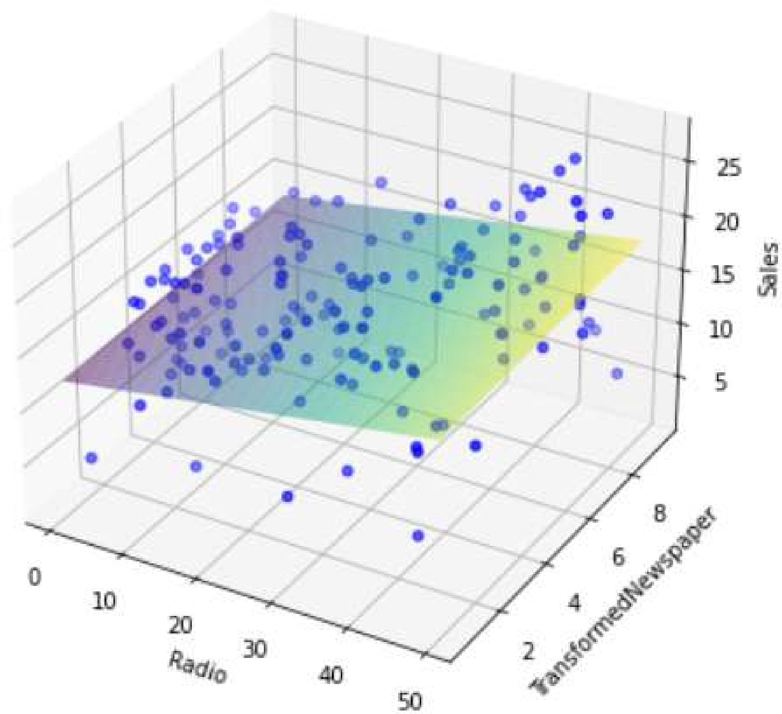
Multiple Linear Regression

A Statistical Approach for Predicting Outcomes

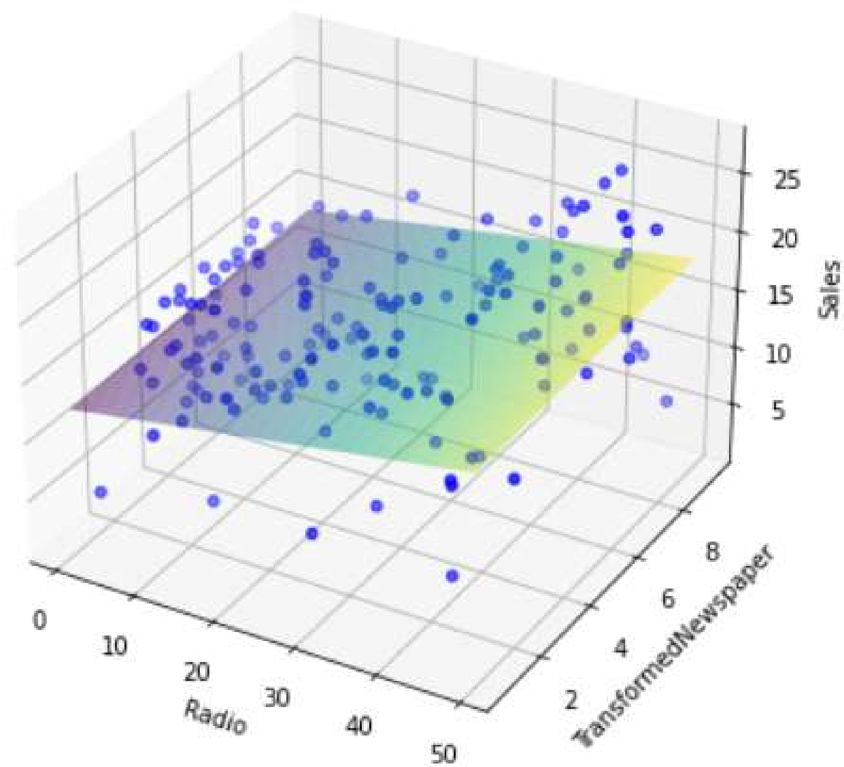


@DataByteSun

What is Multiple Linear Regression?

- **Definition:** Multiple Linear Regression (MLR) is a statistical method used to model the relationship between a dependent variable (y) and two or more independent variables (x_1, x_2, ..., x_k). It extends simple linear regression by considering multiple predictors.

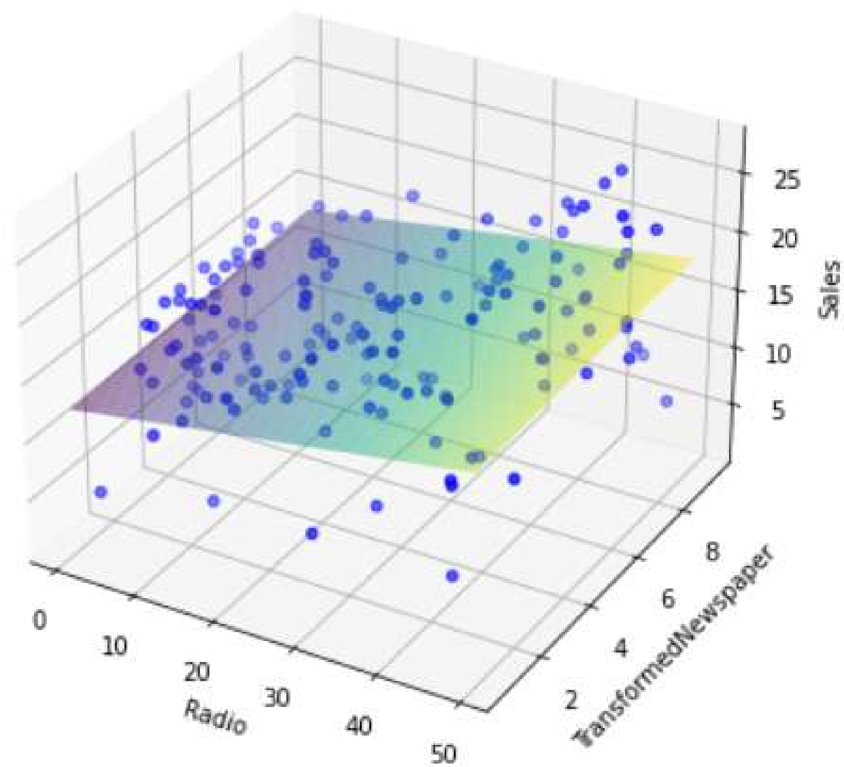
- **Goal:** The goal of MLR is to model the relationship between a dependent variable and multiple independent variables for prediction and data analysis.



@DataByteSun

Real-World Examples

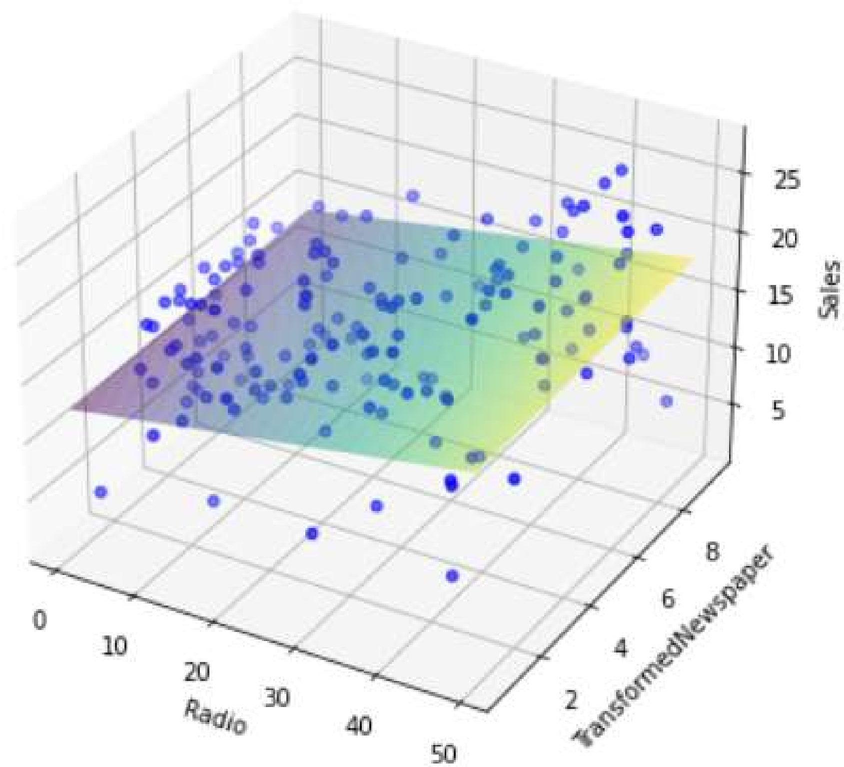
- ❑ Real Estate Pricing: Using factors like square footage, number of bedrooms, and location to predict the price of a house.
- ❑ Sales Forecasting: Analyzing advertising budget, product price, and seasonal trends to predict future sales.
- ❑ Healthcare: Predicting patient outcomes based on factors such as age, weight, and medical history..



@DataByteSun

Key Concepts in Linear Regression

- **Independent Variable (x):** The variable(s) used to make predictions about y.
- **Dependent Variable (y):** The variable we want to predict or understand.
- **Model Assumption:** Assumes a straight-line relationship between the dependent and independent variables.



@DataByteSun

Multiple Linear Regression Equation

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

- Y: predicted value of dependent variable.
- Intercept ($\boldsymbol{\beta_0}$): Intercept (The expected value of y when all x are zero).
- $\boldsymbol{\beta_1 x_1}$: Slope (change in y per unit change in x_1).
- $\beta_n x_n$: The regression coefficient of the last independent variable
- **ϵ** : Error term (differences between actual and predicted y values).

Formula Explanation

Multiple Linear Regression Equation

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Let's say we have n = 2:

$$\beta_0 = \bar{y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Where,

$\bar{X}_1 =$ Mean of X_1

$\bar{X}_2 =$ Mean of X_2

$\bar{y} =$ Mean of y

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)}{n}$$

$$\sum x_2^2 = \sum X_2^2 - \frac{(\sum X_2)}{n}$$

$$\sum x_1 y = \frac{\sum X_1 y - \sum X_1 \sum y}{n}$$

$$\sum x_2 y = \frac{\sum X_2 y - \sum X_2 \sum y}{n}$$

Simple Linear Regression
VS
Multiple Linear Regression

Simple Linear Regression (SLR)

- One dependent variable, **one independent** variable.
- $Y=\beta 0+\beta 1X1+\epsilon$
- Represented as a **straight line** in 2D.
- **Simpler**, easier to analyze.
- Assumes **linearity** and **homoscedasticity.**
- **Easier interpretation** of the slope.

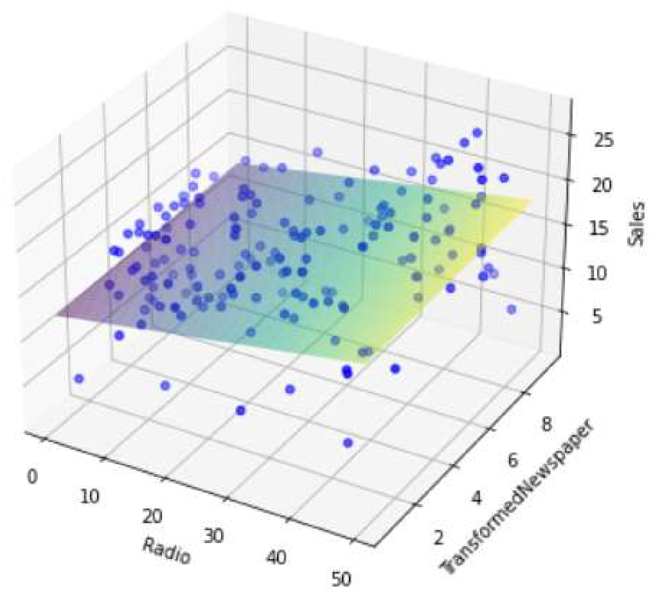
Multiple Linear Regression (MLR)

- One dependent variable, **multiple independent** variables.
- $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$
- Represented as a **hyperplane** in multiple dimensions.
- More **complex**, captures multiple factors.
- Same as SLR, plus checks for **multicollinearity.**
- **More effort** needed for **interpreting** multiple coefficients.

Assumptions of Multiple Linear Regression

Same as Simple Linear Regression

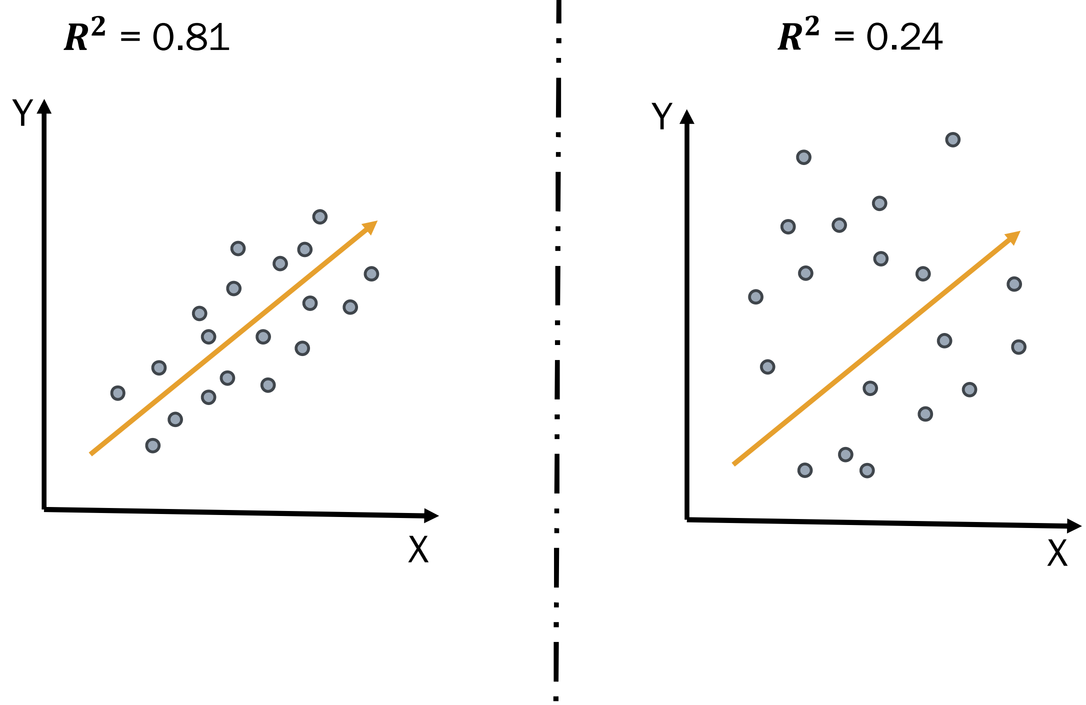
- **Linearity:** Relationship between x and y is linear.
- **Independence:** Observations are independent.
- **Homoscedasticity:** Constant variance of errors.
- **Normality:** Errors are normally distributed.
- **Multicollinearity:** requires that independent variables have low to moderate correlations with one another.



Evaluating Model Fit – R-squared

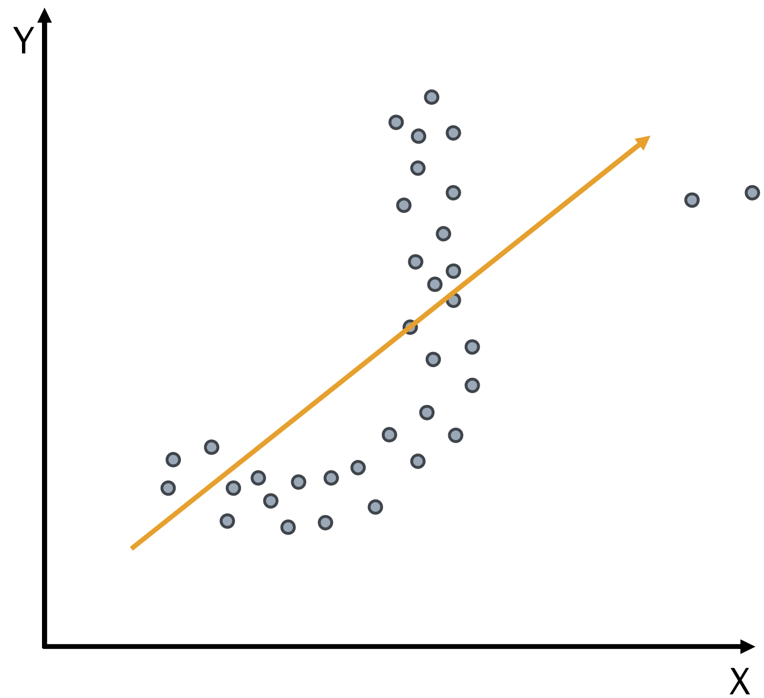
- R-squared (R^2): A metric to assess how well the model fits the data, representing the proportion of variance in y explained by x.
- **Values:** Ranges from 0 to 1, where closer to 1 means a better fit.

An R^2 value closer to 1 suggests a better fit.



Limitations of Linear Regression

- **Multicollinearity:** High correlation among independent variables can affect model stability.
- **Overfitting:** Including too many predictors can result in overfitting and reduce generalizability.
- **Sensitivity to outliers:** Outliers can disproportionately influence results.



@DataByteSun

What We Covered

- ❑ Basic understanding of Multiple Linear Regression.
- ❑ Simple Linear Regression VS Multiple Linear Regression.
- ❑ Assumptions of Multiple Linear Regression.
- ❑ Evaluating Model Fit.
- ❑ Limitations of Linear Regression.

@DataByteSun

Thank You

\- @DataByteSun