

Data and Validation

Machine Learning in Molecular Science

Prof. Michael Shirts

July 22nd, 2024



The importance of good data

- GIGO
 - Garbage in, Garbage out
- Sometimes (often?) data is erroneous
- Need to be careful about what you include in your data set
- For LARGE data sets
- Anomaly detection
 - A problem in unsupervised learning

The Problem: Missing Data?

- How bad is the problem?
 - Let's say we have 10,000 data points, and there are 100 features per point
 - Assume a 3% chance for each feature to be missing for every point
 - The chance that any given data point is complete is $(1-0.03)^{100} = 0.048$
 - Only 4.8% of the points have all features, despite the data being 97% complete!
 - That's leaving a lot of data if we only use ones with all features!

Data imputation

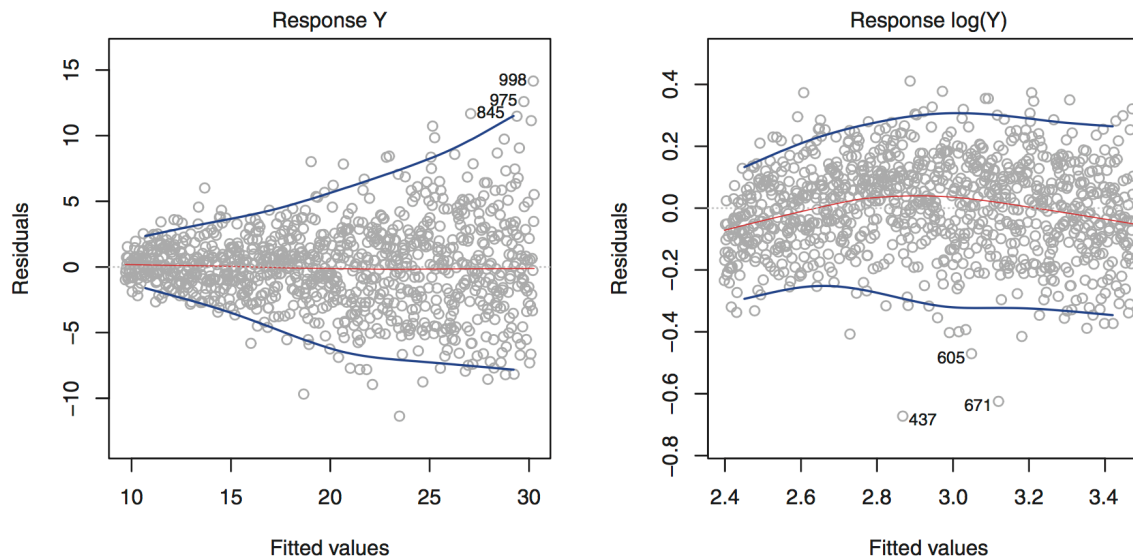
- Strategy
 - Fill in missing data with using "around" it
 - What are the choices?
 - mean of all other choices for that feature
 - Random selection of that feature
 - The mean of "close" data
 - Regression from other inputs
- multiple imputation generates multiple guesses for each missing data point, which can help improve the statistics
- Tools in pandas and scikit-learn to input missing data.

Data Leakage

- Does training data get into the testing set?
- Are some of your features actually surrogates of your labels?
- What is going to happen with ChatGPT 8 when it ends up getting trained on outputs of ChatGPT 4 through 7?

Nonconstant Variance

- This phenomena is known as heteroscedasticity
- One solution: transform data to logarithmic form



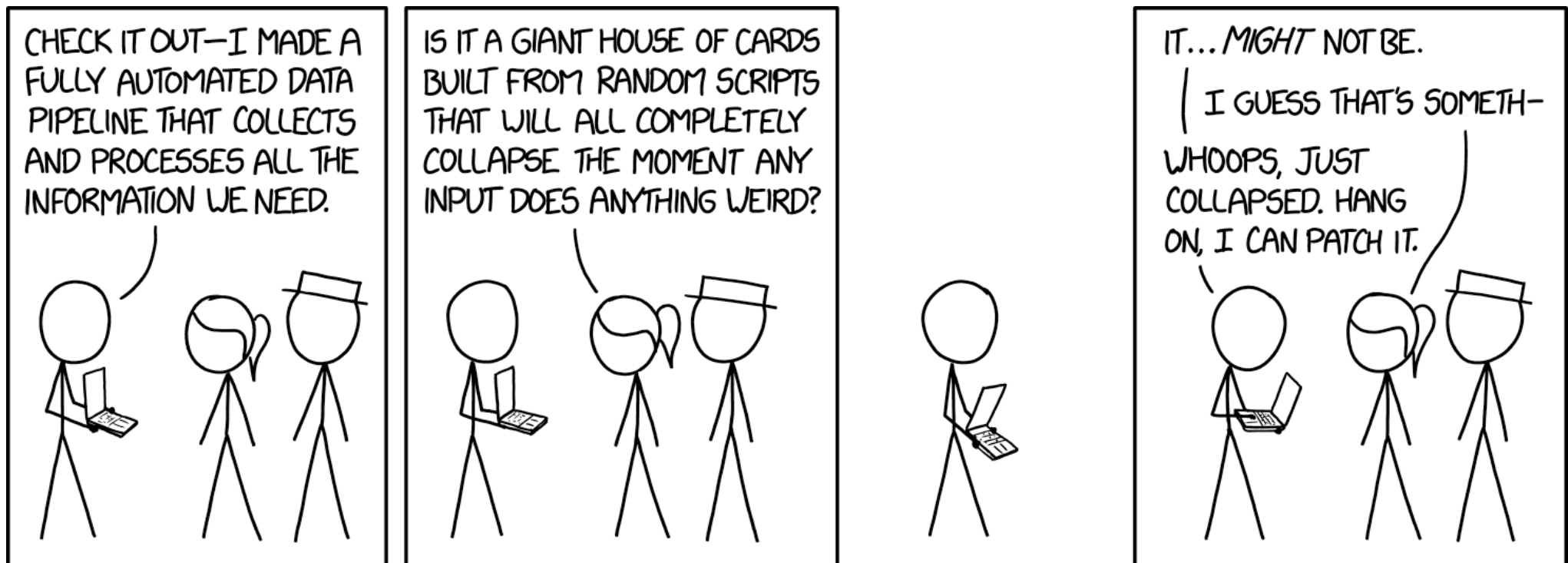
- (Sometimes) Weight the residuals

$$RSS = \sum_i w_i (y_i - \beta_0 - \beta_1 x)^2$$

- Question: should high variance x's be higher in weight or lower in weight?

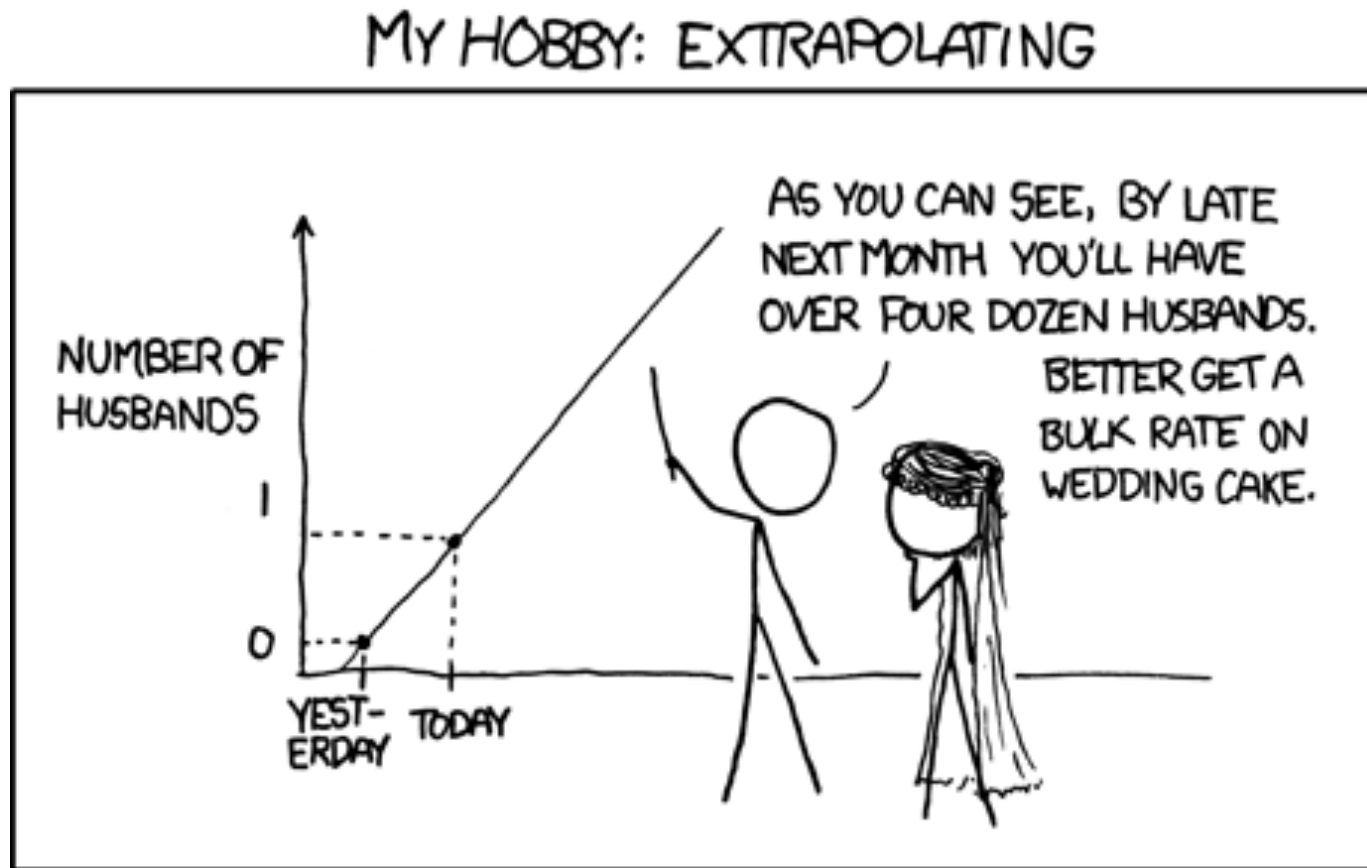
Be careufi

- Two needs:
 - Automating data processing to ensure consistent treatment over the entire data set.
 - Carefully human curation of the data to make sure there's nothing the automation didn't handle.



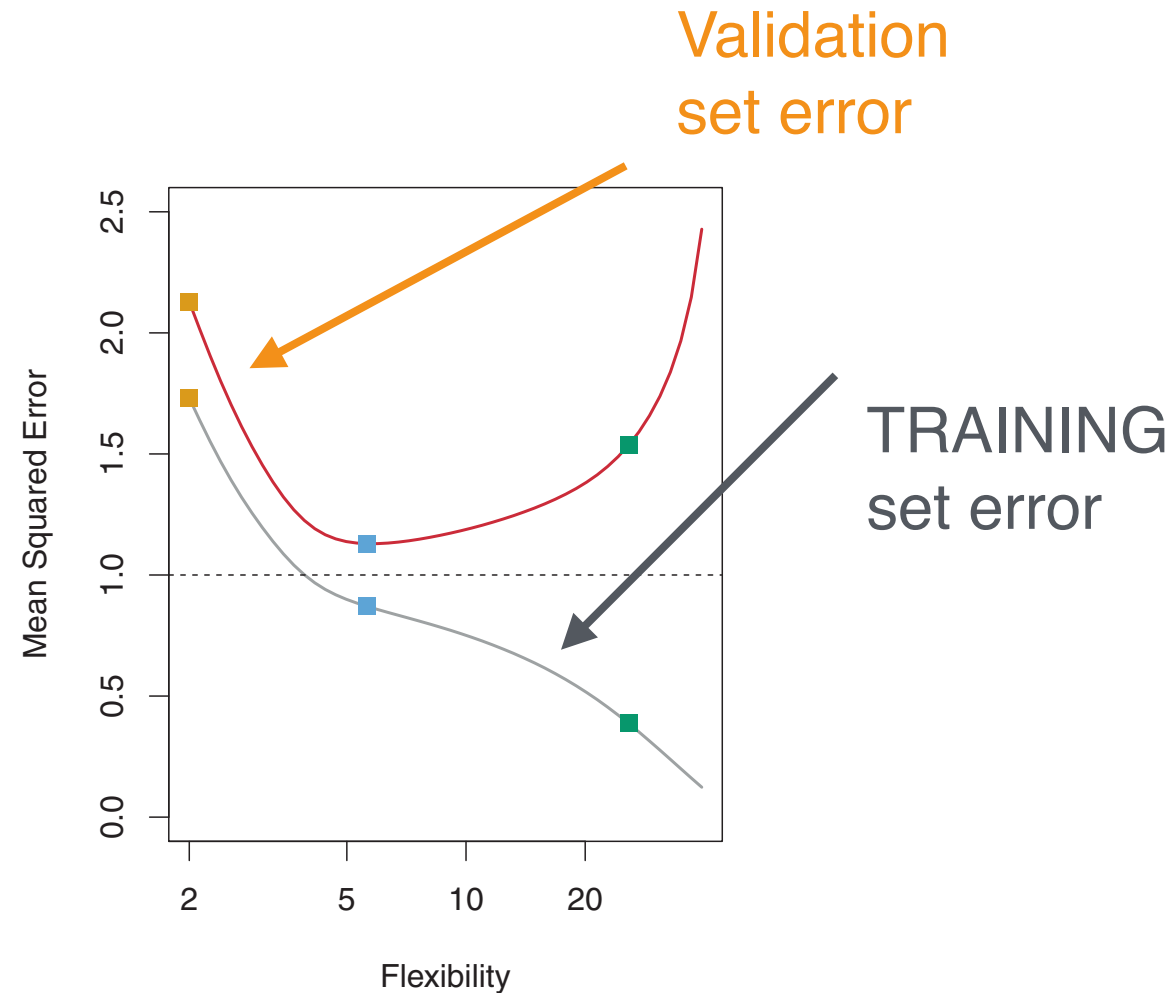
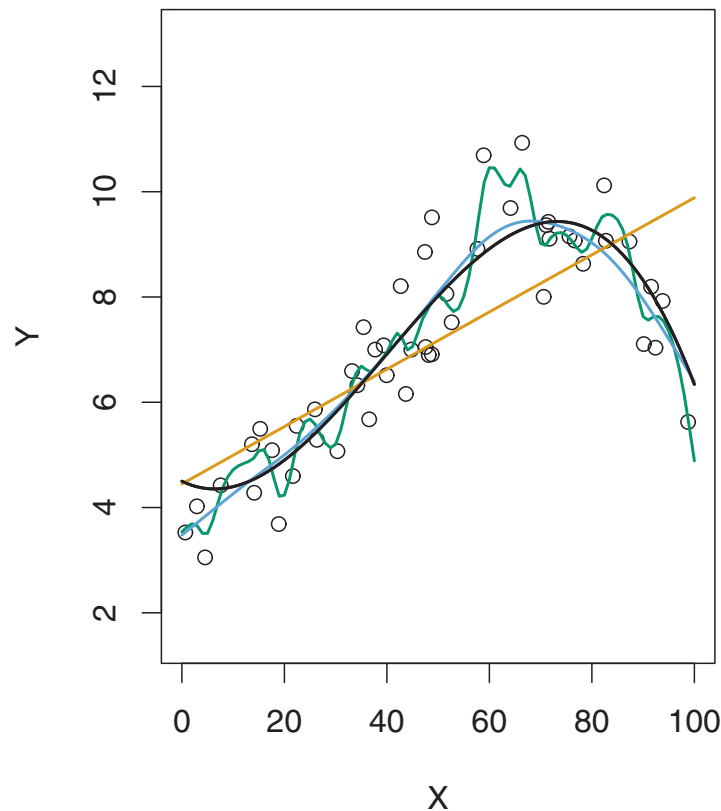
Go to notebook!

Importance of validating models

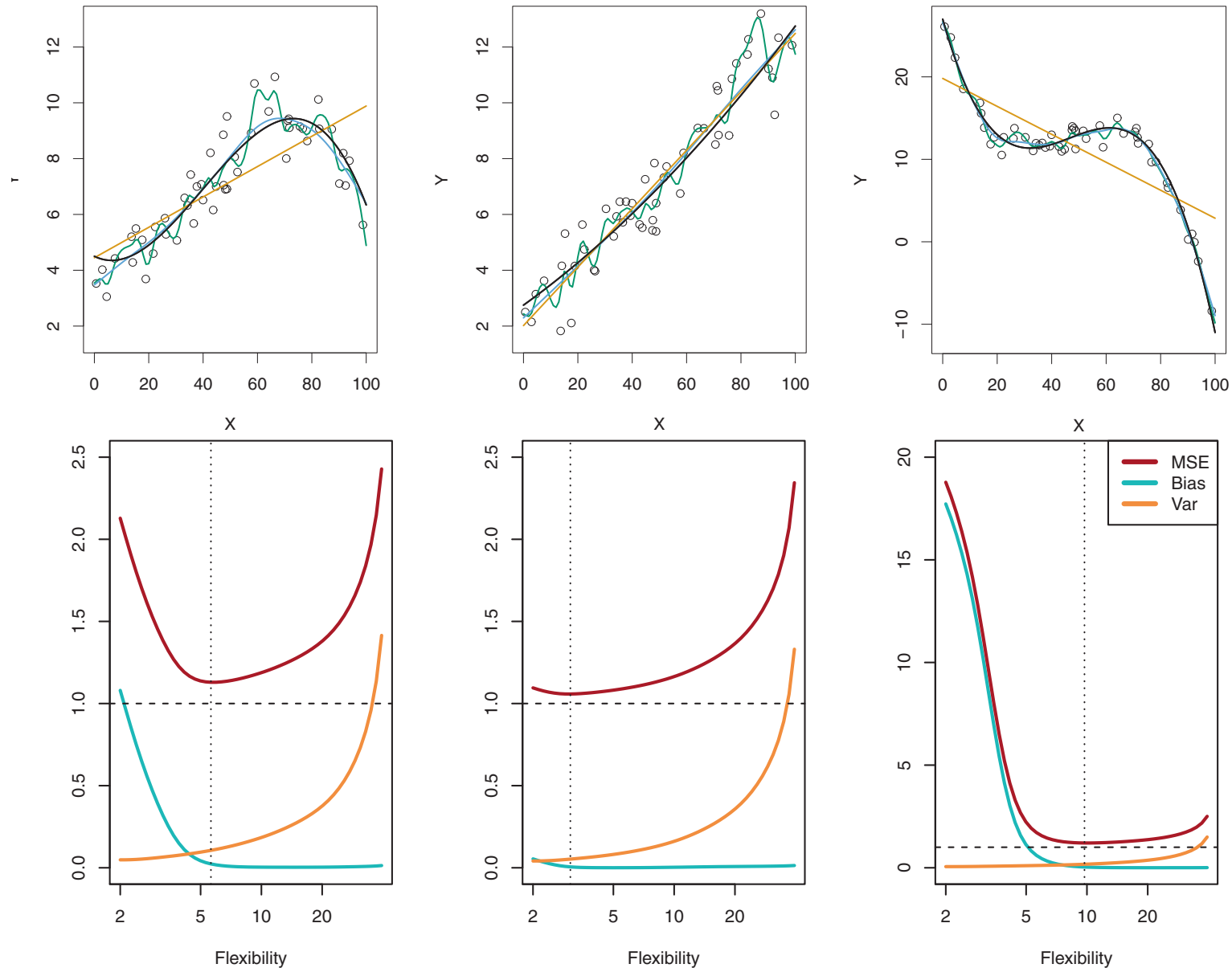


<https://xkcd.com/605/>

Goal is to minimize OUT OF DATASET MSE, not TRAINING MSE



Different types of data have different bias/ variance tradeoffs



Validation

- Basic idea: treat some of your training data as validation data instead.
- Three strategies
 - Train on some, validate on some
 - Leave one out cross-validation
 - K-fold cross-validation

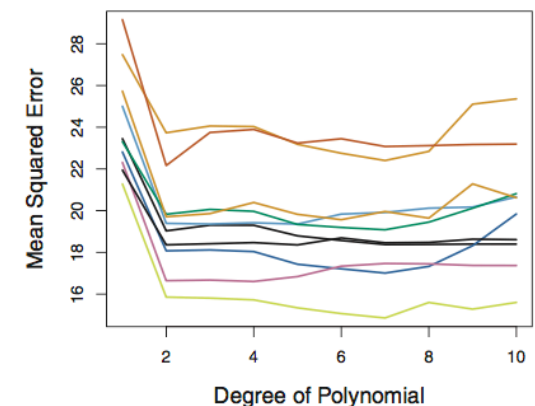
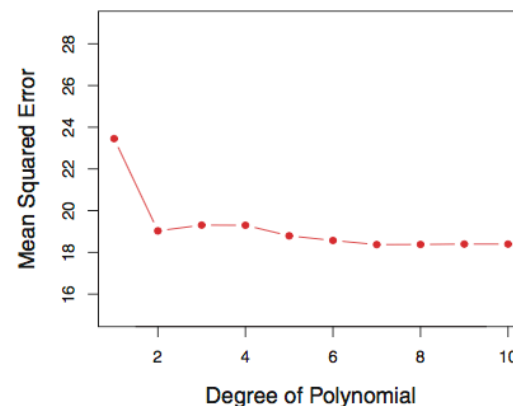
Train some, test some



Train

Test

- Problems:
 - We are leaving some data out of the training
 - How do we make the decisions of how much to do?
 - How do we decide which set of data to reserve for testing?
 - It's a noisy process!



Leave One Out Cross Validation (LOOCV)

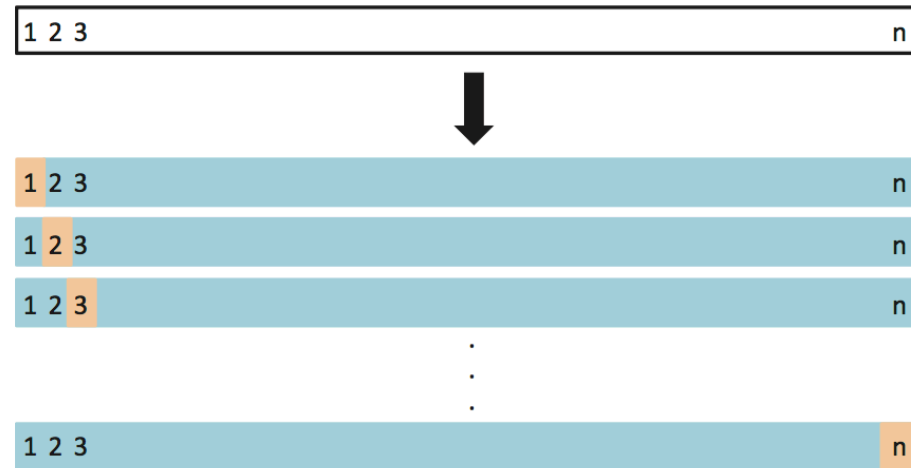


FIGURE 5.3. A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i.$$

Each test is ONE prediction
Report the average over the N tests

K-fold Cross Validation



Test

Train

Example: 5 fold
cross validation



Train

Test

Train



Train

Test

Train



Train

Test

Train



Train

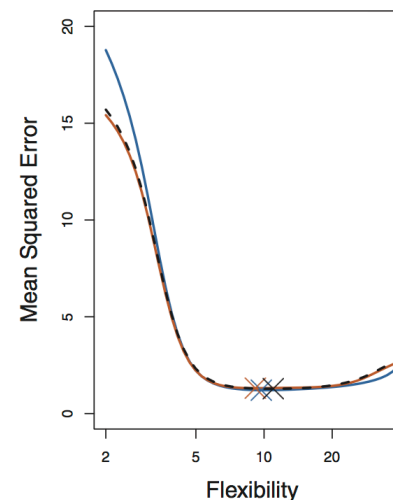
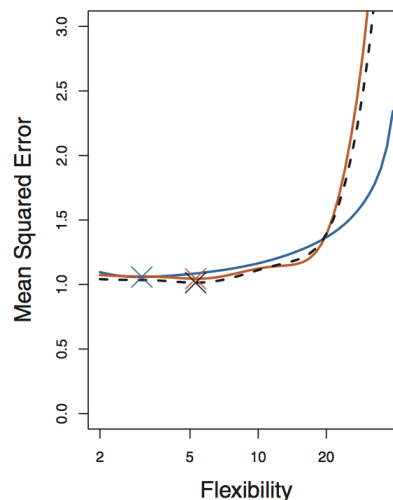
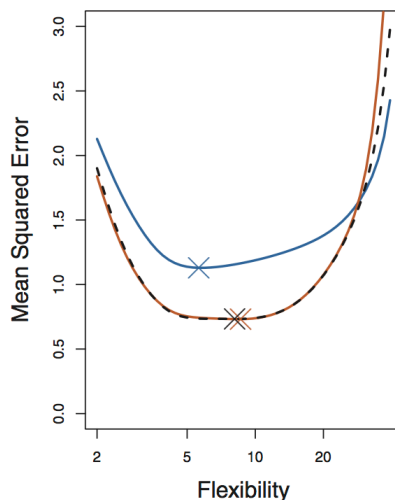
Test

- Average the five training and test MSEs.
- Report the average

$$CV_{(k)} = \frac{1}{K} \sum_{i=1}^K MSE_i$$

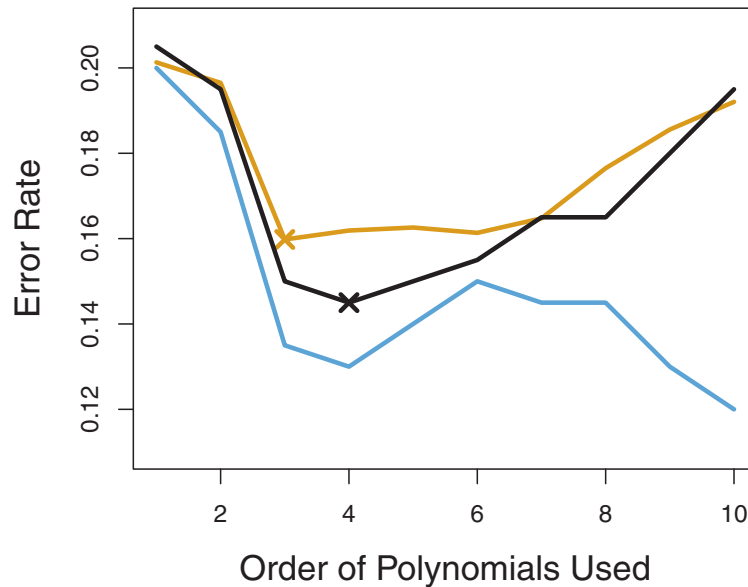
Bias/variance tradeoff: use 5 or 10 folds

- Leave-one-out cross validation is N-fold
 - Always gives the same answer
 - Can be very expensive for large data sets!
- Empirically, 5 fold and 10 fold provide a good bias/variance tradeoff.
- 10 fold is $\sim N/10$ times cheaper than LOOCV!

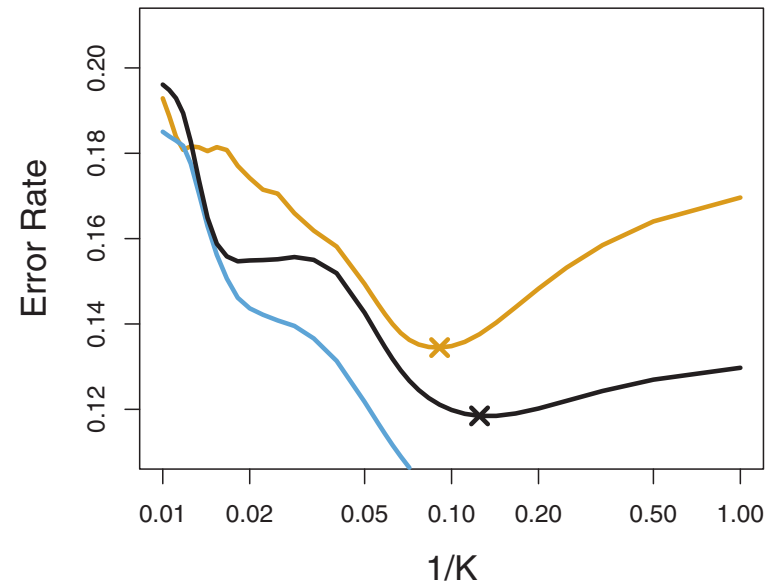


Works for classification as well: Works for all supervised learning

Logistic regression with polynomial boundary



K-nearest neighbors



Blue: training set error

Orange: validation set error

Black: 10-fold cross validation

Reminder:

- Training: Used to train your model
- Validation: Used to estimate out-of-data set performance, in order to tune hyperparameters
- Test: FRESH data, separate from validation to see how the good tuned model is.
- If you are iteratively changing your model based on the test data set, it's actually the validation data set.

To the workbook!