5. Linear Regression

5.1 Introduction

The objective of regression is to determine the value of one or more continuous target variables (t) based on the d-dimensional input vector x. Given a set of N observations x_n, where $n = 1, 2, \ldots, n$, together with the target values y_n, the objective is to predict the value of y for an input x. This scenario can be represented as $y = f(x) + \varepsilon$, i.e. the target variable is a function of input variables x. Uncertainty always exists in the process of determining the value of the target variable and this motivates to represent this as a probability distribution $p(y|x)$.
The simplest form of linear regression is the linear combination of input variables and other useful regression functions linearly combines a fixed set of linear functions of the input variables called basis functions and forms complex functions.

5.2 Linear Model

The simple regression model involves a linear combination of the input variables

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_N x_N \tag{5.1}$$

Where $x_i \in R^d$ and β_i are the parameters of the models. β_0 is called intercept and $\beta_1, \beta_1, \ldots, \beta_N$ are called coefficients of the model. This is referred as linear regression.

$$y = \beta_0 + \sum_{i=1}^{N} \beta_i x_i = X^T \beta \tag{5.2}$$

Therefore, $y = f(x) + \varepsilon$, Where $\varepsilon \sim \mathbb{N}(0, \sigma^2)$, independent, identically distributed(iid) Gaussian nosie with mean 0 and variance σ^2 can be considered as noise or uncertainty. The probabilistic approach of linear regression is to determine the value of the target variable given by the values of the input variables.

$$p(y|x) = p(y|x, \beta) = p(y|x^T \beta, \sigma^2) \tag{5.3}$$

The probability density function of y evaluated at $x^T \beta$ is the likelihood function. Our objective is to estimate the parameters by maximizing the likelihood.

5.3 Simple Linear Regression

Simple regression is designed for a single input or predictor variable and a single output or response variable. The simple linear regression assumes that there exists a linear relationship between X and Y. It can be written as

$$Y = \beta_0 + \beta_1 X \tag{5.4}$$

For example, X may represent income and Y may represent expense. then the simple regression

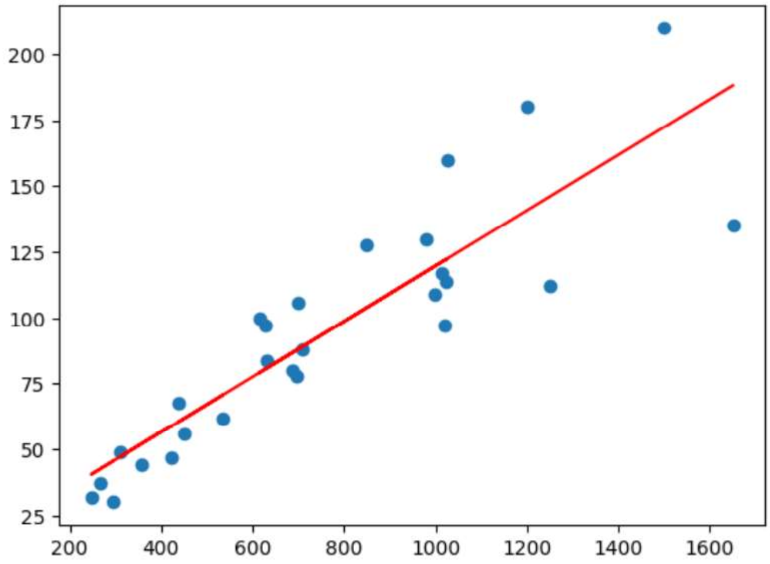

Figure 5.1: Simple Linear Regression

says that expense is regressed onto income by fitting the model

$$Expense = \beta_0 + \beta_1 \times Income \tag{5.5}$$

β_0 and β_1 are unknowns that represent the intercept and slope for the linear model. These intercepts and the slope together are known as parameters or coefficients. We have to find out the values of these parameters using training data. Once we estimate the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ for the linear model, then we can predict the expense for another household using their income based on the linear model

$$Expense = \hat{\beta}_0 + \hat{\beta}_1 \times Income$$

5.3.1 Estimating Coefficients

Many methods exist for determining the coefficients of the linear regression equation. The ordinary least square method finds the parameters by minimizing the squared error. The sum squared error is called residual sum squared error(RSS)

$$RSS = \sum_{i=1}^{N} (Actual - Predicted)^2 \tag{5.6}$$

The linear regression model is represented as a function of input variables $f(x_i) = \beta_0 + \beta_1 x_i$, the sum squared error is

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{5.7}$$

Simple Linear Regression with intercept Only

For the simple linear regression with slope 0 and intercept only, the linear equation becomes $f(x_i) = \beta_0$, and the RSS becomes

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0)^2 \tag{5.8}$$

We have to calculate partial derivative of the RSS with respect to β_0, and equalize it to 0, to get the value of β_0

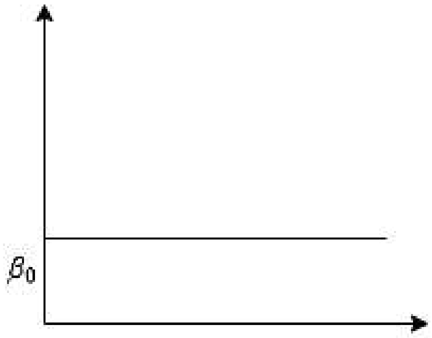

Figure 5.2: Simple linear regression with intercept only

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{N} 2(y_i - \beta_0)(-1) = 0 \tag{5.9}$$

The above equation provides the value of β_0 as follows

$$\beta_0 = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} \tag{5.10}$$

Simple Linear Regression with Slope Only

For the simple linear regression with slope 0 and intercept only, the linear equation becomes $f(x_i) = \beta_1 x_i$, and the RSS becomes

$$RSS = \sum_{i=1}^{N} (y_i - \beta_1 x_i)^2 \tag{5.11}$$

We have to calculate partial derivative of the RSS with respect to β_1, and equalize it to 0, to get the value of β_1

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{N} 2(y_i - \beta_0)(-x_i) = 0 \tag{5.12}$$

The above equation provides the value of β_0 as follows

$$\sum_{i=1}^{N} x_i y_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \Rightarrow \beta_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{5.13}$$

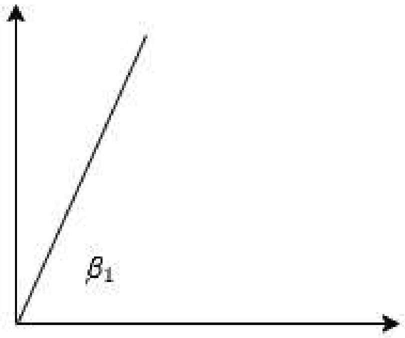

Figure 5.3: Simple Linear Regression with slope only

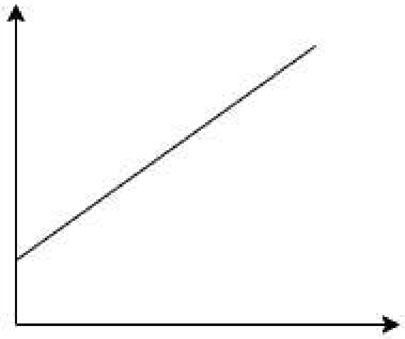

Figure 5.4: Simple Linear Regression with slope only

Simple Linear Regression with Intercept and Slope

For the simple linear regression with slope 0 and intercept only, the linear equation becomes $f(x_i) = \beta_1 x_i$, and the RSS becomes

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{5.14}$$

Take the first order derivative and make them to zero.

$$\frac{\partial RSS}{\partial \beta_0} = \sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \tag{5.15}$$

$$\frac{\partial RSS}{\partial \beta_1} = \sum_{i=1}^{N} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0 \tag{5.16}$$

The value of β_0 and β_1 can be computed from the following equations

$$\beta_0 = \frac{1}{N}(\sum_{i=1}^{N} y_i - \beta_1 \sum_{i=1}^{N} x_i) \tag{5.17}$$

$$\sum_{i=1}^{N} x_i y_i - \beta_0 - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \Rightarrow \beta_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i}{\frac{1}{N}(\sum_{i=1}^{N} x_i)^2 - \sum_{i=1}^{N} x_i^2} \tag{5.18}$$

5.4 Multiple Linear Regression

Simple regression is applicable for the single predictor variable and single response variable. When input consists of more than one predictor variable and one target variable, then the linear regression model is called multiple regression. The input X consists of $d(>1)$ input features $x_1, x_2, ..., x_d$ and N observations. The multiple regression is written as

$$f(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d \tag{5.19}$$

$$f(x_i) = \begin{bmatrix} 1 & x_1 & x_2 & ... & x_d \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_d \end{bmatrix} \tag{5.20}$$

$$Y \approx f(X) = X\mathbb{B} \tag{5.21}$$

where X is input vector and $\mathbb{B}$ is parameters or coefficients vector with size $(d+1)$. The loss function is the total error produced by the model, and can be defined as

$$\mathscr{L} = (Y - X\mathbb{B})^T (Y - X\mathbb{B}) \tag{5.22}$$

Compute first order derivative of the loss function and make it zero to find out the value of $\mathbb{B}$

$$\frac{\partial \mathscr{L}}{\partial \mathbb{B}} = -2X^T (Y - X\mathbb{B}) = 0 \tag{5.23}$$

$$X^T (Y - X\mathbb{B}) = 0 \Rightarrow \mathbb{B} = (X^T X)^{-1} X^T Y \tag{5.24}$$

where X is $N \times (d+1)$ matrix XX^T, and $(XX^T)^{-1}$ produce $(d+1) \times (d+1)$, Y is $N \times 1$ matrix, $X^T Y$ produce $(d+1) \times 1$ matrix, $(X^T X)X^T Y$ provide $(d+1) \times 1$ vector of $(d+1)$ coefficients. X is the input matrix and Y is the target or response variable. Therefore, we can compute the parameters set $\mathbb{B}$ using X and Y.

5.4.1 Maximum Likelihood Estimation

The parameter β that maximizes the likelihood is denoted here as β_{ML}, and this β_{ML} can be obtained as

$$\beta_{ML} \in argmax_\beta p(y|x, \beta) \tag{5.25}$$

As logarithm is a strictly monotonically increasing function, the optimal of a function f is identical to the optimal of $\log f$. Therefore, instead of maximizing the likelihood, apply the log transformation to the likelihood and minimize the negative log-likelihood.
The negative log-likelihood is

$$-\log p(y|x, \beta) = -\log \prod_{i=1}^{N} p(y_i|x_i, \beta) = -\sum_{i=1}^{N} \log p(y_i|x_i, \beta) \tag{5.26}$$

Due to the Gaussian additive noise term, the likelihood is Gaussian, and simplification can be done as follows

$$logp(y_i|x_i,\beta) = -\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2 + constant \tag{5.27}$$

where *constant* includes all terms independent of β.
Therefore, the negative log-likelihood can be written as

$$\begin{aligned}\mathscr{L}(\beta) &= \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 \\ &= \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) \\ &= ||X - Y\beta||^2\end{aligned} \tag{5.28}$$

Where $X = [x_1, x_2, \ldots, x_N]^T \in \mathbb{R}^{N\times d}$is the input training matrix and $y = [y_1, y_2, .., y_N]^T \in \mathbb{R}^N$is the set of target values. We have to find out the parameter β_{ML} that present the global optima of the negative log likelihood function. We find this by computing gradient of $\mathscr{L}(\beta)$ and setting it to 0 and solve the equation of β.

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial\beta} &= \frac{\partial}{\partial\beta}[\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)] \\ &= \frac{1}{2\sigma^2}(Y^TY - 2Y^TX\beta + \beta^TX^TX\beta) \\ &= \frac{1}{\sigma^2}(-Y^TX + \beta^TX^TX) \in \mathbb{R}^{1\times d}\end{aligned} \tag{5.29}$$

The maximum likelihood estimator β_{ML} solves $\frac{\partial\mathscr{L}}{\partial\beta} = 0^T$

$$\begin{aligned}&\frac{\partial\mathscr{L}}{\partial\beta} = 0^T \\ &\iff \beta^TX^TX = Y^TX \\ &\iff \beta^T = (Y^TX)(X^TX)^{-1} \\ &\iff \beta = (X^TX)X^TY\end{aligned} \tag{5.30}$$

5.5 Gradient Method

The Gradient Descent is an iterative optimization algorithm that finds the minimum of a function.

5.5.1 Gradient Descent Algorithm

1. Initialize the slope and intercept $\beta_0 = 0$ and $\beta_1 = 0$. Let $\mathscr{L}$ be the loss function and lr be the learning rate that controls the update of the parameters.
2. Calculate the partial derivative of the loss function with respect to the parameters.

$$\frac{\partial\mathscr{L}}{\partial\beta_0} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1x_i) = 0 \tag{5.31}$$

$$\frac{\partial\mathscr{L}}{\partial\beta_1} = -\frac{2}{N}\sum_{i=1}^{N}2(y_i - \beta_0 - \beta_1x_i)(x_i) = 0 \tag{5.32}$$

3. Update the value of the parameters using the following equation

$$\beta_0 = \beta_0 - lr \times \frac{\partial \mathscr{L}}{\partial \beta_0} \tag{5.33}$$

$$\beta_1 = \beta_1 - lr \times \frac{\partial \mathscr{L}}{\partial \beta_1} \tag{5.34}$$

4. Repeat this process until our loss is very small.