Assignment 2: Team Assignment

Simple Linear Regression

Wikipedia describes linear regression as follows:

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable $y$ and one or more explanatory variables denoted $X$. The case of one explanatory variable is called simple linear regression.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or reduction, linear regression can be used to fit a predictive model to an observed data set of $y$ and $X$ values. After developing such a model, if an additional value of $X$ is then given without its accompanying value of $y$, the fitted model can be used to make a prediction of the value of $y$.
- Given a variable $y$ and a number of variables $X_1, \ldots, X_p$ that may be related to $y$, linear regression analysis can be applied to quantify the strength of the relationship between $y$ and the $X_j$, to assess which $X_j$ may have no relationship with $y$ at all, and to identify which subsets of the $X_j$ contain redundant information about $y$.

Linear regression models are often fitted using the least squares approach.

Simple linear regression is a way to describe a relationship between two variables through an equation of a straight line, called line of best fit, that most closely models the relationship. The following linear regression formula provides the least-squares, best-fit equation for the line:

$$y = a + bx$$

where

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} \quad \text{and} \quad a = \bar{y} - b \bar{x}$$

where $x$-bar and $y$-bar are the means of the $x$’s and $y$’s.
1. Compute the equation for the least-squares, best-fit line through the 5 points \{(2,2), (0,0), (-2,-2), (-1,1), (1,-1)\} shown in the following figure. Draw the best-fit line on the figure.

![Graph with points and line]

Equation of solution line: \( y = 0.6x \)


2. Compute the least-squares error, which is simply the sum of the squares of the \( y \)-differences between each original point and the best-fit regression line.

\[
\text{Least-squares error} = 2 \times (1.6 \times 1.6 + 0.8 \times 0.8) = 2 \times (2.56 + 0.64) = 2 \times 3.20 = 6.40
\]

3. Because some might think the best-fit regression line should have slope = 1 and go through the points (2,2) and (-2,-2), compute the least-squares error to this line to show that it is greater than the least-squares error obtained above.

\[
\text{Least-squares error to this line} = 2 \times (2 \times 2) = 8.0 \text{ so this is not the least-squares best-fit line!}
\]