# Three Ways to Limit Complexity of Linear Regression Models

Unless we have many more training cases than inputs, least squares estimates for regression coefficients may have too high a variance.

Three ways to reduce variance by limiting “complexity” are discussed in the text:

1) Use only a selected subset of the input variables.

2) Find estimates for the regression coefficients by minimizing RSS plus a penalty involving the size of the $\beta$'s (or equivalently, by minimizing RSS subject to a constraint on the size of the $\beta$'s).

3) Replace the original inputs with a smaller set of variables that are linear combinations of the original inputs (equivalently, we find some directions in the input space, and use the projections on those directions).

These methods overlap somewhat: Some penalty methods may set some $\beta$'s to exactly zero, effectively eliminating those input variables. Input selection can be seen as choosing directions restricted to the coordinate axes.

# Ridge Regression

*Ridge regression* adds a penalty of $\lambda \sum_{j=1}^{p} \beta_j^2$ to the residual sum of squares.

The intercept ($\beta_0$) is usually not included in the penalty. This can be done by first centering the inputs and response variables (shifting them to have mean zero over the training set), then fitting a model with no intercept.

We then minimize

$$\begin{aligned}
\text{RSS}(\beta) + \text{Penalty}(\beta) &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \;+\; \lambda \beta^T \beta \\
&= \left( \begin{bmatrix} \mathbf{0}_p \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \sqrt{\lambda}\mathbf{I}_{p\times p} \\ \mathbf{X} \end{bmatrix} \beta \right)^T \left( \begin{bmatrix} \mathbf{0}_p \\ \mathbf{y} \end{bmatrix} - \begin{bmatrix} \sqrt{\lambda}\mathbf{I}_{p\times p} \\ \mathbf{X} \end{bmatrix} \beta \right)
\end{aligned}$$

Where $\mathbf{0}_p$ is a vector of $p$ zeros, and $\mathbf{I}_{p\times p}$ is the $p$ by $p$ identity matrix. This is like pretending there are $p$ extra training cases with responses of zero.

The solution is

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{p\times p})^{-1} \mathbf{X}^T \mathbf{y}$$

The solution always unique if $\lambda > 0$, since $\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{p\times p}$ will be non-singular.

# Why Does Ridge Regression Make Sense?

Ridge regression estimate are biased, but have lower variance than least squares estimates. Why should this be good?

One can prove that there's some $\lambda > 0$ that gives better estimates, but this isn't much help, since we don't know what this $\lambda$ is.

Really, we just think that larger $\beta$'s are less plausible than small $\beta$'s. But how big? We can use cross validation to pick a suitable $\lambda$.

**One problem:** The effect of the penalty changes if we change units for (ie, rescale) some of the inputs — eg, from weight in pounds to weight in kilograms.

**The real solution:** Think hard about how big you might expect each $\beta_i$ to be, and rescale each input so the expected size of of each $\beta_i$ is the same. (The $\beta_i$ rescale in the opposite direction to the inputs.)

**The quick and dirty solution:** Rescale the inputs so they all have standard deviation one. This eliminates arbitrary choices, but is otherwise not especially reasonable.

# The Lasso

The *lasso* penalty is $\lambda \sum_{j=1}^{p} |\beta_j|$.

Again, the intercept ($\beta_0$) is usually not included in the penalty. We can first centre the inputs and response variables, then fit a model with no intercept.

The text expresses the lasso in terms of least squares with a constraint:

$$\hat{\beta}^{\,\text{lasso}} \;=\; \underset{\beta}{\text{argmin}}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta), \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le t$$

For every choice of $t$, there is a choice of $\lambda$ that gives the same result, and vice versa. So picking $t$ and picking $\lambda$ (eg, by cross validation) are equivalent tasks.

However: If you fix $\lambda$, and then let the amount of training data increase, your estimates will converge to the right $\beta$ — eventually, the data overwhelms the penalty. But if you fix $t$, your estimates may never be right.