

Foundation of Machine Learning (CSE4032)

Lecture 03: Simple and Multiple Linear Regression

Dr. Kundan Kumar
Associate Professor
Department of ECE



Faculty of Engineering (ITER)
S'O'A Deemed to be University, Bhubaneswar, India-751030
© 2021 Kundan Kumar, All Rights Reserved

Outline

- 1 Introduction
- 2 Simple regression model
- 3 Multiple regression model
- 4 Multiple regression model from simple univariate regression
- 5 References

Introduction

- In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (X) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (X).
- When there is a **single input variable** (X), the method is referred to as **simple linear regression**; however, when there are **multiple input variables**, literature from statistics often refers to the method as **multiple linear regression**.

Simple Linear Regression Model

Simple Linear Regression

- It is a very straightforward simple linear approach for predicting a quantitative response Y on the basis of a single variable X .
- It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X$$

- **For example**, X may represent TV advertising and Y may represent sales. Then we can regress sales onto TV by fitting the model

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}.$$

- Here, β_0 and β_1 are two **unknown constants** that represent the **intercept** and **slope** terms in the linear model. Together, β_0 and β_1 are intercept-slope known as the **model coefficients** or **parameters**.

Simple regression model

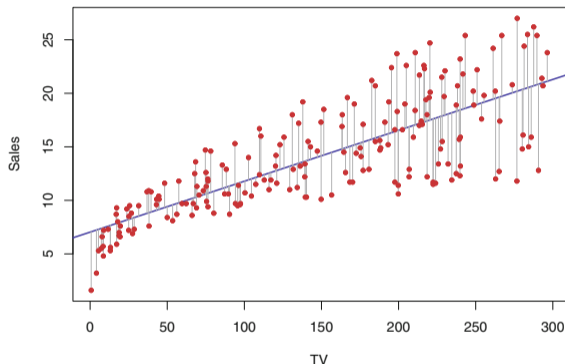
- If we can estimate the model coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, using the training data then we can predict future sales on the basis of a particular value of TV advertising by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where \hat{y} indicates a prediction of Y on the basis of $X = x$.

- **Estimating the Coefficients**
 - Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ represent n observation pairs, each of which consists of a measurement of X and a measurement of Y .
 - Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well—that is, so that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$.
- Most common approach involves minimizing the **least squares criterion**.

Simple regression model



- We define the residual sum of squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2$$

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i th value of X .
- Then $e_i = y_i - \hat{y}_i$ represents the i th residual
- This is the difference between the i th observed response value and the i th response value that is predicted by our linear model.

Simple regression model

- or equivalently as

$$\text{RSS} = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \dots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.
- Using some calculus, one can show that the minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

- Above equation defines the **least squares coefficient estimates** for **simple linear regression**.

Multiple Linear Regression Model

Multiple regression model

- A linear regression model assumes that the regression function $E(Y|X)$ is linear in the inputs X_1, \dots, X_p .
- They are simple and often provide an adequate and interpretable description of how the inputs affect the output.
- An understanding of linear methods is essential for understanding nonlinear ones.
- In fact, many nonlinear techniques are direct generalizations of the linear methods discussed here.

Multiple regression model

- Suppose, we have an input vector $X^T = (X_1, X_2, \dots, X_p)$ and want to predict a real-valued output Y . The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$$

- The linear model either assumes that the regression function $E(Y | X)$ is linear, or that the linear model is a reasonable approximation.
- Here, the β_j 's are unknown parameters or coefficients, and the variables X_j can come from different sources.

Multiple regression model

- The variable X_j can come from different sources
 - Quantitative inputs
 - Transformation of quantitative i/p, for e.g., log, square root, square, exp, etc.
 - Basic expansion, $X_2 = X_1^2$, $X_3 = X_1^3$ leading to a polynomial representation.
 - Numeric or dummy coding of the levels of qualitative i/p.
 - Interaction between variables, e.g., $X_3 = X_1 \cdot X_2$.
- No matter the source of the X_j , the model is linear in the parameters.

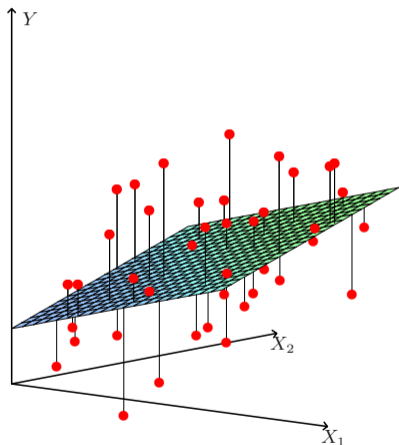
Multiple regression model

- Typically, we have a set of training data $(x_1, y_1) \dots (x_N, y_N)$ from which to estimate the parameters β .
- Each $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ is a vector of feature measurements for the i th case.
- The most popular estimation method is least squares, in which we pick the coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ to minimize the **residual sum of squares**

$$\begin{aligned} \text{RSS}(\beta) &= \sum_{i=1}^N (y_i - f(x_i))^2 \\ &= \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \end{aligned}$$

- Criteria measures the average lack of fit.

Multiple regression model



- From a statistical point of view, this criterion is reasonable if the training observations (x_i, y_i) represent independent random draws from their population. Even if the x_i 's were not drawn randomly, the criterion is still valid if the y_i 's are conditionally independent given the inputs x_i .
- How to minimize the criteria function?

$$X \rightarrow N \times (p + 1)$$

$$Y \rightarrow N \times 1$$

Multiple regression model

- Then we can write

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Differentiating w.r.t β we get

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

- Assuming (for the moment) that \mathbf{X} has **full column rank**, and hence $\mathbf{X}^T\mathbf{X}$ is positive definite, we set the first derivative to zero

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{1}$$

Multiple regression model

- Then, we obtain the unique solution

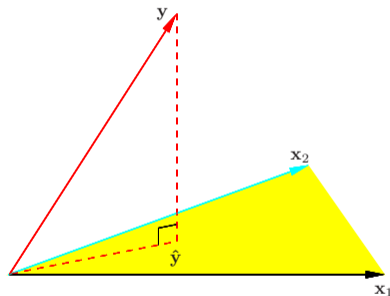
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- The predicted values at an input vector x_0 are given by $\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$; the fitted values at the training inputs are

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

where $\hat{y}_i = \hat{f}(x_i)$.

- The matrix $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is sometimes called the “hat” matrix because it puts the hat on \mathbf{y} .



Multiple regression model

- We minimize $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$ by choosing $\hat{\beta}$ so that the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to this subspace. This orthogonality is expressed in Eq. (1), and the resulting estimate $\hat{\mathbf{y}}$ is hence the orthogonal projection of \mathbf{y} onto this subspace.
- The hat matrix H computes the orthogonal projection, and hence it is also known as a projection matrix.
- If the columns of \mathbf{X} are not linearly independent, so that \mathbf{X} is not of full rank. For example, if two of the inputs were perfectly correlated, (e.g., $\mathbf{x}_2 = 3\mathbf{x}_1$). Then $\mathbf{X}^T\mathbf{X}$ is singular and the least squares coefficients $\hat{\beta}$ are not uniquely defined.
- However, the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ are still the projection of \mathbf{y} onto the column space of \mathbf{X} ; there is just more than one way to express that projection in terms of the column vectors of \mathbf{X} .

Multiple regression model

- **Rank deficiencies** can also occur in signal and image analysis, where the number of inputs p can exceed the number of training cases N . In this case, the features are typically reduced by filtering or else the fitting is controlled by regularization.
- Up to now we have made minimal assumptions about the true distribution of the data. Sampling properties of $\hat{\beta}$ assume y_i are uncorrelated and have constant variance σ^2 , and that the x_i are fixed (non random).

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

Typically one estimates the variance σ^2 by

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

Multiple regression model: drawn inference

- We assume as $Y = f(x) = \beta_0 + \sum_{i=1}^n X_j \beta_j$ is correct model, that means

$$\begin{aligned} Y &= \mathbf{E}(Y \mid X_1, \dots, X_p) + \varepsilon \\ &= \beta_0 + \sum_{j=1}^p X_j \beta_j + \varepsilon \end{aligned}$$

where the error ε is a Gaussian random variable with expectation zero and variance σ^2 , written $\varepsilon \sim N(0, \sigma^2)$.

- It is easy to show that

$$\hat{\beta} \sim N\left(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2\right)$$

Multiple Regression from Simple Univariate Regression

Multiple Regression from Simple Univariate Regression

- The linear model with $p > 1$ inputs is called the **multiple linear regression model**.
- Suppose first that we have a univariate model with no intercept, that is,

$$Y = X\beta + \epsilon$$

- The least squares estimate and residuals are

$$\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2}$$

$$r_i = y_i - x_i \hat{\beta}$$

Multiple Regression from Simple Univariate Regression

- In convenient vector notation, we let $\mathbf{y} = (y_1, \dots, y_N)^T$, $\mathbf{x} = (x_1, \dots, x_N)^T$ and define

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle &= \sum_{i=1}^N x_i y_i, \\ &= \mathbf{x}^T \mathbf{y}\end{aligned}$$

the **inner product** between \mathbf{x} and \mathbf{y} .

- Then we can write

$$\begin{aligned}\hat{\beta} &= \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \\ \mathbf{r} &= \mathbf{y} - \mathbf{x} \hat{\beta}\end{aligned}$$

This simple **univariate regression** provides the building block for **multiple linear regression**.

Multiple Regression from Simple Univariate Regression

- Suppose next that the inputs $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ (the columns of the data matrix \mathbf{X}) are **orthogonal**; that is

$$\langle \mathbf{x}_j, \mathbf{x}_k \rangle = 0 \text{ for all } j \neq k.$$

- Then it is easy to check that the multiple least squares estimates $\hat{\beta}_j$ are equal to $\langle \mathbf{x}_j, \mathbf{y} \rangle / \langle \mathbf{x}_j, \mathbf{x}_j \rangle$ - the **univariate estimates**.
- In other words, when the inputs are orthogonal, they have no effect on each other's parameter estimates in the model.
- Orthogonal inputs occur most often with balanced, designed experiments (where orthogonality is enforced), but almost never with observational data.
- Hence, **we will have to orthogonalize them in order to carry this idea further.**

Multiple Regression from Simple Univariate Regression

- Suppose next that we have an intercept and a single input x . Then the least squares coefficient of x has the form

$$\hat{\beta}_1 = \frac{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} \rangle}$$

where $\bar{x} = \sum_i x_i / N$, and $\mathbf{1} = \mathbf{x}_0$, the vector of N ones.

- We can view the estimate of $\hat{\beta}_1$ as the **result of two applications of the simple regression**. The steps are:
 1. regress \mathbf{x} on $\mathbf{1}$ to produce the residual $\mathbf{z} = \mathbf{x} - \bar{x}\mathbf{1}$;
 2. regress \mathbf{y} on the residual \mathbf{z} to give the coefficient $\hat{\beta}_1$.
- In this procedure, “**regress \mathbf{b} on \mathbf{a}** ” means a simple univariate regression of \mathbf{b} on \mathbf{a} with no intercept, producing coefficient $\hat{\gamma} = \langle \mathbf{a}, \mathbf{b} \rangle / \langle \mathbf{a}, \mathbf{a} \rangle$ and residual vector $\mathbf{b} - \hat{\gamma}\mathbf{a}$. We say that \mathbf{b} is adjusted for \mathbf{a} , or is “orthogonalized” with respect to \mathbf{a} .

Multiple Regression from Simple Univariate Regression

- **Step 1** orthogonalizes \mathbf{x} with respect to $\mathbf{x}_0 = \mathbf{1}$.
- **Step 2** is just a simple univariate regression, using the orthogonal predictors $\mathbf{1}$ and \mathbf{z} .

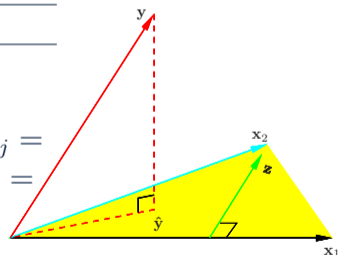
Algorithm: Regression by Successive Orthogonalization.

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$

2. For $j = 1, 2, \dots, p$

Regress \mathbf{x}_j on $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \dots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$

3. Regress \mathbf{y} on the residual \mathbf{z}_p to give the estimate $\hat{\beta}_p$.



Multiple Regression from Simple Univariate Regression

- Note that the inputs $\mathbf{z}_0, \dots, \mathbf{z}_{j-1}$ in step 2 are orthogonal, hence the simple regression coefficients computed there are in fact also the multiple regression coefficients. The result of this algorithm is

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle} \quad (2)$$

- Stated more generally, the j th multiple regression coefficient is the univariate regression coefficient of \mathbf{y} on $\mathbf{x}_{j \cdot 012 \dots (j-1)(j+1) \dots p}$, the residual after regressing \mathbf{x}_j on $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p$: The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of \mathbf{x}_j on \mathbf{y} , after \mathbf{x}_j has been adjusted for $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p$

Multiple Regression from Simple Univariate Regression

- From (2), we also obtain an alternate formula for the variance estimates:

$$\text{Var} \left(\hat{\beta}_p \right) = \frac{\sigma^2}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle} = \frac{\sigma^2}{\|\mathbf{z}_p\|^2}$$

- In other words, the precision with which we can estimate $\hat{\beta}_p$ depends on the length of the residual vector \mathbf{z}_p ; this represents how much of \mathbf{x}_p is unexplained by the other \mathbf{x}_k 's.

Multiple Regression from Simple Univariate Regression

- Above discussed Algorithm is known as the **Gram–Schmidt procedure** for multiple regression, and is also a useful numerical strategy for computing the estimates.
- We can represent step 2 of the Algorithm in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma}$$

where \mathbf{Z} has as columns the \mathbf{z}_j (in order), and $\mathbf{\Gamma}$ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$.

Multiple Regression from Simple Univariate Regression

- Introducing the diagonal matrix \mathbf{D} with j th diagonal entry $D_{jj} = \|\mathbf{z}_j\|$, we get



$$\begin{aligned}\mathbf{X} &= \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma} \\ &= \mathbf{Q}\mathbf{R}\end{aligned}$$

the so-called QR decomposition of \mathbf{X} . Here \mathbf{Q} is an $N \times (p + 1)$ orthogonal matrix, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and \mathbf{R} is a $(p + 1) \times (p + 1)$ upper triangular matrix. The QR decomposition represents a convenient orthogonal basis for the column space of \mathbf{X} . It is easy to see, for example, that the least squares solution is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}, \\ \hat{\mathbf{y}} &= \mathbf{Q}\mathbf{Q}^T\mathbf{y}\end{aligned}\tag{3}$$

Equation (3) is easy to solve because \mathbf{R} , is upper triangular.

References

-  The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Hastie, Tibshirani, and Friedman, Springer.
-  In Introduction to Statistical Learning with Application in R, Second Edition, James, Witten, Hastie, and Tibshirani, Springer.



Thank you!