Linear Regression

(Course: Introduction to Data Science)

Tirtharaj Dash

BITS Pilani, K.K. Birla Goa Campus

tirtharaj@goa.bits-pilani.ac.in

October 24, 2020

Simple Linear Regression I

- Simple linear regression is the most commonly used technique for determining how one variable of interest (the response variable) is affected by changes in another variable (the explanatory variable).
- The term "response" can mean a "dependent" and the term "explanatory" can mean "independent". However, it may be the case that the variable is not truly independent (due to variable interdependency with other variables).

Simple Linear Regression II

- Simple linear regression is used for three main purposes:
  1. To describe the linear dependence of one variable on another
  2. To predict values of one variable from values of another, for which more data are available
  3. To correct for the linear dependence of one variable on another, in order to clarify other features of its variability
- Any line fitted through a set of data points will deviate from each data point to greater or lesser degree.
- The vertical distance between a data point and the fitted line is termed as residual.
- This distance is a measure of the prediction error, in the sense that it is the discrepancy between the actual value of the response variable and the value predicted by the line.

Simple Linear Regression III

- Linear regression determines the best-fit line through a scatter-plot of the data points, such that the sum of squared residuals is minimised (sometimes, the average of the residual is minimised); it is equivalent to minimising the error variance.
- The fit is best refers to a setting where the sum of squared errors is as small as possible. This is the very reason why this method is also called as ordinary least squares (OLS) regression.

Linear regression equations I

Problem definition Given a set of n-points (x_i, y_i), find a fitting line $\beta_0 + \beta_1 x_i = 0$ that can compute the (predicted) response $\hat{y}_i = \beta_0 + \beta_1 x_i$, such that the sum of squared error $\mathcal{L} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is minimised.

An example of data sample:

x	y
3.4	20
6.5	19
10.1	50
5.6	30
7.2	20
⋮	⋮

Linear regression equations II

The goal is to solve for the unknowns βs. For this:

1. We need to set the partial derivatives of $\mathcal{L}$ with respect to the parameters β_0 and β_1
2. Set these to 0. i.e.

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = 0$$
$$\frac{\partial \mathcal{L}}{\partial \beta_1} = 0$$

3. Solve the two equations together to obtain the values for β_0 and β_1

Linear regression equations III

So, here are the **OLS estimators**:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{1}$$

and

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

where, $\overline{x}$ is the mean of x_is and $\overline{y}$ is the mean of y_is. That is:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

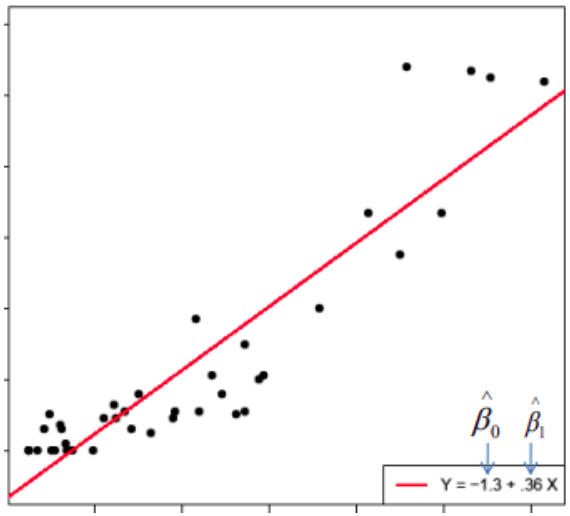
Linear regression equations IV

From Equation , we get:

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

This tells that: The regression line $\hat{\beta}_0 + \hat{\beta}_1 \overline{x} = 0$ goes through the center of the data points $(\overline{x}, \overline{y})$.

The parameter β_1 is called the slope of the fitting line and β_0 is the intercept.



(An example of a fitting line)

Linear regression equations V

We can re-write Equation as:

$$\hat{\beta}_1 = \frac{\text{Sample covariance between } x \text{ and } y}{\text{Sample variance of } x}$$

This tells that: The higher the covariance between x and y, the higher the slope will be.

Linear regression equations VI

In addition, we may want to compute a parameter called coefficient of determination. One way of computing this value is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

In the best case, the numerator (sum of squares of residuals) is 0 and therefore, $R^2 = 1$.

Multivariate Linear Regression I

- The problem of determining a response given multiple explanatory variables.

x_1	x_2	$\ldots$	x_d	y
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Here the predicted response for a data point $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ is given as

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_d x_{i,d}$$

where, $x_{i,j}$ is the value of the jth explanatory variable of the data point $\boldsymbol{x}_i$.

Multivariate Linear Regression II

- The goal is to obtain the parameters $\beta_0, \ldots, \beta_1$ such that the sum squared error is minimised.
- This will require solving a system of $d + 1$ equations:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = 0$$
$$\frac{\partial \mathcal{L}}{\partial \beta_1} = 0$$
$$\vdots$$
$$\frac{\partial \mathcal{L}}{\partial \beta_d} = 0$$

- However, we limit this in this course and instead rely on readily available tools for the same (e.e. Scikit-learn library in Python)

Vectorised Notation:

- Given n data points, each represented by a $(\boldsymbol{x}_i, y_i)$ pair; where $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}]^\top$.
- For mathematical and representational convenience, we include the intercept term of the regression equation within the coefficient term. So, $x_0 = +1$ for all the data points. So, $\boldsymbol{x} = [+1, x_{i,1}, \ldots, x_{i,d}]^\top$.
- Let the coefficients be $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^\top$.
- Now, we can write the linear regression equation as:

$$\hat{y}_i = \boldsymbol{\beta} \cdot \boldsymbol{x}_i; 1 \leq i \leq n$$

- And, the equation of the regression line (called a hyperplane in the multidimensional sense) is:

$$\boldsymbol{\beta} \cdot \boldsymbol{x} = 0$$

Issues

- The primary issue is with the assumption that there is linear relationships among variables. Linear regression model can only represent linear relationships.
- If there is non-linear relationships, it has to be known (or found out) beforehand. Then, each non-linearity or interaction has to be hand-crafted and explicitly given to the model as an input feature.

In practice I

Build a regression model for the following data points:

x_1	x_2	y
1	1	6
1	2	8
2	2	9
2	3	11

Now, predict the response for the data points

- (1,2) the model has seen this data point
- (3,5) the model has not seen this data point

In practice II

```
>>> import numpy as np
>>> from sklearn.linear_model import LinearRegression
>>> X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
>>> y = np.array([6, 8, 9, 11])
>>> reg = LinearRegression().fit(X, y)
>>> print('R^2:',reg.score(X, y))
R^2: 1.0
>>> print('Coefficients:',reg.coef_)
Coefficients: [1. 2.]
>>> print('Intercept:',reg.intercept_)
Intercept: 3.0000000000000018
>>> print('Prediction for [1,2]',reg.predict(np.array([[1, 2]])))
Prediction for [1,2] [8.]
>>> print('Prediction for [3,5]',reg.predict(np.array([[3, 5]])))
Prediction for [3,5] [16.]
```

In practice III

Practice the laboratory assignment available in the notebooks shared in the GitHub lab directory of this course. There are two problems:

- dummy dataset
- Boston house price prediction dataset

Lab link: L06 (Linear Regression)