Definition: Data with more predictors (p) than observations (n), i.e. p>n.

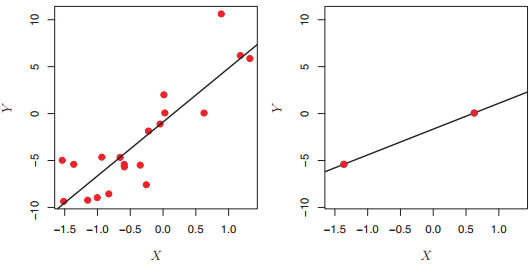
Why it matters: HDD arises frequently in modeling scenarios with expensive or limited samples (like pharmaceutical trials) or many predictors (like tracking an individual’s internet movements). Tech advances have enabled HDD prediction.

How to model HDD: Use subset selection (e.g. forward stepwise), shrinkage (e.g. lasso or ridge regression) or dimension reduction (like principal components regression) to reduce or eliminate unrelated predictors.

Basics

High Dimensional Data (HDD)

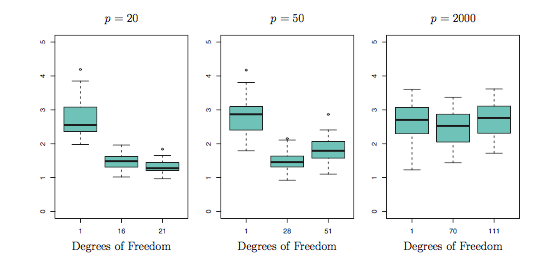
Pitfalls



HDD Challenges: Traditional methods like linear, logistic, and LDA regression don’t work because they will show a perfect fit regardless of whether a true relationship. exists.

HDD Regression: Here we see how the lasso performs as p increases. Only ~20 predictors are truly correlated features, but more are included in the model as p goes up and lasso misses some irrelevant features. Test error usually increases as dimensionality goes up.

An HDD Regression



Interpreting Results from HDD Models

Consider Multicollinearity: Multicollinearity is when 3 or more predictor variables have high correlation with one another. This is another reason sub setting, regularization, and shrinkage are key.

Relevant Predictors Vary: Because there are so many features in a high dimensional setting, models will settle on different predictor variables between test sets. EX. If we are predicting blood pressure for a small number of individuals based on millions of DNA pairs, the relevant pairs will vary between groups (but both sets may still be valid).