**Ridge & LASSO Regression**

**Regularization**

* You have your model ready, you have predicted your output. So why do you need to study regularization? Is it necessary?
* In regularization, what we do is normally we keep the same number of features, but reduce the magnitude of the coefficients.

**Lasso Regression**

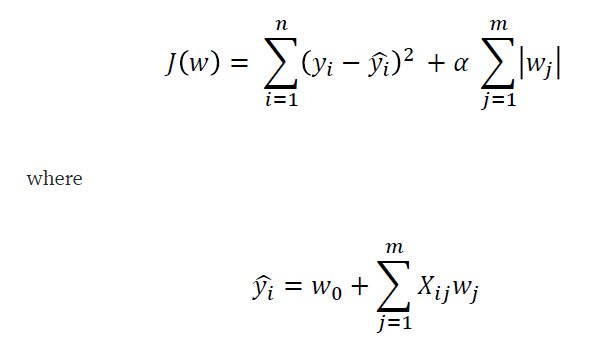
* The acronym “LASSO” stands for **Least Absolute Shrinkage and Selection Operator.**
* Lasso regression is a type of linear regression that uses shrinkage.
* Shrinkage is where data values are shrunk towards a central point, like the mean.
* The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).
* This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

**L1 Regularization**

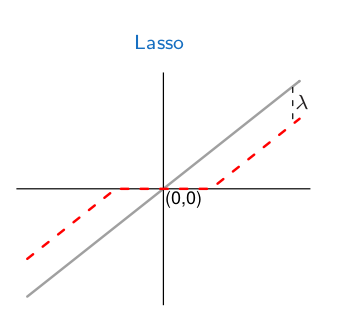
* Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients.
* This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model.
* Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.
* On the other hand, L2 regularization (e.g. Ridge regression) doesn’t result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

**Performing the Regression**

* Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize:



* By increasing the value of the hyperparameter alpha, we increase the regularization strength and shrink the weights of our model.
* Please note that we don’t regularize the intercept term w0. Note also that alpha = 0 corresponds to standard regression analysis.
* Depending on the regularization strength, certain weights can become zero, which makes the LASSO method a very powerful technique for dimensionality reduction.



*The grey lines represent the OLS best fit and the dotted red lines represent the lasso lines.*

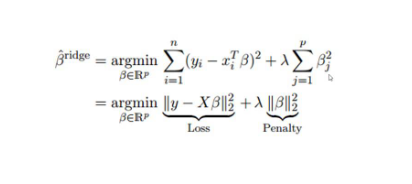
**A tuning parameter, alpha controls the strength of the L1 penalty. λ is basically the amount of shrinkage:**

* When alpha = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
* As alpha increases, more and more coefficients are set to zero and eliminated (theoretically, when alpha = ∞, all coefficients are eliminated).
* As alpha increases, bias increases.
* As alpha decreases, variance increases.

If an intercept is included in the model, it is usually left unchanged.

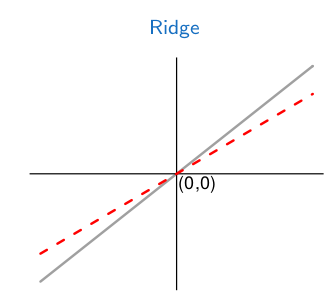
**RIDGE REGRESSION**

* Least squares regression isn’t defined at all when the number of predictors exceeds the number of observations; It doesn’t differentiate “important” from “less-important” predictors in a model, so it includes all of them.
* This leads to [overfitting](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/#overfitting)a model and failure to find unique solutions. Least squares also has issues dealing with multicollinearity in data.
* **Ridge regression avoids all of these problems.** It works in part because it doesn’t require [unbiased estimators](https://www.statisticshowto.com/unbiased/#UE); While least squares produces unbiased estimates, [variances](https://www.statisticshowto.com/probability-and-statistics/variance/)can be so large that they may be wholly inaccurate.
* Ridge regression adds just enough [bias](https://www.statisticshowto.com/what-is-bias/)to make the estimates reasonably [reliable](https://www.statisticshowto.com/reliability-validity-definitions-examples/)approximations to true population values.
* Ridge regression uses a type of [shrinkage estimator](https://www.statisticshowto.com/shrinkage-estimator/) called a ridge estimator. Shrinkage estimators theoretically produce new [estimators](https://www.statisticshowto.com/estimator/)that are shrunk closer to the “true” population parameters. The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

****

## Regularization

* Ridge regression belongs a class of regression tools that use [L2 regularization](https://www.statisticshowto.com/regularization/).
* The other type of regularization, **L1 regularization**, limits the size of the coefficients by adding an L1 penalty equal to the [absolute value](https://www.statisticshowto.com/integer/#abs) of the magnitude of coefficients.
* This sometimes results in the elimination of some coefficients altogether, which can yield sparse models.**L2 regularization**adds an L2 penalty, which equals the square of the magnitude of coefficients.
* All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.
* A [tuning parameter](https://www.statisticshowto.com/tuning-parameter/) (λ) controls the strength of the penalty term. When λ = 0, ridge regression equals least squares regression.
* If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

****

*The grey lines represent the OLS best fit and the dotted red lines represent the ridge lines.*