$$\sum[(y_i - \hat{y}_i)^2] + \lambda \sum_{k=1}^{k} y_i^2$$

Ridge Regression is used to find which of the coefficients had the biggest effect on the observed variable

- **Σ**: A greek symbol that means sum
- **y_i**: The **actual** response value for the i^{th} observation
- **$\hat{y}_i$**: The **predicted** response value
- *λ is the tuning parameter* $\lambda \geq [0, \infty]$
- **Σ**: Ridge penalty

Also note that ridge regression requires the data to be standardized such that each predictor variable has a mean of 0 and a standard deviation of 1.

*least squares regression tries to find coefficient estimates that minimize the sum of squared residuals (RSS) (https://www.statology.org/ridge-regression-in-r/)

As the tuning parameter gets larger the value of the coefficients shrink towards 0.

$$\hat{\beta}(\lambda) = \left(X^T X + \lambda I_{p\times p}\right)^{-1} X_T Y,$$

$$y = f(x_1, x_2, x_3 \ldots x_n)$$

To put it simply we want to know which predictor (x_i) had the greatest affect on the dependent variable y.

[elements of statistical learning Chapter 3.4.1 Shrinkage Methods:]

on the left hand slide is the minimization function. We are trying to best fit the model to the data,

and on the right is the ridge penalty function Ridge regression shrinks the coefficients by imposing a penalty on their size.

$\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage.

In Neural Networks this is known as weight decay.

By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error rate than a full model.

Linear Algebra is required to work through the proofs of Ridge Regression.