Machine Learning and Data Analytics
ME 5013- Fall 2019

Lectures 12

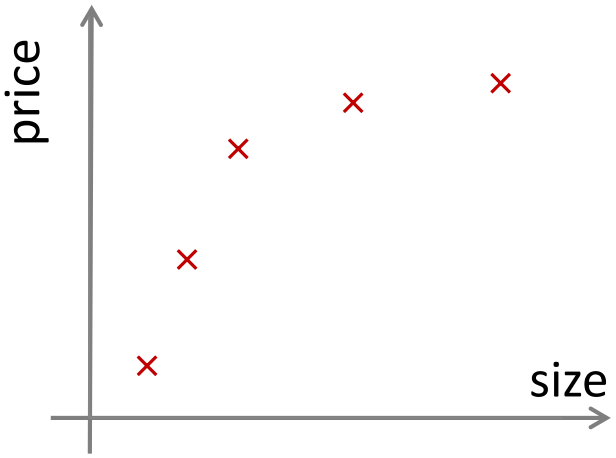
- Model Selection

Adel Alaeddini, PhD
Associate Professor of Mechanical Engineering
Advanced Data Engineering Lab
adel.alaeddini@utsa.edu

Disclosure: the slides are adopted from various online resources

Evaluating your hypothesis

If/when not trained correctly, machine learning models may fails to generalize to new examples not in training set.



$$h_\beta(x) = \beta_0 + \beta_1 x + \cdots + \beta_1 x^4$$

$x_1 =$ size of house
$x_2 =$ no. of bedrooms
$x_3 =$ no. of floors
$x_4 =$ age of house
$x_5 =$ average income in neighborhood
$x_6 =$ kitchen size
$\vdots$
x_{100}

Different classes of model selection

- Subset Selection. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model, i.e. using least squares, on the reduced set of variables.
- Shrinkage. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
- Dimension Reduction. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a model.

Subset Selection: Best Subset Selection

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For k = 1,2,…,p:
   a. Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   b. Pick the best among these $\binom{p}{k}$ models , and call it M_k. Here best can be defined as having the smallest RSS, largest R^2.
3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error, Cp (AIC), BIC, or adjusted R^2.

Note: We often can't examine all possible models, since they are 2^p of them; for example when p = 40 there are over a billion models!

Subset Selection: Forward Selection

- Begin with the *null model* — a model that contains an intercept but no predictors.
- Fit *p* simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

- *Note: It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.*

Subset Selection: Backward Selection

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Note: It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

RSS and R^2 potential issues

- The model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.
- We wish to choose a model with low test error, not a model with low training error. Recall that training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

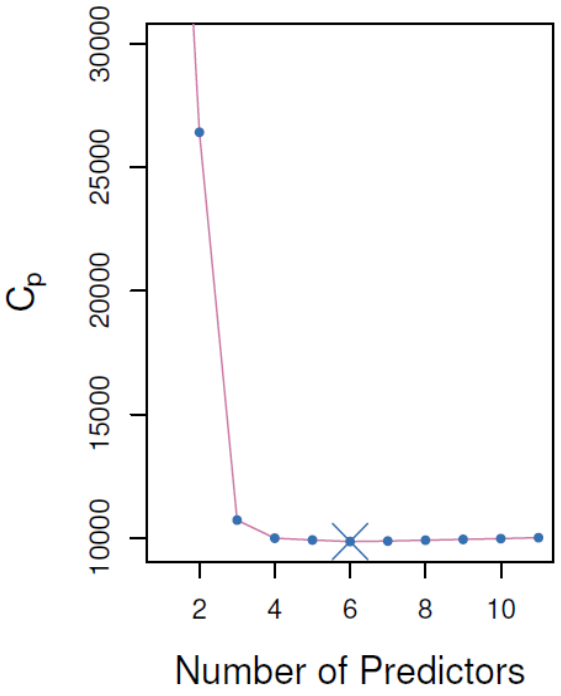
Estimating test error: two approaches

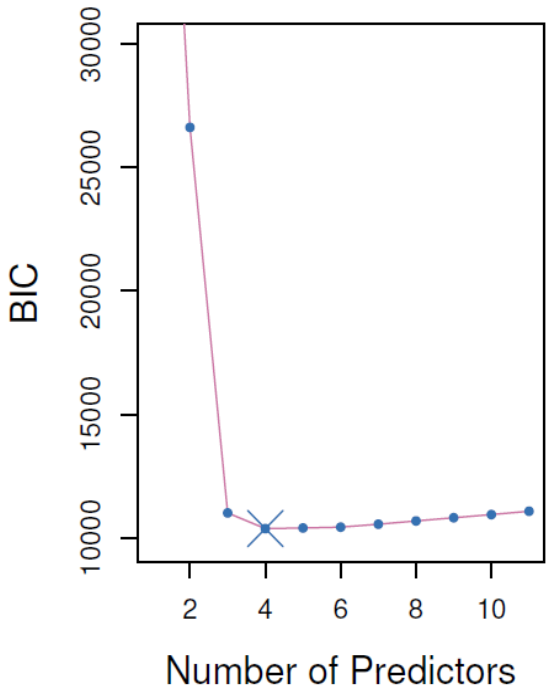
- We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
- We can directly estimate the test error, using either a validation set approach or a cross-validation approach

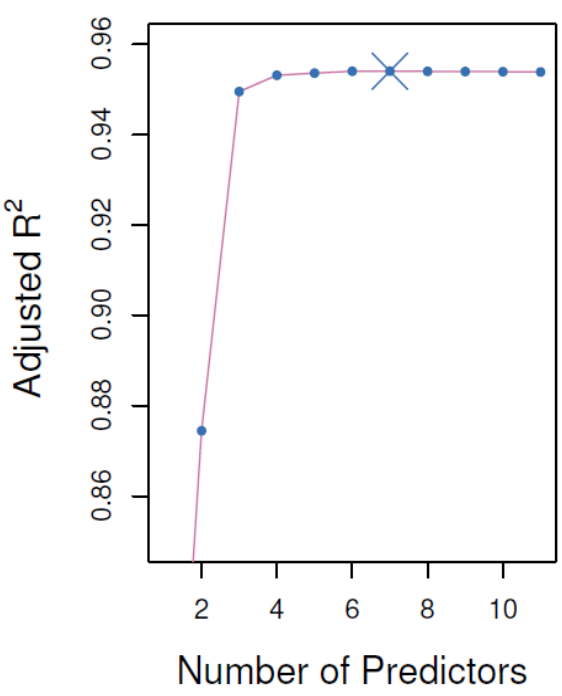
Subset Selection: Cp, AIC, BIC, and Adjusted R^2

- These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
- Mallow's Cp: $$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$
  where d is the total # of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error associated with each response measurement.
- AIC: $$\mathrm{AIC} = -2\log L + 2 \cdot d$$
  Where L is the maximized value of the likelihood function for the estimated model.
- BIC: $$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$
- Adjusted R^2 $$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$

Example: Credit data

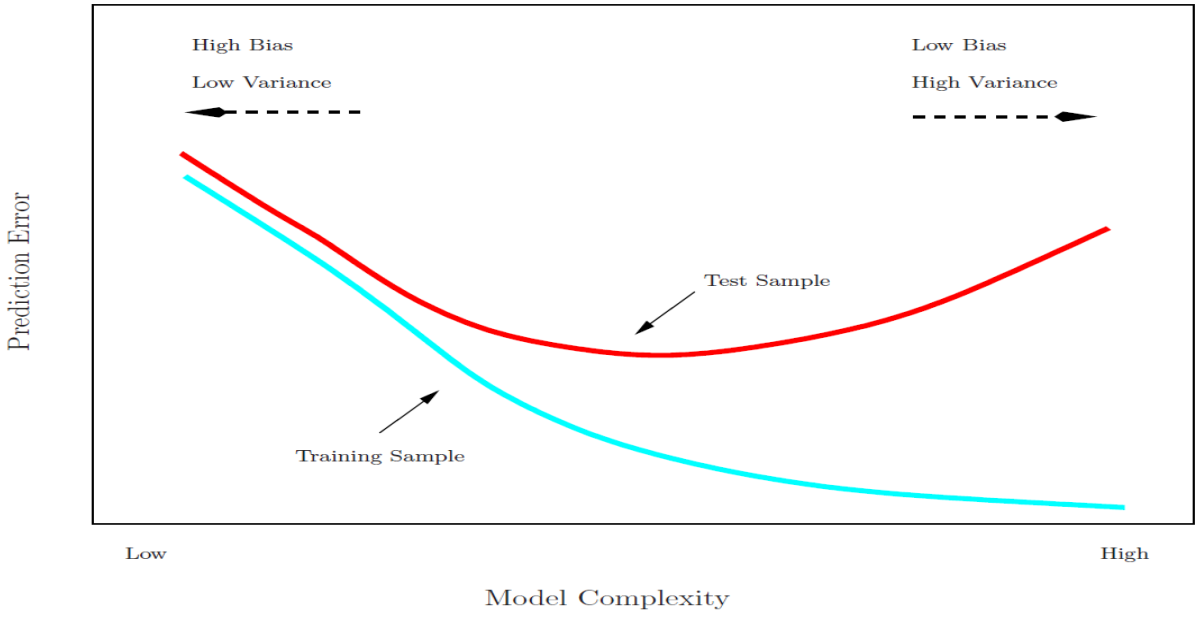






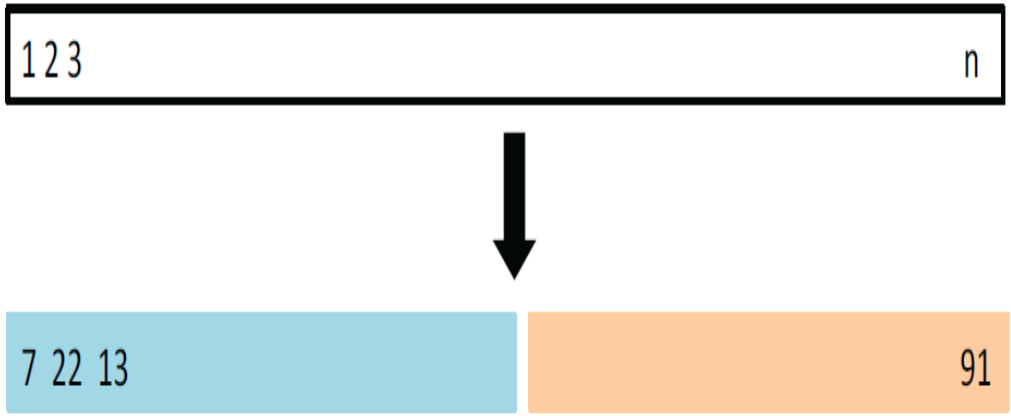
Training Error vs Test Error

- The test error is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the training error can be easily calculated by applying the machine learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.



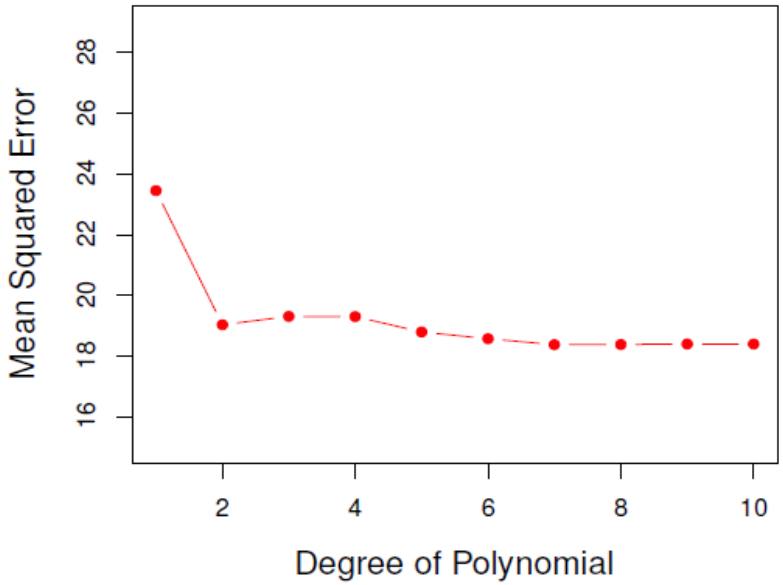
Validation-set approach

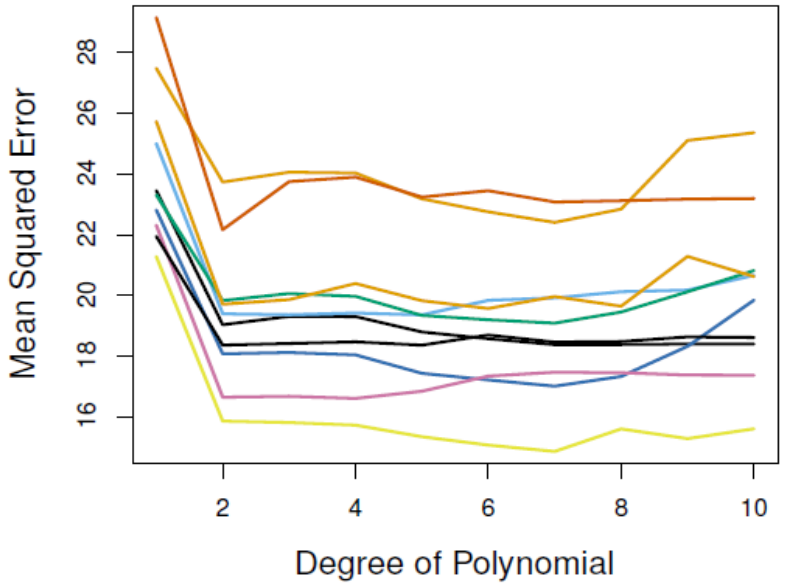
- Randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassication rate in the case of a qualitative (discrete) response



Example: automobile data

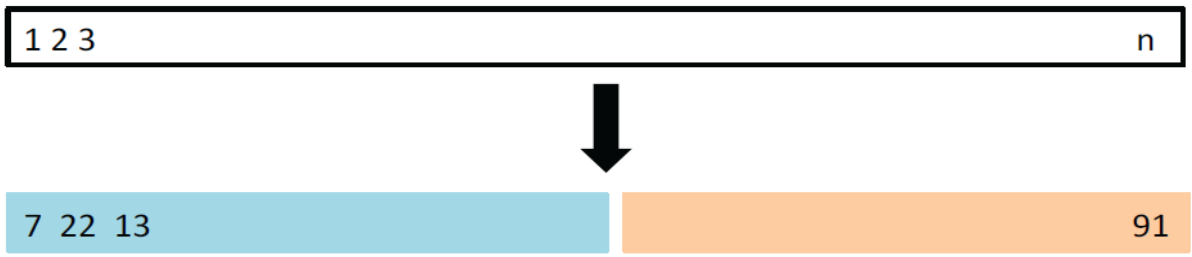
- Want to compare linear vs higher-order polynomial terms in a linear regression
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.





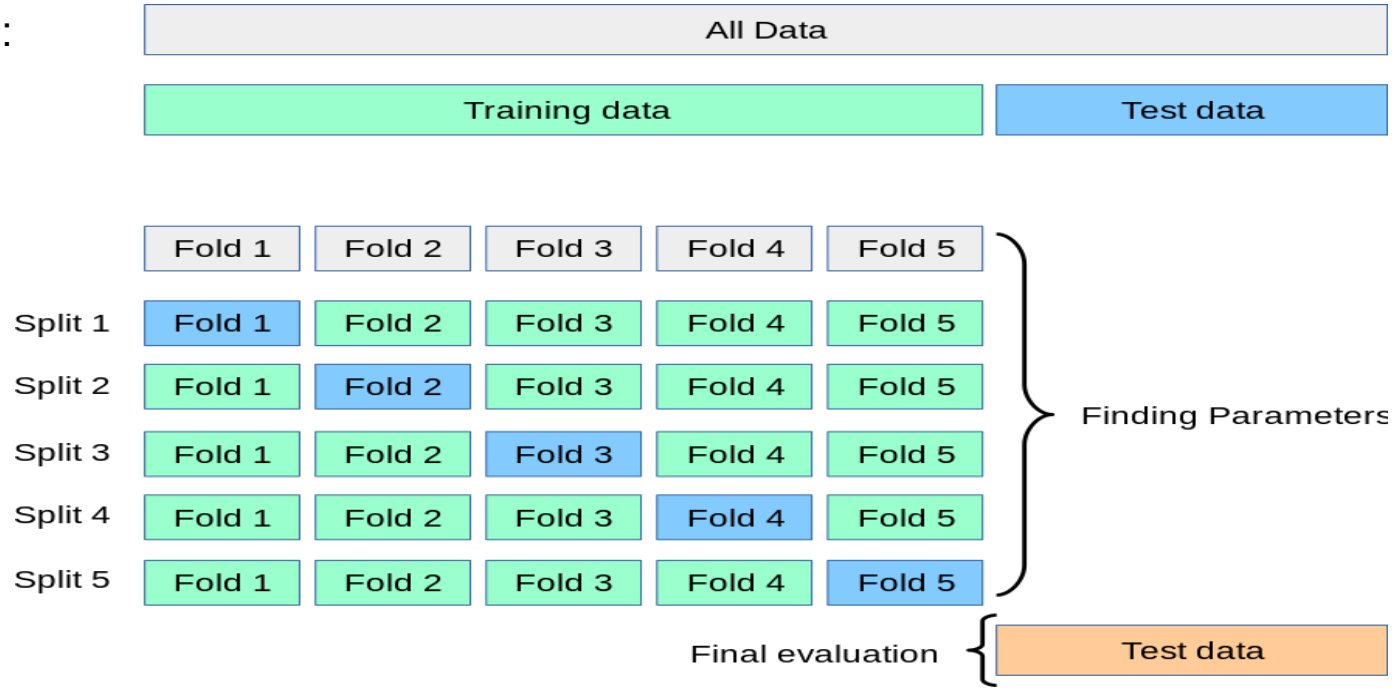
Drawbacks of Validation-set Approach

- the validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations - those that are included in the training set rather than in the validation set - are used to fit the model.
- This suggests that the validation set error may tend to overestimate the test error for the model fit on the entire data set.

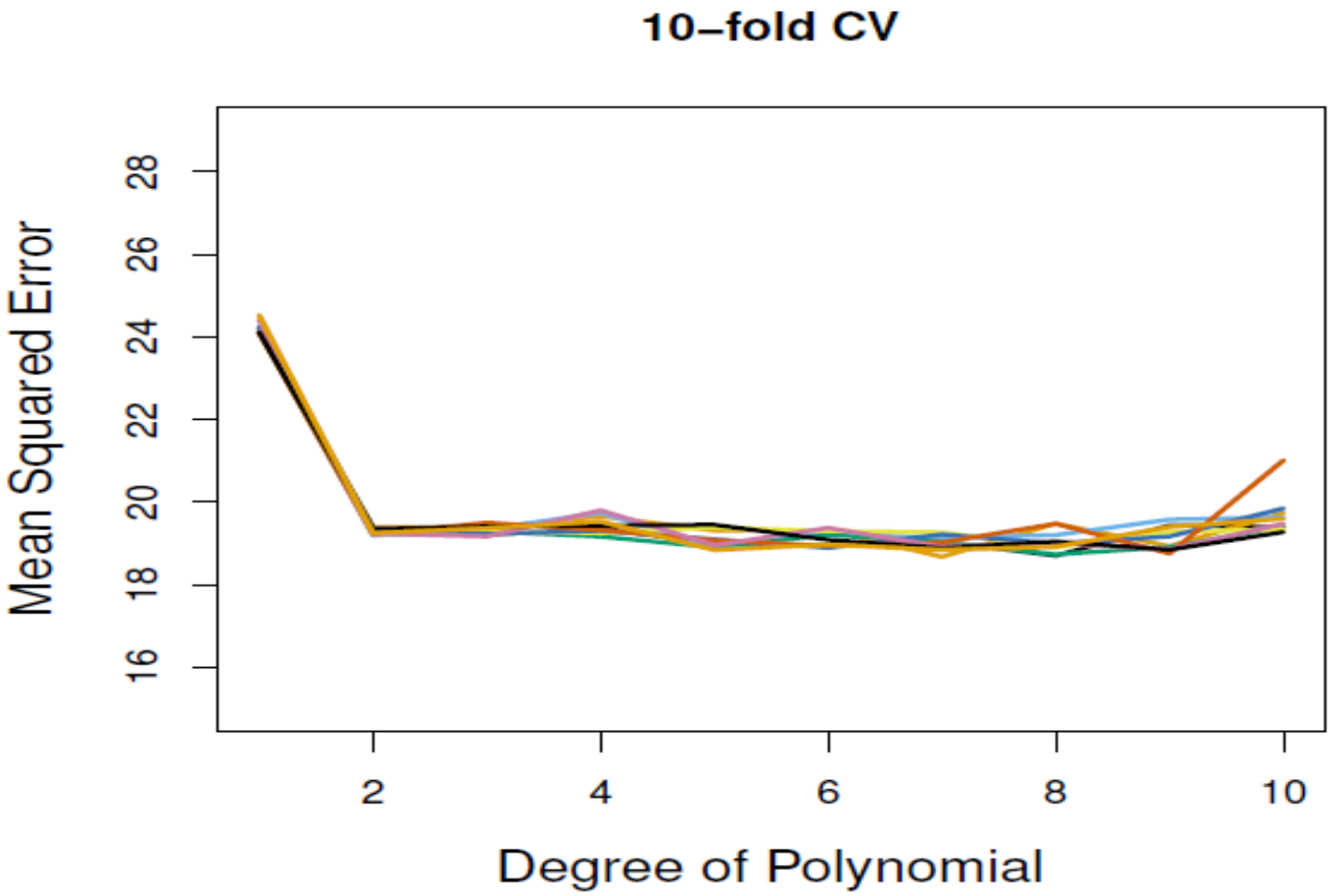


K-fold Cross-validation

- Widely used approach for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1,2,\ldots,K$, and then the results are combined.
- Visualization:



Example: Auto Data



Cross-Validation: Details

Regression

- Let the K parts be $C_1, \ldots, C_k$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k (if N is a multiple of K, then $n_k = \frac{n}{K}$)
- Compute $CV_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$

Where $MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$ and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.

Classification

- We divide the data into K parts be $C_1, \ldots, C_k$, where C_k denotes the indices of the observations in part k. There are nk observations in part k (if N is a multiple of K, then$n_k = \frac{n}{K}$)
- Compute $CV_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} Err_k$

where $Err_k = \sum_{i \in C_k} \frac{\boldsymbol{I}(y_i \neq \hat{y}_i)}{n_k}$

Alternative formulation for $CV_{(k)}$:

$$J_{test}(\beta) = -\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} y_{test}^{(i)} \log h_\beta(x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \log h_\beta(x_{test}^{(i)})$$

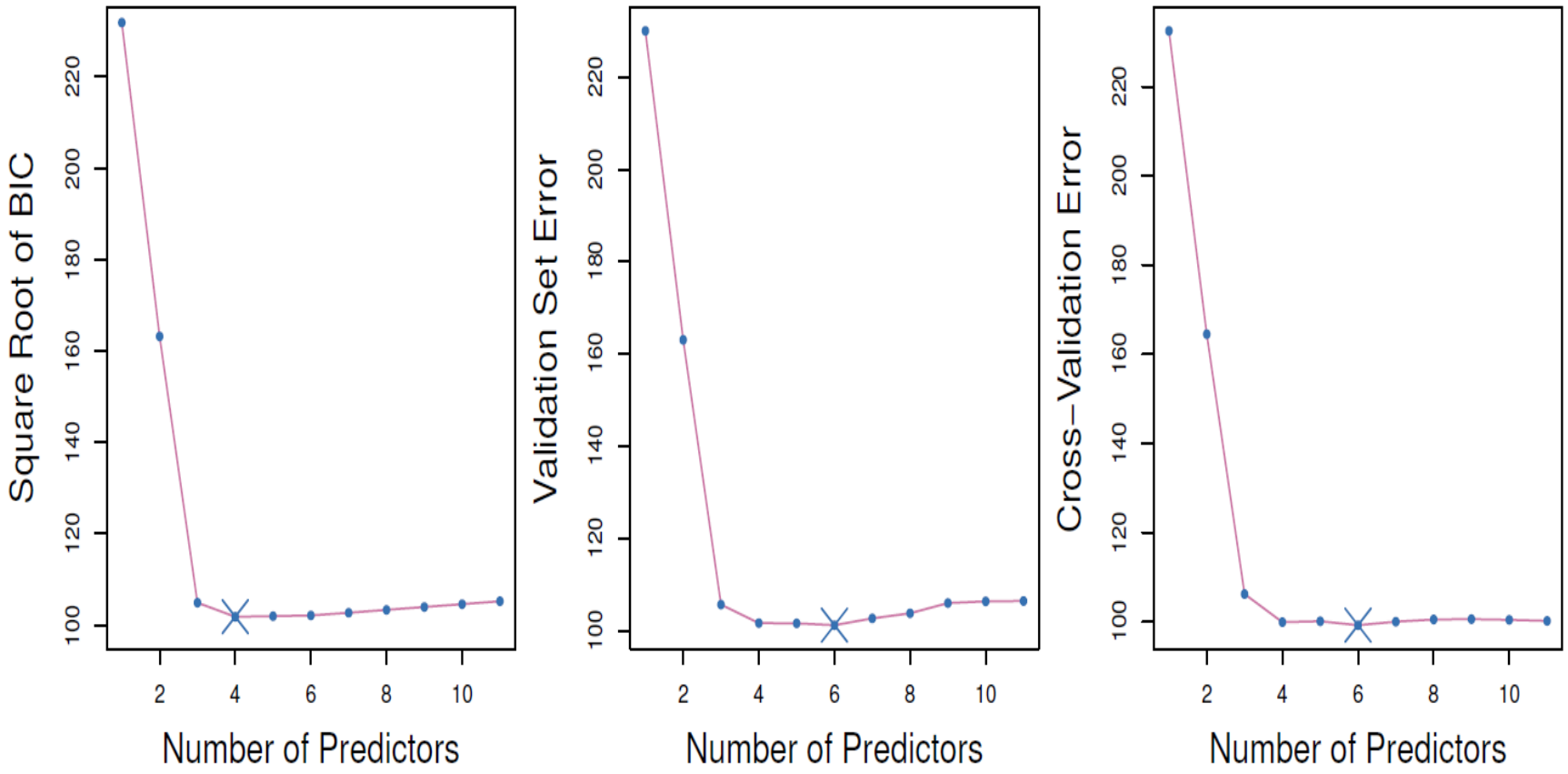
Note:

- Setting K = n yields n-fold or leave-one out cross-validation (LOOCV).
- K = 5 or 10 provides a good compromise for this bias-variance tradeoff.

Summary: Validation and Cross-Validation

- Each of the procedures returns a sequence of models M_k indexed by model size k = 0,1,… Our job here is to select $\hat{k}$. Once selected, we will return model $M_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, Cp, and adjusted R2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

Example: Credit data

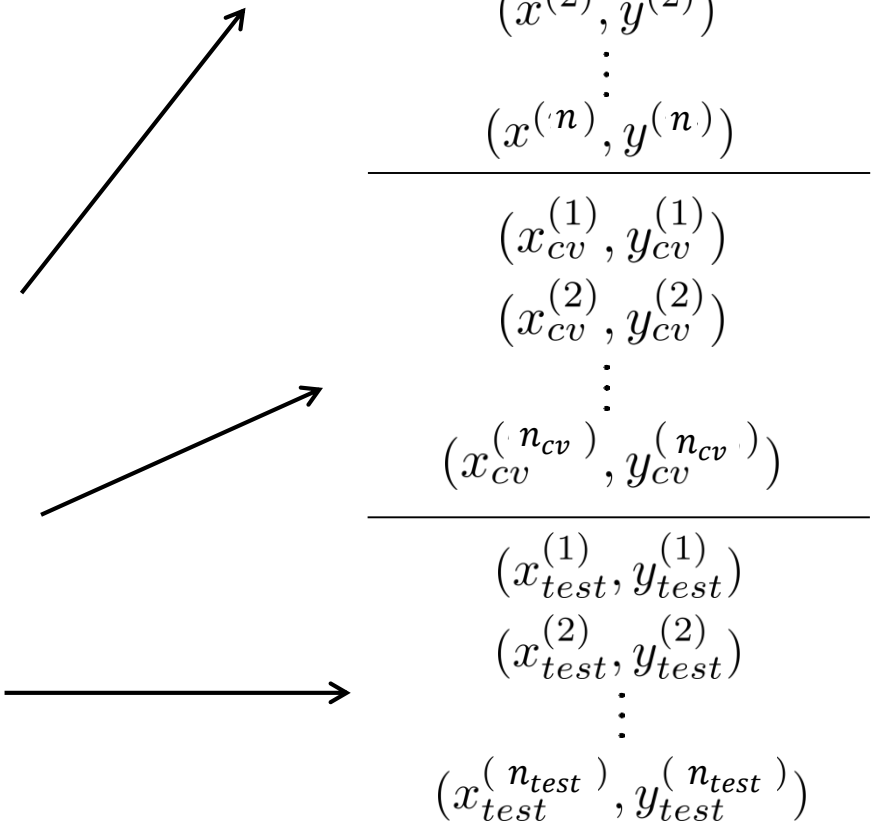


Train/validation/test error

Evaluating your hypothesis

Dataset:

Size	Price
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243



Train/validation/test error

Training error:

$$J_{train}(\beta) = \frac{1}{2\,n}\sum_{i=1}^{n}(h_\beta(x^{(i)}) - y^{(i)})^2$$

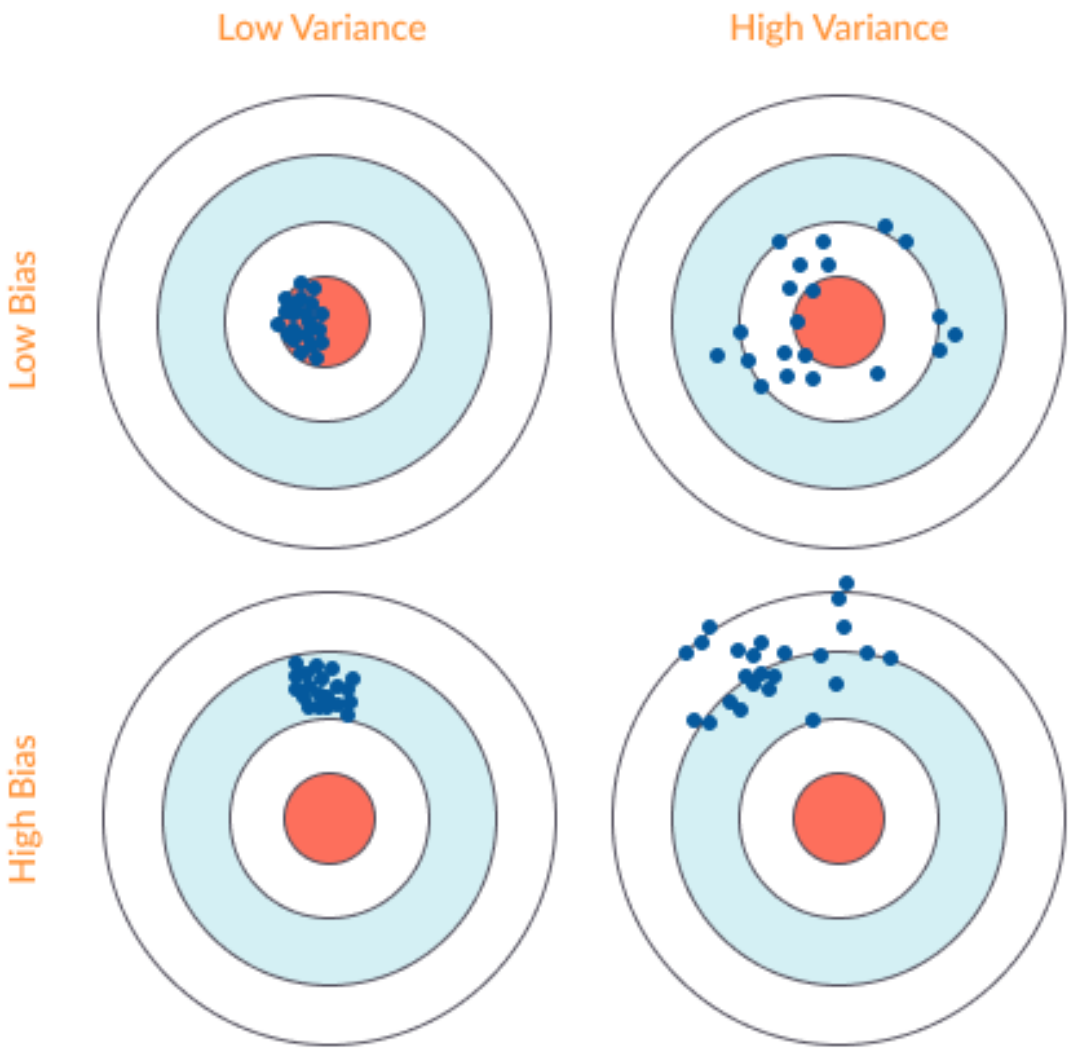
Cross Validation error:

$$J_{cv}(\beta) = \frac{1}{2\;n_{cv}}\sum_{i=1}^{n_{cv}}(h_\beta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

Test error:

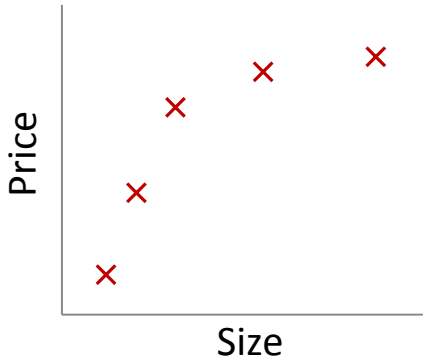
$$J_{test}(\beta) = \frac{1}{2\;n_{test}}\sum_{i=1}^{n_{test}}(h_\beta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

Bias vs. Variance



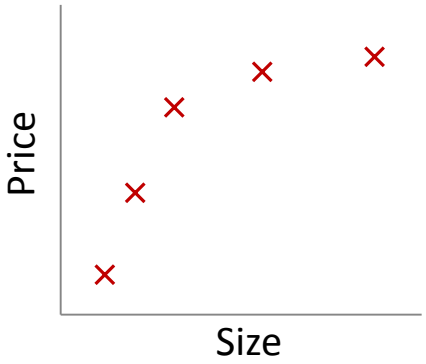
Bias vs. Variance

Bias/variance



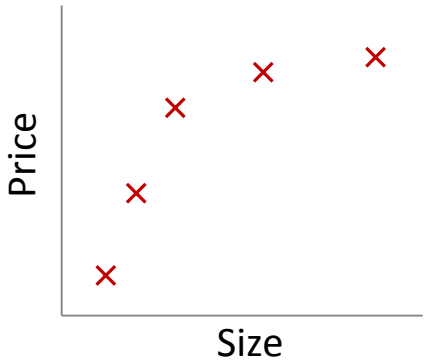
$\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

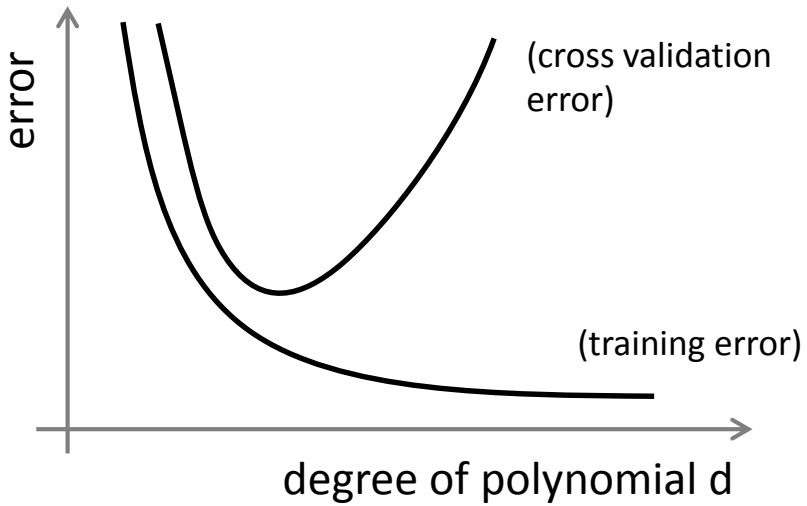


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

Bias vs. Variance

Suppose your learning algorithm is performing less well than you were hoping. Is it a bias problem or a variance problem?



Bias (underfit):

Variance (overfit):