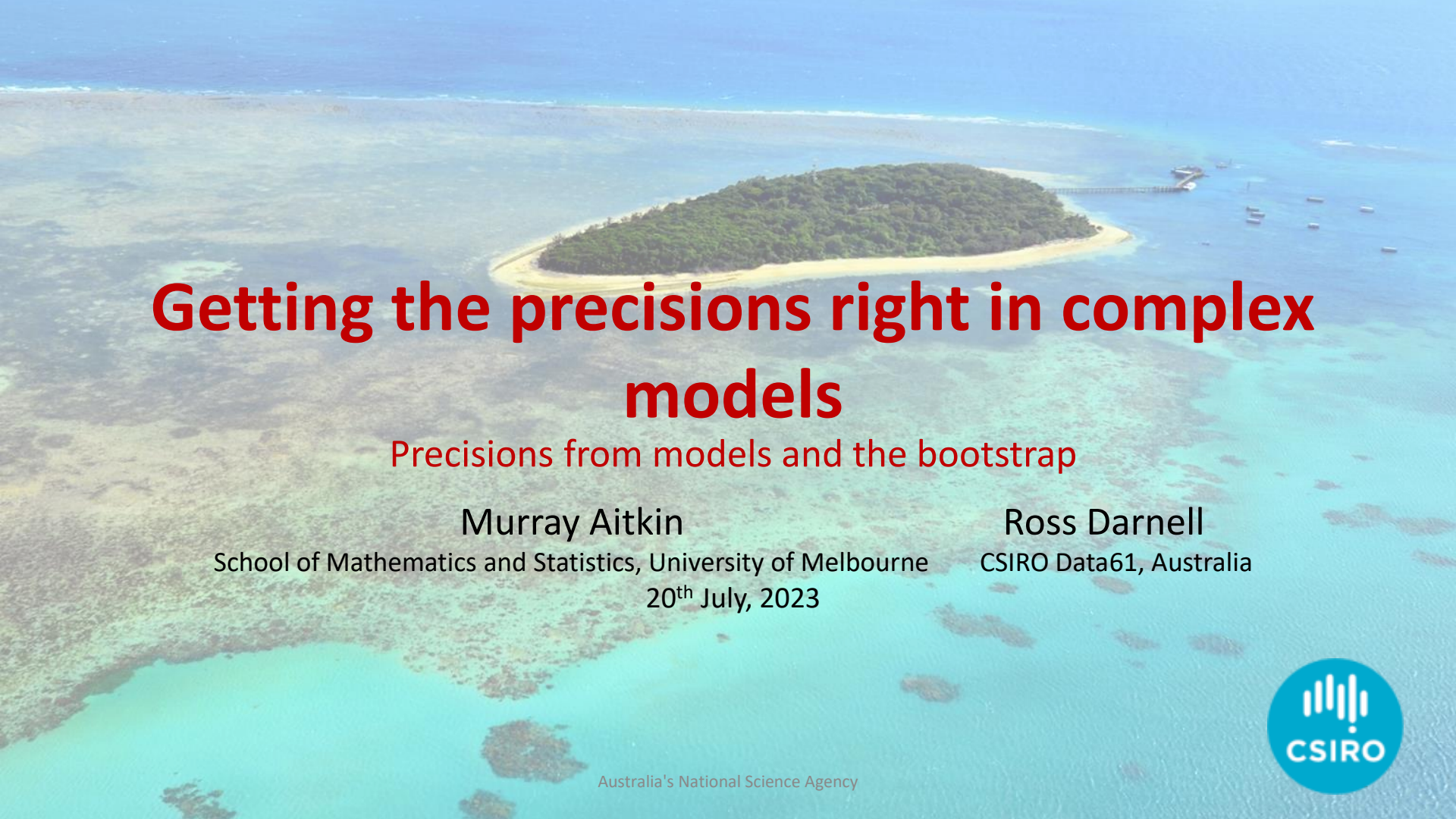


# Getting the precisions right in complex models

## Precisions from models and the bootstrap

Murray Aitkin
School of Mathematics and Statistics, University of Melbourne

Ross Darnell
CSIRO Data61, Australia

20th July, 2023

# The need for precision

A quote from (Cressie 2021) expresses it well:

> In any applied statistics project, the statistician worries about uncertainty and quantifies it by modelling data as realisations generated from a probability space. Another approach to uncertainty quantification is to find similar data sets, and then use the variability of results between these data sets to capture the uncertainty

- The first approach is through probability model-based likelihood analysis, due to R.A. Fisher.
- The second approach is through “model-free” extensions of least squares and repeated sampling, due to J. Neyman and B. Efron.

# A road map slide where we're heading

- Inferential tools for probability and machine learning analyses
- Use of Bootstrap
- The Bayesian bootstrap
- Modelling — splines and polynomials
- Precision modelling using the double GLM
- Complex example - Chlorophyll-a levels in the Great Barrier Reef lagoon.

# Inferential tools in probability model-based analysis

- Standard or generalised regression methods use some version of a model for data $y$ related to covariates $\mathbf{x}$, typically a polynomial in the covariates:

$$y|x_i \sim f(g(\beta' x_i), \phi)$$

for some density or mass function $f$, "link" function $g$ and precision constant $\phi$.

- If the random variation through $f$ also varies with $x$ then $\phi$ has to be modelled as well, typically through another polynomial model $h\big(k(\gamma' x_i)\big)$.

- Given *f, g, h* and *k*, inference is through the likelihood function.

- The approximate validity of the probability model $f$ has to be assessed, frequently from the residuals

$$e_i = y_i - f\big(g\,(\hat{\beta}' x_i), \hat{\phi}\big)\,.$$

# Inferential tools in machine learning

- Regression methods are based on extension of low-degree least squares: higher-level variability is represented by splines – small non-linear terms at break-points where slopes or curvatures change suddenly.
- Precision of the spline or other least squares procedure is assessed through the bootstrap resampling of the data.
- The bootstrap has been recently shown (Aitkin 2022, 13.7 pp. 178-180) to be ineffective in assessing precision.
- We need a different model-free procedure for precision: the Bayesian bootstrap.
- An example shows why….

# A simple example, of incomes

Family incomes in units of 1000 USD, listed in increasing order, of $n = 40$ families sampled randomly from a population of size $648$. The sample mean $\bar{y}$ and variance $s^2$ are **67.1** and **500.9** .

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 26 | 35 | 38 | 39 | 42 | 46 | 47 | 47 | 47 | 52 |
| 53 | 55 | 55 | 56 | 58 | 60 | 60 | 60 | 60 | 60 |
| 65 | 65 | 67 | 67 | 69 | 70 | 71 | 72 | 75 | 77 |
| 80 | 81 | 85 | 93 | 96 | 104 | 104 | 107 | 119 | 120 |

What is the precision of the sample mean income? Is the interval based on the Gaussian model reliable?

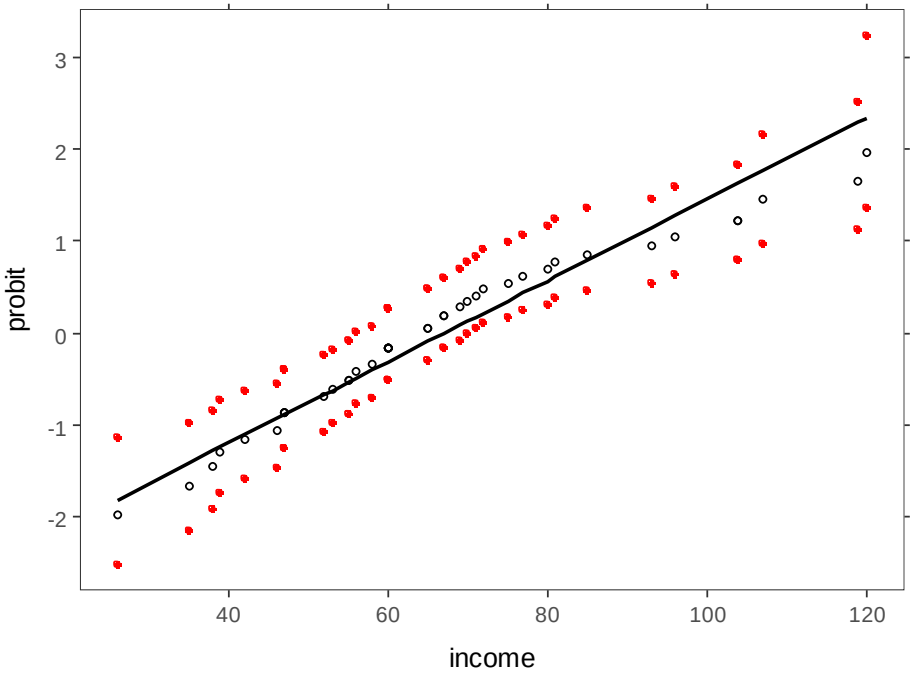
| | | 95% CI | |
|---|---|---|---|
| Type | Estimate | Lower | Upper |
| Population | $\mu$ | | |
| Gaussian model | 67.1 | 60.1 | 74.0 |

# Sample probit scale cdf graph

The figure shows the

- data (circles),
- fitted ML Gaussian distribution (line) and
- 95% credible region for the true cdf (red segments).

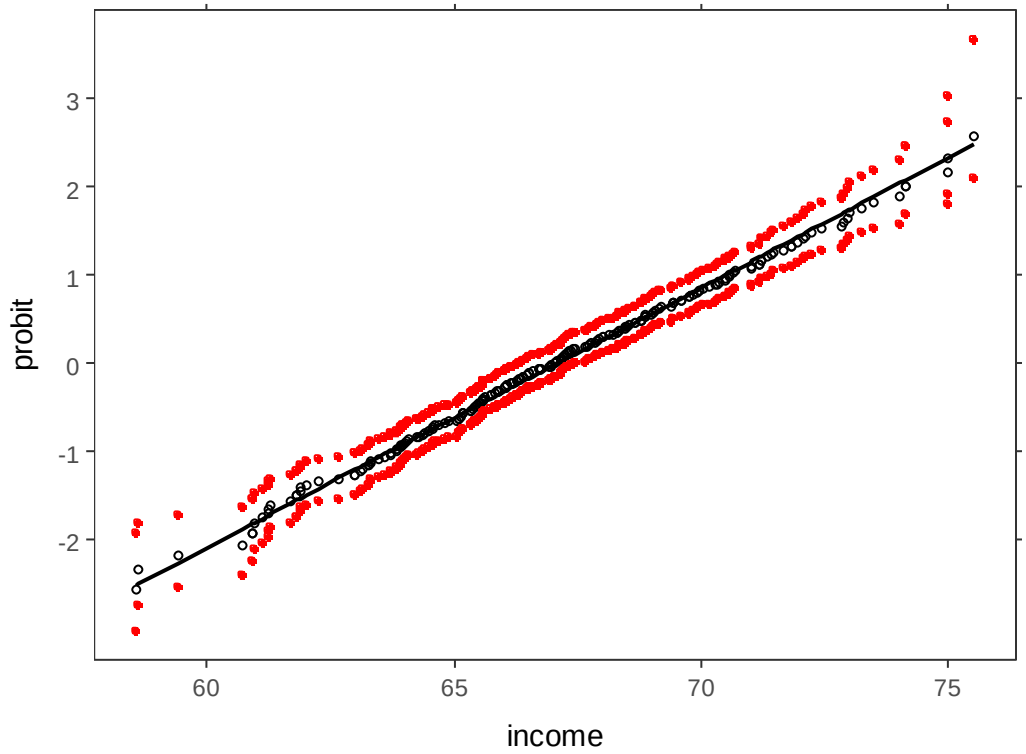
The best-fitting Gaussian cdf lies fully within the 95% credible region, despite the evident curvature.



# Assessing variability through the bootstrap

- The bootstrap is very widely used for assessing the precision of a sample estimate like the sample mean, of a population quantity without a probability model for the data.
- The given sample, called the pseudo-population here, is resampled with replacement a large number $K$ of times.
- For each of the $K$ bootstrap samples we compute the sample mean estimate $\bar{y}_k$ and
- We construct a 95% interval for $\mu$ from the 2.5% and 97.5% quantiles of the ordered sample estimates, or from the variance of the sample means.
- The process is computationally very fast and straightforward.
- We illustrate with $K = 200$.

# Bootstrap means probit scale cdf graph



# Bootstrap analysis

- The asymptotic 95% confidence interval for the pseudo-population mean based on the variance of the bootstrap means is [60.4, 73.7].
- The 95% central confidence interval for the pseudo-population mean based on the 2.5% and 97.5% quantiles of the bootstrap means is [60.9, 74.0].
- These are both similar to, but shorter than, the Gaussian interval of [60.1, 74.0]. The bootstrap sample means are varying randomly, but are varying around the pseudo-population – **sample** – mean $\tilde{\mu}$ of 67.1, not the unknown population mean $\mu$.

| | | 95% CI | |
|---|---|---|---|
| Type | Estimate | Lower | Upper |
| Population | $\mu$ | | |
| Gaussian model | 67.1 | 60.1 | 74.0 |
| Bootstrap variance | 67.1 | 60.4 | 73.7 |
| Bootstrap quantiles | 67.0 | 60.9 | 74.0 |

How do we know the coverage of this interval for μ?
(Aitkin 2022 , §13.7, pp. 178-180) showed that the bootstrap samples are ancillary for the population mean: they convey no usable information about it.

# A fundamental error in the bootstrap argument

- There is no necessary relation between the bootstrap confidence interval and the population mean: it depends on the unknown closeness of the sample mean to the population mean.
- The claim that the bootstrap confidence interval is a confidence interval for the population mean $\mu$ is unfounded: we cannot give a relevant probability statement about the confidence interval coverage.
- Formal likelihood theory shows that the bootstrap samples are irrelevant for inference about $\mu$.(Aitkin 2022 , §13.7, pp. 178-180)
- The uncertainty in the $y$ and hence $\mu$ is recognised in the Bayesian analysis through the multinomial likelihood and conjugate Dirichlet prior and posterior.

# Multinomial population and sample

- We write the $N$ unobserved population values of income as $Y_1^*, \ldots, Y_N^*$..
- We conceptually tabulate these $N$ income values into the $D$ ordered distinct unobserved values $Y_1 < Y_2 < \cdots < Y_I < \cdots < Y_D$,
- with corresponding population counts $N_1, N_2, \ldots, N_I, \ldots, N_D$,
- and population proportions $P_1, P_2, \ldots, P_D$ with $P_I = N_I/N$.
- The population income $Y$ has a multinomial distribution $M(N, P_1, \ldots, P_D)$, with population mean $\mu = \sum_{I=1}^{D} P_I\, Y_I$.
- We tabulate the sample of size $n$ in the same way, into the $d$ ordered distinct observed values $y_1 < \cdots < y_i < \cdots < y_d$,
- with corresponding sample counts $n_1, n_2, \ldots, n_d$,
- sample proportions $p_1, p_2, \ldots, p_d$, and sample mean $\bar{y} = \sum_{i=1}^{d} p_i\, y_i$ .

# The Bayesian solution

- To model-free people, ancillarity may appear irrelevant:
  - They may dismiss the multinomial model and likelihood because the bootstrap is model-free
  - and does not need a model or a likelihood.

- To model-based people, the bootstrap has a model underlying it which should be recognised and incorporated in the analysis.

- How do we proceed with the sample data and just the multinomial model?

- The solution to this difficulty is to be fully Bayesian with the original sample.
- This was done with the Bayesian bootstrap by Don Rubin in 1981.

# The Bayesian bootstrap

To the multinomial likelihood, we add the conjugate Dirichlet prior with indices $a_i$ :

$$Pr[\{n_i\}|\{p_i\}] = \frac{n!}{\prod_{i=1}^{d} n_i\,!} \prod_{i=1}^{d} p^{n_i}$$

$$\pi(\{p_i\}|\{a_i\}) = \frac{a!}{\prod_{i=1}^{d} a_i\,!} \prod_{i=1}^{d} p_i^{a_i-1}$$

where the $a_i$ are the prior parameters and $a = \sum_{i=1}^{n} a_i$. The posterior distribution of the $p_i$ is then Dirichlet with indices $n_i + a_i$:

$$\pi(\{p_i\}|\{n_i\},\{a_i\}) = \frac{(n+a)!}{\prod_{i=1}^{d}(n_i+a_i)\,!} \prod_{i=1}^{d} p_i^{n_i+a_i-1}$$

# Posterior computation

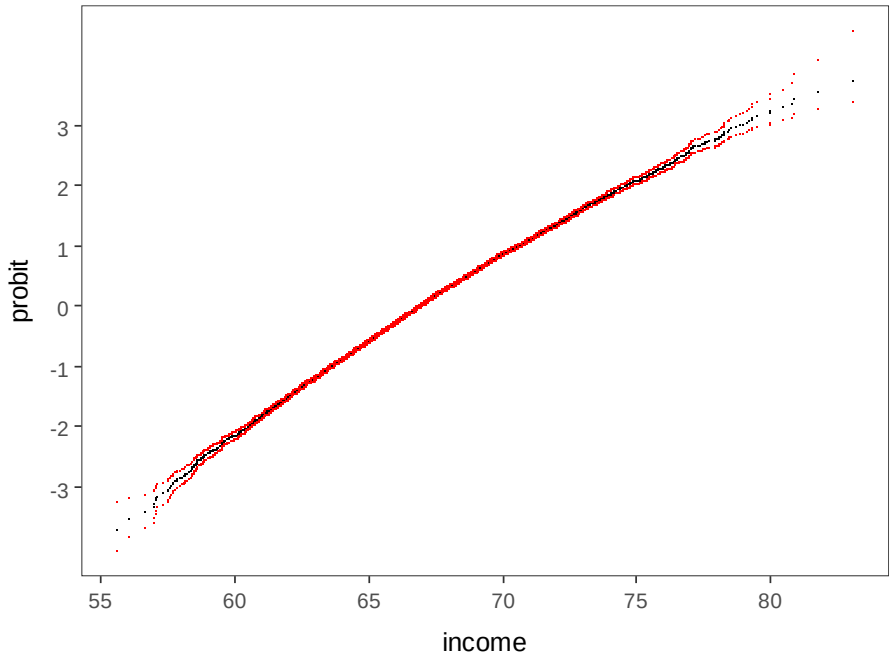
- The reference non-informative prior has prior parameters $a_i = 0$ for all $i$.
- We make $M$ random draws $p_i^{[m]}$ of the $p_i$ from their posterior distribution,
- and substitute them into the mean to give $[M]$ random draws of the mean:

$$\mu^{[m]} = \sum_{i=1}^{d} p_i^{[m]} \; y_i$$

- The figure shows $M = 10{,}000$ random draws, cumulated to give the cdf of the posterior (marginal) distribution of the population mean.
- The 2.5% point is 60.60, the median is 66.91 and the 97.5% point is 74.42.
- $M = 10{,}000$ draws give a very accurate approximation to the true cdf – no smoothing – shown in small dot symbols to prevent crowding.

# Posterior probit of the income population mean 1/2

- The right tail is longer than the left.
- The probit graph for the posterior mean is slightly curved: it is not quite Gaussian.



# Posterior probit of the income population mean 2/2

The posterior median is almost identical to the sample mean.

The 95% central credible interval for the population mean is similar to the Gaussian 95% interval,

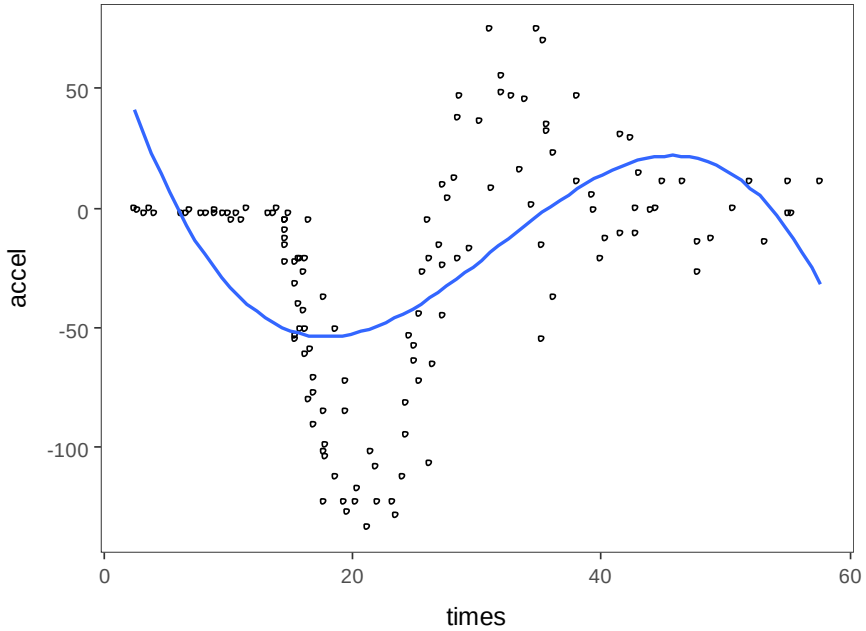
but is shifted slightly to the right, reflecting the curvature of the sample cdf.

| | | 95% CI | |
|---|---|---|---|
| Type | Estimate | Lower | Upper |
| Population | $\mu$ | | |
| Gaussian model | 67.1 | 60.1 | 74.0 |
| Bootstrap variance | 67.1 | 60.4 | 73.7 |
| Bootstrap quantiles | 67.0 | 60.9 | 74.0 |
| Posterior probit | 66.9 | 60.6 | 74.0 |

# Bootstrap extension to regression

- A common approach to “model-free” regression analysis is to obtain the least squares or generalised least squares estimate $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and the fitted function $\widetilde{\boldsymbol{\beta}}'\mathbf{x}$.
- Without a probability model for $\mathbf{y}$, the precision of the fitted function is obtained by bagging the fitted function – bootstrapping it,
- by resampling with replacement a large number of times the observed response and covariate data $(y_i, \mathbf{x}_i)$,
- and redoing the LS or GLS procedure with each resample.
- The variability among the bootstrap resamples of the fitted function is used to provide a (pointwise) confidence region for the true function.
- The region has no valid confidence coverage, as for the bootstrap.

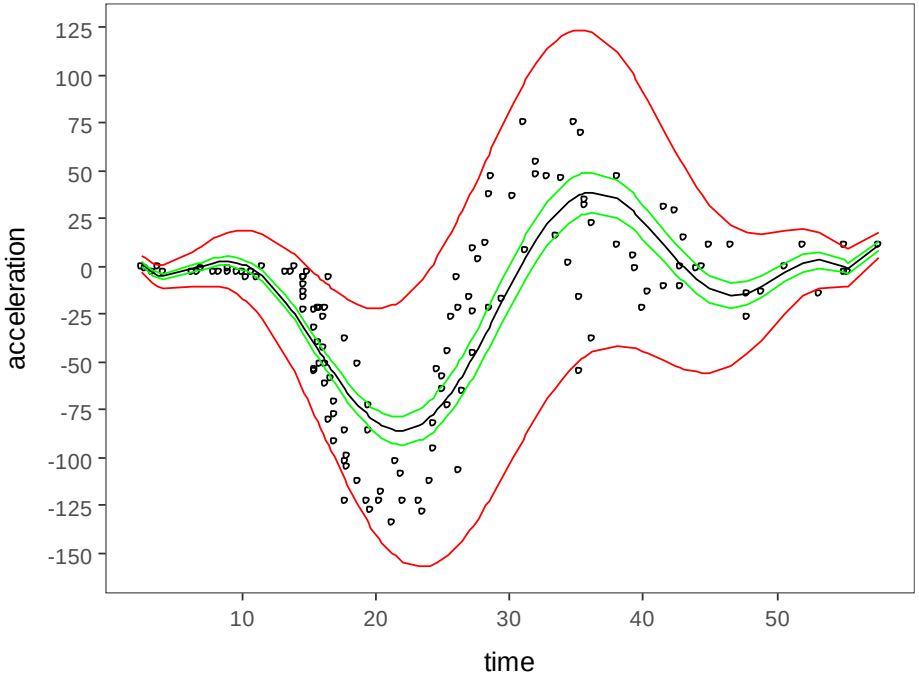
# Machine learning analysis
# The motorcycle data




Machine learners are reluctant to use polynomials of higher order than 2. Splines are preferred instead:

- [Polynomials] are not very flexible in approximating functions with local features such as functions with varying degrees of smoothness at different locations.
- This leads to the introduction of spline functions that allow for more flexibility in function approximation (Fan et al. 2020, 31–32).
- The plot shows the cubic polynomial fit to motorcycle data (Silverman 1985). *Clearly, it does not fit the data very well. Increasing the order of the polynomial fits will help reduce the bias issue, but will not solve the lack of fit issue… This is because the underlying function cannot be economically approximated by a polynomial function.*

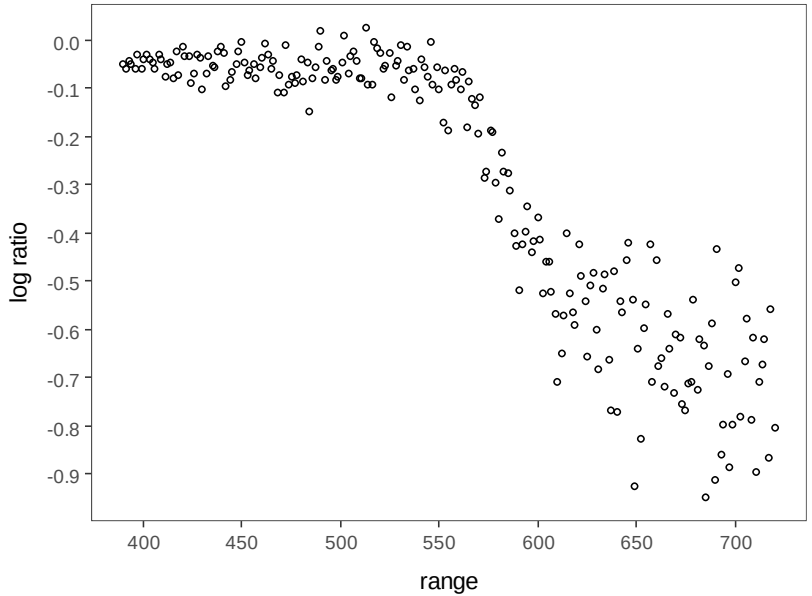
# Polynomial fits to motorcycle data using DGLM




$$E[\text{accel}] = t + t^2 + t^3 + t^5 + t^6 + t^8 + t^{10}$$
$$log(V[\text{accel}]) = t + t^2$$

# Complex data

**Lidar Data from Ruppert, Wand, and Carroll (2003)**




LIDAR is an abbreviation for LIght Detection And Ranging.

- The figure gives the plot of the log of the ratio for the time reflected light to return from two laser sources against the range – the distance the light has travelled before being reflected.
- **Note:** The mean and the variance both fluctuate smoothly but not monotonically as range increases.

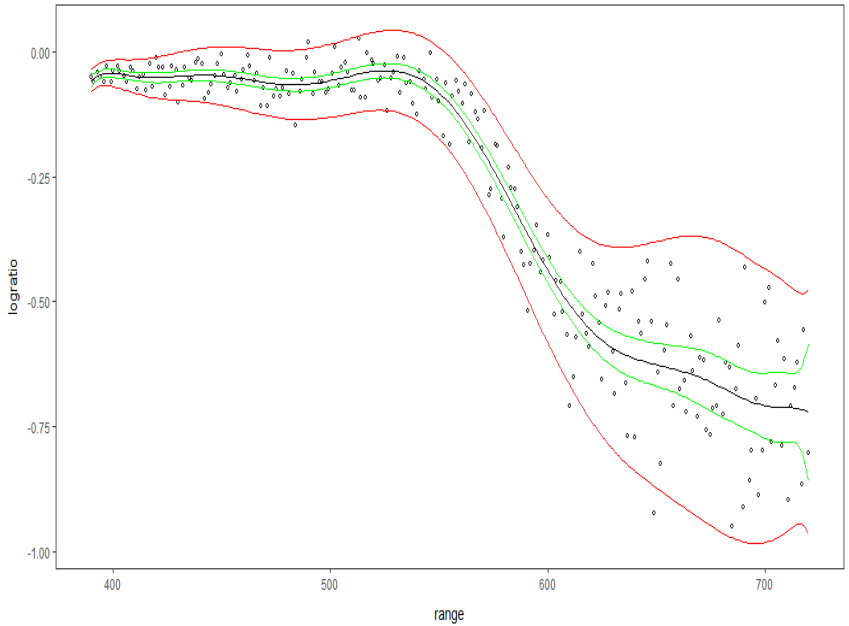
- We need to model both mean and variance independently.

# The double GLM for modelling Gaussian variability

Due to Aitkin (1987) and Smyth (1989). How does it work?

- We model the Gaussian log variance with a polynomial regression model, and
- the mean with a polynomial regression model.
- We alternate the two GLMs to convergence, for either ML or Bayesian posteriors:
  - With a constant variance model, find the well-supported (by LRT) polynomial mean model.
  - With the well-supported mean model, find the well-supported polynomial variance model.
  - With the well-supported polynomial variance model, re-check the well-supported mean model.
  - Repeat if necessary.

# Double GLM analysis



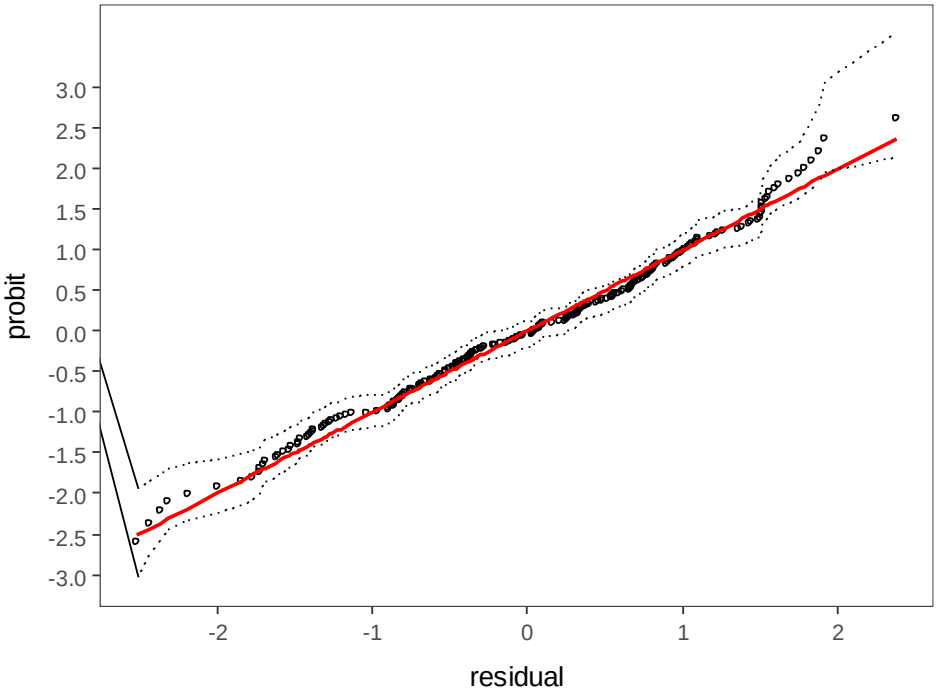

- The DGLM fit requires a tenth-degree polynomial for the mean, and a fourth-degree polynomial for the log variance.
- The fitted model is shown with 95% precision intervals (green curves), and 95% variability bounds (red curves).
- The bounds exclude 7 observations, 3.2% of the data: the Gaussian fit seems to be good.

$$E[\text{logratio}] = t + t^2 + t^3 + t^4 + t^5 + t^6 + t^7 + t^8 + t^9 + t^{10}$$

$$log(V[\text{logratio}]) = t + t^2 + t^3 + t^4$$

# Residual plot from LIDAR data DGLM model

- The probit plot of the standardised residuals shows a good fit to the Gaussian distribution: the cdf in red is entirely within the 95% credible region for the true cdf.



# Comment

Ruppert, Wand, and Carroll (2003) discussed high-degree polynomials for the LIDAR data in their §2.7 but dismissed them:

> The degree-10 fit goes through the data reasonably well but has wiggles that are representative not of any features in the data but rather of high-degree polynomials generally. .. We might use high-degree polynomial models if nothing better were available, but fortunately much better fitting methods are available. (p. 48)

They gave a very detailed discussion of the likelihood ratio test in their §4.8 but did not use it for this example, which would have shown the necessity of the 10-th degree mean function for this data set.

# Spline fit

- Ruppert, Wand, and Carroll (2003) gave in their Chapter 3 a very detailed development of the spline analysis and the effect of the number of knots and the penalty constant (for the penalised least squares analysis) on the appearance of the fitted model, illustrated by many graphs of the LIDAR data.
- Knots locate by eyeballing small changes in slope or gradient of the low-order polynomial which require additional basis functions to model the changes.
- They showed the substantial effect that the penalty constant may have on the degree of “wiggle” of the fitted model, with values of 0, 10, 30 and 1000. Their final fitted model was based on 24 knots and a penalty parameter of 30.
- The clear heterogeneity in the variance was analysed by log variance function spline fitting analogous to the spline mean fitting. This is a parallel to the double GLM fitting of Aitkin.

# Conclusions

- Regression structures of higher polynomial degree than two can be adequately represented by the appropriate degree of polynomial through
  - the likelihood ratio test and
  - the Gram-Schmidt orthogonalisation procedure,
  - which is available in all numerical analysis and many statistical packages.

The double GLM can use the orthogonalisation in both mean and log variance regressions; typically the variance needs a much lower degree model than the mean.

- The spline analysis, while having a strong basis in statistical theory, remains dependent on the eyeball choice of the number of knots and the penalty constant; its precision cannot be formally expressed.

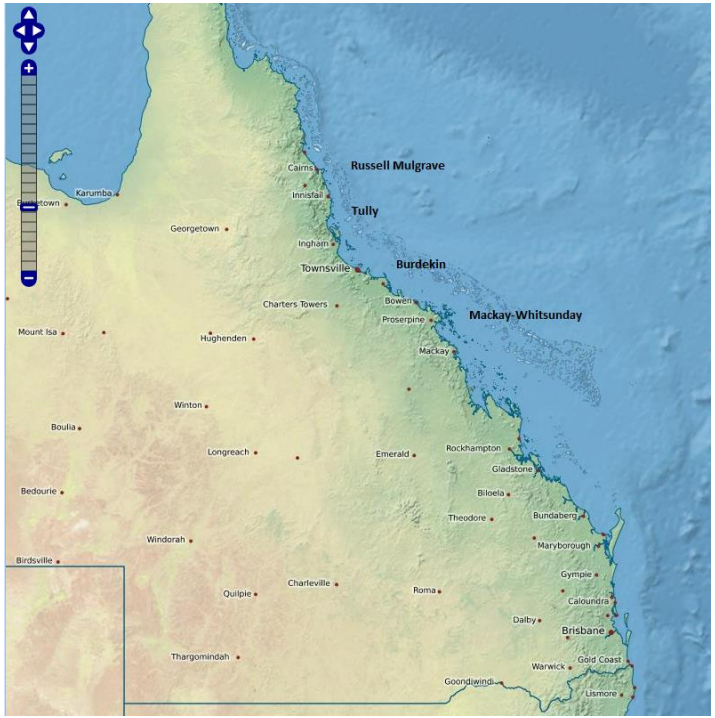
# GBR water quality data (Lloyd-Jones et al. 2022)

The second example is the variability over time, season and region in chlorophyll-a concentration in water samples from four regions of the Great Barrier Reef, to assess any trend or variability in pollutant levels. The published report is in Lloyd-Jones et al. (2022).

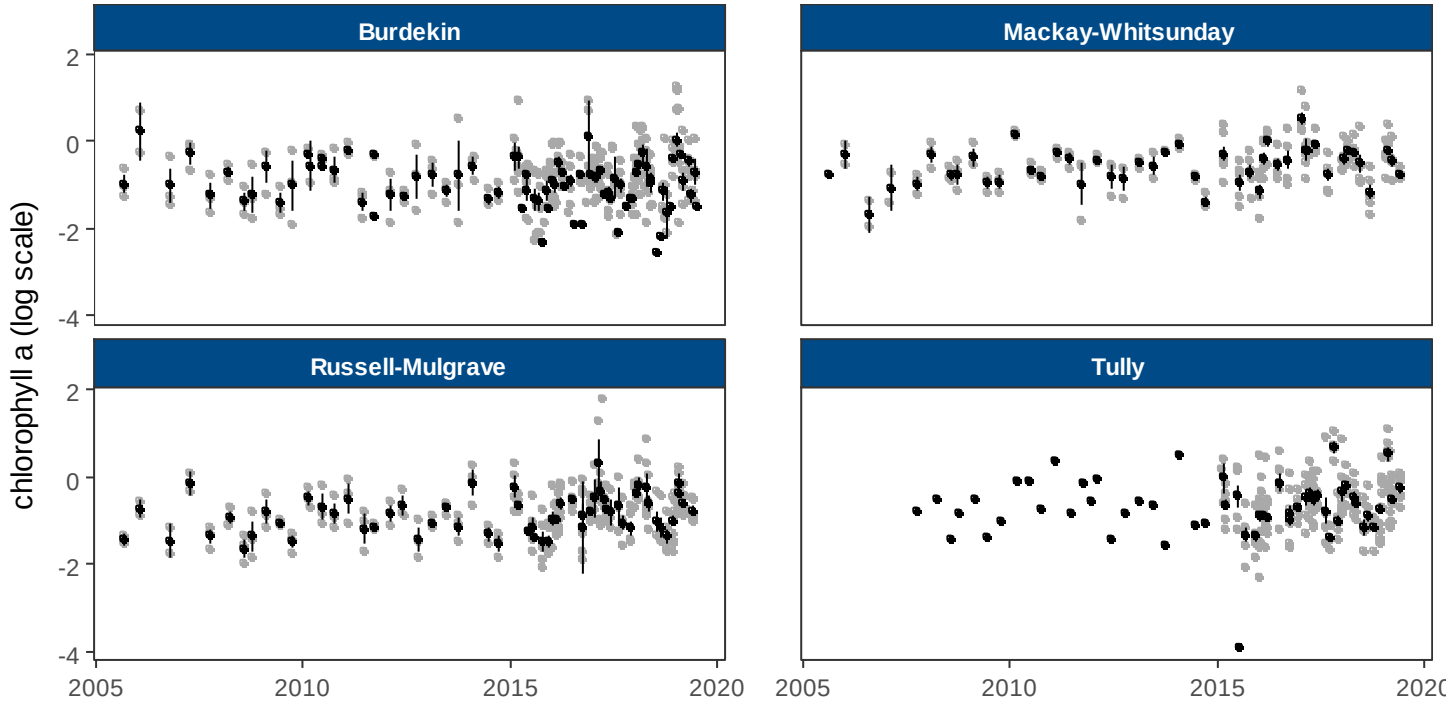
The log of the concentration of chlorophyll-a is the response variable, related to the design and other covariates:

- regions 1–4
- years 2005–2019
- months 1–12
- The sample sizes for each region were small in the early period, and the data set is too small (970) to allow for interactions.
- The among-months variation is modelled by a sine function, allowing among region variations, but
- consistent across years since they are all sites in the same climate and weather area.
- The trend and variability across years are modelled by a double GLM

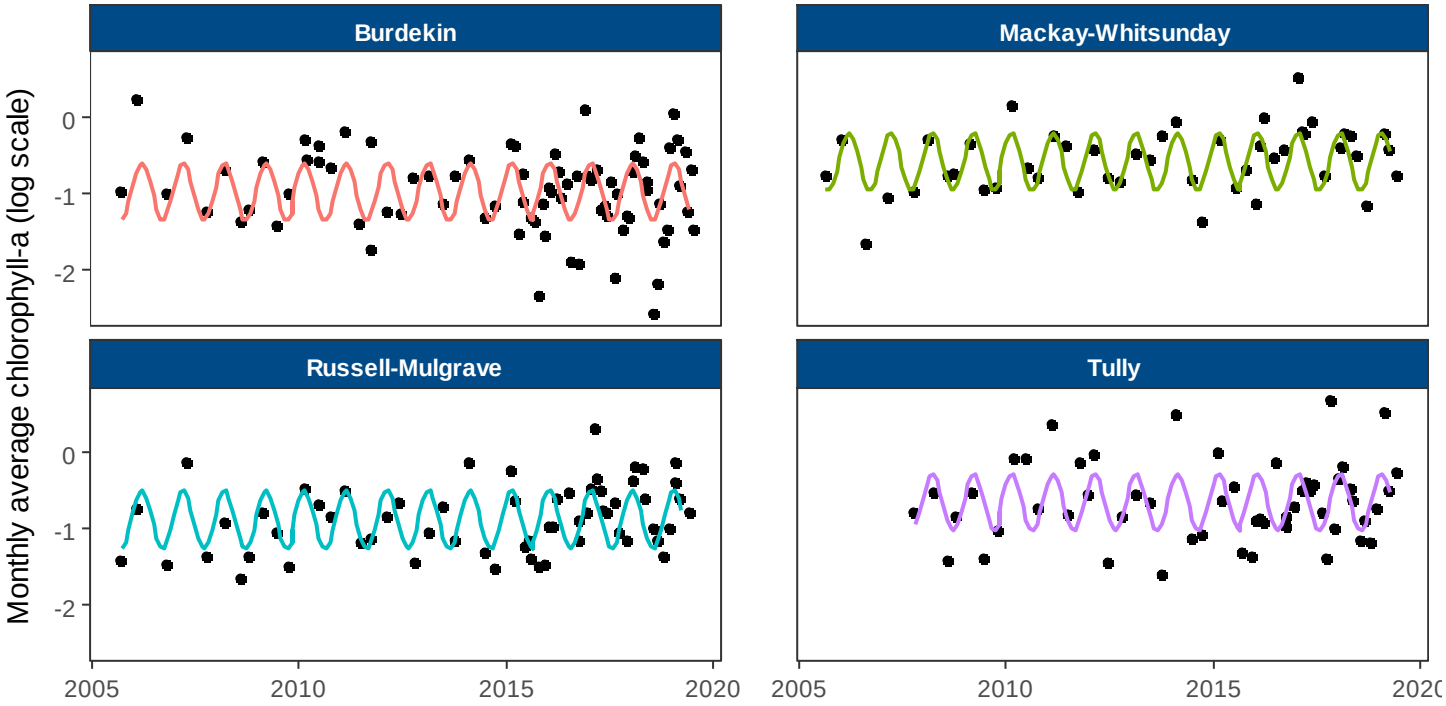
# Location of Regions



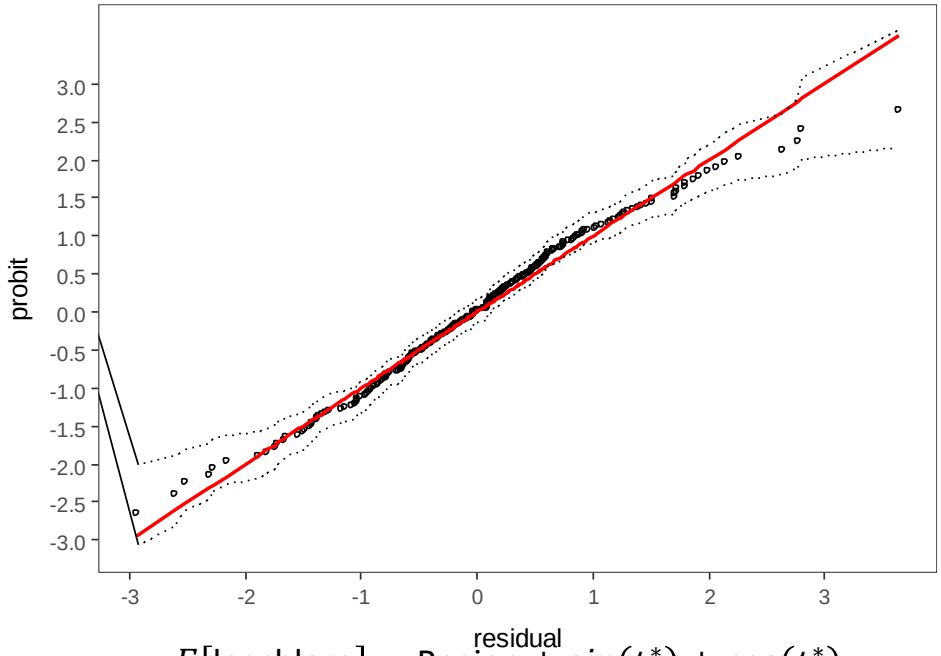
# GBR Chlorophyll-a



# GBR Chlorophyll-a (fitted line using double GLM)



# Probit plot of DGLM residuals




$$E[\text{logchloro}] = \text{Region} + sin(t^*) + cos(t^*)$$
$$log(V[\text{logchloro}]) = \text{Region}$$

- NOTE: When the Gaussian assumption fails the bootstrap or Bayesian bootstrap can be used to estimate precision.

# References


Aitkin, Murray. 1987. “Modelling Variance Heterogeneity in Normal Regression Using GLIM.” *Applied Statistics* 36 (3): 332. https://doi.org/10.2307/2347792.

———. 2022. *Introduction to Statistical Modelling and Inference*. CRC Press LLC.

Cressie, Noel. 2021. “A Few Statistical Principles for Data Science.” *Australian & New Zealand Journal of Statistics* 63 (1): 182–200. https://doi.org/https://doi.org/10.1111/anzs.12324.

Fan, Jianqing, Runze Li, Cun-Hui Zhang, and Hui Zou. 2020. *Statistical Foundation of Data Science*. Taylor & Francis Group.

Lloyd-Jones, L, P Kuhnert, E Lawrence, S Lewis, J Waterhouse, R Gruber, and F Kroon. 2022. “Sampling Re-Design Increases Power to Detect Change in the Great Barrier Reef’s Inshore Water Quality.” Edited by Bijeesh Kozhikkodan Veettil. *PLOS ONE*. Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0271930.

Ruppert, David, M. P. Wand, and R. J. Carroll. 2003. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://doi.org/10.1017/CBO9780511755453.

Silverman, B. 1985. “Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting.” *Journal of the Royal Statistical Society: Series B (Methodological)* 47 (1): 1–21. https://doi.org/10.1111/j.2517-6161.1985.tb01327.x.

Smyth, Gordon K. 1989. “Generalized Linear Models with Varying Dispersion.” *Journal of the Royal Statistical Society. Series B (Methodological)* 51 (1): 47–60. http://www.jstor.org/stable/2345840.

Thank you

- ross.darnell@csiro.au
- murray.aitkin@unimelb.edu.au