Multilinear Regression

Abdelrahman Abdelbaky
Areeg Elkholy
Sama Emad
Shahd Eldansory

November 2024

§1 Introduction

Multilinear regression is a statistical method that relies on two or more independent variables to predict the outcome of the dependent variable. This differs from linear regression, which requires only one independent variable. Hence, multilinear regression allows for a more complex understanding of how multiple factors influence an outcome. Its ultimate purpose is to identify patterns and trends of inputs that will enable analyses and predictions to be made on a target variable. Multilinear regression is extremely helpful in data analysis and predictive modeling; due to its versatility, it enables analysts to quantify the effect of each independent variable while also taking into account the other variables. This is crucial in many fields, including biology, economics, machine learning, and more, where such fields rely heavily on understanding relationships between variables.

§2 Historical Background

The discovery of linear regression started with Francis Galton, who introduced the term "regression to the mean" to describe the likelihood of children to inherit characteristics that are more similar to the population average than their parents. This prepared the path for the development of increasingly complex statistical techniques, including multilinear regression.

By formalizing the mathematical foundations of regression analysis and developing the correlation coefficient, renowned mathematician Karl Pearson, best known for Pearson's Correlation, further improved the field of multilinear regression. Regression was first used as a tool for figuring out how variables relate to one another because of Pearson's work.

Afterwards, Ronald Fisher introduced maximum likelihood estimation and analysis of variance, which helped advance regression procedures. These two tools are also essential to statistical modeling, along with multilinear regression, which is now an essential tool in many different fields.

While social scientists use it to research societal patterns and human behavior, economists use it to forecast financial results and assess market trends. It provides comprehensible models and insights that inform decision-making in machine learning, laying the groundwork for increasingly complex algorithms.

§3 Preliminaries: Linear Algebra Concepts

Before diving into the mathematical framework of multilinear regression, we must establish some key concepts from linear algebra. These will serve as the building blocks for understanding the regression model and its computations.

§3.1 Matrices and Vectors

A matrix is a rectangular array of numbers arranged in rows and columns. Mathematically, an $m \times n$ matrix A is written as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where a_{ij} represents the element in the i-th row and j-th column.

A vector is a special case of a matrix, having only one row or one column: A column vector has dimensions $m \times 1$:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$

§3.2 Matrix Transpose

The **transpose** of a matrix A, denoted by A^T, is a matrix obtained by interchanging the rows and columns of A. Formally, if A is an $m \times n$ matrix with entries $A = [a_{ij}]$, the transpose A^T is an $n \times m$ matrix with entries $A^T = [b_{ij}]$, where $b_{ij} = a_{ji}$ for all i and j. In other words, the element in the i-th row and j-th column of A becomes the element in the j-th row and i-th column of A^T.

Example 3.1

For example:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

§3.3 Matrix Multiplication

The product of two matrices A and B is defined only if the number of columns of A matches the number of rows of B. If A is an $m \times n$ matrix and B is an $n \times k$ matrix, their product $C = AB$ is an $m \times k$ matrix.

The element in the i-th row and j-th column of C, denoted as c_{ij}, is computed as:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Example 3.2

For example:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}, \quad AB = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

§3.4 Identity Matrix

The identity matrix, denoted I_n, is a square matrix of size $n \times n$ where all diagonal elements are 1, and all off-diagonal elements are 0. For any matrix A, multiplying it by the identity matrix does not change A.

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

§3.5 Determinant of a Matrix

The determinant is a scalar value that can be computed from a square matrix and provides important information about the matrix. For example, the determinant tells us whether a matrix is invertible and plays a crucial role in linear transformations.

Definition 3.3

‘

The determinant of a 2×2 matrix is defined as:

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

For larger matrices ($n \times n$), the determinant is computed recursively using cofactor expansion. For a 3×3 matrix A:

$$\det A = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

Properties of the Determinant

- **Invertibility:** A square matrix A is invertible if and only if $\det(A) \neq 0$.
- **Effect of Scaling:** Multiplying a row or column of a matrix by a scalar k multiplies the determinant by k.
- **Transpose:** The determinant of a matrix is equal to the determinant of its transpose: $\det(A) = \det(A^T)$.
- **Product Rule:** For two square matrices A and B of the same size: $\det(AB) = \det(A)\det(B)$.
- **Row Operations:**
  - Swapping two rows of a matrix changes the sign of the determinant.
  - Adding a multiple of one row to another does not change the determinant.

– Scaling a row multiplies the determinant by the same factor.

Example 3.4

Let:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 5 & 6 & 0 \end{bmatrix}$$

The determinant of A is computed as:

$$\det(A) = 1 \cdot \det \begin{bmatrix} 1 & 4 \\ 6 & 0 \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 0 & 4 \\ 5 & 0 \end{bmatrix} + 3 \cdot \det \begin{bmatrix} 0 & 1 \\ 5 & 6 \end{bmatrix}$$

...

§3.6 Inverse of a Matrix

Let $A \in M_{n\times n}$. We call A *invertible* if there is a matrix $B \in M_{n\times n}$ such that

$$AB = BA = I_n.$$

We say that B is an *inverse* of A.

If A is not invertible, then we call it *singular*.

Note that only square matrices are invertible. If $A \in M_{m\times n}$, for $A \cdot B$ and $B \cdot A$ to make sense, we need $B \in M_{n\times m}$. But then $A \cdot B \in M_{m\times m}$ and $B \cdot A \in M_{n\times n}$, so $AB = BA$ forces $m = n$.

The inverse of a matrix is unique. We denote it by A^{-1}.

Let $A \in M_{n\times n}$ and assume there are $B, C \in M_{n\times n}$ such that

$$AB = BA = I_n \quad \text{and} \quad AC = CA = I_n.$$

Then,

$$B = BI_n = B(AC) = (BA)C = I_nC = C.$$

The inverse of a square matrix A, denoted A^{-1}, is defined as the matrix such that:

$$AA^{-1} = A^{-1}A = I$$

Not all matrices are invertible. A matrix A is invertible if and only if it is square and its determinant is not zero.

Example 3.5

$$A = \begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}, \quad A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix}$$

§3.7 Inverse of a 3×3 Matrix

Cofactor Expansion for Computing

In this section, we will introduce the concepts of **Minors** and **Cofactors**, which can be utilized to compute the determinant of any square matrix (of any size).

Minors and Cofactors

Let $A = (a_{ij})$ be a square matrix of size $n \times n$. For each i and j, we define the (i, j)-**minor** of A, denoted by $M_{i,j}$. This is the determinant of the submatrix formed by removing the ith row and the jth column from A.

Next, we define the **cofactors** of A as the "signed minors". Specifically, the (i, j)-**cofactor** of A, denoted by $C_{i,j}$, is given by

$$C_{i,j} = (-1)^{i+j} M_{i,j}.$$

In other words,

$$C_{i,j} = \begin{cases} +M_{i,j}, & \text{if } i+j \text{ is even,} \\ -M_{i,j}, & \text{if } i+j \text{ is odd.} \end{cases}$$

Cofactor Expansion

The value of the determinant of A, denoted by $|A|$, is precisely the cofactor expansion through **any row or any column** of A.

If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix},$$

then

$$\begin{aligned} |A| &= a_{11}C_{1,1} + a_{12}C_{1,2} + a_{13}C_{1,3} + \cdots + a_{1n}C_{1,n} \quad \text{(Row 1)}, \\ |A| &= a_{21}C_{2,1} + a_{22}C_{2,2} + a_{23}C_{2,3} + \cdots + a_{2n}C_{2,n} \quad \text{(Row 2)}, \\ &\vdots \\ |A| &= a_{n1}C_{n,1} + a_{n2}C_{n,2} + a_{n3}C_{n,3} + \cdots + a_{nn}C_{n,n} \quad \text{(Row } n\text{)}, \\ |A| &= a_{11}C_{1,1} + a_{21}C_{2,1} + a_{31}C_{3,1} + \cdots + a_{n1}C_{n,1} \quad \text{(Column 1)}, \\ |A| &= a_{12}C_{1,2} + a_{22}C_{2,2} + a_{32}C_{3,2} + \cdots + a_{n2}C_{n,2} \quad \text{(Column 2)}. \end{aligned}$$

The Adjoint of a Matrix

The cofactor matrix can be utilized to form a new matrix known as **The Adjoint of** A, which can help in determining A^{-1}, if it exists.

The Adjoint of A**:** We define the **adjoint** of A, denoted as $\text{adj}(A)$, to be the transpose of the cofactors matrix. That is,

$$\text{adj}(A) = \begin{bmatrix} C_{1,1} & C_{1,2} & C_{1,3} & \cdots & C_{1,n} \\ C_{2,1} & C_{2,2} & C_{2,3} & \cdots & C_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{n,1} & C_{n,2} & C_{n,3} & \cdots & C_{n,n} \end{bmatrix}^T.$$

Example: Computing the Inverse of a 3×3 Matrix

Given the matrix

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & -1 \\ 3 & 1 & 2 \end{bmatrix},$$

we compute its inverse step by step.

Step 1: Compute the Determinant of A

Using the cofactor expansion along the first row:

$$|A| = 2 \begin{vmatrix} 2 & -1 \\ 1 & 2 \end{vmatrix} - (-1) \begin{vmatrix} 1 & -1 \\ 3 & 2 \end{vmatrix} + 0 \begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix}.$$

Compute the minors:

$$M_{1,1} = \det \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = (2)(2) - (1)(-1) = 4 + 1 = 5,$$
$$M_{1,2} = \det \begin{bmatrix} 1 & -1 \\ 3 & 2 \end{bmatrix} = (1)(2) - (-1)(3) = 2 + 3 = 5,$$
$$M_{1,3} = \det \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} = (1)(1) - (2)(3) = 1 - 6 = -5.$$

Substitute back into the determinant:

$$|A| = 2(5) - (-1)(5) + 0(-5) = 10 + 5 = 15.$$

Thus, $|A| = 15$.

Step 2: Compute the Cofactor Matrix

The cofactor matrix is computed using the formula $C_{i,j} = (-1)^{i+j} M_{i,j}$.

Row 1:

$$C_{1,1} = (-1)^{1+1} M_{1,1} = 5,$$
$$C_{1,2} = (-1)^{1+2} M_{1,2} = -5,$$
$$C_{1,3} = (-1)^{1+3} M_{1,3} = -5.$$

Row 2:

$$M_{2,1} = \det \begin{bmatrix} -1 & 0 \\ 1 & 2 \end{bmatrix} = (-1)(2) - (0)(1) = -2,$$
$$M_{2,2} = \det \begin{bmatrix} 2 & 0 \\ 3 & 2 \end{bmatrix} = (2)(2) - (0)(3) = 4,$$
$$M_{2,3} = \det \begin{bmatrix} 2 & -1 \\ 3 & 1 \end{bmatrix} = (2)(1) - (-1)(3) = 2 + 3 = 5.$$

$$C_{2,1} = (-1)^{2+1} M_{2,1} = -(-2) = 2,$$
$$C_{2,2} = (-1)^{2+2} M_{2,2} = 4,$$
$$C_{2,3} = (-1)^{2+3} M_{2,3} = -5.$$

Row 3:

$$M_{3,1} = \det \begin{bmatrix} -1 & 0 \\ 2 & -1 \end{bmatrix} = (-1)(-1) - (0)(2) = 1,$$
$$M_{3,2} = \det \begin{bmatrix} 2 & 0 \\ 1 & -1 \end{bmatrix} = (2)(-1) - (0)(1) = -2,$$
$$M_{3,3} = \det \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = (2)(2) - (-1)(1) = 4 + 1 = 5.$$

$$C_{3,1} = (-1)^{3+1} M_{3,1} = 1,$$
$$C_{3,2} = (-1)^{3+2} M_{3,2} = -(-2) = 2,$$
$$C_{3,3} = (-1)^{3+3} M_{3,3} = 5.$$

The cofactor matrix is

$$C = \begin{bmatrix} 5 & -5 & -5 \\ 2 & 4 & -5 \\ 1 & 2 & 5 \end{bmatrix}.$$

Step 3: Compute the Adjoint of A

The adjoint is the transpose of the cofactor matrix:

$$\text{adj}(A) = C^T = \begin{bmatrix} 5 & 2 & 1 \\ -5 & 4 & 2 \\ -5 & -5 & 5 \end{bmatrix}.$$

Step 4: Compute the Inverse of A

The inverse of A is given by:

$$A^{-1} = \frac{1}{|A|} \text{adj}(A).$$

Substitute $|A| = 15$ and $\text{adj}(A)$:

$$A^{-1} = \frac{1}{15} \begin{bmatrix} 5 & 2 & 1 \\ -5 & 4 & 2 \\ -5 & -5 & 5 \end{bmatrix}.$$

Thus,

$$A^{-1} = \begin{bmatrix} \frac{5}{15} & \frac{2}{15} & \frac{1}{15} \\ -\frac{5}{15} & \frac{4}{15} & \frac{2}{15} \\ -\frac{5}{15} & -\frac{5}{15} & \frac{5}{15} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{2}{15} & \frac{1}{15} \\ -\frac{1}{3} & \frac{4}{15} & \frac{2}{15} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

§4 Preliminaries: Statistical Concepts

§4.1 Mean

The mean is the average value of a dataset:

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

In regression:

- The mean of Y shows the average outcome,
- The mean of each X shows their average values.

Example 4.1

Suppose:

$$Y = [300K, 400K, 500K], \quad X_1 = [1500, 2000, 2500].$$

The mean values are:

$$\text{Mean of } Y = \frac{300 + 400 + 500}{3} = 400 \text{ K},$$

$$\text{Mean of } X_1 = \frac{1500 + 2000 + 2500}{3} = 2000 \text{ sq.ft..}$$

§4.2 Variance

Variance measures the spread of data around the mean:

$$\text{Variance} = \frac{\sum_{i=1}^{n}(X_i - \text{Mean})^2}{n}.$$

In regression:

- Variance of Y tells how much the dependent variable varies,
- Variance of X shows the variability of predictors.

Example 4.2

For $Y = [300K, 400K, 500K]$ with a mean of $400K$:

$$\text{Variance of } Y = \frac{(300 - 400)^2 + (400 - 400)^2 + (500 - 400)^2}{3} = 6666.67 \text{ K}^2.$$

§4.3 Standard Deviation

Standard deviation is the square root of variance:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}.$$

It measures spread in the same units as the original data.

Example 4.3

Using the previous variance of Y:

$$\text{Standard Deviation of } Y = \sqrt{6666.67} \approx 81.65 \text{ K}.$$

§4.4 P-value

The p-value is an essential topic in statistics and is widely used in regression analysis or any statistical test in order to determine whether the relationship observed in data is due to chance or reflects a true underlying pattern. In simpler terms, the p-value answers the question: *How likely is it that the observed result occurred by random chance?*

Null Hypothesis

The null hypothesis (H_o) is the assumption that there is no effect or no relationship between the variables. In the context of regression, the null hypothesis might state that an independent variable has no effect on the dependent variable.

Interpreting the P-value

The p-value is a measure of evidence against the null hypothesis. Here's how to interpret the p-value:

- **Low p-value (typically $\leq$ 0.05)**: This suggests that the observed effect is statistically significant, meaning it is unlikely to have occurred due to random chance. In the context of regression, this means that the independent variable is likely to have a meaningful relationship with the dependent variable.
- **High p-value ($>$ 0.05)**: A high p-value suggests that the observed effect could easily be due to random chance, which implies that the predictor does not have a statistically significant effect on the dependent variable.

Significance Level

The significance level, denoted as α (alpha), is the threshold against which the p-value is compared. Typically, it's set to 0.05, which means there is a 5% chance of rejecting the null hypothesis if it is actually true. Here's how the p-value is used with the significance level:

- If ($p - value \leq \alpha$) (e.g., 0.05): Reject the null hypothesis. This suggests that the effect (relationship) is statistically significant.
- If ($p - value > \alpha$) (e.g., 0.05): Fail to reject the null hypothesis. This suggests that there is insufficient evidence to conclude that the effect is statistically significant.

§4.5 Residual Sum of Squares (RSS)

The **Residual Sum of Squares** (RSS), also referred to as the **Sum of Squared Residuals** (SSR), is a fundamental concept in regression analysis. It is a measure of the discrepancy between the data and the estimation model. Specifically, RSS quantifies how far off the predictions of a model are from the actual observed values. In simpler terms, it captures the "errors" or differences between the predicted values of the dependent variable and the actual observed values. Understanding RSS is crucial because it provides insight into how well a regression model fits the data.

In a regression model, we are trying to predict a dependent variable Y based on one or more independent variables X. The model produces predicted values $\hat{Y}$ (pronounced as "Y-hat"), which are estimates of the actual values of Y. The difference between each observed value Y_i and the corresponding predicted value $\hat{Y}_i$ is called the residual, denoted as:

$$e_i = Y_i - \hat{Y}_i.$$

The residual represents the error for each data point, showing how much the prediction missed the actual value. To ensure that these errors do not cancel each other out (since some predictions might be above and some below the actual values), we square each residual. This gives us the squared residual for each observation, e_i^2.

The **Residual Sum of Squares (RSS)** is then simply the sum of all these squared residuals:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Mathematically, this formula aggregates the squared differences between the observed and predicted values, giving a single number that summarizes the total amount of error in the model. A smaller RSS indicates that the model's predictions are closer to the actual values, implying a better fit. Conversely, a larger RSS suggests that the model's predictions are far from the actual data, signaling a poor fit.

To understand the significance of RSS, let's consider a simple example. Imagine we are building a model to predict house prices based on their size in square feet. We collect data from three houses:

$$Y = [300K, 400K, 500K], \quad \hat{Y} = [320K, 380K, 510K].$$

For each house, we calculate the residuals:

$$e_1 = 300K - 320K = -20K, \quad e_2 = 400K - 380K = 20K, \quad e_3 = 500K - 510K = -10K.$$

The squared residuals are:

$$e_1^2 = (-20K)^2 = 400K^2, \quad e_2^2 = (20K)^2 = 400K^2, \quad e_3^2 = (-10K)^2 = 100K^2.$$

Thus, the Residual Sum of Squares (RSS) is:

$$RSS = 400K^2 + 400K^2 + 100K^2 = 900K^2.$$

In this case, the RSS value gives us a total measure of how much the predicted prices deviate from the actual prices. A smaller RSS would indicate that our predictions are closer to the real prices, meaning the model is performing well. On the other hand, if the RSS were much larger, it would indicate that our model is not accurately capturing the relationship between house size and price.

§4.6 F-statistic

The F-statistic is another important topic that is also used in multilinear regression as it determines the overall significance of the regression model.

Computing the F-statistic

The F-statistic is computed by comparing the model with all predictors to a model with no predictors. It assesses whether the group of predictors used in the regression model, as a whole, has a statistically significant relationship with the dependent variable.

Mathematically, the F-statistic is given by:

$$F = \frac{\frac{\text{SSR}}{k}}{\frac{\text{SSE}}{n-k-1}}$$

where:

- SSR (Sum of Squares for Regression): The variability explained by the regression model.
- SSE (Sum of Squares for Error): The variability that remains unexplained by the regression model.
- k: The number of predictors (independent variables) in the model.
- n: The total number of data points.

Interpreting the F-statistic

The F-statistic provides a measure of how well the model fits the data compared to a model without any predictors. Here's how to interpret the F-statistic:

- **High F-statistic**: A high F-statistic suggests that the independent variables collectively have a statistically significant relationship with the dependent variable.
- **Low F-statistic**: A low F-statistic indicates that the predictors in the model do not have a meaningful relationship with the dependent variable.

§5 Mathematical Foundations of Multilinear Regression

In many areas of study, we are interested in understanding how one quantity (the dependent variable) changes when other quantities (the independent variables) change. Regression analysis is one of the most common statistical tools that are used to examine the relationship between variables. Regression analysis is a fundamental statistical tool used to examine the relationship between variables.

At its core, regression analysis seeks to model how one or more input variables (often called predictors or independent variables) influence an output variable (known as the dependent variable or response).For example, consider a scenario where we want to predict a student's exam score based on the number of hours they studied and their sleep quality. The exam score is the dependent variable, while study hours and sleep quality are the independent variables.

§5.1 Linear Regression: A Starting Point

Linear regression is the simplest form of regression. It assumes a linear relationship between the independent variables and the dependent variable. In its simplest form, simple linear regression, there is only one independent variable. The relationship can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y: Dependent variable (e.g., exam score),
- x: Independent variable (e.g., hours studied),
- β_0: Intercept, representing the value of y when $x = 0$,
- β_1: Slope, representing the change in y for a one-unit increase in x,
- ϵ: Error term, accounting for factors not captured by the model.

We include the the error term because real-world data is rarely perfectly predictable. Many factors that influence y might not be included in the model. For instance:

- Measurement errors can occur.
- There may be variables influencing y that are unknown or difficult to measure.
- Some randomness is always present in natural and social phenomena.

The error term ϵ represents all these unaccounted-for influences. In a model, our goal is to find the "best" values for β_0 and β_1, minimizing the mismatch between the predicted y (from the formula) and the observed y (from the data).

§5.2 The Multilinear Regression Model

The problem is that in real-world problems, we often have multiple factors that influence the output. For example, a student's exam performance might depend not only on hours studied but also on their sleep quality, prior knowledge, and class attendance. In such cases, multilinear regression is used. Multilinear regression generalizes the concept of simple linear regression to accommodate multiple independent variables. The model is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where:

- y: Dependent variable (e.g., exam score),
- $x_1, x_2, \ldots, x_n$: Independent variable (e.g., hours studied),
- β_0: Intercept, representing the value of y when $x = 0$,
- $\beta_1, \beta_2, \ldots, \beta_n$: Slope, representing the change in y for a one-unit increase in x,
- ϵ: Error term, accounting for factors not captured by the model.

To simplify analysis and computation, we express the model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}$$

§5.3 Least Squares Estimation

The goal of multilinear regression is to estimate the parameter vector $\boldsymbol{\beta}$ such that the residual sum of squares (RSS) is minimized. The RSS is given by:

$$\text{RSS} = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

We can expand the RSS as:

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

Next, expanding this expression:

$$\text{RSS} = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

- $\mathbf{Y}^T\mathbf{Y}$ is a constant term with respect to $\boldsymbol{\beta}$, so it does not affect the minimization process.
- The second term $-2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}$ is linear in $\boldsymbol{\beta}$.
- The third term $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ is quadratic in $\boldsymbol{\beta}$.

Deriving the Normal Equations

To find the optimal $\boldsymbol{\beta}$, we take the derivative of RSS with respect to $\boldsymbol{\beta}$ and set it equal to zero. Differentiating each term:

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

We simplify the equation by canceling the factor of 2:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$$

This is the **normal equation** that must be satisfied for the least-squares solution.

Assuming that $\mathbf{X}^T\mathbf{X}$ is invertible (i.e., the matrix has full rank and is not singular), we can solve for $\boldsymbol{\beta}$ by multiplying both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

This formula provides the least-squares estimate of the regression coefficients.

§5.4 Properties of the Estimator

The least-squares estimator $\boldsymbol{\beta}$ has several desirable properties under the classical assumptions:

Linearity

The estimator $\boldsymbol{\beta}$ is a linear function of $\mathbf{Y}$:

$$\boldsymbol{\beta} = \mathbf{A}\mathbf{Y}, \quad \text{where } \mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Unbiasedness

Under the assumptions of linearity and independence, $\boldsymbol{\beta}$ is an unbiased estimator:

$$\mathbb{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}_{\text{true}}$$

Variance

The variance of $\boldsymbol{\beta}$ is given by:

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

Here, σ^2 is the variance of the error term $\boldsymbol{\epsilon}$, which can be estimated as:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{m - (n+1)}$$

§5.5 Assumptions of Multilinear Regression

For the model and its estimates to be valid, the following assumptions must hold:

- **Linearity:** The relationship between y and x_i is linear.
- **Independence:** The residuals $\boldsymbol{\epsilon}$ are independent.
- **Homoscedasticity:** The variance of $\boldsymbol{\epsilon}$ is constant for all x_i.
- **Normality:** The residuals $\boldsymbol{\epsilon}$ are normally distributed.
- **No Multicollinearity:** The predictors $x_1, x_2, \ldots, x_n$ are not highly correlated.

Example 5.1

Suppose we have three observations ($m = 3$) and two independent variables ($n = 2$). The design matrix $\mathbf{X}$, the observation vector $\mathbf{Y}$, and the coefficients $\boldsymbol{\beta}$ are:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

If $\mathbf{X}^T\mathbf{X}$ is invertible, we can calculate $\boldsymbol{\beta}$ using:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Example: Given data:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix},$$

the normal equations yield:

$$\boldsymbol{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

The calculation gives $\boldsymbol{\beta} = [1, 2]$, meaning the model is $y = 1 + 2x$.

§5.6 What is the Gauss-Markov Theorem?

The greatest classical achievement in statistics is considered to be the Gauss-Markov theorem. Its role in statistics is as significant as the Pythagorean theorem in geometry.

§5.7 Why is it important? Why does it matter?

The Gauss-Markov theorem provides a justification for using Ordinary Least Squares (OLS) in linear regression analysis. OLS is one technique for estimating the parameters of a linear regression. Simply, OLS involves minimizing the sum of the squared residuals (the difference between the observed values and the predicted values) in order to fit a line that represents the relationship between the independent and dependent variables.

Returning to the Gauss-Markov theorem's importance, it guarantees specific conditions under which this technique would be reliable and efficient. Thus, it highlights the importance of adhering to the underlying assumptions to ensure that the OLS estimates are the best and unbiased. Furthermore, it offers numerous commonly utilized practical

applications as well as multiple approaches, including geometry, differential calculus, and other areas.

Assumptions of the Gauss-Markov Theorem

There are five Gauss-Markov assumptions, which are crucial for ensuring that the ordinary least squares (OLS) estimators are the best linear unbiased estimators (BLUE). These assumptions are as follows:

1. **Linearity:** The linear regression model is linear in parameters. Specifically, the model can be written as:
$$y = X\beta + \epsilon$$
where y is the vector of dependent variables, X is the matrix of independent variables, β is the vector of parameters, and ϵ is the vector of error terms. This assumption ensures that the model relationship between the dependent and independent variables is linear in the parameters β, but not necessarily in the variables themselves.

2. **Random Sampling:** The sample taken for the linear regression model must be drawn randomly from the population. This assumption ensures that the sample is representative of the population and that each observation has an equal probability of being included in the sample. This helps eliminate biases in the estimation of parameters and error terms.

3. **Non-Collinearity:** There should be no perfect linear relationship between the independent variables. In other words, the matrix X (the matrix of independent variables) must have full rank. This condition implies that no independent variable can be written as an exact linear combination of other independent variables. If perfect collinearity exists, the inverse of $X'X$ does not exist, and the ordinary least squares estimator cannot be computed.

4. **Zero Conditional Mean:** The expected value of the error terms, given any value of the independent variables, must be zero. Formally, this is written as:
$$E[\epsilon|X] = 0$$
This assumption ensures that the error term is uncorrelated with the independent variables. If this assumption holds, the regression model is unbiased because the errors do not systematically influence the estimates of the coefficients.

5. **Homoscedasticity:** The error terms in the regression should all have the same variance, regardless of the value of the independent variables. This assumption is expressed as:
$$\text{Var}(\epsilon|X) = \sigma^2 I$$
where σ^2 is a constant and I is the identity matrix. Homoscedasticity ensures that the variability of the errors is consistent across all levels of the independent variables. If the variance of the errors differs across observations (heteroscedasticity), the OLS estimators are still unbiased but are no longer efficient.

6. **Normality of Errors (for Inference)** The error terms in the regression model are assumed to follow a normal distribution:
$$\epsilon \sim N(0, \sigma^2).$$

This assumption implies that the errors are distributed with a mean of 0 and a constant variance σ^2.

According to the Gauss-Markov theorem, if these assumptions hold, then the OLS estimator is the Best Linear Unbiased Estimator (BLUE). Therefore, a crucial part of estimating regression coefficients is ensuring the data aligns with these assumptions. If these conditions fail to apply, one can adjust the experimental strategy and select the one that most closely matches the situation.

The Gauss-Markov theorem states that in a linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the matrix of independent variables, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the error term, the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ minimizes the variance among all linear unbiased estimators, provided:

- The errors have zero mean ($\mathbb{E}[\boldsymbol{\varepsilon}] = 0$),
- The errors have constant variance (homoscedasticity),
- The errors are uncorrelated ($\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$).

Example: Suppose you want to predict house prices (y) based on square footage (x_1) and the number of bedrooms (x_2). Using OLS, you estimate the coefficients $\hat{\boldsymbol{\beta}}$. The Gauss-Markov theorem ensures that your estimates have the smallest possible variance compared to any other unbiased method, such as a weighted least squares method (if weights are arbitrary).

§6 Applications

§6.1 Application of Multiple Linear Regression in Analyzing Soccer Team Performance:

This application by Castillo Ramirez et al. (2017) was one that used multiple linear regression (MLR) to understand the relationship between different independent variables and a dependent outcome. Firstly, the authors highlighted that the integration of individual player performance with the team performance as a whole can provide a view of the factors affecting match results. The study utilized multilinear regression to model the relationship between several dependent variables, such as match results, and independent variables, including player statistics and game conditions, which yields many powerful insights on how these factors collectively affect performance. [3] The researchers collected data that encompassed a total of 38 matched played. The study focused on some variables that were hypothesized to impact the result of the match. These independent variables included:

- Minutes Played: The total minutes each player participated in the matches.
- Percentage of Ball Possession: The proportion of time the team maintained control of the ball during matches.
- Shots on Goal: The number of attempts made by the team to score.
- Fouls Committed: The total fouls incurred by the team, which could affect match dynamics.

- Disciplinary Actions: The number of yellow and red cards received by players, which could influence player availability and team performance.

- Diversity of Nationalities: The variety of nationalities represented in the team, which may correlate with team dynamics and performance.

The study then applied multilinear regression analysis to examine the relationship between the dependent variable, which is match results represented by the points earned, and the independent variables. In the study conducted by Castillo Ramirez et al. (2017), the interpretation of results from the multiple linear regression analysis is a critical component that provides insights into the factors influencing the performance of the Chelsea Football Team during the 2014-2015 English Premier League season. The results indicated that certain players and their minutes played during the match had a significant correlation with the points earned. For example, the analysis highlighted specific players, such as Thibaut Courtois and Petr Cech, whose minutes played were associated with the points obtained by the team as shown in the table below:

Table 1
ANOVA tables of the regression model associated with goalkeepers.

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	149,344	2	74,672	67,75	0,0000
Residue	28,656	26	1,10215		
Total	178,0	28			
Variable		*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Thibaut Courtois		0,0236978	0,00242898	9,75627	0,0000
Petr Cech		0,0337486	0,00547692	6,16197	0,0000

R-square (adjusted for g.l.) = 83.282 percent

POINTS OBTAINED = 0.0236978 * (Courtois minutes) + 0.0337486 * (Cech minutes)

Figure 1: Caption describing the table, e.g., Summary of Regression Analysis for Goalkeeper Minutes and Points.

The table illustrates how much of the total variation in points can be explained by the minutes played by the goalkeepers. Key Components of the table and their meaning:

- Sum of Squares: The "Sum of Squares for the Model" (149.344) indicates the extent to which the model (the Courtois and Cech minutes) contributes to the variability of the points. The amount of variability that remains unexplained after taking the model into account is indicated by the "Sum of Squares for Residuals" (28.656).

- F-Statistic: This value (67.75) is obtained by comparing the explained and unexplained variation. A high F-statistic indicates that the model accounts for a significant variability in points, implying that goalkeepers' minutes are crucial.

- P-Value: The P-value (0.0000) indicates how likely it is that the observed relationship occurred by chance. A very low P-value indicates that the association is unlikely to be attributable to random chance, implying that goalkeeper minutes are actually related to points scored.

Based on these values, there is a significant link between Courtois and Cech's minutes played, and the team's points scored. This suggests that as these goalkeepers play more minutes, the team is likely to score more points.

The study also showed that the contribution of defensive and midfield positions was also crucial. This is shown in the tables below.

Table 2
ANOVA tables of the regression model associated with the defenses

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	153,408	2	76,704	77,98	0,0000
Residue	24,5919	25	0,983678		
Total	178,0	27			

Variable	*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Cesar Azpilicueta	0,0270789	0,00249295	10,8622	0,0000
Filipe Luis	0,021897	0,00430342	5,08828	0,0000

R-square (adjusted for g.l.) = 85.6317 percent

POINTS OBTAINED = 0.0270789 * (minutes Azpilicueta) + 0.021897 * (minutes F. Luis)

Table 3
ANOVA tables of the regression model associated with the flyers

Source	*Sum of Squares*	*Gl*	*Squared Mean*	*Reason-F*	*P-value*
Model	152,64	2	76,3202	78,25	0,0000
Residue	25,3595	26	0,975366		
Total	178,0	28			

Variable	*Estimation*	*Standard Error*	*Statistical T*	*P-value*
Eden Hazard	0,0140432	0,0054014	2,59993	0,0152
Nemanja Matic	0,0139743	0,0054288	2,5741	0,0161

R-squared (adjusted for g.l.) = 85.2051 percent

POINTS OBTAINED = 0.0140432 * (Hazard minutes) - 0.0139743 * (Matic minutes)

Figure 2: Caption describing the table, e.g., Summary of Regression Analysis

In summary, the interpretation of this study offers significant statistical analysis and practical implications for team performance which is important for future research. These insights not only contribute to the academic understanding of sports analytics, but also provide actionable recommendations for improving team tactics and outcomes, highlighting the value of multilinear regression in this area.

§6.2 MultiLInear Regression in the Energy Field

In the field of energy prediction, Sias et al. (2024) provide a compelling case for using multilinear regression (MLR) instead of single linear regression (SLR). The authors demonstrate that MLR, which integrates multiple independent variables coming from various energy sources, greatly improves the accuracy of energy consumption projections compared to SLR, which relies on a single input variable (Sias et al., 2024).

Initially, the authors used historical data on energy consumption and supply from a variety of Turkish sources. This dataset contained information about a variety of energy resources, including coal, gas, oil, nuclear, hydro, wind, and solar energy. The data spans a seven-year period, giving a solid basis for study (Sias et al., 2024). The authors created an SLR model that used total energy supply from all sources as a single input variable to estimate energy consumption. This model was used as a basis for comparison to the more complicated MLR model (Sias et al., 2024).

The MLR model was designed to include several independent variables that represent the energy produced by different sources. This approach allowed the model to account for the contributions of multiple energy sources to overall consumption, thus increasing prediction accuracy (Sias et al., 2024). The authors also used the K-Means clustering

approach to increase prediction accuracy even further. This method was used to pre-cluster the data based on regular patterns such daily, weekly, monthly, and quarterly trends.

The performance of the proposed RMLR model was assessed by comparing its predictions to actual historical data. To measure the accuracy of their predictions, the authors calculated a variety of error metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The RMLR model outperformed the SLR model, with a lower error rate of 3.4 percent when utilizing the weekly clustering pattern (Sias et al., 2024).

This study is noteworthy because, in addition to comparing MLR and SLR, the authors also compared their model against other forecasting techniques, including exponential Gaussian Process Regression (GPR), sequential XGBoost, and seq2seq Long Short-Term Memory (LSTM) models. This comparative analysis proved the RMLR model's effectiveness in predicting energy consumption (Sias et al., 2024).

In conclusion, this study shows that the MLR approach, especially when improved with clustering techniques and periodic pattern analysis, greatly increases the accuracy of energy consumption predictions when compared to typical SLR methods. This research not only illustrates the advantages of MLR in energy forecasting but also shows its potential applications in other industries where precise predictions are critical for decision-making and resource management (Sias et al., 2024).

§7 Challenges and Limitations

While multilinear regression is a powerful and widely-used statistical method, it has several limitations that must be considered when applying it in practice:

§7.1 Multicollinearity

One of the main limitations for multilinear regression is multicollinearity. Multicollinearity occurs when two or more predictor variables in the model are highly correlated. This makes it difficult to determine the individual effect of each predictor on the response variable, as the coefficients $\boldsymbol{\beta}$ become unstable and their standard errors increase. This can lead to misleading inferences about the importance of variables. **Example:** If predictors X_1 (advertising spending) and X_2 (marketing spending) are highly correlated, it becomes challenging to assess their individual contributions to sales. **Solution:** Techniques such as removing correlated predictors, using principal component analysis (PCA), or regularization methods like Ridge Regression can mitigate multicollinearity.

§7.2 Overfitting

In the cases when the model is too complex, for example it includes too many predictors relative to the size of the sample, it may capture random noise rather than the underlying trend in the data. Overfitting leads to excellent performance on the training data but poor generalization to new data.

§7.3 Assumptions of the Model

Multilinear regression relies on several key assumptions:

- **Linearity:** The relationship between predictors and the response must be linear. Nonlinear relationships can lead to biased estimates.

- **Independence:** Observations must be independent. Violations can occur in time-series or hierarchical data.
- **Homoscedasticity:** The variance of errors should be constant across all levels of the predictors.
- **Normality:** The residuals are assumed to follow a normal distribution. Deviations can affect hypothesis tests and confidence intervals.

§7.4 Outliers and Influential Points

Outliers and influential points can distort the model by exerting a disproportionate impact on the parameter estimates. These anomalies may arise from errors in data collection or represent rare, meaningful observations.

§7.5 Sensitivity to Data Scaling

Since multilinear regression relies on matrix operations, the scale of predictor variables can affect the numerical stability and interpretability of the model. Variables with larger scales may dominate the regression coefficients.

§7.6 Limited Applicability to Nonlinear Relationships

Multilinear regression assumes a linear relationship between predictors and the response variable. It cannot effectively model interactions or complex, nonlinear patterns without explicitly including additional terms (e.g., polynomial or interaction terms).

YouTube Video Link

Watch the YouTube Video

Google Colab Link

Open Google Colab Notebook

References

[1] Albert Team. "Assumptions of OLS: Econometrics Review." *Albert Resources*, March 1, 2022. https://www.albert.io/blog/key-assumptions-of-ols-econometrics-review/.

[2] Chatterjee, S., and A. S. Hadi. *Regression Analysis by Example*. 5th ed. Wiley, 2015.

[3] Frost, Jim. "Residual Sum of Squares (RSS) Explained." *Statistics By Jim*, February 26, 2024. https://statisticsbyjim.com/regression/residual-sum-of-squares-rss/.

[4] "From Equivalent Linear Equations to Gauss-Markov Theorem." *Journal of Inequalities and Applications*, SpringerOpen, July 13, 2010. https://doi.org/10.1155/2010/259672.

[5] "How OLS Regression Works." *ArcGIS Pro Documentation*. Accessed December 4, 2024. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-ols-regression-works.htm.

[6] "Multiple Linear Regression." *JMP*. Accessed December 4, 2024. https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-multiple-regression.html.

[7] Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill, 2005.

[8] Merriam-Webster. "America's Most Trusted Dictionary." Merriam-Webster. Accessed December 2, 2024. https://www.merriam-webster.com.

[9] Sias, Quota Alief, Rahma Gantassi, Yonghoon Choi, and Jeong Hwan Bae. "Recurrence Multilinear Regression Technique for Improving Accuracy of Energy Prediction in Power Systems." *MDPI*, October 18, 2024. https://www.mdpi.com/1996-1073/17/20/5186.

[10] Sampaio, Vitor. "Understanding Ordinary Least Squares (OLS): The Foundation of Linear Regression." *Medium*, June 2, 2023. https://medium.com/@VitorCSampaio/understanding-ordinary-least-squares-ols-the-foundation-of-linear-regression-1d79bfc3

[11] Sahu, Prashant. "A Comprehensive Guide to OLS Regression: Part-1." *Analytics Vidhya*, November 28, 2024. https://www.analyticsvidhya.com/blog/2023/01/a-comprehensive-guide-to-ols-regression-part-1/.

[12] Montgomery, D. C., E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. 5th ed. Wiley, 2012.

[13] "Numeracy, Maths and Statistics - Academic Skills Kit." Accessed December 4, 2024. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/multiple-regression.html.