4.2.5 Multiple Linear Regression

- Multiple regression is an extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.
- Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- The goal of multiple linear regression (MLR) is to model the **linear** relationship between the explanatory (independent) variables and response (dependent) variable.

  General equation for the probabilistic multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

  where

- y = the value of the dependent variable/ response variable
- β_0 = the regression constant, or intercept
- β_1 = the partial regression coefficient for independent variable 1
- β_2 = the partial regression coefficient for independent variable 2
- β_k = the partial regression coefficient for independent variable k
- k = the number of independent variables.
- The **partial regression coefficient of an independent variable**, β_j, represents *the increase that will occur in the value of y from a one-unit increase in that independent variable if all other variables are held constant.*
- The partial regression coefficients occur because more than one predictor is included in a model.
- the simplest multiple regression model is one constructed with two independent variables, where the highest power of either variable is 1. The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

  The constant and coefficients are estimated from sample information resulting in the following model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

  - Simple regression models yield a line that is fit through data points in the xy plane.

- In multiple regression analysis, the resulting model produces a response surface.
- The procedure for determining formulas to solve for multiple regression coefficients is similar. the formulas are established to meet an objecitve of minising the sum of squares of error for the model.

  Example

  A real estate study was conducted in a major city to investigate the factors influencing the market price of residential homes. Data were collected on 23 houses, recording the market price (in thousands of dollars), the size of each house in square metres, and the age of each house in years. As a data scientist working for a real estate company, you are required to develop a multiple linear regression model to predict the market price of a house based on its size and age.

1. Using the given dataset, perform an exploratory data analysis to summarize and visualize the relationships among the variables.
2. Fit a multiple linear regression model with market price as the dependent variable and the other two variables as predictors.
3. Clearly interpret the regression coefficients, assess the statistical significance of the predictors, and evaluate the overall fit of the model.
4. Use your model to predict the price of a house that is 250 square metres in size and 12 years old, and discuss any assumptions made in the regression analysis.

References

- Black, K., Asafu-Adjaye, J., Khan, N., Perera, N., Edwards, P., & Harris, M. (2007). *Australasian business statistics*. John Wiley & Sons.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.
- Salvatore, D., & Reagle, D. (2002). Statistics and Econometrics, Schaum's Outline Series.