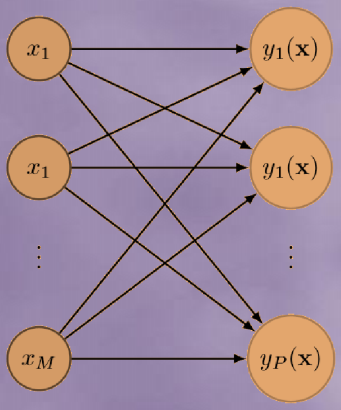

Linear Models

Daniel Khashabi[1]
KHASHAB2@ILLINOIS.EDU

0.1 Simple linear regression model

Here we assume a linear model of regression. Lets consider an auxiliary matrix $\mathbf{X_{aux}} = [\mathbf{X}, 1]^T$. In least square regression we define the regressor as follows:

$$\mathbf{y} = \mathbf{X_{aux}}\beta + \varepsilon$$

Now it's clear that why we defined $\mathbf{X_{aux}}$. The reason is that we want our linear model has an intercept value. The goal is to minimize the following formula for RSS(residual sum of errors):

$$RSS = \sum_{i=1}^{N} (y_i - X_{aux,i}\beta)^2$$

It can be shown that the closed solution to this problem is as follows:

$$\hat{\beta} = \left(\mathbf{X_{aux}^T X_{aux}}\right)^{\mathbf{-1}} \mathbf{X_{aux}^T y} \tag{1}$$

The linear least squares regresion is important in several ways. In addition to its simplicity, the Gauss-Markov Theorem asserts that the least squares estimate of the parameters β have the smallest variance among all linear unbiased estimates, though being “unbiased" doesn't mean to be perfect all the times, meaning that, there may exist a biased estimator that has smaller mean squares error. That is why Ridge regression and Lasso are introduced.

0.2 Subset selection based on AIC and BIC

With subset selection, we retain only a subset of predictor. So least squares is used to predict the result for variables that are retained.

[1] This is part of my notes; to see the complete list of notes check web.engr.illinois.edu/ khashab2/learn.html.

In greedy stepwise *forward* subset-selection, we incrementally consider variables and add them up by considering the goodness criterion. While in *backward* selection we delete predictors that are redundant based on the information ciriteria. Though these criteria are sub-optimal, there are various reasons that are prefered over other methods, especially from computational complexity point of view. Akaike Information Criterion(AIC) and BIC(Bayesian Information Criterion) are among measures for goodness of fit of a statistical model, to make a desired balance between bias and variance. *Backward* and *Forward* subset selection in R, use AIC and BIC as goodness measures.

0.3 Ridge regression with λ chosen by GCV

In order to add more continuity in regressions of linear least square method, we add a shrinkage parameter in the function to be minimized.

$$RSS = \sum_{i=1}^{N} \left(y_i - \beta_0 \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

By adding an extra parameter, λ which helps to regularize the sum of the errors. The closed solution to this minimization is as followoing :

$$\hat{\beta}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

By changing the parameter λ we can get different regularized outputs of our model. For this model we define the *effective degree of freedom*, as a measure of how regularized the variables are, as follwing:

$$edf = \text{tr}\left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\right]$$

This criterion is reasonable; in fact when $\lambda = 0$, i.e. no regularization, and as a result $edf = p$. Whens $\lambda \to \infty$, i.e. full regularized regression, $edf \to 0$.

0.4 Lasso with λ chosen by C_p

in *Lasso* model, similar to shrinkage mode, the estimate is given by:

$$RSS = \sum_{i=1}^{N} \left(y_i - \beta_0 \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

This model lies between the simple least squares model, and ridge shrinkage method. Unlike the ridge regression, there is no direct-way in Lasso to find the closed form for β.

In order to fine tune the parameter λ, the regularization parameter of Lasso, one can use Mallow C_p criterion, without cross-validation.

Ridge method, shrinks all of the predictors, but it shrinks low-variance predictors more, and keeps high-variance predictors(close to orthogonal). Lasso model acts somewhere between ridge regression and best-subset selection method. So it has some qualities of both models.