S21CSE471: Statistical Methods in AI **Spring 2021**

LINEAR REGRESSION

Prepared by:Likhitha Gaddi, Sai Vishwak Gangam, Rahul Valluri

1 Understanding the problem

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$y = c + b * x$$

where y = estimated dependent variable score. c = constant, b = regression coefficient, x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regress and. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

- First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.
- Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional sales income do we get for each additional Rs.1000 spent on marketing?"
- Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

2 Introduction

Given a data set

$$\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}{}^{n}$$

of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where T denotes the transpose, so that ${x_i}^T$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in matrix notation as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Some remarks on notation and terminology:

- $\mathbf{y}$ is a vector of observed values y_i $(i = 1, \ldots, n)$ of the variable called the dependent variable. This variable is also sometimes known as the predicted variable, but this should not be confused with predicted values, which are denoted $\hat{y}$. The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables. Alternatively, there may be an operational reason to model one of the variables in terms of the others, in which case there need be no presumption of causality.
- X may be seen as a matrix of row-vectors $\mathbf{x}_i$ or of n-dimensional column-vectors X_j which are known as regressors, exogenous variables, explanatory variables, covariates, input variables, predictor variables, or independent variables (not to be confused with the concept of independent random variables). The matrix X is sometimes called the design matrix.
- Usually a constant is included as one of the regressors. In particular, $\mathbf{x}_{i0} = 1$ for $i = 1, \ldots, n$. The corresponding element of is called the intercept. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.
- Sometimes one of the regressors can be a non-linear function of another regressor or of the data, as in polynomial regression and segmented regression. The model remains linear as long as it is linear in the parameter vector .
- The values x_{ij} may be viewed as either observed values of random variables X_j or as fixed values chosen prior to observing the dependent variable. Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.
- $\boldsymbol{\beta}$ is a $(p + 1)$ -dimensional parameter vector, where β_0 is the intercept term (if one is included in the model—otherwise $\boldsymbol{\beta}$ is p-dimensional). Its elements are known as effects or regression coefficients (although the latter term is sometimes reserved for the estimated effects). Statistical estimation and inference in linear regression focuses on . The elements of this parameter vector are interpreted as the partial derivatives of the dependent variable with respect to the various independent variables.
- $\boldsymbol{\varepsilon}$ is a vector of values ε_i. This part of the model is called the error term, disturbance term, or sometimes noise (in contrast with the "signal" provided by the rest of the model). This variable captures all other factors which influence the dependent variable y other than the regressors x. The relationship between the error term and the regressors, for example their

correlation, is a crucial consideration in formulating a linear regression model, as it will determine the appropriate estimation method.

Fitting a linear model to a given data set usually requires estimating the regression coefficients $\boldsymbol{\beta}$ such that the error term $\boldsymbol{\varepsilon} = \mathbf{y} - X\boldsymbol{\beta}$ is minimized. For example, it is common to use the sum of squared errors $||\boldsymbol{\varepsilon}||$ as the quality of the fit.

3 Types of Linear Regression

- Simple linear regression 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
- Multiple linear regression : 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
- Logistic regression : 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
- Ordinal regression : 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
- Multinomial regression : 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
- Discriminant analysis : 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam's razor describes the problem extremely well – a simple model is usually preferable to a more complex model.

4 Assumptions

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. Numerous extensions have been developed that allow each of these assumptions to be relaxed (i.e. reduced to a weaker form), and in some cases eliminated entirely. Generally these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model.

The following are the major assumptions made by standard linear regression models with standard estimation techniques:

- ***Weakexogeneity*** : This essentially means that the predictor variables x can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models.
- ***Linearity*** : This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values (see above), linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This technique is used, for example, in polynomial regression, which uses linear regression to fit the response variable as an arbitrary polynomial function (up to a given rank) of a predictor variable. With this much flexibility, models such as polynomial regression often have "too much power", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression(discussed later in the paper).

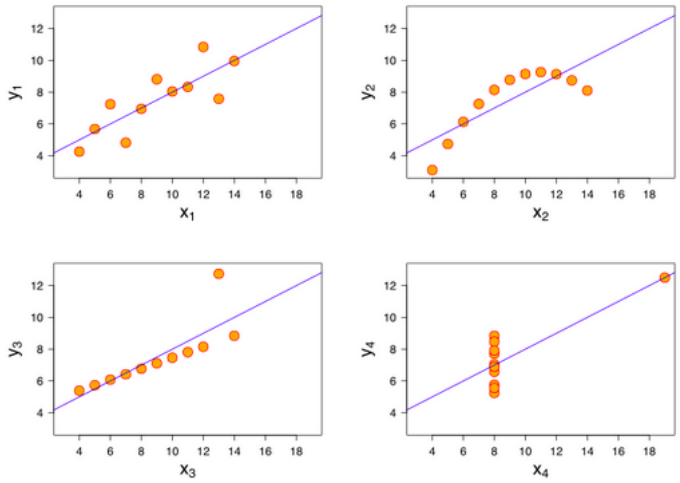
- ***Constant variance*** : This means that different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables. In practice this assumption is invalid if the response variable can vary over a wide scale. In order to check for heterogeneous error variance, or when a pattern of residuals violates model assumptions of homoscedasticity (error is equally variable around the 'best-fitting line' for all points of x), it is prudent to look for a "fanning effect" between residual error and predicted values. This is to say there will be a systematic change in the absolute or squared residuals when plotted against the predictive variables. Errors will not be evenly distributed across the regression line. Heteroscedasticity will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong. Typically, for example, a response variable whose mean is large will have a greater variance than one whose mean is small. For example, a given person whose income is predicted to be Rs.100,000 may easily have an actual income of Rs.80,000 or Rs.120,000 (a standard deviation of around Rs.20,000), while another person with a predicted income of Rs.10,000 is unlikely to have the same Rs.20,000 standard deviation, which would imply their actual income would vary anywhere between -Rs.10,000 and Rs.30,000. (In fact, as this shows, in many cases—often the same cases where the assumption of normally distributed errors fails—the variance or standard deviation should be predicted to be proportional to the mean, rather than constant.) Simple linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity(constant variance) is present.
- ***Independence*** of errors. This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.) Some methods (e.g. generalized least squares) are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.
- ***Lack of perfect multicollinearity*** in the predictors. For standard least squares estimation methods, the design matrix X must have full column rank p; otherwise, we have a condition known as perfect multicollinearity in the predictor variables. This can be triggered by having two or more perfectly correlated predictor variables (e.g. if the same predictor variable is mistakenly given twice, either without transforming one of the copies or by transforming one of the copies linearly). It can also happen if there is too little data available compared to the number of parameters to be estimated (e.g. fewer data points than regression coefficients). In the case of perfect multicollinearity, the parameter vector will be non-identifiable—it has no unique solution. At most we will be able to identify some of the parameters, i.e. narrow down its value to some linear subspace of R^p. See partial least squares regression. Methods for fitting linear models with multicollinearity have been developed; some require additional assumptions such as "effect sparsity"—that a large fraction of the effects are exactly zero. Note that the more computationally expensive iterated algorithms for parameter estimation, such as those used in generalized linear models, do not suffer from this problem.

Beyond these assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
- The arrangement, or probability distribution of the predictor variables x has a major influence on the precision of estimates of . Sampling and design of experiments are highly developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of .

5 Interpretation

A fitted linear regression model can be used to identify the relationship between a single predictor variable x_j and the response variable y when all the other predictor variables in the model are "held fixed".



Specifically, the interpretation of $_j$ is the expected change in y for a one-unit change in x_j when the other covariates are held fixed—that is, the expected value of the partial derivative of y with respect to x_j. This is sometimes called the unique effect of x_j on y. In contrast, the marginal effect of x_j on y can be assessed using a correlation coefficient or simple linear regression model relating only x_j to y; this effect is the total derivative of y with respect to x_j.

Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes (such as dummy variables, or the intercept term), while others cannot be held fixed (recall the example from the introduction: it would be impossible to "hold t_i fixed" and at the same time change the value of ${t_i}^2$).

It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in x_j, so that once that variable is in the model, there is no contribution of x_j to the variation in y. Conversely, the unique effect of x_j can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by x_j. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to x_j, thereby strengthening the apparent relationship with x_j.

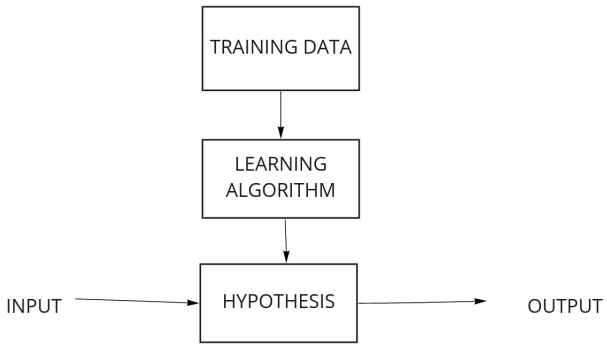
The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study.

The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design. Commonality analysis

may be helpful in disentangling the shared and unique impacts of correlated independent variables.

6 COST FUNCTION

The working of the supervised learning algorithm is shown below:



The equation of hypothesis looks like the following:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

The above equation in general is termed as :

$$h_\theta(x) = \sum_{j=0}^{n} \theta_j x_j; x_0 = 1$$

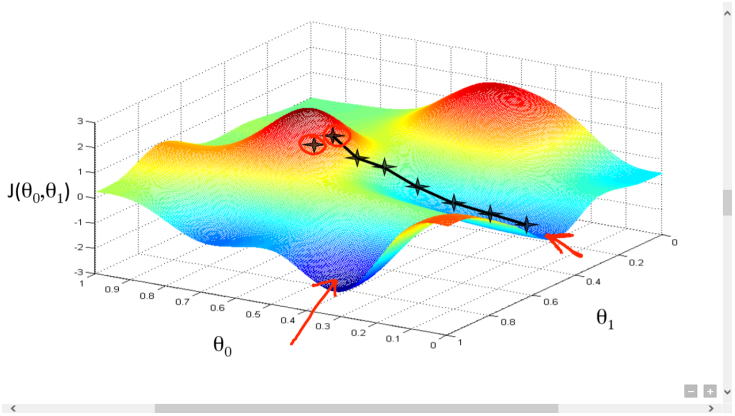
Here θ is n dimensional column vector and x is also n dimensional column vector
Some important terminology is as follows:
θ is called parameters of the learning algorithm
m is called training samples
x is called features/input
y is called output/prediction value
The learning algorithm job is to choose the parameters θ such that it can output accurate value(predicted value) hypothesis.

$$j(x) = 1/2 \sum_{i=0}^{n} (\, h(x^i) - y^i)^2$$

Here above j(x) is the cost function which is used to calculate the squared error between predicted value and actual value and now we need to choose the learning parameter θ in such a way that j(x) is minimized.

7 Gradient Descent

This algorithm is used to find the parameter value θ such that j(x) which is the cost function is minimum.



The first step of this algorithm is to:
1. Take any value of θ(for example a n-dimensional zero vector) and keep changing θ to reduce j(θ).
2. By seeing the above diagram it can be observed in wherever the position you are look at your sideways and go to deepest path because it reduces the j(θ) for small change in θ value.

$$\theta_j := \theta_j - \alpha \frac{\partial(j(\theta))}{\partial\theta}$$

Here α is the learning rate of the algorithm
Now we will be working on calculating $\frac{\partial(j(\theta))}{\partial\theta}$ =

$$\frac{\partial(\,1/2(\,h_\theta(\,x)\,-y\,)^2)}{\partial\theta_j}$$

$$= (h_\theta(\,x)\,-y)\frac{\partial(\,h_\theta(\,x)\,-y)}{\partial\theta_j}$$

$$= (h_\theta(\,x)\,-y)\frac{\partial(\,\theta_0 x_0 + \theta_1 x_1 + ...\theta_n x_n - y)}{\partial\theta_j}$$

$$= (h_\theta(\,x)\,-y)x_j$$

Now the gradient descent equation is:

$$\theta_j := \theta_j - \sum_{i=1}^{n}(h_\theta(\,x^i)\,-y^i)x_j^i$$

7.1 Types Of Gradient Descent

1. **Batch Gradient Descent**

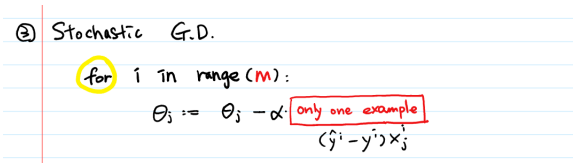
In Batch Gradient Descent, all the training data is taken into consideration to take a single step. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters.
The above equation is an example of batch gradient descent.

$$\theta_j := \theta_j - \sum_{i=1}^{n}(h_\theta(\,x^i)\,-y^i)x_j^i$$

Batch Gradient Descent is great for convex or relatively smooth error manifolds. In this case, we move somewhat directly towards an optimum solution.
-The main disadvantage of batch gradient descent is that it is very slow and it is not feasible for real world data.

2. Stochastic Gradient Descent

Stochastic gradient descent is just like Batch gradient descent except that it looks at only one training sample for each step.



Stochastic gradient descent is useful when there is lot of data and parameters.

8 NORMAL EQUATION

It is not always necessary to run an optimization algorithm to perform linear regression. You can solve a specific algebraic equation ,the normal equation — to get the results directly.

$$J(\theta) = (X\theta - y)^T (X\theta - y)$$

Here X is the design matrix containing m rows with features of the data set .
Now we wil simplify the above equation using matrix transpose identities

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y)$$

$$J(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T (X\theta) + y^T y$$

$$J(\theta) = \theta^T X^T X\theta - 2(X\theta)^T y + y^T y$$

Here parameter θ is unknown so to find out the where the cost function will be minimum we will derive the cost function with θ .

$$\frac{\partial J}{\partial \theta} = 2X^T X\theta - 2X^T y = 0$$

$$X^T X\theta = X^T y$$

Now, assuming that the matrix $X^T X$ is invertible, we can multiply both sides by $(X^T X)^{-1}$ and get:

$$\theta = (X^T X)^{-1} X^T y$$

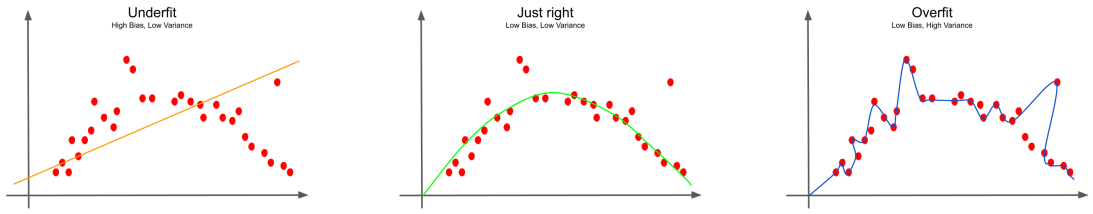
The above equation is known as the normal equation

9 Regularization

Until now, our main aim has been to minimize cost function

$$J(w) := \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

But this definition of cost function has the chance of over-fitting the training data if not careful. This happens because the gradient descent algorithm optimises for minimum difference between model output and dataset labels which force the w values to alter in a way that it forms the just the right line equation for training data. Every feature is given high importance to not miss any nuances of training data which leads to w values being high.



To negate this, we add a new term to the cost function with a new hyperparameter λ . The new cost function we take is:

$$J(w) := \sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda ||w||_p$$

Here, λ is the regularization hyperparameter which maintains the tradeoff between the error term and weights term. The notation $||w_j||_p$ stands for norm of weights. It expands as

$$||w||_p := \left(\sum_{j=1}^{d} |w_j|^p\right)^{\frac{1}{p}}$$

If we start overfitting, the weights become larger. With a high λ value, that is negated as our gradient descent algorithm tries to minimise the cost function. So, if weights grow larger, then the newly added norms of weights grows larger in turn making next iteration weights update to values that are far from global minima. After next iteration, the weights are rectified to 'just the right' values so that they aren't too big to throw itself off global minima and not too small that it gives large error.

In essence one can say that the error term is used to measure the loss of model and the regularization term is used to measure the complexity of model. So, when we put both these terms together in cost function and trying to minimise it, we are actually trying to find the simplest model that is giving the least error.

One thing to note here is that we don't consider the intercept value or bias(represented generally by w_0) while considering the norms because the *regularization* of the bias is taken care of by the error term. When the bias term grows too large, the error term high negative error which throws it off farther away from global minima, which as usual is rectified in next iteration.

9.1 Lasso and Ridge Regression

Lasso and Ridge regression are kinds of linear regression which differ in their regularization terms. For lasso regression, the p value in $||w||p$ is 1. This is called L1 regularization. For ridge regression, the p value is 2 making it being called L2 regularization. So, the formulae for lasso and ridge regression cost functions respectively are:

$$J(w)_{lasso} := \sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda \sum_{j=1}^{d} |w_j|$$

$$J(w)_{ridge} = \sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda \sum_{j=1}^{d} |w_j|^2$$

While ridge regression works more or less like any standard regression, lasso does something interesting. It acts as a feature selector. Let us understand why this is the case.

When we get the derivative of lasso regression cost function as follows

$$J'(w) := \sum_{i=1}^{n} 2x_i(w^T x_i - y_i) + \lambda \sum_{j=1}^{d} \begin{cases} 1, & w_j > 0 \\ -1, & w_j < 0 \end{cases}$$

In case, it is not easy to follow the derivative on regularization term,

$$\frac{\delta|w|}{\delta w} = \begin{cases} 1, & w > 0, \\ -1, & w < 0 \end{cases}$$

And the final weight update formula would look as follows:

$$w_{t+1} := w_t - \alpha \left(\sum_{i=1}^{n} 2x_i(w^T x_i - y_i) + \lambda \sum_{j=1}^{d} \begin{cases} 1, & w_j > 0 \\ -1, & w_j < 0 \end{cases} \right)$$

One can observe from cost derivative that the regularization term increases when more features are greater than zero. So when updating, it subtracts itself from λ. In case $w_j < 0$, it takes $-\lambda$ as regularization term and adds it to w_j. In a way, regularization term in lasso regression works towards making the weight zero. And when it's made zero, it stops regularizing it since we have no condition for $w_j = 0$.

So, while ridge regression tries to penalize weights towards zero, lasso can actually make them zero completely. This makes it work as some form of filter on features by removing the ones that don't contribute to the final value. In fact, lasso is short for Least Absolute Shrinkage and Selection Operator.

Figure 1: Plot showing the intersection of regularization term and loss term in lasso and ridge regression. In lasso, the intersection between diamond contour and elliptical contours mostly occurs on one of the axis, making a parameter null or zero. In ridge, the intersection of circular contour and elliptical contour could occur anywhere on the outer circle.

9.2 Elastic Net Regression

Lasso regression helps in eliminating unnecessary features while ridge regression helps in shrinking the existing features. Elastic Net combines these two provide a robust solution to overfitting problem. The regularization term consists of two terms - L1 and L2 regularization. The cost function will be as follows:

$$J(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda_1 ||w||_1 + \lambda_2 ||w||_2$$

λ_1 controls the penalization on L1 norm while λ_2 controls the penalization on L2 norm.

10 Where Linear regression can and cannot be used

The sensible use of linear regression on a data set requires that four assumptions about that data set be true:

- The relationship between the variables is linear.
- The data is homoskedastic, meaning the variance in the residuals (the difference in the real and predicted values) is more or less constant.

- The residuals are independent, meaning the residuals are distributed randomly and not influenced by the residuals in previous observations. If the residuals are not independent of each other, they're considered to be autocorrelated.
- The residuals are normally distributed. This assumption means the probability density function of the residual values is normally distributed at each x value. I leave this assumption for last because I don't consider it to be a hard requirement for the use of linear regression, although if this isn't true, some manipulations must be made to the model.

Some considerations the business analyst will want to take when using linear regression for prediction and forecasting are:

- ***Scope*** : A linear regression equation, even when the assumptions identified above are met, describes the relationship between two variables over the range of values tested against in the data set. Extrapolating a linear regression equation out past the maximum value of the data set is not advisable.
- ***Spuriousrelationships*** : A very strong linear relationship may exist between two variables that are intuitively not at all related. The urge to identify relationships in the business analyst is strong; take pains to avoid regressing variables unless there exists some realistic reason they might influence each other.

11 REFERENCES

https://towardsdatascience.com/l1-and-l2-regularization-explained-874c3b03f668
https://towardsdatascience.com/l1-and-l2-regularization-explained-874c3b03f668
https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-
https://vigneshmadanan.medium.com/linear-regression-basics-and-regularization-methods-b40359b0aea
https://medium.com/@zxr.nju/the-classical-linear-regression-model-is-good-why-do-we-need-regulari
https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261
https://online.stat.psu.edu/stat508/lesson/5/5.4
https://towardsdatascience.com/difference-between-batch-gradient-descent-and-stochastic-gradient-
https://www.youtube.com/results?search_query=statquest+gradient+descent
https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a