Linear Regression

Linear regression is a powerful statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It is one of the most widely used techniques for predictive modeling and data analysis. This lesson will explore both simple and multiple linear regression, detailing their models, assumptions, and interpretations.

Part 1: Simple Linear Regression

1.1 Model

Simple linear regression involves two variables: one independent variable (predictor) and one dependent variable (response). The relationship between these variables is modeled through a linear equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept of the regression line (the value of Y when $X = 0$).
- β_1 is the slope of the regression line (the change in Y for a one-unit change in X).
- ϵ is the error term, which accounts for the variability in Y that cannot be explained by X.

1.2 Assumptions

Simple linear regression relies on several key assumptions:

1. **Linearity**: The relationship between X and Y must be linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The variance of residual (error) terms should be constant across all values of X.
4. **Normality**: The residuals (errors) of the model should be normally distributed.
5. **No multicollinearity**: As this model involves only one predictor, this condition is inherently satisfied.

1.3 Interpretation

Interpreting the coefficients:

- **Intercept (β_0)**: It represents the expected mean value of Y when all X are 0.
- **Slope (β_1)**: It indicates the expected change in Y for a one-unit change in X.

To evaluate the model, you would typically look at:

- **Coefficient of determination (R^2)**: This statistic measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides an indication of goodness of fit and hence a measure of how well unseen samples are likely to be predicted by the model.

Certainly! Let's solve a simple linear regression problem by hand using theoretical calculations. We'll use a straightforward example with a small dataset to predict house prices based on their size.

Example Problem

Suppose we have a dataset of house sizes and their corresponding selling prices:

House Size (sq ft)	Price (000's $)
1200	200
1500	240
1800	280
2100	310
2500	350

We want to find a linear relationship between the size of a house and its price:

$$\text{Price} = \beta_0 + \beta_1 \times \text{Size}$$

Where:

- β_0 is the intercept.
- β_1 is the slope, indicating the change in price for each square foot increase in size.

Step 1: Calculate the Means

First, calculate the mean of the house sizes and the mean of the prices.

$$\overline{X} = \frac{\sum X_i}{n}$$

$$\overline{Y} = \frac{\sum Y_i}{n}$$

Where:

- X is the house size.
- Y is the price.

Calculations:

$$\overline{X} = \frac{1200 + 1500 + 1800 + 2100 + 2500}{5} = 1820 \text{ sq ft}$$

$$\overline{Y} = \frac{200 + 240 + 280 + 310 + 350}{5} = 276 \text{ (000's \$)}$$

Step 2: Calculate Slope (β_1)

The formula for the slope β_1 of the regression line is:

$$\beta_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

Calculations:

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = (1200 - 1820)(200 - 276) + (1500 - 1820)(240 - 276) + \ldots$$

$$= (-620)(-76) + (-320)(-36) + (-20)(4) + (280)(34) + (680)(74)$$

$$= 47120 + 11520 + 80 + 9520 + 50320 = 123560$$

$$\sum (X_i - \overline{X})^2 = (1200 - 1820)^2 + (1500 - 1820)^2 + \ldots$$

$$= 384400 + 102400 + 400 + 78400 + 462400 = 1028000$$

$$\beta_1 = \frac{123560}{1028000} \approx 0.12$$

Step 3: Calculate Intercept (β_0)

The intercept β_0 is calculated using:

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$\beta_0 = 276 - 0.12 \times 1820 \approx 76.6$$

Step 4: Formulate the Regression Equation

Now, the regression equation is:

$$\text{Price} = 76.6 + 0.12 \times \text{Size}$$

Step 5: Use the Model for Prediction

If you want to predict the price of a house with 2000 sq ft:

$$\text{Price} = 76.6 + 0.12 \times 2000 = 316 \ (000\text{'s } \$)$$

Conclusion

This regression model suggests that for every additional square foot in house size, the price increases by $120.

The model also predicts that a house of size 2000 sq ft would sell for approximately $316,000. This simple calculation illustrates the basic mechanics of forming a linear regression model by hand.

Part 2: Multiple Linear Regression

2.1 Model

Multiple linear regression extends the simple linear regression model by including multiple independent variables. The model is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where:

- $X_1, X_2, \ldots, X_n$ are the independent variables.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables.

2.2 Assumptions

The assumptions of multiple linear regression include all those of simple linear regression, with an additional focus on:

- **No multicollinearity**: Independent variables should not be too highly correlated with each other. This can be tested using variance inflation factors (VIF).
- **Independence of residuals**: There should be no correlation between the residuals. This can be tested using Durbin-Watson statistic.

2.3 Interpretation

Each coefficient in a multiple regression model represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other independent variables constant. This is known as "ceteris paribus" or "all else being equal".

2.4 Model Evaluation

Multiple regression analysis also relies on statistical metrics to evaluate the model's performance:

- **Adjusted R^2**: Similar to R^2, but adjusts for the number of variables in the model.
- **F-statistic**: Used to determine the overall significance of the model.
- **AIC/BIC scores**: Provide a measure of the model's quality and complexity.