LINEAR REGRESSION

Linear regression models assume a linear relationship between the inputs $X_1, X_2, \ ..., \ X_p$ and the output Y. These models are simple and often provide an insight into the effect of input variables on the output variable. Linear models can be expanded to transformations of the input variables, thus making them widely applicable.

1 Single Variable Regression

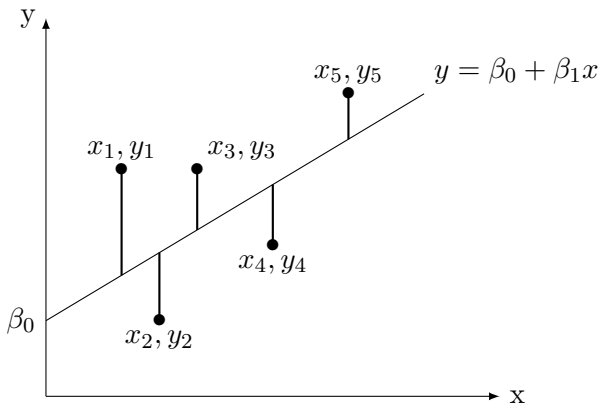
Given training data of the form $(x_1, y_1), (x_2, y_2), \ ..., \ (x_N, y_N)$; the idea is to come up with the best estimate $\hat{y}$ such that,

$$\hat{y} = \beta_0 + \beta_1 x$$

This scheme makes an estimation error at each point,

$$\begin{aligned} \epsilon_i &= y_i - \hat{y}_i \\ &= y_i - (\beta_0 + \beta_1 x) \end{aligned}$$

Generally, sum of squares of all the N errors is minimized to define the best fit. Thus, the objective is to choose β_0 and β_1 such that this Residual Square Sum (RSS) is as small as possible. The errors are visualized below,



Note that as line moves away from the data points the RSS gets bigger since each error becomes bigger. In fact there is no upper bound on RSS, only a lower bound. Hence to get the minimum RSS, equating partial differentials w.r.t. β_0 and β_1 to zero is sufficient without worrying if the extrema is maxima or minima, it is always guaranteed to be a minima.

$$
\begin{aligned}
RSS(\beta_0, \beta_1) &= \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \\
\frac{\partial RSS}{\partial \beta_0} &= -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) \\
\frac{\partial RSS}{\partial \beta_1} &= -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) x_i
\end{aligned}
$$

Equating both to zero,

$$
\begin{aligned}
\beta_0 N + \beta_1 \sum x_i &= \sum y_i \\
\beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i
\end{aligned}
$$

This can be written in matrix form as follows,

$$
\begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}
$$

Multiplying by inverse of the 2×2 matrix on both sides,

$$
\begin{aligned}
\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} &= \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{N \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\
&= \frac{1}{N \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i \\ N \sum x_i y_i - \sum x_i \sum y_i \end{pmatrix}
\end{aligned}
$$

Dividing both numerator and denominator by N^2, β_0 and β_1 can be represented in terms of averages as follows,

$$\beta_0 = \frac{Avg(x^2)Avg(y) - Avg(xy)Avg(x)}{Avg(x^2) - [Avg(x)]^2}$$
$$\beta_1 = \frac{Avg(xy) - Avg(x)Avg(y)}{Avg(x^2) - [Avg(x)]^2} = \frac{Cov(x,y)}{Var(x)}$$

After the calculation of β_0 and β_1, the goodness of fit is measured by correlation,

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

Correlation always resides in the interval $[-1, 1]$. It is 1 or -1 if all the data points (x_i, y_i) lie on a line.

1.1 Input Transformations

- Relationship between x and y is not linear if $y = x^n$. Taking log on both sides, $log(y) = nlog(x)$. Clearly relationship between $log(x)$ and $log(y)$ is linear.
- Relationship between x and y is not linear if $y = a^x$. Taking log on both sides, $log(y) = xlog(a)$. Clearly, relationship between $log(y)$ and x is linear.
- Due to scenarios like this, the inputs are sometimes transformed appropriately before being fed into a linear regression model.

2 Multiple Variable Regression

In the general case, the equation becomes

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Let $\mathbf{X}$ be $N \times (p+1)$ matrix where the first column is all ones and the j^{th} column represents the N values of the input variable x_j. Similarly, let $\mathbf{Y}$ be a vector of the N values of the output variable y. For all the N given observations, it can be written,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

The error vector becomes,

$$\epsilon = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

The residual sum of squares can be represented as,

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \epsilon^T \epsilon \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}^T - \boldsymbol{\beta}^T \mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Taking the derivative w.r.t. $\boldsymbol{\beta}$

$$\begin{aligned} \frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y}^T\mathbf{y}) - \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}) - \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}) + \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) \\ &= 0 - \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Equating the derivative to zero for minimization,

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

These estimators are also known as OLS(Ordinary Least Square) estimators.

3 Bias and Variance

3.1 Definitions

Note that the expectation of a vector $\mathbf{V}$ is defined as,

$$\mathbb{E}[\mathbf{V}] = \begin{pmatrix} \mathbb{E}(v_1) \\ \mathbb{E}(v_2) \\ . \\ . \\ . \\ \mathbb{E}(v_n) \end{pmatrix}$$

The variance of a vector is a little trickier,

$$Var[\mathbf{V}] = \begin{pmatrix} Var(v_1) & Cov(v_1, v_2) & . & . & . & Cov(v_1, v_n) \\ Cov(v_2, v_1) & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ Cov(v_n, v_1) & . & . & . & . & Var(v_n, v_n) \end{pmatrix}$$

3.2 Bias

Let the true population process of $\mathbf{y}$ be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\boldsymbol{\beta}$ is the true population variable and $\mathbf{u}$ is the population error. Least squares estimate as derived above is,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u} \end{aligned}$$

Taking expectation on both sides,

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[\boldsymbol{\beta}] + \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}] \\ &= \boldsymbol{\beta} + \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}] \end{aligned}$$

Assuming $\mathbf{u}$ and $\mathbf{X}$ are independent, the equation can be simplified to,

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{u}]$$

Assuming $\mathbb{E}[\mathbf{u}] = 0$, the equation gets further simplified,

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Under the conditional independence zero mean assumption [1], the least squares estimator is unbiased and on average the estimate is equivalent to the true population parameter.

3.3 Variance

Once again start with the least squares estimate equation,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Taking variance on both sides,

$$Var[\hat{\boldsymbol{\beta}}] = Var[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]$$

Since the variance of the product of a non-stochastic matrix $\mathbf{A}$ and a vector $\mathbf{V}$ is $\mathbf{A}Var(\mathbf{V})\mathbf{A}^T$,

$$\begin{aligned} Var[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ Var[\mathbf{y}]\ ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ Var[\mathbf{y}]\ \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

Assume that $Var[\mathbf{v}] = \sigma^2\mathbf{I}$. This is same as saying that,

- All $y_i's$ are homoscedastic i.e. $Var(y_i) = \sigma^2\ \forall\ i$ [2].
- Distinct $y_i's$ are uncorrelated i.e. $Cov(y_i, y_j) = 0\ \forall\ i \neq j$ [3].

$$Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

4 Gauss-Markov Theorem

This theorem states that under above assumptions [1], [2] and [3], the OLS estimators are BLUE (Best Linear Unbiased Estimators). This means that there are no other unbiased linear estimators which have a lower variance of $\boldsymbol{\beta}$ than OLS.

5 Ridge Regression

A penalty proportional to sum of squares of all β coefficients is added to the residual sum of squares in ridge regression to make sure that coefficients are reasonably bounded and don't grow arbitarily large.

$$\mathbf{Error}(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

The solution to the above ridge regression problem is given below,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

As λ tends to zero, the ridge solution tends to the OLS solution and as λ tends to infinity, the ridge solution tends to zero because of the infinite penalty for non-zero coefficients.

If $\mathbf{X}^T\mathbf{X}$ is non-invertible, OLS has no unique solution, but this problem does not occur with ridge regression as the matrix $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is always guaranteed to be invertible, therfore always ensuring unique solution.

The fundamental idea with ridge regression is to incur some bias in order to reduce variance which may often be desirable in practical situations.

Now, a short excursion on Lagrange multipliers before the introduction to Lasso regression.

6 Lagrange Multipliers

Suppose a function f is to be minimized subject to the constraint that the value of another function g on the same inputs is t. For illustration, consider two inputs x and y with $f(x, y) = x^2 + y^2$, $g(x, y) = 4x + 3y$ and $t = 25$. Hence the problem is posed as follows:

$$\begin{aligned} &minimize \ \&x^2 + y^2 \\ &subject \ to \ \&4x + 3y = 25 \end{aligned}$$

$f(x, y) = x^2 + y^2$ has a circular countour centered at origin. All points on this circle have the same value of f. On the ohter hand, the curve $4x + 3y = 25$ is a line offset from the origin.

One can imagine an expanding circle at origin which finally hits the line. The point where the circle hits the line is the solution as it is the minimum value of f satisfying the constraint $g = 25$.

At this intersection of the circle and the line, their gradients must be aligned to the same direction. More precisely, for some scalar λ,

$$\begin{pmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \end{pmatrix} = \lambda \begin{pmatrix} \frac{\partial g}{\partial x} \\ \\ \frac{\partial g}{\partial y} \end{pmatrix}$$

Note that if one were to minimize $f - \lambda g$, taking the partials and equating them to zero would have have led to the same equation above. Hence it makes no difference to pose the same problem differently as follows,

$$\begin{aligned} &minimize\ \& f - \lambda g \\ &\qquad\qquad\qquad for\ some\ scalar\ \lambda \end{aligned}$$

The function $f - \lambda g$ is called the Langrangian function and λ the Langragial multiplier.

7 Lasso Regression

Instead of ridge regression's penalty which is proportional to the sum of squares of the $\boldsymbol{\beta}$s, Lasso penalizes the sum of moduli of the $\boldsymbol{\beta}$s.

$$\mathbf{Error}(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \lambda \sum |\boldsymbol{\beta}_i|$$

From the discussion of Lagrange mutlipliers, the same problem can be posed as,

$$\begin{aligned} &minimize\ \& \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ &\qquad\qquad\qquad subject\ to\ \sum |\boldsymbol{\beta}_i| \leq t \end{aligned}$$

The graphs of $\sum |\boldsymbol{\beta}_i| \leq t$ are right angle rotated squares centered at origin in two dimensions.

As before, imagine the expanding contours of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ till they hit the constraint graph to deliver the solution. Since the constraint graph has corners, it is highly likely that the expanding contour of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ will hit one of the corners. At that corner, one or more of the $\boldsymbol{\beta}$s will be zero.

Hence, Lasso not only shrinks the $\boldsymbol{\beta}$s but also makes some of them zero introducing sparsity in the solution and only retains only the most important $\boldsymbol{\beta}$s. This feature makes it considerably varied from the ridge regression.