Methods of Applied Statistics I

STA2101H F LEC9101

Week 9

November 17 2021

Today

1. Upcoming events
2. Project
3. GLM Model Selection and Analysis
4. In the News
5. HW 7 and 8 (12.10 – 13.00)

Upcoming

- Monday Nov 22 3.30 Data Science ARES series
  Designing creative courses with students in mind Link

Upcoming

- Monday Nov 22 3.30 Data Science ARES series
  Designing creative courses with students in mind Link

- Friday Nov 19 Toronto Data Workshop Zoom link

- Friday, 19 November 2021, noon - 1pm
  Radu Craiu, Statistical Sciences @ U of T
  Dr. Radu V. Craiu is Professor and Chair of Statistical Sciences at the University of Toronto. His main research interests are in computational methods in statistics, especially, Markov chain Monte Carlo algorithms (MCMC), Bayesian inference, copula models, model selection procedures and statistical genetics.

Project

- Part I 3–5 pages, non-technical
  1. a description of the scientific problem of interest
  2. how (and why) the data being analyzed was collected
  3. preliminary description of the data (plots and tables)
  4. non-technical summary for a non-statistician of the analysis and conclusions

- Part II 3–5 pages, technical
  1. models and analysis
  2. summary for a statistician of the analysis and conclusions

- Part III Appendix

  R script or .Rmd file; additional plots; additional analysis; References

- 40 points total
- Part I:
  description of data and scientific problem 5
  suitability of plots and tables 5
  quality of the presentation 5

clear, non-technical, concise but thorough

- 40 points total

- Part I:
  description of data and scientific problem 5
  suitability of plots and tables 5
  quality of the presentation 5

  clear, non-technical, concise but thorough

- Part II:
  summary of the modelling and methods 5
  suitability and thoroughness of the analysis 10

  justification for choices
  model checks, data checks

- 40 points total

- Part I:
  description of data and scientific problem 5
  suitability of plots and tables 5
  quality of the presentation 5

  clear, non-technical, concise but thorough

- Part II:
  summary of the modelling and methods 5
  suitability and thoroughness of the analysis 10

  justification for choices
  model checks, data checks

- Part III:
  relevance of additional material 5
  complete and reproducible submission 5

Day | Week | Month | Year

Eastern Time

Search

December 2021

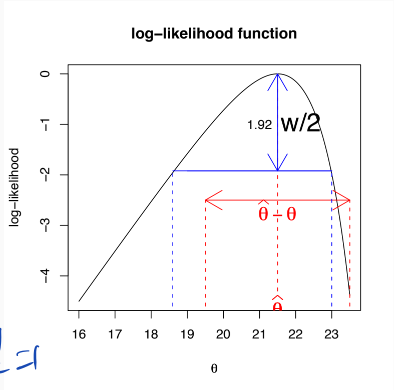
Today

Sun	Mon	Tue	Wed	Thu	Fri	Sat
14	15 RSC Annual CFI reports due Libai + Riccardo 10 AM Riccardo 12 PM 2 more...	16 Linbo Causal 11:30 AM Junhao exam 2:30 PM Comprehensive... 2:30 PM	17 STA2101 10 AM Grad exams 2:30 PM Comprehensive... 3:10 PM 2 more...	18 Yanbo 1 PM Grad exams 2:30 PM Comprehensive... 3:50 PM	19 CANSSI Showc... 11:45 AM	20 RSC AGM 10 AM CANSSI Showc... 11:45 AM
21	22 Harvard 12 PM Office Hour 7 PM	23	24 STA2101 10 AM DSI Research and A... 1 PM FW: DSI Research a... 1 PM Jiaying 2:15 PM Office Hour 4 PM	25 Yanbo 1 PM Seminar 3:30 PM	26 Ruyi Pan 10 AM	27
28 Hanukkah (begins at su...	29 Hanukkah Riccardo 11 AM Heather Battey 12 PM Meeting 3:45 PM EST Office Hour 7 PM	30 Staff meet 2 PM	Dec 1 STA2101 10 AM Office Hour 4 PM	2 Fw: December Boar... 1 PM Yanbo 1 PM December Boar... 1 PM EST Seminar 3:30 PM	3	4
5 Hanukkah	6 Meeting 3:45 PM EST Office Hour 7 PM	7 Statistical Sciences... 3 PM	8 STA2101 10 AM	9 Air Canada fligh... 12:55 PM Yanbo 1 PM Seminar 3:30 PM	10 Ed's workshop	11
12 Air Canada flight... 9:40 AM	13 Acadian Remembrance... DSI Catalyst Grant... 10 AM Meeting 3:45 PM EST Office Hour 7 PM	14 DSI Catalyst Grant... 10 AM	15 Office Hour 4 PM	16 Yanbo 1 PM	17	18
19	20 Meeting 3:45 PM EST	21	22 University Holiday Closure	23 Yanbo 1 PM	24 Christmas Eve	25 Christmas Day

Recap

- likelihood inference, confidence intervals, likelihood ratio tests, model choice via AIC
- binomial response, analysis of deviance, covariate classes, variable selection, goodness-of-fit, deviance residuals
- Poisson response
- see also R Markdown for Nov 3

- model
- maximum likelihood estimate
- confidence intervals
- likelihood ratio statistics
- likelihood ratio confidence intervals



$d=1$

$f(y;\theta)$

mle. $\hat\theta$

$\ell(\theta;y) = \log f(y;\theta)$

$$\frac{\partial}{\partial\theta}\ell(\theta;y) = 0$$

$$\hat\theta = \hat\theta(y)$$

$$\widehat{se}(\hat\theta) = \left\{\sqrt{\left[-\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta;y)\Big|_{\hat\theta}\right]^{-1}}\right\}_{(ii)}$$

$$\widehat{se}(\hat\theta_k) = \{j(\hat\theta)\}^{-1/2}_{kk}$$

$$2\{\ell(\hat\theta) - \ell(\theta)\} = w(\theta) \sim \chi^2_d \text{ under model}$$

- confidence intervals for β_1

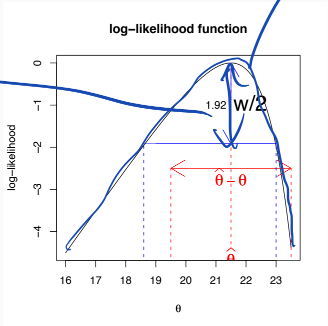
- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- `confint(logitmodcorrect):`
  `( -0.212, -0.024 )`

- profile log-likelihood for single parameters



$\ell_p(\beta_1) = \ell(\beta_1, \hat{\beta}_{(2)}(\beta_1))$ profile

$j(\theta) = -\ell''(\theta)$ $\qquad$ $i(\theta) = E\{-\ell''(\theta)\}$

$\qquad\qquad\qquad\qquad\qquad\qquad = \mathrm{var}_\theta\{\ell'(\theta)\}$ bec. $E_\theta \ell'(\theta) = 0$

Wald ↓ $\qquad$ $\{\ell'(\theta_0)\}^2$ as a var. estimate

$H_0: \theta = \theta_0$: $(\hat\theta - \theta_0)\, j^{1/2}(\hat\theta)$ or $(\hat\theta - \theta_0)\, i^{1/2}(\theta_0)$

or $(\hat\theta - \theta_0)\, i^{1/2}(\hat\theta)$

equiv. "to 1st order"

Score test $\quad \ell'(\theta_0)\, i^{-1/2}(\theta_0)$ ⟵ useful if $d = \dim(\theta)$ is v. large

$\{\theta_{10}, \hat\theta_2(\theta_{10})\}$

Heart data from Nov 3


```
heartmod3 <- update(heartmod, .~ . - behave + dibep )
summary(heartmod3)
```

```
##
## Call:
## glm(formula = chd ~ age + height + weight + sdp + dbp + chol +
##     cigs + dibep, family = binomial, data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.410  -0.435  -0.315  -0.223   2.839
##
## Coefficients:
##              Estimate Std. Error z value         Pr(>|z|)
## (Intercept) -13.55457    2.32014   -5.84 0.00000000515355 ***
## age           0.06477    0.01210    5.35 0.00000008610612 ***
## height        0.01600    0.03312    0.48           0.6289
## weight        0.00782    0.00388    2.02           0.0437 *
## sdp           0.01772    0.00637    2.78           0.0054 **
## dbp          -0.00015    0.01082   -0.01           0.9890
## chol          0.01106    0.00152    7.27 0.00000000000036 ***
## cigs          0.02092    0.00427    4.90 0.00000098078442 ***
## dibep         0.65338    0.14522    4.50 0.00000681630545 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1779.2  on 3141  degrees of freedom
## Residual deviance: 1580.7  on 3133  degrees of freedom
```

$e^{.65} \approx 1.9$

Forensics

The coefficient above for `dibepB` is positive, and significantly so, suggesting an *increase* in risk of `chd` for "type B" relative to "type A". This is not consistent with the coefficients for `behave` above, where both B1 and B2 showed a decrease risk relative to A1. What is going on?

```
xtabs(~ chd + behave, data = wcgs)
```

```
##      behave
## chd     A1   A2   B3   B4
##    no  234 1177 1155  331
##    yes  30  148   61   18
```

no
yes
3154

```
xtabs(~ chd + dibep, data = wcgs)
```

```
##      dibep
## chd      A    B
##    no 1486 1411
##    yes  79  178
```

B3 + B4 = 1486

```
1155 + 331; 1177+234
```

```
## [1] 1486
```

```
## [1] 1411
```

Deviance residuals

`glm.diag; library(SMPracticals)`

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

Deviance residuals

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

Deviance: $2\sum_{i=1}^{n}\{y_i \log\{y_i/n_ip_i(\hat\beta)\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_ip_i(\hat\beta))\}\}$

Dev. $2\{\ell(\tilde p) - \ell(\hat p)\}$

$\rightarrow \tilde p_i = \frac{y_i}{n_i}$

$\rightarrow \hat p_i = \frac{e^{x_i^T\hat\beta}}{1+e^{x_i^T\hat\beta}}$

succ — fail

$O \log \frac{O}{E}$ $\quad O \log \frac{O}{E}$

Observed

Expected (model)

$\sim \chi^2_{n-d}$ $\quad (\beta_1, \dots, \beta_d)$

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-1.4989   -0.7726   -0.1265    0.7997    1.4351
```

Deviance: $2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat\beta)\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat\beta))\}]$

approximately χ^2_{n-q}

$\sum r_{Di}^2 = \text{Dev.}$

$$r_{Di} = \pm\surd(2[y_i \log\{y_i/n_i\hat p_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat p_i)\}])$$

6 0
6 1
6 0
6 2

? $\frac{y_i + .1}{n_i + .1}$ if $y_i = 0$ or n_i ?

... example 10.18: residuals

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

... example 10.18: residuals

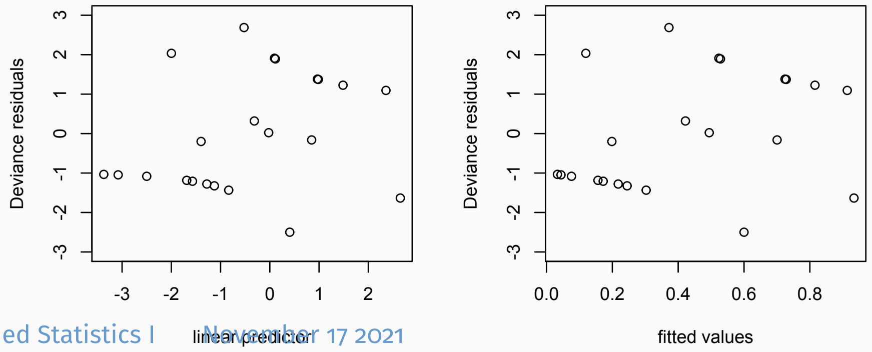
```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

SM

53 rows ↓ 2? Binomials



More on residuals

ELM 6.3; SM 10.4

- there are many versions of residuals

- ?residuals.glm

```
residuals(glm.object, type = c("deviance", "pearson", "working",
"response", "partial"),...
```

X ~~no~~ model heart$residuals ← "working residuals"

More on residuals

- there are many versions of residuals

- `?residuals.glm`
  `residuals(glm.object, type = c("deviance", "pearson", "working", "response", "partial"),...`

- Binomial deviance ≈

$$\sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{n} \left(\frac{y_i - n_i\hat{p}_i}{\widehat{\text{se}}(y_i)}\right)^2 = \sum_{i=1}^{n} r_{Pi}^2$$

Poisson regression

- $y_i \sim Po(\lambda_i)$ $\quad f(y_i; \lambda_i) =$ $\lambda_i^{y_i} e^{-\lambda_i} / y_i!$ $\quad y_i = 0, 1, 2, \ldots$
- $\log(\lambda_i) = x_i^T \beta$ $\quad e^{y_i \log \lambda_i - \lambda_i} = \ell(\lambda_i)$ $\quad \ell(\beta) = \sum_{i=1}^{n} (y_i x_i^T \beta - e^{x_i^T \beta})$
- where's the ϵ? $\quad E(y_i) = \lambda_i = e^{x_i^T \beta}$

$g(E(y_i)) = \log(E y_i) = x_i^T \beta =$ linear pred.

HW 8

$\lambda_i = \beta x_i$

$\ell(\beta) = e^{y_i \log(\beta x_i) - \beta x_i}$

link = identity

Poisson regression

- $y_i \sim Po(\lambda_i)$ $\quad f(y_i; \lambda_i) =$
- $\log(\lambda_i) = x_i^T \beta$
- where's the ϵ? $\quad E(y_i) =$ λ_i $\quad \text{var}(y_i) =$ λ_i

Poisson regression

- $y_i \sim Po(\lambda_i)$ $\quad f(y_i;\lambda_i) =$
- $\log(\lambda_i) = x_i^T\beta$ $\quad$ $E(y_i) = \lambda_i = e^{x_i^T\beta}$
- ~~where's the ϵ?~~ $\quad E(y_i) =$ $\quad \mathrm{var}(y_i) =$ $\quad y_i \quad \hat{E}$
- $L(\beta; y) =$
- $\ell(\beta; y) =$ $\sum_{i=1}^{n}\{y_i x_i^T\beta - e^{x_i^T\beta}\}$
- $\hat\beta =$ $\quad \left.\frac{\partial \ell}{\partial \beta}\right|_{\hat\beta} = 0 \quad : \quad \sum\{y_i x_i^T - e^{x_i^T\hat\beta} x_i^T\} = 0$

$$\Rightarrow \sum y_i x_i^T = \sum e^{x_i^T\hat\beta} x_i^T$$

... Poisson regression

$f(y_i) = y_i^{\lambda_i} e^{-\lambda_i} / y_i!$

- saturated model $y_i \sim Po(\lambda_i)$, $\tilde{\lambda}_i = y_i$
- compare saturated fit to log-linear fit $\ell(\tilde{\lambda}) - \ell(\hat{\lambda}) =$

$\hat{\lambda}_i = e^{x_i^T \hat{\beta}}$ can. link

- *Deviance* $= 2\{\ell(\tilde{\lambda}) - \ell(\hat{\lambda})\} = 2\sum_i \{y_i \log \tilde{\lambda}_i - \tilde{\lambda}_i - (y_i \log \hat{\lambda}_i - \hat{\lambda}_i)\}$

$= 2\sum \{y_i \log(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)\}$

- *Deviance* $\approx$ $2\sum\{O \log \frac{O}{E} - (O - E)\}$

$$X_P^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$$

Pearson's chi-square

- for y_i not "too close" to 0, Deviance or X^2 give a measure of fit of the Poisson regression model $\dot\sim \chi^2_{n-p}$

$$X^2_{Bin} = \sum \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$

Link functions

- generalized linear model: $g\{E(y_i)\} = x_i^T \beta$

- Binomial, $g(p_i) = \log\{p_i/(1 - p_i)\}$
- Poisson, $g(\lambda_i) = \log(\lambda_i)$
- these are mathematically convenient, but might not always be appropriate

Binomial $E(y_i) = p_i$

$\rightarrow \mathrm{var}(y_i) = n_i p_i (1 - p_i)$

$\mathrm{var}(y_i) = V(\mu_i)$

$= V\{E(Y_i)\}$

(2)

Po $E(y_i) = \lambda_i$

$\mathrm{var}(y_i) = \lambda_i$

1) G-of-fit via deviance

Link functions

- generalized linear model: $g\{E(y_i)\} = x_i^T\beta$

- Binomial, $g(p_i) = \log\{p_i/(1-p_i)\}$
- Poisson, $g(\lambda_i) = \log(\lambda_i)$
- these are mathematically convenient, but might not always be appropriate

- example $y_i = 1\{Z_i > 0\}$, $Z_i \sim N(\beta_0 + \beta_1 x_i, 1)$
- $p_i = \mathrm{pr}(y_i = 1 \mid x_i) = 1 - \Phi\{-(\beta_0 + \beta_1 x_i)\} = \Phi(\beta_0 + \beta_1 x_i)$
- $g(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i$

probit link

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx$$

$$L(\beta) = \prod p_i^{y_i}(1-p)^{n_i-y_i} = \prod\{\Phi(\beta_0+\beta_1 x_i)\}^{y_i}\{1-\Phi\}^{\cdots} \quad = \Pr\{N(0,1)\le z\}$$

$$\ell(\beta) = \log = \sum y_i \log \Phi(\quad) + (n_i - y_i)\log\{1-\Phi(\;)\}$$

Link functions

- generalized linear model: $g\{E(y_i)\} = x_i^T\beta$

- Binomial, $g(p_i) = \log\{p_i/(1-p_i)\}$
- Poisson, $g(\lambda_i) = \log(\lambda_i)$
- these are mathematically convenient, but might not always be appropriate

- example $y_i = 1\{Z_i > 0\}$, $\qquad Z_i \sim N(\beta_0 + \beta_1 x_i, 1)$
- $p_i = \mathrm{pr}(y_i = 1 \mid x_i) = 1 - \Phi\{-(\beta_0 + \beta_1 x_i)\} = \Phi(\beta_0 + \beta_1 x_i)$
- $g(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i$ — probit link

- example y_i counts numbers of events over time period t_i — HW8
- $E(y_i) = \beta t_i$ for example, or $\beta_0 + \beta_1 t_i$, or ...
- $g(\lambda_i) = \lambda_i$
- rate models use Poisson with an offset y_i is number of events in time T, $E(y_i) = T\lambda_i$

… Link functions

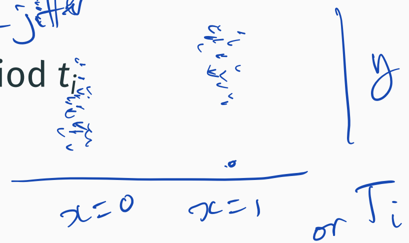
HW8

- example y_i counts numbers of events over time period t_i
- $E(y_i) = \beta t_i$ for example, or $\beta_0 + \beta_1 t_i$, or …
- $g(\lambda_i) = \lambda_i$

$\lambda_i = \beta x_i$ no intercept

glm(y ~ x_i, family = Poisson(link = "?"))

... Link functions



- example y_i counts numbers of events over time period t_i HW8
- $E(y_i) = \beta t_i$ for example, or $\beta_0 + \beta_1 t_i$, or ...
- $g(\lambda_i) = \lambda_i$
- rate models use Poisson with an offset y_i is number of events in time T, $E(y_i) = T\lambda_i$
- $\log(T\lambda_i) = \log(T) + \log(\lambda_i)$, $\lambda_i = x_i^T\beta$
- `glm(y ~ x + offset(log(T)), family = poisson, data = ...)`

Overdispersion

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \text{E}(Y_i) = n_i p_i, \quad \text{Var}(Y_i) = n_i p_i(1 - p_i)$
- variance is determined by the mean

Overdispersion

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i (1 - p_i)$
- variance is determined by the mean

-
```
bmod <- glm(cbind(survive,total-survive) ~ location + period, family = binomial,
            data = troutegg)

summary(bmod)
 Null deviance: 1021.469  on 19  degrees of freedom
## Residual deviance:   64.495  on 12  degrees of freedom
## AIC: 157.03
```

nearly always > 1

$\widehat{\mathrm{var}}\, y_i = \tilde{\phi}\, n_i \hat{p}_i (1 - \hat{p}_i)$

$> n_i \hat{p}_i (1 - \hat{p}_i)$

Overdispersion

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i(1 - p_i)$
- variance is determined by the mean

- ```
  bmod <- glm(cbind(survive,total-survive) ~ location + period, family = binomial,
              data = troutegg)

  summary(bmod)
   Null deviance: 1021.469  on 19  degrees of freedom
  ## Residual deviance:   64.495  on 12  degrees of freedom
  ## AIC: 157.03
  ```
- quasi-binomial: $\mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = \phi n_i p_i(1 - p_i)$
- estimate ϕ?
- usually use $X^2/(n - p)$, where

over-dispersion parameter

$$X^2 = \sum \frac{(y_i - n_i\hat{p}_i)^2}{n\hat{p}_i(1 - \hat{p}_i)}$$

"~~fudge~~"

variance inflation factor

$$\frac{X^2}{n-p} = \tilde{\phi}$$

est. of overdispersion
  ```

Quasibinomial

overdisp.Rmd; overdisp.html

Measures of risk

- see posted handout on case-control studies — coming
- consider for simplicity binomial responses with a single binary covariate:

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 z_i, \quad i = 1, \ldots, n$$

$z_i = 0$: $\quad \text{logit}\, p_i = \beta_0 \quad = \log \frac{p_0}{1-p_0} = \beta_0$

$z_i = 1$: $\quad$ " $= \beta_0 + \beta_1 \quad \log\left(\frac{p_1}{1-p_1}\right) = \beta_1 + \beta_0$

$$\beta_1 = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = \log\left\{\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right\}$$

log odds of $y=1$, when $z=1$ relative to $z=0$

$\beta_1 = 0 \iff p_0 = p_1$

Measures of risk

- see posted handout on case-control studies
- consider for simplicity binomial responses with a single binary covariate:

$$\mathrm{logit}(p_i) \sim \beta_0 + \beta_1 z_i, \quad i = 1, \dots, n$$

- no difference between groups $\iff$ odds-ratio $\equiv 1$

… Measures of risk

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk

... Measures of risk

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if p_1 and p_0 are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

- sometimes p_1/p_0 can be large but if p_1 and p_0 are both small the difference $p_1 - p_0$ might also be very small

- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if p_1 and p_0 are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

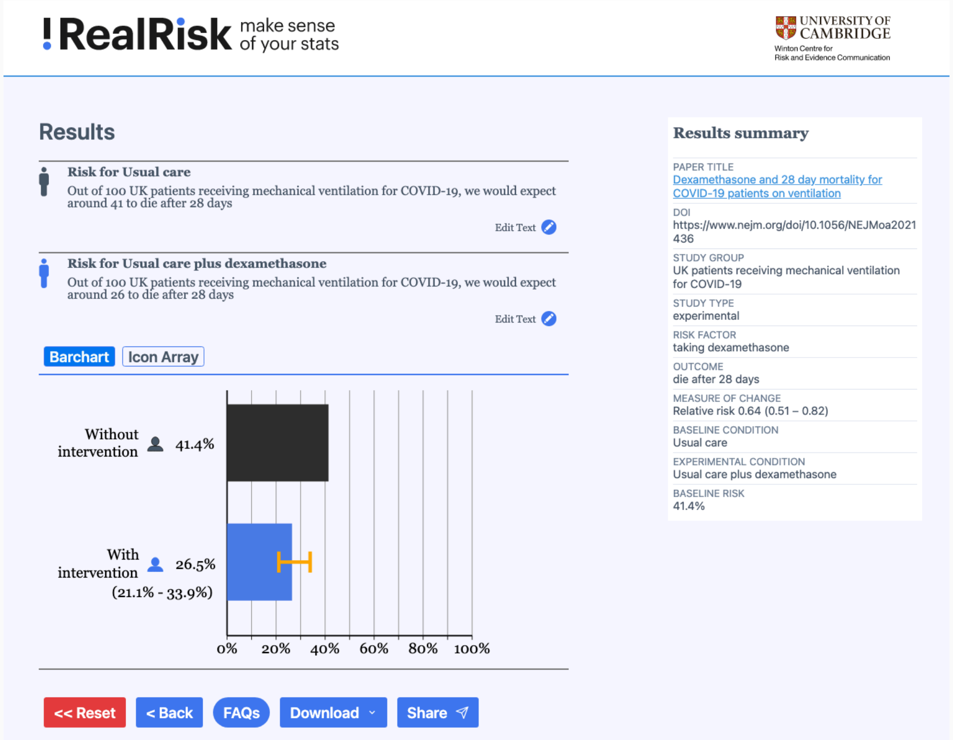
- sometimes p_1/p_0 can be large but if p_1 and p_0 are both small the difference $p_1 - p_0$ might also be very small
- in order to estimate the risk difference we need to know the baseline risk p_0

$\hat{p}_1$ $\hat{p}_0$ (depends on data)

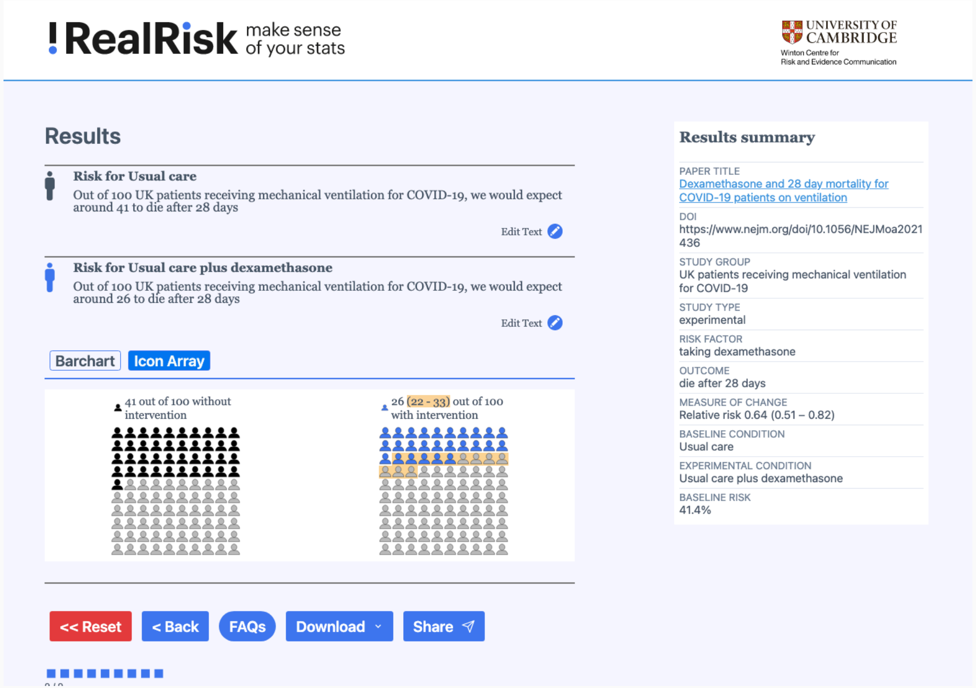
- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if p_1 and p_0 are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

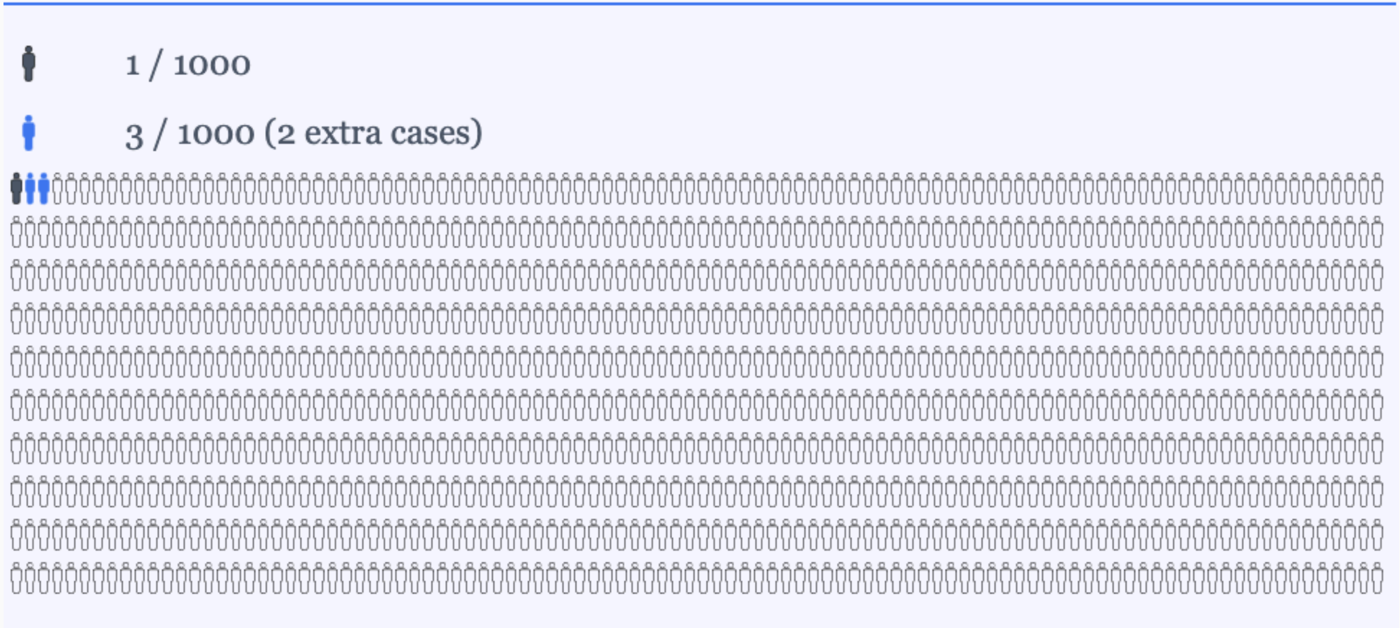
- sometimes p_1/p_0 can be large but if p_1 and p_0 are both small the difference $p_1 - p_0$ might also be very small
- in order to estimate the risk difference we need to know the baseline risk p_0
- bacon sandwiches www.youtube.com/watch?v=4szyEbU94ig
- risk calculator realrisk.wintoncentre.uk/p8



Odds ratio 0.64; baseline risk 41.4%



Odds ratio 0.64; baseline risk 41.4%



Odds ratio 2.91; baseline risk 1/1000

Biostats secret sauce

Whether we sample prospectively or retrospectively, the odds ratio is the same

	Lung cancer	
	1	0
	cases	controls
smoke = 1 (yes)	688	650
smoke = 0 (no)	21	59
	709	709

sample case-control study

$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right)$

retro: $OR = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 2.97$

prosp: $OR = \frac{\{688/(688+650)\}/\{650/(688+650)\}}{21/(21+59)/\{59/(21+59)\}} = \frac{688 \times 59}{650 \times 21} = 2.97$

see "case-control", FELM §2.5,6, SM §10.4.2

Generalized linear models

`glm` has several options for `family`

```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

Generalized linear models

`glm` has several options for `family`

```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

Each of these is a member of the class of generalized linear models

Generalized: distribution of response is not assumed to be normal

Linear: some transformation of $E(y_i)$ is of the form $x_i^T\beta$ link function

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$

$y_i \sim Po(\lambda_i)$

$f(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$

$e^{-\lambda_i \beta} e^{-\lambda_i}$

$e^{-\lambda_i(\beta+1)}$

$f(\lambda_i) = \frac{1}{\Gamma(\alpha)} \lambda_i^{\alpha-1} \beta^{\alpha} e^{-\lambda_i \beta}$

$f(y_i) = \int_0^\infty f(y_i \mid \lambda_i) f(\lambda_i)\, d\lambda_i$

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $E(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i

$\lambda_i = E$(faults in a roll of x_i m.)

β = rate of faults / m.

$\lambda_i = \beta x_i$ not $e^{x_i^T\beta}$

not $e^{\beta_0 + \beta_1 x_i}$

just βx_i

y ~ x - 1, link = "ident"

$\hat{\lambda} = \frac{\bar{y}}{\bar{x}}$ $\widehat{var}(\hat{\lambda}) = \frac{1}{\bar{x}^2}\widehat{var}(\bar{y})$

- $f(y_i; \mu_i, \phi_i) = \exp\{\dfrac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $\mathrm{E}(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^T\beta = \eta_i$ links the n observations together via covariates

$\theta_i \rightarrow \mu_i = E(Y_i) \rightarrow x_i^T\beta$

via $g(\cdot)$

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $E(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^T\beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor

Generalized linear models: theory

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $E(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^T\beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor
- $\text{Var}(y_i \mid x_i) = \phi_i b''(\theta_i) = \phi_i V(\mu_i)$

Generalized linear models: theory

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $E(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^T\beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor
- $\text{Var}(y_i \mid x_i) = \phi_i b''(\theta_i) = \phi_i V(\mu_i)$
- $V(\cdot)$ is the variance function

Examples

- Normal: $f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i^2)\}$

$$= \exp\{\frac{y_i\mu_i - (1/2)\mu_i^2}{\sigma^2} - (1/2)\log\sigma^2 - y_i^2/2\sigma^2 - (1/2)\log\sqrt{(2\pi)}\}$$

$$\phi_i = \sigma^2, \quad \theta_i = \mu_i, \quad b(\mu_i) = \mu_i^2/2, b'(\mu_i) = \mu_i, b''(\mu_i) = 1$$

Examples

- Normal: $f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i^2)\}$

$$= \exp\{\frac{y_i\mu_i - (1/2)\mu_i^2}{\sigma^2} - (1/2)\log\sigma^2 - y_i^2/2\sigma^2 - (1/2)\log\sqrt{(2\pi)}\}$$

$$\phi_i = \sigma^2, \quad \theta_i = \mu_i, \quad b(\mu_i) = \mu_i^2/2, b'(\mu_i) = \mu_i, b''(\mu_i) = 1$$

Examples

- Normal: $f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i^2)\}$

$$= \exp\{\frac{y_i\mu_i - (1/2)\mu_i^2}{\sigma^2} - (1/2)\log\sigma^2 - y_i^2/2\sigma^2 - (1/2)\log\sqrt{(2\pi)}\}$$

$$\phi_i = \sigma^2, \quad \theta_i = \mu_i, \quad b(\mu_i) = \mu_i^2/2, b'(\mu_i) = \mu_i, b''(\mu_i) = 1$$

- Binomial: $f(r_i; p_i) = \binom{m_i}{r_i} p_i^{r_i}(1 - p_i)^{m_i - r_i}; \quad y_i = r_i/m_i$

$$= \exp[m_i y_i \log\{p_i/(1 - p_i)\} + m_i \log(1 - p_i) + \log\binom{m_i}{m_i y_i}]$$

$$\phi_i = 1/m_i, \quad \theta_i = \log\{p_i/(1 - p_i)\}, \quad b(p_i) = -\log(1 - p_i), \quad p_i = E(y_i)$$

Examples

- Normal: $f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i^2)\}$

$$= \exp\{\frac{y_i\mu_i - (1/2)\mu_i^2}{\sigma^2} - (1/2)\log\sigma^2 - y_i^2/2\sigma^2 - (1/2)\log\sqrt{(2\pi)}\}$$

$$\phi_i = \sigma^2, \quad \theta_i = \mu_i, \quad b(\mu_i) = \mu_i^2/2, b'(\mu_i) = \mu_i, b''(\mu_i) = 1$$

- Binomial: $f(r_i; p_i) = \binom{m_i}{r_i} p_i^{r_i}(1 - p_i)^{m_i - r_i}; \quad y_i = r_i/m_i$

$$= \exp[m_i y_i \log\{p_i/(1 - p_i)\} + m_i \log(1 - p_i) + \log\binom{m_i}{m_i y_i}]$$

$$\phi_i = 1/m_i, \quad \theta_i = \log\{p_i/(1 - p_i)\}, \quad b(p_i) = -\log(1 - p_i), \quad p_i = E(y_i)$$

- ELM (p.115) uses $a_i(\phi)$ in place of ϕ_i, later (p.117) $a_i(\phi) = \phi/w_i$;
  SM uses ϕ_i, later (p. 483) $\phi_i = \phi a_i$

Family	Canonical link	Variance function	ϕ_i
Normal	$\eta = \mu$	1	σ^2
Binomial	$\eta = \log\{\mu/(1-\mu)\}$	$\mu(1-\mu)$	$1/m_i$
Poisson	$\eta = \log(\mu)$	μ	1
Gamma	$\eta = 1/\mu$	μ^2	$1/\nu$
Inverse Gaussian	$\eta = 1/\mu^2$	μ^3	ξ

Family	Canonical link	Variance function	ϕ_i
Normal	$\eta = \mu$	1	σ^2
Binomial	$\eta = \log\{\mu/(1-\mu)\}$	$\mu(1-\mu)$	$1/m_i$
Poisson	$\eta = \log(\mu)$	μ	1
Gamma	$\eta = 1/\mu$	μ^2	$1/\nu$
Inverse Gaussian	$\eta = 1/\mu^2$	μ^3	ξ

Gamma: $f(y_i; \mu_i, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} y_i^{\nu-1} \exp(-\frac{\nu}{\mu_i}) y_i$

$$= \exp[-\frac{\nu}{\mu_i} y_i - \nu \log(\frac{1}{\mu_i}) + (\nu - 1)\log(y_i) + \nu\log(\nu) - \log\{\Gamma(\nu)\}]$$

$$= \exp\{\nu(\frac{y_i}{-\mu_i} - \log(\frac{1}{\mu_i}) + (\nu-1)\log(y_i) - \log\Gamma(\nu) + \nu\log(\nu)\}$$

Inference

- $\ell(\beta; y) = \sum\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\}$ $\quad b'(\theta_i) = \mu_i; \quad b''(\theta_i) = V(\mu_i)$
- $g(\mu_i) = g\{b'(\theta_i)\} = \eta_i = x_i^{\mathrm{T}}\beta$
- $\frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \sum \frac{y_i - b'(\theta_i)}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j}$

Inference

- $\ell(\beta; y) = \sum\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\} \qquad b'(\theta_i) = \mu_i; \quad b''(\theta_i) = V(\mu_i)$
- $g(\mu_i) = g\{b'(\theta_i)\} = \eta_i = x_i^{\mathrm{T}}\beta$
- $\frac{\partial\ell(\beta; y)}{\partial\beta_j} = \sum \frac{\partial\ell_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta_j} = \sum \frac{y_i - b'(\theta_i)}{\phi_i}\frac{\partial\theta_i}{\partial\beta_j}$
- $g'(b(\theta_i))b''(\theta_i)\frac{\partial\theta_i}{\partial\beta_j} = x_{ij} = g'(\mu_i)V(\mu_i)$

See Slide 2

Inference

- $\ell(\beta; y) = \sum \{ \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \} \qquad b'(\theta_i) = \mu_i; \quad b''(\theta_i) = V(\mu_i)$
- $g(\mu_i) = g\{b'(\theta_i)\} = \eta_i = x_i^{\mathrm{T}}\beta$
- $\frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \sum \frac{y_i - b'(\theta_i)}{\phi_i} \frac{\partial \theta_i}{\partial \beta_j}$
- $g'(b(\theta_i))b''(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} = x_{ij} = g'(\mu_i)V(\mu_i)$

See Slide 2

- $\frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum \frac{y_i - \mu_i}{\phi_i g'(\mu_i)V(\mu_i)} x_{ij} = \sum \frac{y_i - \mu_i}{a_i\phi g'(\mu_i)V(\mu_i)} x_{ij}$

when $\phi_i = a_i\phi$

Inference

- $\ell(\beta; y) = \sum\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\}$ $\quad b'(\theta_i) = \mu_i; \quad b''(\theta_i) = V(\mu_i)$
- $g(\mu_i) = g\{b'(\theta_i)\} = \eta_i = x_i^{\mathrm{T}}\beta$
- $\frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \sum \frac{y_i - b'(\theta_i)}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j}$
- $g'(b(\theta_i))b''(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} = x_{ij} = g'(\mu_i)V(\mu_i)$ See Slide 2
- $\frac{\partial \ell(\beta; y)}{\partial \beta_j} = \sum \frac{y_i - \mu_i}{\phi_i g'(\mu_i)V(\mu_i)}x_{ij} = \sum \frac{y_i - \mu_i}{a_i\phi g'(\mu_i)V(\mu_i)}x_{ij}$

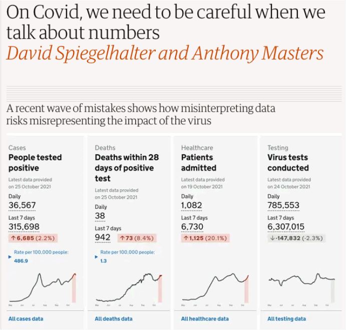
  when $\phi_i = a_i\phi$

- matrix notation:

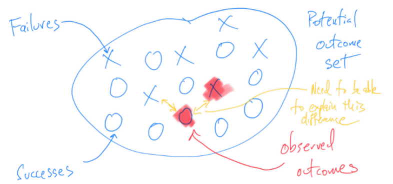
$$\frac{\partial \ell(\beta)}{\partial \beta} = X^{\mathrm{T}}u(\beta), \quad X = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad u = (u_1, \ldots, u_n), \quad u_i = \frac{y_i - \mu_i}{\phi_i g'(\mu_i)V(\mu_i)}$$

In the News

- Guardian, Nov 14 Spiegelhalter & Masters
  "On Covid we need to be careful when we talk about numbers"



- Simply Statistics, Nov 10 Peng
  Thinking about failure in data analysis



Data analysis potential outcome set.

In the News

- Nature Behaviour, Nov Wagenmakers et al.
  Seven steps towards more transparency in statistical practice

nature human behaviour

PERSPECTIVE

https://doi.org/10.1038/s41562-021-01211-8

Check for updates

Seven steps toward more transparency in statistical practice

Eric-Jan Wagenmakers[1]✉, Alexandra Sarafoglou[1], Sil Aarts[2], Casper Albers[3], Johannes Algermissen[4], Štěpán Bahník[5], Noah van Dongen[1], Rink Hoekstra[6], David Moreau[7], Don van Ravenzwaaij[8], Aljaž Sluga[9], Franziska Stanke[10], Jorge Tendeiro[8,11] and Balazs Aczel[12]

We argue that statistical practice in the social and behavioural sciences benefits from transparency, a fair acknowledgement of uncertainty and openness to alternative interpretations. Here, to promote such a practice, we recommend seven concrete statistical procedures: (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discuss their benefits and limitations, and provide guidelines for adoption. Each of the seven procedures finds inspiration in Merton's ethos of science as reflected in the norms of communalism, universalism, disinterestedness and organized scepticism. We believe that these ethical considerations—as well as their statistical consequences—establish common ground among data analysts, despite continuing disagreements about the foundations of statistical inference.