Ridge regression is a technique for reducing variance in multidimensional data introduced by Hoerl and Kennard in 1970[ridge paper]. By applying a penalty term to the variances of each term, bias is introduced to the estimator but the overall variance is reduced and the total error decreased. Ridge regression complements the Bayesian approach to parameter estimation. A lambda term is often called the shrinkage parameter [Stanford slides] controls the size of the coefficients and the amount of regularization, optimizing the the mean square error.

Standard linear regression attempts to find a line orthogonal to all data points but suffers when data points are co-linear, or contain multiple dependant variables [need a citation]. This is unfortunate when trying to regress the Ames housing data because a high value home will contain many desirable features (as well as the converse being true). A general equation for regression would be

Y = XB + e [NCSS software]

“Multicollinearity can create inaccurate estimates of the regression coefficients” [NCSS software] which Hoerl’s shrinkage parameter attempts to minimize.

Ridge regression alleviates this problem

Ridge regression is ideally suited to the Kaggle housing prices challenge because scoring is based on the log mean square error of the submission estimates and this technique is designed to minimize those.

Multi-collinearity

Talk about ames data being cleaned from acticle

Physical constraints --- its all single family homes in Ihiio

Outliers -- identified and removed