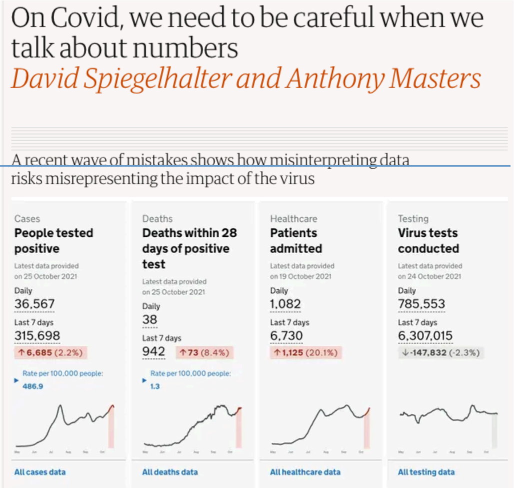
Methods of Applied Statistics I

STA2101H F LEC9101

Week 9

November 17 2021



Today

Start Recording

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

Dr. Sarah Mayes-Tang, U Toronto

- 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

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

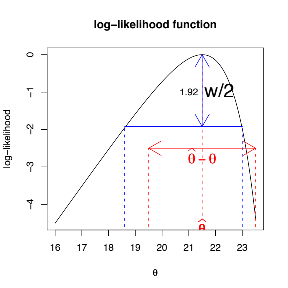
  justification for choices
  model checks, data checks

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

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

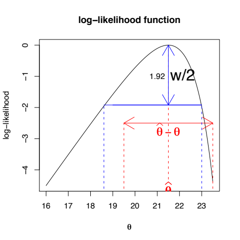


- confidence intervals for β_1

- 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



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 ***
## dibepB        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
```

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
```

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

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

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

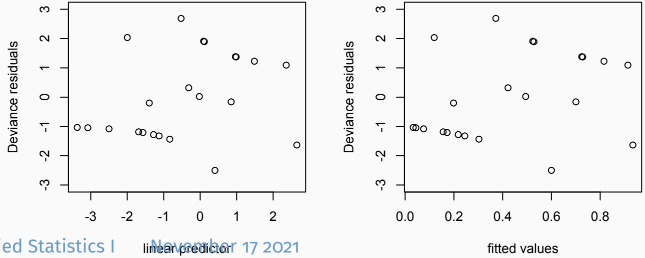
approximately χ^2_{n-q}

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

```
> 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
```



- there are many versions of residuals
- `?residuals.glm`
  `residuals(glm.object, type = c("deviance", "pearson", "working", "response", "partial"),...`
- Binomial deviance $\approx$

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

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

- saturated model $y_i \sim Po(\lambda_i)$, $\quad \tilde{\lambda}_i =$

- compare saturated fit to log-linear fit

- $Deviance = 2\{\ell(\tilde{\lambda}) - \ell(\hat{\lambda})\}$ =

- $Deviance \approx$

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

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

$\dot{\sim} \chi^2_{n-p}$

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 $y_i = 1\{Z_i > 0\}$, $\quad Z_i \sim N(\beta_0 + \beta_1 x_i, 1)$
- $p_i = \mathrm{pr}(y_i = 1 \mid x_i) = 1 - \Phi\{-(\beta_0 + \beta_1 x_i)\} = \Phi(\beta_0 + \beta_1 x_i)$
- $g(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i$ 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 when y_i is number of events in time T, $E(y_i) = T\lambda_i$

… 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 when 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

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i(1 - p_i)$
- 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? over-dispersion parameter
- usually use $X^2/(n - p)$, where

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

Quasibinomial

overdisp.Rmd; overdisp.html

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

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

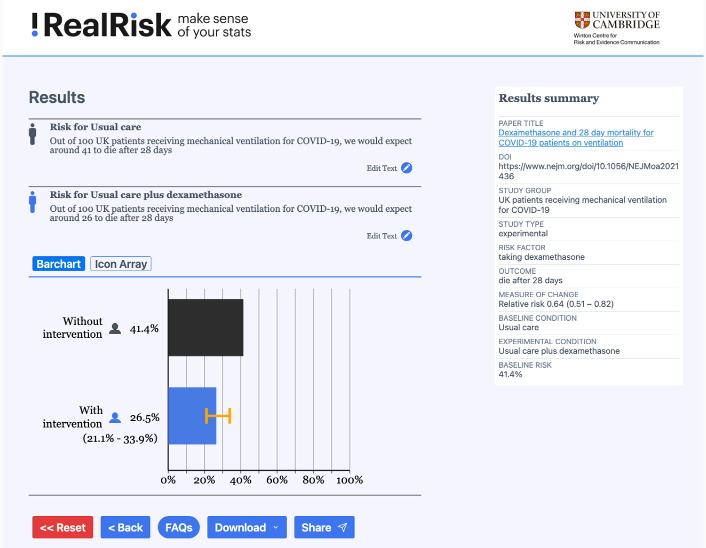
- no difference between groups $\iff$ odds-ratio $\equiv 1$

... Measures of risk

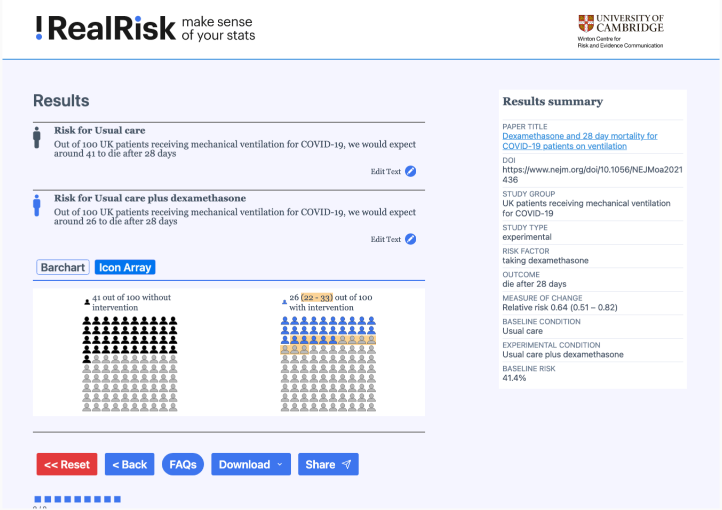
- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if p_1 and p_0 are both small, ($y = 1$ is rare), then

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

- 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

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

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

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")
```

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

- $f(y_i;\mu_i,\phi_i) = \exp\{\dfrac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i;\phi_i)\}$
- $\mathrm{E}(y_i \mid x_i) = b'(\theta_i) = \mu_i$ 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
- $g(\cdot)$ is the link function; η_i is the linear predictor
- $\mathrm{Var}(y_i \mid x_i) = \phi_i b''(\theta_i) = \phi_i V(\mu_i)$
- $V(\cdot)$ is the variance function

Examples

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

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

- Binomial: $f(r_i;p_i) = \binom{m_i}{r_i} p_i^{r_i}(1-p_i)^{m_i-r_i}; \quad y_i = r_i/m_i$
  $= \exp[m_iy_i\log\{p_i/(1-p_i)\} + m_i\log(1-p_i) + \log\binom{m_i}{m_iy_i}]$

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

- 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) $\phi_i = \phi a_i$

… Examples

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

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

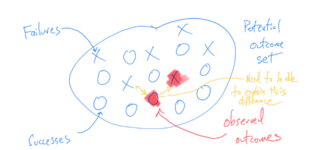
- matrix notation:

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

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.