (logit link): $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$

Interpretation?

- Unit change in x_{jk} yields additive change of logodds by β_k.
- Equivalently, unit change in x_{jk} results in multiplicative change of odds by e^{β_k}.
- In terms of parameter:

$$\pi_j = \frac{\exp\{\beta_0 + \beta_1 x_{j1} + \ldots + \beta_q x_{jq}\}}{1 + \exp\{\beta_0 + \beta_1 x_{j1} + \ldots + \beta_q x_{jq}\}}, \qquad p = q + 1$$

So

$$\frac{\partial}{\partial x_{jk}} \pi_j = \beta_k \pi_j (1 - \pi_j) \quad \text{(logistic equation!)}$$

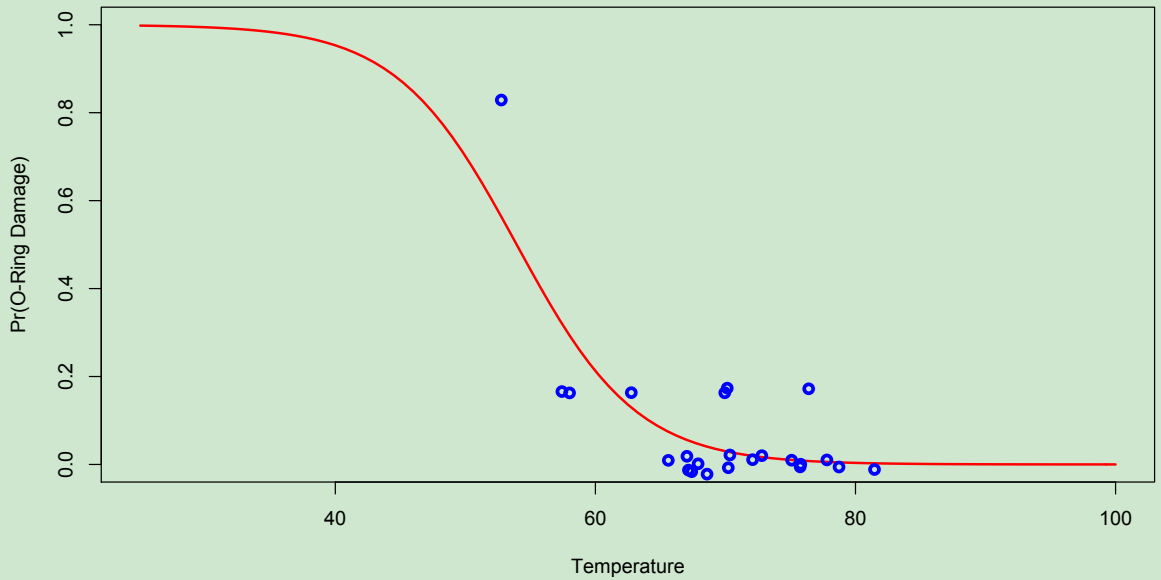
- Thus effects larger when π near $1/2$ than near endpoints of $[0, 1]$.

Example (Challenger Catastrophe, continued)

Logistic regression fit for probability of damage with temp as covariate:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 11.663 - 0.2162 \times t_i$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	−0.2162	0.0532	−4.07	0.0000



Sparseness: covariate classes $\{M_j\}_{j=1}^n$ are "small":

- i.e. n is of the order of N
  - ↪ extreme: continuous covariate, $m_k = 1\ \forall k$, so $n = N$.

Sparseness affects interpretability of deviance:

- In extreme case deviance is only a function of $\hat{\pi}$ (exercise)
  - ↪ No contrast with data! (no information about fit in absolute sense). Similar problems with Pearson statistic. Problems with residuals also.
- $D \sim \chi^2_{N-p}$ breaks down even in non-extreme case, as this requires $m_i \to \infty$ as $N \to \infty$, so small m's can hurt us.
- Deviance reduction is reasonable for comparing nested models, though.
- Interpretability and accuracy of estimators remains the same!
- Rule of thumb: sparseness when $m_k \leq 5$ for several classes.
- A solution: grouping data! (i.e. merge into covariate classes)

Suppose we have data:

x_i	-0.14	2.13	1.11	-0.53	-6.25	-3.29	-0.04	1.07	0.55
Y_i	0	1	1	0	0	0	0	1	1

Logistic regression loglikelihood can be written as:

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n}(1-Y_i)\log(1-\pi_i) \\ &= \sum_{j\in\mathcal{P}} \log\left(\frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}}\right) - \sum_{j\in\mathcal{P}^c} \log\left(1+e^{\beta_0+x_i\beta_1}\right).\end{aligned}$$

where $\mathcal{P} = \{i : x_i > 0\}$. For given β_0, what happens as $\beta_1 \to \infty$?

- Loglikelihood converges to zero! (likelihood converges to 1).
- So MLE does not exist!. Why? The problem is perfect separation.
- $\exists$ hyperplane perfectly separating covariates corresponding to 0's and 1's.
- More likely to occur when p is large relative to n.

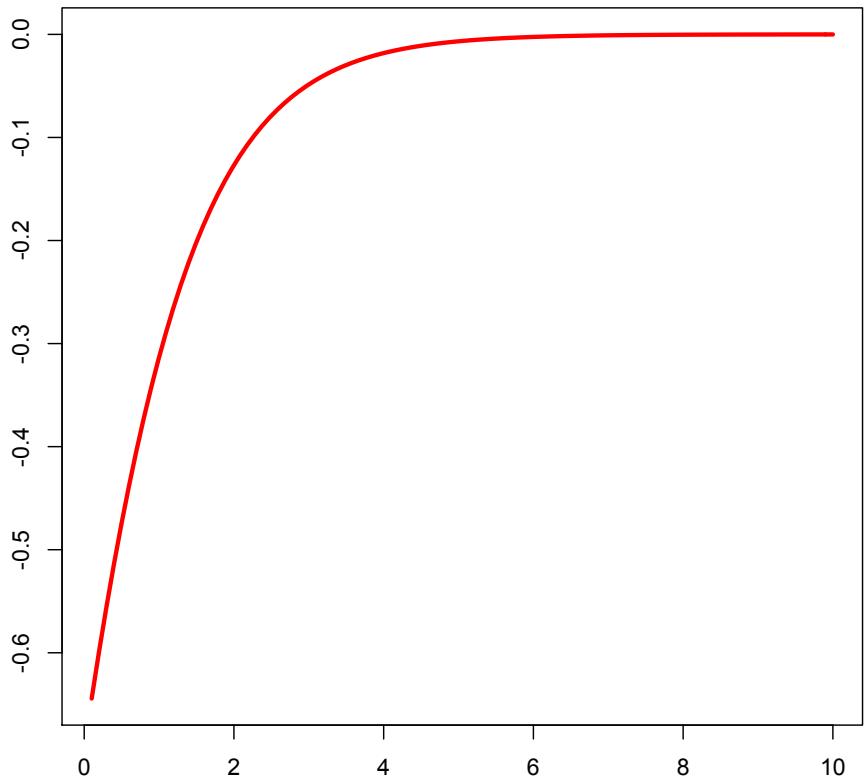


Figure: $\beta_1 \mapsto \ell(\beta_1)$

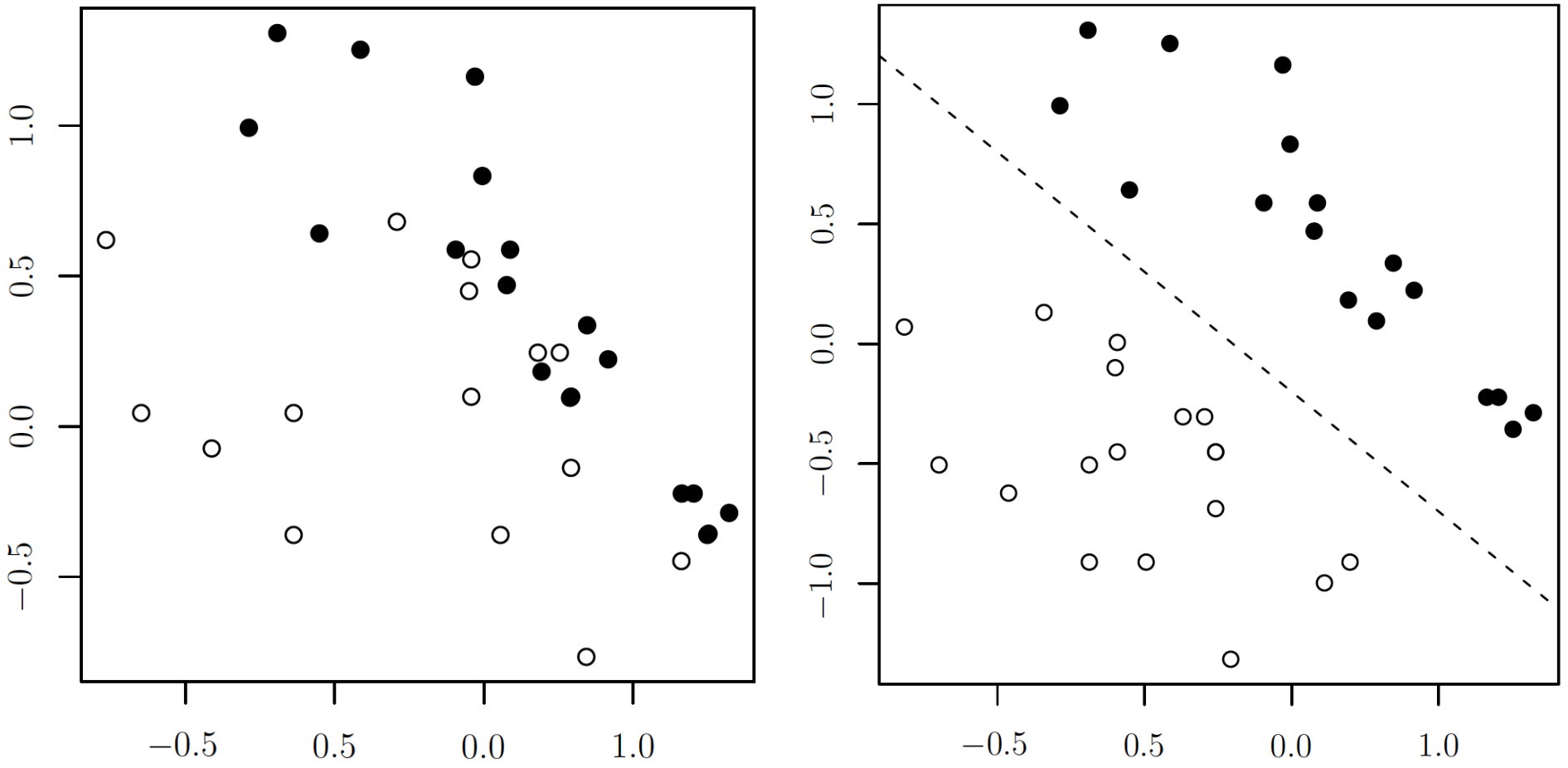


Figure: Overlap (left) vs Complete Separation (right)

We have complete separation when there exists $\gamma \in \mathbb{R}^p$ such that for all i

$$\{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma > 0\} \quad \& \quad \{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma < 0\}$$

Theorem

In the complete separation regime, the logistic regression MLE does not exist, and

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\beta) = 1.$$

Proof.

Let γ be such that $\{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma > 0\} \quad \& \quad \{Y_i = 1 \iff \boldsymbol{x}_i^\top \gamma < 0\}$.

Then we may write the loglikelihood of $t\gamma$ (for some $t > 0$) as

$$\ell(t\gamma) = \sum_{j \in \mathcal{P}} \log \left(\frac{e^{t(\boldsymbol{x}_i^\top \gamma)}}{1 + e^{t(\boldsymbol{x}_i^\top \gamma)}} \right) - \sum_{j \in \mathcal{P}^c} \log \left(1 + e^{t(\boldsymbol{x}_i^\top \gamma)} \right).$$

where $\mathcal{P} = \{i : \boldsymbol{x}_i^\top \gamma > 0\} = \{i : Y_i = 1\}$. The proof is complete upon noting:

1. For $t > 0$, $x_i^\top \gamma > 0 \iff t(x_i^\top \gamma) > 0$ and $x_i^\top \gamma < 0 \iff t(x_i^\top \gamma) < 0$.
2. As $t \to \infty$, $\ell(t\gamma) \to 0$.
3. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\ell(\boldsymbol{\beta}) < 0$ (replace $t(\boldsymbol{x}_i^\top \gamma)$ by $\boldsymbol{\beta}$ above and verify).

□

Ramifications:

- IWLS will fail to converge, with weights converging to zero.
- Standard errors will blow up.
- In a sense a design issue.
  - In Gaussian linear regression, $\boldsymbol{X}$ full rank $\implies$ MLE exists.
  - Binary regression is more subtle, and rank conditions alone do not suffice and instabilities can manifest in ways more subtle than multicollinearity.

Diagnostics and Remedies?

- Often get warning that iterations stopped after maxing out.
- But best keep track both of the likelihood value and the parameter values as the iteration evolves.
- Can remedy by imposing a penalty. Motivates regularised logistic regression:

$$\sum_{i=1}^{n} Y_i(\gamma_0 + \boldsymbol{x}_i^\top \boldsymbol{\gamma}) - \log\left(1 + e^{\gamma_0 + \boldsymbol{x}_i^\top \boldsymbol{\gamma}}\right) + \lambda \|\boldsymbol{\gamma}\|_q^2$$

for $q = 2$ (ridge) or $q = 1$ (lasso). Assuming $\boldsymbol{X}$ has been standardised, and

$$\boldsymbol{\beta}^\top = (\gamma_0, \boldsymbol{\gamma}^\top).$$

Can we at least hope that MLE exists in the overlapping regime?

Theorem (Existence and Uniqueness)

In logistic regression with an intercept term and full rank design, the maximum likelihood estimator uniquely exists if and only if the covariates overlap.

The theorem actually applies more generally to other link functions than logit.

If the model postulates that $\pi_i = g(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, and the design includes an intercept, then overlap is a necessary and sufficient provided that:

1. $-\log(g^{-1}(t))$ and $-\log(1 - g^{-1}(t))$ are convex.
2. $g^{-1}(t)$ is strictly increasing at every t.
3. $0 < g^{-1}(t) < 1$ for all t.

Special case of Bernoulli/Binomial GLM: 2 × 2 Contingency Tables

- How does a single binary covariate affect a binary response?
- Say $x \in \{0, 1\}$ (control/case), $y \in \{0, 1\}$ (failure/success)
- Simple model: individuals are independent, with m_0 and m_1 persons in categories of $x \in \{0, 1\}$ and with success probabilities

$$\pi_0 = \frac{e^{\lambda}}{1 + e^{\lambda}}, \quad \pi_1 = \frac{e^{\lambda+\psi}}{1 + e^{\lambda+\psi}}$$

Yields independent binomial variables

$$W_1 \sim \text{Binomial}(m_1, \pi_1), \quad W_0 \sim \text{Binomial}(m_0, \pi_0)$$

and likelihood

$$L(\psi, \lambda) \propto \frac{e^{(r_0+r_1)\lambda+r_1\psi}}{(1 + e^{\lambda+\psi})^{m_1}(1 + e^{\lambda})^{m_0}}.$$

- Key question: Does treatment affect success probability?
- In mathematics: is it true that $\pi_1 = \pi_0$? If not, by how much do they differ?
- Could consider absolute difference of risks, or probability ratio

$$\pi_1 - \pi_0, \quad \pi_1/\pi_0$$

- More common to consider difference of log odds

$$\psi = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right).$$

- This is natural parameter of exponential family.
- One must be quite careful with 2×2 tables – as the next example will illustrate.

Example (Women and Smoking)

- Data gathered by surveying people on electoral register in 1972–74.habits, ...
- Follow-up study 20 years later: see how many people had died.
- 162 women had smoked before 1972 but had stopped by 1972, and smoking habits were unknown for 18 women; these 180 women were excluded.

Table: Twenty-year survival and smoking status for 1314 women . The smoker and non-smoker columns contain number dead/total (% dead).

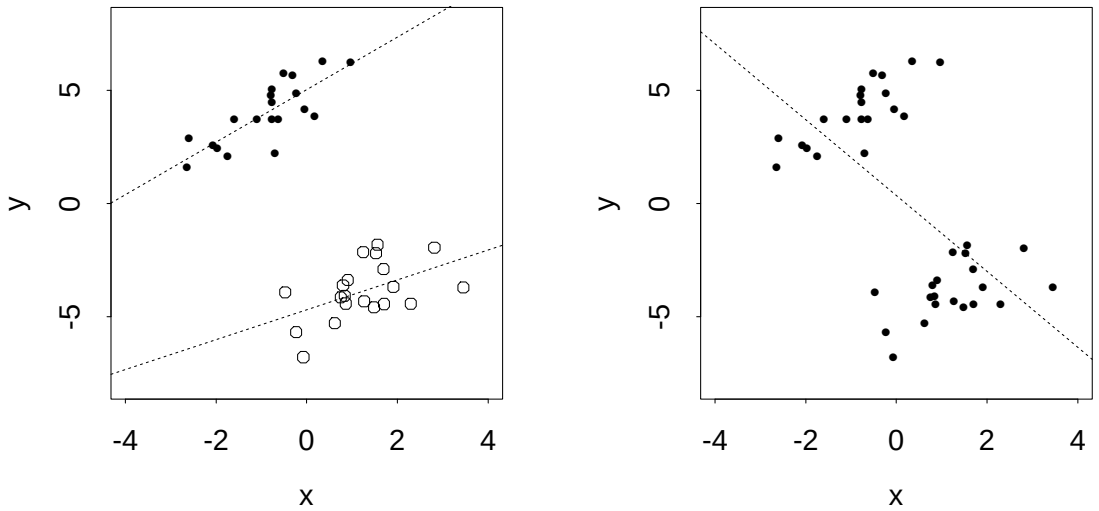
Age (years)	Smokers	Non-smokers
Overall	139/582 (24%)	230/732 (31%)
18–24	2/55 (4%)	1/62 (2%)
25–34	3/124 (2%)	5/157 (3%)
35–44	14/109 (13%)	7/121 (6%)
45–54	27/130 (21%)	12/78 (15%)
55–64	51/115 (44%)	40/121 (33%)
65–74	29/36 (81%)	101/129 (78%)
75+	13/13 (100%)	64/64 (100%)

Example (Women and Smoking, continued)

```
> data(smoking)
> summary(glm(cbind(dead,alive)~smoker,data=smoking,binomial))

Call:
glm(formula = cbind(dead, alive) ~ smoker, family = binomial,
    data = smoking)

..

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78052    0.07962  -9.803  < 2e-16 ***
smoker      -0.37858    0.12566  -3.013  0.00259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.5  on 13  degrees of freedom
Residual deviance: 632.3  on 12  degrees of freedom
```

Figure: Women and Smoking - Simpson's Paradox

- Problem: Smoking seems to reduce the death rate! Why?
- Explanation: Inappropriate marginalisation!
- Interested in how x affects Y, but a third binary variable z is lurking



- Marginalisation over z gives misleading inference about how Y depends on x: $\mathbb{E}[Y|X = x, Z = z]$ increases with x for each z (left panel), but $\mathbb{E}[Y|X = x]$ decreases with x (right panel)

Example (Women and Smoking, continued)

```
> summary(glm(cbind(dead,alive)~age+smoker-1,data=smoking,binomial))

Call:
glm(formula = cbind(dead, alive) ~ age + smoker - 1, family = binomial,
    data = smoking)

..

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
age18-24   -3.8601     0.5939  -6.500 8.05e-11 ***
age25-34   -3.7401     0.3715 -10.067  < 2e-16 ***
age35-44   -2.5190     0.2499 -10.079  < 2e-16 ***
age45-54   -1.7468     0.2157  -8.097 5.62e-16 ***
age55-64   -0.6793     0.1621  -4.190 2.78e-05 ***
age65-74    1.2279     0.1934   6.349 2.17e-10 ***
age75+     23.9472 11293.1430   0.002   0.9983
smoker      0.4274     0.1770   2.414   0.0158 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 902.7701  on 14  degrees of freedom
Residual deviance:   2.3809  on  6  degrees of freedom
```

Figure: Women and Smoking - Age Included

Assume response variables of interest Y_i takes values $y \in \{0, 1, 2, \ldots\}$
– perhaps with upper bound m

↪ depending on sampling scheme/experiment

Three standard models:

- unconstrained responses $Y_i \overset{indep}{\sim} \text{Poisson}(\mu_i)$
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j=1}^{d} Y_j = m$ having multinomial distribution, with probabilities $(\pi_1, \ldots, \pi_d)$ and denominator m.
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j \in I_k} Y_j = m_k$ (for disjoint index parition sets $\{I_k : k = 1, ..., K\}$) having product multinomial.

These models are *very* closely related.

Assume response variables of interest Y_i takes values $y \in \{0, 1, 2, \ldots\}$
– perhaps with upper bound m

↪ depending on sampling scheme/experiment

Three standard models:

- unconstrained responses $Y_i \overset{indep}{\sim} \text{Poisson}(\mu_i)$
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j=1}^{d} Y_j = m$ having multinomial distribution, with probabilities $(\pi_1, \ldots, \pi_d)$ and denominator m.
- constrained responses $(Y_1, \ldots, Y_d)$ subject to $\sum_{j \in I_k} Y_j = m_k$ (for disjoint index parition sets $\{I_k : k = 1, ..., K\}$) having product multinomial.

These models are *very* closely related.

Lemma (Poisson and Multinomial)

Let $Y_1, \ldots Y_d$ be independently distributed as Poisson, with means $\mu_1, ..., \mu_d$, respectively. Then the conditional distribution of

$$(Y_1, ..., Y_d) \text{ given } \sum_{k=1}^{d} Y_i = m$$

is multinomial with denominator m and probabilities $\pi_i = \mu_i / \sum_j \mu_j$.

Proof.

Using Bayes' formula, $\mathbb{P}\left[\bigcap_{i=1}^{d}\{Y_i = y_i\} \,\middle|\, \sum_j Y_j = m\right]$ equals

$$
\begin{aligned}
&= \frac{\mathbb{P}[\sum_j Y_j = m | \bigcap_{i=1}^{d}\{Y_i = y_i\}]\mathbb{P}[\bigcap_{i=1}^{d}\{Y_i = y_i\}]}{\mathbb{P}[\sum_j Y_j = m]} \\
&= \mathbf{1}[\Sigma_{j=1}^{d} y_j = m]\frac{\prod_{j=1}^{d} e^{-\mu_j}\frac{\mu_j^{y_j}}{y_j!}}{e^{-\sum \mu_j}\frac{(\sum \mu_j)^m}{m!}} \\
&= \mathbf{1}[m = \Sigma_{j=1}^{d} y_j]\frac{\prod_{j=1}^{d} e^{-\mu_j}\frac{\mu_j^{y_j}}{y_j!}}{e^{-\sum \mu_j}\frac{(\sum \mu_j)^{\sum y_j}}{m!}} \\
&= \mathbf{1}[\Sigma_{j=1}^{d} y_j = m]\frac{m!}{y_1! \ldots y_d!}\prod_{i=1}^{d}\left(\frac{\mu_i}{\sum_{j=1}^{d}\mu_j}\right)^{y_i}
\end{aligned}
$$

□

Example (Example: Smoking data (Doll and Hill))

Table: Lung cancer deaths in British male physicians. The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Example (Jacamar Data)

Table: Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	Aphrissa boisduvalli N/S/E	*Phoebis argante* N/S/E	*Dryas iulia* N/S/E	*Pierella luna* N/S/E	*Consul fabius* N/S/E	*Siproeta stelenes*† N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Assume that $Y_i \overset{indep}{\sim} \text{Pois}(\mu_i)$

$$\mathbb{P}[Y_i = y] = e^{-\mu}\frac{\mu^y}{y!}, \quad y \in \mathbb{Z}_+, \mu > 0$$

- Exponential family
- Natural parameter $\phi = \log \mu$
- Can fit GLM via some link function $g(\mu)$

$$\underbrace{Y_i}_{\in \mathbb{Z}_+} | x_i \overset{ind}{\sim} \text{Poisson}(\mu_i) \quad \text{such that} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^p, \quad i = 1, ..., n.$$

Log-linear model

Poisson GLM with canonical logarithmic link:

$$\boldsymbol{x}_i^\top \boldsymbol{\beta} = \log \mu_i$$

- Occasionally Y_i counts the events of a Poisson process up to time T_i, so

$$\mathbb{E}[Y_i] = \mu_i = \lambda_i T_i$$

with λ_i the intensity of the process. In this case one sets

$$g(\mu_i) = \log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log T_i$$

- $\log T_i$ is the so-called offset term and is treat as a known constant.

Looks fairly straightforward. What's the big deal?

Earlier lemma suggests intimate relationship with categorical data.

↪ Consider again the binary case, $d = 2$.

$$Y_2 | \{Y_1 + Y_2 = m\} \quad \sim \quad \text{Binomial}\left(m, \pi = \frac{\mu_2}{\mu_1 + \mu_2}\right)$$

- Hence if $\mu_1 = \exp(\gamma + \boldsymbol{x}_1^\top \boldsymbol{\beta})$, $\mu_2 = \exp(\gamma + \boldsymbol{x}_2^\top \boldsymbol{\beta})$,

$$\pi = \frac{\mu_2}{\mu_1 + \mu_2} = \frac{\mu_2 \mu_1^{-1}}{1 + \mu_2 \mu_1^{-1}} = \frac{\exp\{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^\top \boldsymbol{\beta}\}}{1 + \exp\{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^\top \boldsymbol{\beta}\}}.$$

- So we can estimate $\boldsymbol{\beta}$ using either a loglinear model or logistic model
  - ↪ but can't estimate γ from logistic model (lose absolute information)
- This is particularly convenient for fitting more general contingency tables.

- Contingency table entries: count data cross-classified by different categories
  - ↪ Example: jacamar data cross-classify butterflies by

$$6 \text{ species} \times 8 \text{ colours} \times 3 \text{ fates}$$

yielding 144 categories total, each with count $\in \{0, 1, \ldots, 14\}$. Sampling scheme may fix certain totals — in the jacamar data the total for each species and colour is fixed, so reponses are trinomial: (not eaten, sampled, eaten)

Poisson vs multinomial vs product multinomial likelihoods (r=row, c=column):

- Poisson $\left(\prod_{r,c} \left\{ e^{-\mu_{rc}} \frac{\mu_{rc}^{Y_{rc}}}{Y_{rc}!} \right\}\right)$
  - ↪ Just collect data, then arrange into table. Yields poisson distribution for each cell.
- Multinomial $\left(\frac{m!}{\prod_{r,c} Y_{rc}!} \prod_{r,c} \pi_{rc}^{Y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1\right)$
  - ↪ Keep collecting until $m = \sum_{rc} Y_{rc}$ is reached. Yields multinomial distribution for table entries.
- Product multinomial $\left(\prod_r \left\{ \frac{m_r!}{\prod_c Y_{rc}!} \prod_c \pi_{rc}^{Y_{rc}} \right\}, \ \sum_c \pi_{rc} = 1, \forall r\right)$
  - ↪ Fix row totals alone in advance (e.g. fix # of butterflies in each colour/species category). In effect this treats row categories as independent subpopulations, i.e. independent multinomials for table entries of each row.

All three models can be easily fitted using Poisson GLM (with appropriate offsets).

- For multinomial settings, arrange as two-way layout with row totals fixed (single row in multinomial layout, several rows in product multinomial).
  - ↪ In Jacamar data, create new variable `species*colour` with 48 categories – yields 48 rows r, and leaves 3 columns c corresponding to `fate`.
- Model (r, c)-the cell as independent Poisson with mean

$$\mu_{rc} = \exp(\gamma_r + \boldsymbol{x}_{rc}^\top \boldsymbol{\beta})$$

  - γ_r acounts for the overall mean row count.
  - $\boldsymbol{x}_{rc}$ is such that $\sum_c \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} = \beta_{rc}$ which accounts for deviations of the cth column from the overall row count.
  - Interest focuses on $\boldsymbol{\beta}$, not γ_r, so will not worry about identifiability constraints.
- Conditioning on row totals being m_r get (product) multinomial model with probabilities $\{\pi_{rc} : \pi_{rc} \geq 0, \sum_c \pi_{rc} = 1\}$,

$$\pi_{rc} = \frac{\mu_{rc}}{\sum_d \mu_{rd}} = \frac{\exp(\gamma_r + \boldsymbol{x}_{rc}^\top \boldsymbol{\beta})}{\sum_d \exp(\gamma_r + \boldsymbol{x}_{rd}^\top \boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_{rc}^\top \boldsymbol{\beta})}{\sum_d \exp(\boldsymbol{x}_{rd}^\top \boldsymbol{\beta})},$$

  and the $\{\gamma_r\}$ parameters become irrelevant.

Thus multinomial loglikelihood is (up to constants)

$$\ell_{\text{Mult}}\left(\boldsymbol{\beta}; \boldsymbol{y} \middle| \sum_c Y_{rc} = m_r\right) \equiv \sum_{r,c} Y_{rc} \log \pi_{rc}$$

$$= \sum_r \left\{ \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - m_r \log \left(\sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} \right) \right\}.$$

The unconstrained Poisson model, would give loglikelihood (up to constants)

$$\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{r,c} Y_{rc} \log \mu_{rc} - \mu_{rc} = \sum_r \left(M_r \gamma_r + \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - e^{\gamma_r} \sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} \right)$$

where $M_r = \sum_c Y_{rc}$ is not given (i.e. is Poisson random variable). Writing

$$\tau_r = \sum_c \mu_{rc} = e^{\gamma_r} \sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} = \mathbb{E}[M_r]$$

for the row total means and using[1] $\gamma_r = \log \tau_r - \log \left\{ \sum_c \exp(\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}) \right\}$ yields

$$\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \left(\sum_r M_r \log \tau_r - \tau_r \right) + \sum_r \left\{ \sum_c Y_{rc} \boldsymbol{x}_{rc}^\top \boldsymbol{\beta} - M_r \log \left(\sum_c e^{\boldsymbol{x}_{rc}^\top \boldsymbol{\beta}} \right) \right\}.$$

[1] Under identifiability constraints $\boldsymbol{\gamma} \leftrightarrow \boldsymbol{\tau}$ is 1-1 function.

So using Bayes' theorem, we obtain:

$$\boxed{\ell_{\text{Poiss}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \ell_{\text{Poiss}}(\boldsymbol{\tau}; \boldsymbol{m}) + \ell_{\text{Mult}}(\boldsymbol{\beta}; \boldsymbol{Y} | \boldsymbol{M} = \boldsymbol{m})}$$

- Hence inferences on $\boldsymbol{\beta}$ using the multinomial model are equivalent to those based on the Poisson model, provided the row parameters γ_r are included.
- A more detailed calculation shows that the MLE $\hat{\boldsymbol{\beta}}$ and its sampling distribution are identical under the two models.

Example (Smoking Data (Doll and Hill), continued.)

Table: Lung cancer deaths in British male physicians. The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Example (Smoking Data (Doll and Hill), continued.)

- Suppose number of deaths y has Poisson distribution, mean $T\lambda(d,t)$, where T is man-years at risk, d is number of cigarettes smoked daily and t is time smoking (years).
- Log-linear model
  - $\lambda_{rc} = \exp(\gamma_r + \beta_c)$
  - deviance 51.47 on 48 df, one parameter for each row and column
- Model taking into account nature of rows/columns:

$$\lambda(d,t) = \{e^{\gamma_0} + \exp(\gamma_1 + \beta_2 \log d)\} \exp(\beta_3 \log t)$$

- deviance is 59.58 on 59 df with just 4 parameters overall

Table: Parameter estimates (standard errors) for lung cancer data.

	γ_0	γ_1	β_2	β_3
Smokers only	0.96 (25.4)	2.15 (1.45)	1.20 (0.40)	4.50 (0.34)
All data	2.94 (0.58)	1.82 (0.66)	1.29 (0.20)	4.46 (0.33)

Perfect Separation can also affect multinomial regression:

- If any one class is separated from all the rest by a covariate hyperplane, then by reduction to the binary case we can see that the MLE for that class will fail to exist.
- Detection more subtle: now there are $\binom{p}{2}$ cases to determine.
- Simple heuristic: insist that the iterative method used to maximize the likelihood terminate only after both the value of the likelihood function and the parameter vector stop changing.
- Different inference approaches such as extended logistic regression (Clarkson & Jennrich, 1991), bias-reduced ML (Firth, 1993), and exact logistic regression (Mehta & Patel, 1995) can be used that are stable to separation (as long as we know we have a problem!).
- Can also fit penalised loglinear regression with ridge/lasso penalty.