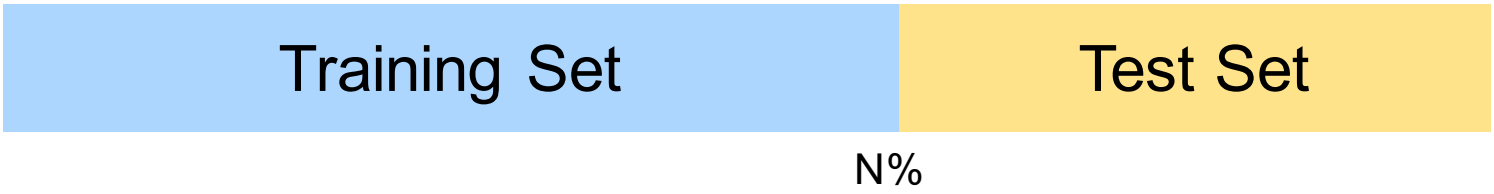
Holdout Sampling

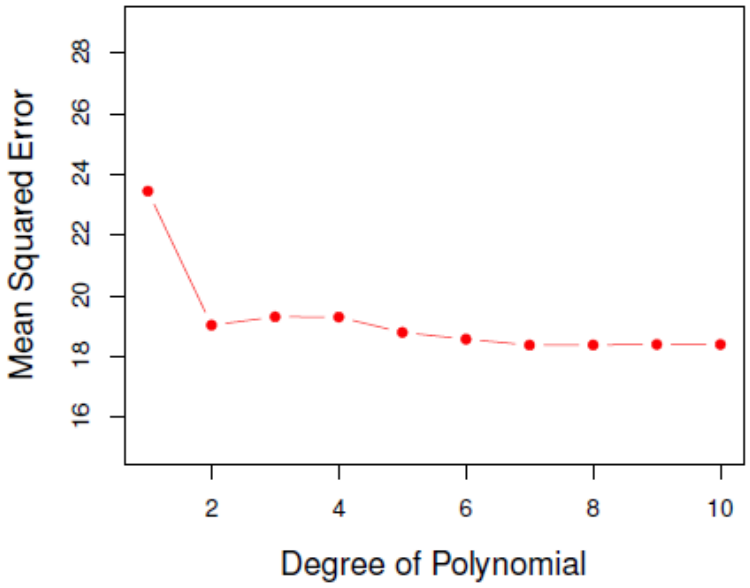
Holdout/Random Splitting

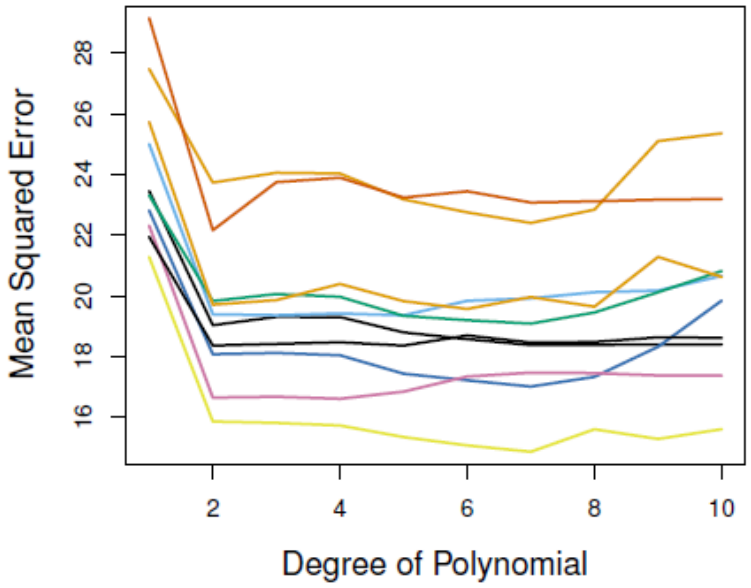


- **Randomly** select **N%** of the data for the **training** set and hold out the remaining data for the test set.
- If N is **too small** (e.g., 50%), then you lose statistical power with the reduced training set, which is problematic with small samples.
- If N is **too large** (e.g., 90%) the resulting training models will be over-fitted.
- Common values for N are between 70% and 80%
- With a single split, the test MSE may be misleading if you get an "unfortunate" split

Random Splitting Illustration

The example below was generated with the **"Auto"** data set in R, predicting **gas mileage** with **horsepower** as the predictor using various polynomial regressions. The **left panel** shows a **single split**, whereas the **right** shows **10** different random **splits**, illustrating the variability you can get from various splits. All models show that the MSE drops is sharply reduced with a **squared regression** and that the MSE does not improve substantially with higher polynomials.





R Tips

`set.seed(1)` → This is a useful command to run before generating random values with the `sample()` function below. Setting the seed to "any" value selects the same observations the next time you draw a random sample. Without a fixed seed you will get slightly different results each time

`set.seed(2)` → Use a different seed if you want to draw a different random sample

`train=sample(392,196)` → Generates a 196 random numbers out of 392 data points (50% sample), which can then be used as an index with `subset=train` or `[train]` or `[-train]` for the test data

`lm.fit=lm(y~x1+x2+etc.,data=dataName,subset=train)` → Fits the model on the training data subset

`mean((y-predict(lm.fit,dataName))^2)` → Calculates the MSE for all the data (i.e., `[-train]`)

`mean((y-predict(lm.fit,dataName))[-train]^2)` → Calculates the MSE for the test data (i.e., `[-train]`) (use [train] to compute the training MSE instead)

KOGOD SCHOOL
of
BUSINESS