Cross-validation

Outline

- The validation set approach
- Leave one out cross validation
- K-fold cross validation
- CV for model selection vs model evaluation
- Cross validation for time series data

Resampling

- Repeatedly and randomly drawing subsets of data from a sample
- Refitting a model (e.g. OLS regression) on these subsets of data reveal information unknown if fitted the model only once, e.g. variability of the fitted model
- Computationally expensive
- Two common resampling methods
    - Cross-validation: model selection and model evaluation
    - bootstrapping: evaluating the variability of an estimate

Cross-validation

- Training error - a measure of the goodness of fit of a model (e.g. OLS regression) to the data used to train the model
- Test error - a measure of how accurate the model is in predicting the response of a new observation
- Mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

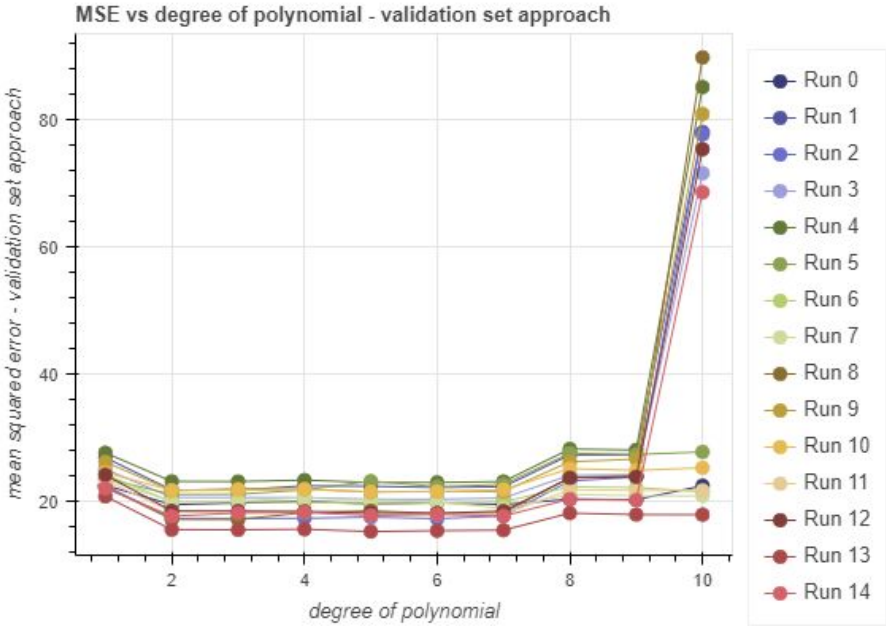
- Cross-validation - a cost effective method that allows for the test error be estimated **independently** of the training error rate.

The validation set approach

- Randomly split the original data into two, a *training set* and a *test set*.
- Fit the OLS model on the training set and predict the responses in the validation set
- Calculate the test MSE (MSE from using the model on the test set)

The validation set approach

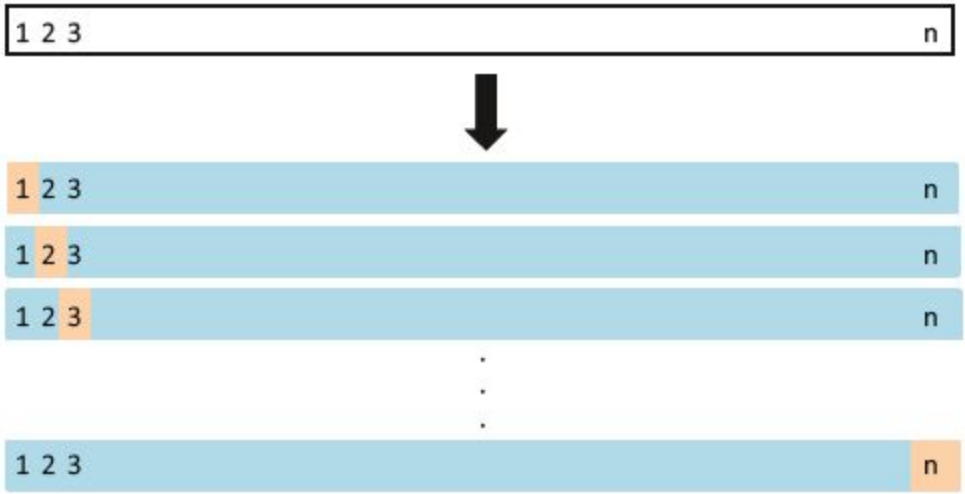
- Pros: conceptually simple and easy to implement
- Cons:
  - Highly variable (in multiple runs)
  - Tend to *overestimate* the test error (because we used only roughly half of the original dataset for training)



Leave one out cross validation (LOOCV)

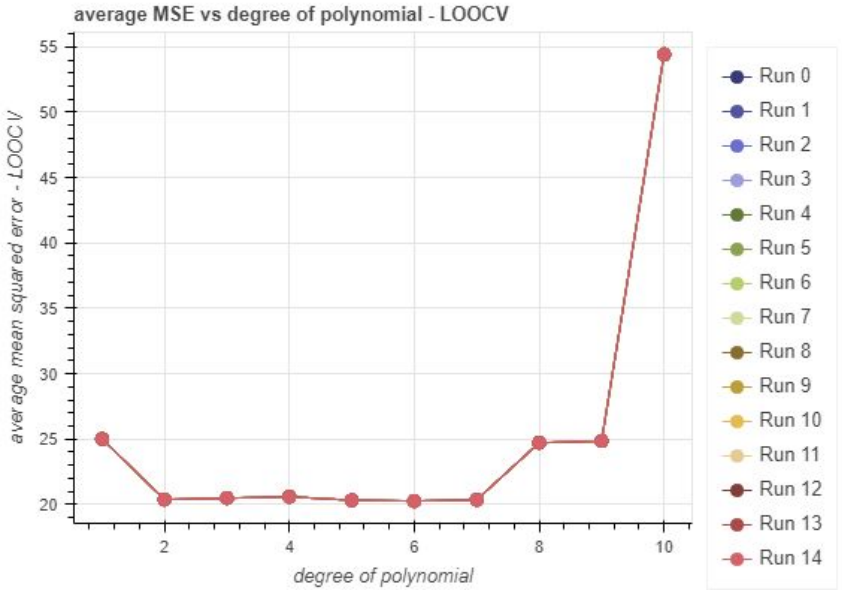
- Use only 1 observation for testing, and fit the OLS regression on the remaining of the original data
- Repeat the procedure n times so that each observation is used for testing once.
- Calculate the test MSE

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$



Leave one out cross validation (LOOCV)

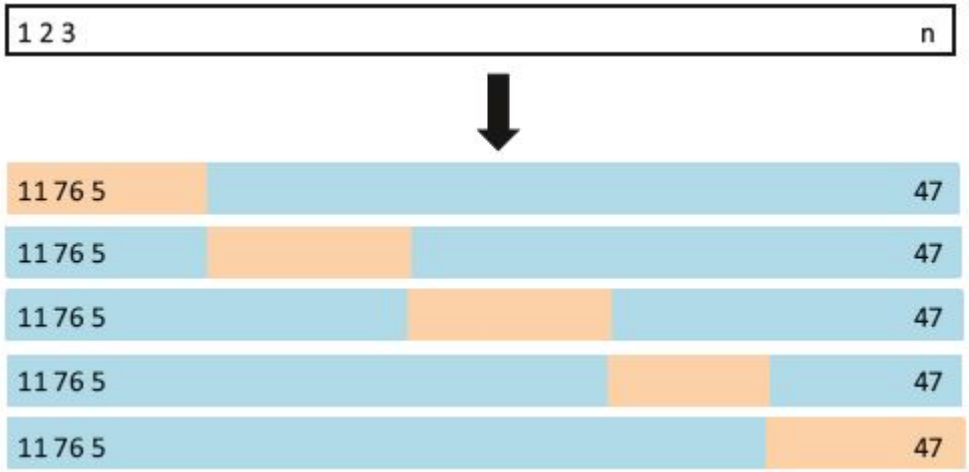
- Pros:
  - Unbiased estimate of the test error (because we used almost all data points for training)
  - Very stable (identical test MSE from multiple runs)!
- Cons:
  - Very time consuming (especially when n is large and/or fitting a complex model)



K-fold cross validation

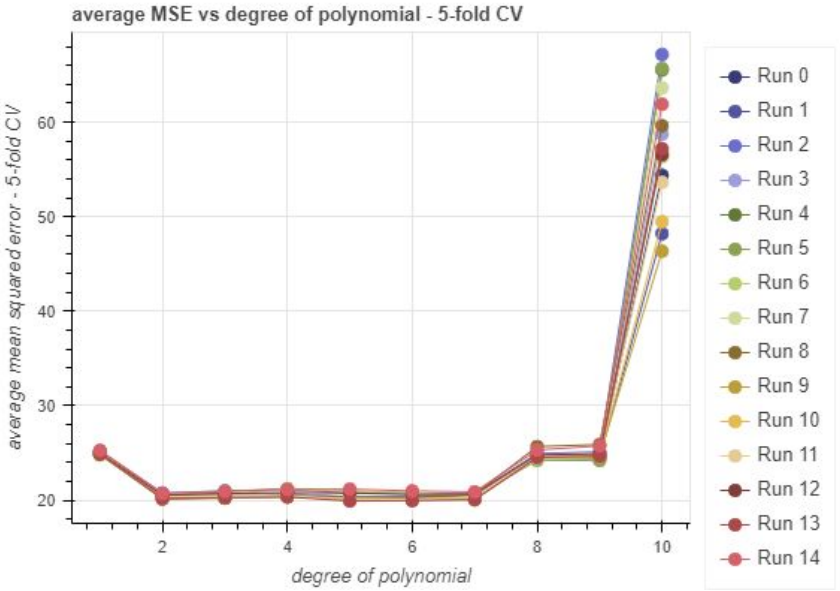
- Randomly divide the original data into k groups (folds)
- Train the model on k-1 folds, use the last fold for testing
- Repeat the procedure k times, so that each fold is used for testing once
- (Repeat the above 3 steps multiple times)
- Calculate the test MSE

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

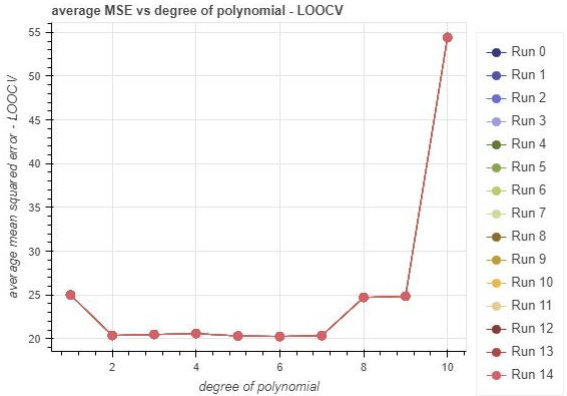


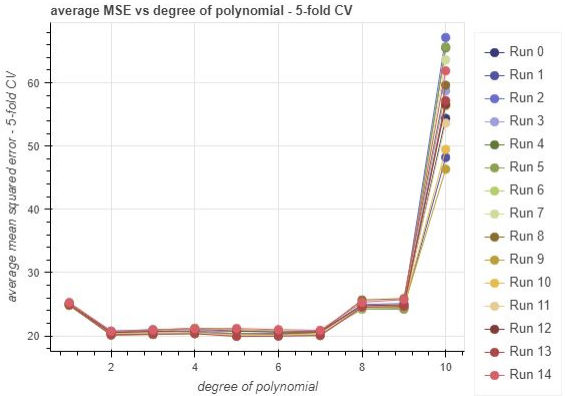
K-fold cross validation

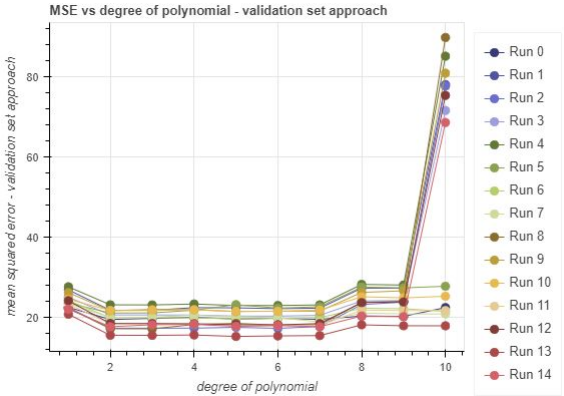
- Pros:
  - Less computationally demanding than LOOCV
- Cons
  - Test error tend to be more biased compared to LOOCV (but much less so compared to validation set approach)



CV for model evaluation vs model selection







Model selection chooses the model with smallest test MSE => LOOCV (being the most stable) allows for the unambiguous choice.

CV for model evaluation vs model selection

Model evaluation estimates the expected range of error in real life applications

- LOOCV: least biased error estimate but with largest variance
- Leave one out: most biased error estimate but with smallest variance (only 1 value of the test error)
- K-fold CV: a balance between bias and variance of the error estimate

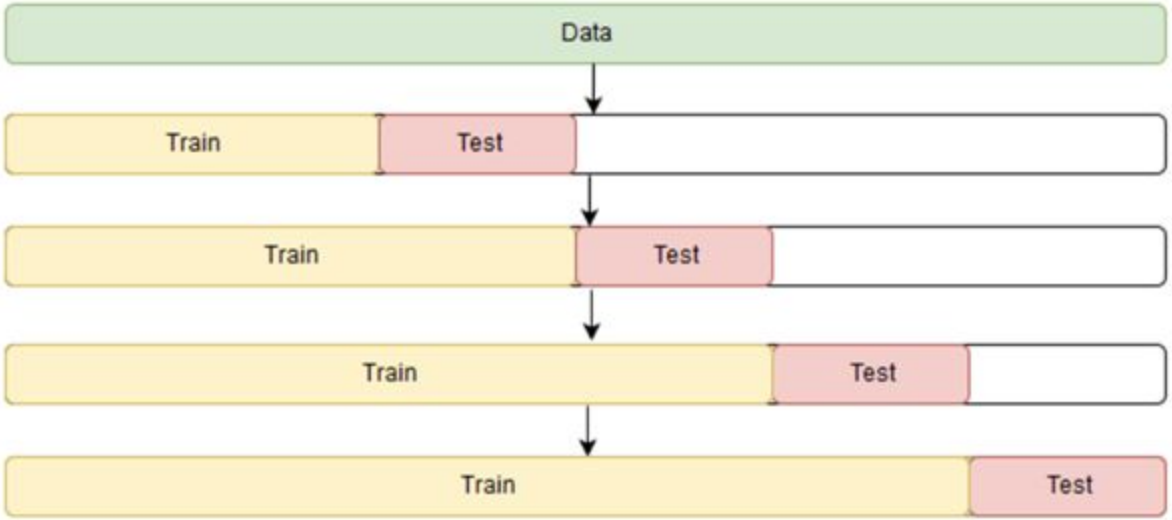
```
MAD distributions
2-fold: average MAD 15.288, stdev MAD 0.896
5-fold: average MAD 15.028, stdev MAD 1.025
10-fold: average MAD 15.085, stdev MAD 1.219
LOO: average MAD 15.079, stdev MAD 12.040

Applying the empirical rule:
2-fold: CV 95pc confidence interval of MAD is [13.497, 17.079]
5-fold: CV 95pc confidence interval of MAD is [12.978, 17.079]
10-fold: CV 95pc confidence interval of MAD is [12.647, 17.523]
LOOCV: 95pc confidence interval of MAD is [-9.001, 39.160]
```

Test error estimate for the regression model mpg ~ horsepower + horsepower^2 by different CV strategies

Cross validation for time series data

Must preserve the chronological order of the data...



https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4
https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

Exercises

- Estimate test MAD for the OLS regression model Sales ~ TV + Radio using different CV strategies