Count on someone who can count!

Statistical Society Australia - Lancaster Lecture

21 October 2020

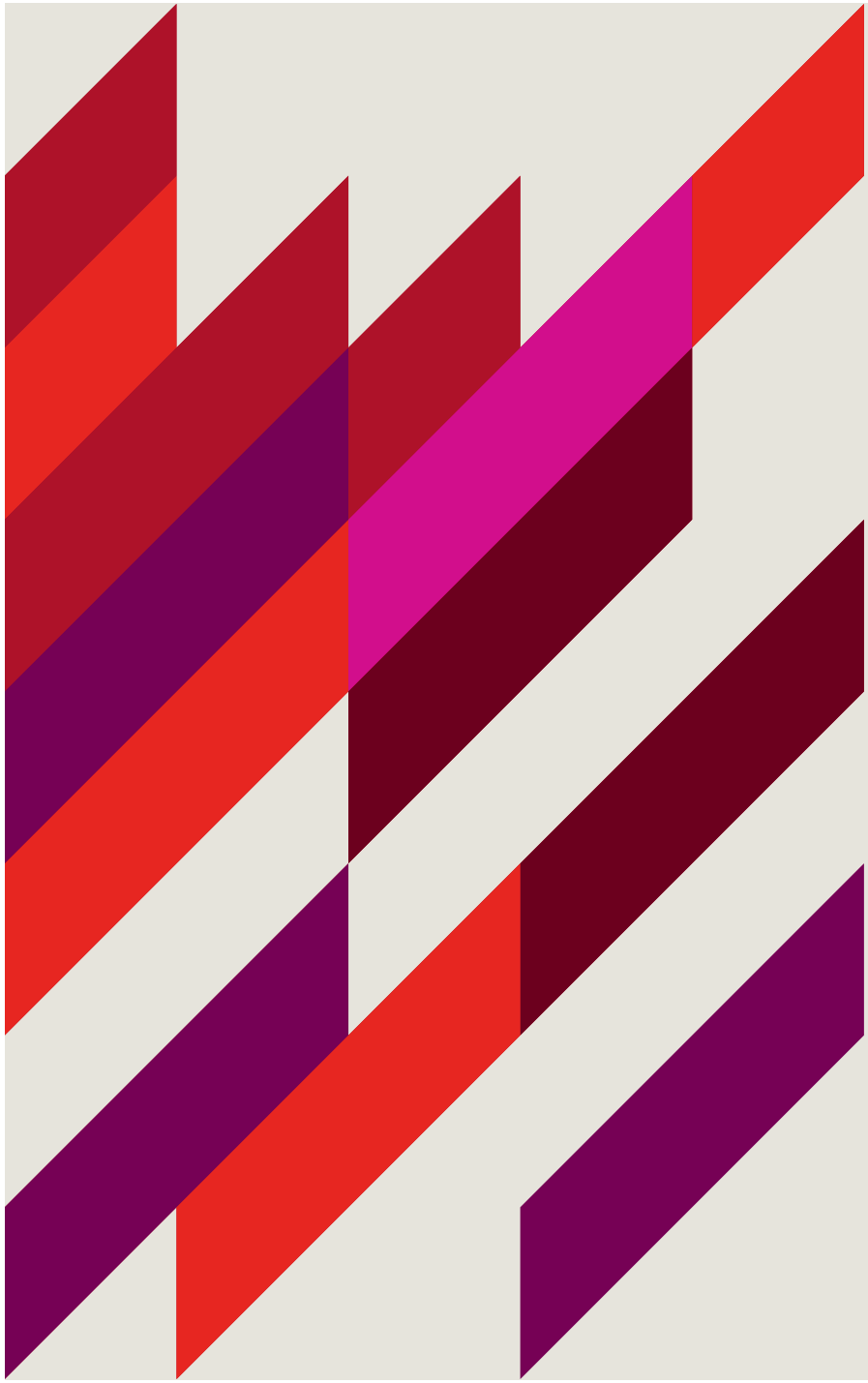
Thomas Fung, Macquarie University

Happy World Statistics Day 2020!

Statistical Society of Australia

Road Map

- Professor H.O. Lancaster
- Conway-Maxwell-Poisson Generalized Linear Models (`mpcmp`)
  - Conway-Maxwell-Poisson Associated Kernel (`compak`) Smoother
- Interpretable Zero-inflated Poisson Regression (`izipr`)

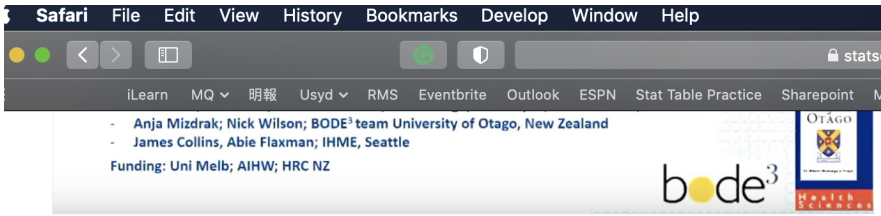


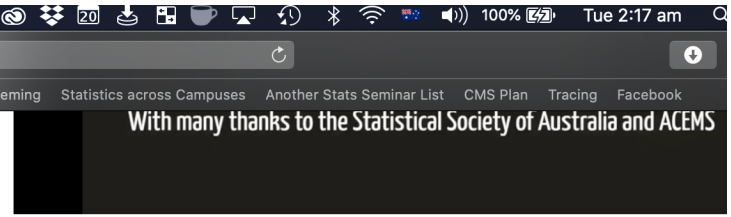
MACQUARIE University
SYDNEY·AUSTRALIA

Professor H.O. Lancaster

Professor Henry Oliver Lancaster

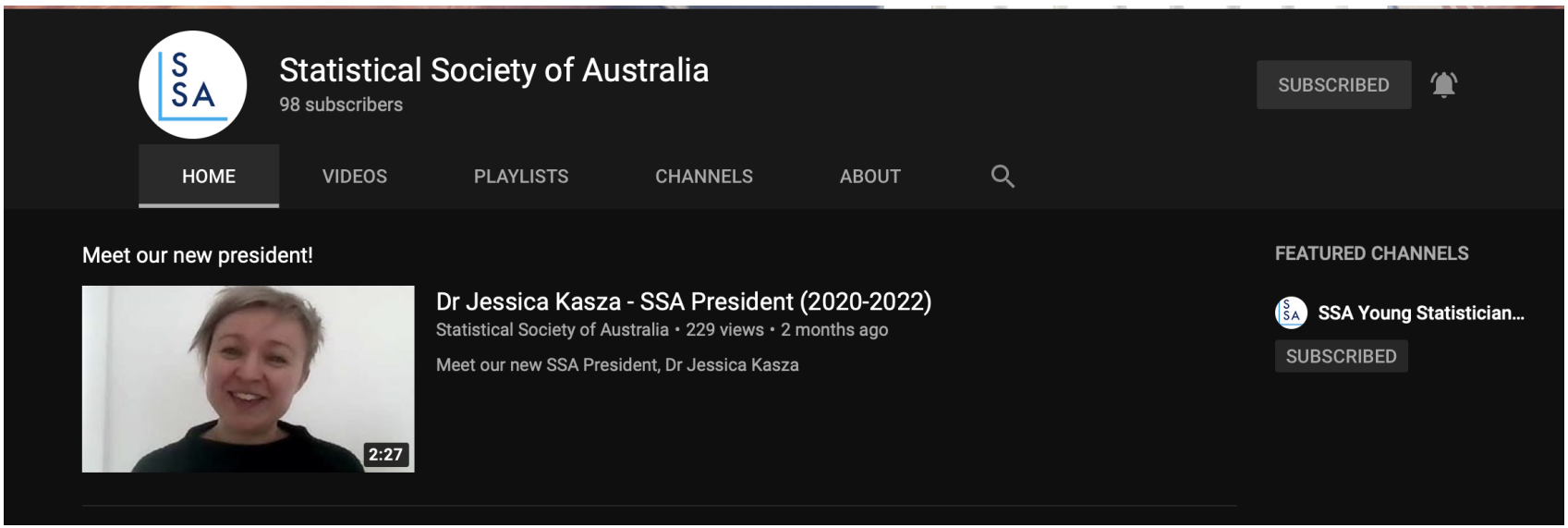
- H.O. Lancaster is a role model for many modern statisticians.
- The "H.O. Lancaster Lecture", commenced in 1979 at a meeting of the SSA NSW Branch.
- Distinguished Professor Kerrie Mengersen gave an excellent Lancaster Lecture last year, reflecting on Prof. Lancaster's extensive research work which culminates in his book *Expectations of Life*.
  - Videos of past seminars are available on the Statistical Society website.
- I would like to use this opportunity to reflect on Prof. Lancaster's contribution to the wider statistical community.





Lancaster lecture: "Expectations of Life": from past to present

The present and future of tidy data


Statistical Society of Australia
98 subscribers
SUBSCRIBED
HOME
VIDEOS
PLAYLISTS
CHANNELS
ABOUT
Meet our new president!
2:27
Dr Jessica Kasza - SSA President (2020-2022)
Statistical Society of Australia • 229 views • 2 months ago
Meet our new SSA President, Dr Jessica Kasza
FEATURED CHANNELS
SSA Young Statistician...
SUBSCRIBED

Professor Henry Oliver Lancaster

- H.O. Lancaster is a role model for many modern statisticians.
- The "H.O. Lancaster Lecture", commenced in 1979 at a meeting of the SSA NSW Branch.
- Distinguished Professor Kerrie Mengersen gave an excellent Lancaster Lecture last year, reflecting on Prof. Lancaster's extensive research work which culminates in his book *Expectations of Life*.
  - Videos of past seminars are available on the Statistical Society website.
- I would like to use this opportunity to reflect on Prof. Lancaster's contribution to the wider statistical community.

NSW Branch and H.O. Lancaster

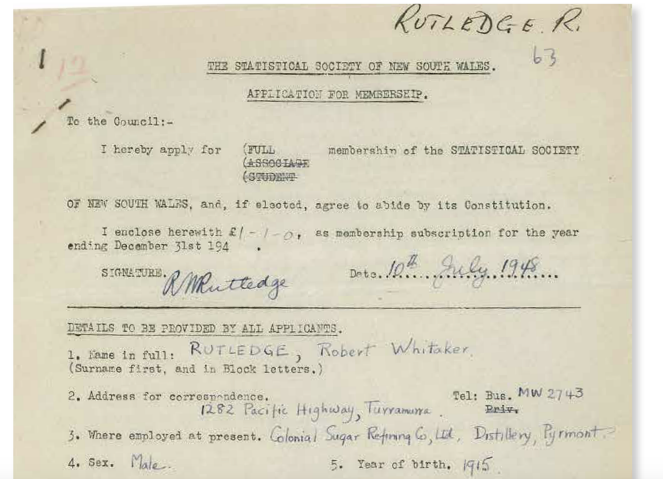
- Statistical Society of Australia was initially founded as the "Statistical Society of New South Wales" in 1947.
- In 1962 it joined forces with the Canberra Statistical Society to create a national Statistical Society.
- Who believes Prof. Lancaster was our first ever president?

NSW Branch and H.O. Lancaster

- Statistical Society of Australia was initially founded as the "Statistical Society of New South Wales" in 1947.
- In 1962 it joined forces with the Canberra Statistical Society to create a national Statistical Society.
- Who believes Prof. Lancaster was our first ever president?
- **He's not!**

NSW Branch and H.O. Lancaster (cont.)

- There were extensive discussions between Stuart Rutherford, Helen Newton Turner, Oliver Lancaster exploring ways in which members of the field of statistics could communicate with each other.



RUTLEDGE R.

THE STATISTICAL SOCIETY OF NEW SOUTH WALES. 63

APPLICATION FOR MEMBERSHIP.

To the Council:-

I hereby apply for (FULL ~~ASSOCIATE~~ ~~STUDENT~~ membership of the STATISTICAL SOCIETY OF NEW SOUTH WALES, and, if elected, agree to abide by its Constitution.

I enclose herewith £1-1-0, as membership subscription for the year ending December 31st 194 .

SIGNATURE. R.W. Rutledge Date. 10th July 1948

DETAILS TO BE PROVIDED BY ALL APPLICANTS.

1. Name in full: RUTLEDGE, Robert Whitaker. (Surname first, and in Block letters.)
2. Address for correspondence. 1282 Pacific Highway, Turramurra. Tel: Bus. MW 2743 ~~Priv.~~
3. Where employed at present. Colonial Sugar Refining Co. Ltd. Distillery, Pyrmont.
4. Sex. Male. 5. Year of birth. 1915.

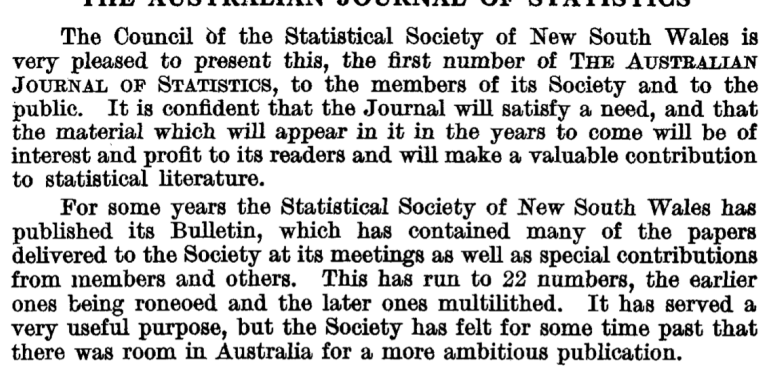
- Helen Newtown Turner was our first President!
  - She was a leading authority on sheep genetics and worked at the McMaster Animal Health Laboratory of the Council for Scientific and Industrial Research (later became a part of CSIRO) for 40 years.

- Prof. Lancaster was the NSW Branch President in 1952–1953 as well as the National President in 1965–1967.

- Prof. Lancaster was also involved in the founding of the Australian Mathematical Society in 1956–1957, and later (1967–1968) served as its President.

A(NZ)JS and H.O. Lancaster

- In 1949, members were treated to the first of many monthly bulletins edited by Prof. Lancaster.
- In 1959, the Bulletin became the first published edition of the Australian Journal of Statistics.
- He was Editor of the journal till 1971.



THE AUSTRALIAN JOURNAL OF STATISTICS

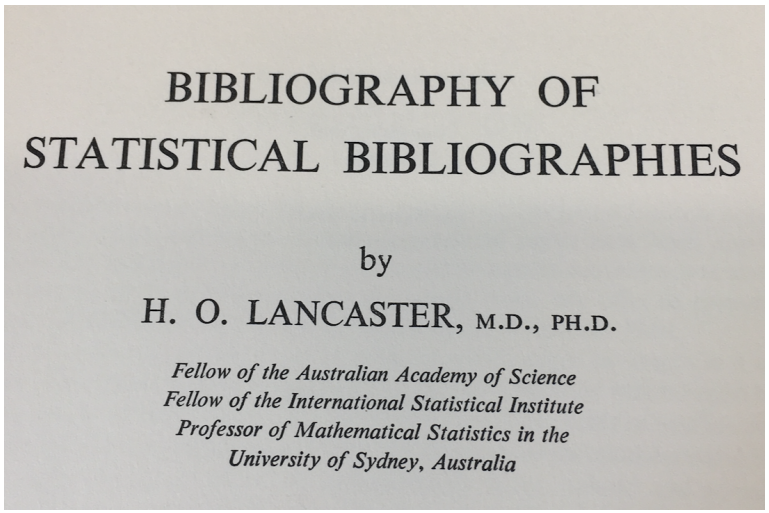
The Council of the Statistical Society of New South Wales is very pleased to present this, the first number of THE AUSTRALIAN JOURNAL OF STATISTICS, to the members of its Society and to the public. It is confident that the Journal will satisfy a need, and that the material which will appear in it in the years to come will be of interest and profit to its readers and will make a valuable contribution to statistical literature.

For some years the Statistical Society of New South Wales has published its Bulletin, which has contained many of the papers delivered to the Society at its meetings as well as special contributions from members and others. This has run to 22 numbers, the earlier ones being roneoed and the later ones multilithed. It has served a very useful purpose, but the Society has felt for some time past that there was room in Australia for a more ambitious publication.

Bibliography and H.O. Lancaster

- Another field that Prof. Lancaster's excelled was Statistical Bibliography.
- The Bibliography alone in *Expectations of Life: A Study in the Demography Statistics and History of World Mortality* is almost 90 pages long and contains (as an estimate) about 2,800 items.
- This compilation was done without the aid of electronic databases.

Bibliography and H.O. Lancaster (cont.)

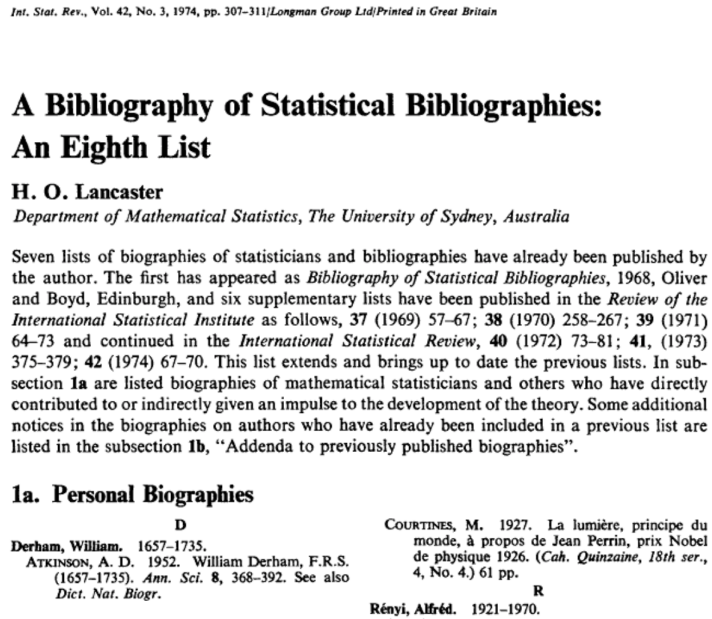


BIBLIOGRAPHY OF
STATISTICAL BIBLIOGRAPHIES

by

H. O. LANCASTER, M.D., PH.D.

Fellow of the Australian Academy of Science
Fellow of the International Statistical Institute
Professor of Mathematical Statistics in the
University of Sydney, Australia



Int. Stat. Rev., Vol. 42, No. 3, 1974, pp. 307–311/*Longman Group Ltd/Printed in Great Britain*

A Bibliography of Statistical Bibliographies: An Eighth List

H. O. Lancaster
Department of Mathematical Statistics, The University of Sydney, Australia

Seven lists of biographies of statisticians and bibliographies have already been published by the author. The first has appeared as *Bibliography of Statistical Bibliographies*, 1968, Oliver and Boyd, Edinburgh, and six supplementary lists have been published in the *Review of the International Statistical Institute* as follows, **37** (1969) 57–67; **38** (1970) 258–267; **39** (1971) 64–73 and continued in the *International Statistical Review*, **40** (1972) 73–81; **41**, (1973) 375–379; **42** (1974) 67–70. This list extends and brings up to date the previous lists. In subsection **1a** are listed biographies of mathematical statisticians and others who have directly contributed to or indirectly given an impulse to the development of the theory. Some additional notices in the biographies on authors who have already been included in a previous list are listed in the subsection **1b**, "Addenda to previously published biographies".

1a. Personal Biographies

D

Derham, William. 1657–1735.
ATKINSON, A. D. 1952. William Derham, F.R.S. (1657–1735). *Ann. Sci.* **8**, 368–392. See also *Dict. Nat. Biogr.*

COURTINES, M. 1927. La lumière, principe du monde, à propos de Jean Perrin, prix Nobel de physique 1926. (*Cah. Quinzaine, 18th ser.*, 4, No. 4.) 61 pp.

R

Rényi, Alfréd. 1921–1970.

- He also published Bibliography of Statistical Bibliographies (1968) and its 21 addenda over the succeeding years.

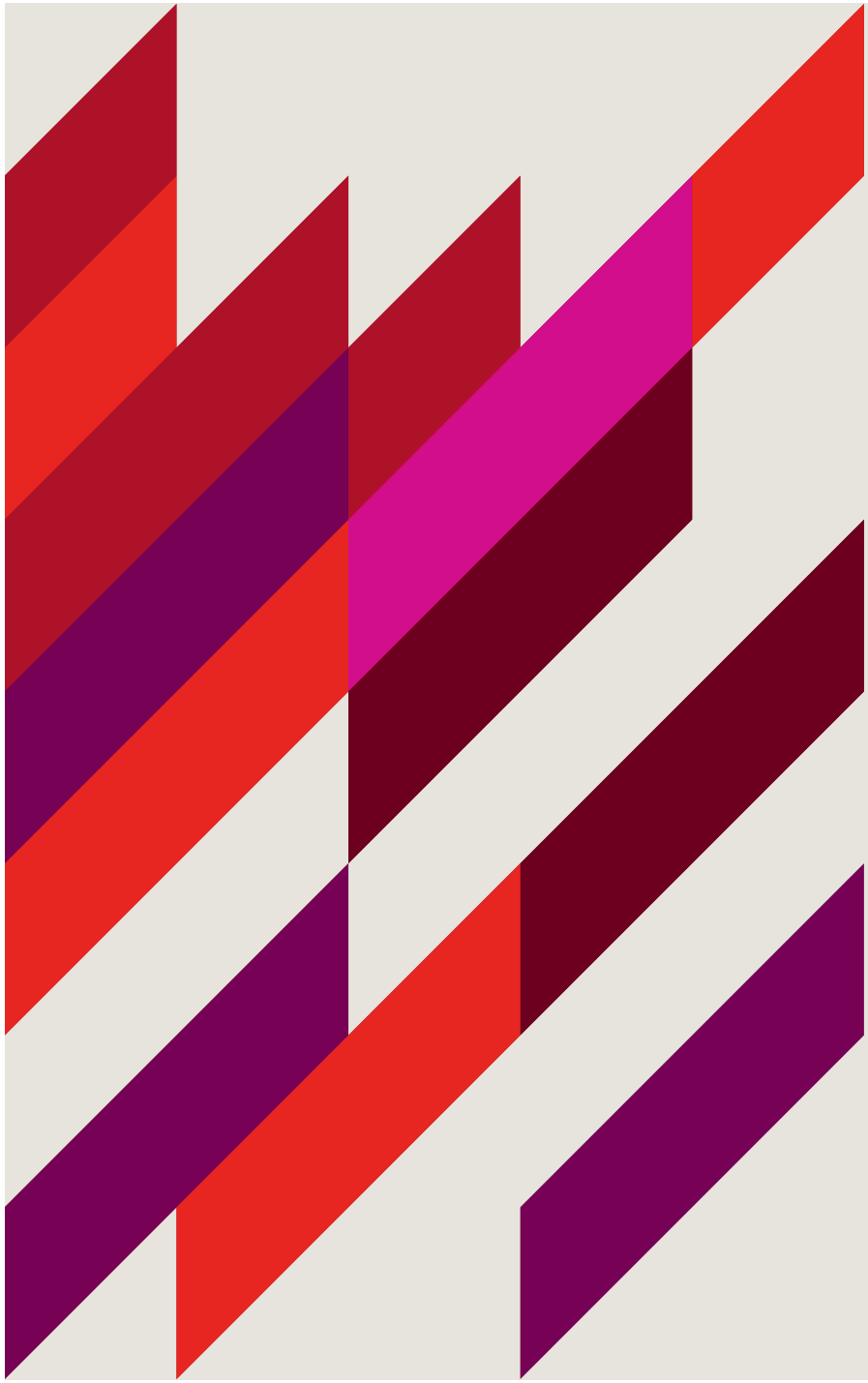
Acknowledgement

- This presentation represents joint research with:

Dr Alan Huang
University of Queensland

Aya Alwan
Macquarie University

Dr Justin Wishart
Displayr



MACQUARIE University
SYDNEY·AUSTRALIA

Conway-Maxwell-Poisson Generalized Linear Models

Conway-Maxwell-Poisson Distributions

MACQUARIE University
SYDNEY·AUSTRALIA

- The CMP distribution was first used by Conway and Maxwell (1962) as a model for a queuing system with dependent service times.

- A random variable is said to have a (standard) CMP distribution with rate parameter λ and dispersion parameter ν if its probability mass function (pmf) is given by

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad y = 0, 1, 2, ...,$$

  where

$$Z(\lambda, \nu) = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y!)^\nu},$$

  is a normalizing constant.

- The CMP includes Poisson ($\nu = 1$), geometric ($\nu = 0$, $\lambda < 1$) and Bernoulli ($\nu \to \infty$ with probability $\lambda/(1 + \lambda)$).

- The CMP distribution does not have closed-form expression for its moments in terms of the parameters λ & ν but satisfy recursive formulas,

$$E(Y^{r+1}) = \begin{cases} \lambda E(Y+1)^{1-\nu}, & r = 0; \\ \lambda \frac{d}{d\lambda} E(Y^r) + E(Y)E(Y^r), & r > 0. \end{cases}$$

- For the first two moments, approximations can be obtained as

$$E(Y) = \frac{\partial \log Z}{\partial \log \lambda} \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu};$$

$$\text{var}(Y) = \frac{\partial^2 \log Z}{\partial (\log \lambda)^2} \approx \frac{1}{\nu} \lambda^{\frac{1}{\nu}} \approx \frac{1}{\nu} E(Y),$$

and they can be particularly accurate for $\nu \leq 1$ or $\lambda > 10^{\nu}$ (see Shmueli et al. (2005)).

- The CMP distribution does not have closed-form expression for its moments in terms of the parameters λ & ν but satisfy recursive formulas,

$$E(Y^{r+1}) = \begin{cases} \lambda E(Y+1)^{1-\nu}, & r = 0; \\ \lambda \frac{d}{d\lambda} E(Y^r) + E(Y)E(Y^r), & r > 0. \end{cases}$$

- For the first two moments, approximations can be obtained as

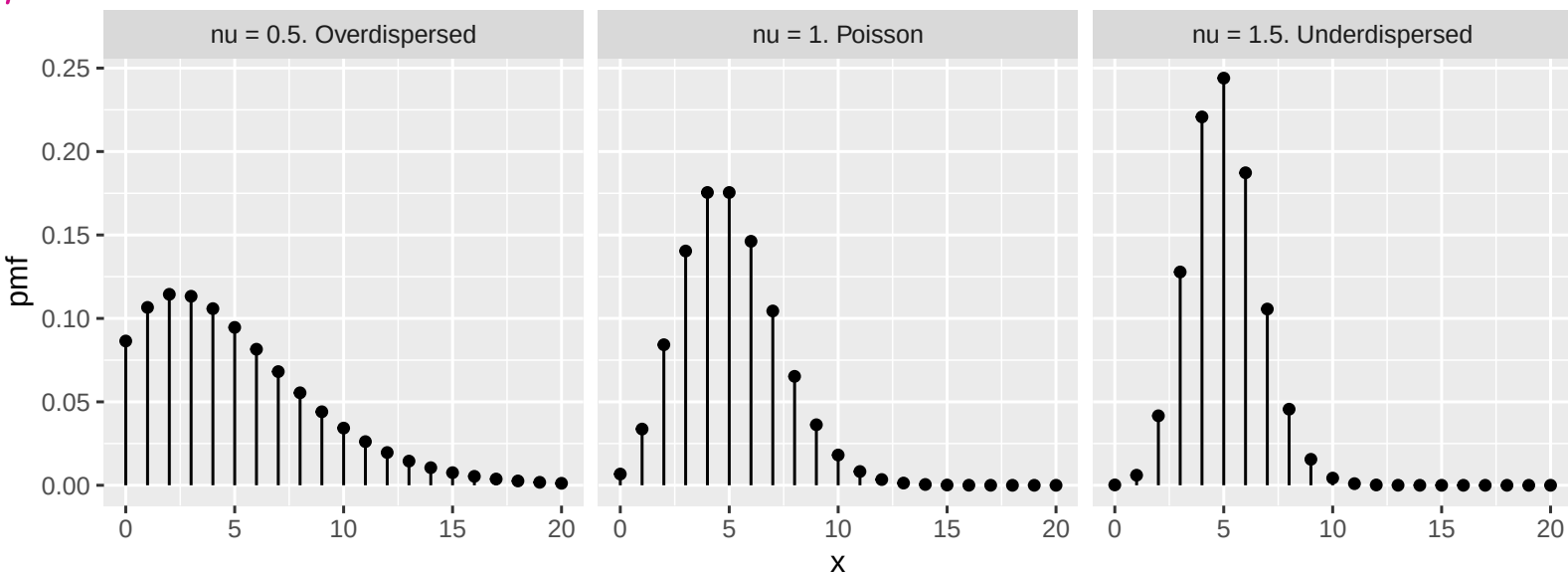
$$E(Y) = \frac{\partial \log Z}{\partial \log \lambda} \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu} \neq \lambda;$$

$$\text{var}(Y) = \frac{\partial^2 \log Z}{\partial (\log \lambda)^2} \approx \frac{1}{\nu} \lambda^{\frac{1}{\nu}} \approx \frac{1}{\nu} E(Y),$$

and they can be particularly accurate for $\nu \leq 1$ or $\lambda > 10^{\nu}$ (see Shmueli et al. (2005)).

- If $\nu < 1$, CMP is overdispersed. In reverse, CMP is underdispersed when $\nu > 1$.

- Here is a plot of the density for a few CMP distributions with **mean** $\mu = 5$:



Conway-Maxwell-Poisson GLM

MACQUARIE University
SYDNEY·AUSTRALIA

- As CMP is one of a few distributions that can handle both under- and over-dispersion, the aim is to extend the GLM formulation to the CMP case so that one can model the relationship between Y and the predictors X.

- Given a set of covariates $X \in \mathbb{R}^q$, Sellers and Shmueli (2010) proposed a GLM for count response Y that can be specified via

$$Y|X \sim CMP(\lambda, \nu), \quad \text{s.t.} \quad \log \lambda = X^\top \beta$$

  where $\beta \in \mathbb{R}^q$ is a vector of regression coefficients.

- This structure forms the basic of the `R` package of `COMPoissonReg` of Sellers, Lotze, and Raim (2017) and `CompGLM` of Pollock (2018).

- This model, however, does not provide a closed-form relationship between $E(Y)$ and the linear predictor, making it incompatible with other commonly used log-linear models

Mean-parametrization

- As it is more convenient and interpretable to model the mean $\mu = E(Y) > 0$ of the distribution directly, Huang (2017) proposed to parametrize the CMP distribution via the mean:

$$P(Y = y|\mu, \nu) = \frac{\lambda(\mu, \nu)^y}{(y!)^\nu} \frac{1}{Z(\lambda(\mu, \nu), \nu)}, \quad y = 0, 1, 2, \ldots,$$

where the rate $\lambda(\mu, \nu)$ is defined as the solution to the mean constraint:

$$\mu = \sum_{y=0}^{\infty} y \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}.$$

- We shall denote this as $\text{CMP}_\mu(\mu, \nu)$ distribution to distinguish it from the original/standard one.

- A GLM that based on CMP_μ can then be specified via
$$Y|X \sim CMP_\mu(\mu(X^\top\beta), \nu),$$
where
$$E(Y|X) = \mu(X^\top\beta) = \exp(X^\top\beta).$$

- Pros:
  - GLM that based on CMP_μ is a genuine GLM, so all the familiar key features of GLMs are retained.
  - Comparable and compatible with other commonly used log-linear regression models for counts.
  - The mean $\mu = \exp(X^\top\beta)$ and the dispersion ν are orthogonal, making it similar in structure to the familiar Negative Binomial regression model for overdispersed counts (in contrast to the rate λ and ν in the standard CMP are not).
  - Easy to incorporate offsets into the model.

- The model can also be extended to allow varying dispersion, i.e. ν itself is modelled via a regression:
$$\nu = \exp(\tilde{X}^\top\gamma),$$
where $\tilde{X}$ is some covariates.

Let's meet the rest of our competition!

- Recall that
$$\mu = E(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu},$$
so Ribeiro et al. (2020) proposed a slightly simpler parametrization of
$$Y|X \sim \text{CMP}\left(\lambda\left(X^\top\beta, \nu\right), \nu\right),$$
where
$$\lambda\left(X^\top\beta, \nu\right) = \left(e^{X^\top\beta} + \frac{(\nu - 1)}{2\nu}\right)^\nu.$$

- We shall denote this as the $\text{CMP}_{\approx\mu}$ distribution.

- This structure forms the basic of the `R` package of `cmpreg` of Ribeiro (2020).

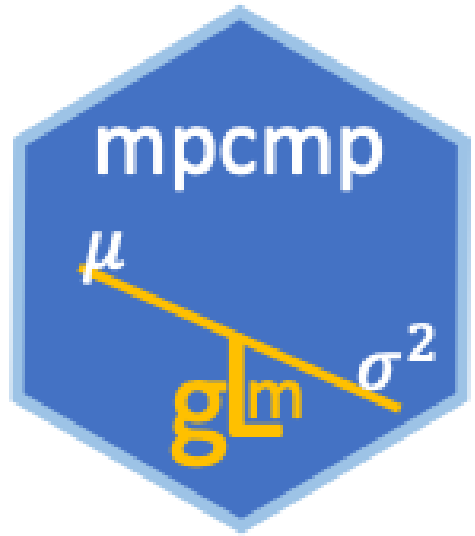
- The problem with this parametrization is that when $\hat{\mu}$ is small, it restricts how small ν can be so that $\lambda > 0$.

- There is also the `DGLMExtPois` (Double Generalized Linear Models Extending Poisson) package of Saez-Castillo, Conde-Sanchez, and Martinez (2020).

- They use the `nloptr` package of Ypma (2017), which is the `R` interface to `NLopt` library, to fit CMP_μ model.

The mpcmp package

- The package **mpcmp** of Fung, Alwan, Wishart and Huang (2020) provides parameter estimates for a **M**ean-**P**arametrized **CMP** log-linear regression and associated standard errors; a LRT for testing data dispersion ($\nu = 1$), and other model diagnostics.



- The optimization is done using Fisher Scoring updates to take advantage of the fact the CMP_μ belongs to the exponential family.

- Notice that this is a constrained optimisation problem as we have to maintain mean constraints at all times, which makes the implementation is a bit more challenging.

Constant Overdispersion Example: `attendance`

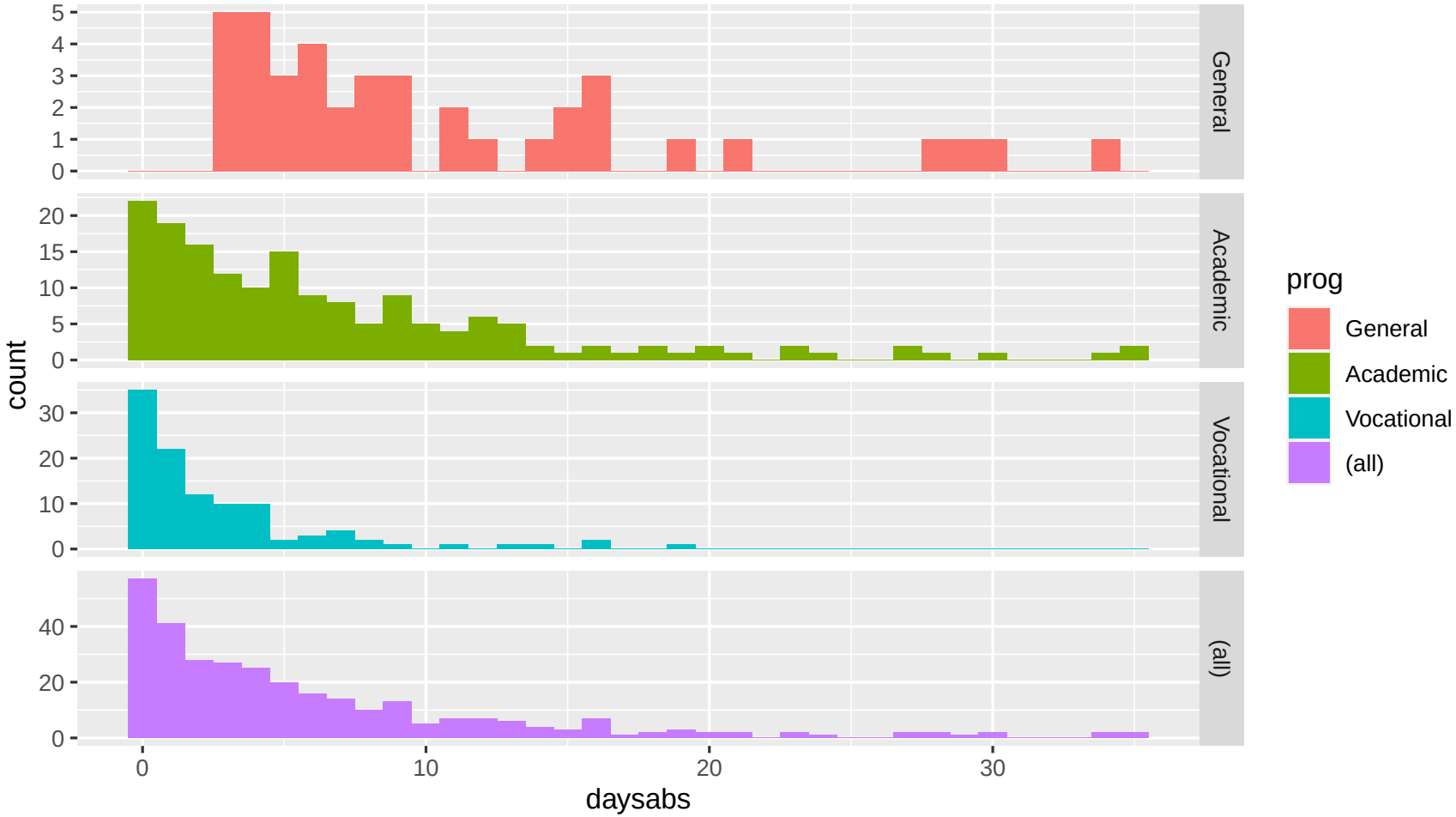


Figure 1: The attendance dataset from `https://stats.idre.ucla.edu/stat/stata/dae/nb_data.dta` examines the relationship between the number of days absent from school and the gender, maths score and academic program of 314 students from two urban high schools.

	CMP	CMP_μ	Neg Bin
Intercept	0.018	2.715	2.707
I(Male)	-0.035	-0.215	-0.211
I(Academic)	-0.050	-0.425	-0.425
I(Vocational)	-0.232	-1.254	-1.253
Math	-0.001	-0.006	-0.006
Disperson	0.021	0.020	1.047

- As `mpcmp` is directly comparable and compatible with other commonly used log-linear regression models for counts, interpreting parameters is straight forward.

- Our model estimates that students in the General program (the reference level) are expected to miss $\exp(+1.254) = 3.504$ times more days of school compared to students in the Vocational program.

	CMP	CMP$_\mu$	Neg Bin
Intercept	0.018	2.715	2.707
I(Male)	-0.035	-0.215	-0.211
I(Academic)	-0.050	-0.425	-0.425
I(Vocational)	**-0.232**	-1.254	**-1.253**
Math	-0.001	-0.006	-0.006
Disperson	0.021	0.020	1.047

- As `mpcmp` is directly comparable and compatible with other commonly used log-linear regression models for counts, interpreting parameters is straight forward.

- Our model estimates that students in the General program (the reference level) are expected to miss $\exp(+1.254) = 3.504$ times more days of school compared to students in the Vocational program.

	CMP	CMP$_\mu$	Neg Bin
Intercept	0.018	2.715	2.707
I(Male)	-0.035	-0.215	-0.211
I(Academic)	-0.050	-0.425	-0.425
I(Vocational)	-0.232	**-1.254**	**-1.253**
Math	-0.001	-0.006	-0.006
Disperson	0.021	0.020	1.047

- As `mpcmp` is directly comparable and compatible with other commonly used log-linear regression models for counts, interpreting parameters is straight forward.

- Our model estimates that students in the General program (the reference level) are expected to miss $\exp(+1.254) = 3.504$ times more days of school compared to students in the Vocational program.

How about $CMP_{\approx\mu}$?

	CMP	CMP_{μ}	Neg Bin
Intercept	0.018	2.715	2.707
I(Male)	-0.035	-0.215	-0.211
I(Academic)	-0.050	-0.425	-0.425
I(Vocational)	-0.232	-1.254	-1.253
Math	-0.001	-0.006	-0.006
Disperson	0.021	0.020	1.047

```
cmpreg::cmp(formula = daysabs~ gender+math+prog,
            data = attendance)
```

```
## Error in optim(par = start, fn = llcmp, method = method, lowe
```

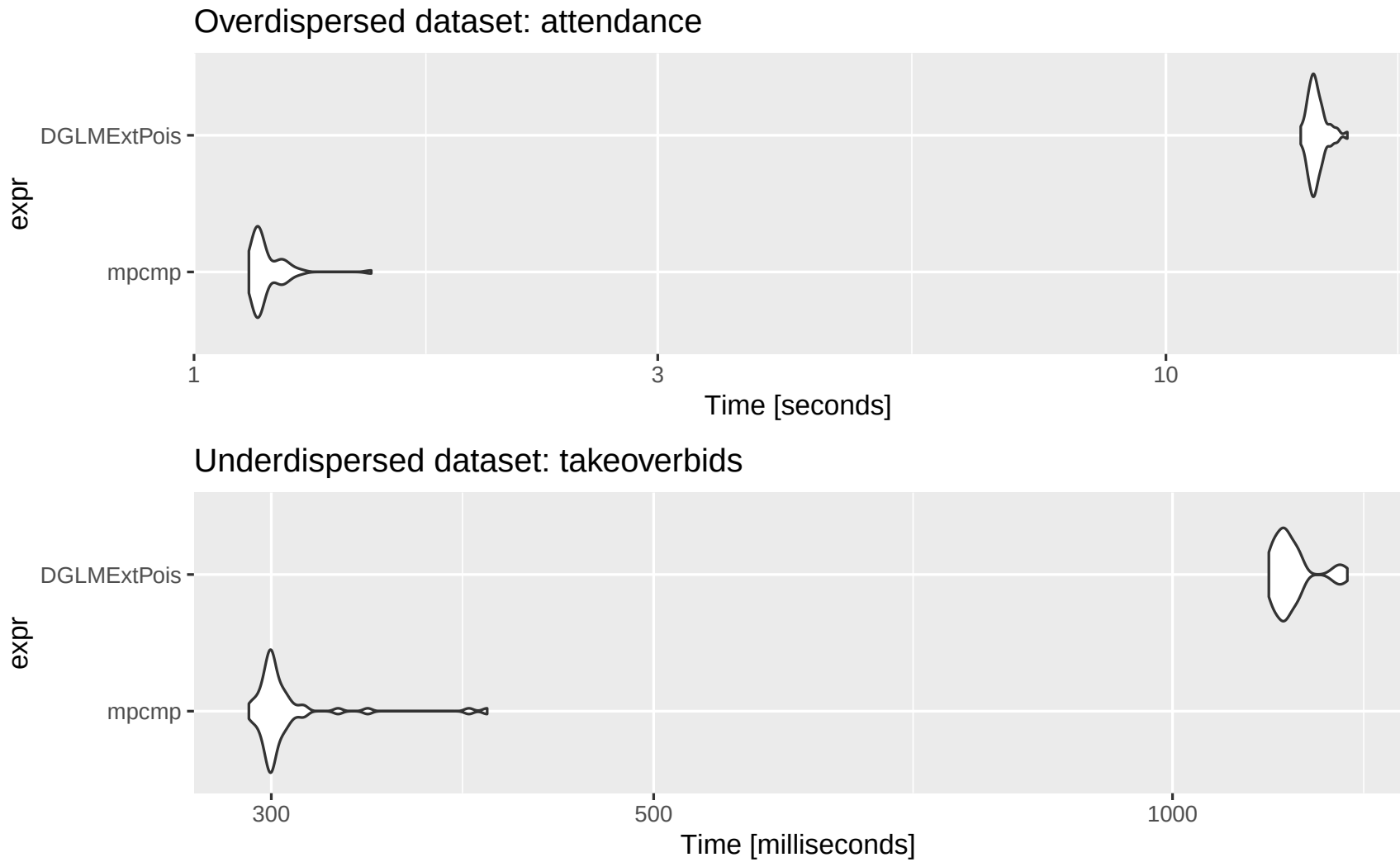
How about $CMP_{\approx\mu}$?

	CMP	CMP_{μ}	Neg Bin
Intercept	0.018	2.715	2.707
I(Male)	-0.035	-0.215	-0.211
I(Academic)	-0.050	-0.425	-0.425
I(Vocational)	-0.232	-1.254	-1.253
Math	-0.001	-0.006	-0.006
Disperson	0.021	**0.020**	1.047

```
cmpreg::cmp(formula = daysabs~ gender+math+prog,
            data = attendance)
```

```
## Error in optim(par = start, fn = llcmp, method = method, lowe
```

How about `DGLMExtPois`?



Analysing the attendance datatset

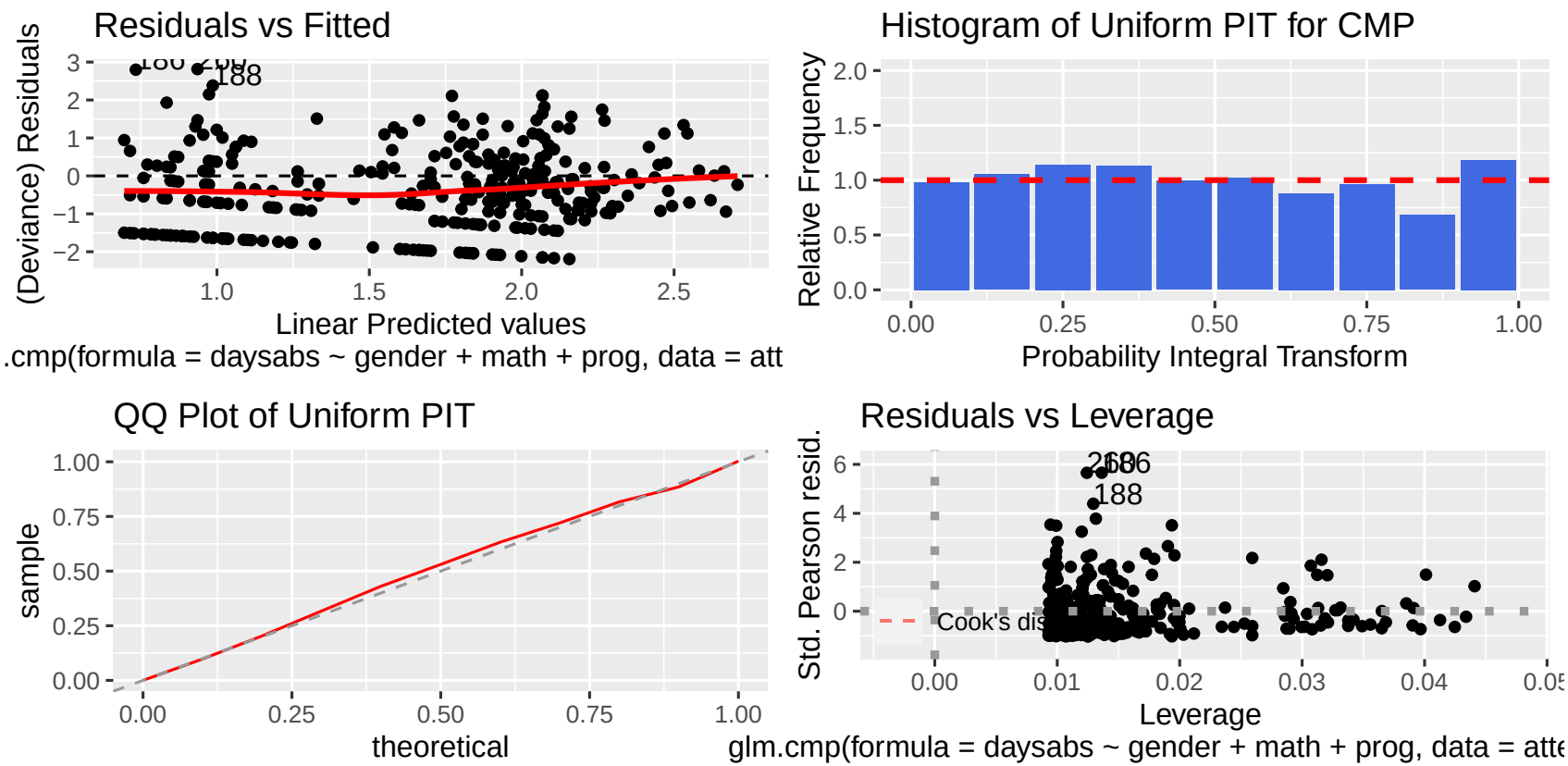
```
summary(M.attendance <- glm.cmp(daysabs~ gender+math+prog, data=attendance))
```

```
##
## Call: glm.cmp(formula = daysabs ~ gender + math + prog, data = attendance)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1925  -1.1166  -0.3973   0.2964   2.8154
##
## Linear Model Coefficients:
##                  Estimate   Std.Err Z value Pr(>|z|)
## (Intercept)     2.714645  0.190407  14.257  < 2e-16 ***
## gendermale     -0.214720  0.117148  -1.833  0.06682 .
## math           -0.006323  0.002386  -2.650  0.00804 **
## progAcademic   -0.425322  0.169524  -2.509  0.01211 *
## progVocational -1.253896  0.189478  -6.618 3.65e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Mean-CMP estimated to be 0.02024)
##
##
##     Null deviance: 455.83  on 313 degrees of freedom
## Residual deviance: 377.44  on 309 degrees of freedom
##
```

Diagnostic

- One of the key features of the `mpcmp` package is that it provides a range of diagnostic plots.

```
autoplot(M.attendance, which = c(1,2,3,8))
```



- The `mpcmp` package also supports `broom` `tidier` method.

```
tidy(M.attendance)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      2.71      0.190       14.3  4.05e-46
## 2 gendermale      -0.215     0.117       -1.83 6.68e- 2
## 3 math            -0.00632   0.00239     -2.65 8.04e- 3
## 4 progAcademic    -0.425     0.170       -2.51 1.21e- 2
## 5 progVocational -1.25       0.189       -6.62 3.65e-11
```

- This means `mpcmp` can take advantage of any enhancement packages that use `broom` such as `modelsummary` package.

	mpcmp	Poisson	neg.bin
(Intercept)	2.715	2.759	2.707
	(0.190)	(0.064)	(0.204)
gendermale	-0.215	-0.242	-0.211
	(0.117)	(0.047)	(0.122)
math	-0.006	-0.007	-0.006
	(0.002)	(0.001)	(0.002)
progAcademic	-0.425	-0.426	-0.425
	(0.170)	(0.057)	(0.182)
progVocational	-1.254	-1.271	-1.253
	(0.189)	(0.078)	(0.200)
Num.Obs.	314	314	314
AIC	1739.0	2640.2	1740.3
BIC	1761.5	2658.9	1762.8
Log.Lik.	-863.513	-1315.089	-864.154

- We also implemented a bunch of commonly used methods for "cmp" objects:

```
##  [1] augment         confint         cooks.distance glance
##  [6] influence       model.matrix    plot           predict
## [11] rstandard       summary         tidy           update
## see '?methods' for accessing help and source code
```

Constant Underdispersion: `takeoverbids`

- A dataset from Cameron & Johansson (1997) that gives the number of bids received by 126 US firms that were successful targets of tender offers during the period 1978-85.

- The dataset comes with a set of explanatory variables such as defensive actions taken by management of target firm, firm-specific characteristics and intervention by federal regulators.

- Under the CMP_μ model, $\hat{\nu} = 1.75$.

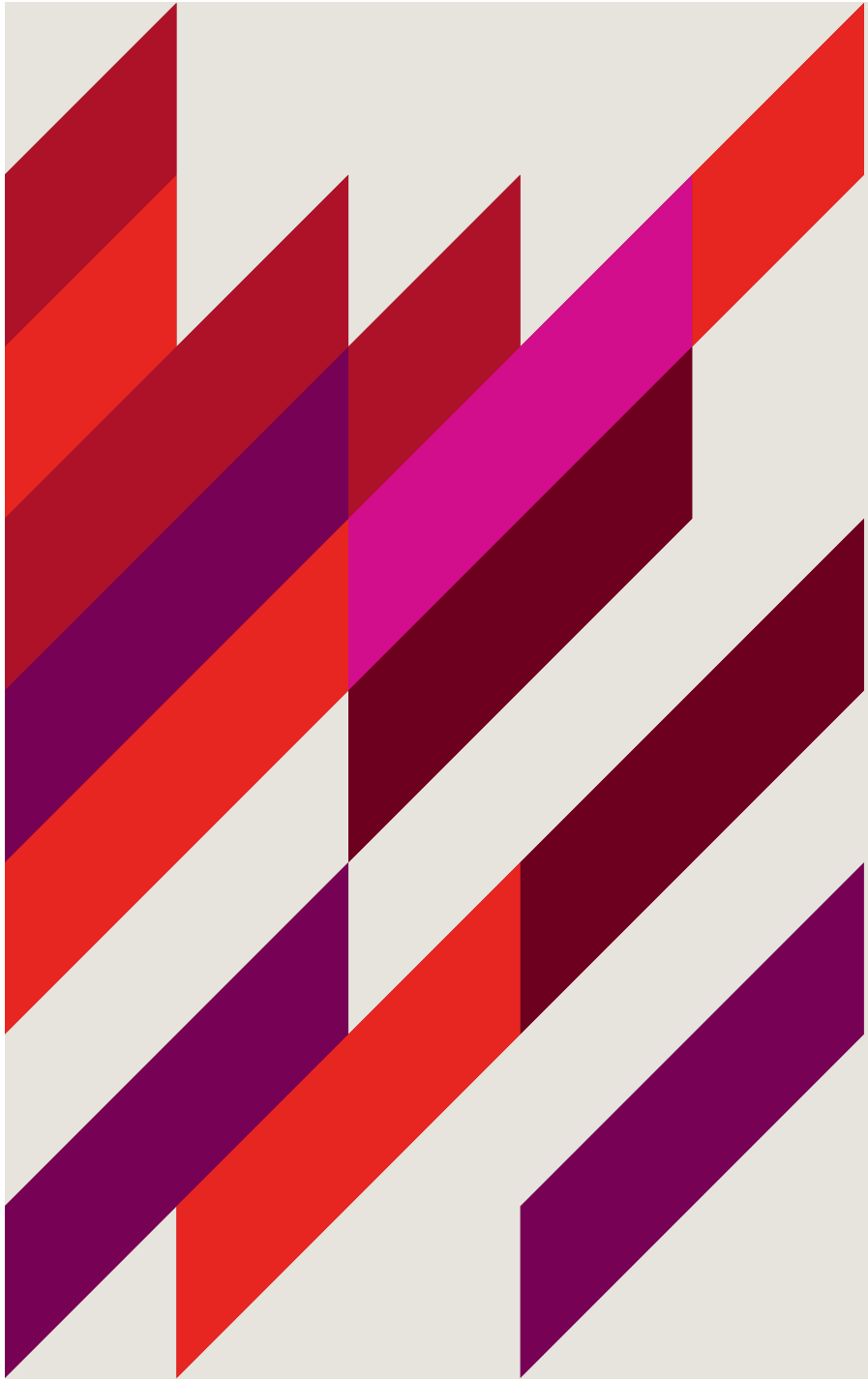
```
##
## Likelihood ratio test for testing nu=1:
##
## Log-Likelihood for Mean-CMP:  -180
## Log-Likelihood for Poisson:  -185
## LRT-statistic:  9.72
## Chi-sq degrees of freedom:  1
## P-value:  0.00182
```

	mpcmp	Poisson
(Intercept)	0.990 (0.435)	0.986 (0.534)
leglrest	0.268 (0.123)	0.260 (0.151)
rearest	-0.173 (0.155)	-0.196 (0.193)
finrest	0.068 (0.174)	0.074 (0.217)
whtknght	0.481 (0.132)	0.481 (0.159)
bidprem	-0.685 (0.308)	-0.678 (0.377)
insthold	-0.368 (0.347)	-0.362 (0.424)
size	0.179 (0.048)	0.179 (0.060)
sizesq	-0.008 (0.002)	-0.008 (0.003)
regulatn	-0.038 (0.130)	-0.029 (0.161)
Num.Obs.	126	126
AIC	382.2	389.9
BIC	413.4	418.3
Log.Lik.	-180.088	-184.948

- All the estimated standard errors are smaller under CMP_{μ}.

Recap

- If you ever consider using a CMP but are unsure of which package you should use, remember to use the one that is mean!
- You can find the `mpcmp` package on `CRAN` as well as from my GitHub page: github.com/thomas-fung/mpcmp.
- Fung, T., Alwan, A., Wishart, J. & Huang, A. (2020). mpcmp: Mean-parametrized Conway-Maxwell-Poisson Regression. R package version 0.3.5.

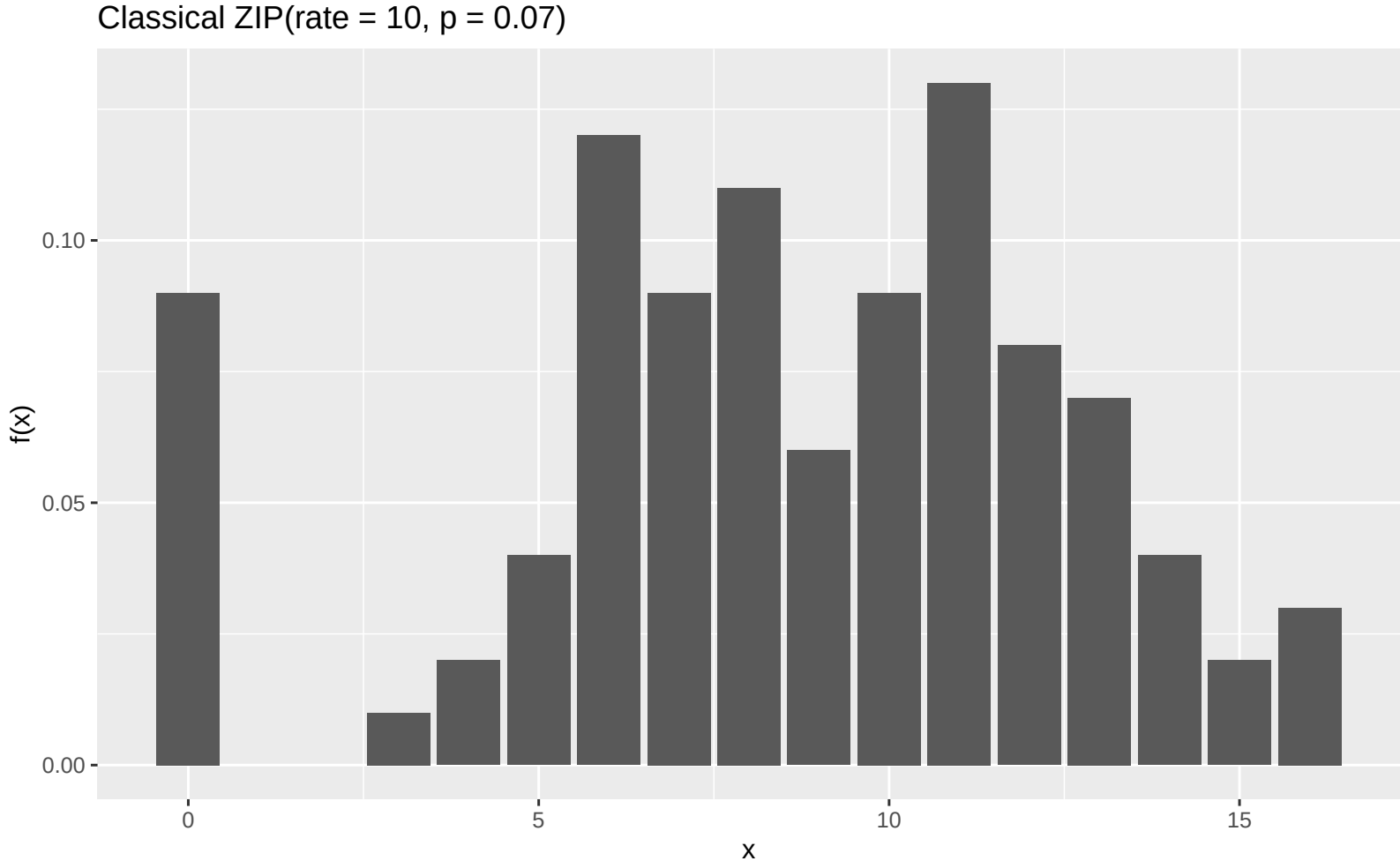


MACQUARIE
University
SYDNEY·AUSTRALIA

Conway-Maxwell-Poisson Associated Kernel (compak) Smoother

Discrete Kernel Smoother

- This represents some joint work with Alan Huang and Lucas Sippel.
- Suppose we have some count data.

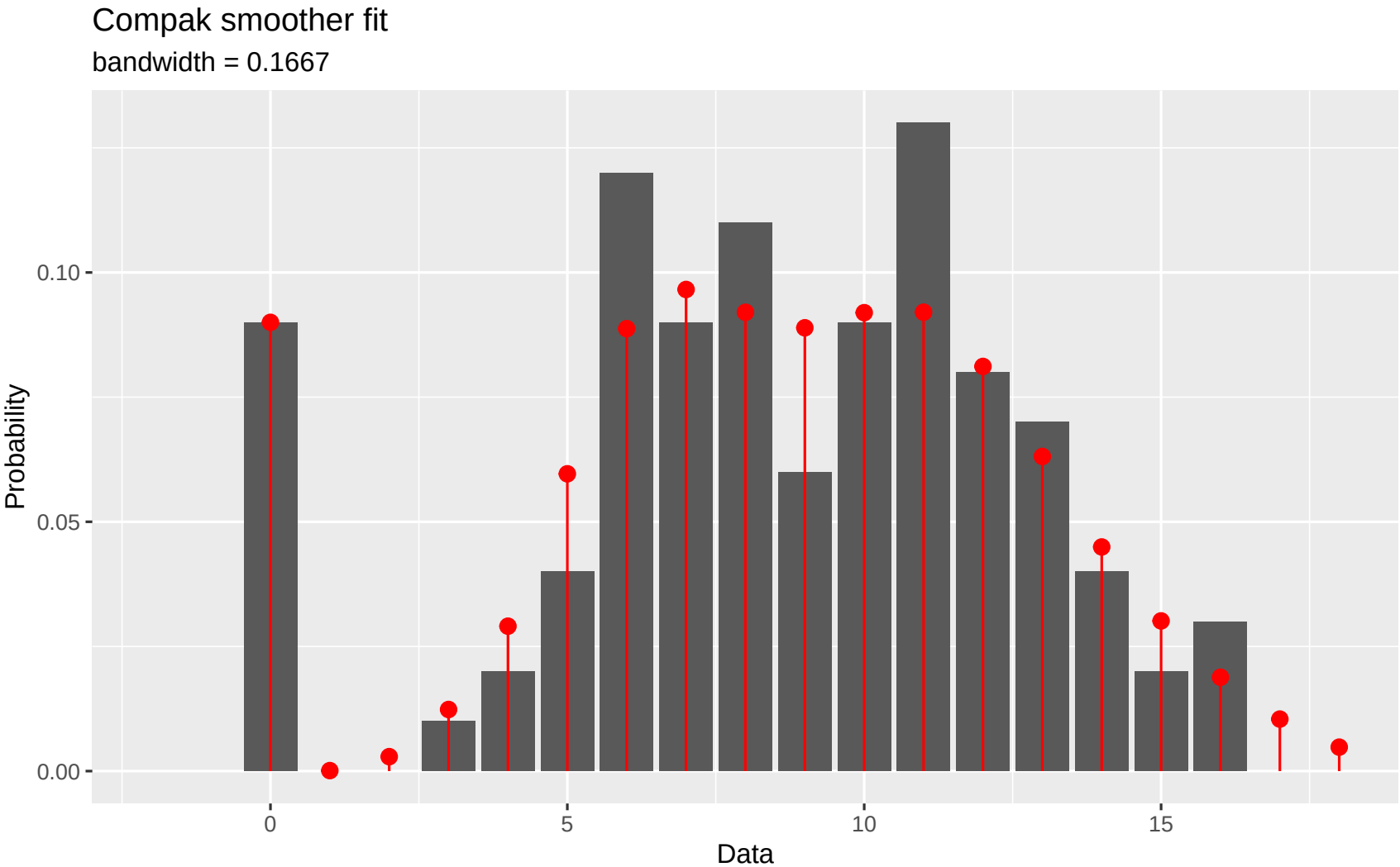


Discrete Kernel Smoother (cont.)

- Kernel smoothers for discrete distributions have not been explored much.
- In this project, we used the CMP_μ distribution to create a consistent second-order discrete associated kernel.
- CMP_μ kernels can be arbitrarily underdispersed, which is the key to obtaining consistency as the sample size increases.
- We called this the `compak` smoother.

Discrete Kernel Smoother (cont.)

- Kernel smoothers for discrete distributions have not been explored much.
- In this project, we used the CMP_μ distribution to create a consistent second-order discrete associated kernel.
- CMP_μ kernels can be arbitrarily underdispersed, which is the key to obtaining consistency as the sample size increases.
- We called this the `compak` smoother.

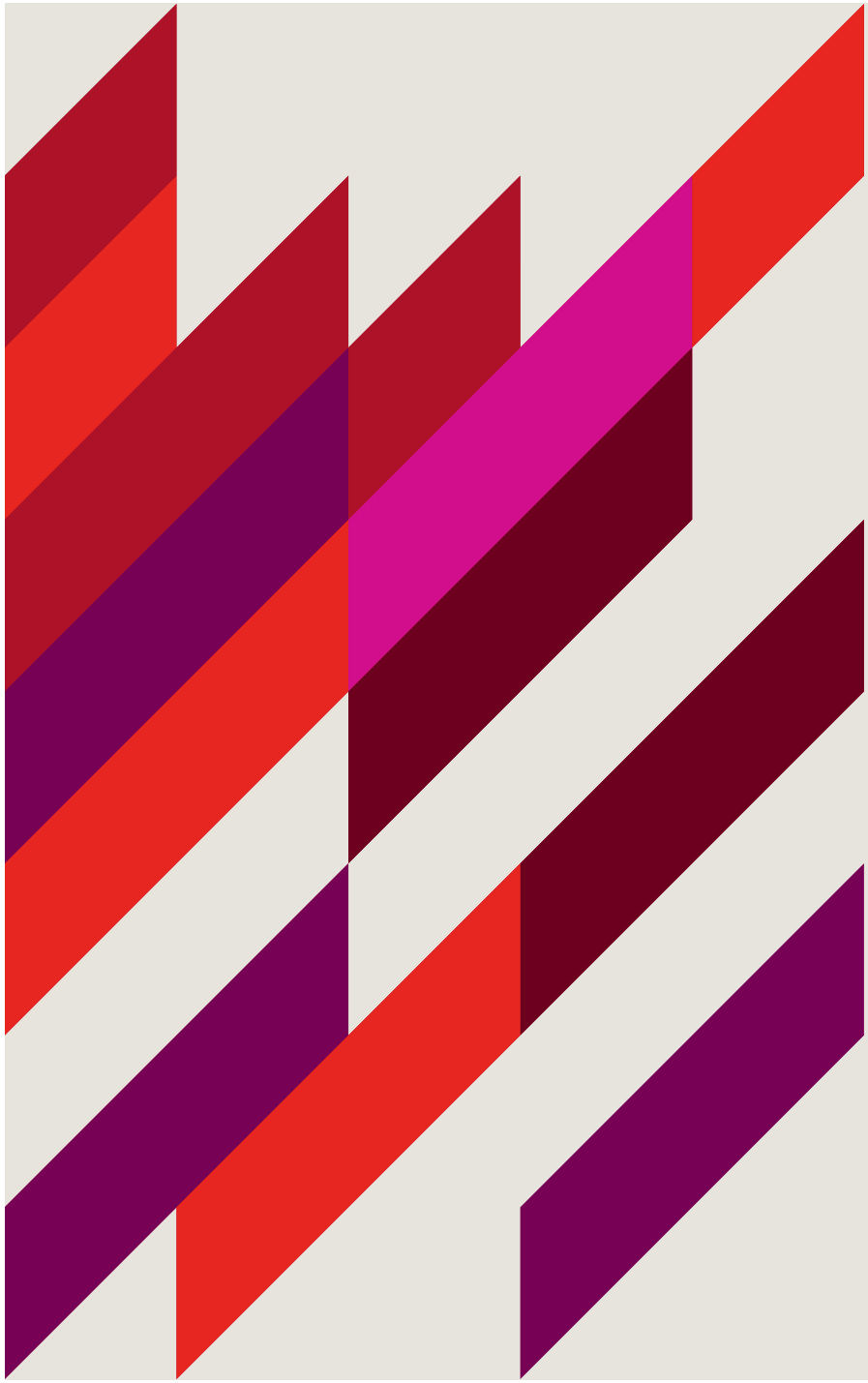

Compak smoother fit
bandwidth = 0.1667
Probability
0.10
0.05
0.00
0
5
10
15
Data

Interested?

- Preprint available on

Huang, A., Sippel, L., & Fung, T. (2020). A consistent second-order discrete kernel smoother. arXiv preprint arXiv:2010.03302.

- The development version of `compak` package is available on my Github page. `https://github.com/thomas-fung/compak`.



MACQUARIE University
SYDNEY·AUSTRALIA

Zero-inflated Poisson

Classical ZIP

- In many count data processes, zero observations occur more frequently than expected from a nominal distribution.
- Perhaps the most well-known model for such scenarios is the zero-inflated Poisson (ZIP) of Lambert (1992).
- ZIP can be constructed via two independent latent variables, namely, $B \sim$ Bernoulli with some probability π of being zero and $P \sim$ Poisson with some rate λ.
- Suppose $Y = BP \sim \mathrm{ZIP}(\pi, \lambda)$, if the corresponding probability mass function is:

$$f(y|\pi,\lambda) = \begin{cases} \pi + (1-\pi)e^{-\lambda}\,, & y = 0; \\ (1-\pi)e^{-\lambda}\frac{\lambda^y}{y!}\,, & y = 1, 2, \ldots \end{cases}$$

with mean $E(Y) = (1-\pi)\lambda$ and variance $\mathrm{var}(Y) = \lambda(1-\pi)(1+\pi\lambda)$.

Classical ZIP (cont.)

- One desirable feature of ZIP is that the latent Bernoulli construction offers an explicit explanation of the excess zeros.
- However, regression (and time-series) models the mean of the observed response Y can only be identified to the product $(1-\pi)\lambda$, using classical ZIP.
- The goodness-of-fit of ZIP models depends crucially on the individual models for π and λ, but this can be not easy to check as neither process is fully observed.

Zero-inflated Poisson exponential families

- Let $f(y|\pi, \lambda)$ be the pmf of $Y \sim \text{ZIP}(\pi, \lambda)$.
- Construct a family $\{f_\theta(y); \theta \in \mathbb{R}\}$ of distributions, indexed by θ, via exponential tilting:
$$f_\theta(y) \propto \exp(\theta y) f(y|\pi, \lambda) \ , \quad \theta \in \mathbb{R}.$$
- Each $f_\theta(y)$ remains a ZIP distribution with new parameters π_θ and λ_θ given respectively by
$$\pi_\theta = \frac{\pi}{\pi + (1-\pi)e^{\lambda(e^\theta - 1)}} \quad \text{and} \quad \lambda_\theta = \lambda e^\theta, \tag{1}$$
- We can set $\lambda \equiv 1$ without loss of generality, and π can then be interpreted as a baseline zero-inflation rate relative to a standard Poisson(1) distribution.

Zero-inflated Poisson exponential families

- Let $f(y|\pi, \lambda)$ be the pmf of $Y \sim \text{ZIP}(\pi, \lambda)$.
- Construct a family $\{f_\theta(y); \theta \in \mathbb{R}\}$ of distributions, indexed by θ, via exponential tilting:
$$f_\theta(y) \propto \exp(\theta y) f(y|\pi, \lambda) \ , \quad \theta \in \mathbb{R}.$$
- Each $f_\theta(y)$ remains a ZIP distribution with new parameters π_θ and λ_θ given respectively by
$$\pi_\theta = \frac{\pi}{\pi + (1-\pi)e^{\lambda(e^\theta - 1)}} \quad \text{and} \quad \lambda_\theta = \lambda e^\theta, \tag{1}$$
- We can set $\lambda \equiv 1$ without loss of generality, and π can then be interpreted as a baseline zero-inflation rate relative to a standard Poisson(1) distribution.

Zero-inflated Poisson exponential families (cont.)

MACQUARIE University
SYDNEY·AUSTRALIA

- For mathematical convenience, we are going to use the baseline odds of zero-inflation:
$$\nu = \pi/(1-\pi),$$
- Similar to CMP, we prefer to parametrize the distribution with the mean, so we set
$$\mu = E(Y_\theta) = (1-\pi_\theta)\lambda_\theta.$$
- We then write the distribution as $\text{ZIP}_\nu(\mu)$.

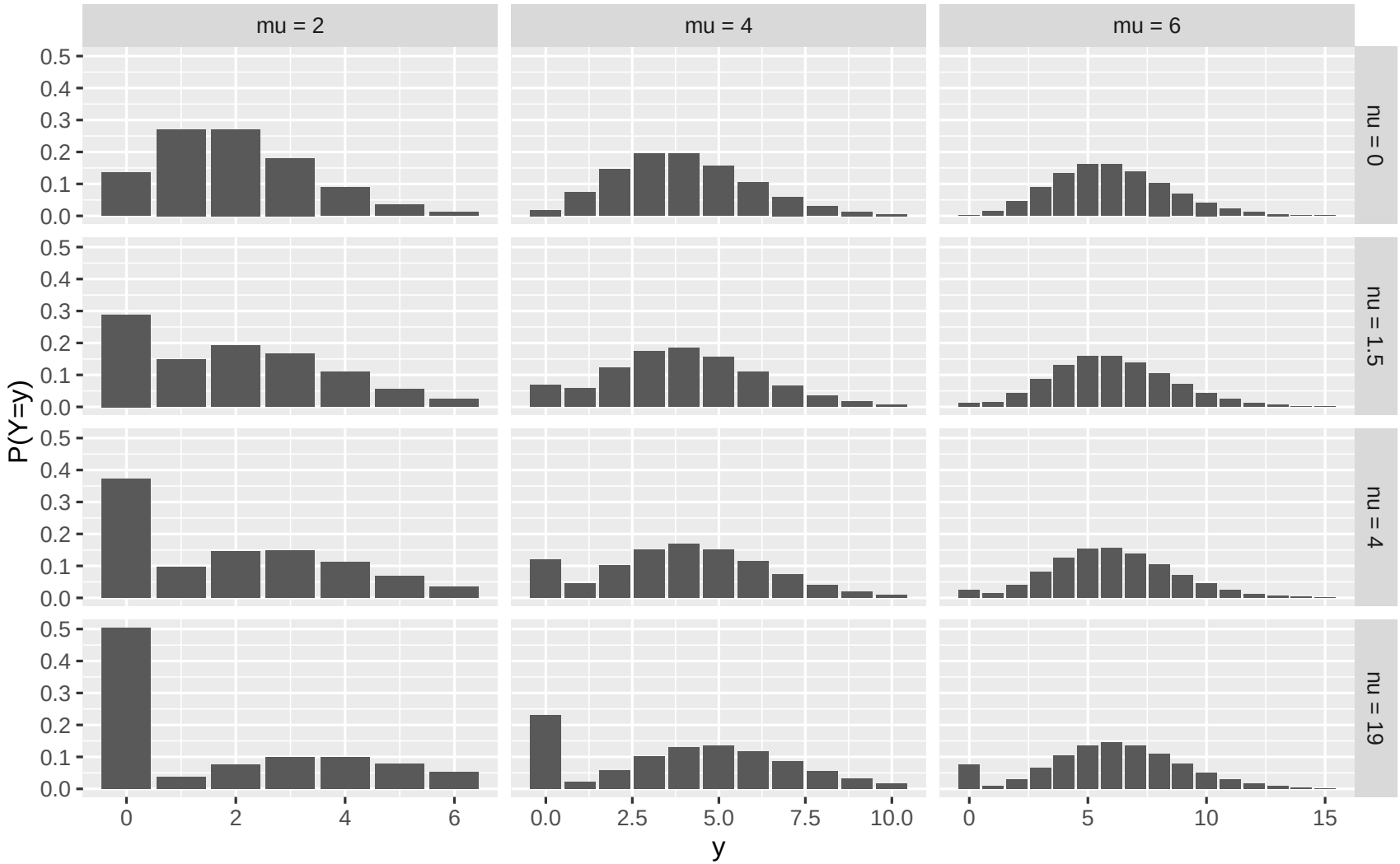
Behind the scene

- The canonical parameter θ can then be defined as a function of μ via
$$\theta = \log\left(\mu + W\left\{\nu e^{1-\mu}\mu\right\}\right),$$
where $W\{.\}$ is the Lambert-W function.

- Then the $\text{ZIP}_\nu(\mu)$ family has zero-inflation and Poisson rate parameters given respectively by
$$\pi(\mu) = \frac{W(\mu;\nu)}{\mu + W(\mu;\nu)} \qquad \text{and} \qquad \lambda(\mu) = \mu + W(\mu;\nu),$$
where $W(\mu;\nu) = W\{\nu e^{1-\mu}\mu\}$.

- The variance function of $\text{ZIP}_\nu(\mu)$ is
$$V(\mu,\nu) = \mu(1 + W(\mu;\nu)).$$

Some pmf of $\text{ZIP}_\nu(\mu)$



Regression models using ZIP_ν families

- Using the zero-inflated Poisson exponential families allows us to construct simple, interpretable regression models for zero-inflated data, such as a log-linear model.

- This can be considered a zero-inflated analogue to the log-linear Negative-Binomial and CMP models for dispersed counts.

- Given a response Y and a vector of covariates $X \in \mathbb{R}^q$, we can define a ZIP_ν regression model via

$$Y|X \sim \text{ZIP}_\nu(\mu(X^\top \beta)) \ , \ \text{for some } \nu \in [0, \infty)$$

where

$$E(Y|X) = \mu(X^\top \beta)$$

for some mean function $\mu(\cdot)$ and a vector of parameters $\beta \in \mathbb{R}^q$.

`bioChemist` dataset

- The dataset contains the number of articles produced by 915 graduate students in biochemistry during the last three years of their PhD, along with some information on the graduates
  - gender,
  - marital status,
  - the number of kids under 5,
  - how prestigious the department is,
  - mentor's publication record over the same period.

- We fitted the classical ZIP using the `zeroinfl()` in the `pscl` package as well as our algorithm.

	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
(intercept)	0.621	0.070	−0.576	0.330	0.357	0.069
genderF	−0.209	0.063	0.109	0.280	−0.229	0.063
married	0.105	0.071	−0.354	0.318	0.157	0.070
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046
mentor	0.018	0.002	−0.134	0.042	0.025	0.002
ν			—		0.572	0.088
log.Lik	−1605.0 (−1605.8)				−1613.1	
AIC	3230.0 (3225.5)				3238.4	
BIC	3278.2 (3259.2)				3267.3	

	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
(intercept)	0.621	0.070	−0.576	0.330	0.357	0.069
genderF	−0.209	0.063	0.109	0.280	−0.229	0.063
married	0.105	0.071	−0.354	0.318	0.157	0.070
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046
mentor	0.018	0.002	−0.134	0.042	0.025	0.002
ν			—		0.572	0.088
log.Lik	−1605.0 (−1605.8)				−1613.1	
AIC	3230.0 (3225.5)				3238.4	
BIC	3278.2 (3259.2)				3267.3	

How about interpretation?

	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046

- For the classical zero-inflated Poisson model, model interpretation requires two-steps.
- Each additional kid under 5 years old is associated with an increase in the log-odds of being in the subpopulation that *did not have the opportunity to produce a paper* of $0.219 \approx 24\%$ increase in odds.
- Given a graduate is in the other subpopulation that *have the opportunity to produce paper(s)* then each additional kid under 5 years old is associated with a decrease in the expected number of papers by a factor of $\exp(-0.143) = 0.87$, i.e. 13% decrease.

How about interpretation?

	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
kid5	−0.143	0.047	**0.219**	0.196	−0.190	0.046

- For the classical zero-inflated Poisson model, model interpretation requires two-steps.
- Each additional kid under 5 years old is associated with an increase in the log-odds of being in the subpopulation that *did not have the opportunity to produce a paper* of **0.219** $\approx$ 24% increase in odds.
- Given a graduate is in the other subpopulation that *have the opportunity to produce paper(s)* then each additional kid under 5 years old is associated with a decrease in the expected number of papers by a factor of $\exp(-0.143) = 0.87$, i.e. 13% decrease.

How about interpretation?

	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046

- For the classical zero-inflated Poisson model, model interpretation requires two-steps.
- Each additional kid under 5 years old is associated with an increase in the log-odds of being in the subpopulation that *did not have the opportunity to produce a paper* of $0.219 \approx 24\%$ increase in odds.
- Given a graduate is in the other subpopulation that *have the opportunity to produce paper(s)* then each additional kid under 5 years old is associated with a decrease in the expected number of papers by a factor of $\exp(-0.143) = 0.87$, i.e. 13% decrease.

How about interpretation? (cont.)

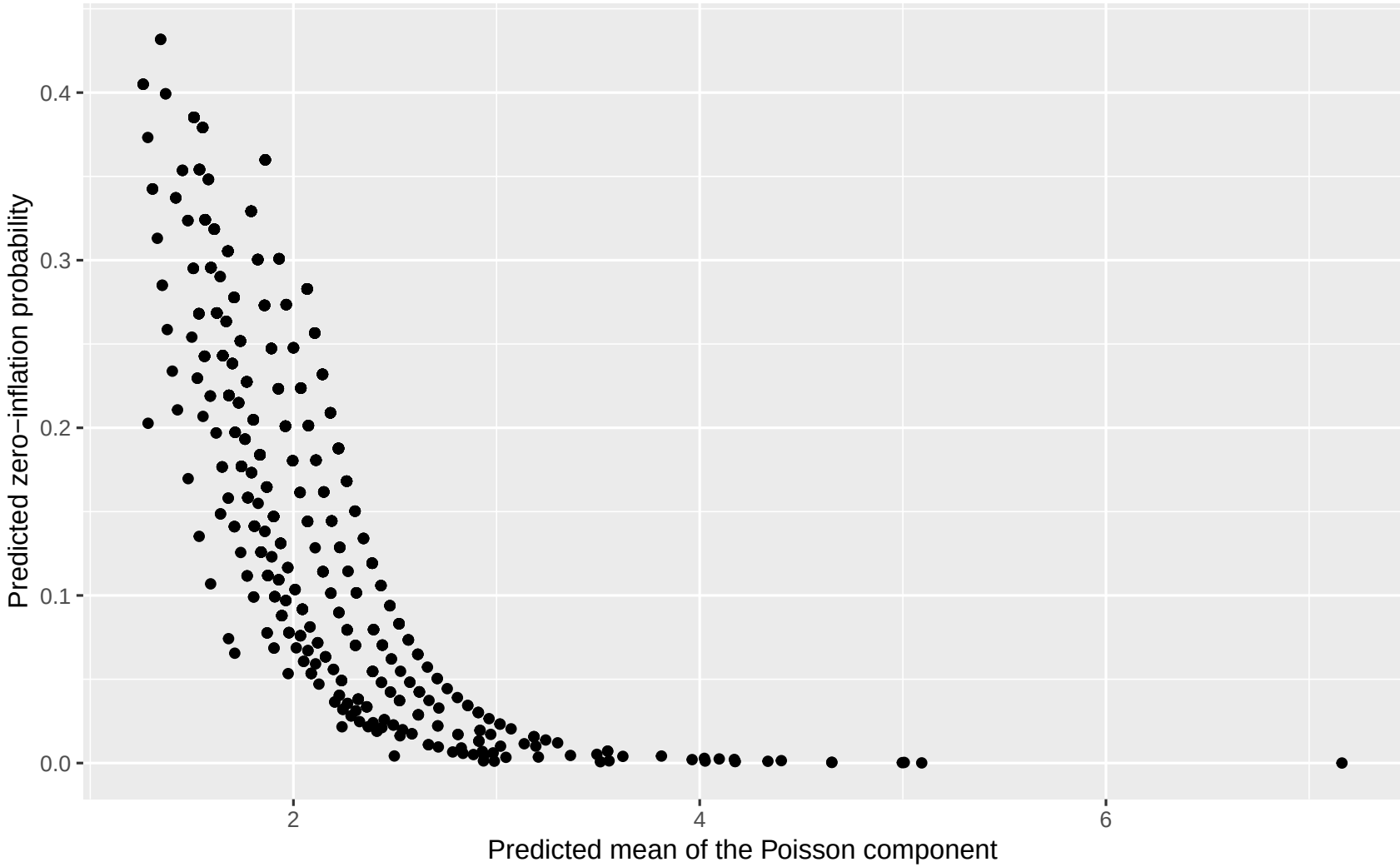
	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046

- The effect of each additional kid under 5 years old is a multiplicative factor of $\exp(-0.190) = 0.82$ to the expected number of papers produced.
- This value has already been adjusted for zero-inflation.

Predicting the zero-inflation component

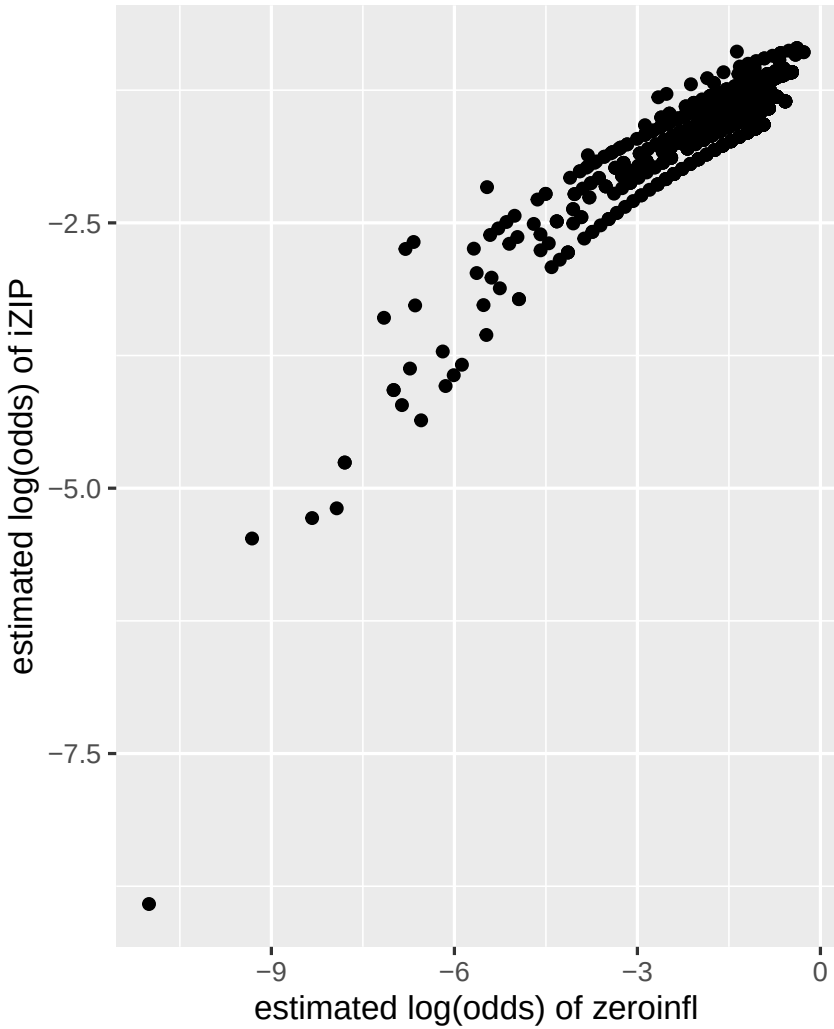
	ZIP model				ZIP_ν model	
	Poisson component		Bernoulli component			
coefficient	estimate	se	estimate	se	estimate	se
(intercept)	0.621	0.070	−0.576	0.330	0.357	0.069
genderF	−0.209	0.063	0.109	0.280	−0.229	0.063
married	0.105	0.071	−0.354	0.318	0.157	0.070
kid5	−0.143	0.047	0.219	0.196	−0.190	0.046
mentor	0.018	0.002	−0.134	0.042	0.025	0.002
ν			—		0.572	0.088

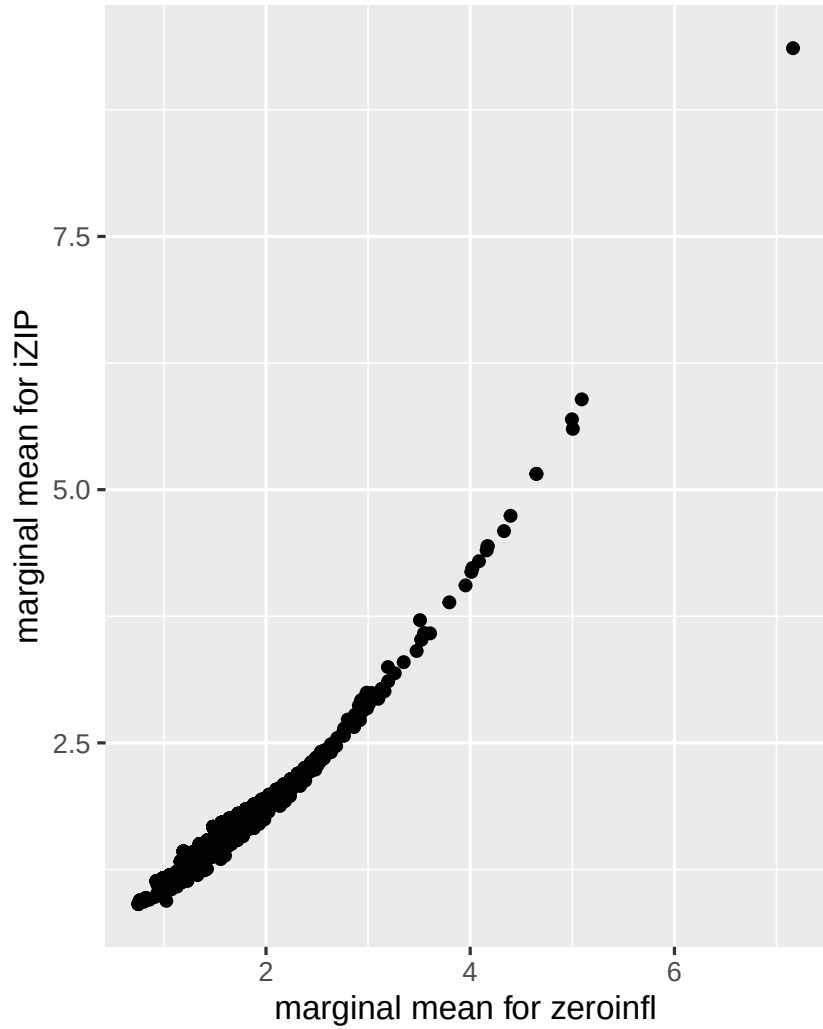
- Notice that all variables with a positive effect on the Poisson component of the classical ZIP model had a negative effect on the Bernoulli component.
- In other words, as the expected number of papers produced increases, the probability of being in the “do not have opportunity to write a paper" (i.e., zero-inflation) subpopulation tends to decrease, and vice versa.



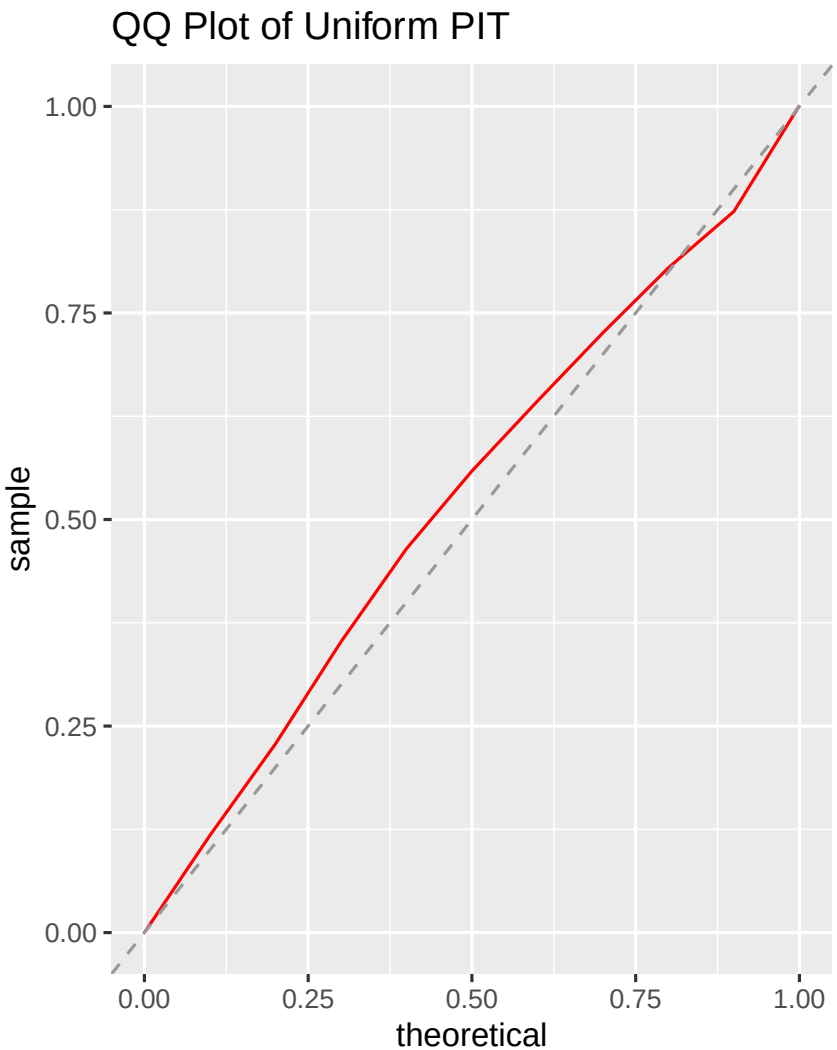
- The strong negative relationship here provides a clear example of how constant zero-inflation can be unrealistic in practice.
  - But this is the assumption used in some time series model for counts.

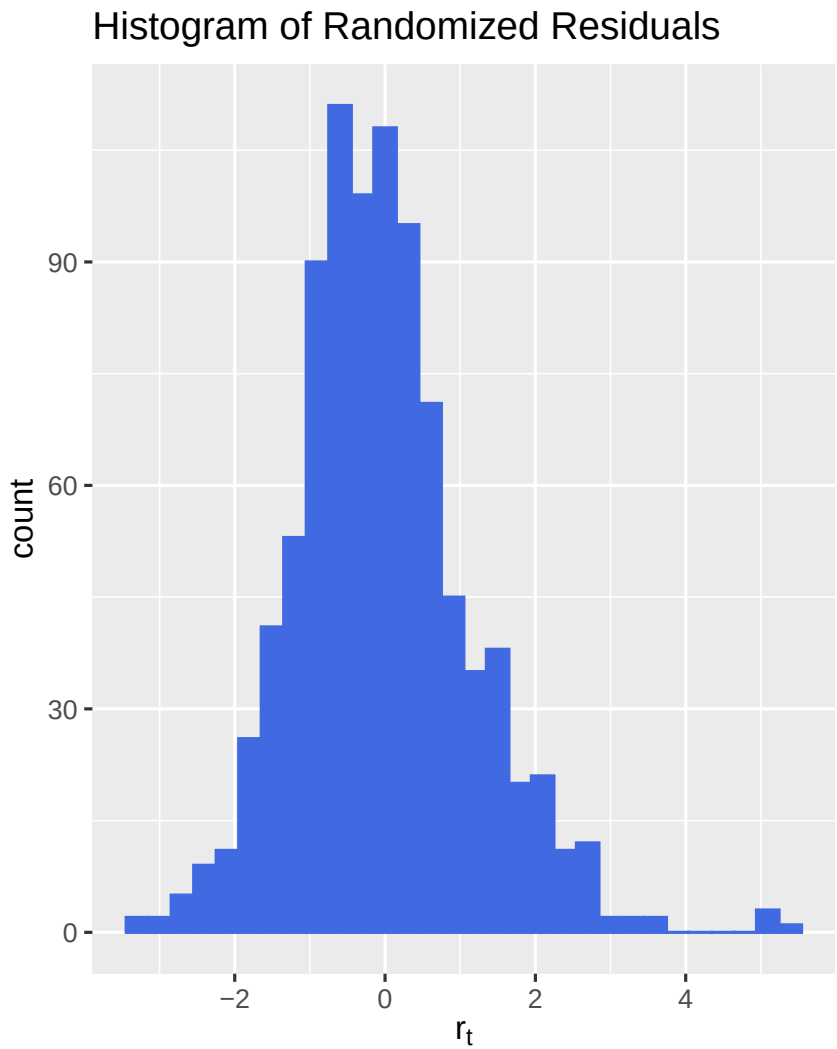
Predicting the subpopulation





Diagnostic


QQ Plot of Uniform PIT
sample
theoretical
0.00
0.25
0.50
0.75
1.00


Histogram of Randomized Residuals
count
r_t
0
30
60
90
−2
0
2
4

Recap

- The proposed model handles the zero-inflation here without the need to specify an explicit model for the zero-inflation process.
- In exchange of some goodness-of-fit, we gain some interpretability and able to carry out extra diagnostics test over the standard ZIP model.

Acknowledgement: 2020 NSW Branch Council

- Mr. Runze Alex Xu
- Dr. Boris Beranger
- Dr. Munir Hiabu
- Dr. Clara Grazian
- Dr. Nicole de la Mata
- Dr. Gordana Popovic
- Mr. Ardalan Mirzaei
- Dr. Luca Maestrini
- Dr. Mark Donoghoe
- Dr. Matias Quiroz
- Prof. Matt Wand

References I

Conway, R. W., and W. L. Maxwell. 1962. "A queuing model with state dependent service rates." New York :

Huang, Alan. 2017. "Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts." *Statistical Modelling* 17 (6): 359–80. https://doi.org/10.1177/1471082X17697749.

Lambert, Diane. 1992. "Zero-inflated poisson regression, with an application to defects in manufacturing." *Technometrics* 34 (1): 1–14. https://doi.org/10.1080/00401706.1992.10485228.

Pollock, Jeffrey. 2018. "CompGLM: Conway-Maxwell-Poisson GLM and Distribution Functions." https://cran.r-project.org/package=CompGLM.

Ribeiro, Eduardo E. 2020. *Cmpreg: Reparametrized Com-Poisson Regression Models*.

References II

Ribeiro, Eduardo E., Walmes M. Zeviani, Wagner H. Bonat, Clarice G. B. Demetrio, and John Hinde. 2020. "Reparametrization of COM–Poisson regression models with applications in the analysis of experimental data." *Statistical Modelling* 20 (5): 443–66. https://doi.org/10.1177/1471082X19838651.

Saez-Castillo, Antonio Jose, Antonio Conde-Sanchez, and Francisco Martinez. 2020. *DGLMExtPois: Double Generalized Linear Models Extending Poisson Regression*. https://CRAN.R-project.org/package=DGLMExtPois.

Sellers, Kimberly F., and Galit Shmueli. 2010. "A flexible regression model for count data." *Annals of Applied Statistics* 4 (2): 943–61. https://doi.org/10.1214/09-AOAS306.

Sellers, Kimberly, Thomas Lotze, and Andrew Raim. 2017. "COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression." https://cran.r-project.org/package=COMPoissonReg.

References III

Shmueli, G., T. P Minka, J. B Kadane, S. Borle, and P. Boatwright. 2005. *Applied Statistics* 54 (1): 127–42. https://doi.org/10.1111/j.1467-9876.2005.00474.x.