Statistical Learning and Data Mining

Module 5: Model Selection

Semester 2, 2017

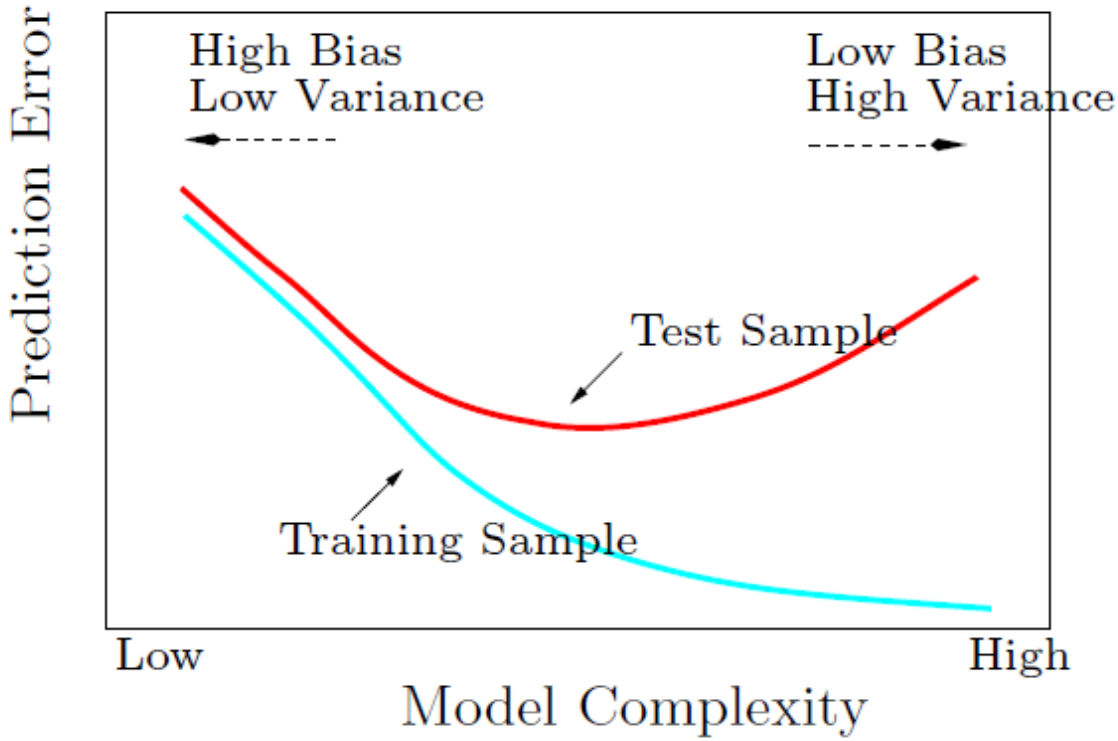
Discipline of Business Analytics, The University of Sydney Business School

Module 5: Model Selection

1. Validation set
2. Cross validation
3. Optimism
4. Analytical criteria
5. Comparison of model selection methods
6. Application
7. Limitations of model selection

Introduction

In Module 1, we introduced the fundamental concept of the bias-variance trade-off for estimation.



The bias-variance trade-off (key concept)

- Increasing model complexity brings higher flexibility and therefore lower bias. However, this comes at a cost of higher variance: there is higher sample variability when estimating complex models. Overfitting can be a problem.

- Decreasing model complexity leads to lower variance. However, simpler models may not be sufficiently flexible to capture the underlying patterns in the data, leading to higher bias.

The bias-variance trade-off

To review, we use the training data to estimate the additive error model

$$Y = f(\boldsymbol{x}) + \varepsilon$$

leading to an estimator $\widehat{f}(\boldsymbol{x}_0)$ for the regression function at given input point $\boldsymbol{x}_0$, $f(\boldsymbol{x}_0)$.

We can decompose the estimation error as

$$\begin{aligned} E(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0))^2 &= \Big[E(\widehat{f}(\boldsymbol{x}_0)) - f(\boldsymbol{x}_0)\Big]^2 + E([\widehat{f}(\boldsymbol{x}_0) - E(\widehat{f}(\boldsymbol{x}_0))]^2) \\ &= \text{Bias}^2[\widehat{f}(\boldsymbol{x}_0)] + \text{Var}[\widehat{f}(\boldsymbol{x}_0)] \end{aligned}$$

Examples

Linear regression. Adding predictors increases model complexity. Least squares estimates have high variability when the number of inputs is large, but excluding relevant predictors leads to bias.

Review: Module 2 slides 32 and 33.

Examples

KNN regression. Reducing the number of neighbours increases model complexity. Closer neighbours means lower bias. However, averaging fewer observations increases variance.

Review: Module 3 slide 13.

Model selection (key concept)

Model selection methods estimate the generalisation performance of a model based on the training data, allowing us to choose between models with different degrees of complexity.

We select the model that is estimated to have the best predictive ability.

Approaches to model selection

Validation set. Randomly split the training set into two, one part for training the model, and validation set for selecting the model complexity.

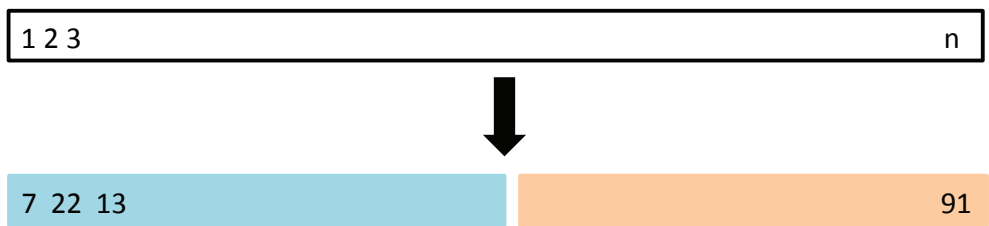
Resampling methods. Estimate generalisation performance by generating multiple splits of the training data.

Analytical criteria. Use analytical results to penalise training performance to account for overfitting.

Validation set

Validation set (key concept)

In the **validation set** approach, we randomly split the training data into a training set and a validation set. We select the model with the best predictive performance in the validation set.



Typically, we use 50-80% of the data for the training set.

Training, validation, and test split

1. Estimate different models on the training data.
2. Predict the observations in the validation set.
3. Select the model with best validation set performance.
4. Re-estimate the selected model by combining the training and validation sets.
5. Predict the test data with the selected model.

Validation set

The validation set approach has serious limitations when the size of the training data is not large. The model may not have enough data to train on, and there may not be enough cases in the validation set to reliably estimate generalisation performance.

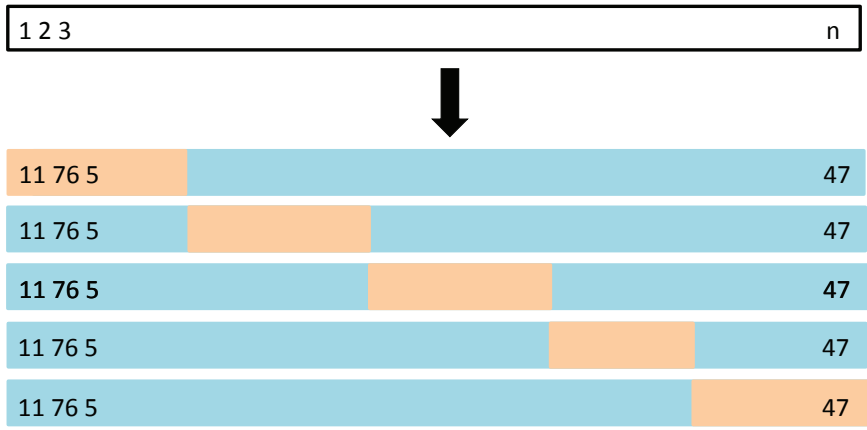
We turn instead to resampling methods.

Cross validation

Cross validation

Cross validation methods are based multiple random training/validation set splits. Unlike in the validation set approach, each observation gets a turn at being predicted.

K-fold cross-validation (key concept)



The idea of K-fold cross validation is simple:

1. We randomly split the training sample into K **folds** of roughly equal size.
2. For each fold $k \in \{1, \ldots, K\}$, we estimate the model on all other folds combined, and use k as the validation set.
3. The cross validation error is the average error across the K validation sets.

K-fold cross-validation

5-fold and 10-fold CV. $K = 5$ or $K = 10$ folds are common choices for cross validation.

Leave one out cross validation (key concept). If we set $K = N$, this is called leave one out cross validation, or **LOOCV**. For each observation i, we train the model on all other observations, and predict i.

Leave one out CV

Algorithm Leave one out CV for regression

1: **for** i=1:N **do**
2: Assign observation i to the validation set.
3: Assign observations $1, \ldots, i-1, i+1 \ldots, N$ to the training set $\mathcal{D}_{-i}$.
4: Estimate the model using the training set $\mathcal{D}_{-i}$.
5: Compute the prediction $\widehat{f}^{-i}(\boldsymbol{x}_i)$.
6: Compute the squared error $(y_i - \widehat{f}^{-i}(\boldsymbol{x}_i))^2$.
7: **end for**

8: Compute the leave-one-out MSE:

$$\text{MSE}_{\text{CV}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{f}^{-i}(\boldsymbol{x}_i))^2$$

LOOCV vs K-fold cross-validation

LOOCV. Approximately unbiased estimator of the expected prediction error. However, it can have high variance in some settings (since the training sets are very similar for every prediction) and a high computational cost (except in special cases).

K-fold. Lower computational cost and may have lower variance. However, it is subject to bias since the training sets are smaller than N.

Cross validation: recommendations

One standard deviation rule. Pick the simplest model within one standard deviation of the model with the lowest cross validated errors.

Choice of K. There are no general guidelines for choosing K since the trade-off between variance, bias, and computational cost is highly context specific. The variance of LOOCV tends to be relatively low with stable estimators such as linear regression.

Many predictors. When there are many predictors, pre-screening based on the entire training set may result in misleading CV.

Leave one out CV for linear regression

For a linear regression estimated by OLS, we can use a shortcut compute the leave-out-one errors without having to re-estimate the model. Given the OLS fitted values

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

we can show that the leave-one-out MSE is

$$\text{MSE}_{\text{CV}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f}^{-i}(\boldsymbol{x}_i))^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1 - H_{ii}}\right]^2,$$

where H_{ii} is the ith diagonal element of the hat matrix $\boldsymbol{H}$.

Generalised cross validation

The previous method applies to many situations in which we have fitted values of the type

$$\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}.$$

The **generalised cross validation** method approximates the leave one out MSE as

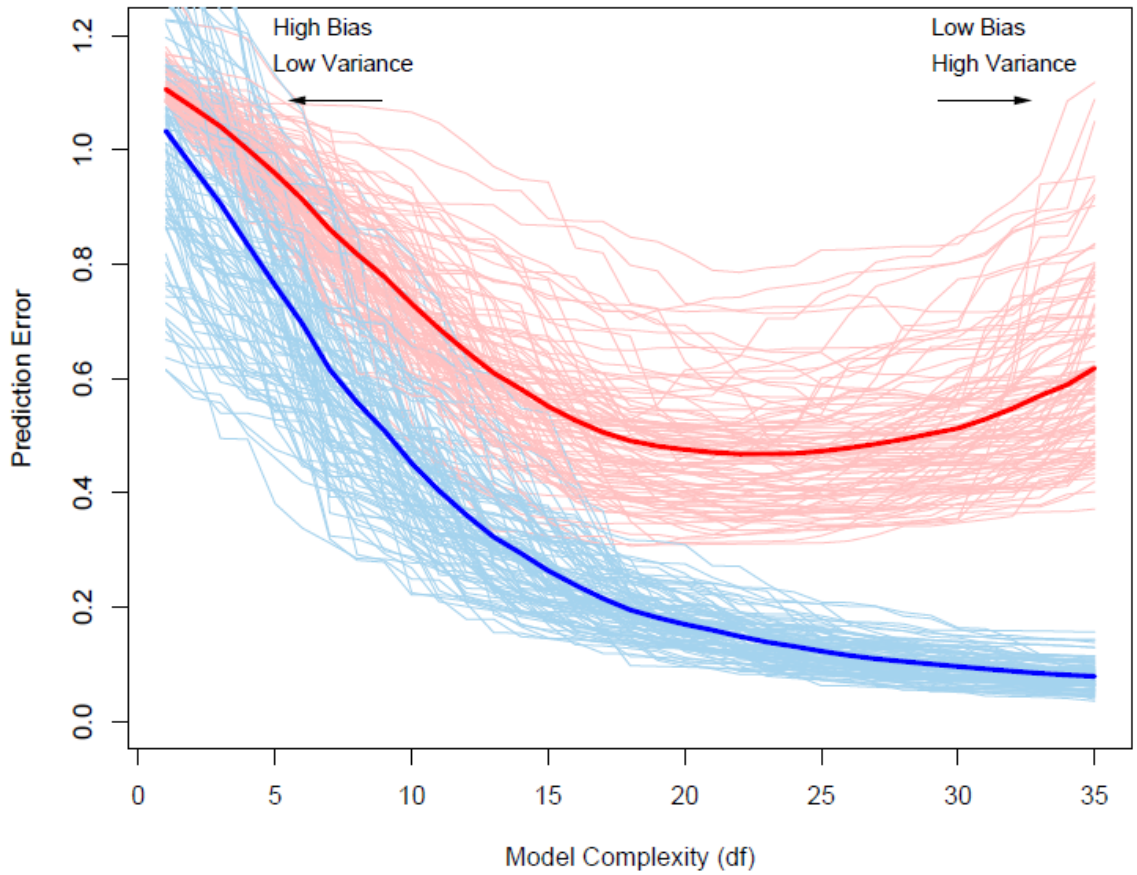
$$\mathsf{MSE}_{\mathsf{GCV}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1 - \mathsf{tr}(\boldsymbol{S})/N}\right]^2,$$

where $\mathsf{tr}(\boldsymbol{S})$ is the trace of $\boldsymbol{S}$ (the sum of the elements in its diagonal). GCV can be computationally convenient in some settings.

Optimism

Optimism

Our objective in this section is to develop a better understanding of overfitting. This discussion will inform our understanding of analytical criteria in the next section.



Optimism

We define the **training error** as the empirical loss for the training data,

$$\overline{\text{err}}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, \widehat{f}(\boldsymbol{x}_i)).$$

We focus on our standard regression setting,

$$\overline{\text{err}}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2,$$

which is RSS$/N$ for linear regression estimated by least squares.

Optimism

The **expected prediction error** (EPE) is

$$\begin{aligned}
\text{EPE}(\boldsymbol{x}_0) &= E_{\mathcal{D}}\left[\left(Y_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \\
&= E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_0) + \varepsilon_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \\
&= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0)\right)^2\right],
\end{aligned}$$

where $\boldsymbol{x}_0$ is fixed.

Optimism

Now, consider the estimation error for a training case i,

$$
\begin{aligned}
E_{\mathcal{D}}\left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2 &= E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) + \varepsilon_i - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] \\
&= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] - 2E_{\mathcal{D}}\left[\widehat{f}(\boldsymbol{x}_i)\varepsilon_i\right] \\
&= \sigma^2 + E_{\mathcal{D}}\left[\left(f(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] - 2\mathsf{Cov}(\widehat{f}(\boldsymbol{x}_i), \varepsilon_i)
\end{aligned}
$$

Unlike in the EPE, the last term appears because the estimator $\widehat{f}(\boldsymbol{x}_i)$ is a function of $\mathcal{D}$, which includes training case i itself.

Optimism

Averaging over the data, the expected value of the training error is

$$\begin{aligned} E\left[\text{err}_{\mathcal{D}}\right] &= \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left[\left(Y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2\right] \\ &= \sigma^2 + \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left(f(x_i) - \widehat{f}(\boldsymbol{x}_i)\right)^2 \\ &\quad - \frac{2}{N}\sum_{i=1}^{N} \text{Cov}(\widehat{f}(\boldsymbol{x}_i), \varepsilon_i). \end{aligned}$$

Because of the last term, the training error is not a good estimate of the expected prediction error.

Optimism

We define the out of sample error as

$$\begin{aligned}\overline{\mathsf{Err}}_{\mathsf{out}} &= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(Y_i^0 - \widehat{f}(x_i)\right)^2\right], \\ &= \sigma^2 + \frac{1}{N}\sum_{i=1}^{N} E_{\mathcal{D}}\left[\left(f(x_i) - \widehat{f}(x_i)\right)^2\right],\end{aligned}$$

where $Y_i^0 = f(\boldsymbol{x}_i) + \varepsilon_i^0$ indicates an independent case for a given $\boldsymbol{x}_i$.

Optimism

The **optimism** of the training error is

$$\overline{\mathsf{Err}}_{\mathsf{out}} - E(\overline{\mathsf{Err}}_{\mathcal{D}}) = \frac{2}{N} \sum_{i=1}^{N} \mathsf{Cov}(\widehat{f}(x_i), \varepsilon_i)$$

The more we overfit, the higher $\mathsf{Cov}(\widehat{f}(x_i), \varepsilon_i$ will be, increasing the optimism.

Example: linear regression

For the linear regression model, we can show that

$$\text{Optimism} = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}\left(\widehat{f}(x_i), \varepsilon_i\right) = \frac{2}{N}\sigma^2(p+1)$$

Interpretation:

- The larger the sample size (N), the harder it is to overfit.
- The larger the variance of the errors (σ^2), the larger the overfitting.
- The optimism is proportional to the number of predictors.

Analytical criteria

Analytical criteria

Analytic criteria provide estimate the generalisation error based on theoretical arguments. They have the form:

$$\textbf{criterion} = \textbf{training loss} + \textbf{penalty for number of parameters}$$

Mallow's C_p statistic

The **Mallow's** C_p statistic applies to linear regression. It directly implements the recipe suggested by our calculation of the optimism:

$$C_p = \frac{\text{RSS}}{N} + \frac{2}{N}\widehat{\sigma}^2(p+1),$$

In the formula, $\widehat{\sigma}^2$ is an estimate of variance of the errors based on the largest model under consideration.

Mallow's C_p statistic

We select the model with the lowest C_p. To compare two specifications,

$$\Delta C_p = \text{MSE}_1 - \text{MSE}_2 + \frac{2}{N}\widehat{\sigma}^2(p_1 - p_2).$$

Review: Log-Likelihood (key concept)

Let $y_1, \ldots, y_N$ be a sample for which each observation has density $p(y_i|\boldsymbol{x}_i; \theta)$, where θ is a $d-$dimensional parameter vector. The log-likelihood is

$$\begin{aligned} L(\theta) &= \log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i; \theta) \\ &= \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i; \theta) \end{aligned}$$

The maximum likelihood estimate $\widehat{\theta}$ is the value of θ that maximises the log-likelihood for the training data.

Akaike Information Criterion (key concept)

The **Akaike information criterion** (**AIC**) applies to models estimated by maximum likelihood.

$$\text{AIC} = -2L(\widehat{\theta}) + 2d,$$

where $L(\widehat{\theta})$ is the maximised log-likelihood and d is the number of estimated parameters. We select the model with the lowest AIC.

Akaike Information Criterion

- The AIC is one of the most popular and versatile strategies for model selection.
- The formula follows the in-sample performance plus penalty for complexity structure.
- The AIC has a rigorous theoretical justification which we not address here. However, keep in mind that it is an asymptotic approximation ($N \to \infty$).

AIC for linear regression

In the special case of comparing linear regression specifications under Gaussian errors, the AIC simplifies to (up to proportionality and ignoring constant terms in the log-likelihood):

$$\text{AIC} = \log(\text{RSS}/N) + 2\ \frac{(p+2)}{N},$$

The number of parameters is $d = p + 2$ because the parameter vector includes the constant and the variance of errors. Note that this is different from the formula in the book, which is a simplification.

Relation between Mallow's C_p and the AIC

For a linear regression with Gaussian errors and known variance:

$$\text{AIC} = \text{constant} + \frac{1}{\sigma^2}\left(\text{RSS}/N + 2\sigma^2\frac{(p+1)}{N}\right),$$

which compares to

$$C_p = \text{RSS}/N + \frac{2}{N}\sigma^2(p+1).$$

Hence, the AIC and C_p lead to the same decision in this case. For practical purposes, the AIC and C_p are regarded as the same for linear regression.

Bayesian information criterion

The **Bayesian information criterion** (**BIC**) also applies to models estimated by maximum likelihood.

$$\text{BIC} = -2L(\widehat{\theta}) + \log(N)d$$

We select the model with the lowest BIC.

Bayesian information criterion

- The BIC formula is comparable to the AIC case, but with 2 penalty factor replaced by $\log(N)$. Hence, the BIC penalises complexity more heavily when $N \geq 8$. The BIC has a very different theoretical justification to the AIC.

- The BIC is an asymptotic approximation to a Bayesian approach to model selection.

BIC: Gaussian linear regression case

In the special case of a linear regression with Gaussian errors, the BIC simplifies to (ignoring constant terms)

$$\text{BIC} = N \log(\text{RSS}/N) + \log(N)(p+2)$$

If we assume that the variance of the errors is known, we have instead

$$\text{BIC} = \frac{N}{\sigma^2}\left[\text{RSS}/N + \log(N)\frac{(p+1)}{N}\right]$$

In this case the BIC is proportional to to AIC and C_p, but with a $\log(N)$ penalty factor instead of 2.

Comparison of model selection methods

Model selection properties

Consistency. In a collection of models that includes the correct model, the probability that the model selection criterion chooses the correct one approaches one when $N \to \infty$.

Efficiency. The selected model predicts as well as the theoretically best model under consideration in terms of expected loss when $N \to \infty$.

It is not possible to combine these properties (Claeskens and Hjort, 2008, Section 4.9).

Properties of model selection methods

LOOCV, AIC, and Mallow's C_p. Efficient but not consistent. The efficiency follows because they construct unbiased estimators of the test error. However, they select models that are strictly more complex than the true model when $N \to \infty$.

BIC. Consistent under some conditions but not efficient. It often chooses models that are too simple because of its heavier penalty on complexity.

LOOCV, AIC and C_p

- LOOCV, AIC, and C_p are equivalent when $N \to \infty$. They will pick the same model in practice when N is large.
- In finite samples, we can view AIC and C_p as theoretical approximations to LOOCV.
- The advantage of AIC and C_p over LOOCV is mainly computational. CV should be preferred to AIC when the assumptions of the model (e.g., constant error variance) are likely to be wrong.
- LOOCV is universally applicable, while this is not the case for AIC and C_p.

Application

Equity premium prediction data

Quarterly data from Goyal and Welch (2008), updated to 2015.

```
Response: quarterly S&P 500 returns minus treasury bill rate

Predictors (lagged by one quarter):
1. dp         Dividend to price ratio
2. dy         Dividend yield
3. ep         Earnings per share
4. bm         Book-to-market ratio
5. ntis       Net equity expansion
6. tbl        Treasury bill rate
7. ltr        Long term rate of return on US bods
8. tms        Term spread
9. dfy        Default yield spread
10.dfr        Default return spread
11.infl       Inflation
12.ik         Investment to capital ratio

Number of observations: 275 (1947-2015)
```

Complete subset regressions

- Suppose that we want to use a linear regression to predict the equity premium. One option is to include all the $p = 12$ available predictors and estimate the model by OLS. However, the data are very noisy and this will lead to overfitting.

- The complete subset regressions (CSR) method is a simple and easy to understand algorithm that we can use to reduce overfitting.

- The CSR method predicts the response by taking a simple average of the predictions produced by linear regressions for all $\binom{p}{k}$ possible combinations of $k < p$ predictors.

Complete subset regressions

Algorithm Complete subset regressions

1: Set k.

2: Generate all the $S = \binom{p}{k}$ possible predictor subsets of size k.

3: **for** subset in subsets **do**

4: Construct $N \times (k+1)$ design matrix $\boldsymbol{X}_{\text{subset}}$ based on the predictors included in the subset.

5: Estimate the model by OLS based on $\boldsymbol{X}_{\text{subset}}$ and denote the coefficients as $\widehat{\boldsymbol{\beta}}_{\text{subset}}$.

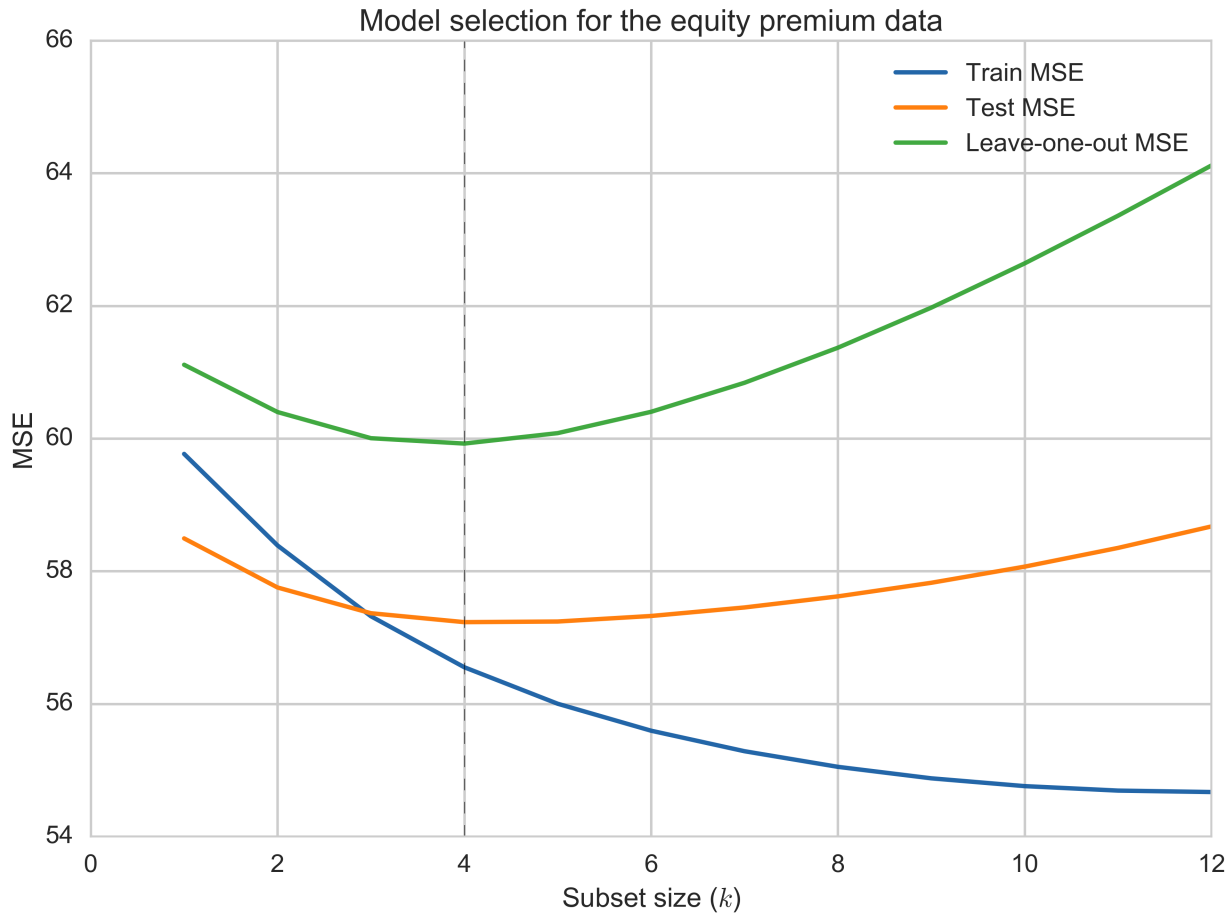
6: **end for**

7: The prediction for a new input vector $\boldsymbol{x}_0$ is

$$\widehat{y}_0 = (1/S) \sum_{\text{subsets}} \boldsymbol{x}_{0\text{subset}}^T \widehat{\beta}_{\text{subset}}$$

Equity premium prediction

We use leave-one-out cross validation to select the optimal model complexity (k) for the CSR method.



Equity premium prediction

- The optimal subset size according to leave-one-out cross validation is $k = 4$. That corresponds to the value of k with the lowest test MSE. The AIC also leads to the choice of $k = 4$.

- This data is characterised by a low signal-to-noise ratio. The evidence in the literature shows that the predictably of the equity premium is low.

- Using all the inputs leads to very poor predictions: the R^2 score is only 0.014 in this configuration (compared to 0.04 for CSRs and a training R^2 of 0.10).

Limitations of model selection

Limitations of model selection

- Standard statistical inference is no longer valid after model selection.
- The reason is that standard inference assumes a fixed model, whereas model selection will by definition pick the specific model that best fits the sample. This will lead to optimistic estimates of sample variation based on the chosen model.
- In our context, the only way around this difficulty would be data splitting: using one part of the sample for model selection, and another for inference.

Limitations of model selection

Model selection is an important tool in your data analysis process, but should not be a replacement to model building through EDA, diagnostics, and domain knowledge.

Review questions

- How does model selection relate to the bias-variance trade-off?
- What is a validation set? How is it different from a test set?
- What is K-Fold cross validation? Describe how it works.
- What is the one standard deviation rule?
- What is the Akaike Information Criterion?
- Why is it incorrect to conduct statistical inference after model selection (using the same data)?