MODEL SELECTION AND STATISTICAL SIGNIFICANCE

Reporting the Best Science

Erik Kusch

erik.kusch@au.dk

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

24/02/2021

1. Model Selection
   - (adjusted) R^2
   - Mallow's C_p
   - Akaike Information Criterion (AIC)
   - Bayesian Information Criterion (BIC)
   - Receiver-Operator Characteristic (ROC)
2. Model Validation
   - Cross-Validation
   - Bootstrap
3. Building Models
   - Subset Selection
   - Shrinkage Methods
4. Statistical Significance
   - The p-value Conundrum
   - Alternatives
5. Summary
   - What Now?

1 Model Selection
- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristic (ROC)

2 Model Validation
- Cross-Validation
- Bootstrap

3 Building Models
- Subset Selection
- Shrinkage Methods

4 Statistical Significance
- The p-value Conundrum
- Alternatives

5 Summary
- What Now?

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

What? Why? How?

What - *Bias-Variance Trade-Off*:

- Trade-off between smooth and flexible models:
  - **Bias**: *error* that is *introduced by modelling a data/real life problem by a much simpler model*
  - **Variance**: *how much* $\hat{f}$ *(estimated mapping function of predictors and responses) would change (vary) if the training data set were to be changed*
- Simple models: High bias, low variance $\rightarrow$ **under-fitting**
- Complex models: Low bias, high variance $\rightarrow$ **over-fitting**

Why - to identify *Best Model*

- Finding the optimal trade-off between bias and variance allows for *most reliable* analyses

How - *Model Selection Criteria*:

- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristics (ROCs)
- ...

R^2

In `R`: `summary(...)$r.squared` with `...` being a regression object

Definition: Proportion of variation in Y that can be explained by regression using predictor(s) X. Values bound between 0 and 1.
Does **not penalize complex models**! **Large R^2** values do **not necessarily imply a good model**.

Calculation:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\frac{1}{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}(y_i - \overline{y}_i)^2} \quad (1)$$

TSS $\sum_n (y_i - \overline{y})^2$

RSS $\sum_{i=1} (y_i - \hat{y}_i)^2$

n Number of samples

Also called **Coefficient of Determination**.

R^2

In `R`: `summary(...)$r.squared` with `...` being a regression object

Definition: Proportion of variation in Y that can be explained by regression using predictor(s) X. Values bound between 0 and 1.
Does **not penalize complex models**! **Large R^2** values do **not necessarily imply a good model**.

Calculation:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\frac{1}{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}(y_i - \overline{y}_i)^2} \quad (1)$$

TSS $\sum(y_i - \overline{y})^2$

RSS $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

n Number of samples

Also called **Coefficient of Determination**.

Adjusted R^2

In `R`: `summary(...)$adj.r.squared` with `...` being a regression object

Definition: Proportion of variation in Y that can be explained by regression using predictor(s) X. Values bound between 0 and 1.
Does **penalize complex models**! **Increases** only if a **predictor** is **significant** and can **improve the model fit**.

Calculation:

$$R^2_{adj} = 1 - \frac{\frac{1}{n-p-1}(y_i - \hat{y}_i)^2}{\frac{1}{n}(y_i - \overline{y}_i)^2} = R^2 - (1 - R^2)\frac{p}{n-p-1} \quad (2)$$

TSS	$\sum(y_i - \overline{y})^2$
RSS	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
n	Number of samples
p	Number of parameters

The larger p is relative to n, the larger the adjustment will be.

Adjusted R^2

In `R`: `summary(...)$adj.r.squared` with `...` being a regression object

Definition: Proportion of variation in Y that can be explained by regression using predictor(s) X. Values bound between 0 and 1.
Does **penalize complex models**! **Increases** only if a **predictor** is **significant** and can **improve the model fit**.

Calculation:

$$R^2_{adj} = 1 - \frac{\frac{1}{n-p-1}(y_i - \hat{y}_i)^2}{\frac{1}{n}(y_i - \overline{y}_i)^2} = R^2 - (1 - R^2)\frac{p}{n-p-1} \qquad (2)$$

TSS	$\sum(y_i - \overline{y})^2$
RSS	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
n	Number of samples
p	Number of parameters

The larger p is relative to n, the larger the adjustment will be.

Mallow's C_p

In `R`: `Cp()` in `CombMSC` package

Definition: Estimate of test mean squared error of regression model fit using *ordinary least squares*.
Does **penalize complex models**!

Calculation:

$$C_p = \frac{1}{n}\left(RSS + 2p\hat{\sigma}^2\right) \tag{3}$$

RSS	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
n	Number of samples
p	Number of parameters
σ^2	Estimate of the variance of the error ε

Akaike Information Criterion (AIC)

In `R`: `AIC()` in base `R`

Definition: Estimate of test mean squared error of regression model fit using *maximum likelihood estimation*.
Does **penalize complex models**!

Calculation:

$$AIC = 2p + 2ln(L(\hat{\theta})) \tag{4}$$

p Number of parameters
$L(\hat{\theta})$ Maximum value of model likelihood function

For the standard linear model ($Y = \beta_0 + \sum_{j=1}^{p}(\beta_j X_j) + \varepsilon$) with Gaussian errors, maximum likelihood and least squares are the same thing leading to

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2p\hat{\sigma}^2\right) \tag{5}$$

Akaike Information Criterion (AIC)

In `R`: `AIC()` in base `R`

Definition: Estimate of test mean squared error of regression model fit using *maximum likelihood estimation*.
Does **penalize complex models**!

Calculation:

$$AIC = 2p + 2ln(L(\hat{\theta})) \tag{4}$$

p Number of parameters
$L(\hat{\theta})$ Maximum value of model likelihood function

For the standard linear model ($Y = \beta_0 + \sum_{j=1}^{p}(\beta_j X_j) + \varepsilon$) with Gaussian errors, maximum likelihood and least squares are the same thing leading to

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2p\hat{\sigma}^2\right) \tag{5}$$

Bayesian Information Criterion (BIC)

In `R`: `BIC()` in base `R`

Definition: Estimate of test mean squared error of regression model fit using *maximum likelihood estimation*.
Generally **penalizes complex models** more than other metrics!

Calculation:

$$BIC = ln(n)p + 2ln(L(\hat{\theta})) \tag{6}$$

n	Number of samples
p	Number of parameters
$L(\hat{\theta})$	Maximum value of model likelihood function

For the standard linear model ($Y = \beta_0 + \sum_{j=1}^{p}(\beta_j X_j) + \varepsilon$) with Gaussian errors we get:

$$BIC = \frac{1}{n}\left(RSS + ln(n)p\hat{\sigma}^2\right) \tag{7}$$

Bayesian Information Criterion (BIC)

In `R`: `BIC()` in base `R`

Definition: Estimate of test mean squared error of regression model fit using *maximum likelihood estimation*.
Generally **penalizes complex models** more than other metrics!

Calculation:

$$BIC = ln(n)p + 2ln(L(\hat{\theta})) \tag{6}$$

n	Number of samples
p	Number of parameters
$L(\hat{\theta})$	Maximum value of model likelihood function

For the standard linear model ($Y = \beta_0 + \sum_{j=1}^{p}(\beta_j X_j) + \varepsilon$) with Gaussian errors we get:

$$BIC = \frac{1}{n}\left(RSS + ln(n)p\hat{\sigma}^2\right) \tag{7}$$

Receiver-Operator Characteristic (ROC)

In `R`: `ROC()` in `Epi` package

Definition: Multiple metrics estimating classification accuracy.
Highlights **Trade-Off** between **Sensitivity** (rate of true positives) and **Specificity** (rate of true negatives)

Calculation:

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

TN	Number of true negative assignments
FP	Number of false positive assignments
TP	Number of true positive assignments
FN	Number of false negative assignments

The AUC of the ROC curve is indicative of how well the model performs overall. Higher scores represent better accuracy.

Receiver-Operator Characteristic (ROC)

In `R`: `ROC()` in `Epi` package

Definition: Multiple metrics estimating classification accuracy.
Highlights **Trade-Off** between **Sensitivity** (rate of true positives) and **Specificity** (rate of true negatives)

Calculation:

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

TN	Number of true negative assignments
FP	Number of false positive assignments
TP	Number of true positive assignments
FN	Number of false negative assignments

The AUC of the ROC curve is indicative of how well the model performs overall. Higher scores represent better accuracy.

1 Model Selection
- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristic (ROC)

2 Model Validation
- Cross-Validation
- Bootstrap

3 Building Models
- Subset Selection
- Shrinkage Methods

4 Statistical Significance
- The p-value Conundrum
- Alternatives

5 Summary
- What Now?

What? Why? How?

What - *Asses Model Inference*:

- How well do our models predict outcomes Y given inputs X?

Why - to quantify how much we *trust our models*

- Placing a lot of trust in a non-validated model can have terrible consequences
- Comparing how much to trust different models can help us chose the better model or weigh predictions according to accuracy

How - *Model Validation*:

- Training/Test Data Approach
- Leave-One-Out Cross-Validation (LOOCV)
- k-Fold Cross-Validation (k-fold CV)
- Bootstrap
- ...

What? Why? How?

What - *Asses Model Inference*:

- How well do our models predict outcomes Y given inputs X?

Why - to quantify how much we *trust our models*

- Placing a lot of trust in a non-validated model can have terrible consequences
- Comparing how much to trust different models can help us chose the better model or weigh predictions according to accuracy

How - *Model Validation*:

- Training/Test Data Approach
- Leave-One-Out Cross-Validation (LOOCV)
- k-Fold Cross-Validation (k-fold CV)
- Bootstrap
- ...

What? Why? How?

What - *Asses Model Inference*:

- How well do our models predict outcomes Y given inputs X?

Why - to quantify how much we *trust our models*

- Placing a lot of trust in a non-validated model can have terrible consequences
- Comparing how much to trust different models can help us chose the better model or weigh predictions according to accuracy

How - *Model Validation*:

- Training/Test Data Approach
- Leave-One-Out Cross-Validation (LOOCV)
- k-Fold Cross-Validation (k-fold CV)
- Bootstrap
- ...

What? Why? How?

What - *Asses Model Inference*:

- How well do our models predict outcomes Y given inputs X?

Why - to quantify how much we *trust our models*

- Placing a lot of trust in a non-validated model can have terrible consequences
- Comparing how much to trust different models can help us chose the better model or weigh predictions according to accuracy

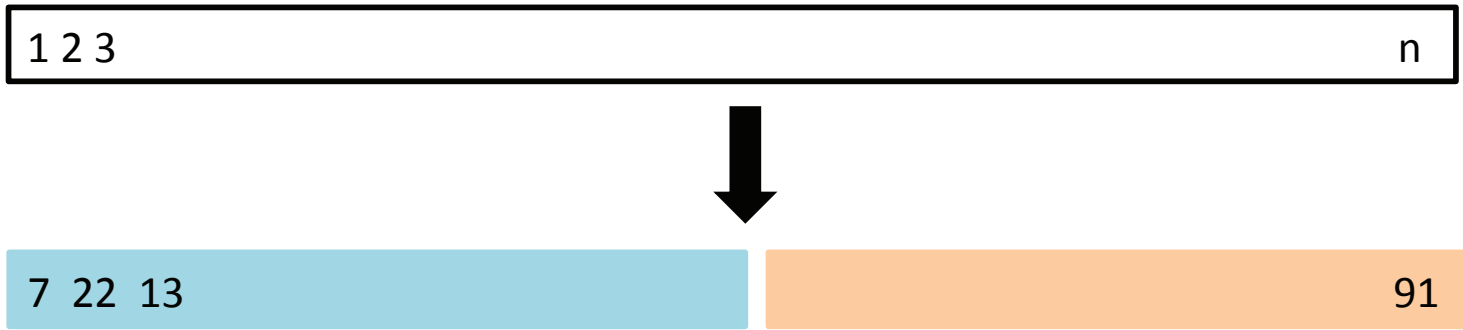
How - *Model Validation*:

- Training/Test Data Approach
- Leave-One-Out Cross-Validation (LOOCV)
- k-Fold Cross-Validation (k-fold CV)
- Bootstrap
- ...

Training/Test Data

Procedure:

1. Randomly split the data into **training** and **test** (also known as *hold-out*) **parts**.
2. Use the training part to build each possible model.
3. For each model, use the test part to calculate the test error rate.
4. Choose the model that gave the lowest test error rate.



Training/Test Data

Procedure:

1. Randomly split the data into **training** and **test** (also known as *hold-out*) **parts**.
2. Use the training part to build each possible model.
3. For each model, use the test part to calculate the test error rate.
4. Choose the model that gave the lowest test error rate.

Drawbacks:

- The **test error can be highly variable** on different sampling splits.
- Only part of the observations are used to fit the model (training data). Statistical methods tend to have **higher bias** when trained on fewer observations.

Also known as **Validation Data Cross-Validation**.

Leave-One-Out Cross-Validation (LOOCV)

Procedure:

1. Split data into **training** ($n-1$ observations) and **test** (1 observation) **parts**.
2. For i in $1, ..., n$:
   1. Fit the model on training part and obtain $\hat{y}_1$ for x_1 in the test part.
   2. Compute the corresponding test error, denoted as MSE_i.
3. Compute the final MSE for the each candidate model: $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$

Leave-One-Out Cross-Validation (LOOCV)

Procedure:

1. Split data into **training** ($n-1$ observations) and **test** (1 observation) **parts**.
2. For i in $1, ..., n$:
   1. Fit the model on training part and obtain $\hat{y}_1$ for x_1 in the test part.
   2. Compute the corresponding test error, denoted as MSE_i.
3. Compute the final MSE for the each candidate model: $CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$

Advantages over the validation set approach:

- Far **less bias**. Tends not to overestimate the test error rate as much as the validation set approach does.
- Performing LOOCV multiple times will **always yield the same results** - there is **no randomness in the training/validation set splits**.

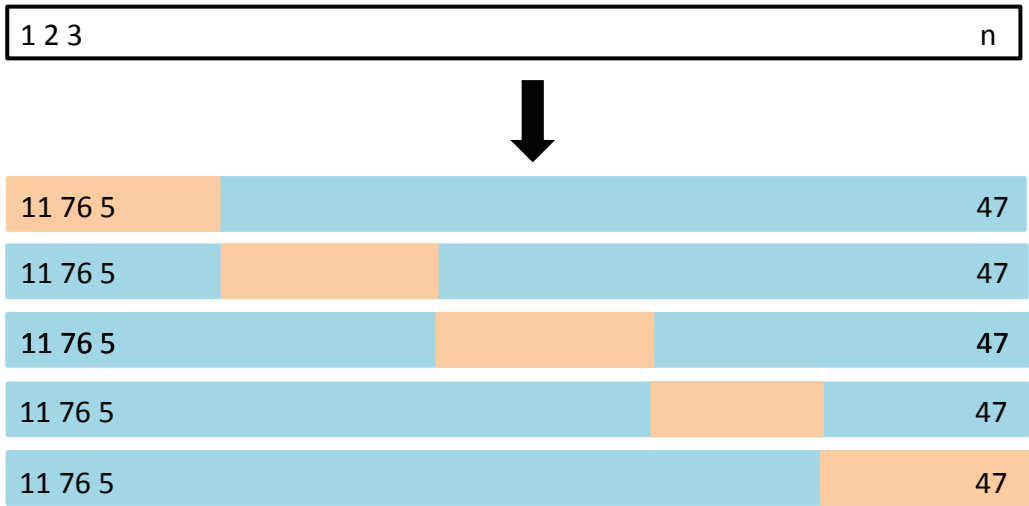
Drawbacks:

- **Computational intensity** (every model needs to be fit $n-1$ times)!

k-Fold Cross-Validation (k-fold CV)

Procedure:

1. For each candidate model:
   1. Fit model on $K-1$ training parts, compute error (MSE) on the test part.
   2. Repeat above step K times for different test parts resulting in $MSE_1, ..., MSE_k$.
   3. Calculate the corresponding k-fold CV value: $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$
2. Choose the model with the lowest $CV_{(k)}$



k-Fold Cross-Validation (k-fold CV)

Procedure:

1. For each candidate model:
   1. Fit model on $K-1$ training parts, compute error (MSE) on the test part.
   2. Repeat above step K times for different test parts resulting in $MSE_1, ..., MSE_k$.
   3. Calculate the corresponding k-fold CV value: $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$
2. Choose the model with the lowest $CV_{(k)}$

Advantage over LOOCV:

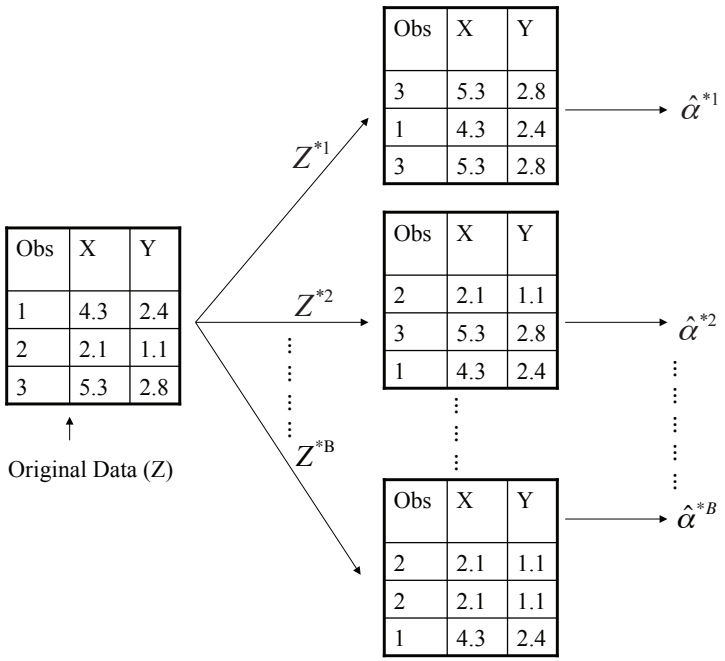
- Much **less computationally expensive**!

LOOCV is k-fold CV with $k = n$.

Bootstrap

Procedure:

1. Treat the observed sample $x = (x_1, x_2, ..., x_n)$ as population.
2. Obtain bootstrap sample $x^* = (x_1^*, x_2^*, ..., x_n^*)$ by resampling with replacement.
3. Repeat above step B times to receive B bootstrap samples, build models for each sample and estimate model parameters.



Bootstrap

Procedure:

1. Treat the observed sample $x = (x_1, x_2, ..., x_n)$ as population.
2. Obtain bootstrap sample $x^* = (x_1^*, x_2^*, ..., x_n^*)$ by resampling with replacement.
3. Repeat above step B times to receive B bootstrap samples, build models for each sample and estimate model parameters.

Advantages:

- Very **flexible** in its application to different methods.
- Allows **assessments of parameter uncertainty**.

Bootstrap estimates of a sampling distribution are analogous to histogram: one constructs a histogram of the available sample to obtain an estimate of the shape of the density function.

1 Model Selection
- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristic (ROC)

2 Model Validation
- Cross-Validation
- Bootstrap

3 Building Models
- Subset Selection
- Shrinkage Methods

4 Statistical Significance
- The p-value Conundrum
- Alternatives

5 Summary
- What Now?

Best Subset Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p$:
   1. Fit all $p_{(k)} = \frac{n!}{k!(n-k)!}$ models that contain exactly k predictors.
   2. Pick the best among these $p_{(k)}$ models, and call it M_k.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Low RSS or a high R^2 indicates a model with a **low training error**, whereas a good model is characterized by a low **test error rate**.

Advantages:

- **Simple** and conceptually appealing approach.

Drawbacks:

- **Suffers from computational limitations** and becomes computationally unfeasible for $p > 40$.

Best Subset Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p$:
   1. Fit all $p_{(k)} = \frac{n!}{k!(n-k)!}$ models that contain exactly k predictors.
   2. Pick the best among these $p_{(k)}$ models, and call it M_k.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Low RSS or a high $\mathbf{R^2}$ indicates a model with a **low training error**, whereas a good model is characterized by a low **test error rate**.

Advantages:

- **Simple** and conceptually appealing approach.

Drawbacks:

- **Suffers from computational limitations** and becomes computationally unfeasible for $p > 40$.

Best Subset Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p$:
   1. Fit all $p_{(k)} = \frac{n!}{k!(n-k)!}$ models that contain exactly k predictors.
   2. Pick the best among these $p_{(k)}$ models, and call it M_k.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Low RSS or a high $\mathbf{R^2}$ indicates a model with a **low training error**, whereas a good model is characterized by a low **test error rate**.

Advantages:

- **Simple** and conceptually appealing approach.

Drawbacks:

- **Suffers from computational limitations** and becomes computationally unfeasible for $p > 40$.

Best Subset Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p$:
   1. Fit all $p_{(k)} = \frac{n!}{k!(n-k)!}$ models that contain exactly k predictors.
   2. Pick the best among these $p_{(k)}$ models, and call it M_k.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Low RSS or a high $\mathbf{R^2}$ indicates a model with a **low training error**, whereas a good model is characterized by a low **test error rate**.

Advantages:

- **Simple** and conceptually appealing approach.

Drawbacks:

- **Suffers from computational limitations** and becomes computationally unfeasible for $p > 40$.

Forward Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p-1$:
   1. Consider all $p-k$ models that one predictor to M_k.
   2. Choose the best among these $p-k$ models, and call it M_{k+1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Forward Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p-1$:
   1. Consider all $p-k$ models that one predictor to M_k.
   2. Choose the best among these $p-k$ models, and call it M_{k+1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Forward Selection

Let M_0 denote the null model, which contains no predictors.

1. For $k = 1, 2, ..., p-1$:
   1. Consider all $p-k$ models that one predictor to M_k.
   2. Choose the best among these $p-k$ models, and call it M_{k+1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Backward Selection

Let M_p denote the full model, which contains p predictors.

1. For $k = p-1, p-2, ..., 1$:
   1. Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   2. Choose the best among these k models, and call it M_{k-1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Backward Selection

Let M_p denote the full model, which contains p predictors.

1. For $k = p-1, p-2, ..., 1$:
   1. Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   2. Choose the best among these k models, and call it M_{k-1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Backward Selection

Let M_p denote the full model, which contains p predictors.

1. For $k = p-1, p-2, ..., 1$:
   1. Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   2. Choose the best among these k models, and call it M_{k-1}.
2. Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Advantages over Best Subset Selection:

- **Reduced computational expense**. Only considers $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models instead of 2^p.

Drawbacks:

- **Not guaranteed to find the best possible model** out of all 2^p models containing subsets of the p predictors.

Shrinkage - What Do I Use It For?

Shrinking extreme values towards a central value results in a **better estimate** of the true mean.

Why?

- More stable parameter estimates (less extreme outliers considered)
- Reduction of sampling and non-sampling errors

Disadvantages

- Erroneous estimates if population has atypical mean. Knowing when this is the case is difficult.
- Possible introduction of bias.
- Shrunk models may fit new data worse than original models would.

How?

- Fitting a model with all p predictors
- Shrink estimated coefficients towards zero relative to the least squares estimates

Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform *variable selection*.

Shrinkage - What Do I Use It For?

Shrinking extreme values towards a central value results in a **better estimate** of the true mean.

Why?

- More stable parameter estimates (less extreme outliers considered)
- Reduction of sampling and non-sampling errors

Disadvantages

- Erroneous estimates if population has atypical mean. Knowing when this is the case is difficult.
- Possible introduction of bias.
- Shrunk models may fit new data worse than original models would.

How?

- Fitting a model with all p predictors
- Shrink estimated coefficients towards zero relative to the least squares estimates

Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform *variable selection*.

Shrinkage - What Do I Use It For?

Shrinking extreme values towards a central value results in a **better estimate** of the true mean.

Why?

- More stable parameter estimates (less extreme outliers considered)
- Reduction of sampling and non-sampling errors

Disadvantages

- Erroneous estimates if population has atypical mean. Knowing when this is the case is difficult.
- Possible introduction of bias.
- Shrunk models may fit new data worse than original models would.

How?

- Fitting a model with all p predictors
- Shrink estimated coefficients towards zero relative to the least squares estimates

Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform *variable selection*.

Shrinkage - What Do I Use It For?

Shrinking extreme values towards a central value results in a **better estimate** of the true mean.

Why?

- More stable parameter estimates (less extreme outliers considered)
- Reduction of sampling and non-sampling errors

Disadvantages

- Erroneous estimates if population has atypical mean. Knowing when this is the case is difficult.
- Possible introduction of bias.
- Shrunk models may fit new data worse than original models would.

How?

- Fitting a model with all p predictors
- Shrink estimated coefficients towards zero relative to the least squares estimates

Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform *variable selection*.

Shrinkage - What Do I Use It For?

Shrinking extreme values towards a central value results in a **better estimate** of the true mean.

Why?

- More stable parameter estimates (less extreme outliers considered)
- Reduction of sampling and non-sampling errors

Disadvantages

- Erroneous estimates if population has atypical mean. Knowing when this is the case is difficult.
- Possible introduction of bias.
- Shrunk models may fit new data worse than original models would.

How?

- Fitting a model with all p predictors
- Shrink estimated coefficients towards zero relative to the least squares estimates

Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform *variable selection*.

Ridge Regression

The ridge regression coefficient estimates, $\hat{\beta}^R$, are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{10}$$

Equation 10 **trades off two different criteria**:

- Coefficient estimates that fit the data well, by **making the RSS small**.
- The **shrinkage penalty** ($\lambda \sum_j \beta_j^2$) is small when $\beta_0, \beta_1, ..., \beta_p$ are close to zero, thus the shrinking penalty forces the estimates of β_j towards zero.

The **tuning parameter** λ controls the relative impact of these two terms on the regression coefficient estimates. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero (decreased variance but increased bias).

Ridge Regression

The ridge regression coefficient estimates, $\hat{\beta}^R$, are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{10}$$

Equation 10 **trades off two different criteria**:

- Coefficient estimates that fit the data well, by **making the RSS small**.
- The **shrinkage penalty** ($\lambda \sum_j \beta_j^2$) is small when $\beta_0, \beta_1, ..., \beta_p$ are close to zero, thus the shrinking penalty forces the estimates of β_j towards zero.

The **tuning parameter** λ controls the relative impact of these two terms on the regression coefficient estimates. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero (decreased variance but increased bias).

Ridge Regression

The ridge regression coefficient estimates, $\hat{\beta}^R$, are the values that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{10}$$

Equation 10 **trades off two different criteria**:

- Coefficient estimates that fit the data well, by **making the RSS small**.
- The **shrinkage penalty** ($\lambda \sum_j \beta_j^2$) is small when $\beta_0, \beta_1, ..., \beta_p$ are close to zero, thus the shrinking penalty forces the estimates of β_j towards zero.

The **tuning parameter** λ controls the relative impact of these two terms on the regression coefficient estimates. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero (decreased variance but increased bias).

The Lasso

The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

The β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso.

The penalty $|\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly 0 when the tuning parameter λ is sufficiently large.

The lasso **performs variable selection**.

$\rightarrow$ Models generated from the lasso (also referred to as *sparse models*) are generally much easier to interpret than those produced by ridge regression.

The Lasso

The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

The β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso.

The penalty $|\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly 0 when the tuning parameter λ is sufficiently large.

The lasso **performs variable selection**.

$\rightarrow$ Models generated from the lasso (also referred to as *sparse models*) are generally much easier to interpret than those produced by ridge regression.

The Lasso

The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

The β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso.

The penalty $|\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly 0 when the tuning parameter λ is sufficiently large.

The lasso **performs variable selection**.

$\rightarrow$ Models generated from the lasso (also referred to as *sparse models*) are generally much easier to interpret than those produced by ridge regression.

The Lasso

The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} (\beta_j x_{i,j}) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

The β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso.

The penalty $|\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly 0 when the tuning parameter λ is sufficiently large.

The lasso **performs variable selection**.

$\rightarrow$ Models generated from the lasso (also referred to as *sparse models*) are generally much easier to interpret than those produced by ridge regression.

Ridge vs. Lasso

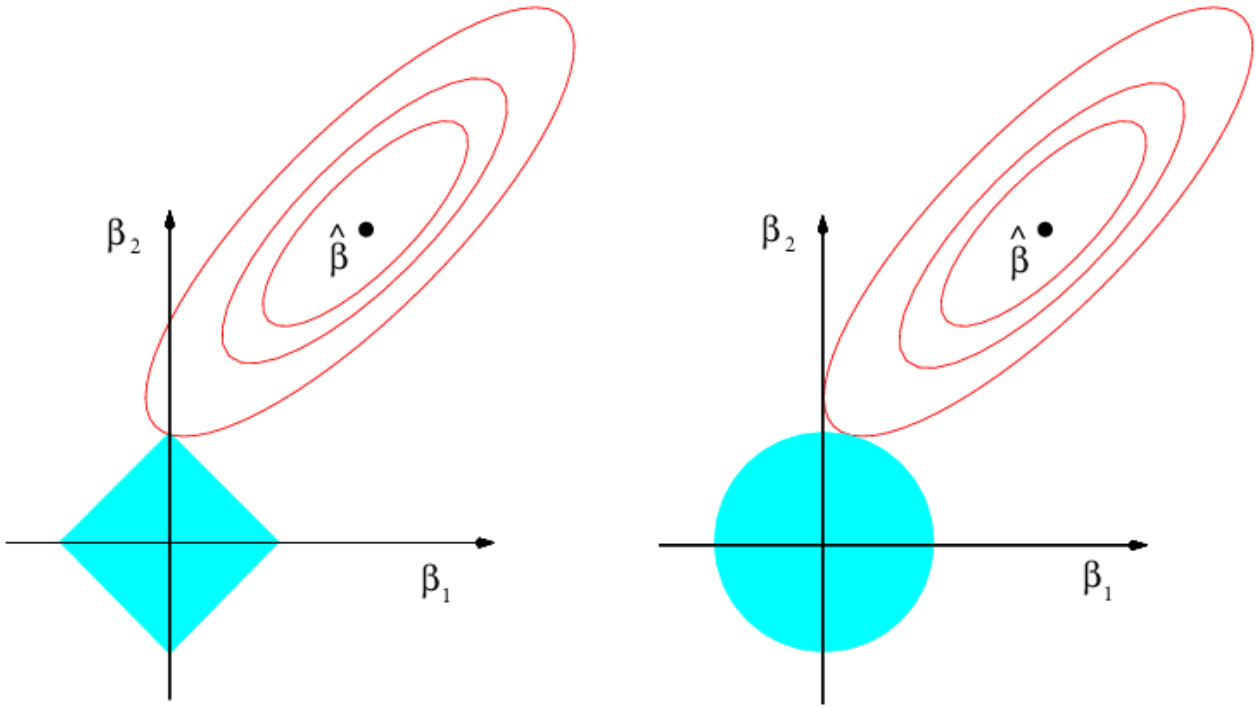


Figure 1: Error and constrain functions of the lasso and ridge regression: Both plots present a situation where $p = 2$. Contours of the error and constraint functions for the lasso (**left**) and ridge regression (**right**). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

1 Model Selection
- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristic (ROC)

2 Model Validation
- Cross-Validation
- Bootstrap

3 Building Models
- Subset Selection
- Shrinkage Methods

4 Statistical Significance
- The p-value Conundrum
- Alternatives

5 Summary
- What Now?

The p-value Conundrum

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the p-value will always be below the $.05$ cut-off, no matter the magnitude of the effect

The p-value Conundrum

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the p-value will always be below the $.05$ cut-off, no matter the magnitude of the effect

The p-value Conundrum

"The p-value is the probability of randomly obtaining an effect at least as extreme as the one in your sample data, given the null hypothesis."

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of p does not yield any information about the strength of an observed effect

Mathematical Quirks

- It varies strongly from sample-to-sample (depending on statistical power of the set-up)
- If the sample size is big enough, the p-value will always be below the $.05$ cut-off, no matter the magnitude of the effect

Effect sizes

"A measure of the magnitude of a statistical effect within the data (i.e. values calculated from test statistics)."

$\sim$ Nakagawa & Cuthill (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. Biological Reviews.

- **Intuitive** to interpret and often what we are interested in
- Three types for most situations:
  - r statistics (correlations)
  - d statistics (comparisons of values)
  - OR (odds ratio) statistics (risk measurements)
- These are **point estimates**
- Need to be reported alongside some information of credibility
- These are usually *standardized* thus enabling meta-studies

In `R`: https://cran.r-project.org/web/packages/compute.es/compute.es.pdf and https://cran.r-project.org/web/packages/effsize/effsize.pdf

Confidence Intervals

"Confidence intervals (CIs) answer the questions: 'How strong is the effect' and 'How accurate is that estimate of the population effect'."

$\sim$ Halsey (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biology Letters.

- **Intuitive** to interpret
- Answers the questions we are most interested in
- Does not require additional information of statistical certainty
- Combines **point estimates** and **range estimates**
- Removes some of the pressure of the *"file drawer problem"*
- Shares the same mathematical framework as the p-value calculation
- Especially useful in **data visualization**

In R, many functions come with in-built ways of establishing CIs.

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a indicator of model fit.

$\sim$ Burnham et al. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology.

- Used for **model selection and comparison**
- Lower AICs indicate better model fit
- One can establish contrasting models adhering to different hypothesis and identify which model suits the data best
- A proper hypothesis selection tool
- Model selection often comes with some degree of uncertainty
- Can be misused in step-wise model building procedures

In `R`, most model outputs can be assessed using the `AIC()` function.

Bayes Factor

" The minimum Bayes factor is simply the exponential of the difference between the log-likelihoods of two competing models."

$\sim$ Goodman (2001). Of P-Values and Bayes: A Modest Proposal. Epidemiology.

- **Intuitive** to interpret (Bayes Factor of 1/10 means that our study decreased the relative odds of the null hypothesis being true tenfold)
- Uses prior information to establish expected likelihoods thus enabling a progression in science

In R: https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf or direct Bayesian Statistics using JAGS or STAN (for example)

1 Model Selection
- (adjusted) R^2
- Mallow's C_p
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Receiver-Operator Characteristic (ROC)

2 Model Validation
- Cross-Validation
- Bootstrap

3 Building Models
- Subset Selection
- Shrinkage Methods

4 Statistical Significance
- The p-value Conundrum
- Alternatives

5 Summary
- What Now?

Summary

1. *"Which model explains my data better?"* $\rightarrow$ **Model Comparison**
   - Measures of fit, which penalise complex models (e.g.: Adjusted R^2, Mallow's C_p)
   - Information criteria (e.g. AUC, BIC, ROC)
   - Practice *model comparison* not *model selection*
2. *"How good is my model at predicting things?"* $\rightarrow$ **Model Validation**
   - Cross-Validation
   - Boot-Strapping
3. *"Which parameters should my model include?"* $\rightarrow$ **Model Building**
   - Model comparison/Lasso for subset selection
   - Shrinkage for robust parameter estimates
4. *"What do I report?"* $\rightarrow$ **Statistical Significance**
   - Don't use p-values!
   - Report Intervals and Effect Sizes.

Where do we go from here?

"Treat statistics as a science, and not a recipe"

$\sim$ Andrew Vickers

"The numbers are where the scientific discussion should start, not end!"

$\sim$ Regina Nuzzo