Overfitting and Selection Bias in Model Selection

Mark Wolters

UBC/SFU Joint Student Seminar

24-May-2006

OUTLINE

Introduction

Selection Procedures and Selection Criteria

Problem 1: Overfitting

- Tendency to select models with unnecessary complexity

Problem 2: Selection Bias

- Selection process introduces bias into coefficient estimates

Are There Solutions?

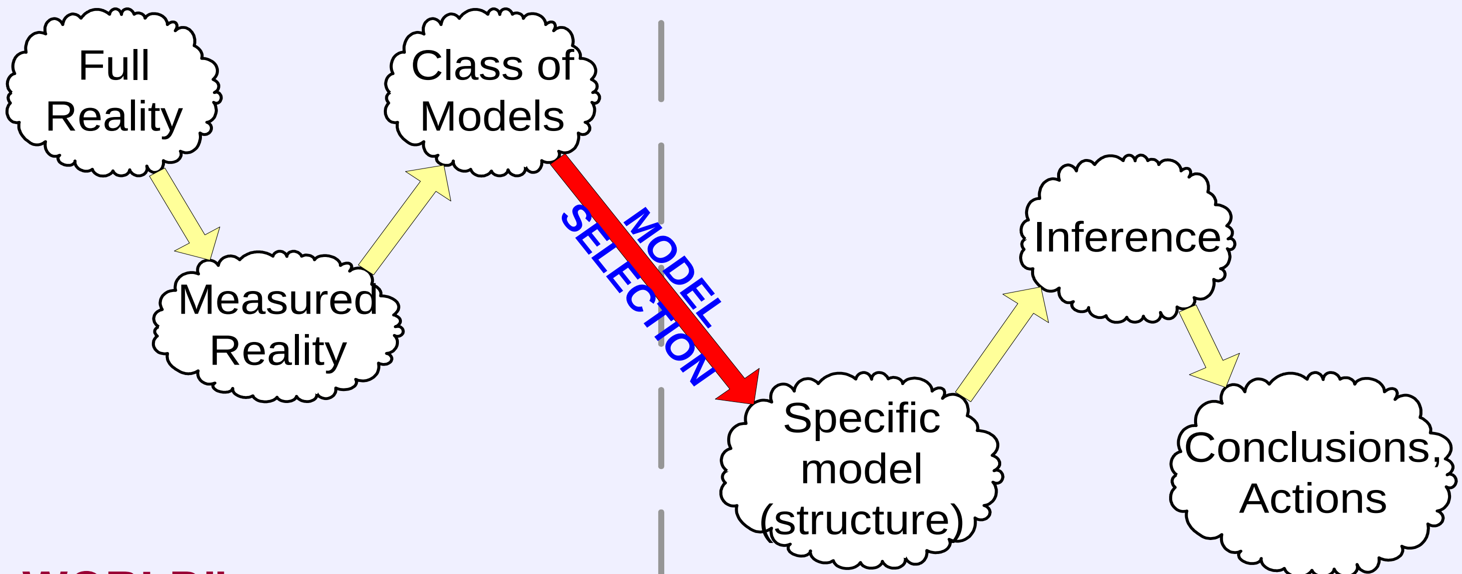
Take-Home Messages

Introduction

INTRODUCTION

Where does model selection fit in?

- Consider sequence of simplifications in data analysis:



"REAL WORLD"

-Define problem

-Choice of predictors, response

-Possible relationships among variables

-Measurement systems

-Experimental design

"IDEAL WORLD"

-Parameter estimation

-Subject-matter interpretation

INTRODUCTION

Model selection is the point at which the real world is left behind for good.

After model selection:

- The universe is divided into “important” and “nonexistent”
- The nature of the relationship between variables is fixed.

During subsequent analysis:

- Make some confidence intervals…
- All conclusions are ***conditional on model truth***.

INTRODUCTION

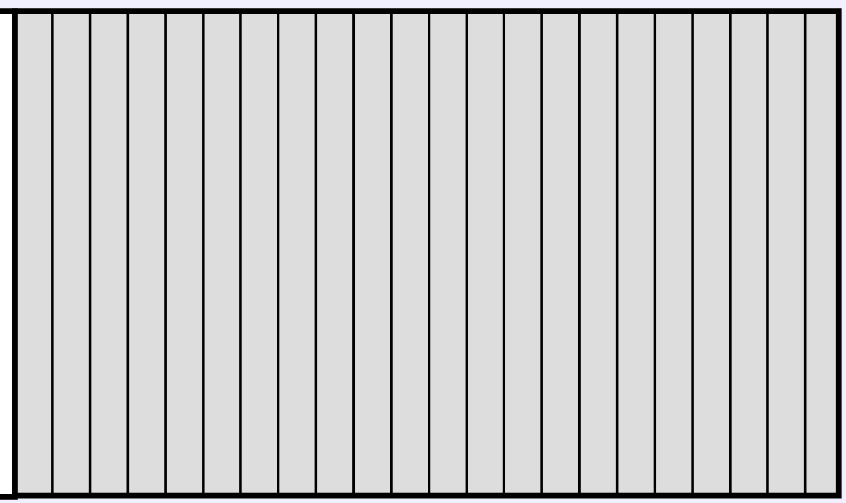
Motivating example

12-run Plackett-Burman design with interactions (PB12)

- Industrial screening experiments.
- Suspect a few main effects and a few interactions may be active.
- By including interactions:
  - introduce complex aliasing.
  - ***finding active factors becomes a model selection problem***.

X =

1	1	1	1	1	1	1	1	1	1	1	1
1	-1	1	-1	1	1	1	-1	-1	-1	1	-1
1	-1	-1	1	-1	1	1	1	-1	-1	-1	1
1	1	-1	-1	1	-1	1	1	1	-1	-1	-1
1	-1	1	-1	-1	1	-1	1	1	1	-1	-1
1	-1	-1	1	-1	-1	1	-1	1	1	1	-1
1	-1	-1	-1	1	-1	-1	1	-1	1	1	1
1	1	-1	-1	-1	1	-1	-1	1	-1	1	1
1	1	1	-1	-1	-1	1	-1	-1	1	-1	1
1	1	1	1	-1	-1	-1	1	-1	-1	1	-1
1	-1	1	1	1	-1	-1	-1	1	-1	-1	1
1	1	-1	1	1	1	-1	-1	-1	1	-1	-1



11 main effects at 2 levels

55 two-way interactions

12 runs

INTRODUCTION

Terms and Notation—PB12 design

- ***Full matrix*** **X**. $\underset{12\times 67}{\mathbf{X}}$
- ***True coefficients***, β (mostly zeros). $\underset{67\times 1}{\boldsymbol{\beta}_{full}}$
- ***Response vector***, **y**. $\underset{n\times 1}{\mathbf{y}}$
- ***Model size***, p.
  - number of variables w/o intercept.
- ***Model matrix***, **M**. $\underset{12\times(p+1)}{\mathbf{M}}$
  - formed by selecting p columns from **X**
- Standard linear regression model. $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$
  - σ^2 is ***residual variance***

INTRODUCTION

12-run Plackett-Burman design

- Full model not estimable.
- How many different models are possible?
- Consider only models respecting ***effect heredity***:

Model Size	# of Models
2	165
3	1,705
4	15,510
5	125,202
6	902,649
7	5,893,800

Even considering only models respecting heredity, model sets quickly become huge.

Selection Procedures and Selection Criteria

SELECTION PROCEDURES & CRITERIA

Elements of a model selection procedure:

- ***Criterion***
  - How do we measure whether one model is better than another?
- ***Search method***
  - How do we find good models?
- ***Information processing***
  - How do we use the results of the search?

Why is model selection difficult?

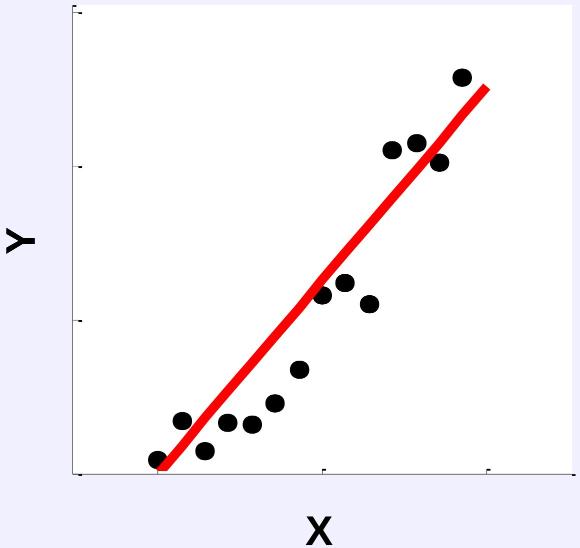
- ***Model selection uncertainty***
  - best model is subject to sampling variability.
- ***Large model sets***
- ***Many possible criteria***
- ***Hard to compare models of different sizes***

SELECTION PROCEDURES & CRITERIA

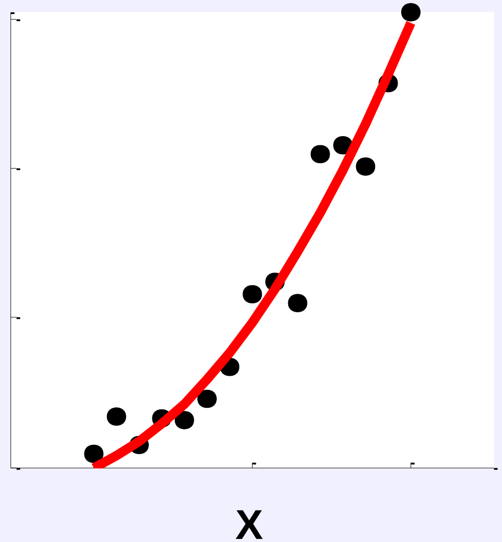
Need to measure model goodness… So what is a good model?

- Simple example: three models for response Y with predictor X:

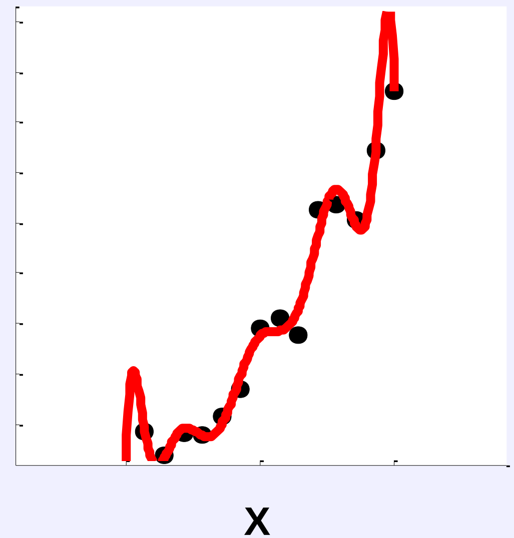
M1: $Y = \beta_1 + \beta_2 X$



M2: $Y = \beta_1 + \beta_2 X + \beta_3 X^2$



M3: 10^{th} degree poly.



How to choose between these?

-Goodness of fit?

-Physical basis?

-Study objectives?

-Statistical decision rule?

Unnecessary complexity?

Essence of model selection problem:
when is extra complexity justified?

SELECTION PROCEDURES & CRITERIA

Importance of predictive power:

- Goodness-of-fit isn't useful in itself.
- "Parsimony" isn't useful in itself:

(predictive power)	**+**	**(no simplicity)**	**=**	**POTENTIALLY USEFUL**
(predictive power)	**+**	**(simplicity)**	**=**	**POTENTIALLY USEFUL**
(no predictive power)	**+**	**(simplicity)**	**=**	**DANGEROUS**

- ***Adequate predictive power is essential***.
- Practical considerations may justify trading predictive power for simplicity.
- "Principle of parsimony" misunderstood?

There are no parsimonious models, only parsimonious modellers.

SELECTION PROCEDURES & CRITERIA

Some important selection criteria

- Why so many different criteria?
  - "Good model" is subjective concept.
  - Difficult problem; many proposals.

Residual Sum of Squares (RSS):

$$RSS = \left(\mathbf{y} - \mathbf{y}\right)^T \left(\mathbf{y} - \mathbf{y}\right)$$

- Purely goodness-of-fit based
- Proportional to maximized log likelihood
- Likelihood can be interpreted as evidence.
- Problem: always gets better as parameters added.

Mallows' C_p:

$$C_p = \frac{RSS}{\sigma^2} - \left(n - 2k\right)$$

- Estimate of standardized total MSE of **y**.
- (n-2k) term penalizes extra parameters.

SELECTION PROCEDURES & CRITERIA

Some important selection criteria (cont'd)

Akaike Information Criterion (AIC)

- Based on estimate of Kullback-Liebler discrepancy between model and truth.
- Balance between likelihood and parameter penalty.

$$AIC = -2\ln\left(L\left(\boldsymbol{\beta}, \sigma^2 \middle| \mathbf{y}\right)\right) + 2K$$

Many others, and variants, exist.

Problem 1: Overfitting

OVERFITTING

When is a good model not a good model?

Definition:

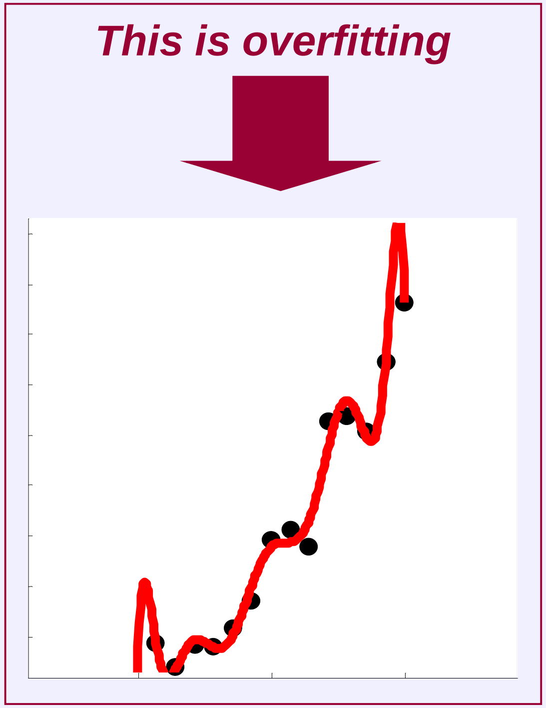
- Choosing a model with unnecessary complexity.
- Usually refers to selecting a model that:
  - Includes all the truly-active variables.
  - Also includes spurious variables.

Causes of overfitting

- Criteria tend to prefer larger models.
- There are many more larger models.

Results of overfitting

- Fit is “too good”
- Poor predictive performance



OVERFITTING

SIMULATION 1: Overfitting in PB12 model selection

- True model has size 3. Active factors: (**1, 2, 1*3**)
- True model: E[**y**] = 1 + $\mathbf{X}_1$ + $\mathbf{X}_2$ + $\mathbf{X}_{1*3}$ (all coefficients equal 1.0)

Consider 4 sets of models:

• TRUE	the true model	(**1, 2, 1*3**)
• M1	a specific overfitted model	(**1, 2, 1*3, 1*4**)
• OVER	all models overfitted by 1 variable	(27 models)
• OTHER	all other models	(17187 models)

Use AIC_c as model selection criterion

Select from all models of size 3 or 4 (exhaustive search)

Repeat model selection for 250 simulated y's

- Count how many times models in each set get selected.

OVERFITTING

SIMULATION 1 results:

Proportion of selections from each group:

True model vs. single overfitted model:

	TRUE	M1
$\sigma = 0.5$	0.98	0.02
$\sigma = 1.0$	0.96	0.04

True model vs. 27 overfitted models and 17187 wrong models:

	TRUE	OVER	OTHER
$\sigma = 0.5$	0.33	0.66	0.01
$\sigma = 1.0$	0.12	0.33	0.55

MODEL SELECTION UNCERTAINTY

- **Criterion does well at choosing the true model vs. single bad option.**
- **But high number of options results in sub-optimal choices over whole model set.**
- **Increasing residual variance makes matters much worse.**

Problem 2: Selection Bias

SELECTION BIAS

What is selection bias?

Using the same data for model selection and parameter estimation introduces bias into coefficient estimates.

Why?

- Regression coefficients unbiased only if model is ***given*** and ***true***.
- Well-fitting models tend to have larger coefficients; hence ***selection procedures prefer models with large coefficients***.

Magnitude of bias depends on:

- **Selection procedure.**
- **Experimental design.**
- **The nature of the truth.**

Usual effects of selection bias:

- **Coefficients too large.**
- **Variance estimates too small.**
- **Confidence interval coverage poor.**

SELECTION BIAS

SIMULATION 2: selection bias in PB12 experiment

- Same active factors as before: (**1, 2, 1*3**)
- True model: $E[\mathbf{y}] = 1 + \mathbf{X}_1 + 0.75\mathbf{X}_2 + 0.5\mathbf{X}_{1*3}$
- Residual variance: $\sigma^2 = 1$.

Use AIC_c as model selection criterion

Select from all models of size 3 or 4 (exhaustive search)

Repeat model selection for 1000 simulated y's

Simulation output:

- Distribution of β, σ estimates based on best model.
- True coverage of standard 95% confidence intervals.

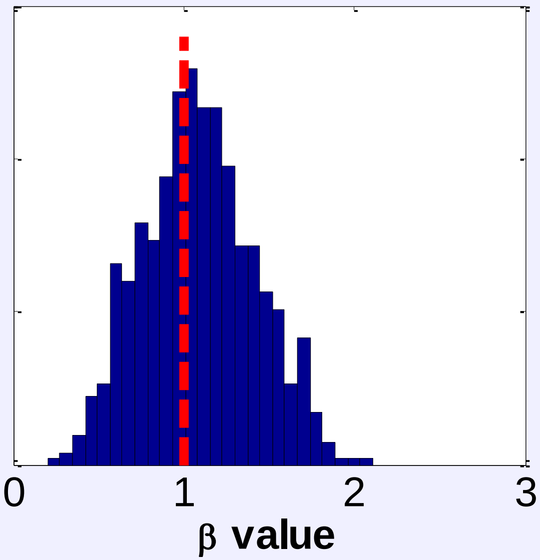
SELECTION BIAS

SIMULATION 2 results

Reminder: $\sigma^2 = 1$

Red line = true value

Coefficient of Variable 1



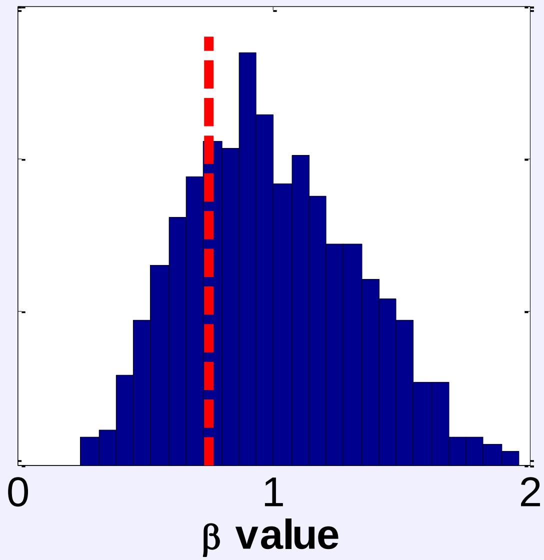
Times in best model: 727

True value = 1.0

Mean estimate: 1.09

bias: 0.31*se(β)

Coefficient of Variable 2



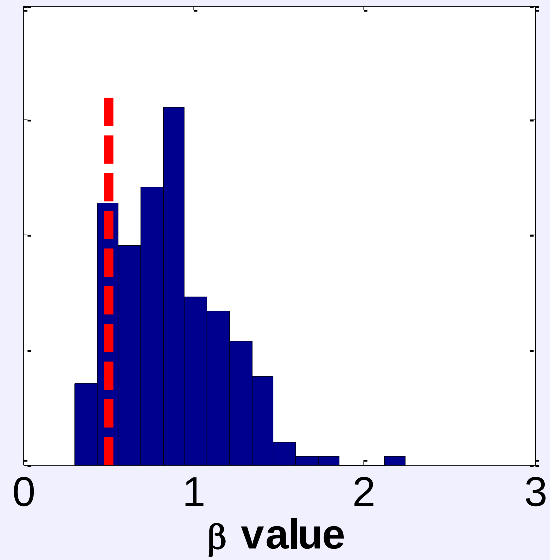
Times in best model: 652

True value = 0.75

Mean estimate: 1.00

bias: 0.82*se(β)

Coefficient of Variable 1*3



Times in best model: 243

True value = 0.5

Mean estimate: 0.86

bias: 1.18*se(β)

SELECTION BIAS

SIMULATION 2 results (cont'd)

True coverage of 95% t-intervals on each coefficient:

	Var 1	Var 2	Var 1*3
Proportion of intervals containing true value	0.65	0.58	0.56

Estimates of residual variance in selected models:

	True value	Average estimate over selected models
σ^2 value	1.0	0.25

SELECTION BIAS

Notes on selection bias

- Selection bias has the potential to totally invalidate inference.
- Severity of problem usually difficult to work out theoretically.
- In general, expect worse problem:
  - When many models in close competition
  - When effect sizes are small

Are There Solutions?

ARE THERE SOLUTIONS?

Overfitting and selection bias are natural consequences of this style of data analysis.

- Awareness of risk is first step.
- Conservative, iterative, learning approach will help.

<u>To combat overfitting:</u>

- Consider multiple models; report multiple models.
- Model averaging and/or Bayesian approach.

<u>To combat selection bias:</u>

- Incorporate information from selection procedure into inference.
  - (open research area?)
- Resolve model selection issue in preliminary stages of study.
- Use subject-matter knowledge to restrict model sets.
- Model averaging will help.

Take-Home Messages

TAKE-HOME MESSAGES

Recommendations if doing this sort of model building:

- ***Give model selection due attention, or risk invalid inference***.

- ***Think carefully about relative importance of GOF, simplicity, and predictive power in your specific case***.

- ***For huge model sets,***
  - ***Particular choice of selection criterion not that important.***
  - ***Best-ranked model in any trial likely not the true best.***
  - ***Inference from a single model is perilous***.

- ***Simulations are invaluable in exploring the issues in specific cases***.

Supporting Slides

SUPPORTING SLIDES

What is "truth" really like?

- Simulations usually have several large β's and the rest exactly zero.
  - Assumes truth can actually be described by a linear model with the chosen predictors.
  - True factors probably "small," but not exactly zero. Truth probably never like this; but sometimes close enough?
- Assumption of normal, independent, homoscedastic errors is key for regression setting.
  - Truth likely not that simple.
- Key question: is truth ***close enough*** to these ideals to make modelling worth while?
- ***Claim: deviations from the ideal will make overfitting and selection bias worse.***

SUPPORTING SLIDES

SIMULATION 2—results for only when the true model was selected (124 cases):

True coverage of 95% t-intervals on each coefficient:

	Var 1	Var 2	Var 1*3
Proportion of intervals containing true value	0.77	0.81	0.77

Average estimated coefficients:

	Var 1	Var 2	Var 1*3
True value	1	0.75	0.5
Mean estimate	1.05	0.79	0.72

SUPPORTING SLIDES

Data produced from linear relationship:

M1: $Y = \beta_1 + \beta_2 X$

M2: $Y = \beta_1 + \beta_2 X + \beta_3 X^2$

M3: 10^{th} degree poly.