Regularization

What is Regularization?

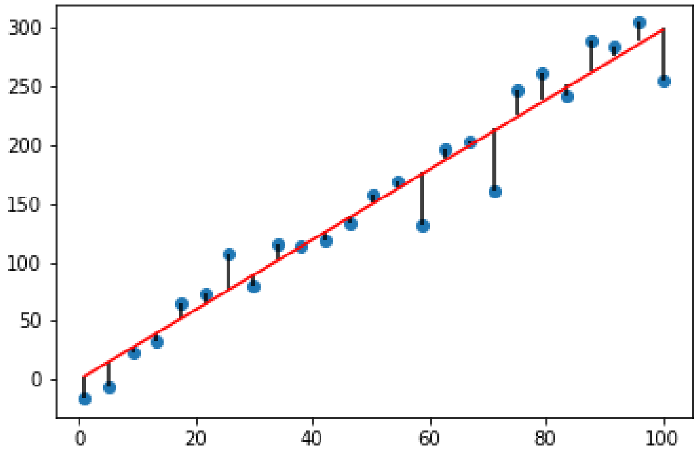
Regularization is used to help with overfitting when datasets are large (i.e., large number of predictors or a large number of samples).

Regularization methods add a penalty term to the regression that works to shrink the regression model coefficient estimates towards zero.

We will be discussing two regularization methods: Ridge Regression and Lasso Regression.

Ridge Regression

Recall from linear regression that we want to fit a line that is as close as possible to our data points. Linear regression uses the **least squares criterion** to measure “closeness”.



Residual

$$e_i = y_i - \hat{y}_i$$

Residual Sum of Squares

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Least squares chooses intercept & coefficients to minimize the RSS

Ridge Regression

In ridge regression, the coefficients are estimated by minimizing a slightly different criterion:

$$RSS + \alpha \sum_{j=1}^{p} \beta_j^2$$

Called the **shrinkage penalty**. It is small when coefficients are close to zero.

In the second term, α is a tuning parameter that is determined by the user. This parameter controls the weight that you want to give to each term in the criterion.

This is called L2 regularization since the shrinkage penalty is the sum of the square of the coefficients.

Lasso Regression

In lasso regression, the coefficients are estimated by minimizing the following criterion:

$$RSS + \alpha \sum_{j=1}^{p} |\beta_j|$$

This uses an L1 shrinkage penalty that allows some of the coefficients to be equal to zero.

Ridge regression will shrink all coefficients towards zero but it will never make any of the coefficients equal to zero. Sometimes, we may want to build a model that excludes some of the predictors. In this case, we would use lasso regression.

Look at Example

Ridge vs. Lasso

How do we choose which method to use?

- Generally, we want to use Lasso when we have a large number of predictors, many of which we expect to have coefficients that are very small or zero.
- Ridge regression works best when the response is a function of many predictors that have coefficients that are about the same.
- In practice, it is difficult to know this information about the predictors. Therefore, it is often recommended to perform cross validation to choose.

Choosing the Tuning Parameter

How do we choose the value of the tuning parameter?

One common way is to use cross validation. Simply choose a grid of α and loop through them, performing cross validation each time. Then choose the value of α that corresponds to the smallest CV error.

Scaling

You must scale your data before using these regularization methods.

In Python, this can be done in the Lasso or Ridge by setting normalize to True:

```
lassoreg = Lasso(alpha=alpha,normalize=True, max_iter=1e5)
```

This is required because the penalty term weighs each predictor equally so we want to make sure each predictor is in the same range.

Methods of Regression

Standard-

Min-Max-

Robust-

Validation methods