Linear Regression

Linear regression is the cornerstone of modern statistics by which a descriptive analyst fits a straight line to data. Let (x_i, y_i), for $i = 1, \cdots n$ be the data points and a and a_0 the coefficients of the regression line $\hat{y} = ax + a_0$. Here x is the vector of features (also known as independent variables, covariates, or explanatory variables) and a the vector of coefficients. The coefficient a_0 is a constant.

The key question is how to estimate the coefficients. The best line should perhaps minimize the prediction error, $y - \hat{y}$. However, this can be done in different ways each with its advantages and limitations which is why to one task in machine learning (such as classification, clustering, etc) there are different algorithms.

Dataset

We will use multiple linear regression to estimate the stock index (as the response variable or dependent variable) of a fictitious economy by using two features:

- Interest Rate
- Unemployment Rate

Therefore, the linear regression model is:

$$\text{Stock Index} = a_1 \text{Interest Rate} + a_2 \text{Unemployment Rate} + a_0.$$

- **Minimize the sum of absolute deviations:** Another way of thinking about minimizing the prediction errors is to minimize the sum of absolute errors:

$$\min_{a_0, a_1, a_2} \sum_{i=1}^{n} |y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0)| \tag{2}$$

This is an unconstrained nonlinear program too! But one can "linearize" this program which makes it possible to use a linear programming algorithm to estimate the coefficients. To do so, define $z_i = |y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0)|$ for $i = 1, \cdots n$. So now we can rewrite the original program in (2), as

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z_i = |y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0)| \qquad i = 1, \cdots n \end{aligned}$$

Although the objective function is now linear, we have nonlinear constraints. To linearize the constraints, we can substitute $z_i = |y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0)|$ with two constraints: $z_i \geq y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0)$ AND $z_i \geq -(y_i - (a_1 x_i^1 + a_2 x_i^2 + a_0))$.

Table 1: Dataset

Year	Month	Interest Rate	Unemployment Rate	Stock Index
2017	12	2.75	5.3	1464
2017	11	2.50	5.3	1394
2017	10	2.50	5.3	1357
2017	9	2.50	5.3	1293
2017	8	2.50	5.4	1256
2017	7	2.50	5.6	1254
2017	6	2.50	5.5	1234
2017	5	2.25	5.5	1195
2017	4	2.25	5.5	1159
2017	3	2.25	5.6	1167
2017	2	2.00	5.7	1130
2017	1	2.00	5.9	1075
2016	12	2.00	6.0	1047
2016	11	1.75	5.9	965
2016	10	1.75	5.8	943
2016	9	1.75	6.1	958
2016	8	1.75	6.2	971
2016	7	1.75	6.1	949
2016	6	1.75	6.1	884
2016	5	1.75	6.1	866
2016	4	1.75	5.9	876
2016	3	1.75	6.2	822
2016	2	1.75	6.2	704
2016	1	1.75	6.1	719