MATH 569 Statistical Learning

Part VI: Model Assessment and Selection

Maggie Cheng

ILLINOIS INSTITUTE OF TECHNOLOGY

Fig 7.1

red light curve: conditional test error on one training set Err_τ
red solid curve: expected test error $E[Err_\tau]$
blue light curve: training error for one training set $\overline{err}$
blue solid curve: expected training error $E[\overline{err}]$

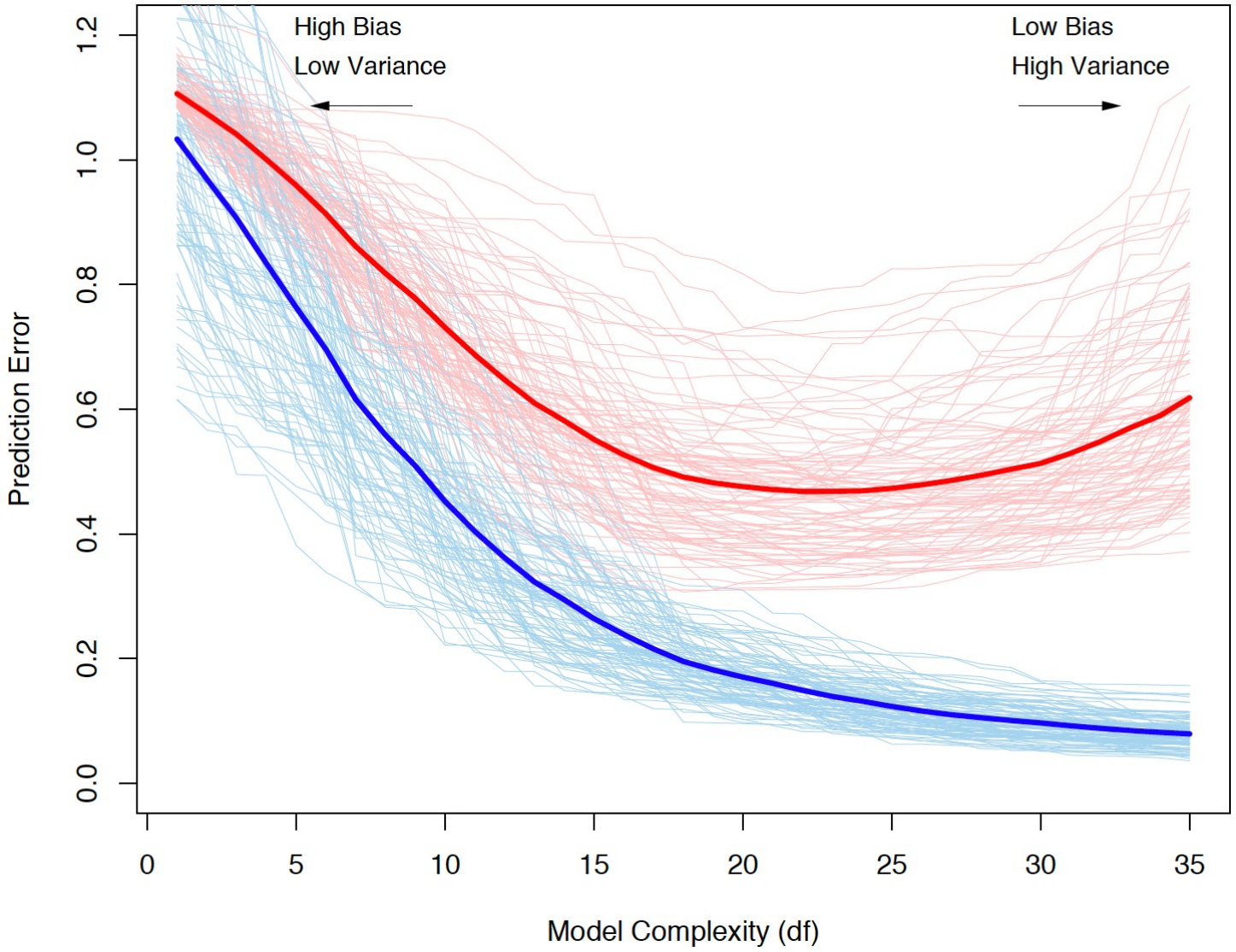


Fig 7.2 Behavior of bias and variance

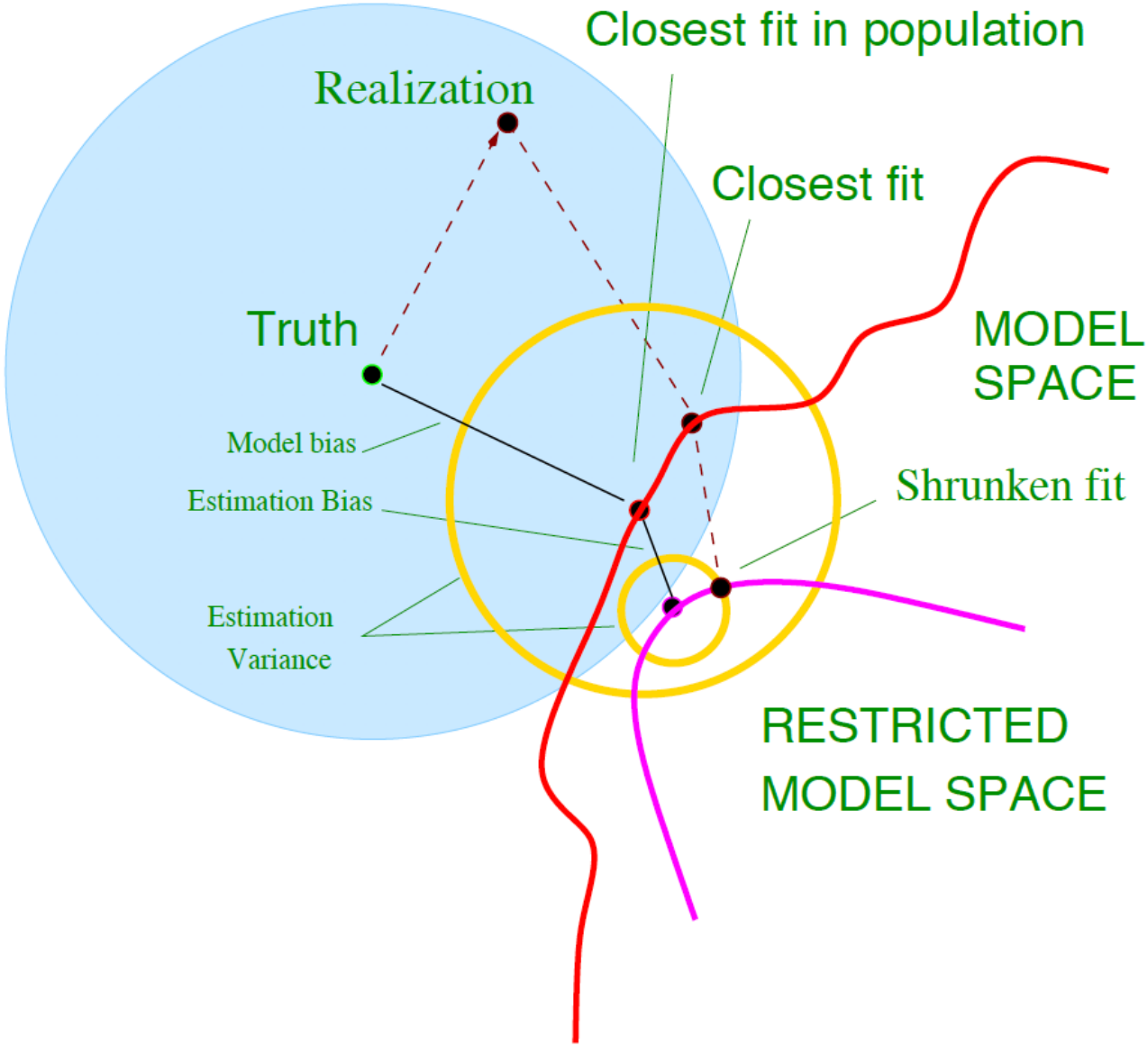


Fig 7.3

Expected prediction error (orange)
squared bias (green)
variance (blue)

Left: orange= green + blue
Right: not equal

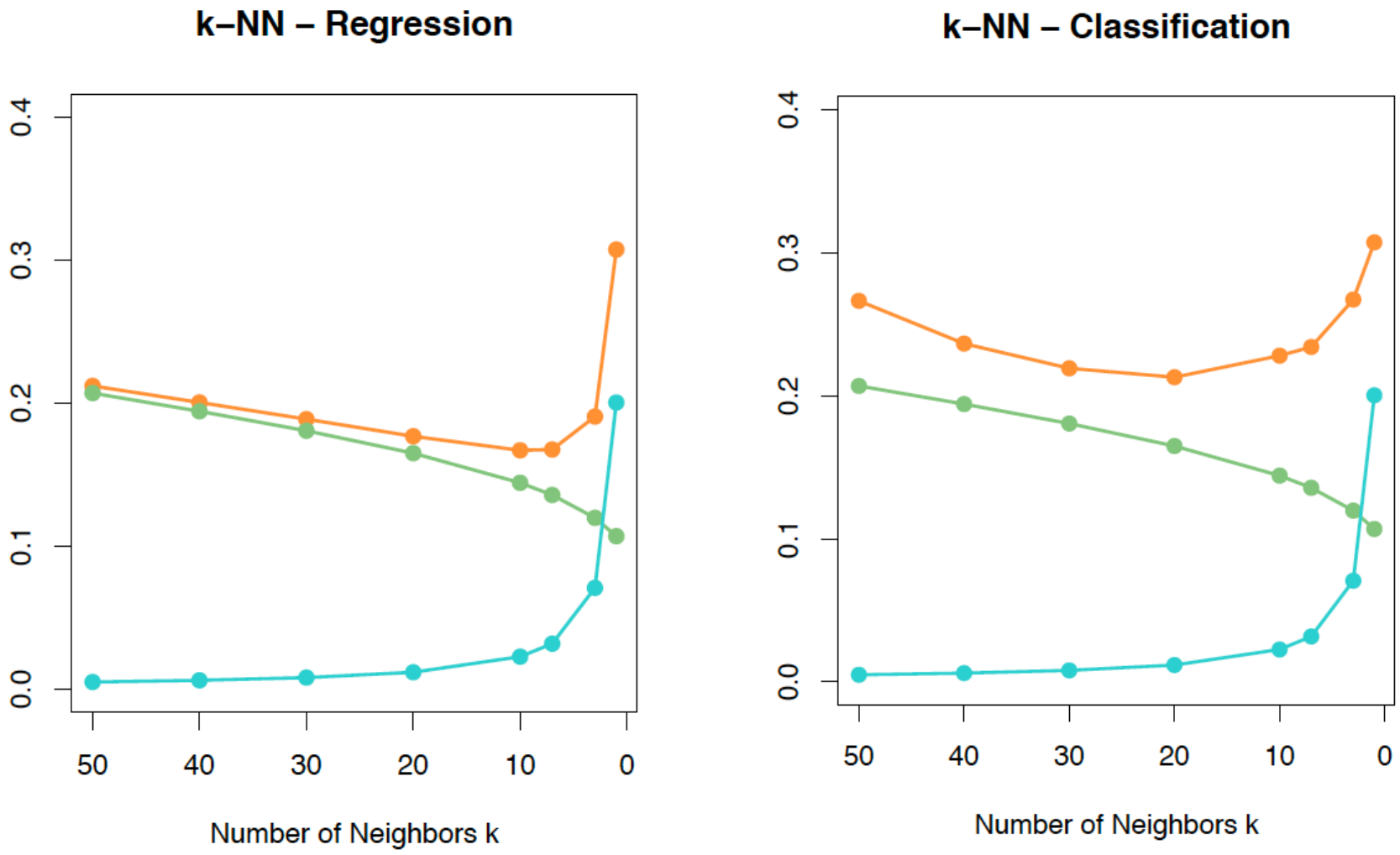
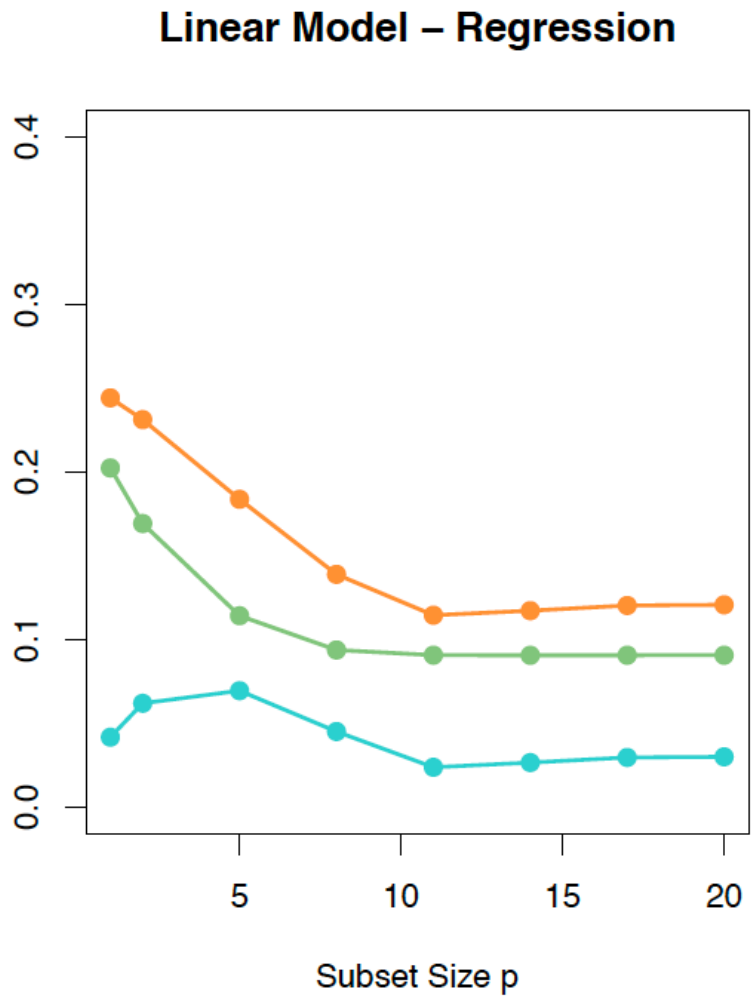
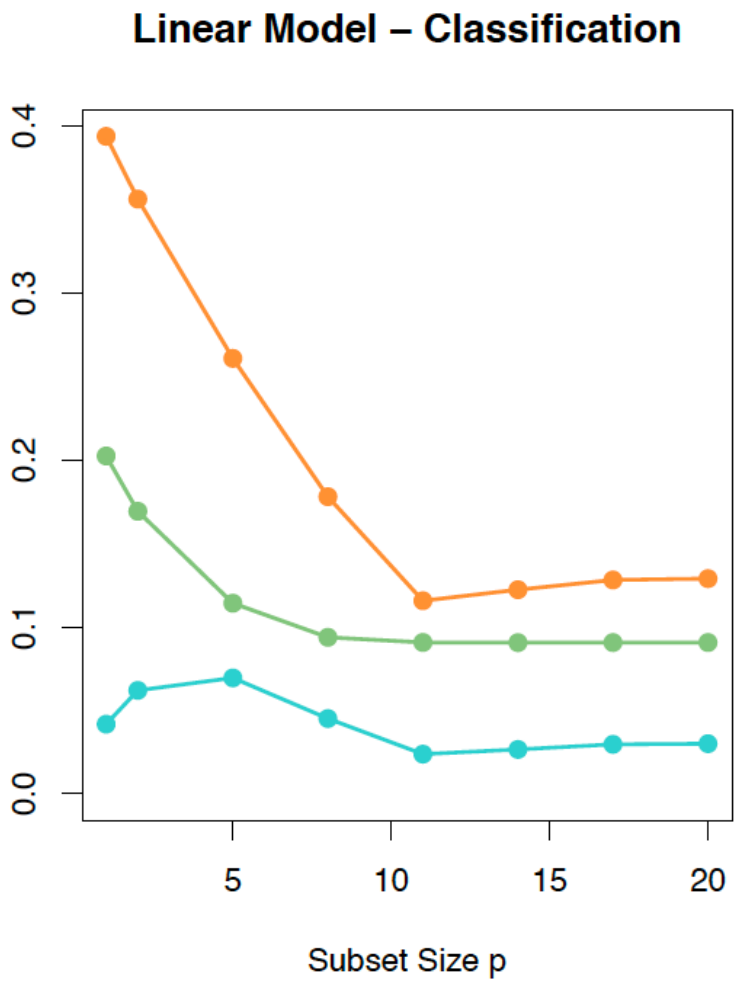


Fig 7.3

Expected prediction error (orange)
squared bias (green)
variance (blue)

Left: orange= green + blue
Right: not equal





Training-set-size Bias for Cross Validation

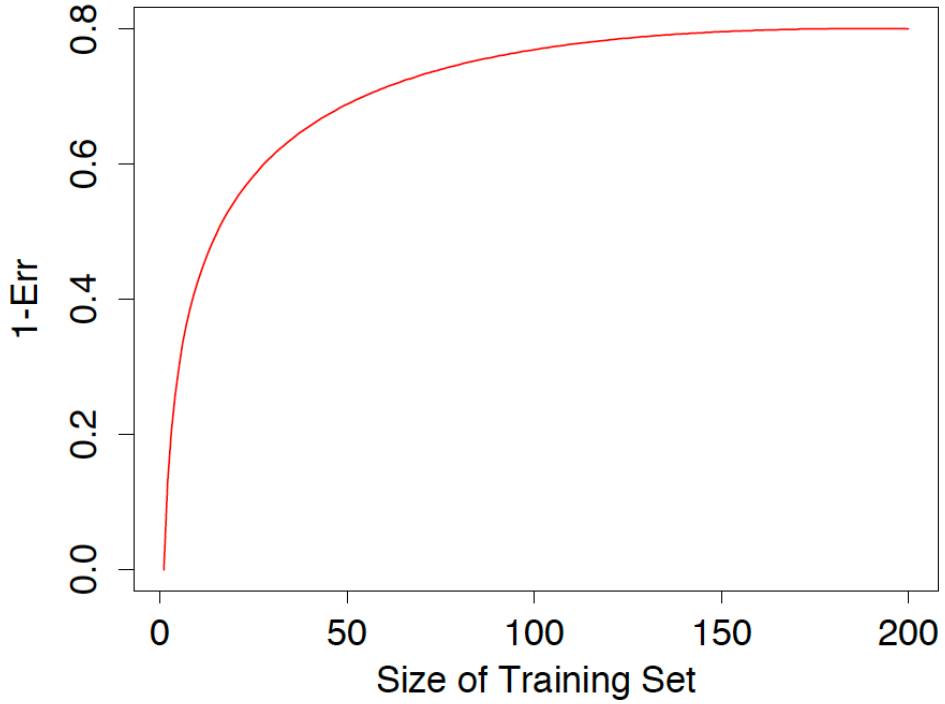


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of* $1 - \text{Err}$ *versus the size of the training set* N*. With a dataset of* 200 *observations,* 5*-fold cross-validation would use training sets of size* 160*, which would behave much like the full set. However, with a dataset of* 50 *observations fivefold cross-validation would use training sets of size* 40*, and this would result in a considerable overestimate of prediction error.*

The wrong and right way to do cross validation

Wrong way

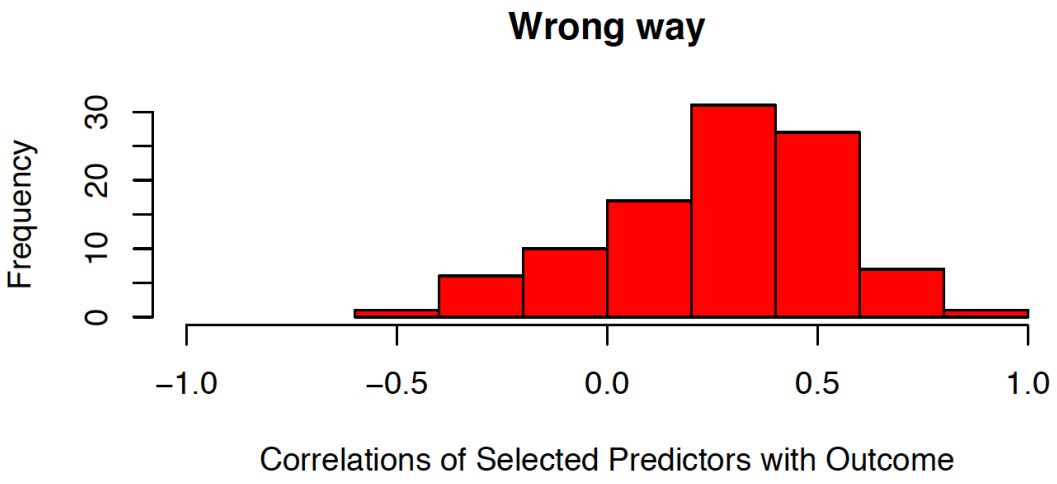
1. Screen the predictors: Select predictors based on all samples
2. Build a model: Use the selected predictors to build the model
3. Perform cross-validation: Use cross validation to estimate the unknown tuning parameters and estimate the prediction error of the final model

Right way

1. Divide the samples into K folds
2. For each fold k=1, …, K
    a) Find a subset of good predictors using all but the k-th fold samples
    b) Use this subset of predictors to build a model using all but the k-th fold samples
    c) Use the model to predict the outcome for the samples in the k-th fold

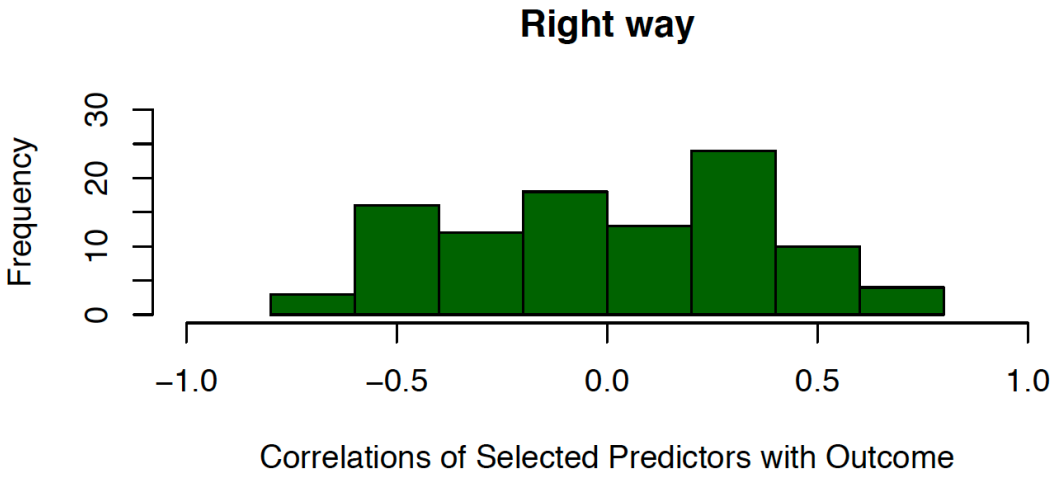
The wrong and right way to do cross validation

Fig 7.10



Ave(correlation)=0.28

What's the problem: the good predictors are chosen after seeing all samples.



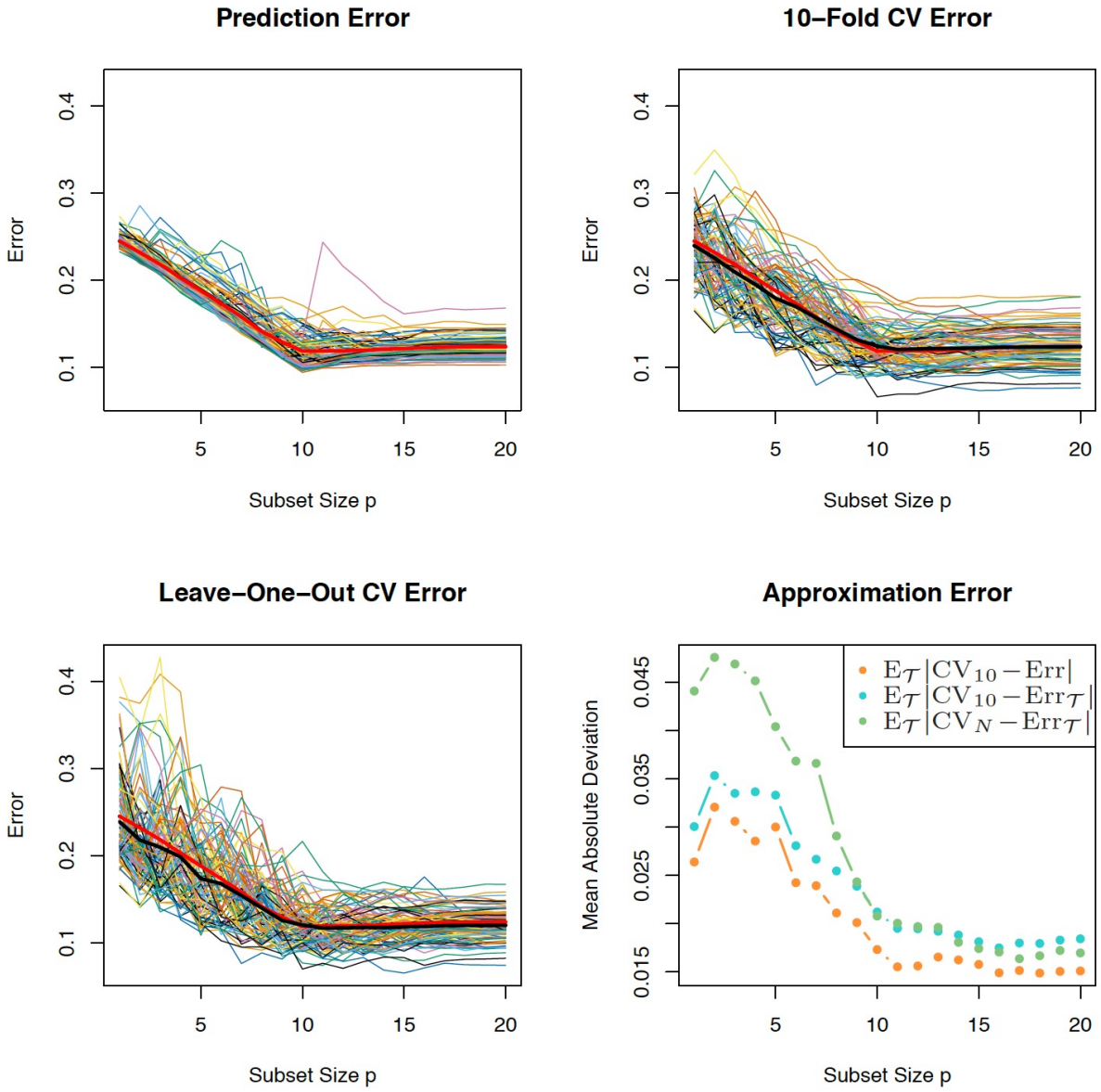
Ave(correlation)=0

Samples must be left out before any selection or screening is applied that uses labels.

One exception: unsupervised screening that does not use label.

What does k-fold C.V. error really estimate--- Conditional or Expected Test Error?

Fig 7.14



Thick red: Err

Thick black: $E_\tau[CV_K]$

Schematic of the bootstrap process

Fig 7.12