Variable and model selection

Francisco Rodríguez-Sánchez

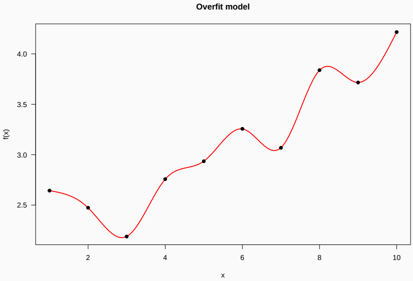
https://frodriguezsanchez.net

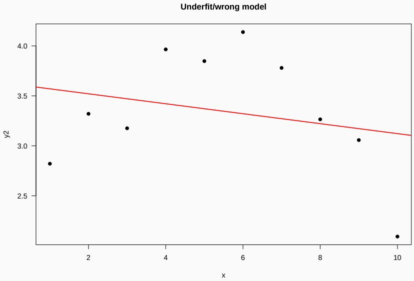
Overfitting and balanced model complexity

- On one hand, we want to **maximise fit**.

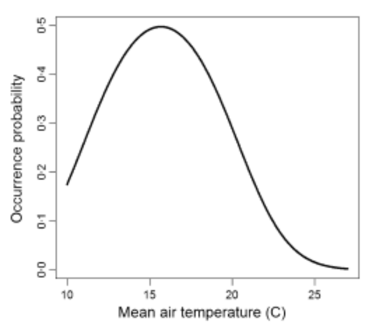
- On one hand, we want to **maximise fit**.
- On the other hand, we want to **avoid overfitting** and overly complex models.

Overfitting and balanced model complexity



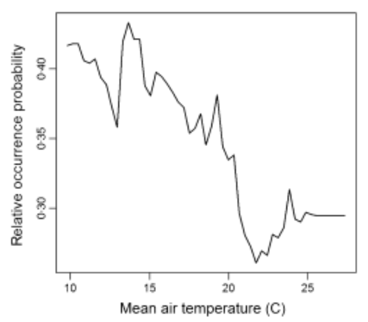


GLMM



Wenger & Olden (2012)

Random forests



Overfitted models will work badly on new data

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out…)
- **Information Criteria**:

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)
- **Information Criteria**:
    - AIC

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)
- **Information Criteria**:
  - AIC
  - BIC

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)
- **Information Criteria**:
  - AIC
  - BIC
  - DIC

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)
- **Information Criteria**:
  - AIC
  - BIC
  - DIC
  - WAIC...

Evaluating models' predictive accuracy

- **Cross-validation** (k-fold, leave one out...)
- **Information Criteria**:
    - AIC
    - BIC
    - DIC
    - WAIC...
- All these methods have flaws!

AIC (Akaike Information Criteria)

$$AIC = -2 * LogLikelihood + 2K$$

- First term: **model fit**

AIC (Akaike Information Criteria)

$$AIC = -2 * LogLikelihood + 2K$$

- First term: **model fit**
- **K = number of parameters** (penalisation for model complexity)

AIC (Akaike Information Criteria)

$$AIC = -2 * LogLikelihood + 2K$$

- First term: **model fit**
- **K = number of parameters** (penalisation for model complexity)
- Lower is better

AIC (Akaike Information Criteria)

$$AIC = -2 * LogLikelihood + 2K$$

- First term: **model fit**
- **K = number of parameters** (penalisation for model complexity)
- Lower is better
- AIC biased towards complex models.

AIC (Akaike Information Criteria)

$$AIC = -2 * LogLikelihood + 2K$$

- First term: **model fit**
- **K = number of parameters** (penalisation for model complexity)
- Lower is better
- AIC biased towards complex models.
- AICc recommended with 'small' sample sizes (n/p < 40). But see Richards 2005

Problems of IC

- No information criteria is panacea: all have problems.

Problems of IC

- No information criteria is panacea: all have problems.
- They estimate *average* out-of-sample prediction error. But errors can differ substantially within dataset.

Problems of IC

- No information criteria is panacea: all have problems.
- They estimate *average* out-of-sample prediction error. But errors can differ substantially within dataset.
- Sometimes better models rank poorly (e.g. see Gelman et al. 2013). Combine with **thorough model checks**.

So which variables should enter my model?

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.

Choosing predictors

- Choose variables based on background knowledge, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of reasonable candidate models.

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)
    - If |r| > 0.5 - 0.7, consider leaving one variable out, but keep it in mind when interpreting model results.

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)
  - If |r| > 0.5 - 0.7, consider leaving one variable out, but keep it in mind when interpreting model results.
  - Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)
    - If |r| > 0.5 - 0.7, consider leaving one variable out, but keep it in mind when interpreting model results.
    - Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
    - Many methods available, e.g. sequential, ridge regression...

Choosing predictors

- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)
  - If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
  - Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
  - Many methods available, e.g. sequential, ridge regression...
  - Measurement error can seriously complicate things (Biggs et al 2009; Freckleton 2011)

Choosing predictors

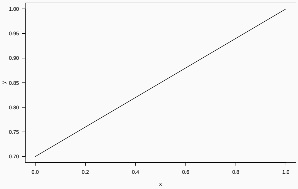
- Choose variables based on **background knowledge**, rather than throwing plenty of them in a fishing expedition.
- Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- Number of variables **balanced with sample size** (e.g. at least 10 - 30 obs per param)
- Assess collinearity between predictors (Dormann et al 2013)
    - If |r| > 0.5 - 0.7, consider leaving one variable out, but keep it in mind when interpreting model results.
    - Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
    - Many methods available, e.g. sequential, ridge regression...
    - Measurement error can seriously complicate things (Biggs et al 2009; Freckleton 2011)
- For predictors with large effects, **consider interactions**.

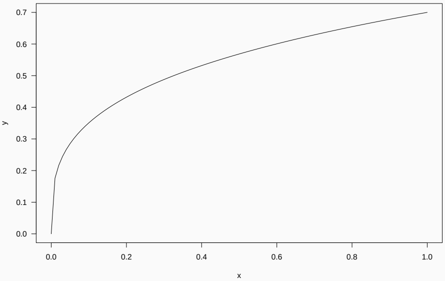
Think about the shape of relationships

y ~ x + z

Really? Not everything has to be linear! Actually, it often is not.

Think about shape of relationship.





Removing predictors

Stepwise regression has many problems

- Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? J. Animal Ecology.

Stepwise regression has many problems

- Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? J. Animal Ecology.
- Mundry & Nunn (2009) Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. Am Nat.

Stepwise regression has many problems

- Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? J. Animal Ecology.
- Mundry & Nunn (2009) Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. Am Nat.
- This includes `stepAIC` (e.g. Dahlgren 2010; Burnham et al 2011; Hegyi & Garamszegi 2011).

Other common bad practices

- Testing bivariate relationships before building multivariable model

Heinze & Dunkler 2016

Other common bad practices

- Testing bivariate relationships before building multivariable model
- Removing non-significant predictors

Heinze & Dunkler 2016

Removing predictors?

- Always **keep 'core' predictors** (based on previous knowledge)

Heinze et al 2018

Removing predictors?

- Always **keep 'core' predictors** (based on previous knowledge)
- If ratio sample size/number of predictors is low (<10 EPP), avoid variable selection (too unstable)

Heinze et al 2018

Removing predictors?

- Always **keep ‘core’ predictors** (based on previous knowledge)
- If ratio sample size/number of predictors is low (<10 EPP), avoid variable selection (too unstable)
- If performing variable selection, always **assess stability** (bootstrap, etc)

Heinze et al 2018

Summary

1. Choose meaningful variables

Summary

1. Choose meaningful variables
    - Beware collinearity

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets
    - Don't assume linear effects: think about appropriate functional relationships

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets
    - Don't assume linear effects: think about appropriate functional relationships
    - Consider interactions for strong main effects

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets
    - Don't assume linear effects: think about appropriate functional relationships
    - Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets
    - Don't assume linear effects: think about appropriate functional relationships
    - Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check fitted models thoroughly

Summary

1. Choose meaningful variables
    - Beware collinearity
    - Keep good n/p ratio
2. Generate global model or (small) set of candidate models
    - Avoid stepwise and all-subsets
    - Don't assume linear effects: think about appropriate functional relationships
    - Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check fitted models thoroughly
5. Always report effect sizes