Evaluating Frequentist Model Averaged Confidence Intervals

Paul Kabaila, Alan Welsh* & Waruni Abeysekera

*Mathematics Sciences Institute
Australian National University

Department of Mathematics and Statistics
La Trobe University

Model Averaging

Attempts to incorporate model uncertainty into inference by using multiple models $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ for a fixed, finite J.

In frequentist model averaging, we **define a weight** $\hat{w}_j$ for each model, usually by exponentiating minus AIC, BIC or some other information criterion.

- Buckland et al. (1997) and Burnham & Anderson (2002) **averaged point estimates and tried to estimate the standard error of the estimator**. The distribution theory is not correct (Claeskens & Hjort, 2008, p207) but the coverage is fine in some simulations.
- The Hjort & Claeskens (2003) approach is essentially the **same as constructing the standard confidence interval from the full model** (Kabaila & Leeb, 2006; Wang & Zou, 2013).
- Fletcher and Turek (2011) and Turek and Fletcher (2012) **averaged the equations defining the endpoints** of confidence intervals.

Fletcher-Turek Profile Intervals (MPI)

A $1-\alpha$ level profile likelihood confidence interval for a parameter θ in a model $\mathcal{M}_j$ is obtained by computing the signed-root log-likelihood ratio for θ under $\mathcal{M}_j$,

$$R_j(\theta) = \text{sgn}(\hat{\theta}_j - \theta)[2\{\ell_j(\hat{\theta}_j, \hat{\lambda}_j) - \ell_j(\theta, \hat{\lambda}_{j\theta})\}]^{1/2},$$

and then solving for the lower and upper endpoints of the interval the two equations obtained by equating the normal cumulative distribution function evaluated at the signed-root log likelihood ratio to $1-\alpha/2$ and $\alpha/2$, respectively.

When we have models $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, MPI confidence intervals for θ, with nominal coverage $1-\alpha$, are obtained by solving for the endpoints a weighted average of the respective endpoint equations for each model:

$$\sum_{j=1}^{J} \hat{w}_j \Phi\{R_j(\theta)\} = \alpha/2 \quad \text{and} \quad \sum_{j=1}^{J} \hat{w}_j \Phi\{R_j(\theta)\} = 1-\alpha/2.$$

Turek-Fletcher Tail-Area Intervals (MATA)

A $1-\alpha$ level tail area confidence interval for a parameter θ in a model $\mathcal{M}_j$ is obtained in the same way by replacing the signed-root log-likelihood ratio by the t ratio

$$T_j(\theta) = (\hat{\theta}_j - \theta)/\text{se}(\hat{\theta}_j)$$

and solving the two equations obtained by equating $G_{\nu_j}\{T_j(\theta)\}$ to $1-\alpha/2$ and $\alpha/2$, where G_{ν_j} is the cumulative distribution function of the distribution of $T_j(\theta)$ under model $\mathcal{M}_j$ (often the Student t distribution with ν_j degrees of freedom).

When we have models $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, MATA confidence intervals for θ, with nominal coverage $1-\alpha$, are obtained by solving for the endpoints a weighted average of the respective endpoint equations for each model:

$$\sum_{j=1}^{J} \hat{w}_j G_{\nu_j}\{T_j(\theta)\} = \alpha/2 \quad \text{and} \quad \sum_{j=1}^{J} \hat{w}_j G_{\nu_j}\{T_j(\theta)\} = 1-\alpha/2.$$

Evaluation

Compare the coverage and expected length properties of intervals. Actually, we should evaluate **minimum coverage probabilities** and **maximum expected lengths** to characterise performance over unknown nuisance parameters.

- Simulations cover only a limited set of values of the unknown nuisance parameters and the conclusions apply only to these settings.
- Variability in simulation results complicates finding bounds on coverage or expected length, particularly when there are a large number of nuisance parameters.
- Consider **simple cases** (e.g. two nested regression models) where we can do exact calculations to evaluate the properties of the confidence intervals both in particular settings and uniformly over unknown nuisance parameters.

Cloud seeding example

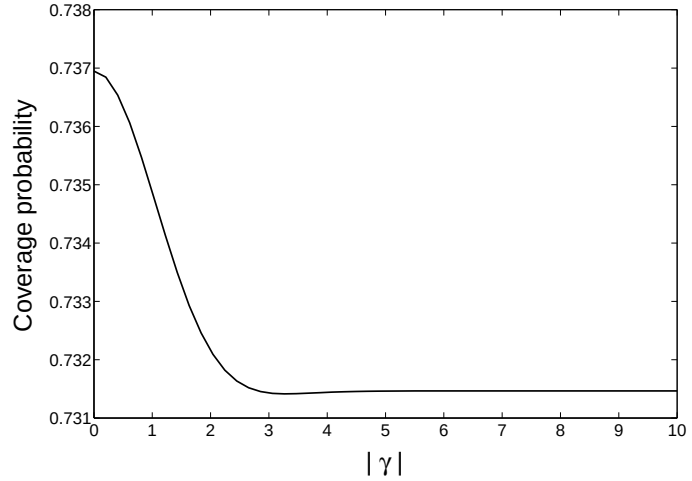
Biondini, Simpson and Woodley (1997): $n = 33$ observations from an experiment to compare seeding against a control treatment.

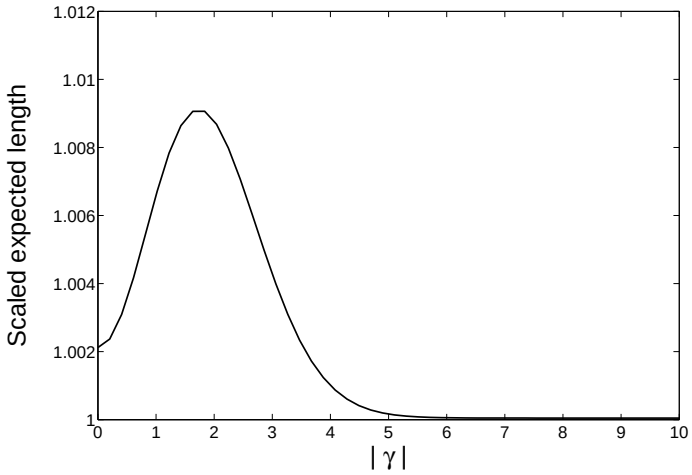
The **response variable** is the floating target rainfall volume.

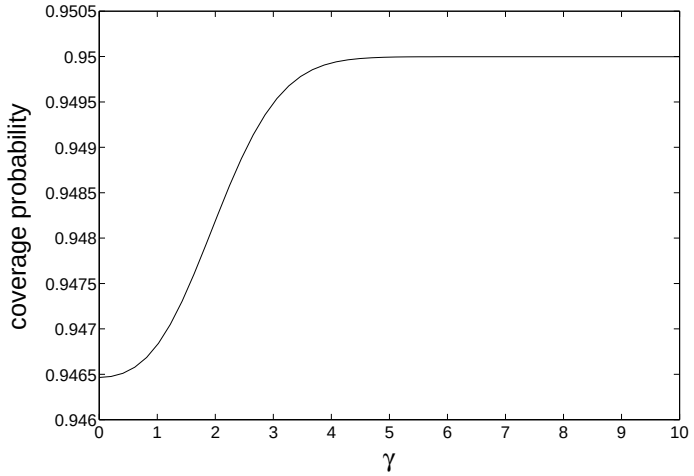
Then **explanatory variables** are the treatment indicator, 5 main effects (including seedability), 5 squared effects and the 10 interactions between the five main effects so that p, the dimension of the regression parameter vector, is 22.

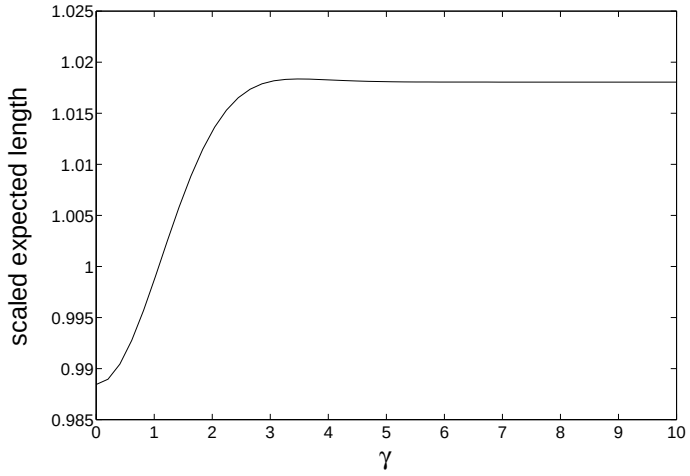
The goal is to construct a 95% confidence interval for θ, the **expected response when cloud seeding is used minus the expected response under random control when all the other explanatory variables are the same**.

$\mathcal{M}_2$ is the full model ($p = 22$) and $\mathcal{M}_1$ is the submodel excluding the squared seedability term. The expected length is scaled by the expected length of the standard confidence interval with the minimum coverage.



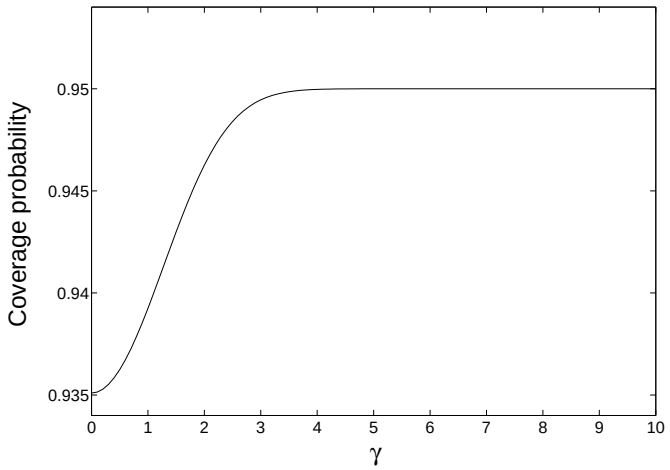






For MPI, the coverage probability is close to 0.7315 for all γ rather than the nominal 0.95 and the scaled expected length is close to one for all γ. Therefore, **MPI is actually similar to the standard 0.7315 confidence interval for θ**. (For the cloud seeding example, the poor minimum coverage of MPI is due to the value of $p/n = 2/3$ not being small.)

For MATA, the coverage probability of MATA is close to 0.95 for all γ with a minimum coverage probability 0.9465 and the scaled expected length is close to one for all γ. Therefore, **MATA is similar to the standard 0.95 confidence interval for θ under $\mathcal{M}_2$**.



MPI is uniformly worse and **MATA is slightly better in coverage** than the **confidence interval constructed after selecting between models $\mathcal{M}_1$ and $\mathcal{M}_2$ the model with smaller AIC and ignoring the selection process.**

Conclusion

An ideal confidence interval should have **minimal coverage equal to its nominal coverage** and, to show a benefit of model selection, have **scaled expected length** that

(a) is substantially less than 1 under $\mathcal{M}_1$;

(b) has a maximum value that is not much larger than 1; and

(c) is close to 1 if the data happens to strongly contradict the model $\mathcal{M}_1$.

This is evidently difficult to achieve.

Performing well in a simple evaluation situation does not mean that a model averaging procedure will always perform well; we also need to explore other situations, such as other models.

Details and references in Kabaila et al (2015) *Scand. J. Statist.* —DOI: 10.1111/sjos.12163

Model averaging over two regression models

Model $\mathcal{M}_2$: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is a random n-vector of responses, $\mathbf{X}$ is a known $n \times p$ model matrix with $p < n$ linearly independent columns, $\boldsymbol{\beta}$ is an unknown p-vector parameter and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, with σ^2 an unknown positive parameter.

Model $\mathcal{M}_1$: $\mathcal{M}_2$ with $\tau = \mathbf{c}^\top\boldsymbol{\beta} - t = 0$, where $\mathbf{c}$ is a specified nonzero p-vector that is linearly independent of $\mathbf{a}$ and t is a specified number.

Parameter of interest: $\theta = \mathbf{a}^\top\boldsymbol{\beta}$, where $\mathbf{a}$ is a specified nonzero p-vector.

Estimators: Let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$ and $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$ be the usual unbiased estimator of σ^2. Set $\hat{\theta} = \mathbf{a}^\top\hat{\boldsymbol{\beta}}$ and $\hat{\tau} = \mathbf{c}^\top\hat{\boldsymbol{\beta}} - t$. Define $v_\theta = \mathrm{Var}(\hat{\theta})/\sigma^2 = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a}$ and $v_\tau = \mathrm{Var}(\hat{\tau})/\sigma^2 = \mathbf{c}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{c}$.

Important quantities: the known correlation $\rho = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{c}/(v_\theta v_\tau)^{1/2}$ between $\hat{\theta}$ and $\hat{\tau}$ and the scaled unknown parameter $\gamma = \tau/\left(\sigma v_\tau^{1/2}\right)$.