Model Selection and Parameter Tuning

University of California, San Diego
Instructor: Armin Schwartzman

The need for model selection

□ After learning about polynomial regression, and regression after expansion in a basis (or more generally, in a dictionary), we now know that we can augment the model (almost) at will as long as at least one of the predictor variables is numerical.

□ However, expanding the model too much will lead to overfitting. This happens when instead of approaching $\mathbb{E}(y|\mathbf{x})$ (estimation) we start to approach the noisy data points (interpolation).

In particular, if the predictor observations $x_1, \dots, x_n \in \mathbb{R}$ are all distinct, it is possible to exactly interpolate the observations with a polynomial of degree at most n.

□ We assume here that we have a satisfactory model in that no curvature is revealed in diagnostic plots. We are now in a position were we want to prune the model back, meaning, remove some variables from the model.

Stepwise selection

□ We start with stepwise methods which are easy to motivate.

□ These classical methods proceed in a greedy fashion, including or removing one variable at a time. They key is that, at each stage, they compare models with the same number of parameters, and use an F-test to gage the statistical significance of a variable.

□ The main variants are:

▷ Forward stepwise selection

▷ Backward stepwise selection

▷ Hybrid stepwise selection

□ *Remark:* The sequential ANOVA procedure for choosing the degree of a polynomial model resembles and but is not exactly forward selection.

Forward stepwise selection

0. Set a threshold F_{IN} (=4 by default in R).

1. Start with the intercept $J_0 = \emptyset$.

2. Let SSE_k be the residual sum of squares (RSS) for

$$y = \beta_0 + \sum_{j \in J_k} \beta_j x_j + \varepsilon$$

3. For each $\ell \notin J_k$, let $\text{SSE}_k(\ell)$ be the RSS for

$$y = \beta_0 + \sum_{j \in J_k} \beta_j x_j + \beta_\ell x_\ell + \varepsilon$$

4. Define

$$F_k(\ell) = \frac{\text{SSE}_k - \text{SSE}_k(\ell)}{\text{SSE}_k(\ell)/(n-k-2)}$$

5. If $\max_\ell F_k(\ell) < F_{\text{IN}}$, stop and return J_k;
   otherwise $J_{k+1} = J_k \cup \{\arg\max_\ell F_k(\ell)\}$ and continue.

Backward stepwise selection

0. Set a threshold F_{OUT} (=4 by default in R).

1. Start with the full model $J_0 = \{1, \ldots, p\}$. (This requires $p < n$.)

2. Let SSE_k be the RSS for

$$y = \beta_0 + \sum_{j \in J_k} \beta_j x_j + \varepsilon$$

3. For each $\ell \in J_k$, let $\text{SSE}_k(\ell)$ be the RSS for

$$y = \beta_0 + \sum_{j \in J_k, j \neq \ell} \beta_j x_j + \varepsilon$$

4. Define

$$F_k(\ell) = \frac{\text{SSE}_k(\ell) - \text{SSE}_k}{\text{SSE}_k/(n-p+k-1)}$$

5. If $\min_\ell F_k(\ell) > F_{\text{OUT}}$, stop and return J_k;
   otherwise $J_{k+1} = J_k \setminus \{\arg\min_\ell F_k(\ell)\}$ and continue.

Hybrid stepwise selection

□ Set two thresholds $F_{\mathrm{IN}} \geq F_{\mathrm{OUT}}$ (both =4 by default in R). Then alternate between a forward and a backward step.

□ If $F_{\mathrm{IN}} < F_{\mathrm{OUT}}$, then the algorithm will loop endlessly!

The purpose of model selection

□ Although these stepwise methods make intuitive sense, we would like to formally define an objective for model selection:

What makes a model a good one?

□ Measuring the quality of fit by R^2 alone leads to overfitting, as we always end up choosing the full (most complex) model.

□ If interpretation is the main goal, then dropping near-linearly dependent variables may be satisfactory. (This may be done by inspecting the VIFs.)

□ If prediction is the main goal, then we need a way to compare the prediction abilities of different models. This is what we will formalize in these slides.

What follows is largely borrowed from (Hastie, Tibshirani and Friedman, 2009).

Comparing models

□ We are given an i.i.d. sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, assumed to be a realization of a regression model with additive error:

$$y = f(\mathbf{x}) + \varepsilon,$$

where (in a simple setting) $\mathbf{x}$ and ε are independent, and

$$\mathbb{E}(\varepsilon) = 0, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 < \infty.$$

In general, not much is known about f. All we have is the sample.

□ We want to compare all the linear models of the form

$$f_J(\mathbf{x}) = \sum_{j \in J} \beta_j x_j$$

where J is a subset of $\{1, \ldots, p\}$.

□ The predictor variables in $\mathbf{x}$ could be the result of a polynomial or spline expansion based on one or several original variables, and one of them could be an intercept, e.g., $x_1 \equiv 1$.

Least squares

□ We fit all the models by least squares. The same model fitted by a different method would effectively yield a different procedure for predicting y from $\mathbf{x}$.

□ Therefore, let

$$\widehat{f}_J(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_J^\top \mathbf{x}$$

be the model fitted on the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ by least squares.

Prediction error

□ We want to choose $J \subset \{1, \ldots, p\}$ expected prediction error, defined as

$$\mathrm{EPE}_J = \mathbb{E}_{\mathrm{data}}\, \mathbb{E}_{\mathrm{new}} \left[\left(y_{\mathrm{new}} - \widehat{f}_J(\mathbf{x}_{\mathrm{new}})\right)^2\right],$$

where

$$y_{\mathrm{new}} = f(\mathbf{x}_{\mathrm{new}}) + \varepsilon_{\mathrm{new}},$$

is a new observation not used in fitting the model, meaning $(\mathbf{x}_{\mathrm{new}}, y_{\mathrm{new}})$ is not part of the data.

The expectation is over $(\mathbf{x}_{\mathrm{new}}, y_{\mathrm{new}})$ and the data $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ used to fit the model f_J.

□ Other names: test error or generalization error.

Residual sum of squares and overfitting

□ Define the averaged squared error of model J on the data itself:

$$\mathrm{Err}_J = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{f}_J(\mathbf{x}_i))^2 = \frac{1}{n} \mathrm{SSE}_J$$

Note that $R_J^2 = 1 - \frac{n\mathrm{Err}_J}{\mathrm{SS}_Y}$.

□ This can be seen as an estimate of EPE_J, where:

1. The new observation is uniformly sampled from the data.
2. The expectation over multiple datasets is not taken.
   (We only have one dataset anyway.)

This is doing as if the empirical distribution were the population distribution.

However, using the $(\mathbf{x}_i, y_i)$'s as new observations is not justified when the same dataset was used for fitting the model.

□ Consequently, Err_J dramatically underestimates EPE_J, and minimizing it over all the models leads to overfitting in general — unless the number of observations far exceeds the number of variables.

Training, validation and test sets

□ With a large amount of data, the dataset would be divided into:

1. Training set – used to fit each model.
2. Validation set – used to estimate the prediction error of each model.
3. Final test set – used to estimate the prediction error of the chosen model.

□ The last step is optional and not needed for model selection per se.

Training, validation and test sets

□ We use the training set to fit all the models (those we are interested in comparing), obtaining:

$$\widehat{f}^{\text{train}}_J(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}^{\text{train}}_J, \qquad \widehat{\boldsymbol{\beta}}^{\text{train}}_J = \arg\min_{\mathbf{b}_J} \sum_{i \in \text{Train}}^{n} (y_i - \mathbf{b}_J^\top \mathbf{x}_i)^2.$$

□ We use the validation set to estimate the prediction error of each model

$$\widehat{\text{Err}}_J = \frac{1}{|\text{Val}|} \sum_{i \in \text{Val}} (y_i - \widehat{f}^{\text{train}}_J(\mathbf{x}_i))^2,$$

and choose the model that minimizes that

$$\hat{J} = \arg\min_J \widehat{\text{Err}}_J.$$

□ We note that $\widehat{\text{Err}}_{\hat{J}}$ is biased downward for $\mathbb{E}(\text{EPE}_{\hat{J}})$, since the selection that lead to $\hat{J}$ used the validation test in the process. Hence, we use the final test set to estimate $\mathbb{E}(\text{EPE}_{\hat{J}})$ by

$$\frac{1}{|\text{Test}|} \sum_{i \in \text{Test}} (y_i - \widehat{f}^{\text{train}}_{\hat{J}}(\mathbf{x}_i))^2.$$

Alternative methods when the data is scarce

□ In many situations, data is scarce / limited and other methods are used:

- ▷ Re-sampling: cross-validation, bootstrap.
- ▷ Analytical: adjusted R^2, C_p, AIC, BIC, MDL, SRM.

□ *Note.* In what follows, we implicitly focus on one given model J and drop the subscript J. This amounts to removing all the variables not in J. Let $d = |J|$, the number of variables in model J.

Cross-validation

□ K-fold cross-validation (CV) divides the data into K blocks. The model is fitted on $K-1$ blocks and tested on the remaining block. Each block plays the role of validation data in turn. This results in K different estimates for the prediction error. Their average is the cross-validation estimate.

□ Here is a pictorial description with $K = 5$ (common in practice):

	Block 1	Block 2	Block 3	Block 4	Block 5
Round 1	VALIDATE	TRAIN	TRAIN	TRAIN	TRAIN
Round 2	TRAIN	VALIDATE	TRAIN	TRAIN	TRAIN
Round 3	TRAIN	TRAIN	VALIDATE	TRAIN	TRAIN
Round 4	TRAIN	TRAIN	TRAIN	VALIDATE	TRAIN
Round 5	TRAIN	TRAIN	TRAIN	TRAIN	VALIDATE

Note that the training set is of size $n_K := (1 - \frac{1}{K})n < n$.

□ Let $\mathrm{EPE}(m)$ denote the expected prediction error (for a given model) based on fitting the model on data of size m.

K-fold CV is unbiased for $\mathrm{EPE}(n_K)$ and $\mathrm{EPE}(n_K) \geq \mathrm{EPE}(n)$.

PRESS and Generalized Cross-Validation (GVC)

□ When $K = n$, the method is called leave-one-out CV or prediction residual error sum of squares (PRESS).

□ Recall that the fitted values returned by the model are $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the corresponding hat matrix.

□ PRESS takes the form:

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - H_{ii}}\right)^2$$

where $\widehat{y}_{(i)} = \widehat{\boldsymbol{\beta}}_{(i)}^\top \mathbf{x}_i$ and $\widehat{\boldsymbol{\beta}}_{(i)}$ is the fit when excluding observation i.

This comes from the formula

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{v^TA^{-1}u + 1}$$

□ GCV approximates PRESS by:

$$\mathrm{GCV} = \frac{\mathrm{Err}}{(1 - \mathrm{trace}(\mathbf{H})/n)^2}$$

Bias-Variance Decomposition

□ Other methods for model selection do not resample the data, but rather go directly for an analytical estimator of EPE.

□ We already saw that the R^2 (or residual sum of squares on the data itself) is not a good criterion. This is because it does not take into consideration the complexity of the model.

□ All reasonable criteria implement a bias / variance trade-off:

- ▷ Bias – the model has to be complex enough to capture the complexity of the underlying functional relationship relating y to $\mathbf{x}$.
- ▷ Variance – the model has to be simple enough that we can estimate it reliably with the sample we are provided with.

□ In general

$$\text{model complexity} \nearrow \qquad \text{bias} \searrow \qquad \text{variance} \nearrow$$

□ Choosing a model amounts to balancing the squared bias and the variance.

Bias-Variance Decomposition

□ Suppose $(\mathbf{x}_0, y_0)$ is a new observation, where $\mathbf{x}_0$ is fixed and $y_0 = f(\mathbf{x}_0) + \varepsilon_0$.

□ The EPE at $\mathbf{x}_0$ can be decomposed into:

$$\begin{aligned}
\mathbb{E}\left((y_0 - \widehat{f}(\mathbf{x}_0))^2\right) &= \operatorname{Var}(\varepsilon_0) + \mathbb{E}\left[(\widehat{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2\right] \\
&= \sigma^2 + \left[\mathbb{E}\,\widehat{f}(\mathbf{x}_0) - f(\mathbf{x}_0)\right]^2 + \mathbb{E}\left[(\widehat{f}(\mathbf{x}_0) - \mathbb{E}\,\widehat{f}(\mathbf{x}_0))^2\right] \\
&= \sigma^2 + \operatorname{Bias}^2(\widehat{f}(\mathbf{x}_0)) + \operatorname{Var}(\widehat{f}(\mathbf{x}_0))
\end{aligned}$$

□ σ^2 is the irreducible error. It does not depend on the sample size.

□ For a linear model fitted by least squares, and assuming that the predictors are given:

$$\mathbb{E}\left((y_0 - \widehat{f}(\mathbf{x}_0))^2\right) = \sigma^2 + \left(f(\mathbf{x}_0) - (\mathbb{E}\,\widehat{\boldsymbol{\beta}})^T\mathbf{x}_0\right)^2 + \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$$

where $\mathbf{X}$ is the design matrix with ith row vector $\mathbf{x}_i$.

In-sample error

□ Let

$$y_i^{\mathrm{new}} = f(\mathbf{x}_i) + \varepsilon_i^{\mathrm{new}}$$

denote a new response at $\mathbf{x}_i$.

□ Define the in-sample error as the expected average error over *new* observations at the *same* design points:

$$\mathrm{Err}_{\mathrm{in}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathrm{data}}\, \mathbb{E}_{\mathrm{new}} (y_i^{\mathrm{new}} - \widehat{f}(\mathbf{x}_i))^2$$

□ This is very close to EPE, and coincides with it when the $\mathbf{x}_i$'s are on a regular grid and these are all the locations (in $\mathbf{x}$) that we are interested in (e.g., in signal processing).

Optimism

□ The optimism is defined as

$$\mathrm{op} = \mathrm{Err}_{\mathrm{in}} - \mathbb{E}(\mathrm{Err}) = \frac{2}{n}\sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{y}_i)$$

Hence, the more we overfit, the larger the optimism.

□ For a linear model with d variables (counting the intercept, if there is one),

$$\sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{y}_i) = d\sigma^2,$$

□ We estimate the in-sample error by

$$\widehat{\mathrm{Err}_{\mathrm{in}}} = \mathrm{Err} + \widehat{\mathrm{op}}$$

where $\widehat{\mathrm{op}}$ is estimated by plugging in an estimator for σ^2. Note that Err is already an average, so we estimate its expectation by itself.

Mallow's C_p

□ Assume a linear model with d variables.

□ Mallow's C_p is the estimate of the in-sample error:

$$C_p = \mathrm{Err} + 2\,\frac{d}{n}\,\widehat{\sigma}_*^2$$

where $\widehat{\sigma}_*^2$ is often chosen as the estimate from the full model.

□ Sometimes Mallow's C_p is defined as

$$C_p = \frac{\mathrm{SSE}}{\widehat{\sigma}_*^2} + 2d - n$$

Akaike Information Criterion (AIC)

□ Assume a general parametric model $y|\mathbf{x} \sim \phi_{\theta(\mathbf{x})}$, meaning y given $\mathbf{x}$ has density $\phi_{\theta(\mathbf{x})}$ (with respect to some dominating measure).

For a linear model with standard assumptions, $\phi_{\theta(\mathbf{x})}$ is normal with mean $\mathbf{b}^\top\mathbf{x}$ and variance σ^2, so that $\theta(\mathbf{x}) = \{\mathbf{b}^\top\mathbf{x}, \sigma^2\}$ (unless σ^2 is known).

□ The Akaike information criterion (AIC) is defined as

$$\mathrm{AIC} = -2 \sum_{i=1}^{n} \log \phi_{\widehat{\theta}(\mathbf{x}_i)}(y_i) + 2d,$$

where $\widehat{\theta}$ is the MLE for θ based on a sample $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, and d is a measure of model complexity or effective number of parameters.

□ For a linear model with i.i.d. normal errors and d variables (incl. intercept),

$$\mathrm{AIC} \quad = \quad n \log(\mathrm{Err}) + 2d + \mathrm{constant}$$

If the variance σ^2 is known

$$\mathrm{AIC} = \frac{n}{\sigma^2}\mathrm{Err} + 2d + \mathrm{constant}$$

in which case the AIC criterion is equivalent to Mallow's C_p.

Bayesian Information Criterion (BIC)

□ Bayesian information criterion (BIC) is similar to AIC:

$$\mathrm{BIC} = -2 \sum_{i=1}^{n} \log P_{\widehat{\theta}(\mathbf{x}_i)}(y_i) + (\log n)\ d$$

□ BIC penalizes model complexity more than AIC.

□ In theory, BIC is consistent while AIC is not, as the sample size increases.

□ In practice, BIC seems to select overly simple models. (From hearsay...)

□ BIC coincides with minimum description length (MDL), a criterion arising from information theory.

Beyond linear models

□ The estimates that we discussed so far (data splitting into training and validation sets; cross-validation; the bootstrap) apply to any procedure for estimating the underlying regression function (and beyond).

□ For example, cross-validation (and GCV) are commonly used for selecting tuning parameters, for example when using smoothing splines. In that case, we are comparing models f_λ for several choices of tuning parameter λ, which now plays the role of J above.