

Model comparison: Confirmatory and Exploratory analysis

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline





Principles for statistical ecology

Received: 23 January 2023 | Accepted: 13 November 2023

DOI: 10.1111/2041-210X.14270

REVIEW

Methods in Ecology and Evolution












Four principles for improved statistical ecology

Gordana Popovic¹ | Tanya Jane Mason^{2,3} | Szymon Marian Drobnik^{4,5} |
Tiago André Marques^{6,7} | Joanne Potts⁸ | Rocío Joo⁹ | Res Altwegg¹⁰ |
Carolyn Claire Isabelle Burns¹¹ | Michael Andrew McCarthy¹² | Alison Johnston¹³ |
Shinichi Nakagawa⁴ | Louise McMillan¹⁴ | Kadambari Devarajan^{15,16} |
Patrick Leo Taggart¹⁷ | Alison Wunderlich¹⁸ | Magdalena M. Mair^{19,20} |
Juan Andrés Martínez-Lanfranco²¹ | Malgorzata Lagisz⁴ | Patrice Pottier⁴

Poor practice

- ▶ Hypothesising after results are known (HARKing)
- ▶ Not reporting non-significant results
- ▶ Misinterpreting non-significant results
- ▶ Providing insufficient detail on methods and analysis
- ▶ P-hacking

Four principles

1.  Define a focused research question
 -  Plan sampling
 -  Plan analysis
2.  Develop a model
 -  Think about the distribution
 -  Generally consider data properties (e.g., dependence)
3.  Report methods in detail
 -  Report findings in detail
 -  Ensure reproducibility

A protocol for data exploration

Methods in Ecology and Evolution



Methods in Ecology and Evolution 2010, **1**, 3–14

doi: 10.1111/j.2041-210X.2009.00001.x

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur^{*1,2}, Elena N. Ieno^{1,2} and Chris S. Elphick³

Highly cited. Perhaps because this is a difficult subject?

A protocol: steps

- 1) Look for outliers
- 2) Check constant variance
- 3) Check normality
- 4) Consider excess zeros
- 5) Explore collinearity of covariates
- 6) What shape does $y = f(x)$ take?
- 7) Interactions
- 8) Account for study design

Typical steps in a real analysis

Before analysis

1. Determine research question
2. Collect data
3. *Plan analysis*

Analysis

- 1) Explore the data
- 2) Fit a model
- 3) Fit another model
- 4) **Determine which model is better**
- 5) Check model assumptions, if violated return to 3)
- 6) Report all results

What makes a good statistical model?

For prediction or for inference?

- ▶ A model is usually only good at one thing
 - ▶ Prediction
 - ▶ Inference
- ▶ But it can do both

What makes a good statistical model?

For prediction or for inference?

- ▶ A model is usually only good at one thing
 - ▶ Prediction
 - ▶ Inference
- ▶ But it can do both

This is a matter for **calibration**

What makes a good statistical model?

One that helps to answer your research question.

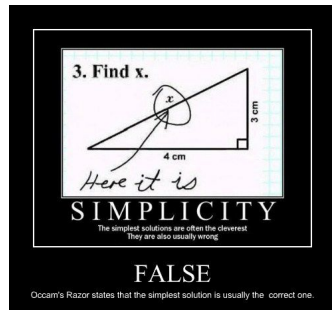
- ▶ Accurately represents the data generating process
- ▶ Not too difficult to interpret
- ▶ Is robust

Principle of Parsimony

The simplest explanation is often the correct one.

A simpler model might

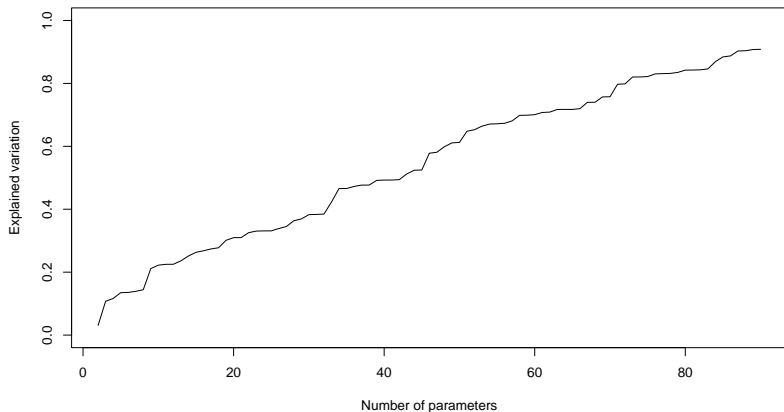
- ▶ Make fewer assumptions
- ▶ Be easier to interpret
- ▶ Be less prone to overfitting
- ▶ **Inadequately accommodate properties of data**



If two competing models fit the data equally well, we continue with the simpler model.

The problem of model complexity

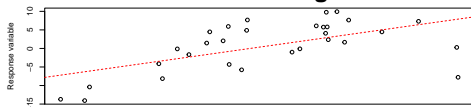
A model will always fit better if you add a parameter



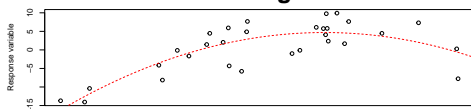
Model complexity



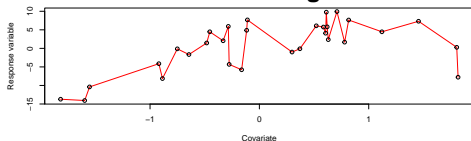
Underfitting



Just right



Overfitting



Example: Baseball game attendance

Attendance to baseball games



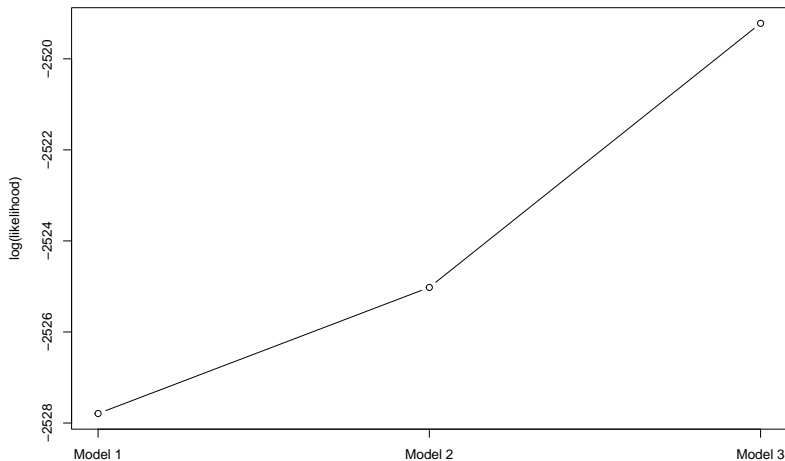
Baseball attendance: data from yesterday

Data from Cochran (2002)

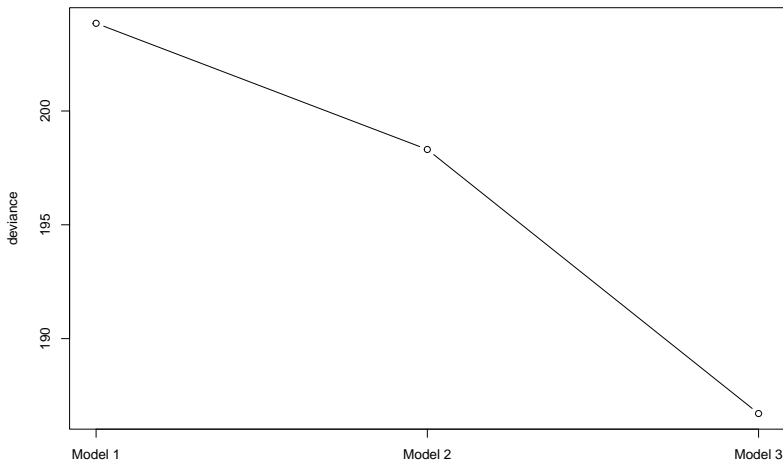
- ▶ 838 games
- ▶ Many variables: team, league, division, year, runs scores, wins
- ▶ Response variable: attendance

franchise	league	division	year	attendance	runs.scored	runs.allowed	wins	losses	games.behind
BAL	AL	EAST	69	1062069	779	517	109	53	0.0
BOS	AL	EAST	69	1833246	743	736	87	75	22.0
CLE	AL	EAST	69	619970	573	717	62	99	46.5
DET	AL	EAST	69	1577481	701	601	90	72	19.0
NYA	AL	EAST	69	1067996	562	587	80	81	28.5
WAS	AL	EAST	69	918106	694	644	86	76	23.0
CAL	AL	WEST	69	758388	528	652	71	91	26.0

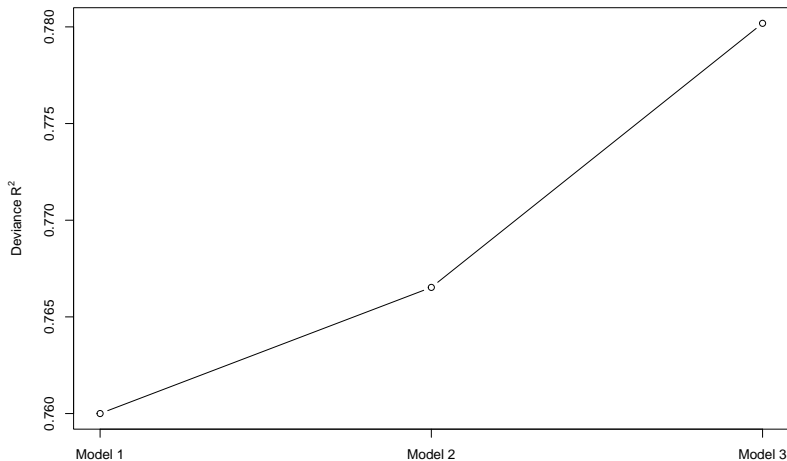
Baseball attendance: log-likelihoods



Baseball attendance: deviances



Baseball attendance: explained variance



Confirmatory vs. Exploratory analysis



Confirmatory: test hypothesis

Exploratory: find a good model



Usually indicated by the research question.

Hypothesis testing

- 1) Define null hypothesis
- 2) Define alternative hypothesis
- 3) Choose a test statistic (e.g., deviance in GLMs)
- 4) Calculate its distribution under the null hypothesis
- 5) Check if the observed test statistic falls within the null distribution
- 6) Accept/Reject

Likelihood ratio test

Is improved fit due to noise or is the alternative model actually better?

Procedure

- ▶ Fit two models: M_0 with k parameters and M_1 with r
- ▶ Calculate likelihood ratio $\Lambda = \log \left(\frac{\mathcal{L}(\mathbf{y}; \Theta_0)_{M_0}}{\mathcal{L}(\mathbf{y}; \Theta_1)_{M_1}} \right)$
- ▶ $\mathcal{L}(\mathbf{y}; \Theta_0)_{M_0} \leq \mathcal{L}(\mathbf{y}; \Theta_1)_{M_1}$
- ▶ $-2\Lambda \sim \chi^2(k_1 - k_0)$ under the null
- ▶ $p \geq 0.05$ difference in likelihood is due to sampling

LRT approximation assumptions

- ▶ $n \rightarrow \infty$
- ▶ Θ_0 contained in Θ_1 : nested models
- ▶ The true parameter is in the interior of the parameter space
- ▶ Model is “identifiable”
- ▶ Hessian matrix is sufficiently close to the Fisher information
- ▶ y_i are independent

These assumptions may fail, especially in models more complex than GLMs

Alternatively: LRT by simulation.

LRT and deviance

$$\Lambda = \frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}}_{M0}) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_{M1})}{\phi} \quad (1)$$

Example: Lizard habitat preference

- ▶ Data originally by Schoener (1970)
- ▶ Counts of two species of lizard in Jamaica



Figure 2: wikipedia.org: grahami



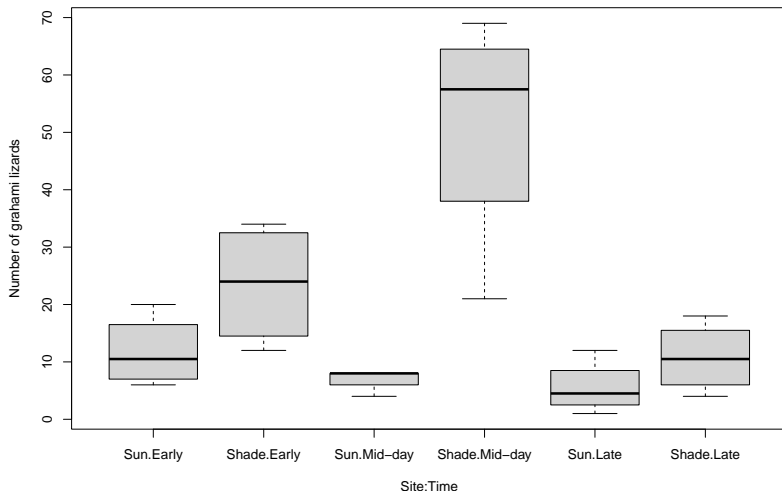
Figure 3: wikipedia.org: opalinus

Lizards: the data

Dataset also covered in McCullagh and Nelder (1989)

	Site	Diameter	Height	Time	grahami	opalinus
1	Sun	D ≤ 2	H < 5	Early	20	2
2	Sun	D ≤ 2	H < 5	Mid-day	8	1
3	Sun	D ≤ 2	H < 5	Late	4	4
4	Sun	D ≤ 2	H ≥ 5	Early	13	0
5	Sun	D ≤ 2	H ≥ 5	Mid-day	8	0
6	Sun	D ≤ 2	H ≥ 5	Late	12	0
7	Sun	D > 2	H < 5	Early	8	3
8	Sun	D > 2	H < 5	Mid-day	4	1
9	Sun	D > 2	H < 5	Late	5	3
10	Sun	D > 2	H ≥ 5	Early	6	0
12	Sun	D > 2	H ≥ 5	Late	1	1
13	Shade	D ≤ 2	H < 5	Early	34	11
14	Shade	D ≤ 2	H < 5	Mid-day	69	20
15	Shade	D ≤ 2	H < 5	Late	18	10
16	Shade	D ≤ 2	H ≥ 5	Early	31	5
17	Shade	D ≤ 2	H ≥ 5	Mid-day	55	4
18	Shade	D ≤ 2	H ≥ 5	Late	13	3
19	Shade	D > 2	H < 5	Early	17	15
20	Shade	D > 2	H < 5	Mid-day	60	32
21	Shade	D > 2	H < 5	Late	8	8

Example: Lizards



Lizards: interaction

Null hypothesis: There is no interaction of Time and Site

Fit M_0 :

```
nmodel <- glm(cbind(ghahami, opalinus)~Time+Site,
               data = lizards, family="binomial")
```

Alternative hypothesis: There is an interaction of Time and Site

Fit M_1 :

```
amodel <- update(nmodel, formula = . ~ Time*Site)
```

Lizards: interaction

```
##
## Call:
## glm(formula = cbind(ghrami, opalinus) ~ Time + Site + Time:Site,
##      family = "binomial", data = lizards)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.24071    0.47037   4.764 1.9e-06 ***
## TimeMid-day       0.06188    0.87818   0.070  0.9438
## TimeLate        -1.22911    0.62585  -1.964  0.0495 *
## SiteShade       -1.16315    0.51296  -2.268  0.0234 *
## TimeMid-day:SiteShade 0.07270    0.91343   0.080  0.9366
## TimeLate:SiteShade  0.69387    0.70486   0.984  0.3249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

Lizards: LRT

```
(Lambda <- 2*(logLik(amodel)-logLik(nmodel)))
```

```
## 'log Lik.' 1.098943 (df=6)
```

```
k <- attr(logLik(amodel),"df")-attr(logLik(nmodel),"df")
pchisq(Lambda, k,lower.tail=FALSE)
```

```
## 'log Lik.' 0.5772548 (df=6)
```

We reject the alternative hypothesis.

Lizards: LRT simulation

```

Lambdas <- NULL
for(i in 1:1000){
  ynew <- as.matrix(stats::simulate(nmodel))
  nmodel2 <- glm(ynew~Time+Site,
                 data = lizards, family="binomial")
  amodel2 <- update(nmodel2, formula = .~Time*Site)
  # Store test statistic
  Lambdas <- c(Lambdas,
               2*(logLik(amodel2)-logLik(nmodel2)))
}
# if <0.05 our test statistic in the tail.
sum(Lambdas>Lambda)/1000
  
```

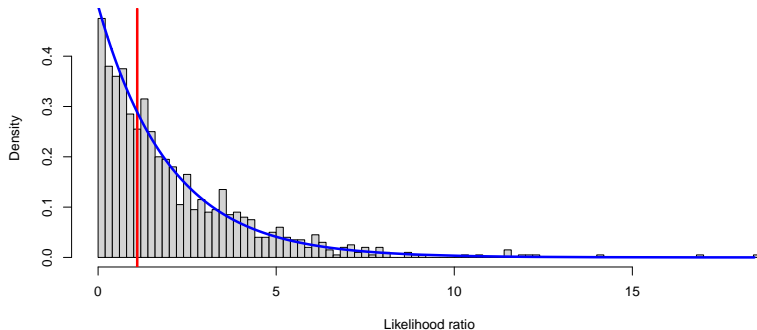
```
## [1] 0.598
```


Lizards: LRT simulation

Red: Observed statistic

Blue: $\chi^2(2)$

Histogram of Lambdas



Information criteria

A different paradigm:

Find the best model amongst a set of models.

Best:

- ▶ Penalise complexity (number of parameters)
- ▶ By fit (likelihood)

Most commonly:

- 1) AIC: Akaike's Information Criterion (Akaike 1974)
- 2) BIC: Bayesian Information Criterion (Schwarz 1978)

Lower = better

Akaike's Information Criterion

$$\text{AIC} = -2\mathcal{L}(\mathbf{y}; \Theta) + 2k \quad (2)$$

- ▶ Penalizes model complexity
- ▶ (approximately) Measures information loss to the true data generating process
- ▶ Asymptotically

AIC tends to select too complex models with little data. Finite sample correction (Sugiura 1978):

$$\text{AIC}_c = \text{AIC} + \frac{2k(k+1)}{n-k-1} \quad (3)$$

Find the model that predicts best

Bayesian Information Criterion

$$\text{BIC} = -2\mathcal{L}(\mathbf{y}; \Theta) + k \log(n) \quad (4)$$

So the penalty is different.

Find the model closest to the "true" model

Connection of AIC and LRT

Rule of thumb: difference of 2 points means a model is better

$$\begin{aligned}
 \Delta\text{AIC} &= \text{AIC}_{M_1} - \text{AIC}_{M_0} \\
 &= 2\mathcal{L}(\mathbf{y}; \Theta_0) - 2\mathcal{L}(\mathbf{y}; \Theta_1) + 2k_1 - 2k_0 \quad (5) \\
 &= -2\Lambda + 2(k_1 - k_0)
 \end{aligned}$$

So AIC with a rule of $= 2$ can be seen as a more liberal LRT (Sutherland et al. 2023)

The cult of (A)IC

Presentation by Mark Brewer

"Always use (A)IC for model comparison"

My perspectives

- ▶ Use common sense
- ▶ Do not blindly test all models ("dredging")
- ▶ Use model comparison techniques in moderation



Don't take the "best" model paradigm too seriously

Freedman's paradox

Just by chance, predictors with no relationship to the response will be selected.

Lizards: interaction selection

```
##           df      AIC
## nmodel    4 119.7237
## amodel     6 122.6248
```

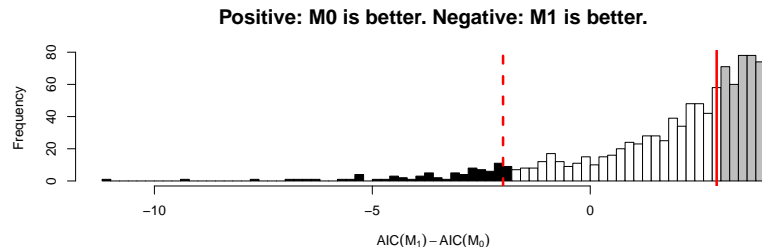
```
##           df      AICc
## nmodel    4 121.9460
## amodel     6 127.8748
```

```
##           df      BIC
## nmodel    4 124.2657
## amodel     6 129.4378
```


ΔAIC by simulation under the simpler model

Solid: observed ΔAIC

Dashed: $-2 \Delta AIC$



Black: We would incorrectly select the more complex model (7%)

White: We would consider the models equivalent (34%)

Gray: Under a $2 \Delta AIC$ rule we would select the more simpler model (59%)

Δ AIC by simulation under the additive model

Models with a more extreme test statistic: 0.607

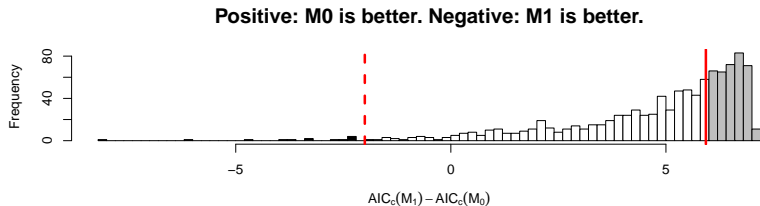
Models with more than 2 AIC difference: 0.591

93% of the time we would have selected the simpler model

ΔAIC_c by simulation under the simpler model

Solid: observed ΔAIC_c

Dashed: $-2 \Delta AIC_c$



Black: We would incorrectly select the more complex model (1%)

White: We would consider the models equivalent (40%)

Gray: Under a $2 \Delta AIC$ rule we would select the more simpler model (89%)

Δ AICc by simulation under the simpler model

```
## Models with more than 2 AIC difference: 0.591
```

```
## Models with more than 2 AICc difference: 0.885
```

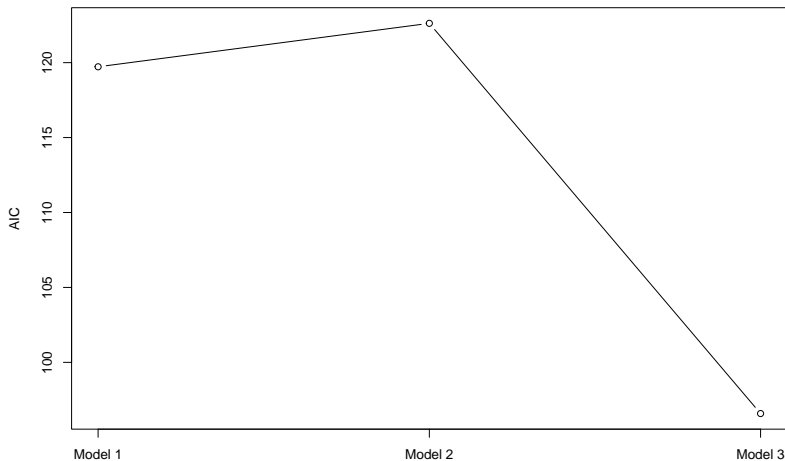
99% of the time we would select the simpler model.

Lizards: interaction and height

```

model1 <- glm(cbind(grahami, opalinus)~Time+Site,
              data = lizards, family="binomial")
model2 <- update(model1, formula = . ~ Time*Site)
model3 <- update(model2, formula = ~.+Height)
  
```

Lizards: AIC plotted



Omitted variable bias

Occurs if we omit a variable, i.e., we have the model:

$$g\{\mathbb{E}(y_i|x_{i1}, x_{i2})\} = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 \quad (6)$$

but we fit without x_{i2} . The consequence is especially clear for linear models:

$$y_i = \alpha + x_{i1}\beta_1 + \epsilon_i, \quad \text{with } \epsilon_i = x_{i2}\beta_2 + \epsilon_i^* \quad (7)$$

- ▶ ϵ_i may be correlated with x_{i1}
- ▶ Residual variance changes with x_{i1}
- ▶ Causes bias in parameter estimates and incorrect standard errors

Multimodel inference

What to do when you have multiple models that are equally good?

- ▶ The simpler model
- ▶ Do not “model average” for inference
- ▶ Report multiple models

Summary

- ▶ Hypothesis testing and information criteria: two different paradigms
- ▶ Do not mix them
- ▶ Do not fall for “the best model”
- ▶ The full model is often a good model
- ▶ Keep things simple
- ▶ See also [Murtaugh \(2014\)](#) and Burnham and Anderson, and a lot of others