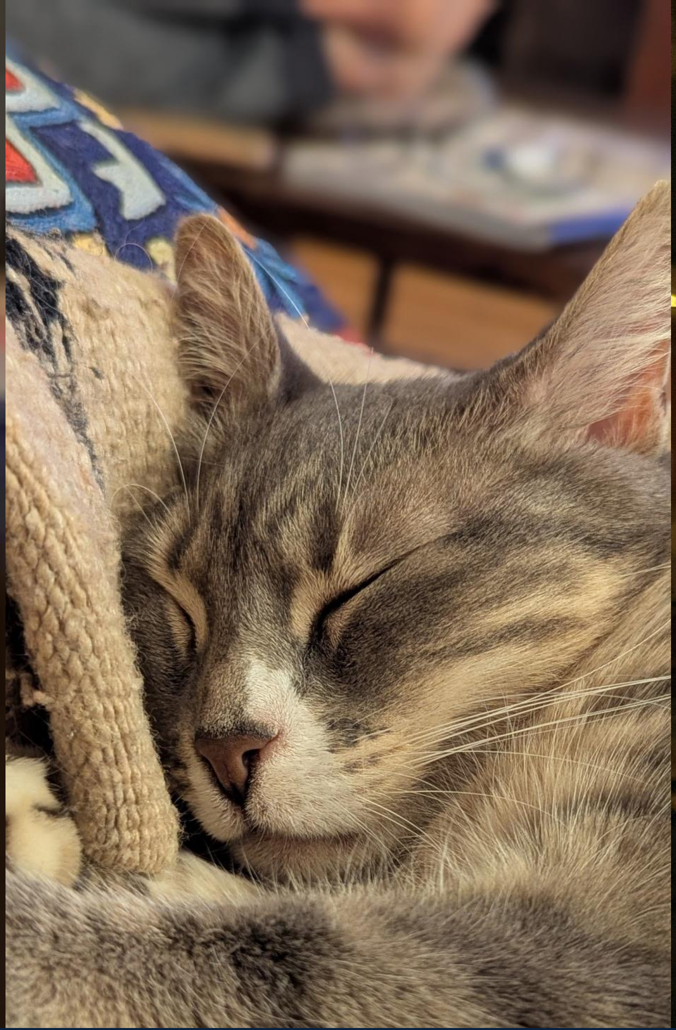
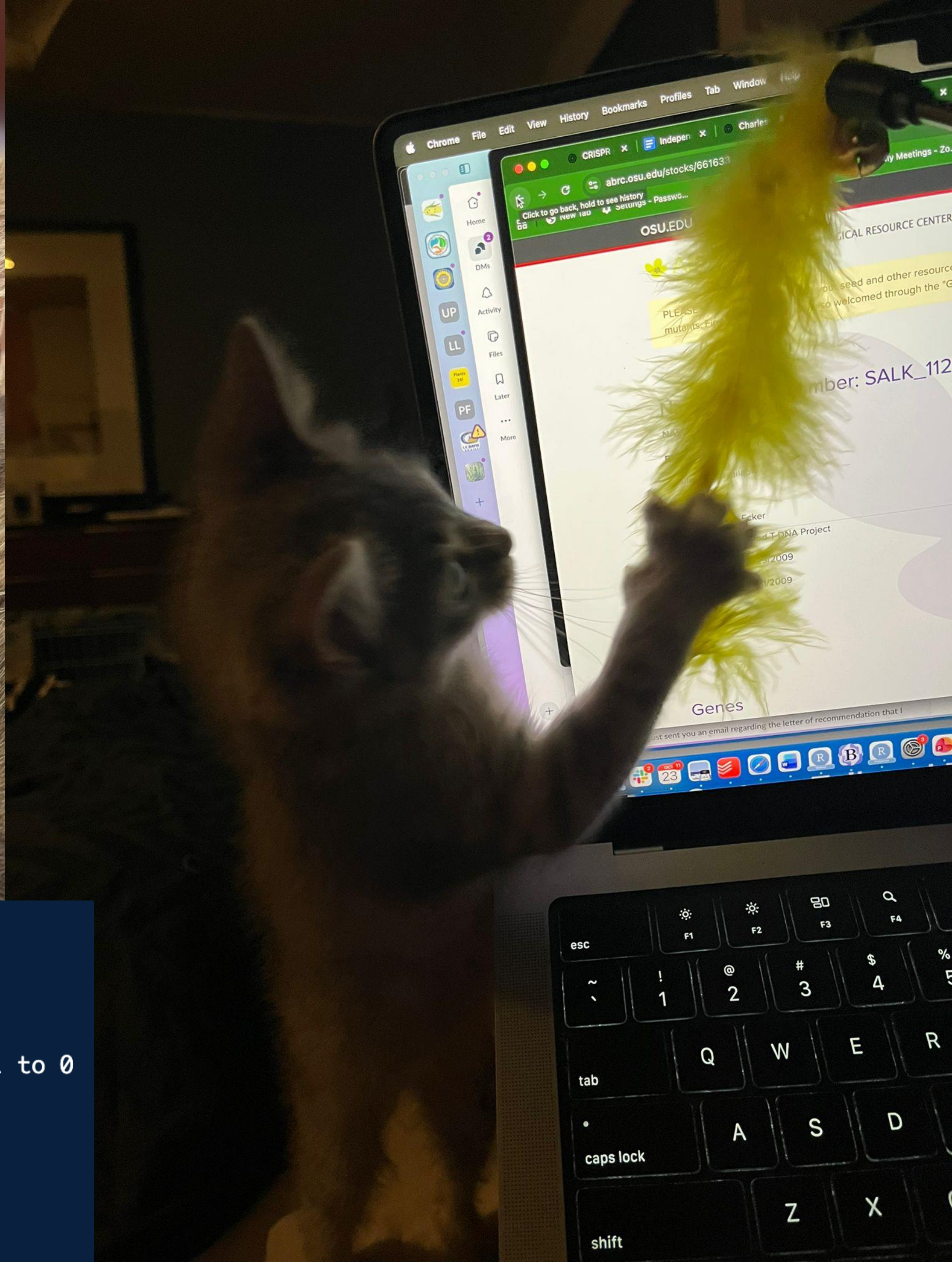
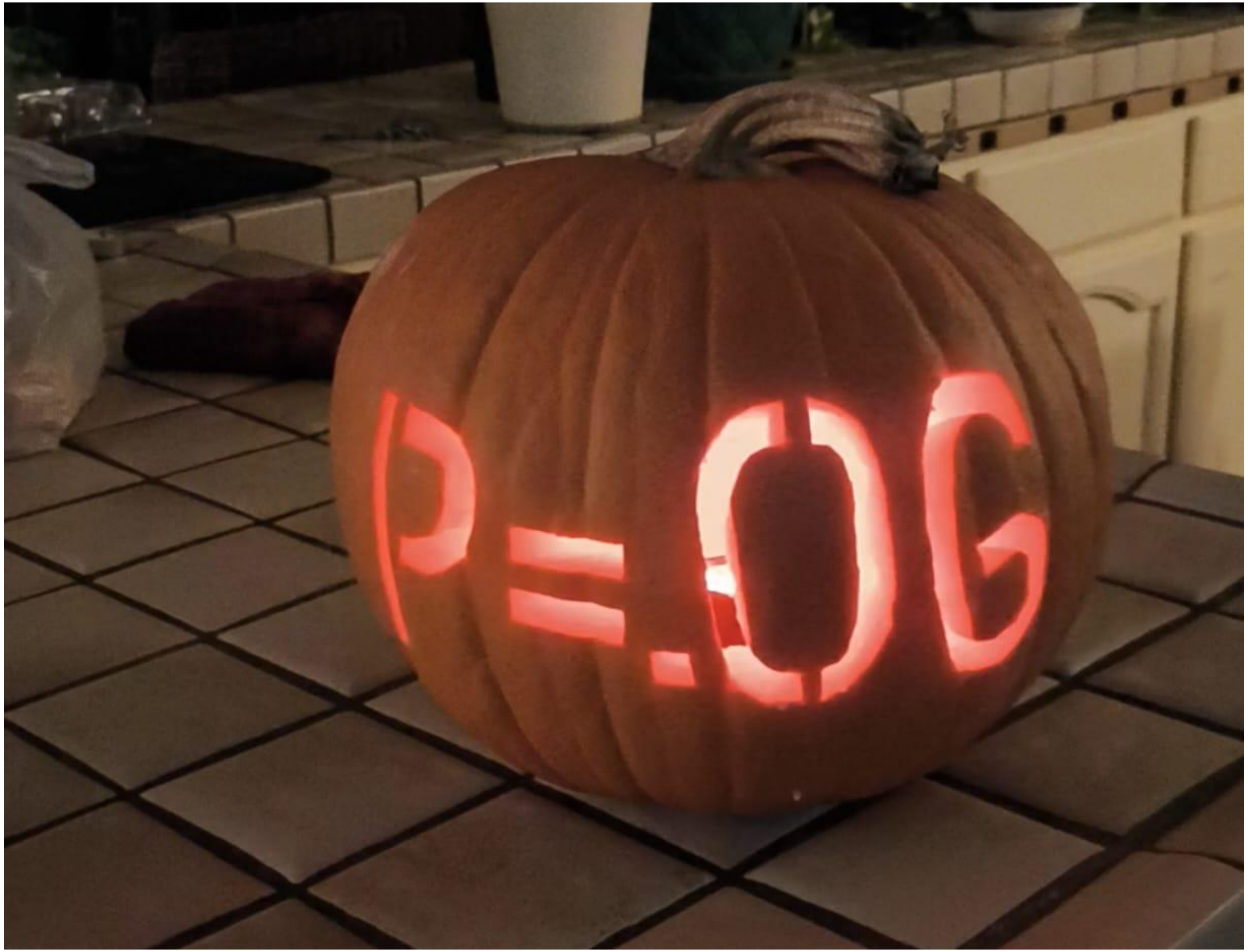




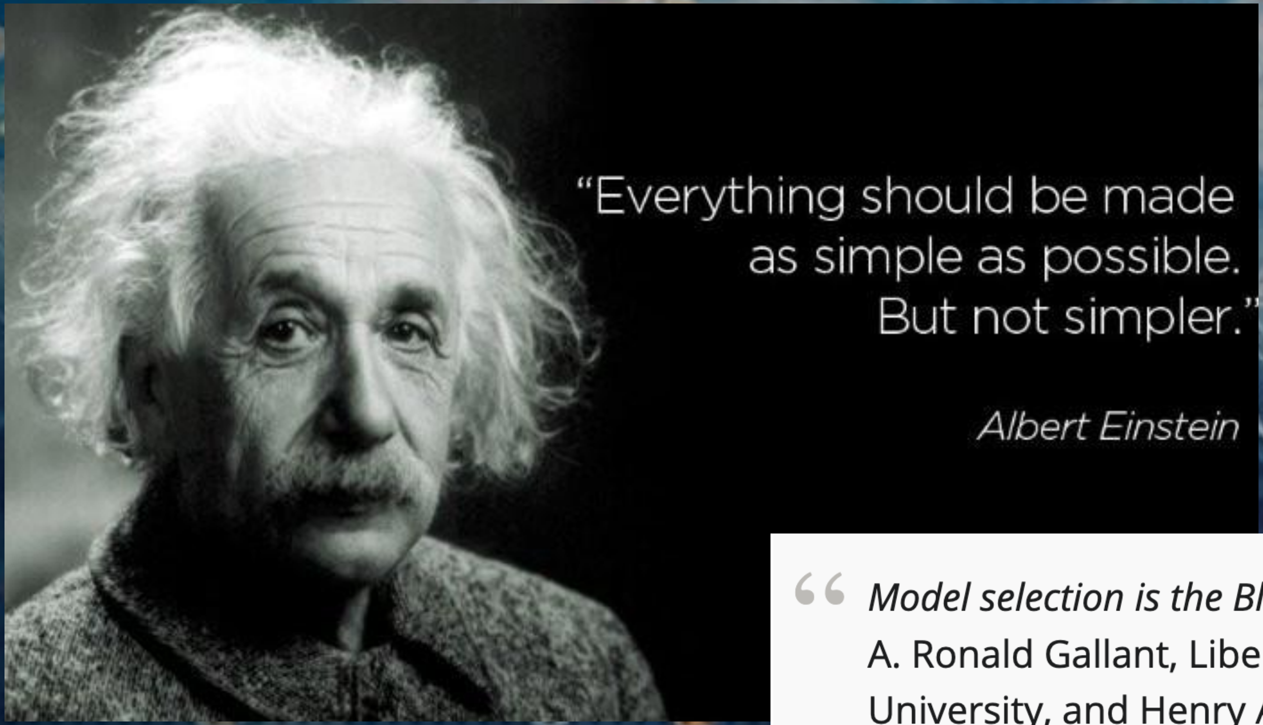
```
	One Sample t-test

data:  cuteness
t = 104.84, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.75474 10.13111
sample estimates:
mean of x
 9.942927
```



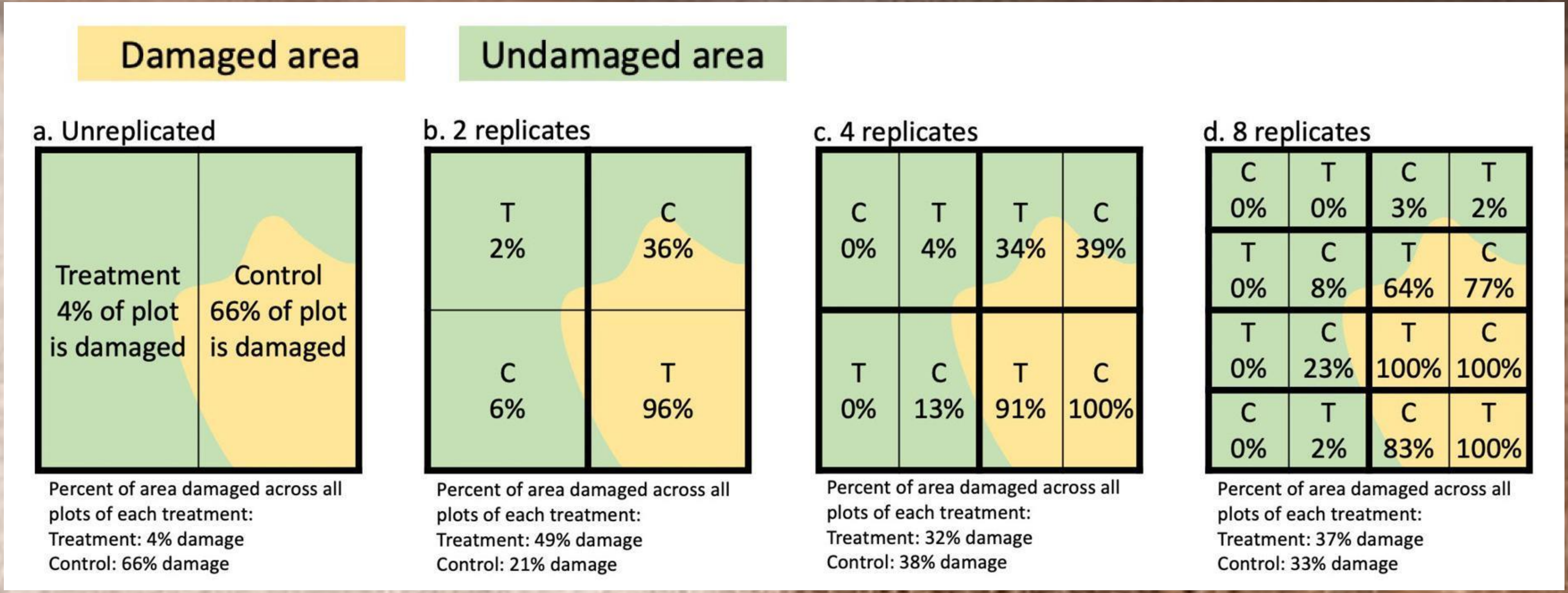
VITOR GONÇALVES DA SILVA (he/him)
Agricultural Engineer, M.Sc
PhD Candidate in AgEng

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Model Selection



> *Model selection is the Black Hole of Statistics*.
> A. Ronald Gallant, Liberal Arts Professor of Economics, Pennsylvania State University, and Henry A. Latane Distinguished Professor (emeritus) of Economics, UNC-Chapel Hill (*personal communication*)



extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. Use a priori covariates without overfitting
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. Handle confounding & multicollinearity
5. Decide random vs fixed effects sensibly

extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. **Choose a defensible model that respects the design**
2. Use a priori covariates without overfitting
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. Handle confounding & multicollinearity
5. Decide random vs fixed effects sensibly

1. Design is key!

Identify factors

Treatment (usually fixed)

Block (nuisance; often random effects)

Plan before data peek

Pre-declare covariates & key interactions

(if your design includes a block have in it your model)

extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. **Use *a priori* covariates without overfitting**
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. Handle confounding & multicollinearity
5. Decide random vs fixed effects sensibly

2. *A Priori* Covariates and interactions

Include covariates for a reason – you don't have to include every possible interaction term between them, only if meaningful

Hierarchy principle: if include interaction A*Z, also include A and Z

extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. Use *a priori* covariates without overfitting
3. **Avoid non-identifiable terms (e.g., n≤1 per combo)**
4. Handle confounding & multicollinearity
5. Decide random vs fixed effects sensibly

3. Observations per Factorial Combination

Rule: If **n ≤ 1 per cell**, interactions at that level aren't testable
(Residual df → 0; saturated model)

```
> set.seed(1)
> dat <- expand.grid(Treatment = c("A","B"), Genotype = c("G1","G2"))
> dat$y <- rnorm(nrow(dat))  # exactly 1 obs per TreatmentxGenotype
>
> m <- lm(y ~ Treatment * Genotype, data = dat)
> summary(m)        # Residual df = 0 (saturated); SEs and tests are not meaningful

Call:
lm(formula = y ~ Treatment * Genotype, data = dat)

Residuals:
ALL 4 residuals are 0: no residual degrees of freedom!

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.6265        NaN     NaN      NaN
TreatmentB              0.8101        NaN     NaN      NaN
GenotypeG2             -0.2092        NaN     NaN      NaN
TreatmentB:GenotypeG2   1.6208        NaN     NaN      NaN

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA

> anova(m)          # F-tests not available with 0 residual df
Analysis of Variance Table

Response: y
                   Df  Sum Sq Mean Sq F value Pr(>F)
Treatment           1 2.62603 2.62603     NaN    NaN
Genotype            1 0.36148 0.36148     NaN    NaN
Treatment:Genotype  1 0.65676 0.65676     NaN    NaN
Residuals           0 0.00000     NaN
Warning message:
In anova.lm(m) : ANOVA F-tests on an essentially perfect fit are unreliable
```

```
> table(dat$Genotype, dat$Treatment)

    A B
 G1 1 1
 G2 1 1
```

!! Not enough reps to do interaction !!

extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. Use *a priori* covariates without overfitting
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. **Handle confounding & multicollinearity**
5. Decide random vs fixed effects sensibly

4. Multicollinearity & Confounding

Complete confounding: e.g., each Block contains only one Treatment → cannot separate effects
Symptoms: rank deficiency, NA coefficients,

```
set.seed(1)
n_per_level <- 4
Block <- gl(3, n_per_level, labels = paste0("B",1:3))
Treatment <- gl(3, n_per_level, labels = paste0("T",1:3))    # PERFECT confounding: each block = one treatment
y <- rnorm(length(Block))
dat <- data.frame(y, Block, Treatment)

m_bad <- lm(y ~ Block + Treatment, data = dat)
summary(m_bad)      # Coefficients for one factor become NA (rank deficiency)
alias(m_bad)        # Shows exact aliasing between Block and Treatment
```

```
> summary(m_bad)     # Coefficients for one factor become NA (rank deficien

Call:
lm(formula = y ~ Block + Treatment, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01251 -0.49546  0.03429  0.49499  1.05976 

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29892    0.37444   0.798    0.445
BlockB2     -0.66347    0.52953  -1.253    0.242
BlockB3      0.05348    0.52953   0.101    0.922
TreatmentT2       NA         NA      NA       NA
TreatmentT3       NA         NA      NA       NA

Residual standard error: 0.7489 on 9 degrees of freedom
Multiple R-squared:  0.2018,    Adjusted R-squared:  0.02443 
F-statistic: 1.138 on 2 and 9 DF,  p-value: 0.3627
```

extension.oregonstate.edu/

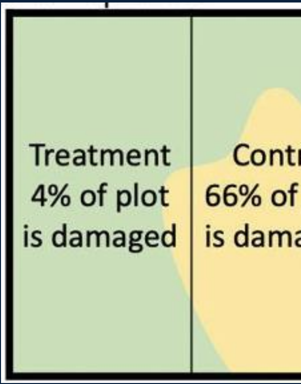
Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. Use *a priori* covariates without overfitting
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. **Handle confounding & multicollinearity**
5. Decide random vs fixed effects sensibly

4. Multicollinearity & Confounding

Complete confounding: e.g., each Block contains only one Treatment → cannot separate effects
Symptoms: rank deficiency, NA coefficients,

```
> summary(m_bad)     # Coefficients for one factor become NA (rank deficien

Call:
lm(formula = y ~ Block + Treatment, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01251 -0.49546  0.03429  0.49499  1.05976 

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29892    0.37444   0.798    0.445
BlockB2     -0.66347    0.52953  -1.253    0.242
BlockB3      0.05348    0.52953   0.101    0.922
TreatmentT2       NA         NA      NA       NA
TreatmentT3       NA         NA      NA       NA

Residual standard error: 0.7489 on 9 degrees of freedom
Multiple R-squared:  0.2018,    Adjusted R-squared:  0.02443 
F-statistic: 1.138 on 2 and 9 DF,  p-value: 0.3627
```




```
set.seed(1)
n_per_level <- 4
Block <- gl(3, n_per_level, labels = paste0("B",1:3))
Treatment <- gl(3, n_per_level, labels = paste0("T",1:3))    # PERFECT confounding: each block = one treatment
y <- rnorm(length(Block))
dat <- data.frame(y, Block, Treatment)

m_bad <- lm(y ~ Block + Treatment, data = dat)
summary(m_bad)      # Coefficients for one factor become NA (rank deficiency)
alias(m_bad)        # Shows exact aliasing between Block and Treatment
```

extension.oregonstate.edu/

Model Selection for Experiments e.g. (with Blocks)

1. Choose a defensible model that respects the design
2. Use *a priori* covariates without overfitting
3. Avoid non-identifiable terms (e.g., n≤1 per combo)
4. Handle confounding & multicollinearity
5. **Decide random vs fixed effects sensibly**

5. Random vs fixed effects

Treatment: fixed (specific levels of interest)

Blocks: usually random (sampled nuisance structure; generalization)

Genotypes:

Fixed: only care about those specific lines

Random: represent a population; want variance components/BLUPs

```r
# Fixed BLOCKS (ANOVA-style)
m_fixed_block <- lm(y ~ Treatment + Block, data = dat)

# Random BLOCKS
# install.packages("lme4")
library(lme4)
m_rand_block  <- lmer(y ~ Treatment + (1|Block), data = dat)
```

Let yield on plot i in site j be

$$y_{ij} = \underbrace{X_{ij}\beta}_{\text{fixed effects (e.g., N rate, variety)}} + \underbrace{u_j}_{\text{site effect}} + \underbrace{\varepsilon_{ij}}_{\text{residual noise}}$$

with

$$u_j \sim \mathcal{N}(0, \sigma^2_{\text{site}}), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

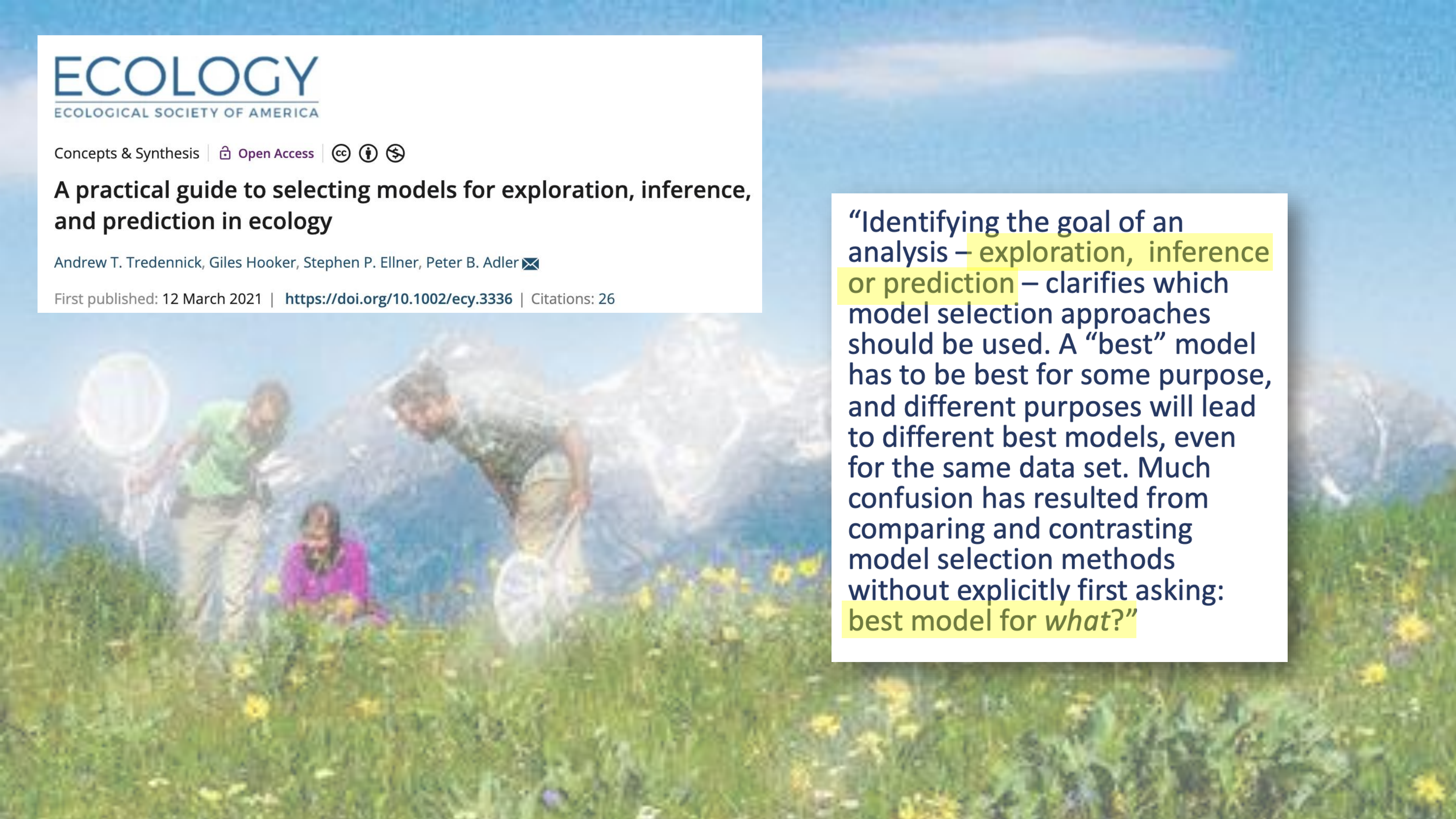
independent.

- u_j is a **latent (unobserved) random effect** for site j.
- We **do not** treat each u_j as its own free parameter (that would be fixed effects). Instead, we assume all sites are draws from a shared distribution with variance σ^2_{site}.

Given estimated variances $\hat{\sigma}^2_{\text{site}}, \hat{\sigma}^2$ and fixed effects $\hat{\beta}$, the **BLUP** (best linear unbiased predictor) of site j's effect is

$$\hat{u}_j = \underbrace{\frac{\hat{\sigma}^2_{\text{site}}}{\hat{\sigma}^2_{\text{site}} + \hat{\sigma}^2/n_j}}_{\text{shrinkage weight } w_j} \times \underbrace{(\bar{r}_j)}_{\text{site's mean residual}}, \quad \text{where} \quad \bar{r}_j = \frac{1}{n_j}\sum_i (y_{ij} - X_{ij}\hat{\beta}).$$

extension.oregonstate.edu/



ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Concepts & Synthesis | Open Access

A practical guide to selecting models for exploration, inference, and prediction in ecology

Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, Peter B. Adler

First published: 12 March 2021 | https://doi.org/10.1002/ecy.3336 | Citations: 26

"Identifying the goal of an analysis – exploration, inference or prediction – clarifies which model selection approaches should be used. A "best" model has to be best for some purpose, and different purposes will lead to different best models, even for the same data set. Much confusion has resulted from comparing and contrasting model selection methods without explicitly first asking: best model for *what*?"

Observations

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

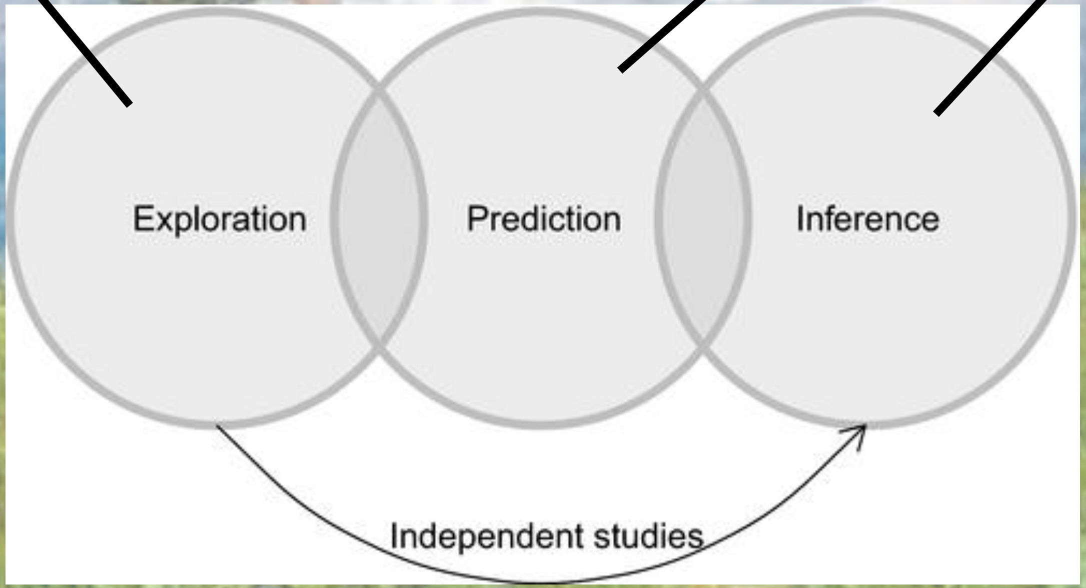
Response

Predictor variable "effects"

Prediction is about $\hat{\mathbf{y}}$

The goal of exploration is to describe patterns in the data and generate hypotheses about nature.

Inference is about $\hat{\beta}$



ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA
Concepts & Synthesis | Open Access

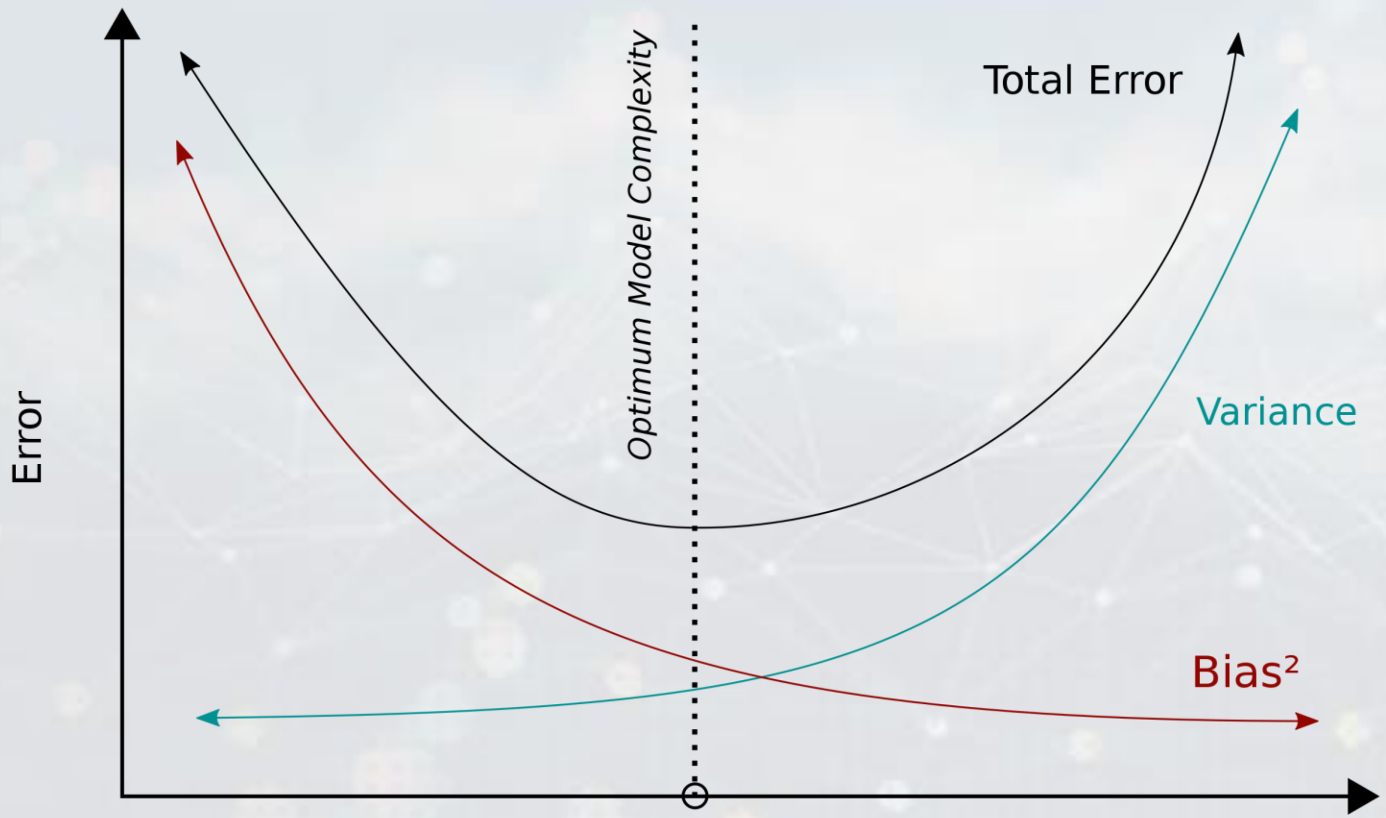
A practical guide to selecting models for exploration, inference, and prediction in ecology

Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, Peter B. Adler

First published: 12 March 2021 | https://doi.org/10.1002/ecy.3336 | Citations: 26

Bias: "accuracy" – how close the model's average prediction is to the observed values in modeled dataset.

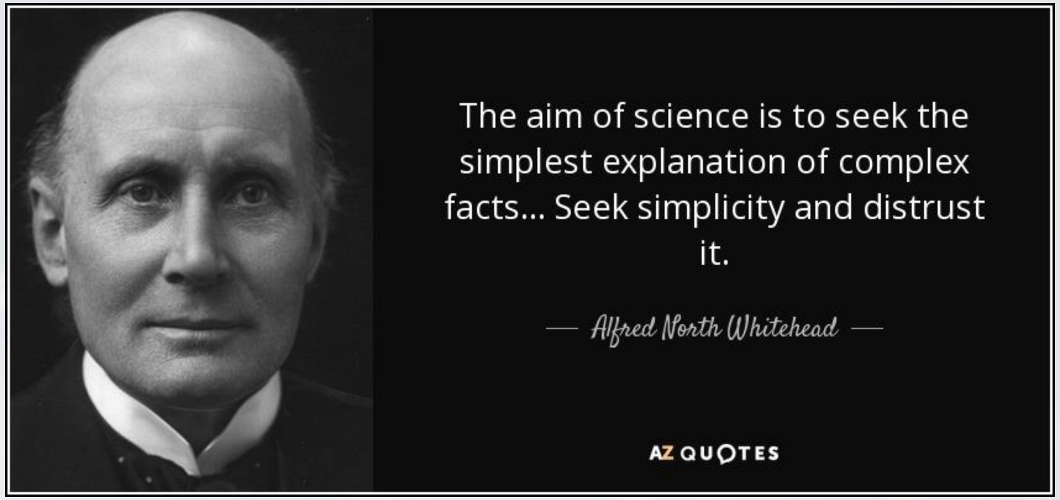
Variance: "consistency" – how much the model's predictions change for different training sets.



lm(Y ~ X)

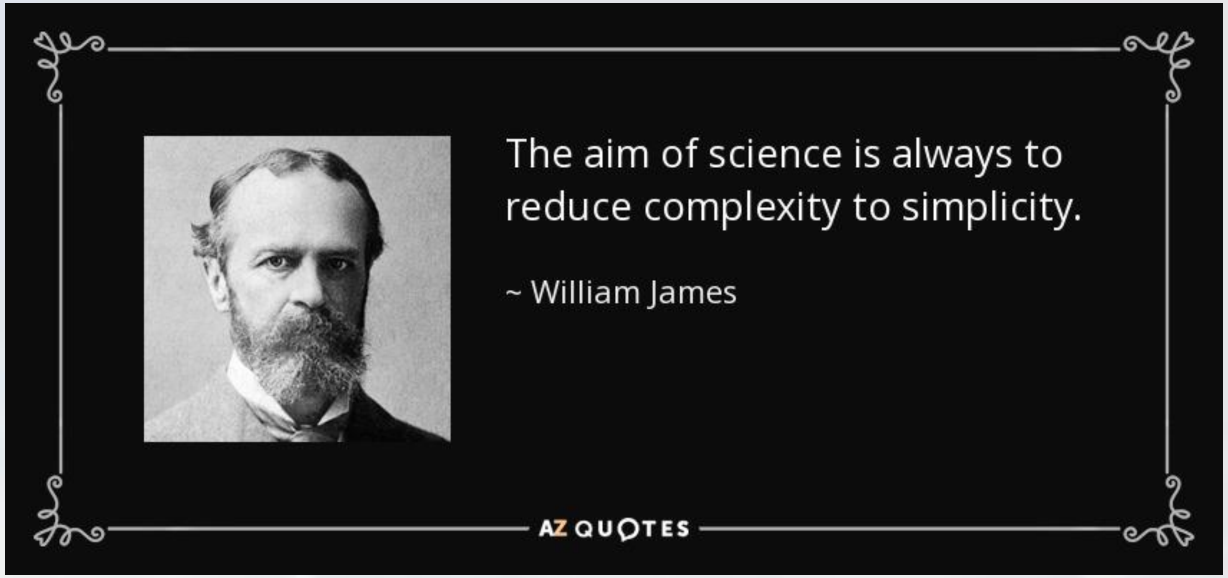
lm(Y ~ X + X2 + X3)

lm(Y ~ X+X2+X3+X*X3*X2+poly(X3, 4)....)



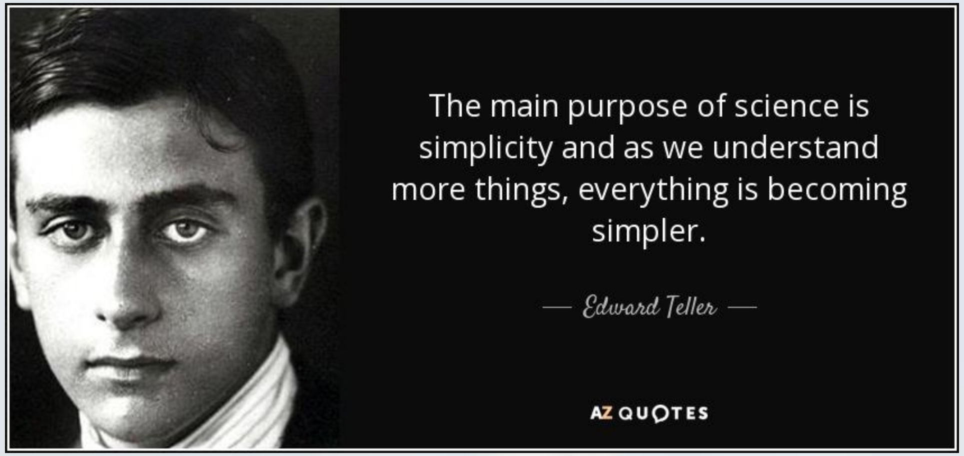
Objectives for the next 2 weeks

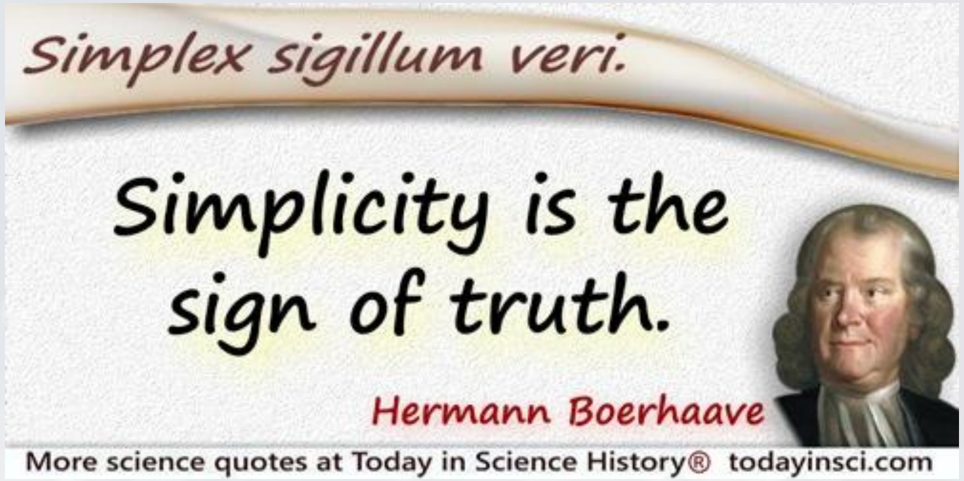
- Why we perform model selection
- Overview of some of the main methods with their pros and cons
- How to implement them in R
- Model selection for different types of analysis



Model selection is a controversial topic where there is not always an agreed upon "correct approach".

Note: when it comes to model selection, many best practices are field specific – I encourage you to consult the literature for your discipline





Goals of Model Selection

1. Aim:

- Identify a model capturing key relationships between response and explanatory variables without being overly complex.
- Often referred to as the "MAM" or **minimum adequate model**.

2. Balance Between Variance and Bias:

- Avoid models that underfit or overfit the data.
- Limit the use of interaction terms unless necessary.
- Minimize the number of parameters.

Things to Keep in Mind

1. Marginality (main effects vs interactions):

- Main effects need to be prioritized before interactions in the model formula.
- If an interaction term is found to be significant, the main effects should also be included in the model.
- When retaining an interaction term in a model, the significance of the main effects is not separately assessed.

2. Handling Missing Data:

- Be aware of instances and locations of missing data.
- Comparing two models is not advised if they rely on different data subsets.

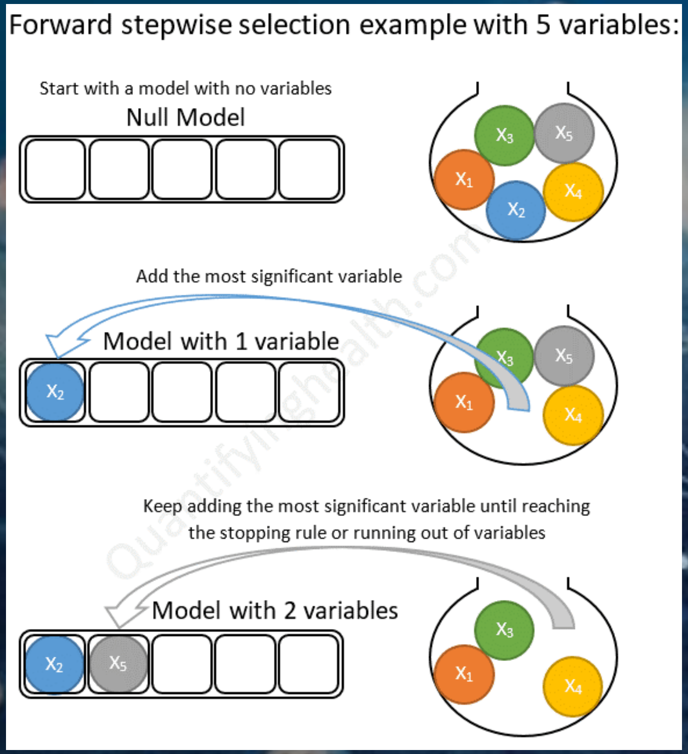
Things to Remember:

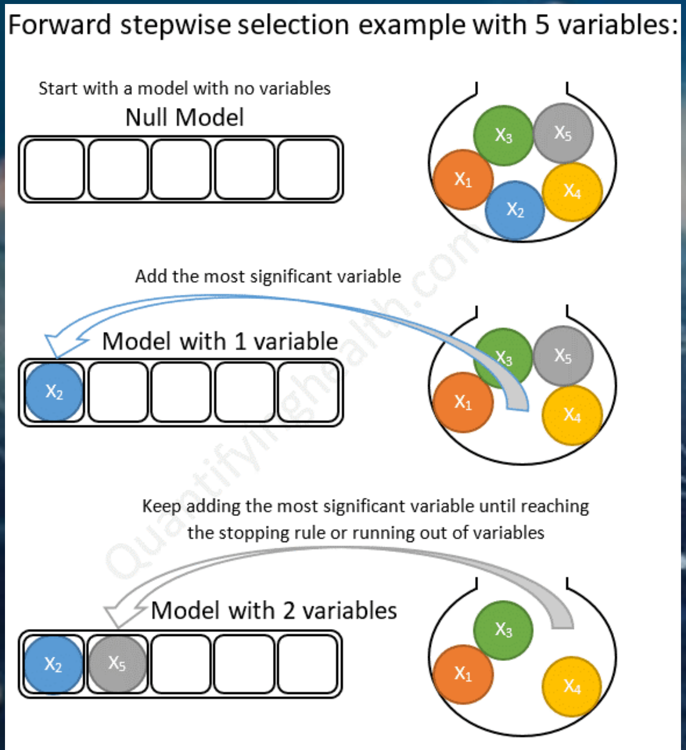
Automated Model Selection in R:

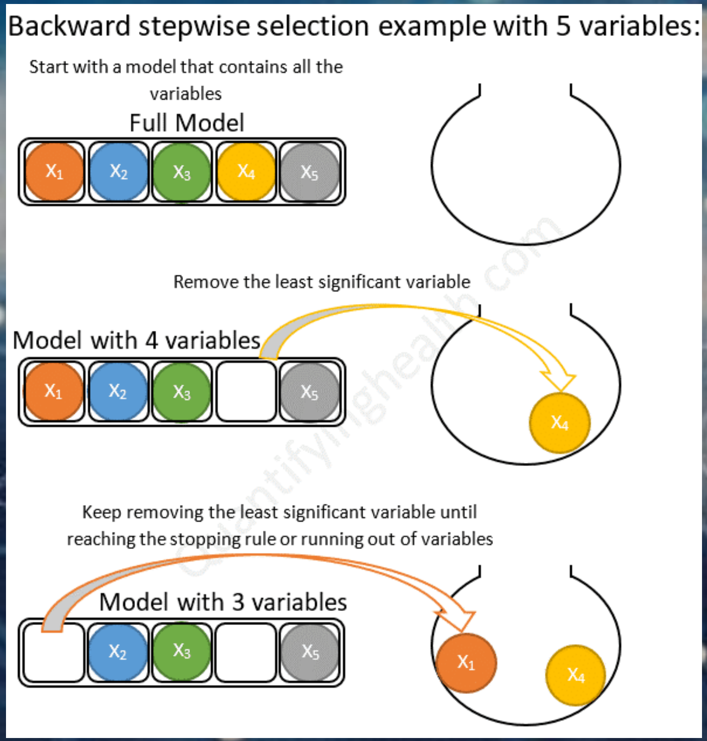
- Automated methods are widely used and simple to apply (e.g. stepAIC)
- They prioritize statistical criteria for choosing model terms.

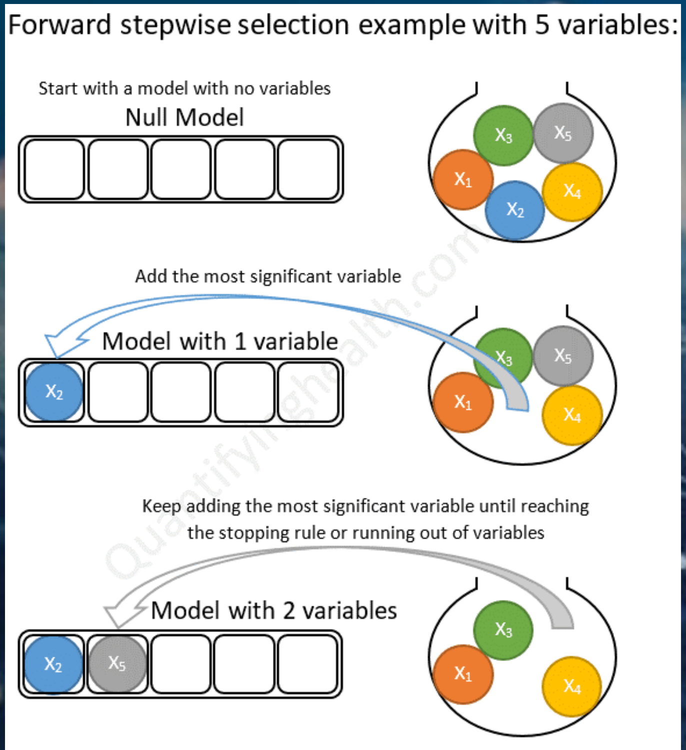
Caveats:

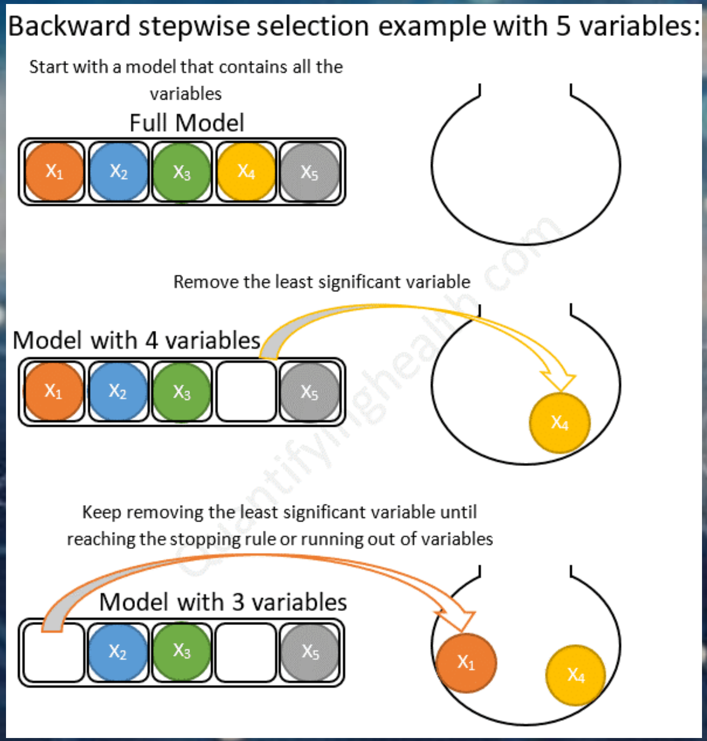
- Sometimes essential design components ("block") are vital to keep.
- Approach automated methods with caution.
- Manual methods can be useful in specific scenarios.
- Some automated tools allow for customization to make smarter choices.

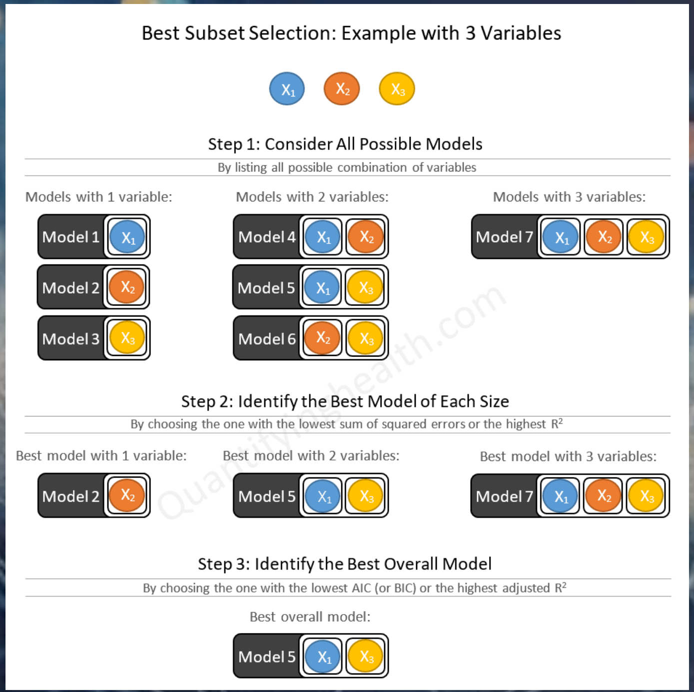

Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
Add the most significant variable
Model with 1 variable
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X_1
X_2
X_3
X_4
X_5


Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
Add the most significant variable
Model with 1 variable
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X1
X2
X3
X4
X5


Backward stepwise selection example with 5 variables:
Start with a model that contains all the variables
Full Model
Remove the least significant variable
Model with 4 variables
Keep removing the least significant variable until reaching the stopping rule or running out of variables
Model with 3 variables
X1
X2
X3
X4
X5


Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
Add the most significant variable
Model with 1 variable
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X1 X2 X3 X4 X5


Backward stepwise selection example with 5 variables:
Start with a model that contains all the variables
Full Model
Remove the least significant variable
Model with 4 variables
Keep removing the least significant variable until reaching the stopping rule or running out of variables
Model with 3 variables
X1 X2 X3 X4 X5

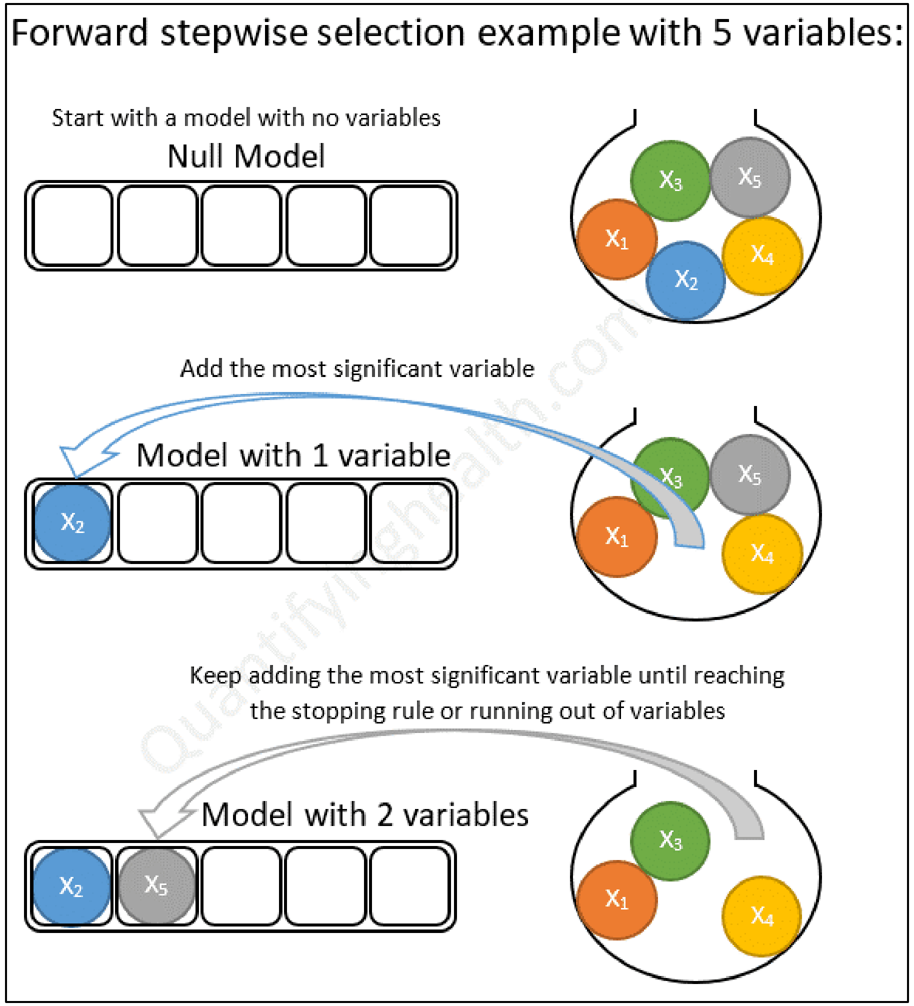

Best Subset Selection: Example with 3 Variables
Step 1: Consider All Possible Models
By listing all possible combination of variables
Models with 1 variable:
Models with 2 variables:
Models with 3 variables:
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Step 2: Identify the Best Model of Each Size
By choosing the one with the lowest sum of squared errors or the highest R2
Best model with 1 variable:
Best model with 2 variables:
Best model with 3 variables:
Step 3: Identify the Best Overall Model
By choosing the one with the lowest AIC (or BIC) or the highest adjusted R2
Best overall model:
Model 5

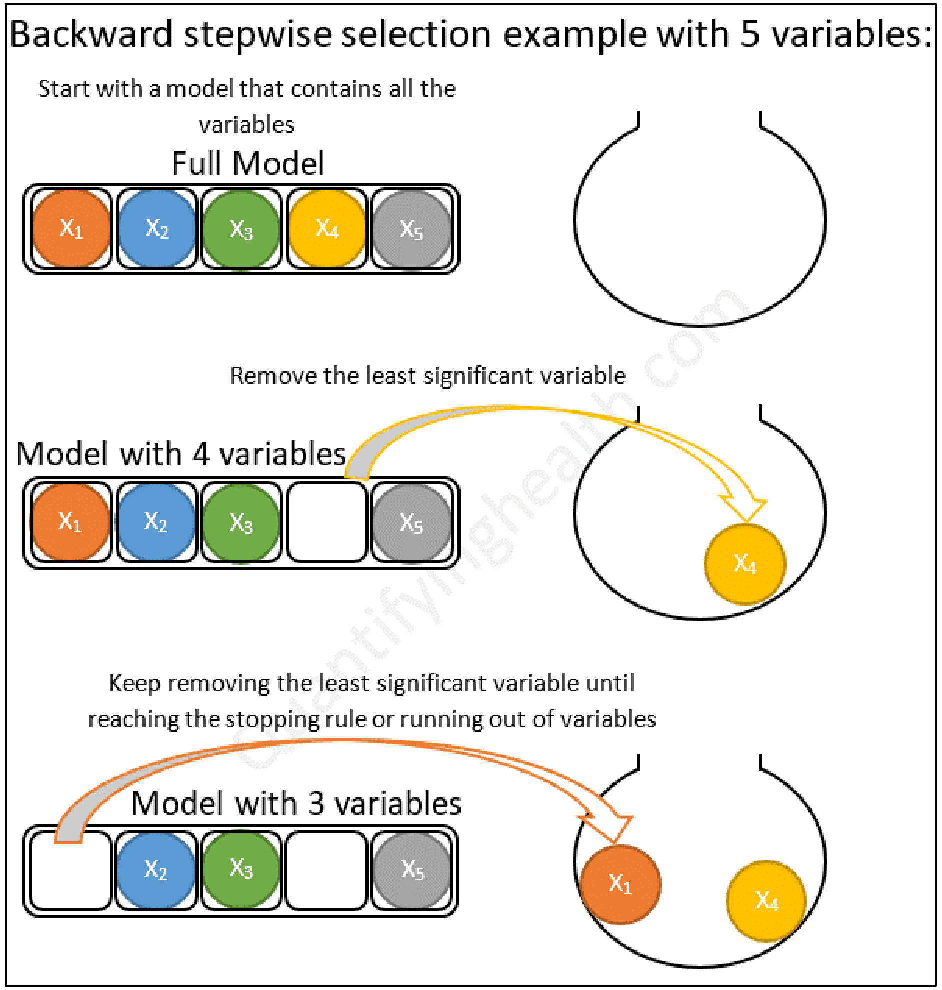
Stepwise regression methods

lm(y~1, data)

lm(y~x2, data)

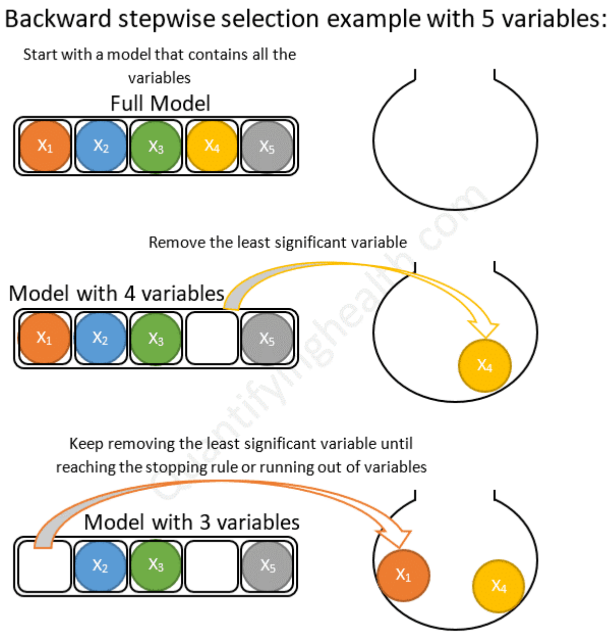
lm(y~x2+x5, data)

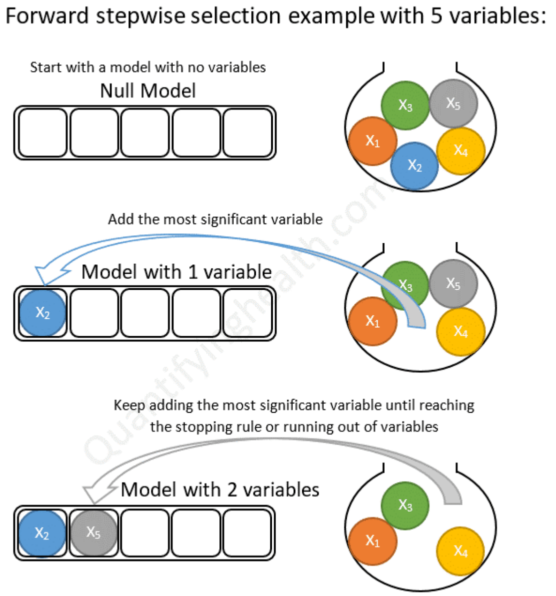




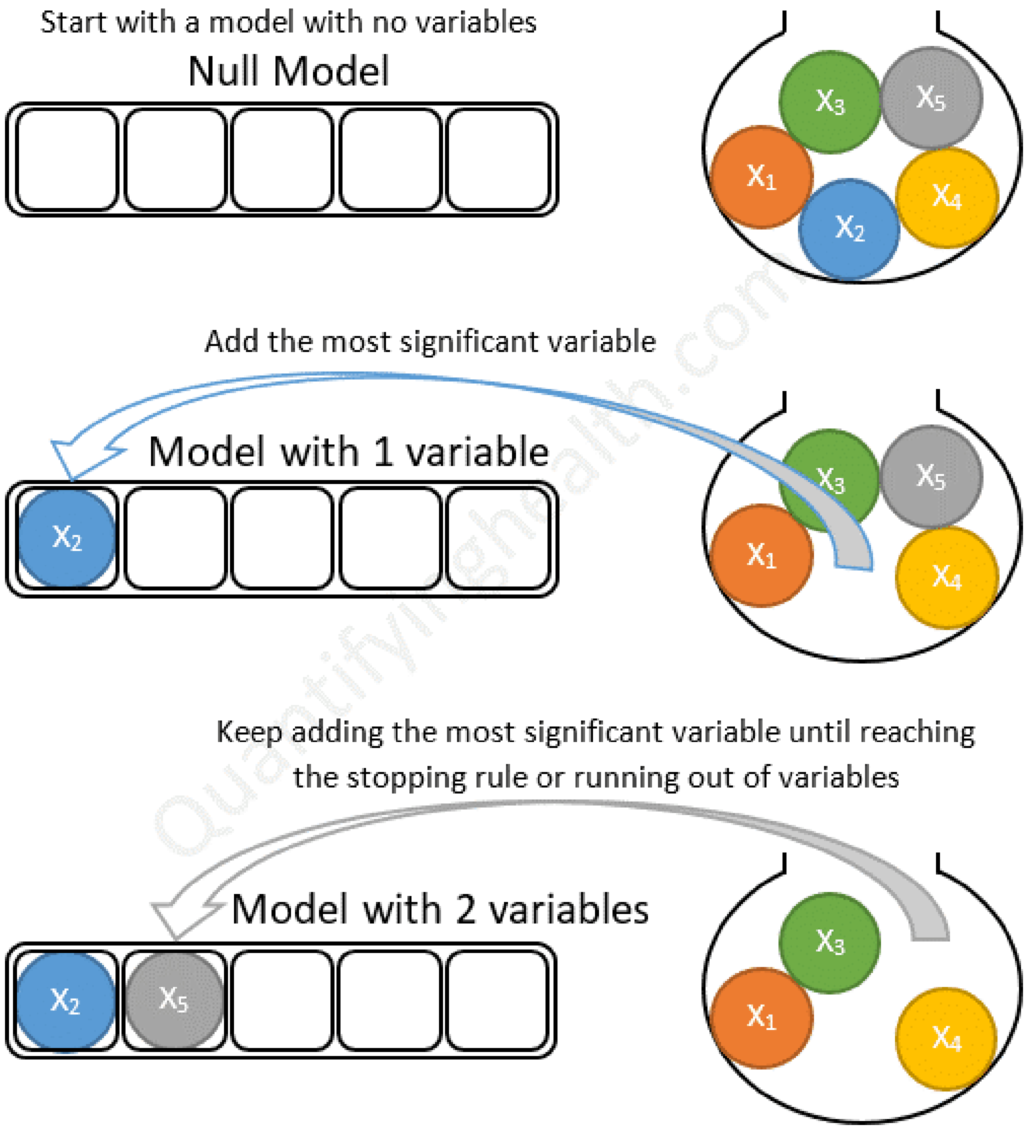
Stepwise regression methods

- Begin with either a full model or a null model.
- Adjust by removing or adding variables.
  - Options include backward or forward selection.
  - A combination of both forward and backward methods is available.
- Implement a stopping criteria to ascertain when the optimal model is identified.





Forward stepwise selection example with 5 variables:



Forward Selection

•**Start**: Initiate with a model without predictors (only the intercept).

•**Process**:

- On every iteration, include the predictor that most enhances the model.
- Each step integrates an additional predictor.
- A predictor, once added, remains in the model.

•**End**: Cease when adding more predictors no longer boosts the model.

•**Condition**: This method is applicable when the number of observations (n) is less than the number of predictors (p).

Example:

Step 1: null model

```
Null.Model <- lm(y ~ 1)  # Model with no predictors, only intercept
```

Step 2: (Introduce the best predictor, say x4)

```
Model.S1 <- lm(y ~ x4)
```

Add **x4**

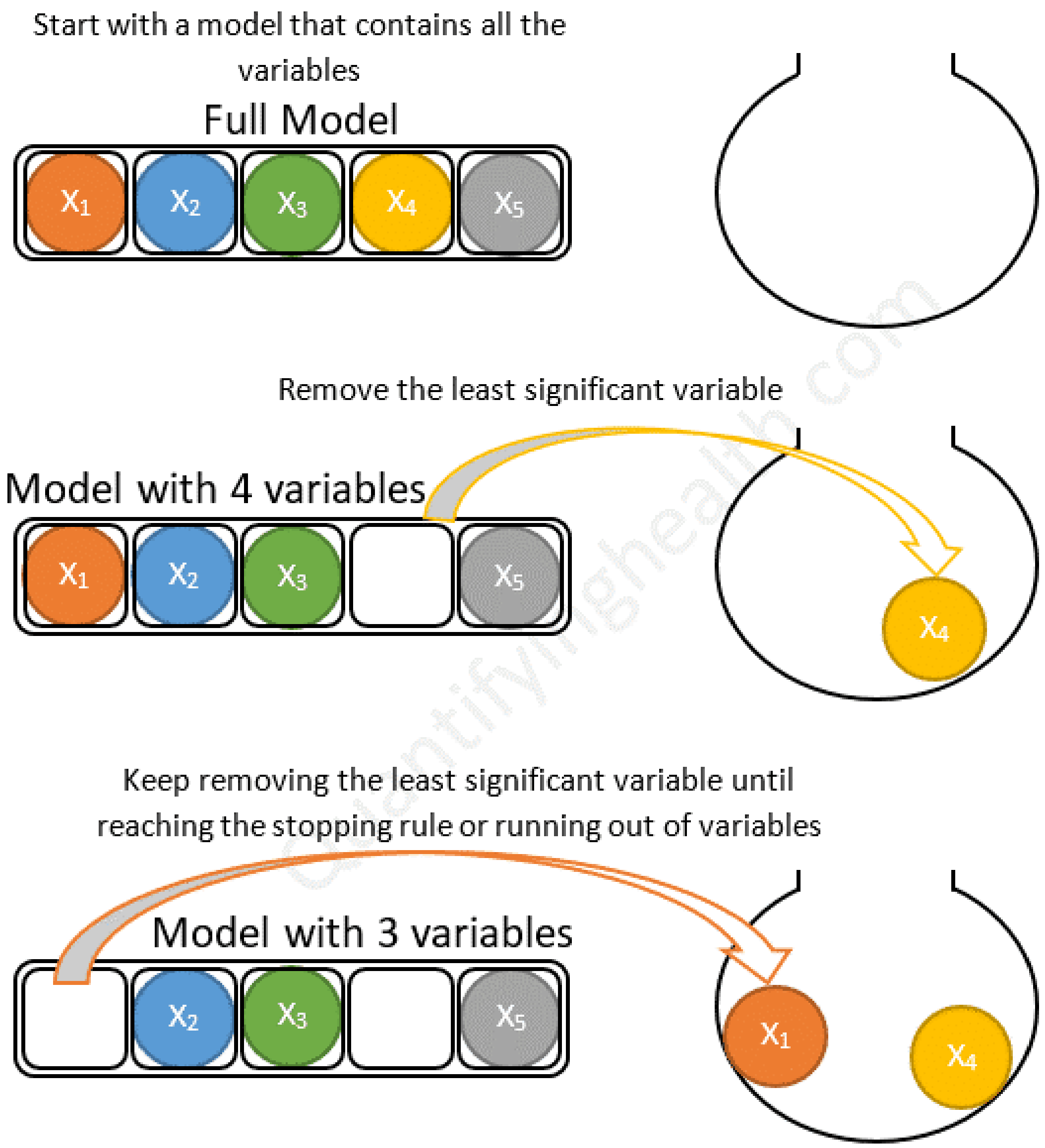
Step 3: (Add the next best predictor, say x1, along with x4)

```
Model.S2 <- lm(y ~ x4 + x1)
```

Add **x1**

… continue adding one predictor at a time, until the model no longer benefits from the inclusion of more predictors.

Backward stepwise selection example with 5 variables:



Backward Selection

- **Start**: Use a full model with all predictors.
- **Process**:
  - At every stage, eliminate the predictor that contributes the least to the model.
  - Each step reduces the predictors by one.
  - A predictor, once removed, doesn't re-enter the process.
- **End**: Stop when removing predictors no longer enhances the model.
- **Condition**: Applicable only when the number of observations (n) exceeds the number of predictors (p).

Example:

Step 1: full model

```
Full.Model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)
```

Step 2: remove least significant predictor

```
Model.S1 <- lm(y ~ x1 + x2 + x4 + x5 + x6)
```

Removes **x3**

Step 3: remove least significant predictor

```
Model.S2 <- lm(y ~ x2 + x4 + x5 + x6)
```

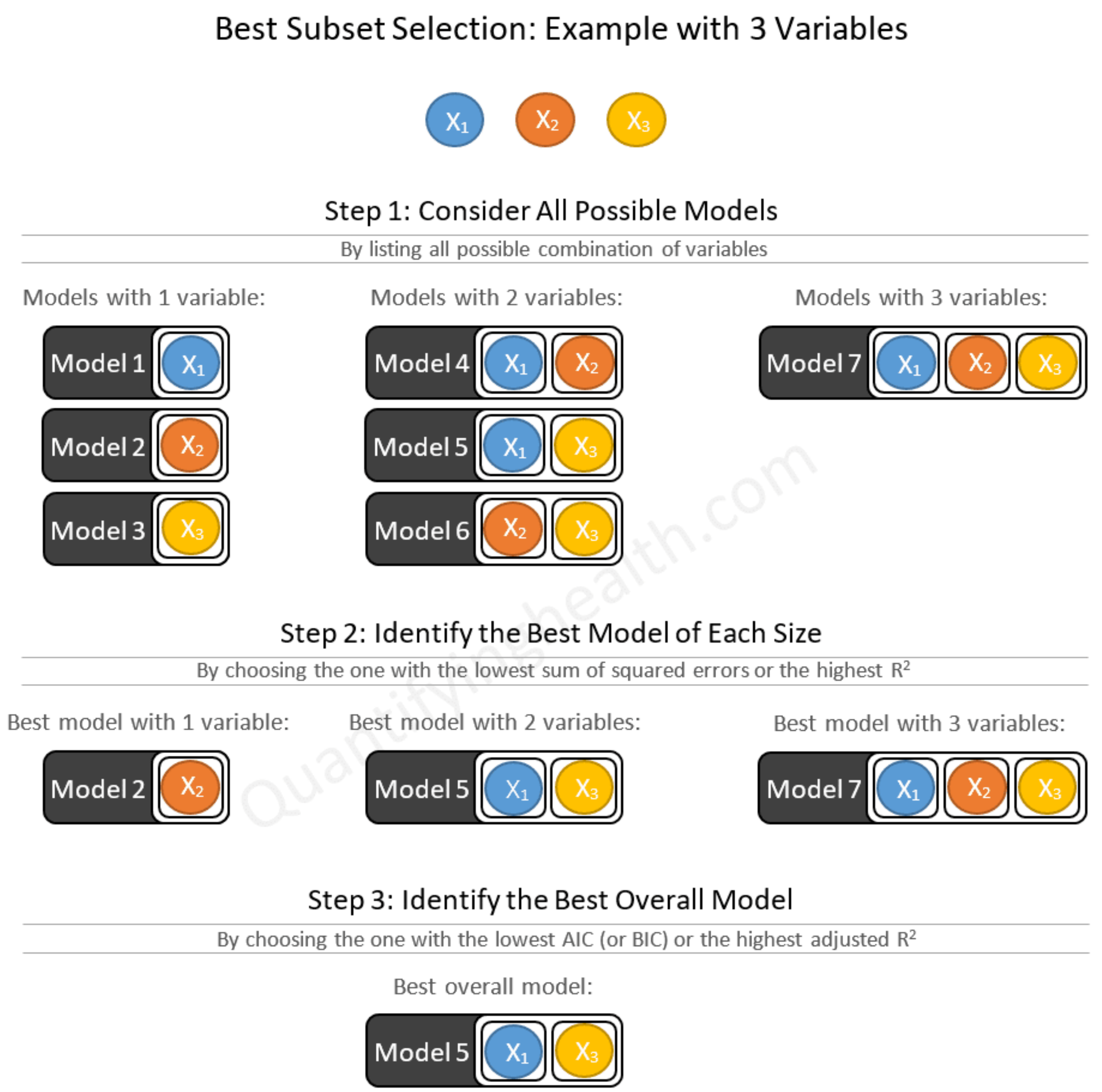
Removes **x1**

...continue until all predictors are significant

Issues with Stepwise Methods

- **Popularity**: Stepwise methods aren't as favored today; specifics vary by field.
- **Goal Issues**: The methods aim to pinpoint the "best" model, but data often doesn't support such a confident choice.
- **Algorithmic Concerns**:
    - Approach (be it forward or backward).
    - Order in which parameters are added or removed.
    - The count of potential parameters. All these can influence the final model choice.
- **Hypothesis Testing**: One stepwise regression can lead to numerous hypothesis tests.
- **Model Comparison Limitations**:
    - For nested models, e.g., y ~ x1 + x2 + x3 vs. y ~ x1 + x3.
    - Not for non-nested ones, like y ~ x1 + x4 vs. y ~ x5 + x8.
- **Literature Reference**: Whittingham et al., 2006 discussed the use of stepwise modeling in certain fields.

All subsets search



All subsets search

- Perform an exhaustive search of all possible combinations of variables
- Use a criteria to rank and compare models

How many models is this?

Ignoring interactions and nonlinearities (polynomials):

If we have p variables, there are 2^p models

- with 10 variables ~ 1000 models
- with 20 variables ~ 1 million models
- with 50 variables ~ 1 x 10^15 models

Most approaches in R implemented on a personal computer will fail ~ 30 predictors

All subsets search

- Perform an exhaustive search of all possible combinations of variables
- Use a criteria to rank models

This approach has been criticized as “data-dredging” or “fishing” and is only recommend for exploratory analyses

““Let the computer find out” is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting (Burnham and Anderson, 2002)."

Construct a set of Candidate Models (Hypothesis testin

```
> anova(fit.1,fit.2,fit.3,fit.4,fit.5)
Analysis of Variance Table

Model 1: wage ~ age
Model 2: wage ~ poly(age, 2)
Model 3: wage ~ poly(age, 3)
Model 4: wage ~ poly(age, 4)
Model 5: wage ~ poly(age, 5)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   2998 5022216
2   2997 4793430  1    228786 143.59 <2e-16 ***
3   2996 4777674  1     15756   9.89 0.0017 **
4   2995 4771604  1      6070   3.81 0.0510 .
5   2994 4770322  1      1283   0.80 0.3697
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Construct a set of Candidate Models (Hypothesis testing)

- Develop a series of hypothesis-driven models that capture the processes and patterns you are interested in testing
- These hypotheses and the models related to them should be based on detailed knowledge of your system that came from prior research, literature reviews, and conversations with collaborators
- Compare the models in this limited set to find out which one/ones have the most support

How to evaluate models?

- Null hypothesis statistical testing
- Information theoretic approaches
- Prediction error coupled with cross validation

ANOVA vs. Information Criteria vs. Cross Validation in Model Selection

- **ANOVA (Analysis of Variance., i.e. hypothesis testing)**
  - Focus: Tests the overall fit of a statistical model.
  - Application: Compares nested models, i.e., models that differ by one or more predictor variables.
  - Criteria: Relies on significance testing (p-values).
  - Limitations: Only suitable for comparing hierarchically nested models.

ANOVA vs. Information Criteria vs. Cross Validation in Model Selection

- **ANOVA (Analysis of Variance., i.e. hypothesis testing)**
  - Focus: Tests the overall fit of a statistical model.
  - Application: Compares nested models, i.e., models that differ by one or more predictor variables.
  - Criteria: Relies on significance testing (p-values).
  - Limitations: Only suitable for comparing hierarchically nested models.
- **Information Criteria: AIC (Akaike Information Criterion) and related (AICc and BIC)**
  - Focus: Measures the goodness of fit and complexity of a model simultaneously.
  - Application: Can compare non-nested models.
  - Criteria: Smaller AIC values indicate better-fitting models.
  - Strength: Balances model fit and model complexity.
  - Limitations: Provides a relative measure, i.e., only useful when comparing multiple models.

ANOVA vs. Information Criteria vs. Cross Validation in Model Selection

- **ANOVA (Analysis of Variance., i.e. hypothesis testing)**
  - Focus: Tests the overall fit of a statistical model.
  - Application: Compares nested models, i.e., models that differ by one or more predictor variables.
  - Criteria: Relies on significance testing (p-values).
  - Limitations: Only suitable for comparing hierarchically nested models.
- **Information Criteria: AIC (Akaike Information Criterion) and related (AICc and BIC)**
  - Focus: Measures the goodness of fit and complexity of a model simultaneously.
  - Application: Can compare non-nested models.
  - Criteria: Smaller AIC values indicate better-fitting models.
  - Strength: Balances model fit and model complexity.
  - Limitations: Provides a relative measure, i.e., only useful when comparing multiple models.
- **Cross Validation**
  - Focus: Assesses a model's predictive performance on unseen data.
  - Application: Splits the dataset into training and testing subsets multiple times.
  - Criteria: Uses metrics like RMSE (for regression) or accuracy (for classification) to evaluate performance.
  - Strength: Helps prevent overfitting by validating the model on different data subsets.
  - Limitations: Computationally expensive, especially with large datasets or complex models.

Null Hypothesis Testing for Model Comparison

Use F-tests or log likelihood ratio tests to compare nested models

Implemented in R using anova() function

Reduced model: m1 <- lm(plant height ~ shading + moisture + nitrogen)

Full model: m2 <- lm(plant height ~ shading + moisture + nitrogen + shading:moisture)

anova(m1, m2) # compare the reduced and full model

	Df	RSS	Df	Sum of Sq	F	Pr(>F)
m1 (reduced)	243	1083				
m2 (full)	244	1164	1	19.51	4.46	0.0215**

H0 = the two models are equally useful for predicting the outcome

Ha = the larger model is significantly better than the smaller model

If $p < 0.05$, accept Ha. The full model is better than the reduced model. It has more explanatory power

If $p > 0.05$ (or the chosen threshold for significance), use the simplified model

Information Theoretic (IT) Approaches

As scientists, we often care about:

“what is the weight of evidence for a number of competing hypotheses?”

These models aren’t necessarily nested, so we need methods to compare non-nested models.

Example of non-nested study: which factors shape the diversity of the gut microbiome of lemurs?

Hyp 1: social factors

Hyp 2: environmental/abiotic factors

Overview of Information Theoretic Approaches for Model Selection

Common IT Criteria for Model Selection

- **AIC (Akaike Information Criterion)**
  - Balances model fit and model complexity
  - Lower AIC indicates a better model
- **BIC (Bayesian Information Criterion)**
  - Similar to AIC but penalizes complexity more heavily
  - Favors simpler models, especially with larger datasets
- **AICc (Corrected Akaike Information Criterion)**
  - Modification of AIC for small sample sizes
  - Includes a correction factor to account for bias
  - Tends to select more parsimonious models than AIC with small data

$$\boxed{AIC = -2 \log(model|data) + 2K}$$

$$\boxed{BIC = -2 \log(model|data) + \log(n)*K}$$

$$\boxed{AICc = AIC + (2K(K+1))/(n-K-1)}$$

Key Principles

- **Parsimony**: Preference for simpler models to prevent overfitting
- **Trade-off**: Balance between model complexity and goodness of fit

Applications and Examples

- Widely used in statistics, machine learning, and various scientific fields
- Example: Model selection in linear regression, time series analysis, etc.

Advantages of IT Approaches

- Objective criteria for model comparison
- Encourages simplicity, aiding interpretability and generalization

•**Limitations and Considerations**

- May not always align with domain-specific goals or constraints
- Performance can depend on sample size and data characteristics
- AICc is particularly recommended when sample size is small relative to the number of parameters

AIC (Akaike Information Criterion)

$$AIC = -2 \log (model|data) + 2K$$

K = total number of parameters in the model

Log-likelihood of the model given the data. Measures the model's fit (or lack of fit) to the observed data

Penalty that adjusts for the number of variables in the model to discourage overfitting

- AIC is an estimate of the information lost between the fitted model and the true model. It balances information loss due to both bias and variance.
- The value of AIC tells us nothing about the quality of the model. It only reflects the quality of a model relative to the other models. It is a comparative tool.
- Models with lower AIC are preferred

Delta Δ AIC

Differences in AIC or delta AIC are pivotal for ranking models

We measure the difference in AIC between the top model (with the lowest AIC) and all other models in the set

We often retain a set of top models that should be considered, rather than focusing on a single model

Δ AIC from the top model

0-2 substantial support

3-7 modest support

8-11 relatively little support

> 11 essentially no support

AICc: Modification for Small Sample Size

$$AICc = AIC + (2K(K+1))/(n-K-1)$$

- When sample size is small, AIC tends to select models with too many parameters -> overfitting
- AICc includes an additional bias correction term
- Should be applied when sample sizes are small
  - when $n/k < 40$ (n = number of samples, k = number of parameters)
- As sample size increases, AICc converges to AIC -> many people recommend always using AICc

BIC (Bayesian Information Criterion)

$$BIC = -2 \log(\text{model}|\text{data}) + \log(n) * K$$

Equation is very similar to AIC

- swapped out 2*K for log (n)*K
- n = number of samples/observations

Model with the lowest BIC is considered "best"

BIC will generally select for smaller models than AIC because it places a higher penalty on additional variables

Things to Keep in Mind about IT Approaches

- IT approaches allow models to be ordered from 'best" to "worse"
- They are only valid relative to the set of considered models
- They don't tell you how good your best model is at explaining the patterns in the data
  - Important to compare models that are supported by biological knowledge
  - Include a null model with only the intercept

> Users should keep in mind the hazards that a "thoughtless approach" of evaluating all possible models poses. Although this procedure is in certain cases useful and justified, it may result in selecting a spurious "best" model, due to the model selection bias. *"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting* (Burnham and Anderson, 2002).

Review | TRENDS in Ecology and Evolution Vol.19 No.2 February 2004

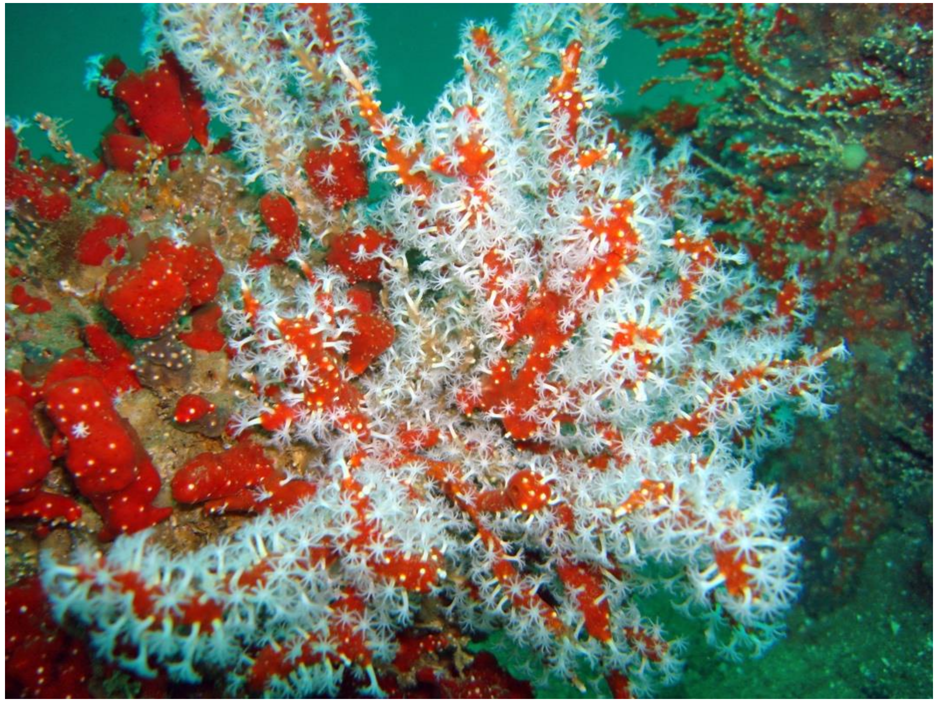
ELSEVIER

Model selection in ecology and evolution

Jerald B. Johnson[1] and Kristian S. Omland[2]

Table I. Commonly used model selection methods

Model selection method	Calculation[a]	Elements	Refs
Adjusted R^2	$R^2_{adj} = 1 - \dfrac{RSS/n - p - 1}{\sum(y_i - \bar{y})^2/n - 1}$	Fit	[7]
Likelihood ratio test	$\mathrm{LRT} = -2\{\ln[L(\hat{\theta}_p \mid y)] - \ln[L(\hat{\theta}_{p+q} \mid y)]\} \sim \chi^2_q$	Fit and complexity	[7]
Akaike information criterion (AIC)	$\mathrm{AIC} = -2\ln[L(\hat{\theta}_p \mid y)] + 2p$	Fit and complexity	[3]
Small sample unbiased AIC (AIC_c)	$\mathrm{AIC}_c = -2\ln[L(\hat{\theta}_p \mid y)] + 2p\left(\dfrac{n}{n-p-1}\right)$	Fit and complexity (with bias correction term for small sample size)	[3]
Schwarz criterion (BIC)	$\mathrm{SC} = -2\ln[L(\hat{\theta}_p \mid y)] + p \cdot \ln(n)$	Fit, complexity, and sample size	[10]



Dataset:

	SA	depth	temp	salinity
1	25.30443	4.901916	7.647184	3.559808
2	21.86207	1.941074	9.691845	4.535293
3	22.67535	5.018254	9.213907	5.385524
4	24.14200	2.645673	8.489021	3.249306
5	26.12916	4.130543	8.484586	2.809092
6	22.51604	2.587241	8.248697	5.260266

Backward (AIC):

```
Call:
lm(formula = SA ~ depth + temp + salinity + depth:salinity +
    temp:salinity, data = coral)
```

Forward (AIC):

```
Call:
lm(formula = SA ~ temp + depth, data = coral)
```

All subsets (AICc):

```
Call:
lm(formula = SA ~ depth + salinity + temp + depth:salinity +
    salinity:temp + 1, data = coral)
```

All subsets (BIC):

```
Call:
lm(formula = SA ~ depth + temp + 1, data = coral)
```

All models result:

Coefficients

Change in AICc between models

”Probability this is best model”

```
Global model call: lm(formula = SA ~ (depth + temp + salinity)^2, data = coral)
---
Model selection table
```

	(Int)	dpt	sln	tmp	dpt:sln	dpt:tmp	sln:tmp	df	logLik	AICc	delta	weight
48	-4.60600	1.5090	3.08100	2.604	-0.1975		-0.2506	7	-535.926	1086.3	0.00	0.221
6	8.01000	0.6898		1.570				4	-539.237	1086.6	0.32	0.188
40	-1.00600	0.7109	2.21500	2.562			-0.2438	6	-537.231	1086.8	0.49	0.173
64	-2.62500	0.9071	3.25200	2.356	-0.2136	0.074770	-0.2630	8	-535.732	1088.1	1.75	0.092
16	4.78900	1.4490	0.79810	1.574	-0.1888			6	-538.042	1088.4	2.11	0.077
8	7.99000	0.6869	0.02850	1.561				5	-539.216	1088.7	2.36	0.068
22	7.92400	0.7110		1.580		-0.002372		5	-539.237	1088.7	2.40	0.066
56	-0.02886	0.4244	2.25800	2.454		0.032130	-0.2489	7	-537.193	1088.8	2.53	0.062
32	5.86700	1.1830	0.82320	1.442	-0.1957	0.032910		7	-538.004	1090.5	4.16	0.028
24	7.83600	0.7248	0.02889	1.578		-0.004249		6	-539.215	1090.8	4.46	0.024
5	10.77000			1.564				3	-555.434	1117.0	30.65	0.000
7	10.63000		0.12330	1.523				4	-555.079	1118.3	32.01	0.000
39	4.46800		1.63900	2.216			-0.1687	5	-554.237	1118.7	32.40	0.000
4	20.09000	0.6244	0.55080					4	-591.335	1190.8	104.52	0.000
12	18.45000	1.0280	0.96110		-0.1001			5	-591.118	1192.5	106.17	0.000
2	22.11000	0.6785						3	-597.076	1200.3	113.94	0.000
3	22.24000		0.62590					3	-600.240	1206.6	120.26	0.000
1	24.77000							2	-607.192	1218.4	132.12	0.000

Models ranked by AICc(x)

Keeping multiple models

Coefficients

Change in AICc between models

Global model call: lm(formula = SA ~ (depth + temp + salinity)^2, data = coral)

Model selection table

	(Int)	dpt	sln	tmp	dpt:sln	dpt:tmp	sln:tmp	df	logLik	AICc	delta	weight
48	-4.60600	1.5090	3.08100	2.604	-0.1975		-0.2506	7	-535.926	1086.3	0.00	0.221
6	8.01000	0.6898		1.570				4	-539.237	1086.6	0.32	0.188
40	-1.00600	0.7109	2.21500	2.562			-0.2438	6	-537.231	1086.8	0.49	0.173
64	-2.62500	0.9071	3.25200	2.356	-0.2136	0.074770	-0.2630	8	-535.732	1088.1	1.75	0.092
16	4.78900	1.4490	0.79810	1.574	-0.1888			6	-538.042	1088.4	2.11	0.077
8	7.99000	0.6869	0.02850	1.561				5	-539.216	1088.7	2.36	0.068
22	7.92400	0.7110		1.580		-0.002372		5	-539.237	1088.7	2.40	0.066
56	-0.02886	0.4244	2.25800	2.454		0.032130	-0.2489	7	-537.193	1088.8	2.53	0.062
32	5.86700	1.1830	0.82320	1.442	-0.1957	0.032910		7	-538.004	1090.5	4.16	0.028
24	7.83600	0.7248	0.02889	1.578		-0.004249		6	-539.215	1090.8	4.46	0.024
5	10.77000			1.564				3	-555.434	1117.0	30.65	0.000
7	10.63000		0.12330	1.523				4	-555.079	1118.3	32.01	0.000
39	4.46800		[illegible]	[illegible]			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	20.09000	0.6244	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	18.45000	1.0280	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	22.11000	0.6785						3	-597.076	1200.3	113.94	0.000
3	22.24000		0.62590					3	-600.240	1206.6	120.26	0.000
1	24.77000							2	-607.192	1218.4	132.12	0.000

Models ranked by AICc(x)

```
delta6.exp<-subset(dredge.aicc, delta<=6,recalc.weights=FALSE)
```

Rule:
ΔAICc < 6

Keeping multiple models

Coefficients

Change in AICc between models

```
Global model call: lm(formula = SA ~ (depth + temp + salinity)^2, data = coral)
---
Model selection table
```

	(Int)	dpt	sln	tmp	dpt:sln	dpt:tmp	sln:tmp	df	logLik	AICc	delta	weight
48	-4.60600	1.5090	3.08100	2.604	-0.1975		-0.2506	7	-535.926	1086.3	0.00	0.221
6	8.01000	0.6898		1.570				4	-539.237	1086.6	0.32	0.188
40	-1.00600	0.7109	2.21500	2.562			-0.2438	6	-537.231	1086.8	0.49	0.173
64	-2.62500	0.9071	3.25200	2.356	-0.2136	0.074770	-0.2630	8	-535.732	1088.1	1.75	0.092
16	4.78900	1.4490	0.79810	1.574	-0.1888			6	-538.042	1088.4	2.11	0.077
8	7.99000	0.6869	0.02850	1.561				5	-539.216	1088.7	2.36	0.068
22	7.92400	0.7110		1.580		-0.002372		5	-539.237	1088.7	2.40	0.066
56	-0.02886	0.4244	2.25800	2.454		0.032130	-0.2489	7	-537.193	1088.8	2.53	0.062
32	5.86700	1.1830	0.82320	1.442	-0.1957	0.032910		7	-538.004	1090.5	4.16	0.028
24	7.83600	0.7248	0.02889	1.578		-0.004249		6	-539.215	1090.8	4.46	0.024
5	10.77000							[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	10.63000							[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
39	4.46800		1.63900	2.216			-0.1687	5	-554.237	1118.7	32.40	0.000
4	20.09000	0.6244	0.55080					4	-591.335	1190.8	104.52	0.000
12	18.45000	1.0280	0.96110		-0.1001			5	-591.118	1192.5	106.17	0.000
2	22.11000	0.6785						3	-597.076	1200.3	113.94	0.000
3	22.24000		0.62590					3	-600.240	1206.6	120.26	0.000
1	24.77000							2	-607.192	1218.4	132.12	0.000

Models ranked by AICc(x)

```
delta2.exp<-subset(dredge.aicc, delta<=2, recalc.weights=FALSE)
```

Rule:
ΔAICc < 2

Keeping multiple models

Coefficients

"Probability this is best model"

```
Global model call: lm(formula = SA ~ (depth + temp + salinity)^2, data = coral)
---
Model selection table
```

	(Int)	dpt	sln	tmp	dpt:sln	dpt:tmp	sln:tmp	df	logLik	AICc	delta	weight
48	-4.60600	1.5090	3.08100	2.604	-0.1975		-0.2506	7	-535.926	1086.3	0.00	0.221
6	8.01000	0.6898		1.570				4	-539.237	1086.6	0.32	0.188
40	-1.00600	0.7109	2.21500	2.562			-0.2438	6	-537.231	1086.8	0.49	0.173
64	-2.62500	0.9071	3.25200	2.356	-0.2136	0.074770	-0.2630	8	-535.732	1088.1	1.75	0.092
16	4.78900	1.4490	0.79810	1.574	-0.1888			6	-538.042	1088.4	2.11	0.077
8	7.99000	0.6869	0.02850	1.561				5	-539.216	1088.7	2.36	0.068
22	7.92400	0.7110		1.580		-0.002372		5	-539.237	1088.7	2.40	0.066
56	-0.02886	0.4244	2.25800	2.454		0.032130	-0.2489	7	-537.193	1088.8	2.53	0.062
32	5.86700	1.1830	0.82320	1.442	-0.1957	0.032910		7	-538.004	1090.5	4.16	0.028
24	7.83600	0.7248	0.02889	1.578		-0.004249		6	-539.215	1090.8	4.46	0.024
5	10.77000			1.564				3	-555.434	1117.0	30.65	0.000
7	10.63000		0.12330	1.523				4	-555.079	1118.3	32.01	0.000
39	4.46800	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	20.09000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	18.45000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	22.11000	0.6785						3	-597.076	1200.3	113.94	0.000
3	22.24000		0.62590					3	-600.240	1206.6	120.26	0.000
1	24.77000							2	-607.192	1218.4	132.12	0.000

```
Models ranked by AICc(x)
```

```
weight95.exp<-subset(dredge.aicc, cumsum(weight) <= .95)
```

Rule:
Cumulative sum of weight > 0.95

Averaging multiple models

```
Global model call: lm(formula = SA ~ (depth + temp + salinity)^2, data = coral)
---
Model selection table
```

	(Int)	dpt	sln	tmp	dpt:sln	dpt:tmp	sln:tmp	df	logLik	AICc	delta	weight
48	-4.60600	1.5090	3.08100	2.604	-0.1975		-0.2506	7	-535.926	1086.3	0.00	0.221
6	8.01000	0.6898		1.570				4	-539.237	1086.6	0.32	0.188
40	-1.00600	0.7109	2.21500	2.562			-0.2438	6	-537.231	1086.8	0.49	0.173
64	-2.62500	0.9071	3.25200	2.356	-0.2136	0.074770	-0.2630	8	-535.732	1088.1	1.75	0.092
16	4.78900	1.4490	0.79810	1.574	-0.1888			6	-538.042	1088.4	2.11	0.077
8	7.99000	0.6869	0.02850	1.561				5	-539.216	1088.7	2.36	0.068
22	7.92400	0.7110		1.580		-0.002372		5	-539.237	1088.7	2.40	0.066
56	-0.02886	0.4244	2.25800	2.454		0.032130	-0.2489	7	-537.193	1088.8	2.53	0.062

"Probability this is best model"

Rule:
Cumulative sum of weight > 0.95

```
weight95.exp<-subset(dredge.aicc, cumsum(weight) <= .95)
```

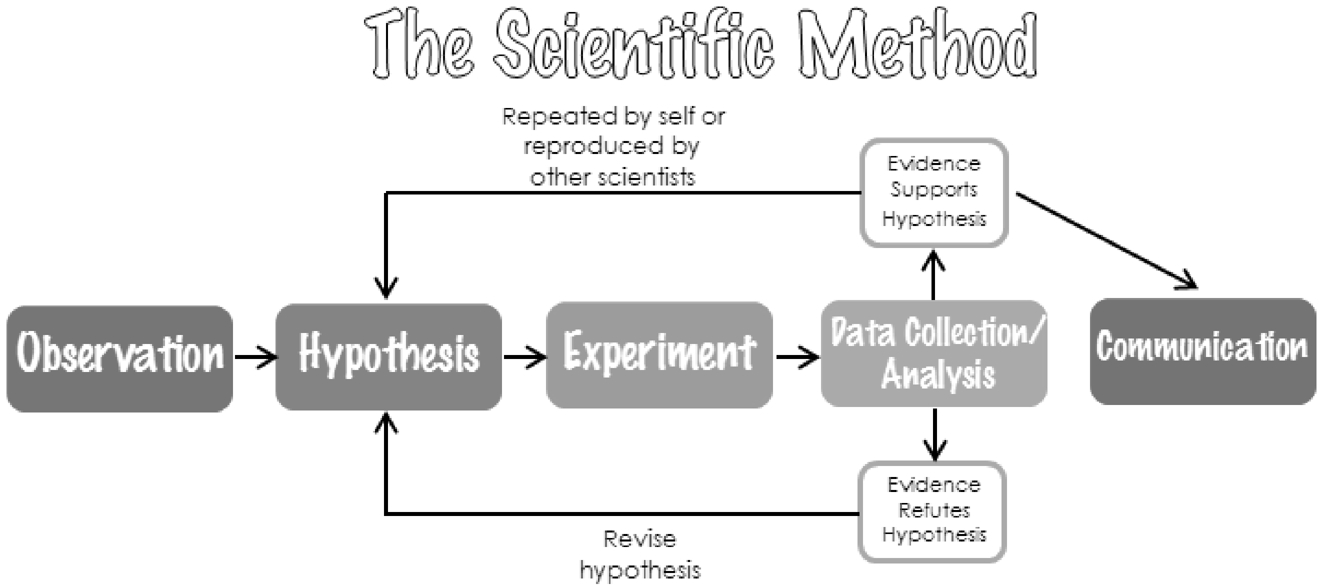
```
avgmod.95 <- summary(model.avg(weight95.exp))
```

Model-averaged coefficients:
(full average)

	Estimate	Std. Error	Adjusted SE	z value	Pr(>|z|)	
(Intercept)	1.590045	6.749103	6.762924	0.235	0.81412	
depth	0.951203	0.711846	0.714384	1.332	0.18302	
salinity	1.654541	1.622693	1.625290	1.018	0.30868	
temp	2.127162	0.662603	0.664174	3.203	0.00136	**
depth:salinity	-0.082184	0.126647	0.126898	0.648	0.51722	
salinity:temp	-0.144876	0.155288	0.155571	0.931	0.35172	
depth:temp	0.009229	0.062305	0.062573	0.147	0.88274	

(conditional average)

	Estimate	Std. Error	Adjusted SE	z value	Pr(>|z|)	
(Intercept)	1.59005	6.74910	6.76292	0.235	0.81412	
depth	0.95120	0.71185	0.71438	1.332	0.18302	
salinity	2.26223	1.49181	1.49567	1.513	0.13040	
temp	2.12716	0.66260	0.66417	3.203	0.00136	**
depth:salinity	-0.19961	0.12457	0.12519	1.594	0.11084	
salinity:temp	-0.25032	0.12357	0.12418	2.016	0.04383	*
depth:temp	0.03958	0.12428	0.12485	0.317	0.75126	



Parameter	Exploration	Inference	Prediction
Purpose	generate hypotheses	test hypotheses	forecast the future accurately
Priority	thoroughness	avoid false positives	minimize error
A priori hypotheses	not necessary	essential	not necessary, but may inform model specification
Emphasis on model selection	important	minimal	important
Key statistical tools	any	null hypothesis significance tests	AIC; regularization; machine learning; cross-validation; out-of-sample validation
Pitfalls	fooling yourself with over-fitted models with spurious covariate effects	misrepresenting exploratory tests as tests of a priori hypotheses	failure to rigorously validate prediction accuracy with independent data

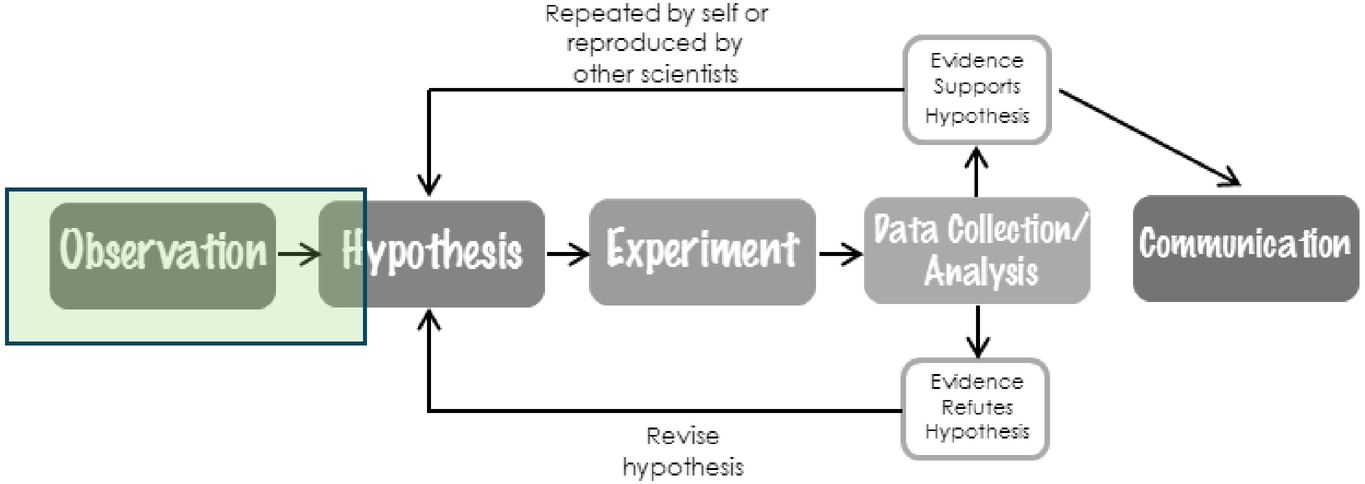
slido

Please download and install the
Slido app on all computers you use

What analyses do expect to use in your analyses?

ⓘ Start presenting to display the poll results on this slide.

The Scientific Method



Parameter	Exploration	Inference	Prediction
Purpose	generate hypotheses	test hypotheses	forecast the future accurately
Priority	thoroughness	avoid false positives	minimize error
A priori hypotheses	not necessary	essential	not necessary, but may inform model specification
Emphasis on model selection	important	minimal	important
Key statistical tools	any	null hypothesis significance tests	AIC; regularization; machine learning; cross-validation; out-of-sample validation
Pitfalls	fooling yourself with over-fitted models with spurious covariate effects	misrepresenting exploratory tests as tests of a priori hypotheses	failure to rigorously validate prediction accuracy with independent data

Exploratory Model Selection

Goal: describe the patterns in the data and generate hypotheses for future testing

Main trade-off: being thorough vs the need to avoid spurious relationships

Approach: lots of options

Important things to keep in mind:

- Avoid claims of confirmation in your result
- Be clear about your objectives
- Propose hypotheses based on your findings, but emphasize that you haven’t tested these hypotheses yet
- Don’t use a dataset for exploration to find a subset of variables, then write hypotheses based on these variables and use the same dataset to test the hypotheses. This may seem obvious, but it happens all the time.

Exploratory model selection process described in Tredennick et al. 2021

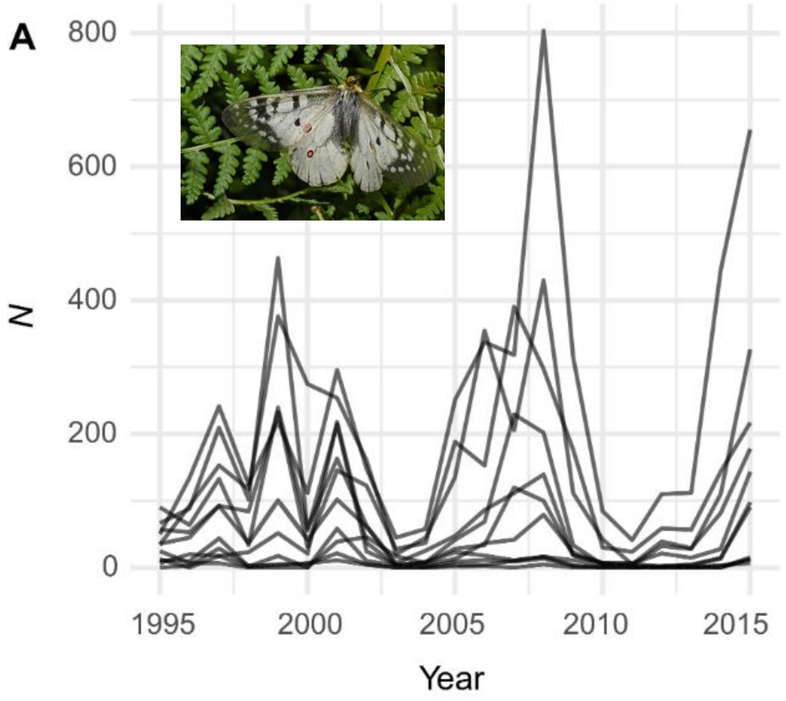
Objective: Which weather variables are associated with population growth rates of *Parnassius smintheus* (a butterfly)?

Starting point:
Possible predictors: 96 weather variables
Response: population growth rate of *P. smintheus*
20 years of data

Step 1:
Calculated the correlations between population growth rate and each of the 96 weather variables
Visually examined all variables with absolute value > 0.3
Outcome: kept 15 weather variables with high correlations with pop growth

Step 2:
Fit a model that included 15 weather variables from Step 1
Used stepwise NHST to remove variables and reduce the model
Used adjusted p-values to compensate for multiple hypothesis testing
Outcome: model with 6 variables remained

End point:
The six remaining variables are hypotheses that need to be tested using independent data



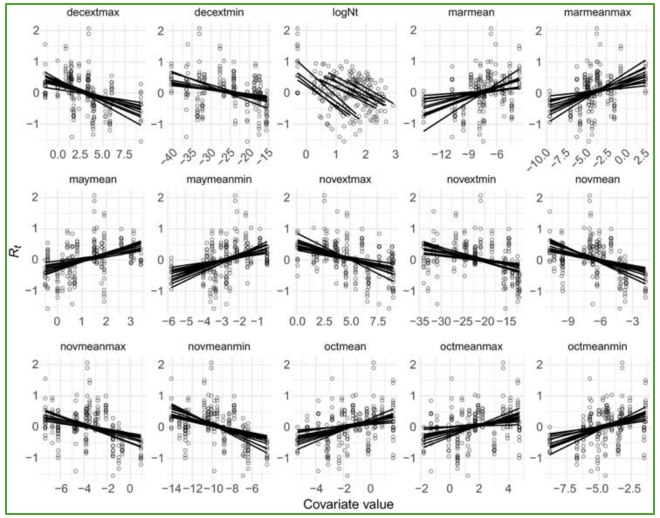
96 potential predictor variables

Calculate correlation with growth rate

|ρ| > 0.3

Filter based on correlation

15 predictors



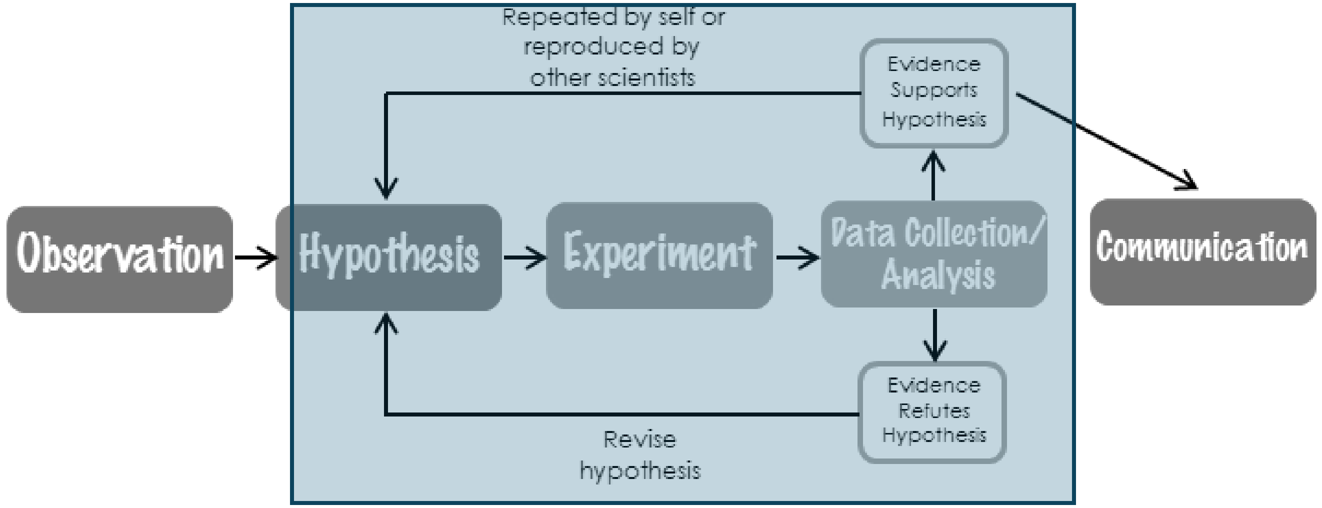
Filter individual predictors using drop1() in R

Covariate	Estimate	SE	*t*	drop1 *P*	Include	
(Intercept)	−0.10	0.30	−0.33			
decextmax	0.05	0.02	2.72	0.00	yes	yes
decextmin	−0.02	0.00	−5.17	0.00	yes	yes
logNt	−0.45	0.05	−8.30	0.00	yes	yes
marmeanmax	−0.04	0.01	−2.46	0.01	yes	no
maymean	0.23	0.04	5.92	0.00	yes	yes
novextmax	−0.07	0.01	−5.34	0.00	yes	yes
novmeanmax	−0.03	0.02	−1.71	0.00	yes	yes
octmeanmin	0.05	0.02	3.16	0.05	yes	no

After considering multiple testing?

"hypotheses that need to be tested using independent data"

The Scientific Method



Parameter	Exploration	Inference	Prediction
Purpose	generate hypotheses	test hypotheses	forecast the future accurately
Priority	thoroughness	avoid false positives	minimize error
A priori hypotheses	not necessary	essential	not necessary, but may inform model specification
Emphasis on model selection	important	minimal	important
Key statistical tools	any	NHST null hypothesis significance tests	AIC; regularization; machine learning; cross-validation; out-of-sample validation
Pitfalls	fooling yourself with over-fitted models with spurious covariate effects	misrepresenting exploratory tests as tests of a priori hypotheses	failure to rigorously validate prediction accuracy with independent data

Model selection for inference process described in Tredennick et al. 2021

Hypothesis: Extreme high temperatures during early winter reduce *P. smintheus* population growth rate, but only in years of low snow fall

Defined early winter as Nov and Dec
Averaged max temp over these two months and averaged amount of snowfall per month

Created two models:
1) Full: which included interaction between the max temp and snow fall
2) Reduced/alternative: excluding interaction

Used null-hypothesis significance testing to compare the two models
- anova(full model, reduced model) in R

Results suggested that the interaction should be retained in the model
Reported the results for the full model
Because the interaction was important, they did not test the significance of the main effects

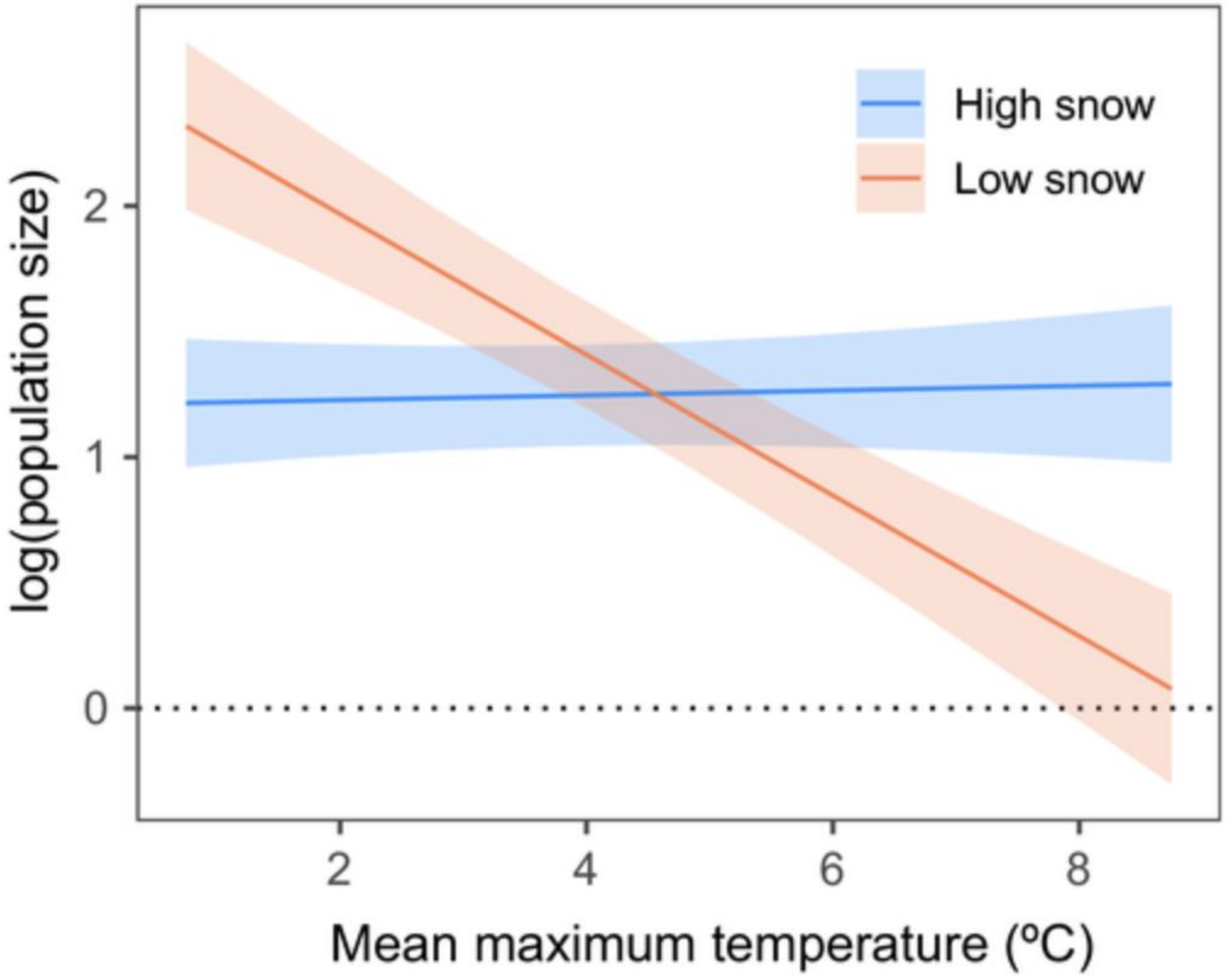
Model 1 (full)
Growth rate ~ Temperature * Snow

Model 2 (reduced)
Growth rate ~ Temperature + Snow

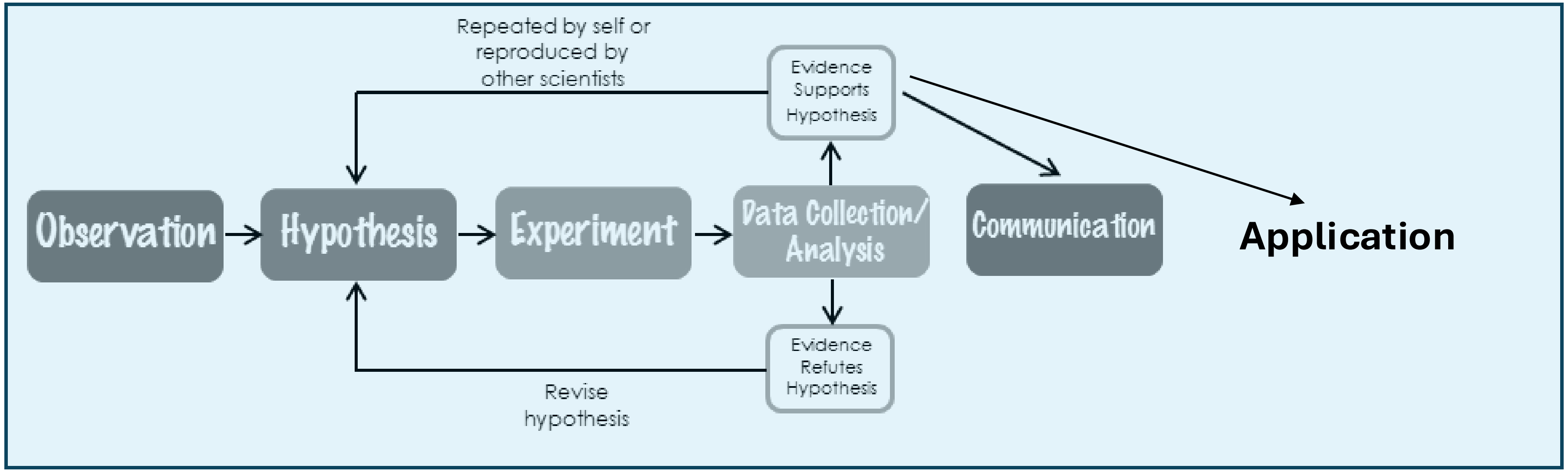
ANOVA
Should interaction be kept in model?

$(\chi^2 = 25.37, \text{df} = 1, \boxed{P < 0.001})$

Model 1 (full)
Growth rate ~ Temperature * Snow



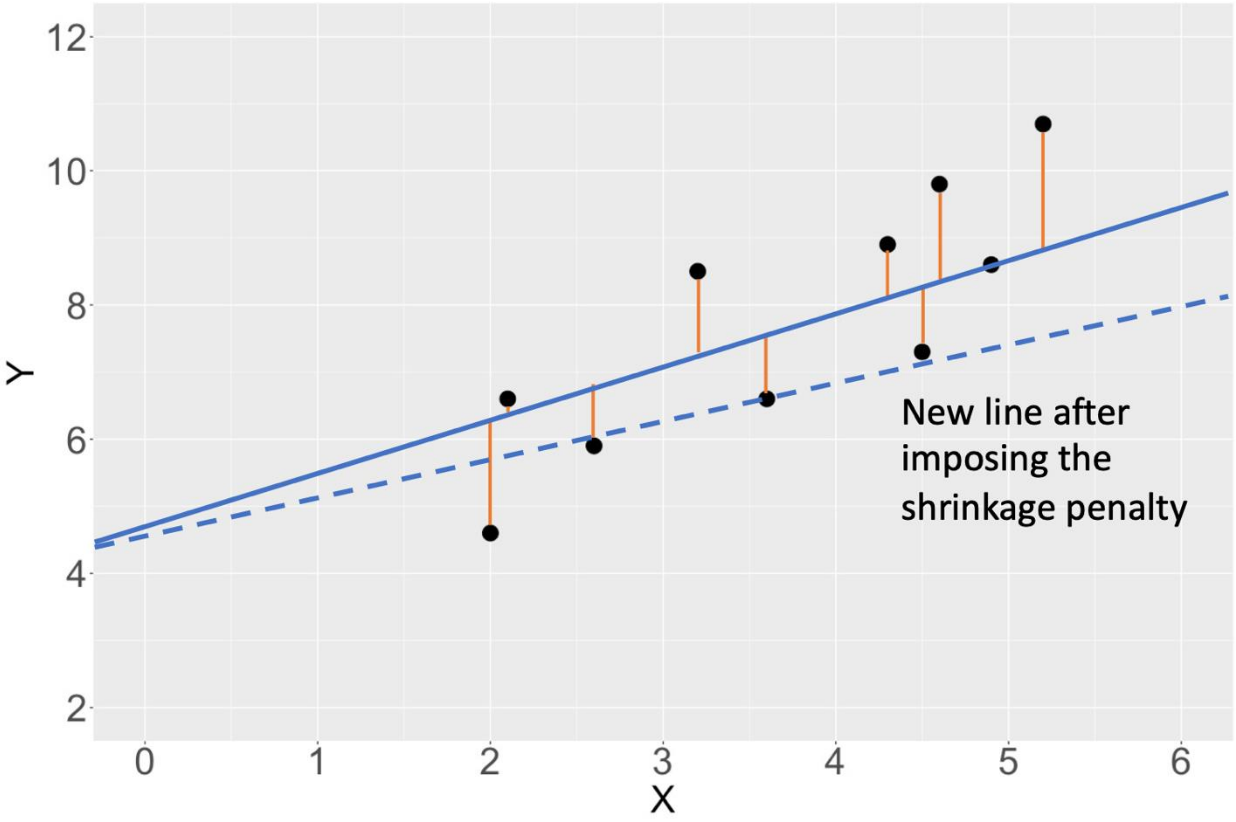
The Scientific Method

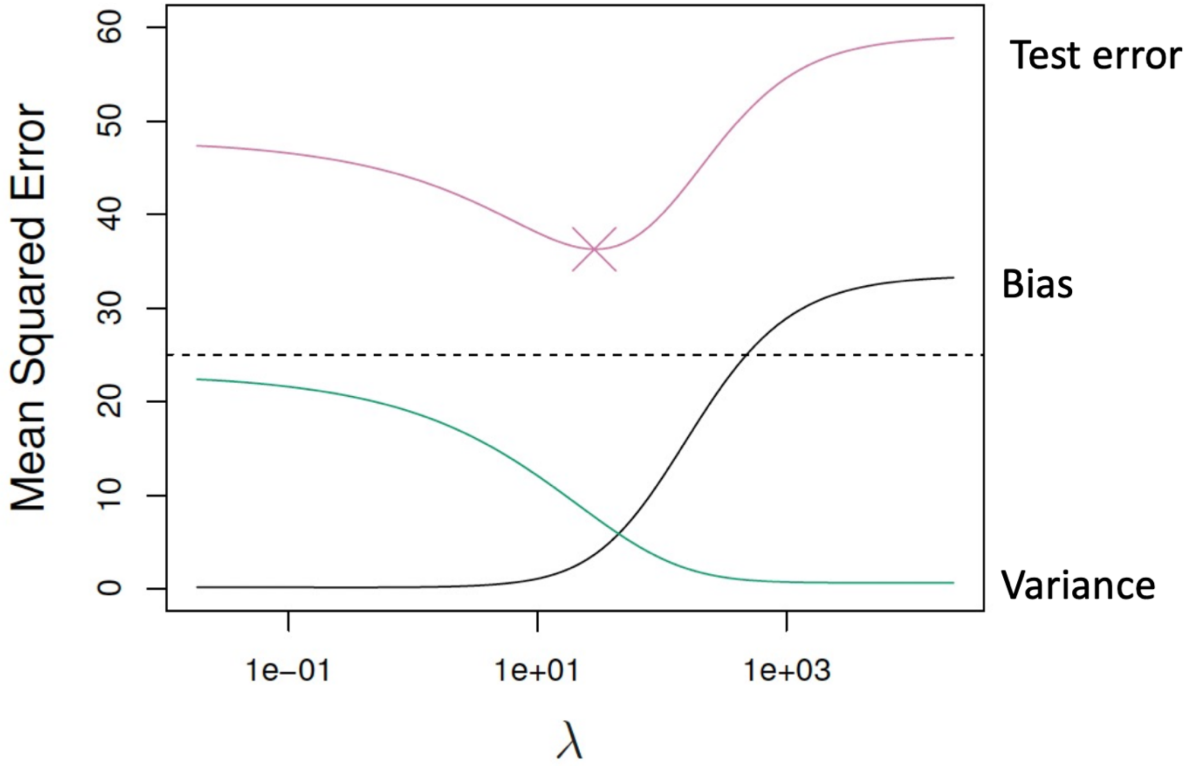


Parameter	Exploration	Inference	Prediction
Purpose	generate hypotheses	test hypotheses	forecast the future accurately
Priority	thoroughness	avoid false positives	minimize error
A priori hypotheses	not necessary	essential	not necessary, but may inform model specification
Emphasis on model selection	important	minimal	important
Key statistical tools	any	null hypothesis significance tests	AIC; regularization; machine learning; cross-validation; out-of-sample validation
Pitfalls	fooling yourself with over-fitted models with spurious covariate effects	misrepresenting exploratory tests as tests of a priori hypotheses	failure to rigorously validate prediction accuracy with independent data

Regularization

Reducing slopes can give better accuracy to predict future data





K-fold cross validation

Mixed effect models (more later this quarter)

$$Y_i = \beta_0 + \beta_1 \text{Genotype}_1 + \beta_2 \text{Genotype}_2 + \beta_3 \text{Genotype}_3 + \beta_1 \text{Block}_1 + \beta_2 \text{Block}_2 + \epsilon_i$$

```
lm(Yield ~ Block + Genotype, data = data)
```

AIC = **1102.502**

$$Y_i = \beta_0 + \beta_1 \text{Genotype}_1 + \beta_2 \text{Genotype}_2 + \beta_3 \text{Genotype}_3 + u_{\text{Block}_j} + \epsilon_{ij}$$

```
lmer(Yield ~ (1 | Block)+Genotype , data = data)
```

AIC= **1094.867**

Let yield on plot i in site j be

$$y_{ij} = \underbrace{X_{ij}\beta}_{\text{fixed effects (e.g., N rate, variety)}} + \underbrace{u_j}_{\text{site effect}} + \underbrace{\varepsilon_{ij}}_{\text{residual noise}}$$

with

$$u_j \sim \mathcal{N}(0, \sigma^2_{\text{site}}), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

independent.

- u_j is a **latent (unobserved) random effect** for site j.
- We **do not** treat each u_j as its own free parameter (that would be fixed effects). Instead, we assume all sites are draws from a shared distribution with variance σ^2_{site}.

Say each site has $n_j = 3$ plots. After fitting the fixed effects, site A's average residual is $\bar{r}_A = +1.3$ Mg/ha (above the global line).

Suppose REML gives $\hat{\sigma}^2_{\text{site}} = 0.50$ and $\hat{\sigma}^2 = 0.90$.

Shrinkage weight for site A:

$$w_A = \frac{0.50}{0.50 + 0.90/3} = \frac{0.50}{0.50 + 0.30} = \frac{0.50}{0.80} = 0.625.$$

BLUP for site A:

$$\hat{u}_A = 0.625 \times 1.3 \approx 0.81 \text{ Mg/ha}.$$

- **Not** the raw $+1.3$ (that would be the fixed-effect-like estimate with no pooling).
- **Pulled toward 0** because we have few plots per site and nontrivial residual noise.

If site B had $n_B = 30$ plots, its weight would be $0.50/(0.50 + 0.90/30) \approx 0.94$, i.e., **much less shrinkage**.

Given estimated variances $\hat{\sigma}^2_{\text{site}}, \hat{\sigma}^2$ and fixed effects $\hat{\beta}$, the **BLUP** (best linear unbiased predictor) of site j's effect is

$$\hat{u}_j = \underbrace{\frac{\hat{\sigma}^2_{\text{site}}}{\hat{\sigma}^2_{\text{site}} + \hat{\sigma}^2/n_j}}_{\text{shrinkage weight } w_j} \times \underbrace{(\bar{r}_j)}_{\text{site's mean residual}}, \quad \text{where} \quad \bar{r}_j = \frac{1}{n_j}\sum_i (y_{ij} - X_{ij}\hat{\beta}).$$