"If a particular model (parametrization) does not make biological sense, this is reason to exclude it from the set of candidate models, particularly in the case where causation is of interest. In developing the set of candidate models, one must recognize a certain balance between keeping the set small and focused on plausible hypotheses, while making it big enough to guard against omitting a very good a priori model."

Burnham and Anderson, 2002, Model Selection and Multimodel Inference: A Practical Information-theoretic Approach

Introduction to model selection in R

Timothée Bonnet

May 30, 2019

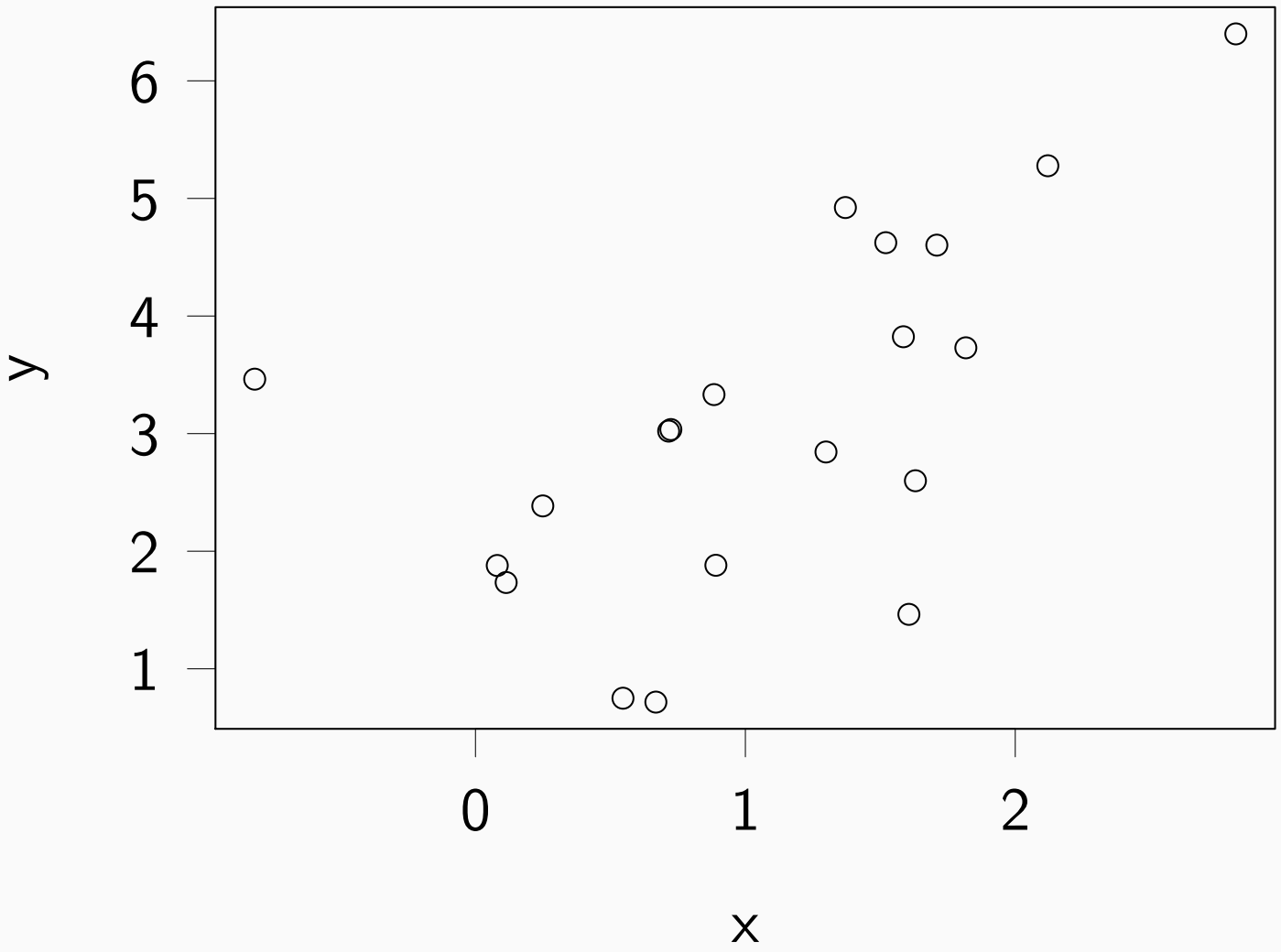
BDSI / RSB

Why model selection?

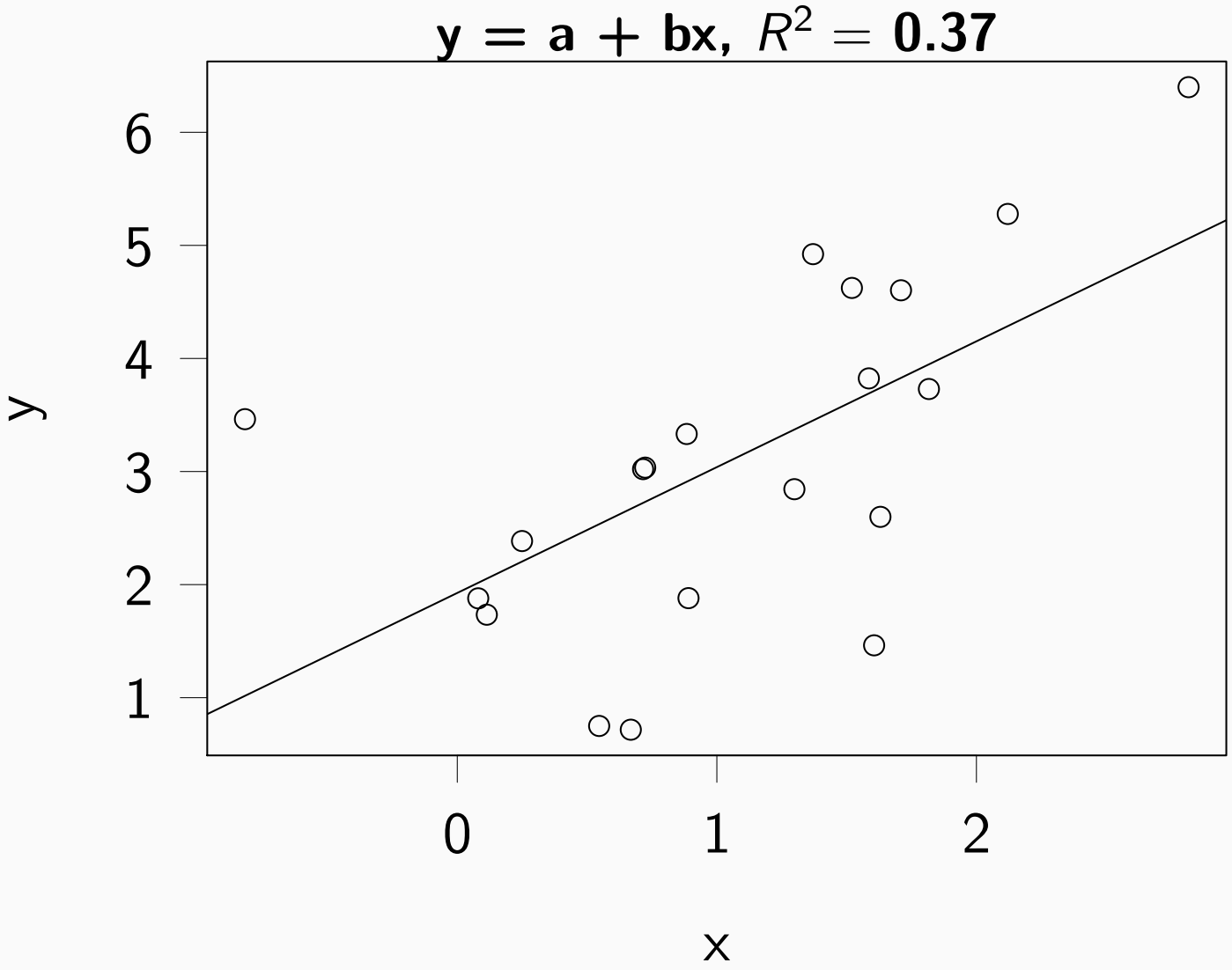
Information criteria vs. stepwise selection

Model selection and causal inference

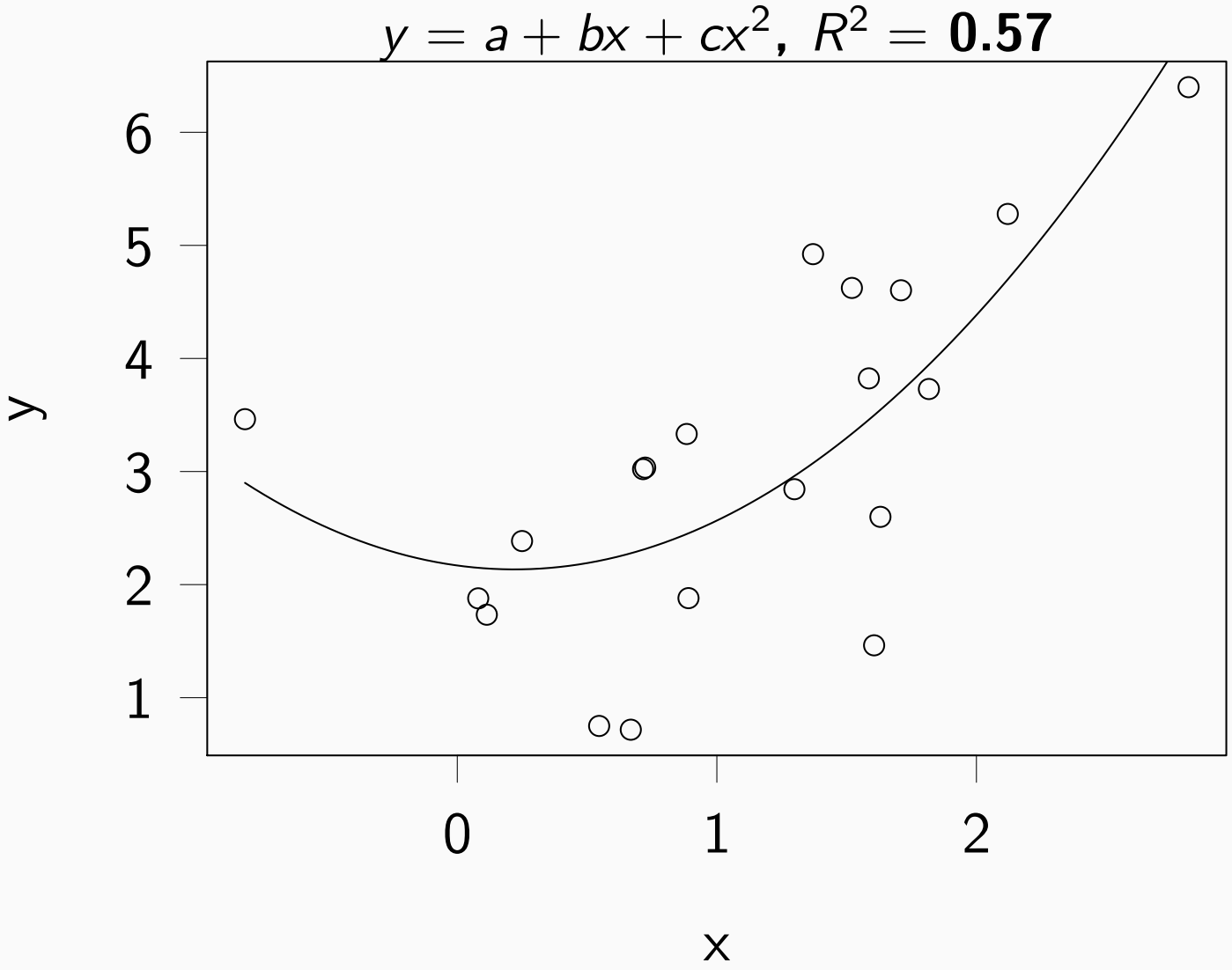
Overfitting



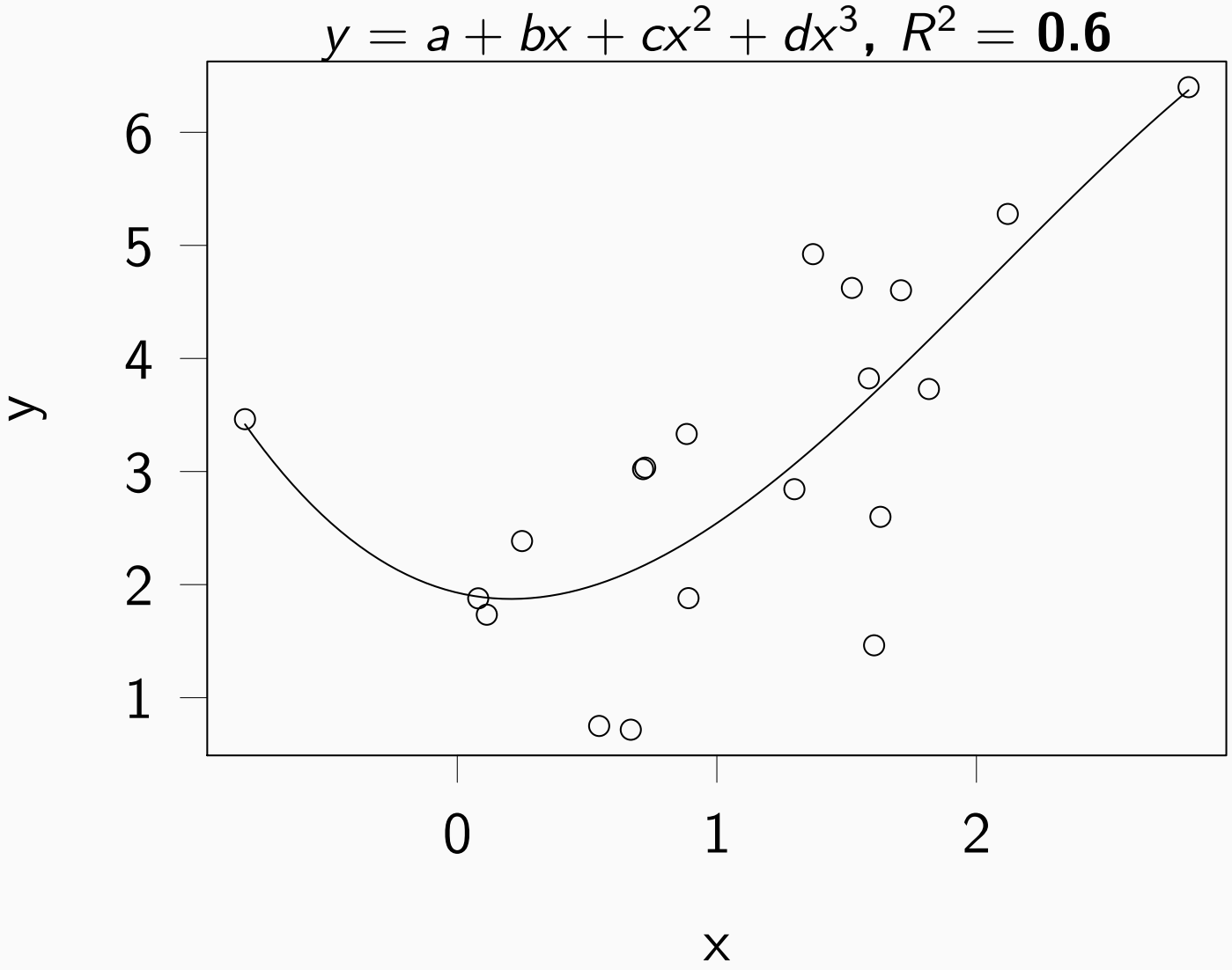
Overfitting



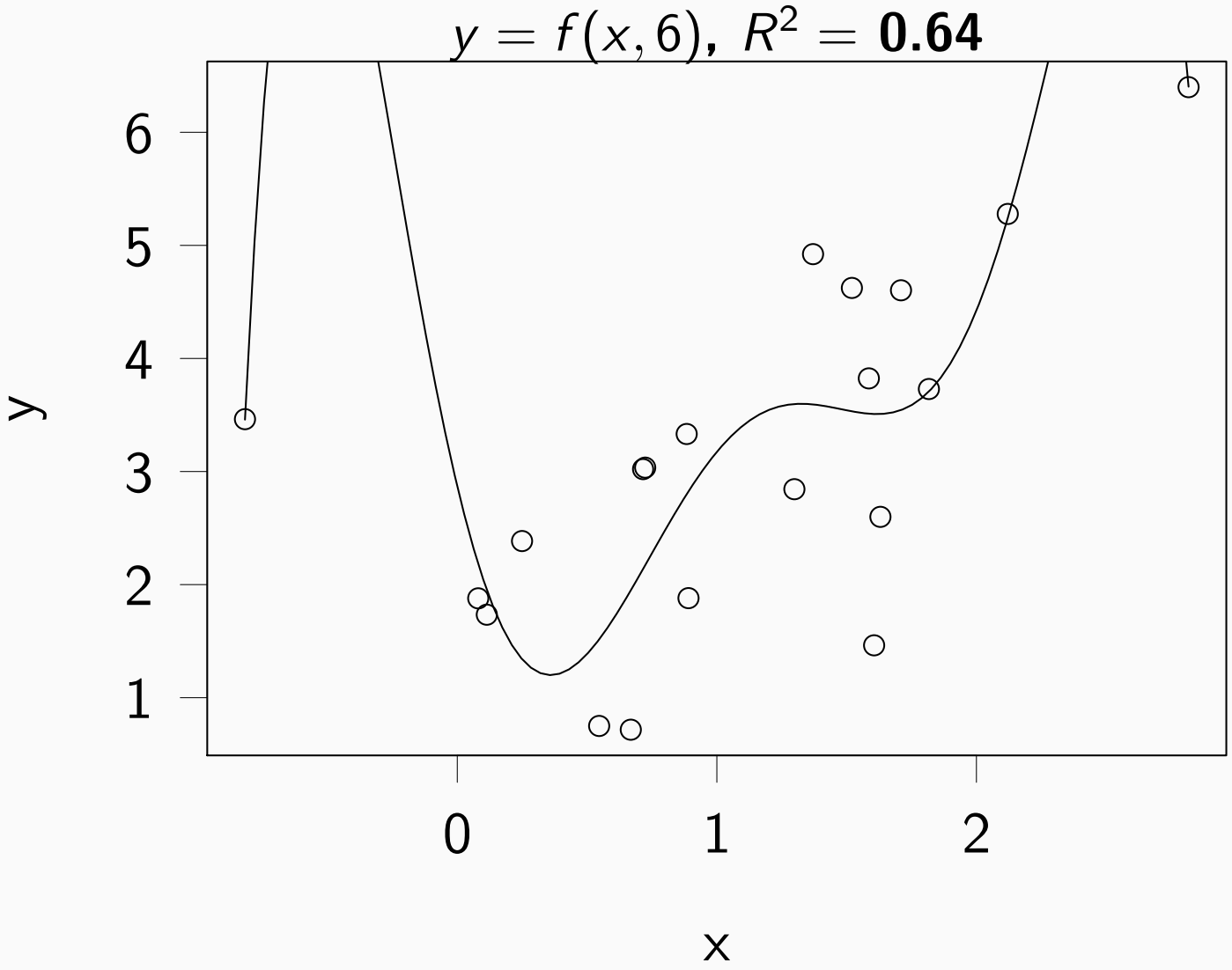
Overfitting



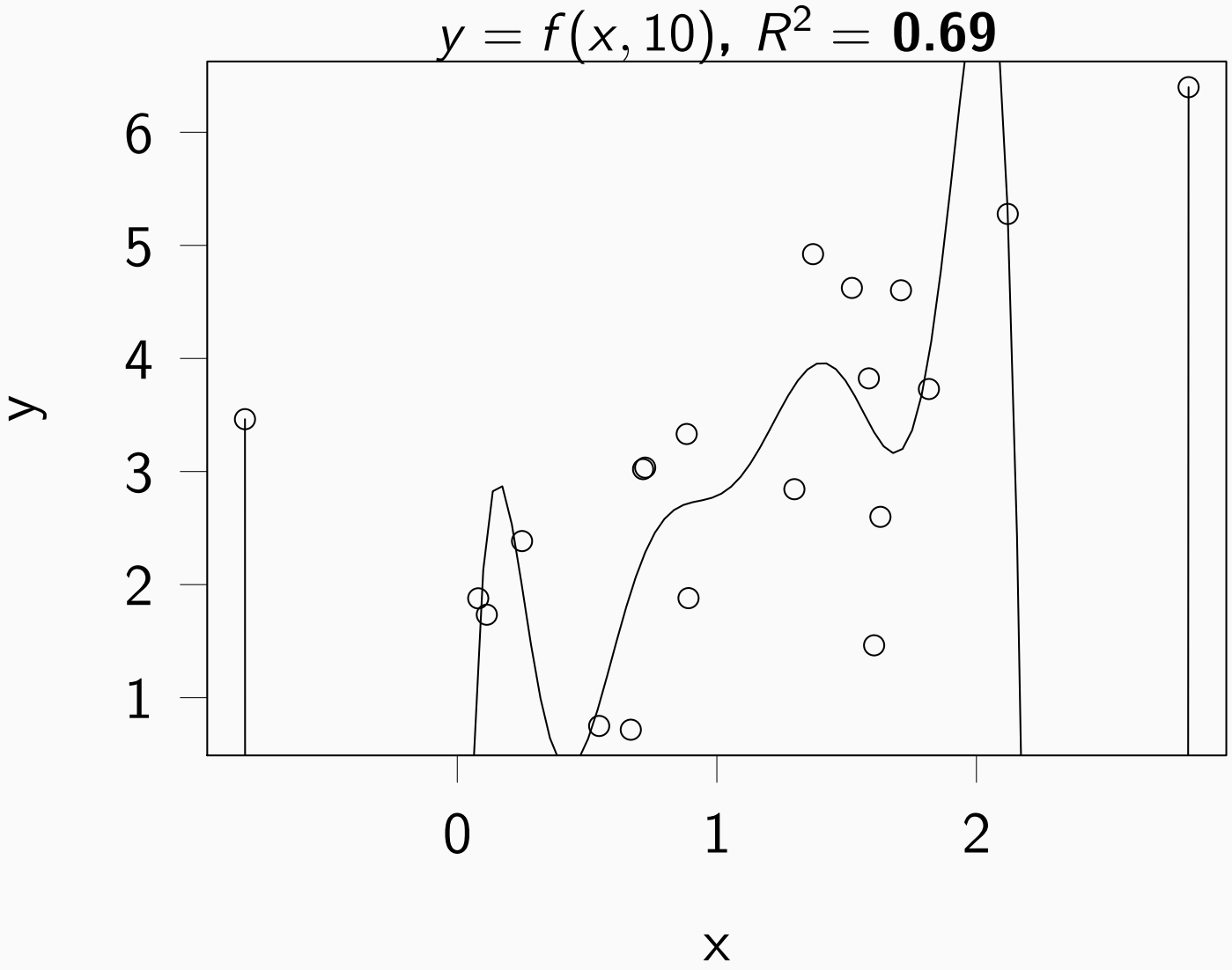
Overfitting



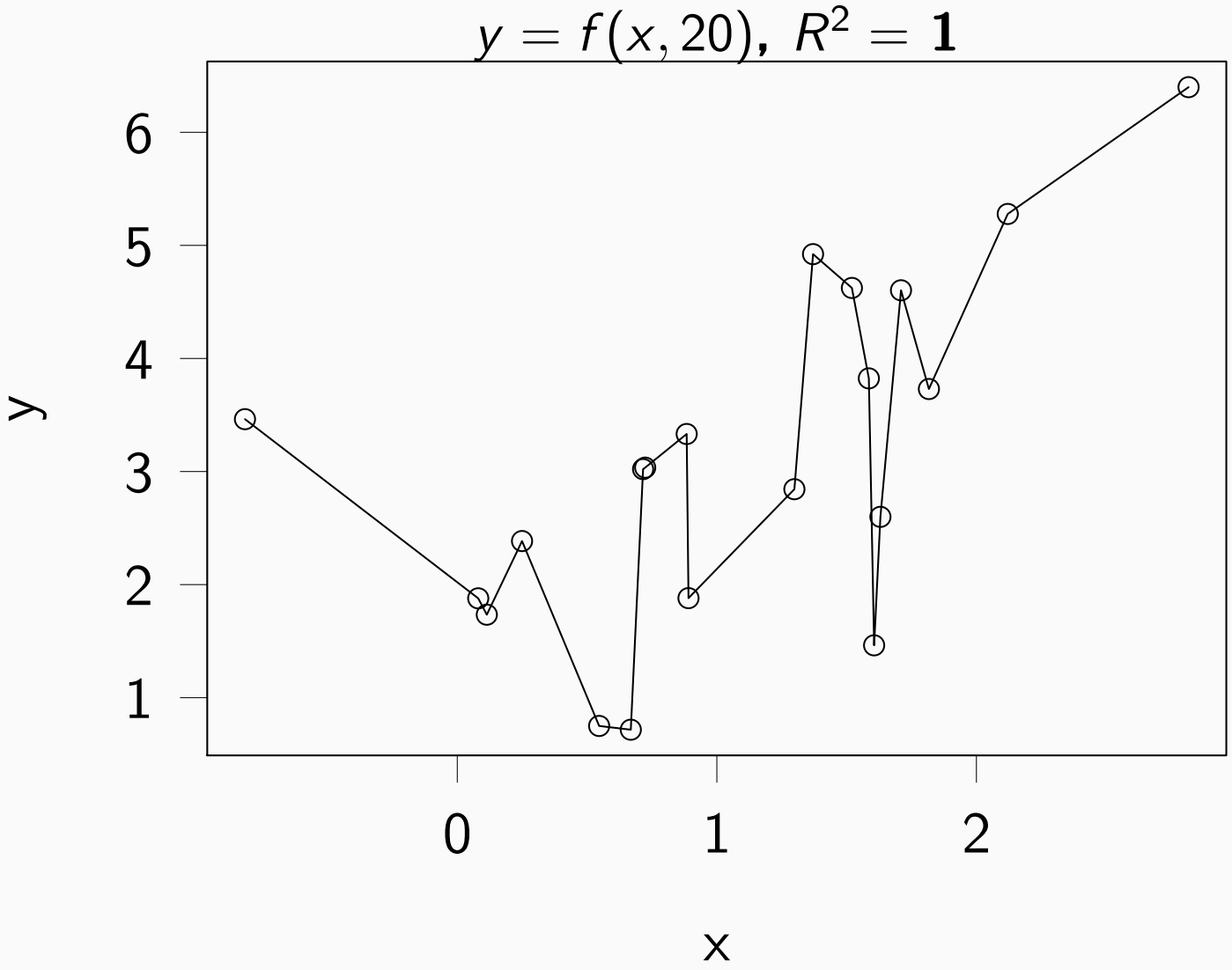
Overfitting



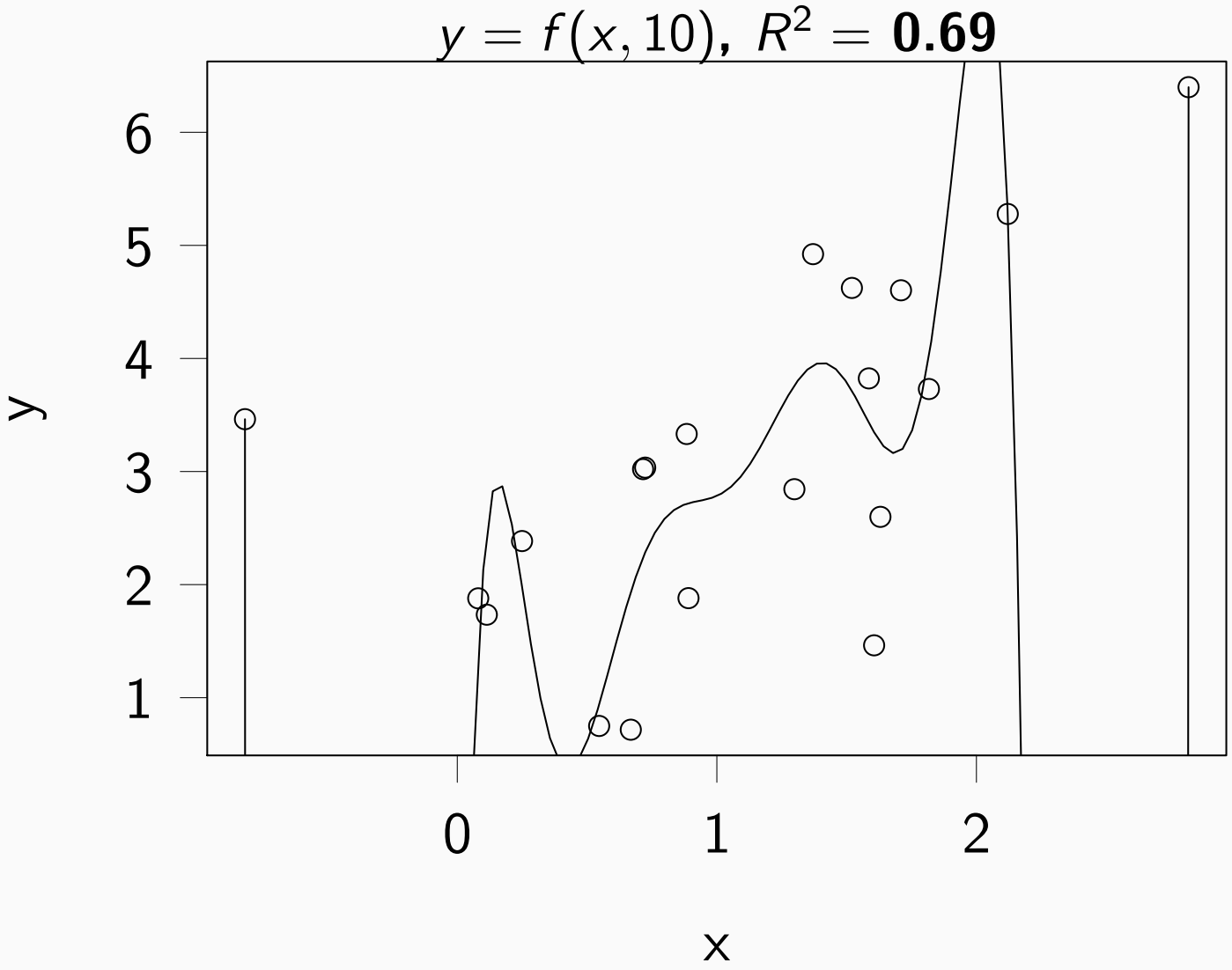
Overfitting



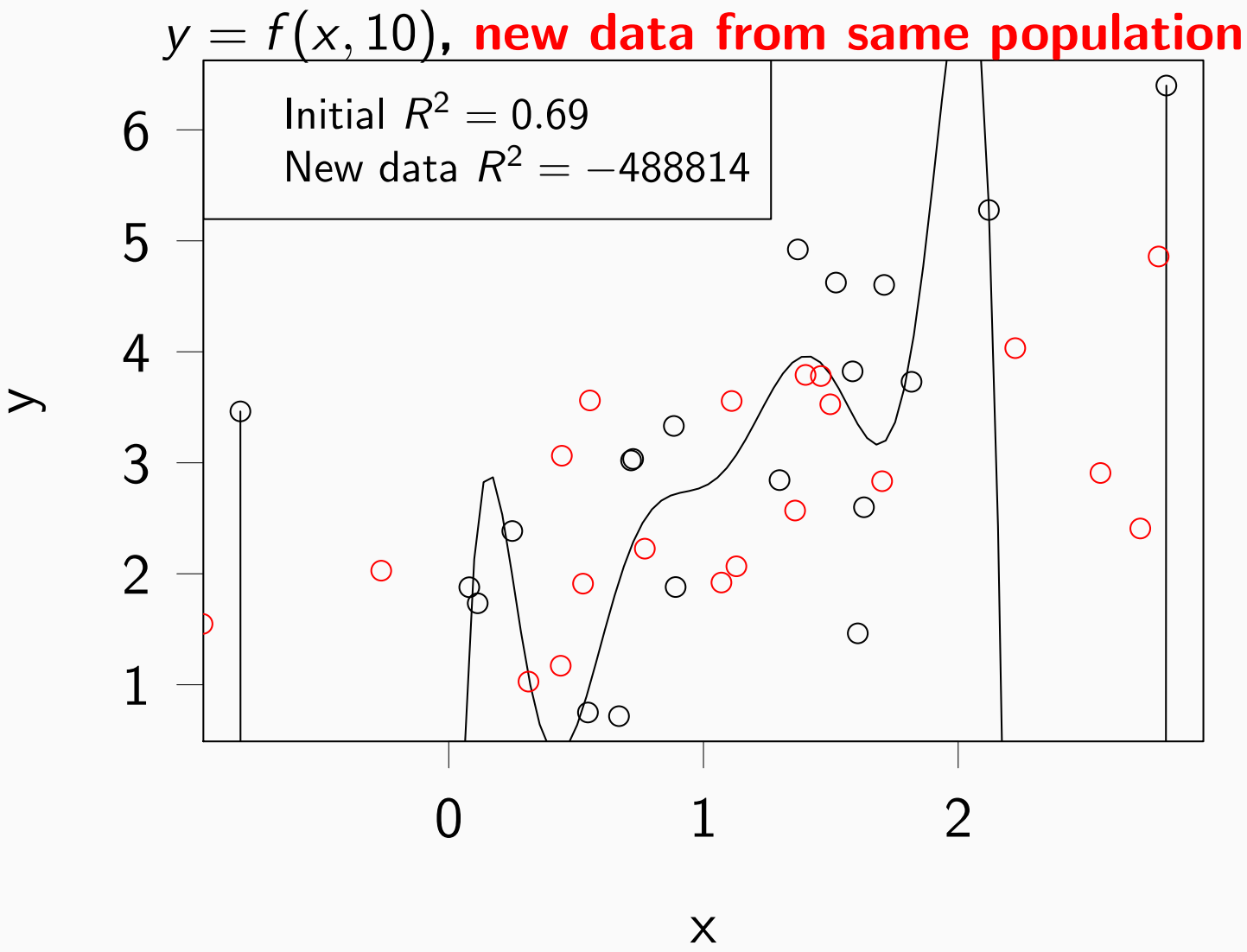
Overfitting



Overfitting: What is wrong? Part 1



Overfitting: What is wrong? Part 1



Overfitting: What is wrong? Part 2

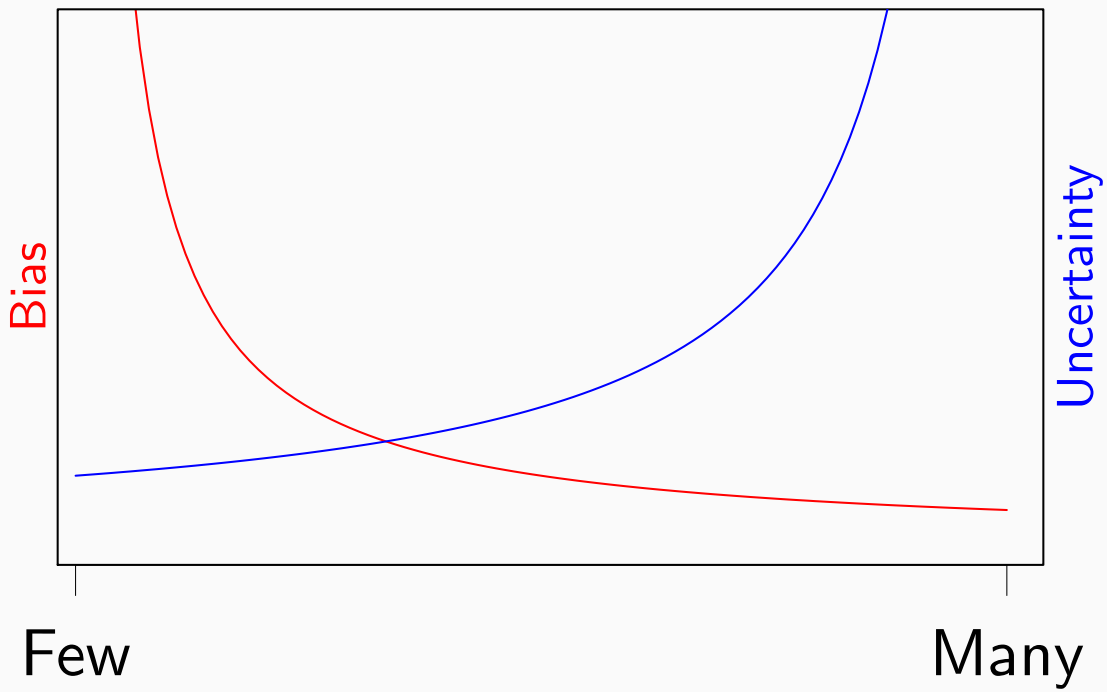
	Estimate	Std. Error	t value	p.value
(Intercept)	-6.06	8.44	-0.72	0.49
poly(x, 10, raw = TRUE)1	149.34	155.39	0.96	0.36
poly(x, 10, raw = TRUE)2	-855.96	901.62	-0.95	0.37
poly(x, 10, raw = TRUE)3	1984.17	2164.20	0.92	0.38
poly(x, 10, raw = TRUE)4	-1647.96	1928.23	-0.85	0.41
poly(x, 10, raw = TRUE)5	-1070.35	1116.87	-0.96	0.36
poly(x, 10, raw = TRUE)6	3444.36	3796.44	0.91	0.39
poly(x, 10, raw = TRUE)7	-3113.36	3503.16	-0.89	0.40
poly(x, 10, raw = TRUE)8	1416.27	1611.35	0.88	0.40
poly(x, 10, raw = TRUE)9	-328.50	376.33	-0.87	0.41
poly(x, 10, raw = TRUE)10	30.84	35.50	0.87	0.41

Overfitting: What is wrong

Very good fit, but model USELESS

- Worse prediction than just taking the mean
- No biological interpretation
- Huge standar error, large p-values

“Less bias = More uncertainty”



What is the best compromise?

What is the best compromise?

Stepwise regression, one of multiple approaches:
Start from a simple model

1. Add covariates, once at the time, fit the models
2. Keep the "most significant" covariate
3. If no covariate is significant, stop

Try stepwise regression

```
datsub1 <- read.csv("datsub1.csv")
str(datsub1)
```

Try stepwise regression

```
datsub1 <- read.csv("datsub1.csv")
str(datsub1)
```

Stepwise selection part 1:

```
summary(lm(y ~ x1, data = datsub1))
summary(lm(y ~ x2, data = datsub1))
summary(lm(y ~ x3, data = datsub1))
summary(lm(y ~ x4, data = datsub1))
```

Try stepwise regression

```
datsub1 <- read.csv("datsub1.csv")
str(datsub1)
```

Stepwise selection part 1:

```
summary(lm(y ~ x1, data = datsub1))
summary(lm(y ~ x2, data = datsub1))
summary(lm(y ~ x3, data = datsub1))
summary(lm(y ~ x4, data = datsub1))
```

Stepwise selection part 2:

```
summary(lm(y ~ x2 + x1, data = datsub1))
summary(lm(y ~ x2 + x3, data = datsub1))
summary(lm(y ~ x2 + x4, data = datsub1))
```

Try stepwise regression

```
datsub1 <- read.csv("datsub1.csv")
str(datsub1)
```

Stepwise selection part 1:

```
summary(lm(y ~ x1, data = datsub1))
summary(lm(y ~ x2, data = datsub1))
summary(lm(y ~ x3, data = datsub1))
summary(lm(y ~ x4, data = datsub1))
```

Stepwise selection part 2:

```
summary(lm(y ~ x2 + x1, data = datsub1))
summary(lm(y ~ x2 + x3, data = datsub1))
summary(lm(y ~ x2 + x4, data = datsub1))
```

Stepwise selection part 3:

```
summary(lm(y ~ x2 + x4 + x1, data = datsub1))
summary(lm(y ~ x2 + x4 + x3, data = datsub1))
```

Try stepwise regression

Based on this approach the best model is:

```
summary(lm(y ~ x2 + x4, data = datsub1))
```

Compare this to the full model:

```
summary(lm(y ~ x1 + x2 + x3 + x4, data = datsub1))
```

Which is best at predicting y ? (you can use the function `predict()` or look at the R-squared in `summary()`)

Try stepwise regression

The full model explains more variation in y in `datsub1`. But can we trust the full model for new data?

Try stepwise regression

The full model explains more variation in y in `datsub1`. But can we trust the full model for new data?

The dataset `datsub2` comes from the same population. Make predictions of y in `datsub2` based on the best and the full models, and compare them to the true values of y in `datsub2`. You could use the function `predict`, `plot`, and `cor`.

Start for the full model:

```
datsub2 <- read.csv("datsub2.csv")
predictedy <- predict(lm(y ~ x1 + x2 + x3 + x4, data = datsub1),
                      newdata = datsub2)
```

Summary: Why model selection

- Adding predictors increases fit to the response, in the current data
- But too many predictors:
  - DECREASE fit in new data (from the same population)
  - Hinder biological interpretation
  - Increases esimtation uncertainty (larger SE and p-values)
- Model selection aims to balance fit and generalisation

Why model selection?

Information criteria vs. stepwise selection

Model selection and causal inference

Model selection aims to balance fit and generalisation: How?

Stepwise regression is an option. But not very good.

Load `threepreddat.csv` and apply the previous stepwise regression method to model selection. What is the best model to predict the response?

Model selection aims to balance fit and generalisation: How?

Stepwise regression is an option. But not very good.

Load `threepreddat.csv` and apply the previous stepwise regression method to model selection. What is the best model to predict the response?

Compare your best model to:

```
summary(lm(response ~ pred2 + pred3, data = threepreddat))
```

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Instead, use Information Criteria

- Akaike information criterion (AIC), invented in the 1970s, maximizes prediction

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Instead, use Information Criteria

- Akaike information criterion (AIC), invented in the 1970s, maximizes prediction
- Later BIC, DIC, TIC. . . maximizes different aspects of model performance

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Instead, use Information Criteria

- Akaike information criterion (AIC), invented in the 1970s, maximizes prediction
- Later BIC, DIC, TIC. . . maximizes different aspects of model performance
- AIC $= 2\times$Number of parameters - $2\times$ log(model likelihood)

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Instead, use Information Criteria

- Akaike information criterion (AIC), invented in the 1970s, maximizes prediction
- Later BIC, DIC, TIC. . . maximizes different aspects of model performance
- AIC = 2×Number of parameters - 2× log(model likelihood)
- Smaller is better

Information criteria

Do NOT use stepwise regression for model selection

- Different version of stepwise regression often disagree
- Only nested models are compared
- Sometimes the best models cannot be discovered stepwise

Instead, use Information Criteria

- Akaike information criterion (AIC), invented in the 1970s, maximizes prediction
- Later BIC, DIC, TIC. . . maximizes different aspects of model performance
- AIC $= 2\times$Number of parameters - $2\times$ log(model likelihood)
- Smaller is better
- Only relative measure, no absolute meaning

AIC in R

Find the best AIC-model from `threepreddat`. For instance:

```
AIC(lm(response ~ pred1, data = threepreddat))
```

```
[1] 151.1339
```

AIC best practice:

False positives happen $\rightarrow$

if you compare many models, "spurious" one can be best by pure chance

AIC best practice:

False positives happen $\rightarrow$

if you compare many models, "spurious" one can be best by pure chance

Minimize the risks:

- Compare only models that make biological sense (e.g., avoid complex interactions)

AIC best practice:

False positives happen $\rightarrow$

if you compare many models, "spurious" one can be best by pure chance

Minimize the risks:

- Compare only models that make biological sense (e.g., avoid complex interactions)
- If you know enough, pre-select small set of models representing competing hypotheses

AIC best practice:

False positives happen $\rightarrow$

if you compare many models, "spurious" one can be best by pure chance

Minimize the risks:

- Compare only models that make biological sense (e.g., avoid complex interactions)
- If you know enough, pre-select small set of models representing competing hypotheses
- **Confirmation:** AIC-model selection on half your dataset, then fit best model on the second half

AIC best practice:

False positives happen $\rightarrow$

if you compare many models, "spurious" one can be best by pure chance

Minimize the risks:

- Compare only models that make biological sense (e.g., avoid complex interactions)
- If you know enough, pre-select small set of models representing competing hypotheses
- **Confirmation:** AIC-model selection on half your dataset, then fit best model on the second half
- Acknowledge results are inconlusive is AIC-difference below 2 (or below 5)

Practice: form groups of 2-3 people

Load `VoleWeight.csv`. We want to understand what factors explain variation in individual body weight. Think about what models could make sense. Use half the dataset for AIC comparison of those models. Check model performance on the second half.

Summary: Information criteria vs. stepwise selection

- Stepwise

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models
  - **does NOT find the "best" model**

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models
  - **does NOT find the "best" model**
- AIC

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models
  - **does NOT find the "best" model**
- AIC
  - Compare all models, nested or not

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models
  - **does NOT find the "best" model**
- AIC
  - Compare all models, nested or not
  - finds model best at predicting the response

Summary: Information criteria vs. stepwise selection

- Stepwise
  - Different approches give inconsistent results
  - compares only nested models
  - **does NOT find the "best" model**
- AIC
  - Compare all models, nested or not
  - finds model best at predicting the response
  - works better if competing models are pre-selected

Why model selection?

Information criteria vs. stepwise selection

Model selection and causal inference

Competing hypotheses

How does respiration rate scale with body mass in mammals? For a while researchers fought over two ideas, respiration could increase as a function of $mass^{2/3}$ or as a function of $mass^{3/4}$; while maybe the shape of the animals played a role. Let's find out!.

Load `metabo.csv` and compare models through AIC selection.

NB: you can fit exponents of a predictors using the function `I()`. For instance, for the exponent 0.5 of x:

```
lm(y ~ I(x^(1/2)))
```

Useful tools

Package `MuMIn` useful for model selection

```
install.packages("MuMIn") library(MuMIn)
```

Useful tools

Package `MuMIn` useful for model selection

```
install.packages("MuMIn") library(MuMIn)
```

1. AICc

- AIC is biased for small sample size
- AICc ("second-order AIC") when sample size / number of parameters is less than 40
- `MuMIn::AICc()`

Useful tools

Package `MuMIn` useful for model selection

```
install.packages("MuMIn") library(MuMIn)
```

1. AICc

- AIC is biased for small sample size
- AICc ("second-order AIC") when sample size / number of parameters is less than 40
- `MuMIn::AICc()`

2. dredge

- Automate model selection
- Convenient BUT many competing models, some may not make sense
- `MuMIn::dredge()`

Try automated model selection

Try to use `dredge()`, with selection based on `AICc`, to automate model selection on the vole data.

For some reason you first need to run:

```
options(na.action="na.fail")
```

Do you find the same result?

If you get bored, go to next slide for a challenge!

Challenge

In general, when a response variable is independent of the response there is a 5% probability to find a p-value below 0.05 (that's a false positive). Let's see for ourselves that it does not work if we do model selection first!

The code below creates one dataset with a response variable independent of the "mainpredictor" (the thing we want to test) and of control variables. We fit a full model, test many control variable combinations with `dredge()` and extract the p-value for "mainpredictor". Create a for-loop to look at the frequency of significant p-values!

```
nobs <- 60
mainpredictor <- rnorm(nobs)
control1 <- rnorm(nobs) ; control2 <- rnorm(nobs)
control3 <- rnorm(nobs) ; control4 <- rnorm(nobs)
control5 <- rnorm(nobs) ; response <- rnorm(nobs)

mfull <- lm(response ~ mainpredictor +
              control1*control2*control3 + control4*control5)

modall <- dredge(mfull, fixed = "mainpredictor")
summary(get.models(modall, 1)[[1]])$coefficients[2,4]#the pvalue
```

“There is madness in our methods”

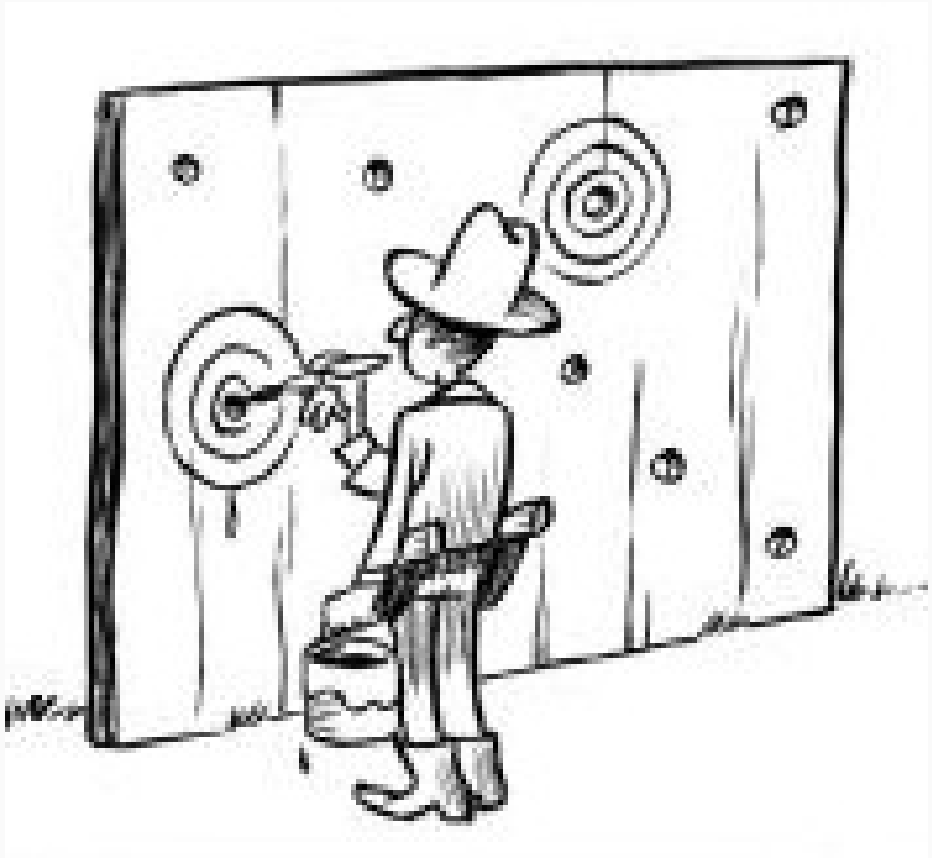


Figure 1: Null-hypothesis testing after model selection ©Dirk Jan-Hoek

https://methodsblog.com/2015/11/26/madness-in-our-methods/

Summary: Model selection and inference

- AIC/AICc best for exploratory / predictive models
- AIC/AICc alone can be used for causal inference if models are all meaningful competing biological hypotheses
- After AIC/AICc selection p-values are wrong
- Correct p-values and standard errors MUST be computed on new data (Confirmatory model)

Want more practice?

Use AIC/AICc to:

- Tell what drives the increase in number of babies in `babies.csv`
- What drives bird abundance (ABUND) in `loyn.csv`
- **Challenge:** How do differences in AIC between nested models scale with p-values for the extra predictor?

Want to know more?

Very good, but very pro:

- Burnham and Anderson, 2002, Model Selection and Multimodel Inference: A Practical Information-theoretic Approach

AIC is not always the best choice:

- **"Your goal matters in the choice between AIC, BIC, p-values...:"** Aho & al. A graphical framework for model selection criteria and significance tests: Refutation, confirmation and ecology. Methods in Ecology and Evolution. 2017;8:47–56.
- **"Careful when combining estimates from different models"**: Cade. Model averaging and muddled multimodel inferences. Ecology. 2015;96(9):2370–82.

Before you leave:

1. Write one thing you liked and one you disliked on a sticky note
2. Be sure you signed the presence sheet, especially if you want credit for the HDR career development framework
3. Leave your email address if you want to join the Slack channel
4. Past workshops at `https://timotheenivalis.github.io/RSB-R-Stats-Biology/`

Thank you!!