5.1 Cross-Validation

Dr. Lauren Perry

Test and Training Data

- In practice, we typically want to use all of our data to build our models.
  - But if we do this directly, we only capture the *training error rate*, which may be inflated due to overfitting.
- Cross-validation gives us an option for using all of our data to train the models *and* getting a test error rate!

We will discuss the methods in this section using linear regression as an example. However, they can be used on any of the models discussed in this book.

The Validation Set Approach

This is how we've been working with test and training data.

- Randomly divide the data into parts: training and test (validation).
- Build the model on the training set.
- Examine model fit using the test set.

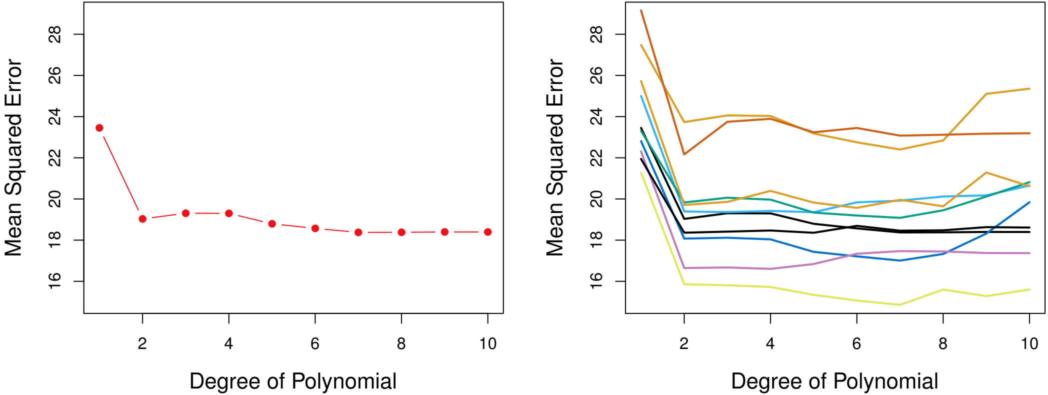
The Validation Set Approach: Auto Data

```r
train.ind <- sample(1:nrow(Auto), floor(0.8*nrow(Auto)))
auto.train <- Auto[train.ind, ]
auto.test  <- Auto[-train.ind, ]
mod1 <- lm(mpg ~ poly(horsepower, 3), data=auto.train)

preds <- predict(mod1, auto.test)
mean((preds - auto.test$mpg)^2)
```

```
## [1] 16.84336
```

The Validation Set Approach: Auto Data



The Validation Set Approach

Problems:

1. Estimates of error rate can be highly variable depending on which observations are included in the training set.
2. Only a subset of observations are used to train the model, which may result in worse estimates.

Leave-One-Out Cross-Validation (LOOCV)

Cross-validation is similar to the validation set approach, but it addresses the problems mentioned in the previous slide.

Idea:

- Hold one observation (x_1, y_1) back and train the model on the remaining $n - 1$ observations.
- Make a prediction $\hat{y}_1$ for the excluded observation and find its mean squared error

$$\text{MSE}_1 = (y_1 - \hat{y}_1)^2$$

- Repeat the procedure n times in total, holding back a different observation each time.

Leave-One-Out Cross-Validation (LOOCV)

- Then the LOOCV estimate for the test MSE is

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

- For least squares linear or polynomial regression, we can find $\mathrm{CV}_{(n)}$ without having to do extra work:

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

  - Here, h_i is the leverage.
  - Residuals for high-leverage points are inflated in such a way that this formula holds.

Leave-One-Out Cross-Validation (LOOCV)

Advantages:

- Far less bias than validation set approach.
  - Less likely to overestimate test error rate.
- No variability from randomly selecting training data.

Disadvantages:

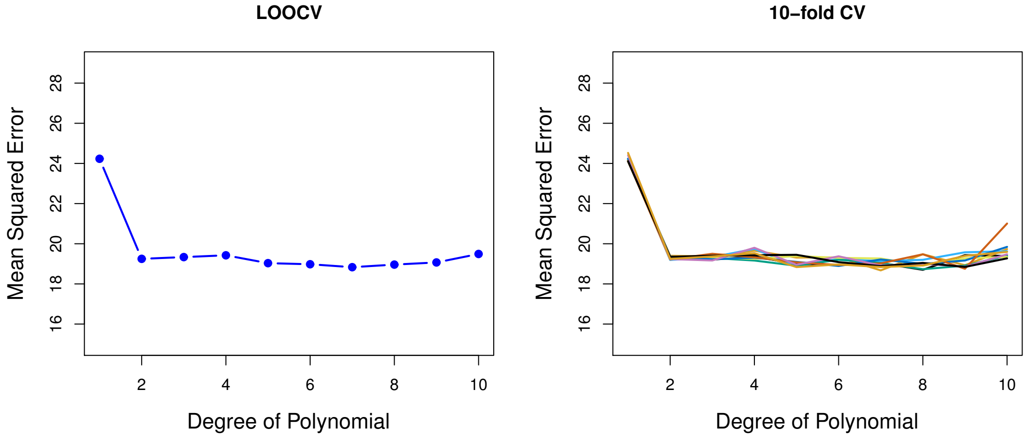
- Can be very expensive to implement.
  - Model needs to be fit n times.

LOOCV: Auto Data

```r
library(boot)
mod2 <- glm(mpg ~ poly(horsepower, 3), data=Auto)
cv.err <- cv.glm(Auto, mod2)
cv.err$delta
```

```
## [1] 19.33498 19.33448
```

LOOCV: Auto Data



k-Fold Cross-Validation

This is an alternative to LOOCV.

- The data is divided into k groups, or *folds*, of approximately equal size.
- The first fold is treated as a test (validation) set and the method is fit on the remaining $k-1$ folds.
- The MSE is computed on the observations in the held-out fold.
- Repeat k times.

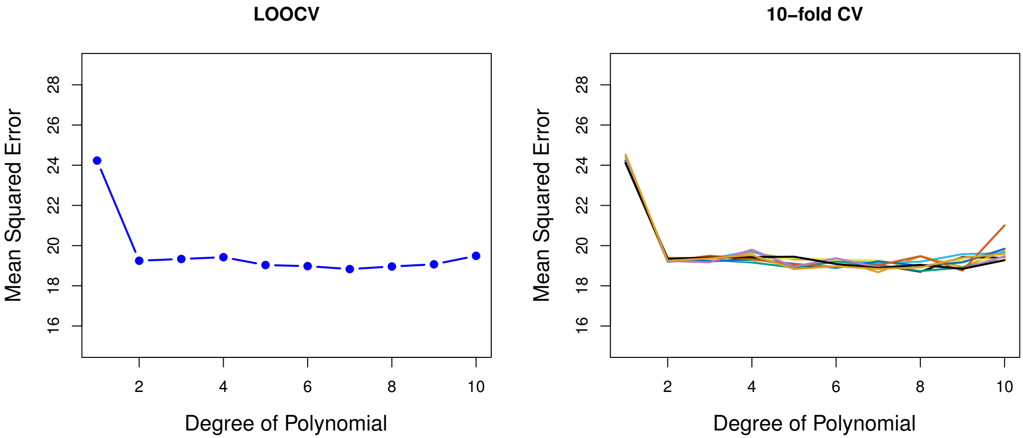
Then

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

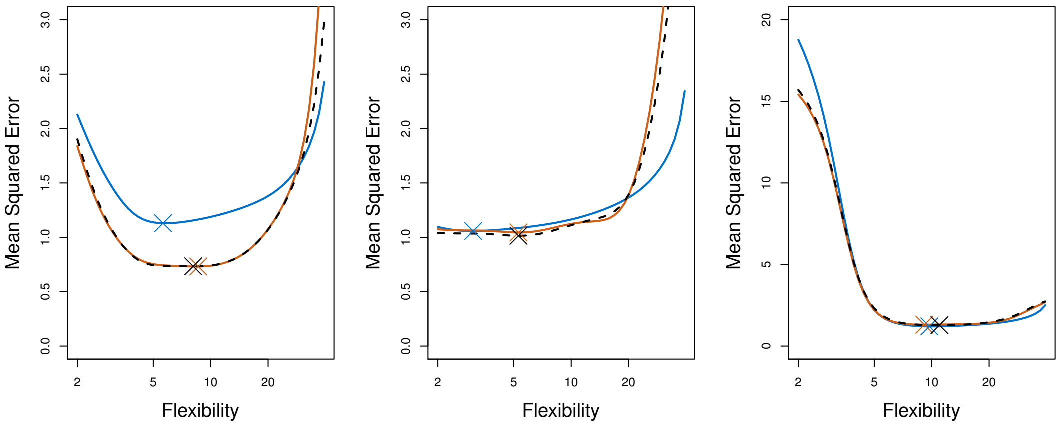
k-Fold Cross-Validation

- LOOCV is a special case of k-fold cross-validation where $k = n$.
- In practice, we usually set $k = 5$ or $k = 10$.
  - Computational advantage (fewer models to fit).
  - Good balance for bias-variance trade-off.

k-Fold Cross-Validation



RHS: nine different 1-fold CV estimates for the `Auto` data set. Notice that there is some variability, but it's not substantial.



True test MSE shown in blue (simulated data). LOOCV estimates shown as black dashed curve. Orange curve shows 10-fold CV.

In some cases, we may only be interested in the location of the minimum point in the estimated MSE curve (for choosing a model or method).

Advantages of *k*-Fold Cross Validation: Bias-Variance Trade-Off

Bias:

- Validation set approach can lead to overestimates of test error rate (uses only part of data).
- LOOCV gives approximately unbiased estimates of test error (uses almost all of data).
- 5- or 10-fold cross-validation has some intermediate amount of bias.

Advantages of k-Fold Cross Validation: Bias-Variance Trade-Off

Variance

- LOOCV has a higher variance than k-fold CV with $k < n$.
- Each LOOCV model fit is trained on almost all of the data.
    - Therefore its outputs are highly correlated with each other.
- In contrast, 5- or 10-fold CV results in models that are less correlated with each other (overlap between training sets is smaller).
- The mean of many highly correlated quantities will have higher variance than the mean of many quantities that are not so highly correlated.
    - Thus the test error estimate for LOOCV tends to have higher variance than that of 5- or 10-fold CV.

Cross-Validation for Classification Problems

Now, instead of examining MSE, we examine the number of misclassified observations.

For LOOCV, this looks like

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

where $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$

LOOCV: Default Data

```r
library(boot)
mod3 <- glm(default ~ ., data=Default, family="binomial")
cv.err <- cv.glm(Default, mod3, K=10)
cv.err$delta # error rate
```

```
## [1] 0.02140122 0.02139571
```