Generalized linear models II

Sofia Olhede

November 24, 2020

1. Special examples
2. Sparseness
3. Separation I

Bernoulli Observations

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

Binomial Observations

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.

Bernoulli versus Binomial

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 $[0, 1]$.

Design matrix vs Outcomes

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)

Jittered residuals

Brillinger & Preisler (1983), Brillinger (1996) suggest the use of "jittered quantile residuals" for binary responses (a.k.a *quantile residuals*).

Idea:

- If $Y \sim F(y;\theta)$ continuous, then $U := F(Y;\theta)$ has a uniform distribution on $(0,1)$.
- If $\hat{\theta} \simeq \theta \implies \hat{U} := F(Y|\hat{\theta}) \overset{d}{\approx} \text{Unif}(0,1)$.
- Hence, obtain quantile residuals $R_i = \Phi^{-1}(F(Y_i;\hat{\theta})) \overset{d}{\approx} N(0,1)$.

In discrete case must be a little more careful:

- let $\hat{U}_i^{(1)} \sim \text{Unif}[0,\hat{\pi}_i]$ and $\hat{U}_i^{(2)} \sim \text{Unif}[\hat{\pi}_i,1]$.
- Jittering randomisation: $\hat{U}_i = \hat{U}_i^{(1)} Y_i + \hat{U}_i^{(2)}(1-Y_i)$
- And so $\Phi^{-1}(\hat{U}_i) \overset{d}{\approx} N(0,1)$ distribution, if fit is good and model correct.

May now construct some plots, e.g. normal probability plots, plots against covariates, plots against "linked fits".

Other aspects of the design matrix

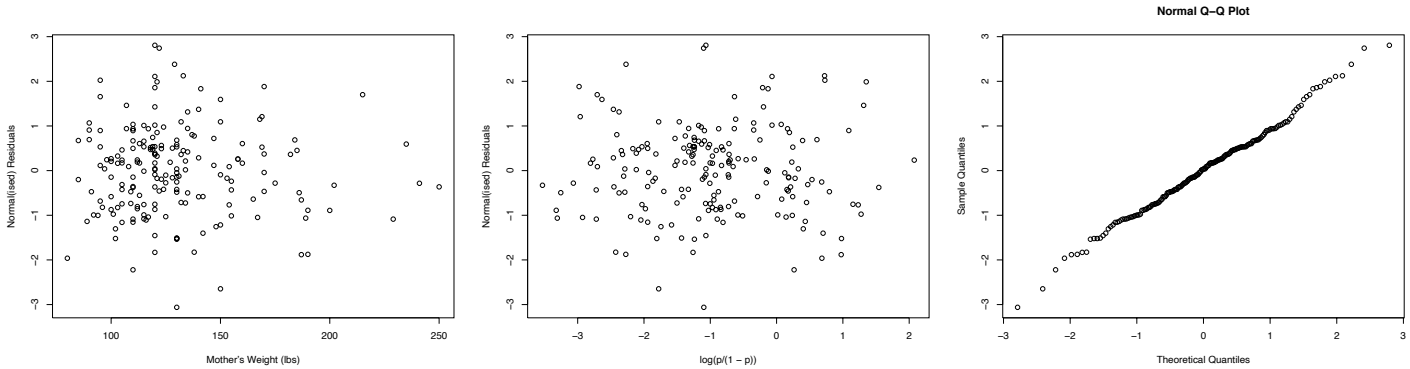


Figure: Jittered Quantile Residuals for Birth Weight Data

Separation

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:

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

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.

Separation II

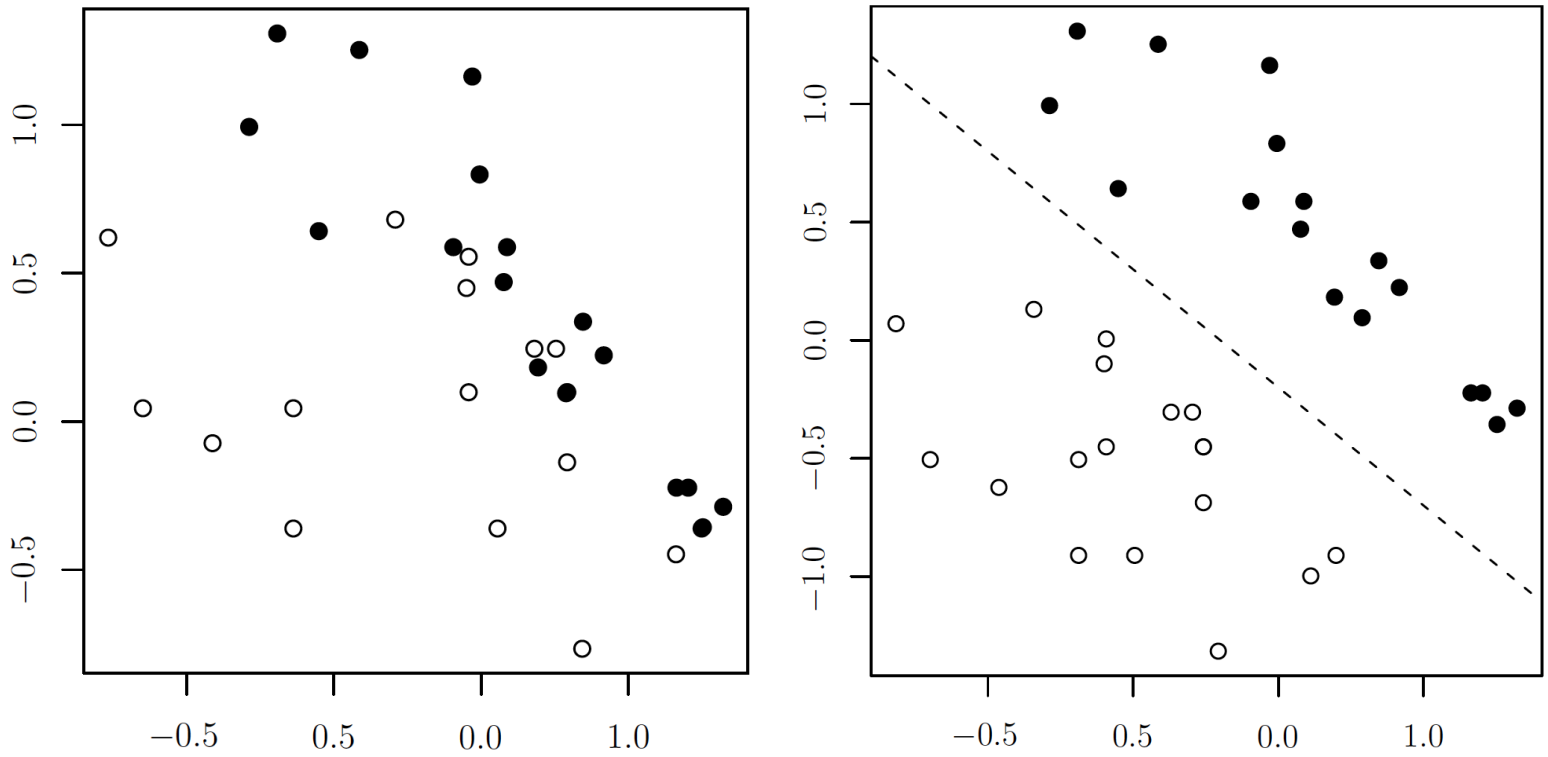


Figure: Overlap (left) vs Complete Separation (right)

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

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

Separation III

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 $t > 0$, $\boldsymbol{x}_i^\top \gamma > 0 \iff t(\boldsymbol{x}_i^\top \gamma) > 0$ and $\boldsymbol{x}_i^\top \gamma < 0 \iff t(\boldsymbol{x}_i^\top \gamma) < 0$.
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).

□

Separation IV

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:

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

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

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

Overlap

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.

Contingency table

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

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

Contingency table II

Table: Notation for 2×2 table.

	Success	Failure	Total
Case	R_1	$m_1 - R_1$	m_1
Control	R_0	$m_0 - R_0$	m_0
Total	$R_1 + R_0$	$m_1 + m_0 - R_1 - R_0$	$m_1 + m_0$

Contingency table III

- 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.

Count Data I

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

Proof.

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

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

Count Data II

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

$\updownarrow$

Poisson GLM with canonical logarithmic link:

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

Count Data III

- 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?

Count Data IV

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.

Count Data V

- 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.

Count Data VI

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.

Count Data VII

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[8] $\gamma_r = \log \tau_r - \log\left\{\sum_c \exp(\boldsymbol{x}_{rc}^\top \boldsymbol{\beta})\right\}$ yields

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

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

Count Data VII

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.

Count Data VIII

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.

Nonparametric relationships with x_i

So far we have discussed the following setup:

$$Y_i \mid \boldsymbol{x}_i^\top \overset{ind}{\sim} \text{Dist}[\phi_i] \rightarrow \begin{cases} \phi_i = g(\boldsymbol{x}_i^\top) = \boldsymbol{x}_i^\top \boldsymbol{\beta}), \\ \boldsymbol{\beta} \in \mathbb{R}^p, \end{cases}$$

with $\boldsymbol{\beta}$ to be estimated from data, e.g.

- Dist $= \mathcal{N}(\mu_i, \sigma^2)$ and $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (Gaussian linear regression)
- Dist $\in \mathrm{ExpFamily}(\phi_i)$ and $\phi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$) (GLM)

Would now like to extend model to a more flexible dependence:

$$Y_i \mid \boldsymbol{x}_i^\top \overset{ind}{\sim} \text{Dist}[\phi_i] \rightarrow \begin{cases} \phi_i = g(\boldsymbol{x}_i^\top), \\ g \in \mathcal{F} \subset L^2(\mathbb{R}^p) \text{ (say)}, \end{cases}$$

with $g : \mathbb{R}^p \rightarrow \mathbb{R}$ unknown, to be estimated given data $\{(Y_i, \boldsymbol{x}_i^\top)\}_{i=1}^n$.

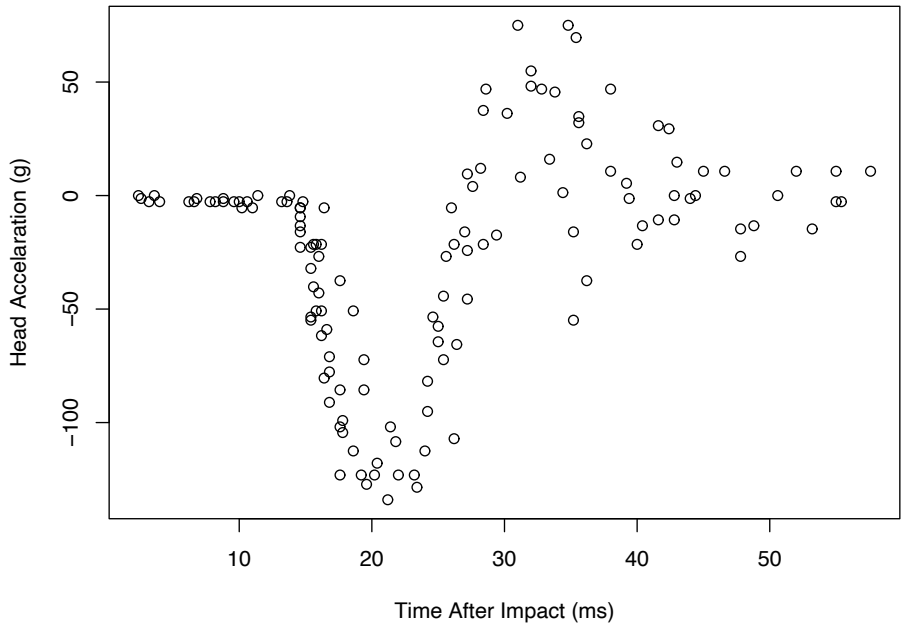
- A *nonparametric* problem (parameter ∞-dimensional)!
- How to estimate g in this context?
- $\mathcal{F}$ is usually assumed to be a class of smooth functions (e.g., C^k).

Nonparametric relationships with x_i

Start from simplest problem:

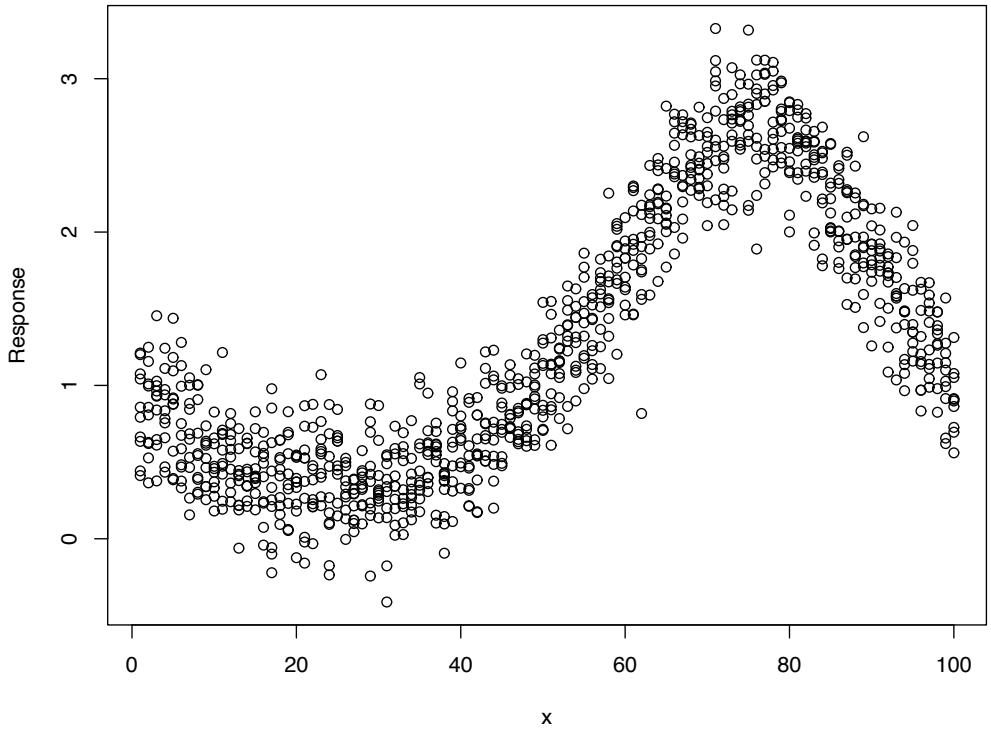
$$\left.\begin{array}{l}\text{Dist} \equiv \mathcal{N}(\mu_i, \sigma^2) \\ x_i \in \mathbb{R}\end{array}\right\} \Longrightarrow \quad Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Figure: Motorcycle Accident Data



Nonparametric relationships with x_i

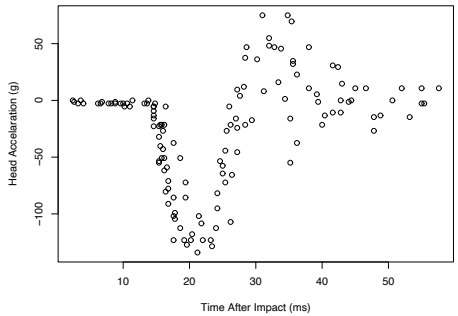
- Ideally: large sample plus multiple x_i with same value (many large covariate classes):



- Then average Y_i's at each covariate class and interpolate ...
- But this is never the case in practice... ...

Nonparametric relationships with x_i

- Usually unique x_i distinct:



- Here is where the smoothness assumption comes in
- Since have distinct value for each x_i, need to borrow information from nearby ...
- ... use continuity!!! (or even better, *smoothness*)

- Recall: A function $g : \mathbb{R} \to \mathbb{R}$ is *continuous* if:

$$\forall\, \epsilon > 0\ \exists\, \delta > 0 :\ |x - x_0| < \delta \implies |g(x) - g(x_0)| < \epsilon.$$

- So maybe average Y_i's corresponding to x_i's in a δ-neighbourhood of x.
- Motivates the use of a kernel smoother ...

Nonparametric relationships with x_i

Naive idea: $\hat{g}(x_0)$ should be the average of Y_i-values with x_i's "close" to x_0.

$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{|x_i - x_0| \leq \lambda\}} \sum_{i=1}^{n} Y_i \mathbf{1}\{|x_i - x_0| \leq \lambda\}.$$

A weighted average! Choose other weights? Kernel estimator:

$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{\lambda}\right)} \sum_{i=1}^{n} Y_i K\left(\frac{x_i - x_0}{\lambda}\right) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i Y_i.$$

- K is a kernel function.
  - ↪ E.g. standard Gaussian pdf, $K(x) = \varphi(x)$.
- λ is the bandwidth parameter
  - ↪ small λ gives local behaviour, large λ gives global behaviour
- Choice of K not so important, choice of λ very important.
- The resulting fitted values are linear in the responses, i.e., $\hat{\boldsymbol{Y}} = \boldsymbol{S}_\lambda \boldsymbol{Y}$, where the smoothing matrix $\boldsymbol{S}_\lambda$ depends on $x_1, \ldots, x_n$, K, and λ. Analogous to a projection matrix in linear regression, but $\boldsymbol{S}_\lambda$ is not a projection.

Nonparametric relationships with x_i

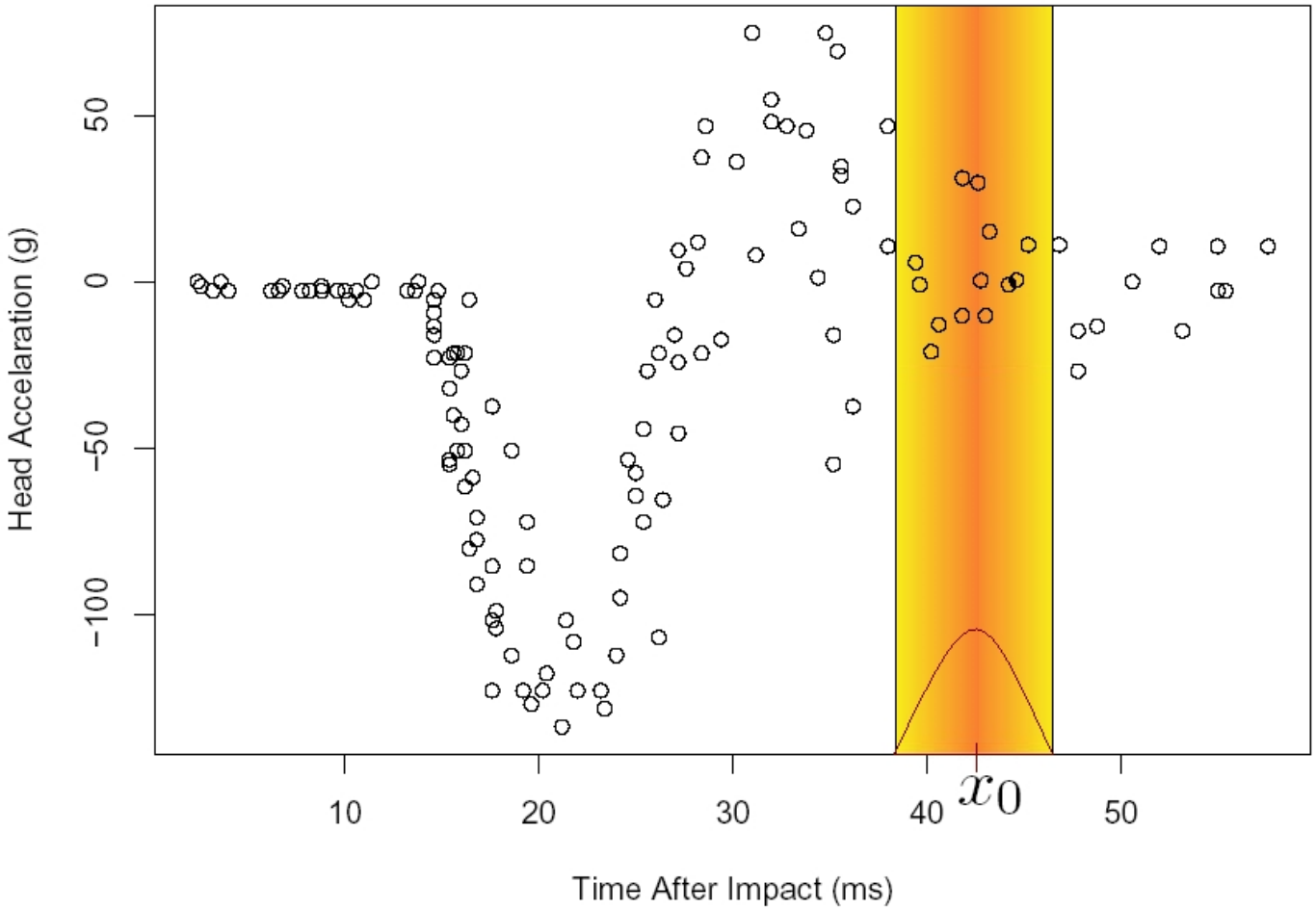


Figure: Visualising a kernel smoother at work

Nonparametric relationships with x_i

```
> plot(time,accel,xlab="Time After Impact (ms)",ylab="Head Accelaration (g)")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=0.7))
> lines(ksmooth(time,accel,kernel="normal",bandwidth=5),col="red")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=10),col="blue")
```

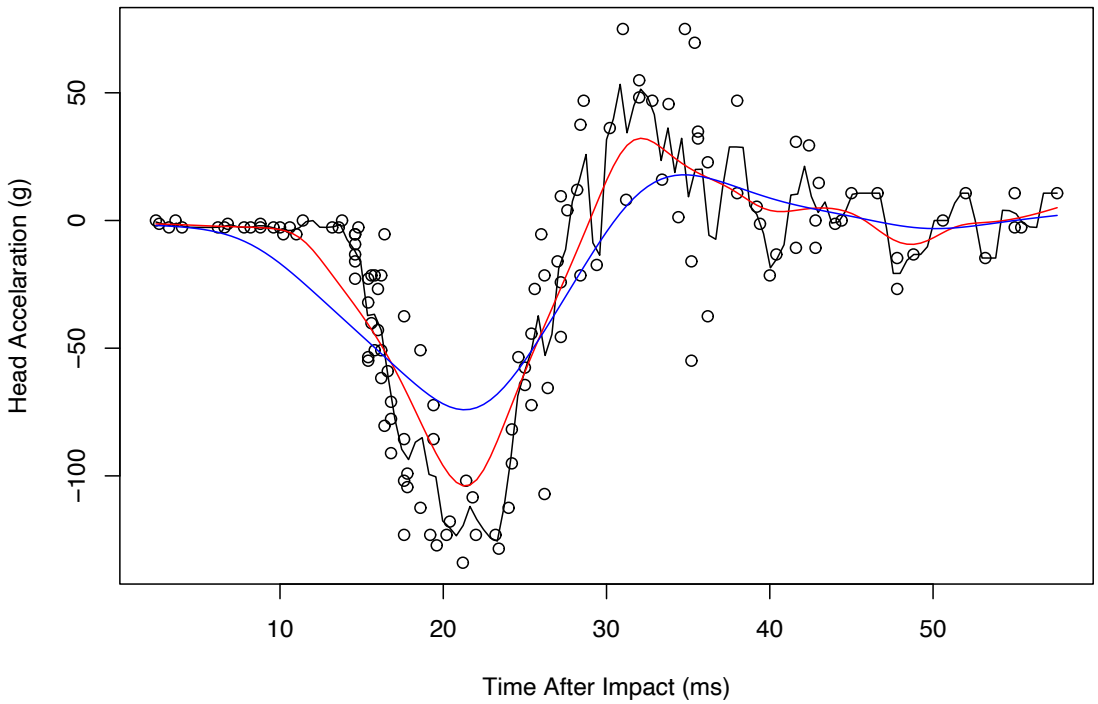


Figure: Motorcycle data kernel smooth for varying bandwidths