# Model Selection Tutorial 2: Problems With Using AIC to Select a Subset of Exposures in a Regression Model

Daniel F. Schmidt and Enes Makalic

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

Work in Progress
15th November 2011

# Content

# Overview of Talk

- Testing for association between exposures and outcomes
  - Can be framed as model selection
- Nest and non-nested regression problems
- Akaike's Information Criterion
  - Properties
- Problems with AIC in non-nested setting
- Alternatives

# Problem Setting

- The following is common in epidemiological studies
- We have a group of $n$ subjects
- On each subject $i$ we have measured ...
  - an outcome of interest, $y_i$
  - $q < n$ exposure variables, $x_{i,1}, \ldots, x_{i,q}$

# Problem Setting

- The following is common in epidemiological studies
- We have a group of $n$ subjects
- On each subject $i$ we have measured ...
  - an outcome of interest, $y_i$
  - $q < n$ exposure variables, $x_{i,1}, \ldots, x_{i,q}$
- We wish to explore the relationship between exposures and outcome
- A common tool is linear regression

# Linear Regression Model (1)

- Linear regression model for explaining data $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ is the full design matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ are the unknown parameter coefficients
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ are i.i.d. Gaussian variates

# Linear Regression Model (1)

- Linear regression model for explaining data $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ is the full design matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ are the unknown parameter coefficients
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ are i.i.d. Gaussian variates
- Only a subset of exposures $\mathbf{X}$ are associated with $\mathbf{y}$
- Task: Determine which exposures, if any, are associated with $\mathbf{y}$

# Linear Regression Model (2)

- Let $\gamma \subset \{1, 2, \ldots, q\}$ denote which exposures are in design submatrix $\mathbf{X}_\gamma$
- Linear model indexed by $\gamma \in \Gamma$

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - Set of all candidate subsets $\Gamma$
  - $\mathbf{X} = (\mathbf{x}_{\gamma_1}, \mathbf{x}_{\gamma_2}, \ldots, \mathbf{x}_{\gamma_{|\gamma|}})$ is the design sub-matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{|\gamma|})'$ is the unknown parameter vector
  - Total number of unknown parameters is $k_\gamma = |\gamma| + 1$

# Linear Regression Model (2)

- Let $\gamma \subset \{1, 2, \ldots, q\}$ denote which exposures are in design submatrix $\mathbf{X}_\gamma$
- Linear model indexed by $\gamma \in \Gamma$

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \tau \mathbf{I}_n)$$

  - Set of all candidate subsets $\Gamma$
  - $\mathbf{X} = (\mathbf{x}_{\gamma_1}, \mathbf{x}_{\gamma_2}, \ldots, \mathbf{x}_{\gamma_{|\gamma|}})$ is the design sub-matrix
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{|\gamma|})'$ is the unknown parameter vector
  - Total number of unknown parameters is $k_\gamma = |\gamma| + 1$
- Example
  - $q = 10$, $\gamma = \{2, 3, 6, 10\}$, $\mathbf{X}_\gamma = (\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_{10})$
  - $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' \in \boldsymbol{\Theta}_\gamma$
  - $k_\gamma = 5$

# Nested Models

- Concepts of nested and non-nested sets of models are *very* important
- We can partition the complete set of models $\Gamma$ into subsets

$$\Gamma = \Gamma_1 \cup \Gamma_2 \ldots ...$$

  where $\Gamma_j$ is the set of all models with $j$ free parameters
- A sequence of models is nested iff :
  - There is only one model with $j$ parameters, for all $j$
  - Models with $k$ parameters can exactly represent all models with $j < k$ parameters

# Nested Models, Example

- Polynomial regression
  - Consider predicting outcome variable, say $y_i = \text{height}_i$, from an exposure variable, say $\text{age}_i$
  - Let

$$\begin{aligned} x_{i,1} &= 1 \\ x_{i,2} &= \text{age}_i \\ x_{i,3} &= \text{age}_i^2 \\ x_{i,4} &= \text{age}_i^3 \end{aligned}$$

    and so on ...
  - Frame as problem of order selection

$$\Gamma_0 = \{\}, \Gamma_1 = \{1\}, \Gamma_2 = \{1,2\}, \Gamma_3 = \{1,2,3\}, \Gamma_4 = \{1,2,3,4\}$$

$\Rightarrow$ The above forms a nested sequence

# Nested Models, Example

- By setting relevant $\beta$ coefficients to zero ...
  - ... a linear model can represent a constant model
  - ... a quadratic can represent a linear and a constant model
  - ... and so on
  - All models can represent the empty model (no coefficients)
- e.g., a cubic

$$y_i = \beta_4 x_i^3 + \beta_3 x_i^2 + \beta_2 x_i + \beta_1$$

with $\beta_4 = \beta_3 = 0$ is effectively a linear model

# Non-Nested Models, Example

- Alternatively frame the problem as selecting individual polynomial terms

$$
\begin{aligned}
\Gamma_0 &= \{\} \\
\Gamma_1 &= \{1\}, \{2\}, \{3\}, \{4\} \\
\Gamma_2 &= \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\} \\
\Gamma_3 &= \{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\} \\
\Gamma_4 &= \{1,2,3,4\}
\end{aligned}
$$

  $\Rightarrow$ The above is *not* a nested model class. There are ...

  - 4 models with 1 parameter
  - 6 models with 2 parameters
  - 4 models with 3 parameters

- e.g., model $\{3,4\}$ *cannot* represent models $\{1\}$ or $\{2\}$

# Epidemiology

- Nested regression modelling requires a naturally ordering of the exposures
- In epidemiology we usually have no such ordering
  $\Rightarrow$ we are usually working in a non-nested setting !
- For example, we may have measured a subject's
  - Age
  - Sex
  - Marital Status
  - Smoking Status
  - Genotypes
  - etc...
- What is a "natural" ordering for those variables ?

# Model Selection

- Selection of exposures is a specific instance of the problem of model selection
- We have observed data $\mathbf{y} = (y_1, \ldots, y_n)$
- Have a candidate set, $\Gamma$, of models to explain the data
  - Some models more complex than others
  - Want to find model that justifiably best fits the data
- More complex models (more parameters) usually fit better
  - Is the extra fit warranted, or due to random variation ?
- Use model selection criterion to choose between models
  - One method is Akaike Information Criterion (AIC)

# Statistical Models

- The word "model" denotes a conditional probability distribution, $p(\mathbf{y}|\boldsymbol{\theta}_\gamma)$
  - $\gamma$ is a symbol denoting a model from the set $\Gamma$
  - $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$ are the parameters associated with the model
- e.g., linear regression model, assuming normal noise, with $\gamma = \{2, 5\}$ specifies

$$y_i|x_{i,2},\, x_{i,5},\, \boldsymbol{\theta}_\gamma \sim N\left(\beta_1 x_{i,2} + \beta_2 x_{i,5},\, \tau\right)$$

where $\boldsymbol{\theta}_\gamma = (\beta_1, \beta_2, \tau)$, and $N(\cdot)$ is the normal distribution

# The "true" model

- The final concept we need is that of the "true" model
- Let
  - $\gamma_* \in \Gamma$ denote the true model
  - $\boldsymbol{\theta}_*$ denote the true parameters in model $\gamma_*$
- By assuming a "true" model, we are assuming the data is generated by the conditional distribution represented by this model
- This is an important assumption for many model selection criterion

# Content

# Fitting Models

- Use maximum likelihood to fit models to data

$$\hat{\boldsymbol{\theta}}_\gamma = \underset{\boldsymbol{\theta}_\gamma \in \Theta_\gamma}{\arg\max} \{p(\mathbf{y}|\boldsymbol{\theta}_\gamma)\}$$

- Need to measure how close a fitted model is to the truth
  - One possibility is Kullback–Leibler (KL) divergence
  - Let

$$d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma) = 2\mathrm{E}_{\boldsymbol{\theta}_*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)\right]$$

  - The KL divergence is given by

$$2\Delta(\boldsymbol{\theta}_*||\hat{\boldsymbol{\theta}}_\gamma) = \underbrace{d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma)}_{cross\ entropy} - \underbrace{d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*)}_{entropy}$$

- Ideally choose the model with the smallest KL divergence
  $\Rightarrow$ Relies on knowledge of the truth

# Akaike's Information Criterion (AIC) (1)

- Akaike's idea was to *estimate* the expected KL divergence for each model
  $\Rightarrow$ Select model with lowest expected KL divergence
- It turns out that

$$-2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)$$

  acts as a biased estimate of the cross-entropy risk

$$\mathrm{E}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_\gamma)\right]$$

- Select model, $\hat{\gamma}_{\mathrm{AIC}}$, that minimises

$$\mathrm{AIC}(\gamma) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2k_\gamma$$

  where $k_\gamma$ is the number of parameters in model $\gamma$

# Akaike's Information Criterion (AIC) (2)

- For linear models, select the set of exposure $\gamma$ that minimises

$$\mathrm{AIC}(\gamma) = n \log \hat{\tau}(\mathbf{y}; \gamma) + 2|\gamma|$$

where

$$\begin{aligned} \hat{\tau}(\mathbf{y}; \gamma) &= \frac{1}{n}(y - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma))'(y - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma)) \\ \hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma) &= (\mathbf{X}_\gamma' \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma' \mathbf{y} \end{aligned}$$

are the maximum likelihood estimates.

# Properties of AIC (1)

- For nested models, the AIC has several favourable properties
- Under certain conditions (model $\gamma$ contains truth, nested models)

$$\mathrm{E}\left[\mathrm{AIC}(\gamma)\right] = \mathrm{E}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_\gamma)\right] + o_n(1)$$

  where $o_n(1)$ denotes a quantity that vanishes as $n \to \infty$.
- In words: the AIC is an asympotically unbiased estimate of cross-entropy risk

# Properties of AIC (2)

- The AIC is asymptotically efficient
- This means, that (roughly) if
  - The truth is *not* in the set of candidate models $\Gamma$, and
  - the set of candidate models is nested, and is growing with $n$

  then

$$\mathbb{P}(\hat{\gamma}_{\text{AIC}} = \gamma_{\text{BEST}}) \rightarrow 1 \text{ as } n \rightarrow \infty$$

  where

$$\gamma_{\text{BEST}} = \arg\min_{\gamma \in \Gamma} \left\{ d(\hat{\boldsymbol{\theta}}_{\gamma}, \hat{\boldsymbol{\theta}}_{*}) \right\}$$

- In words: under the above conditions, the AIC asympotically selects the best fitting model in terms of prediction error
- Example: Polynomial regression where the truth is not a polynomial

# Properties of AIC (3)

- The AIC is not consistent
- This means, that even if
  - The truth is in the set of candidate models, i.e. $\gamma_* \in \Gamma$, and
  - the set of candidate models does not grow with $n$

  then

$$\mathbb{P}(\hat{\gamma}_{\text{AIC}} \neq \gamma_*) \rightarrow c \text{ as } n \rightarrow \infty$$

  where $c > 0$ is a non-zero constant depending on the set of candidate models
- In words: regardless of the amount of data we obtain, the AIC will have a non-zero probability of erroneously included exposures that are not associated with the outcome

# Content

# Main Result (1)

- Main result concerns non-nested model sequences
- Assume usual AIC regularity conditions
- Setup
  - Starting from the true model $\gamma_*$ with no parameters
  - Choosing between candidate models with $k$ parameters
  - Let

$$a_m = \log p(\mathbf{y}|\boldsymbol{\theta}_*) - \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m), \quad m \in \Gamma_k$$

    be the improvement in fit for the candidate models.
  - We assume the $a_m$ are all independent random variates
- Selecting best $k$ model parameter is equivalent to

$$\hat{m}_k = \underset{m \in \Gamma_k}{\arg\max} \{a_m\}$$

# Main Result (2)

## Theorem

$$2\mathrm{E}_{\boldsymbol{\theta}^*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] + 2\alpha(k, q_k) = \mathrm{E}_{\boldsymbol{\theta}^*}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] + o_n(1)$$

*where*

$$\alpha(k, q_k) = \mathrm{E}_{\chi^2_k}\left[\max\left\{z_1, \ldots, z_{q_k}\right\}\right]$$

*and*

- $z_1, \ldots, z_{q_k}$ *are iid* $\chi^2_k$ *variates with* $k$ *degrees of freedom*
- $q_k = |\Gamma_k|$ *is the number of candidate models with* $k$ *parameters*

# Implications (1)

- What does it mean ?
- If model class is *nested*, $q_k = 1$ and $\alpha(k, q_k) = k$
  $\Rightarrow$ we recover the regular AIC
- If $q_k > 1$, $\alpha(k, q_k) > k$
  - regular AIC underestimates expected cross-entropy
  - as $\alpha(k, q_k)$ does not depend on $n$, this bias will remain even as $n \to \infty$

  $\Rightarrow$ Increased likelihood of overfitting depends on $q_k$ !

# Implications (2)

- How serious is it ?
- Consider a simple regression setup
  - We have $q$ orthogonal covariates

$$\Gamma_1 = \{1\}, \{2\}, \ldots, \{q\}$$

  - The "true" model includes no covariates, i.e., $\gamma_* = \{\}$
  - Use AIC to decide whether to prefer a model in $\Gamma_1$ to the true, empty model $\gamma_*$
- Using our Theorem, we can determine probability of overfitting when using AIC

# Implications (2, cont'd)

TABLE: Probability of Overfitting

| $q_1$ | $\mathbb{P}$(overfit) |
|---|---|
| 1 | 0.157 |
| 2 | 0.290 |
| 3 | 0.402 |
| 4 | 0.496 |
| 5 | 0.575 |
| 8 | 0.746 |
| 10 | 0.819 |
| 15 | 0.923 |
| 25 | 0.986 |
| 50 | 0.999 |
| 100 | 1.000 |

# Implications (3)

- What if the $a_m$ variates are not independent ?
  - e.g., if several regression models include the same covariates their $a_m$ variates will be correlated
- In this case our Theorem acts as an upper-bound
- How tight this bound is in many usual model selection settings is a topic for future research
- The theorem implies an unfortunate conclusion
  $\Rightarrow$ the problem is greatest when exposures are orthogonal !

# Implications (4)

- Forward stepwise model building with hypothesis testing
  - Including exposure with smallest $p$-value to pass some nominal "significance" level
- AIC can be viewed in this light:
  - Twice drop in likelihood, $2a_m$, asymptotically $\chi^2_1$ distributed
  - Performs $\chi^2_1$-test at $\approx 16\%$ significance level
- But, we always consider exposure with biggest drop, i.e.,

$$\hat{m}_k = \underset{m \in \Gamma_k}{\arg\max} \{a_m\}$$

- Data is *dictating* order in which we consider exposures
  $\Rightarrow$ "Null" distribution is actually maximum of $\chi^2_1$ variates
- So the "$p$-values" derived from $\chi^2$ test can be (very) wrong !

# Alternatives (1)

- Several possible alternatives
- In our paper, we suggest a threshold for inclusion of

$$\mathrm{E}\left[a_m\right] + 1 = (1/2)\alpha(1, r) + 1$$

  where $r$ is the number of exposures not included in the model
- Is based on expected improvement in fit due to inclusion of best unassociated exposure
  - Performed well on a difficult problem (wavelet denoising)
  - Can be conservative

# Alternatives (2)

- Minimum Message Length (MML) regression criterion *(Schmidt & Makalic, 2009)*

$$\begin{aligned}\mathrm{MML}_u(\gamma) &= \left(\frac{n-k_\gamma}{2}\right)\log\hat{\tau}_{\mathrm{MML}}(\mathbf{y};\gamma) + \frac{k_\gamma}{2}\log\left(\frac{||\mathbf{y}||^2}{k_\gamma}\right) \\ &\quad -\frac{1}{2}\log(k_\gamma+1) - \log\Gamma\left(k_\gamma+1\right) - \log\Gamma\left(q-k_\gamma+1\right)\end{aligned}$$

where

$$||\mathbf{y}||^2 = \sum_{i=1}^{n} y_i^2, \quad \hat{\tau}_{\mathrm{MML}}(\mathbf{y};\gamma) = \frac{n\,\hat{\tau}(\mathbf{y};\gamma)}{n-k_\gamma}$$

  - $k_\gamma$ is the number of exposures in the model
  - $q$ is the total number of candidate exposures
  - $\Gamma(\cdot)$ is the Gamma function
- Powerful properties combining BIC and AIC

# Alternatives (3)

- Covariance Inflation Criterion (CIC) *(Tibshirani & Knight)*
- Forward stepwise regression based on false discovery rate (FDR) *(Benjamini & Gavrilov)*
- "Extended" Bayesian Information Criterion *(Chen & Chen)*
- LASSO coupled with AIC *(Zou, Hastie and Tibshirani)*
  - AIC score once again a good estimate of expected cross-entropy
- And many more ...

# Example (1)

- Conclude with a simple example
- Data generated from polynomial

$$y_i = 2x_i^5 + 0.8x_i^3 + 9.4x^2 - 5.7x - 2 + \varepsilon_i$$

where

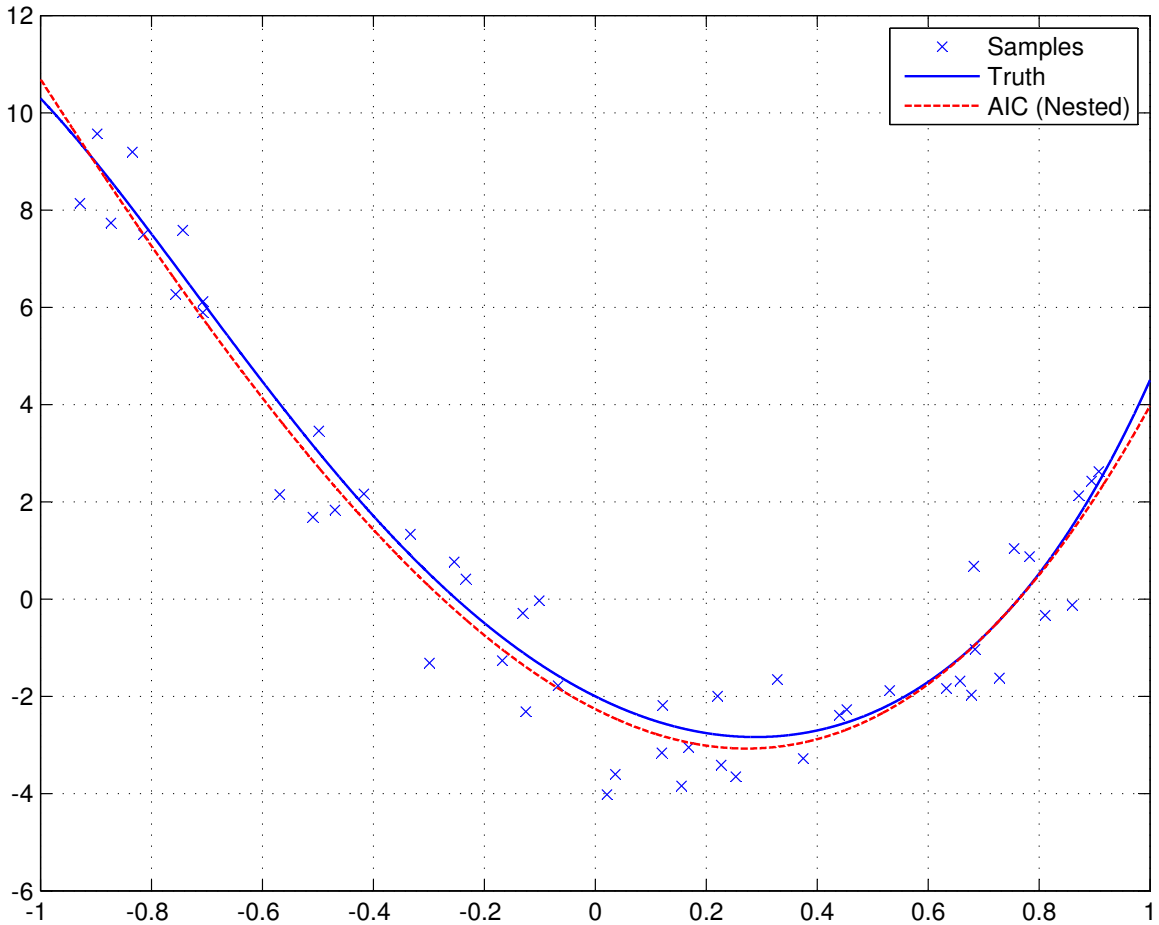
$$\varepsilon_i \sim N(0, 1)$$

- Generated $n = 100$ random samples
- Fitted models using
  - AIC with nested structure
  - AIC with non-nested (all subsets) structure
  - MML with non-nested (all subsets) structure

# Example (2)

- AIC (nested structure)

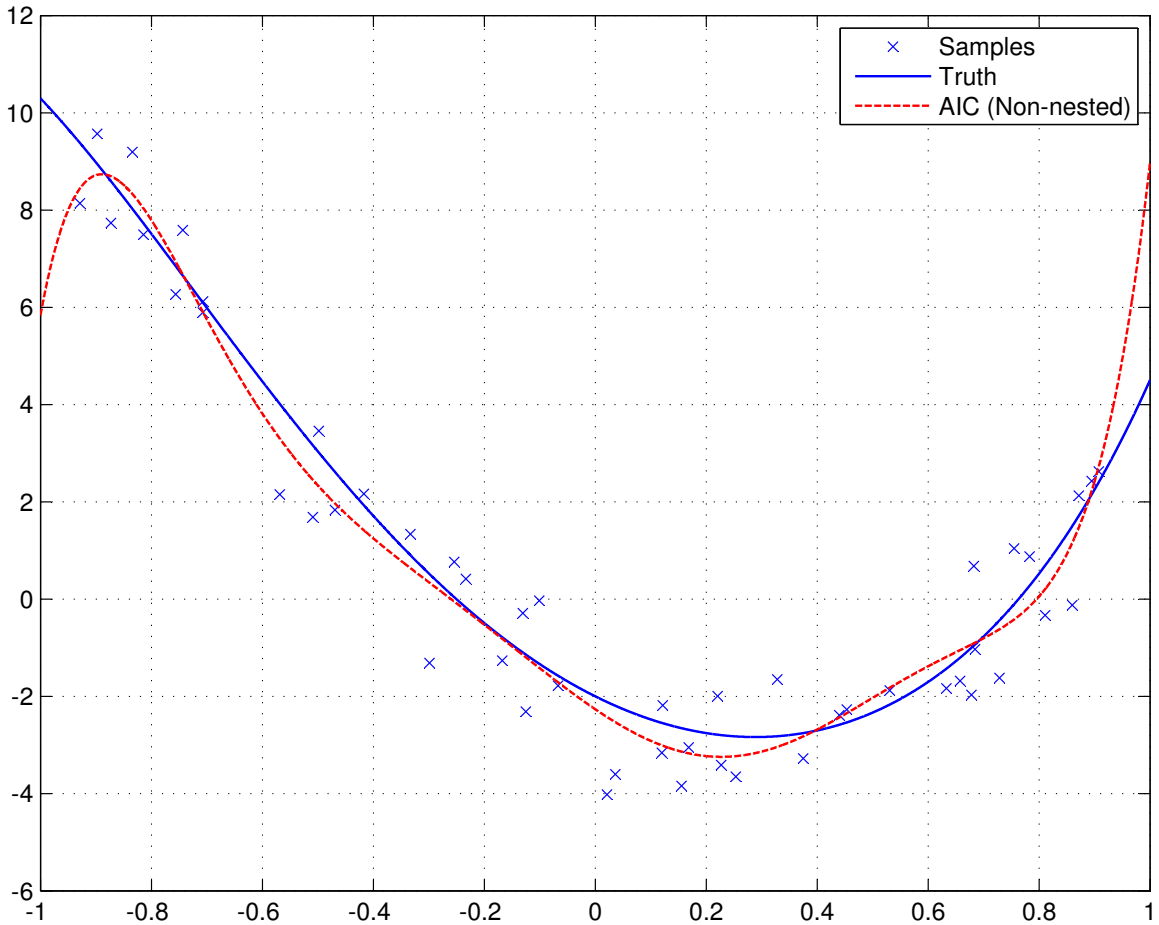
$$y_i = 2.42x_i^3 + 9.6x_i^2 - 5.8x_i - 2.26$$



# Example (3)

- AIC (non-nested structure)

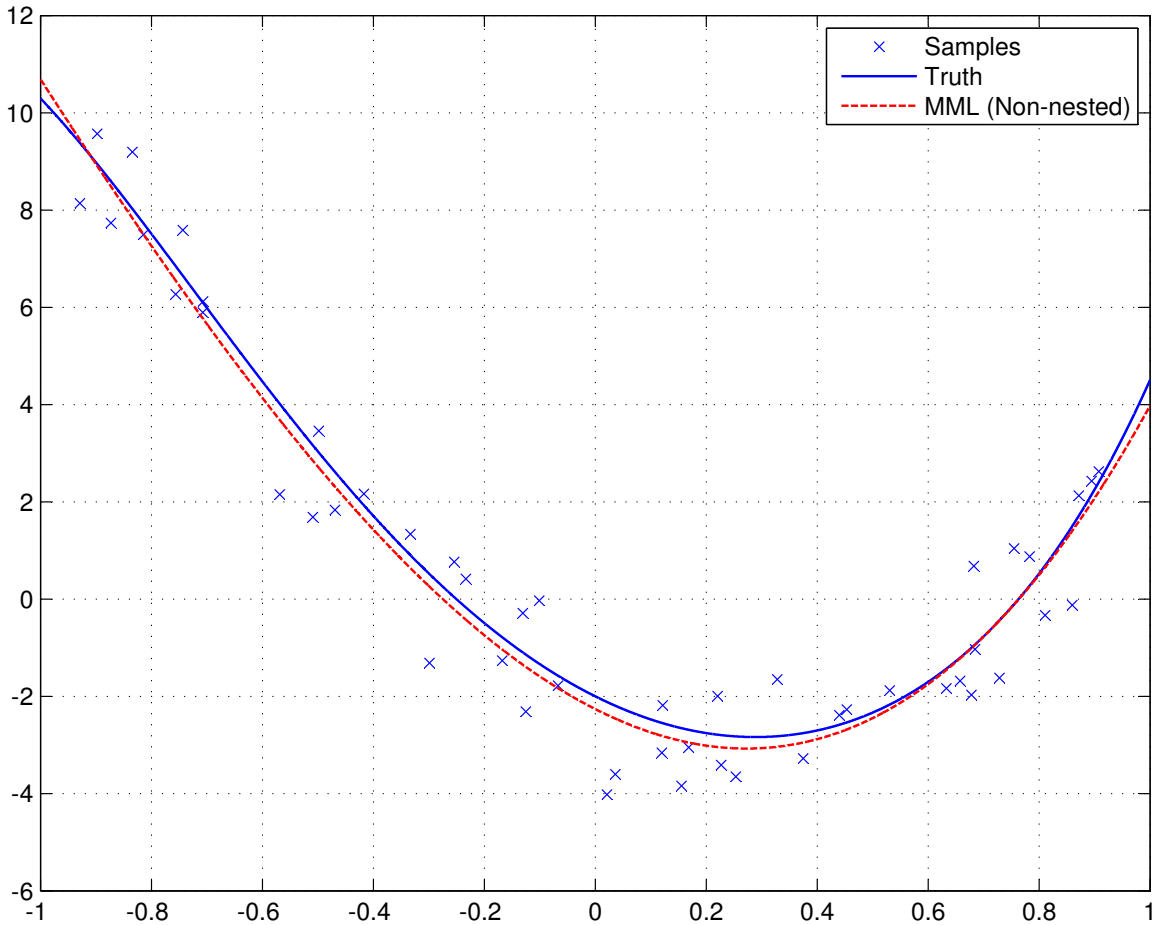
$$y_i = 48.19x_i^7 - 65.75x_i^5 + 26.86x_i^3 + 9.66x_i^2 - 7.73x_i - 2.26$$



# Example (4)

- MML (non-nested structure)

$$y_i = 2.42x_i^3 + 9.6x_i^2 - 5.8x_i - 2.26$$



# Conclusion

- Examined the use of the Akaike Information Criterion (AIC) in the case of non-nested model selection
- AIC can perform badly in non-nested settings due to bias
  - Depends on the number of models under consideration
  - Leads to excess "overfitting"

  $\Rightarrow$ The bias cannot be overcome even as $n \to \infty$
- There exists other procedures more suitable in this setting

# References


- Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716–723
- Kullback, S., Leibler, R.A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, Vol. 22, pp. 79–86
- Schmidt, D.F., Makalic, E. The Behaviour of the Akaike Information Criterion When Applied to Non-nested Sequences of Models. *Lecture Notes in Artificial Intelligence*, 2010, Vol. 6464, pp. 223–232
- Schmidt, D.F., Makalic, E. Invariant MML Linear Regression. *Lecture Notes in Artificial Intelligence*, 2009, Vol. 5866, pp. 312–321
- Tibshirani, R. Knight, K. The Covariance Inflation Criterion for Adaptive Model Selection. *Journal of the Royal Statistical Society, Series B*, 1999, Vol. 61, pp. 529–546
- Chen J., Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 2008, Vol. 95, pp. 759–771
- Benjamini Y., Gavrilov Y. A simple forward selection procedure based on false discovery rate control. *Annals of Applied Statistics*, 2009, Vol. 3, pp. 179–198
- Zou H., Hastie T., Tibshirani R. On the “degrees of freedom” of the LASSO. *The Annals of Statistics*, 2007, Vol. 35, pp. 2173–2192
- Linhart, H., Zucchini, W. Model Selection. Wiley, New York, 1986