

Chapter 7 - Model Assessment and Selection

Geir Storvik and Aliaksandr Hubin

April 21, 2022

- Assume we have several possible models
- How to choose the best?
- How to evaluate the chosen model?

- Regression: $Y = f(\mathbf{x}) + \varepsilon$
- $\hat{Y} = \hat{f}(\mathbf{x})$ based on $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$
- Assume prediction of Y for new $\mathbf{x}$
- Note: $Y = f(\mathbf{x}) + \varepsilon$ is **independent** of $\hat{Y}$!
 - $\Rightarrow \varepsilon$ is **independent** of $\hat{Y}$!

$$\begin{aligned}\text{MSE} &= E[(\hat{Y} - Y)^2] \\ &= E[(\hat{Y} - f(\mathbf{x}) - \varepsilon)^2] \\ &= E[(\hat{Y} - f(\mathbf{x}))^2] - 2E[(\hat{Y} - f(\mathbf{x}))\varepsilon] + E[(\varepsilon)^2] \\ &= E[(\hat{Y} - f(\mathbf{x}))^2] + \sigma^2\end{aligned}$$

- Evaluation of model requires estimate of MSE.

$$\varepsilon \sim N(0, \sigma^2)$$
$$E\{(\hat{Y} - f(\mathbf{x}))\} E\{\varepsilon\}$$

Further decomposition

- We have

$$E[(\hat{Y} - f(\mathbf{x}))^2]$$

$$= E[(\hat{Y} - E[\hat{Y}] + E[\hat{Y}] - f(\mathbf{x}))^2]$$

$$= E[(\hat{Y} - E[\hat{Y}])^2] + E[(E[\hat{Y}] - f(\mathbf{x}))^2] +$$

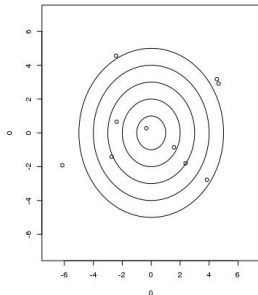
$$2E[(\hat{Y} - E[\hat{Y}])(E[\hat{Y}] - f(\mathbf{x}))] \rightarrow 0$$

$$= \text{Var}[\hat{Y}] + (E[\hat{Y}] - f(\mathbf{x}))^2$$

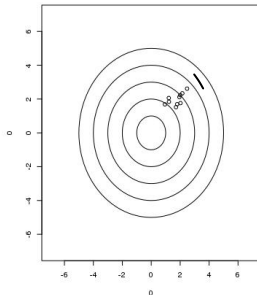
$$= \text{Variance} + \text{bias}^2$$

$$\begin{aligned} & E\{\hat{Y} - E\{\hat{Y}\}\} \\ &= E\{\hat{Y}\} - E\{E\{\hat{Y}\}\} \\ &= 0 \end{aligned}$$

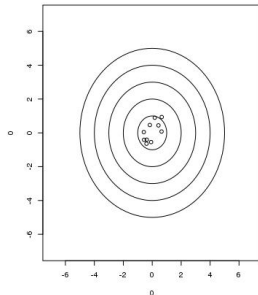
Stor varians, liten forventningsskjevhet



Liten varians, stor forventningsskjevhet



Liten varians, liten forventningsskjevhet



- $\text{MSE}(\mathbf{x}) = E[(\hat{Y}(\mathbf{x}) - Y(\mathbf{x}))^2]$ only related to **one** $\mathbf{x}$
 - Changed notation to emphasize dependence on $\mathbf{x}$
- In practice: Want to do predictions on **many** future $\mathbf{x}$'s
- We do not know these, but assume they are drawn from some distribution $p(\mathbf{x})$
- More natural to look at

$$\int_{\mathbf{x}} \text{MSE}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = E^{p(\mathbf{x})}[E[(\hat{Y}(\mathbf{x}) - Y(\mathbf{x}))^2]]$$

- Estimation of

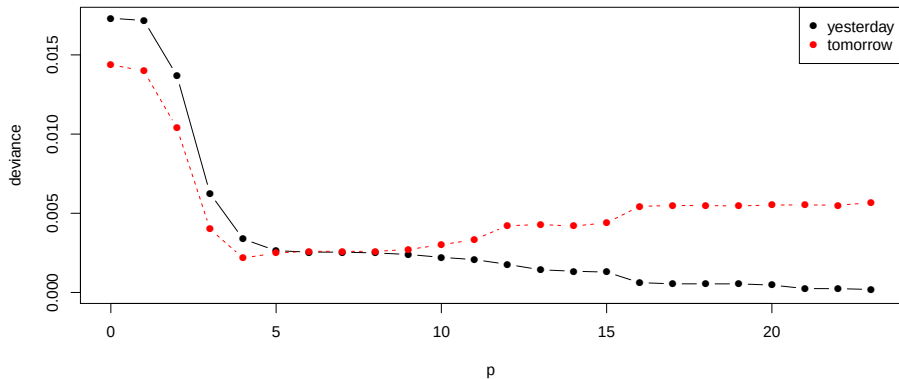
$$\int_{\mathbf{x}} \text{MSE}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = E^{p(\mathbf{x})} [E[(\hat{Y}(\mathbf{x}) - Y(\mathbf{x}))^2]]$$

- $p(\mathbf{x})$ unknown, but we can use the approximation

$$\begin{aligned} \int_{\mathbf{x}} \text{MSE}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} &\approx \frac{1}{n} \sum_{i=1}^n \text{MSE}(\mathbf{x}_i) \\ &= \frac{1}{n} \sum_{i=1}^n E[(\underbrace{\hat{Y}(\mathbf{x}_i)} - Y(\mathbf{x}_i))^2] \end{aligned} \quad \text{unbiased}$$

- "Obvious estimate": $\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$
- $Y(\mathbf{x})$ **independent** of $\hat{Y}(\mathbf{x})$
 - $\hat{Y}(\mathbf{x})$ based on $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$
 - $(Y, \mathbf{x})$ are **independent new data**
- y_i **dependent** of $\hat{y}_i$
- Leads to **underestimation** or **optimism**

Illustration: Yesterday-Today data `data Yesterday.R`



- Main problem: We need a **separate** data set for evaluation
- Possible approaches
 - One **training set** and one **test set**
 - Cross-validation
 - Correcting for underestimation/optimism
 - Criteria based on information

- Divide the full dataset **randomly** into two disjoint datasets
 - **Training set**: Used to **estimate/train** the model
 - **Test set**: Used to **Evaluate** the model
- See prototype example earlier
- **Strengths:**
 - Independent dataset for evaluation
 - Easy to apply
 - Easy to understand
- **Weaknesses:**
 - Reduces the dataset used for training the model
 - Might not be a problem if huge amounts of data are available
 - If the same test set is used to choose between models, the final error measure will be too optimistic
 - For $X_1, \dots, X_p$ all have $E[X_i] = \mu$, $\hat{\mu} = \min\{X_1, \dots, X_p\}$ will have $E[\hat{\mu}] < \mu$.

- Divide the full dataset **randomly** into three disjoint datasets
 - **Training set**: Used to **estimate/train** the model 50%,
 - **Validation set**: Used to **select** model 25%,
 - **Test set**: Used to **Evaluate** the **final** model 25%.
- Weakness: Require large datasets

- We want the **the training set** as large as possible
- Do also need a test set
- Cross validation:
 - Predict y_i by $\hat{y}_{-i}$ (all observations except $(y_i, \mathbf{x}_i)$)
 - Estimate MSE by

$$\widehat{\text{MSE}}_{LOO} = \sum_{i=1}^n (y_i - \hat{y}_{-i})^2$$

- Strengths
 - Using (almost) all data for training
 - Using all data as test
- Weaknesses
 - Using (slightly) different models for each prediction
 - Computer-intensive



- Divide the data into K groups
- For an observation y_i that belongs to group k_i , construct $\hat{y}_i^K$ by using data from all groups **except group k_i**
- Estimate MSE by



$$\widehat{\text{MSE}}_{\text{CV-K}} = \sum_{i=1}^n (y_i - \hat{y}_{-i}^K)^2$$

- Strengths (compared to LOO-CV)
 - Only need to train K models, less computer-intensive
 - All observations still used as tests
- Weaknesses
 - Using **more** different models for each prediction
 - Illustration: `credit data` and `new credit data` `credit train test.R`, `credit train test new.R`

- What if we have sparse data, and/or constrained computational power and/or difficult to divide data set?
- Within-sample estimate, $RMSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ known to **underestimate**
- Note, for regression, Gaussian noise

$$\underbrace{\log L(\hat{\theta})}_{\text{underestimate}} = \sum_{i=1}^n [-0.5 \log(2\pi) - 0.5 \log(\hat{\sigma}^2) - 0.5(y_i - \hat{y}_i)^2]$$
$$= \text{Const.} - 0.5n \log(\hat{\sigma}^2) - 0.5 \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- overestimate*
- In general: $\log L(\hat{\theta})$ will **underestimate** $\log L(\theta^{true})$
 - Can we correct for this?
 - Information based criteria:

$$IC = \underbrace{-2 \log L(\hat{\theta})}_{\text{underestimate}} + \underbrace{\text{penalty}(p)}_{\text{overestimate}}, \quad p = |\theta|$$

- Different choices of penalty identifies different **model selection** criteria
- Penalty should increase with p

- $\hat{\theta}$ is maximum likelihood estimate
- $\hat{\theta}_0$ is ML estimate under $H_0 : g_j(\theta) = 0, j = 1, \dots, q$

$$D = 2[\log L(\hat{\theta}) - \log L(\hat{\theta}_0)] \stackrel{H_0}{\approx} \chi_q^2$$

Likelihood ratio statistics

$$\text{P-value} \approx \Pr(\chi_q^2 > w)$$

- $E[\chi_q^2] = q$, so under H_0 ,

$$2 \log L(\hat{\theta}) \approx 2 \log L(\hat{\theta}_0) + q$$

- We would like the difference to be **significantly larger** in order to choose the larger model, so penalty should be **larger** than p .

- Akaike (1973) suggested minimizing **divergence** (non symmetric distance) between **true** model $p_*(\cdot)$ and **assumed** $p(\cdot; \theta)$

$$KL(p_*(\cdot), p(\cdot; \theta)) = E_{p_*} \left\{ \log \frac{p_*(Y)}{p(Y; \theta)} \right\}$$

$$= \underbrace{E_{p_*}[\log p_*(Y)]}_{\text{const}} - E_{p_*}[\log p(Y; \theta)] \xrightarrow{\theta} \min$$

where KL is the **Kullback-Leibler divergence**

- Measure of divergence between the distribution of future data and that predicted by the model
- Minimizing KL corresponds to maximizing $E_{p_*}[\log p(Y; \theta)] \xrightarrow{\theta} \max$
- $p_*(\cdot)$ is unknown but we can use the **estimate**

$$E_{p_*}[\log p(Y; \theta)] \approx p(\mathbf{y}; \theta) \stackrel{\text{ind}}{=} \frac{1}{n} \sum_{i=1}^n \log p(y_i; \theta)$$

- Unbiased for any fixed θ , but

$$\sum_{i=1}^n \log p(y_i; \hat{\theta}) \xrightarrow{\hat{\theta}} \max$$

become **overestimated** when $\hat{\theta}$ maximizes $\frac{1}{n} \sum_{i=1}^n \log p(y_i; \theta)$ (the ML estimate)

$$p(y) = \sum_{m \in M} p(y|m) p(m)$$

- Under regularity assumptions:

$$AIC = -2 \log p(y; \hat{\theta}) + 2p$$

is an approximately unbiased estimate of $E_{p_*}[\log p(Y; \theta)]$

- Alternatives

$$AIC_c = -2 \log p(y; \hat{\theta}) + 2p + \frac{2p(p+1)}{n-(p+1)}$$

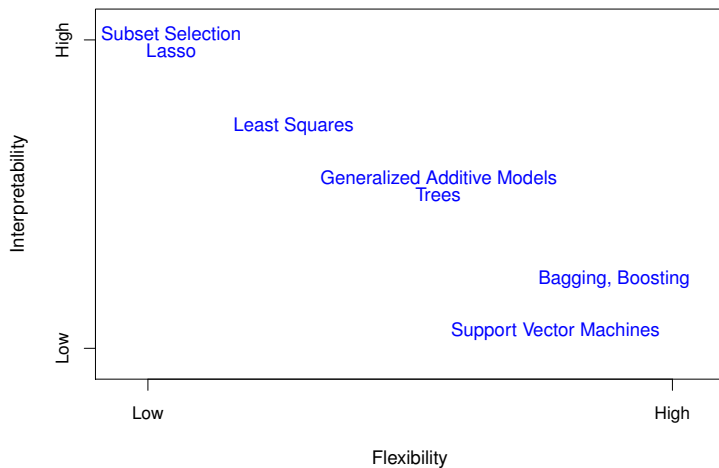
$$BIC/SIC_c = -2 \log p(y; \hat{\theta}) + \log(n)p$$

$$HQ_c = -2 \log p(y; \hat{\theta}) + cp \log \log n (c \geq 2)$$

$$p(m|y) \propto$$

$$\frac{p(y|m)p(m)}{\sum_{m' \in M} p(y|m')p(m')} = \frac{e^{-\frac{1}{2} BIC_m} p(m)}{\sum_{m' \in M} e^{-\frac{1}{2} BIC_{m'}} p(m')}$$

Overview of different methods



Selection criteria

- $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is based on q covariates.
- $R^2 = 1 - \frac{RSS}{TSS}$ too optimistic.
- $R^2_{adj} = 1 - \frac{RSS/(n-q-1)}{TSS/(n-1)}$

$$E \left[\frac{RSS}{n-q-1} \right] = \sigma^2$$

if linear model is true .

- Mallows C_p : ($\hat{\sigma}^2$ is an estimate of σ^2 based on all p covariats)

$$C_p = \frac{1}{n} (RSS + 2q\hat{\sigma}^2)$$

$$\bullet \text{ AIC} = -2 \log(L(\hat{\theta})) + 2|\theta|$$

$\rightarrow \min$

- Linear models: $\text{AIC} = \text{Const} + n \log(\hat{\sigma}^2) + 2(q+2)$ where $\hat{\sigma}^2$ is an estimate within the model of interest
- Will typically include too many variables

$$\bullet \text{ BIC} = -2 \log(L(\hat{\theta})) + \log(n)|\theta|$$

- Linear models: $\text{BIC} = \text{Const} + n \log(\hat{\sigma}^2) + \log(n)(q+2)$ where $\hat{\sigma}^2$ is an estimate within the model of interest
- Penalizes more complex models harder
- AIC and BIC can also be used outside linear models.

- Can be used for non-nested models (in contrast to likelihood ratio tests)
- Work directly on the data at hand
 - Can also be used for **dependent** data
- Less models to fit/train
- Rely more heavily on model assumptions than cross-validation

$$2^q \downarrow > 1000$$

- Assume **set** of models obtained by selection from a set of explanatory variables

$$S = \{1, x_1, x_2, \dots, x_q\}$$

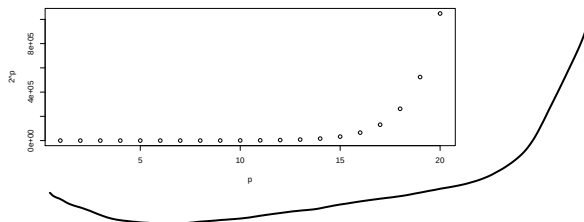
- Automatic procedure for variable selection:
 - Select the subset that minimizes AIC (or a similar criterion)
- In total 2^q possible subsets. Impossible to consider all
- Stepwise (forward) selection:**
 - Start with $S_0 = \{1\}$
 - Find x_j such that $S_1 = \{1, x_j\}$ is the best among all $\{1, x_k\}$ for $k = 1, \dots, q$
 - Find x_k such that $S_1 = \{1, x_j, x_k\}$ is the best among all $\{1, x_j, x_l\}$ for $l \neq j$
 - $\vdots$
- Stop when no improvement in AIC
- Alternative: Backwards selection
- Main problem: Computationally costly

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
 2. For $k = 0, \dots, p - 1$:
 - (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
 - (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

- Search through $\sum_{k=0}^{p-1} (p - k) = 1 + p(p+1)/2$ models
- Can also be used if $p \geq n$, stops at models of size $n - 1$.
- Can also give better **statistical** properties due to less chances of overfitting
- Also possible with backwards selection but then only if $n > p$.
- Also possible with hybrid approaches, switching between going backwards and forwards
- Illustration: `Hitters data Hitters.R`

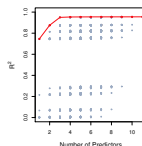
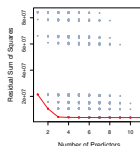
- With p variables there are 2^p possible models



- Search through **all** models

Algorithm 6.1 Best subset selection

- Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
- For $k = 1, 2, \dots, p$:
 - Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
- Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .



- Comparison** based on $p + 1$ models.

- R_{adj}^2 , C_p , AIC og BIC:
 - Use **all** data, but **correct for** overfitting within training data by including penalty term.
 - Based on assumed model
 - Usually fast to compute
- Cross-validation
 - Can be used more generally
 - Although more difficult with dependent/complex structures
 - Although AIC/BIC easier to use when **dependence/complex** structures between observations
 - Can not use all data when fitting/training
 - Validation based on (slightly) different models than those one would use in the end