Regression

Introduction

- Regression analysis is a statistical method used for the estimation of relationships between a dependent variable and one or more independent variables.
- It allows us to understand how the value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.
- **Importance of Regression in Machine Learning:**
- **Predictive Analysis:** Regression models are pivotal in predicting and forecasting, enabling machines to identify trends and make predictions based on historical data.
- **Decision Making:** Helps in making informed decisions by understanding the relationships between variables. For instance, businesses can forecast sales, revenue, and customer trends.
- **Feature Understanding:** It provides insights into the importance of different variables on the outcome, facilitating feature selection and optimization in model development.

Understanding Variables in Regression

- **Dependent Variables (DV):**
- **Definition:** The dependent variable, also known as the response or outcome variable, is what you aim to predict or explain. It's called 'dependent' because its values depend on the influences of other variables.
- **Characteristics:** In regression analysis, the DV is typically continuous and measured on an interval or ratio scale. However, in logistic regression, the DV can be categorical.
- **Examples:** Sales figures (in dollars), patient blood pressure levels, or test scores.
- **Independent Variables (IV):**
- **Definition:** Independent variables, also known as predictors or explanatory variables, are the inputs or factors that you hypothesize to have an impact on the dependent variable.
- **Characteristics:** IVs can be continuous, categorical, or even binary. The choice and treatment of IVs depend on the type of regression analysis being conducted.
- **Examples:** Marketing budget, medication dosage, or hours spent studying.
- **Interplay between DV and IV:**
- The essence of regression is to explore and model the relationship between these variables, aiming to understand how changes in the IVs affect changes in the DV.

The Goals of Regression Analysis

- **Predictive Modeling:**
- One of the primary goals of regression analysis is to create a model that can predict the value of the dependent variable based on known values of independent variables. This is invaluable in forecasting future events or conditions.
- **Causal Inference:**
- Regression allows researchers and analysts to infer causal relationships between variables. By observing how the dependent variable changes as the independent variables are varied, one can draw conclusions about causal effects, though caution must be exercised to account for confounding factors.
- **Variable Relationship Exploration:**
- Understanding the form and strength of the relationship between the dependent variable and each independent variable. This includes determining whether relationships are linear or non-linear and identifying interaction effects between variables.
- **Optimization:**
- In many cases, regression analysis is used to find the optimal conditions for desired outcomes. For businesses, this could mean determining the optimal price point for a product or the most efficient use of resources.
- **Hypothesis Testing:**
- Regression analysis is often employed to test theories or hypotheses about relationships between variables. For example, one might use regression analysis to test if and how education level affects income.

Introduction to Linear Regression

- Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
- The primary goal is to predict the value of the dependent variable based on the values of the independent variables, assuming that the relationship between them is linear.

- **Mathematical Formula:**
- The equation of a linear regression model is represented as:

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

- - Where:
- - y is the dependent variable,
- - x_1, x_2, ..., x_n are independent variables,
- - beta_0 is the y-intercept,
- - beta_1, beta_2, ..., beta_n are the coefficients of the independent variables,
- - epsilon is the error term.

Simple vs. Multiple Linear Regression

- **Simple Linear Regression:**
- Involves only one independent variable to predict the dependent variable.
- The formula simplifies to: $y = \beta_0 + \beta_1 x_1 + \epsilon$
- Used to understand the relationship between two variables and is straightforward to model and interpret.

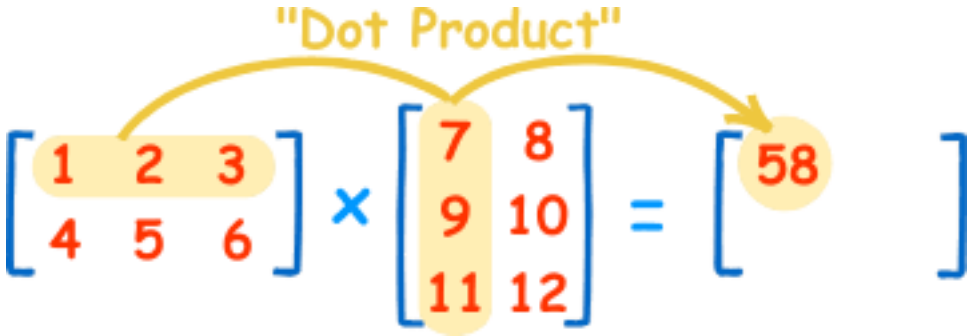
- **Multiple Linear Regression:**
- Incorporates two or more independent variables to predict the dependent variable.
- It allows for the examination of the simultaneous effect of several variables on the dependent variable.
- While providing a more comprehensive model, it requires careful consideration to avoid multicollinearity and to ensure model interpretability.
- **Choosing Between Simple and Multiple Linear Regression:**
- The choice depends on the research question and the complexity of the dataset.
- Simple linear regression is a good starting point for understanding the relationship between two variables.
- Multiple linear regression is suited for more complex questions involving multiple factors influencing the outcome.
- **Practical Applications:**
- Simple linear regression could be used to predict sales based on advertising spend, while multiple linear regression might analyze the impact of advertising spend, seasonality, and price changes on sales.

Example

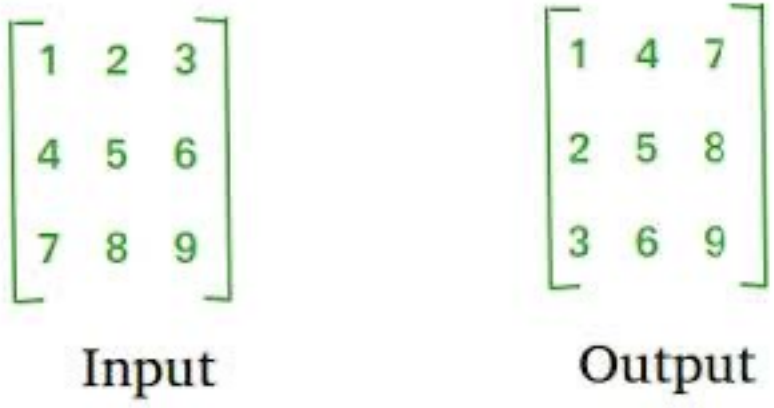
- *x* (Feature): 1, 2, 3
- *y* (Target): 2, 6, 14
- **Construct the Matrix *X***: $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$
- **Target Vector *y***: $\mathbf{y} = \begin{bmatrix} 2 \\ 6 \\ 14 \end{bmatrix}$

- Calculating Coefficients Using the Normal Equation: $\beta = (X^T X)^{-1} X^T y$

- Matrix Multiplication



- Matrix Transpose



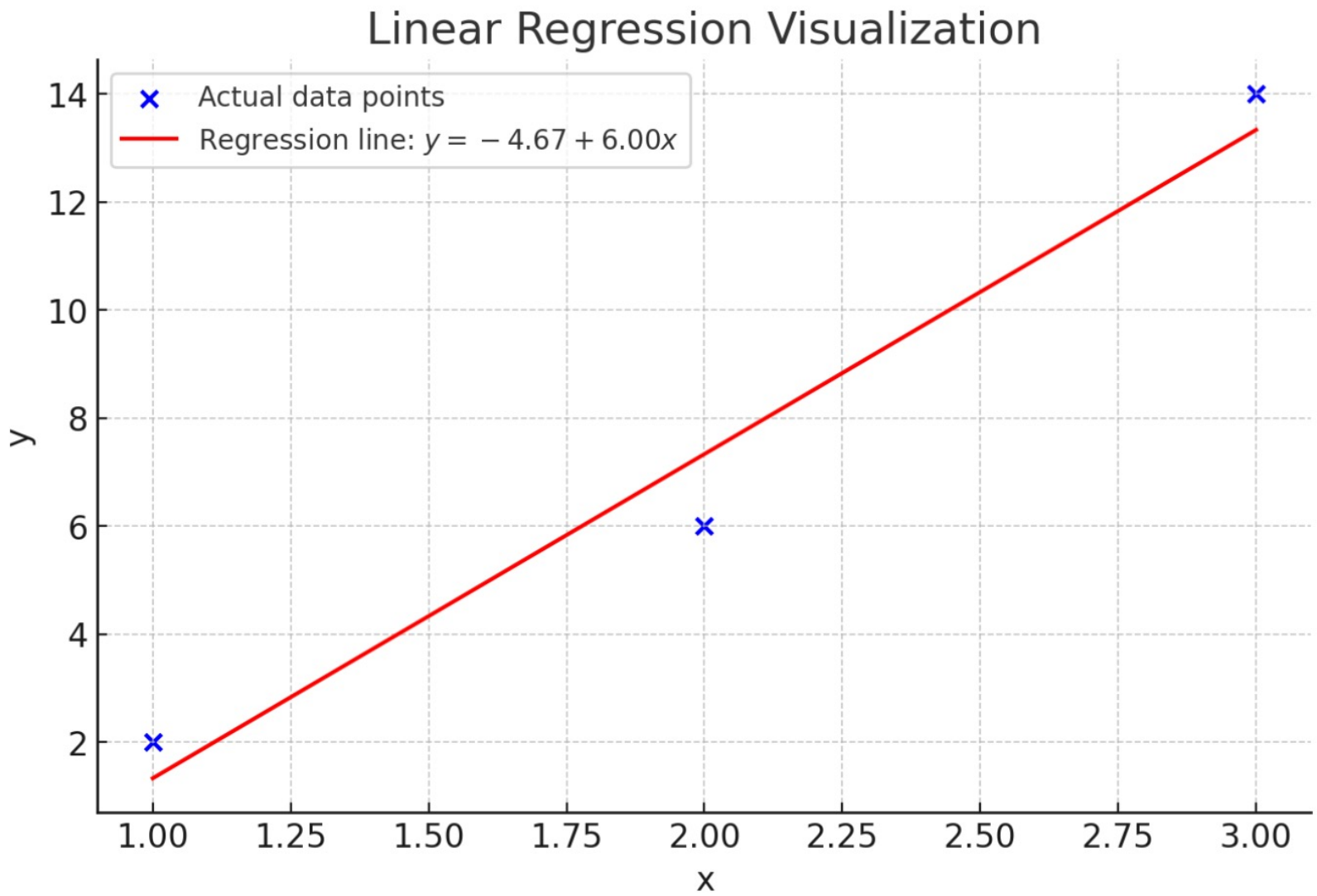
Matrix Inverse

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

- $X^T X$ $\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$
-
- $(XT\ X)^{-1}$ (Inverse of ($X^T X$) $\begin{bmatrix} 2.333 & -1.000 \\ -1.000 & 0.500 \end{bmatrix}$
- $X^T y$ $\begin{bmatrix} 22 \\ 56 \end{bmatrix}$

- Finally, $(XT\ X)^{-1}\ X^T y$

- Coefficients β $\begin{bmatrix} -4.667 \\ 6.000 \end{bmatrix}$
- These are the regression coefficients where β_0 = -4.667 is the intercept and β_1 = 6.000 is the slope.
- These calculations yield a linear model y=−4.667+6x


Linear Regression Visualization
Actual data points
Regression line: $y = -4.67 + 6.00x$
14
12
10
8
6
4
2
y
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
x

Understanding the Coefficients (β)

- β_0: This is the intercept of the line. It represents the value of *y* when *x*=0.
- β_1 : This is the slope of the line. It indicates how much *y* changes for a one-unit increase in *x*.
- **Equation of the Line**
- From the coefficients we computed earlier (β_0=−4.67and β_1 =6), the equation of the regression line can be expressed as:
- *y*=−4.67+6*x*

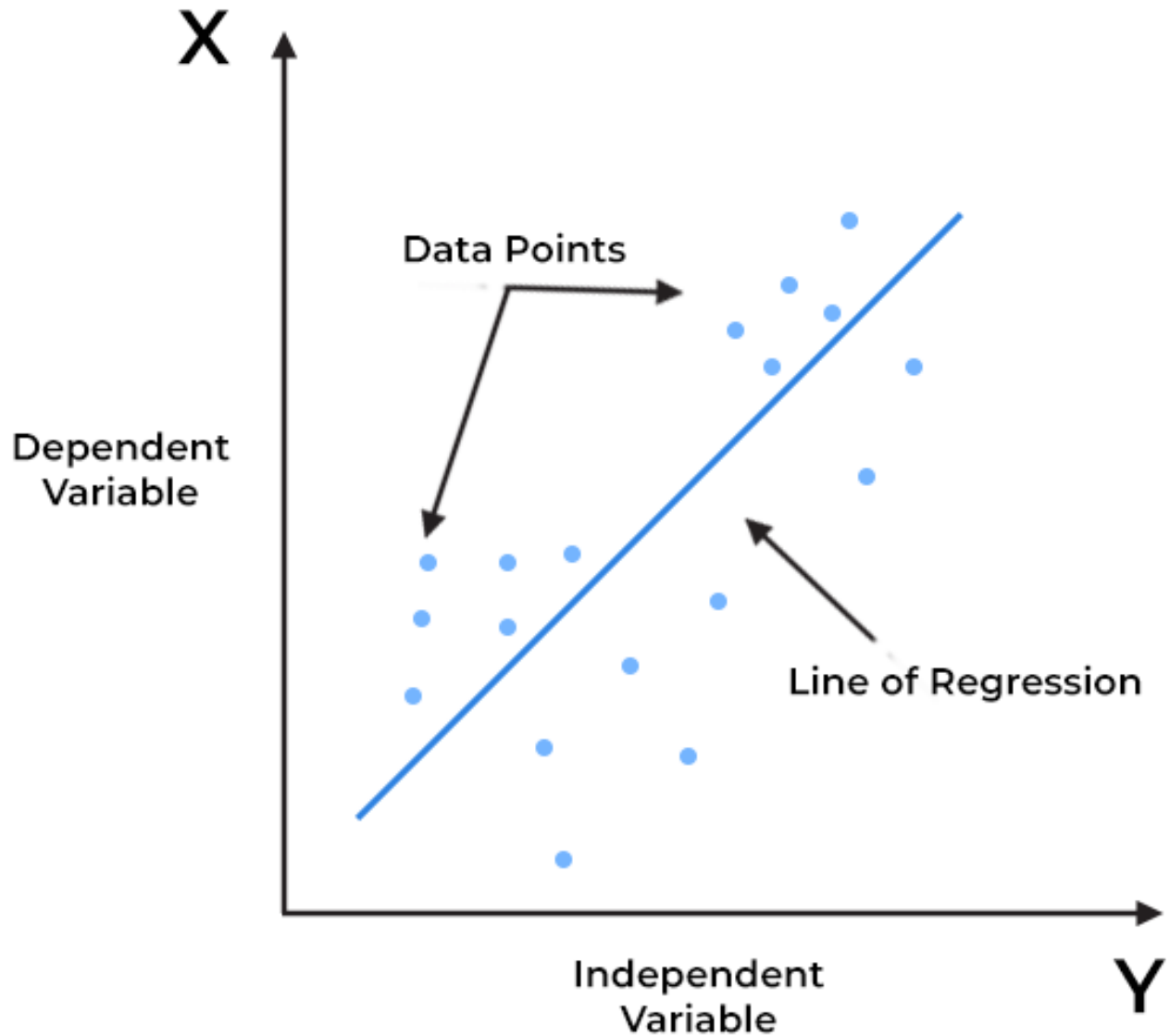
- **Plotting the Line**
- Here’s how the line gets plotted using the equation:

1.Calculate *y* values:

1. For each *x* in the dataset, substitute into the equation to get corresponding *y* values. For instance, $y=-4.67+6\times1=1.33$.

2.Draw the Line:

1. The line is drawn by connecting the points calculated from the equation across the range of *x* values. This line is intended to fit as closely as possible to the actual data points, minimizing the distance between the predicted values and the observed values.


X
Data Points
Dependent
Variable
Line of Regression
Independent
Variable
Y

Evaluating Linear Regression Models

- **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)**
- MSE measures the average of the squares of the errors—i.e., the average squared difference between the estimated values and the actual value.
- Calculation: $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- y_i is the actual value, $\hat{y}_i$ is the predicted value, and *n* is the number of observations.
- A lower MSE indicates a better fit of the model to the data.

Root Mean Squared Error (RMSE):

- RMSE is the square root of the mean squared error, providing a measure of the **magnitude** of the error.
- Calculation: $RMSE=\sqrt{MSE}$
- RMSE is in the same units as the dependent variable and is particularly useful for comparing the prediction errors of different models or datasets.
- Like MSE, a lower RMSE value indicates a model that better fits the dataset.
- **Choosing Between MSE and RMSE:**
- Both MSE and RMSE are widely used to evaluate the performance of regression models, with RMSE being **more interpretable** due to being in the same units as the response variable.
- The choice between MSE and RMSE can depend on the specific context of the problem and whether you wish to heavily penalize larger errors (as MSE does due to its squaring effect).
- **Considerations:**
- It's essential to use these metrics in conjunction with other evaluation metrics and visualizations to get a comprehensive understanding of model performance.
- Evaluating a model solely on a single metric might not reveal potential issues such as overfitting or underfitting.

See Lesson 3.pynb

- https://github.com/sanadv/MLCourse

- Code cell
- `3.1 Linear Regression`

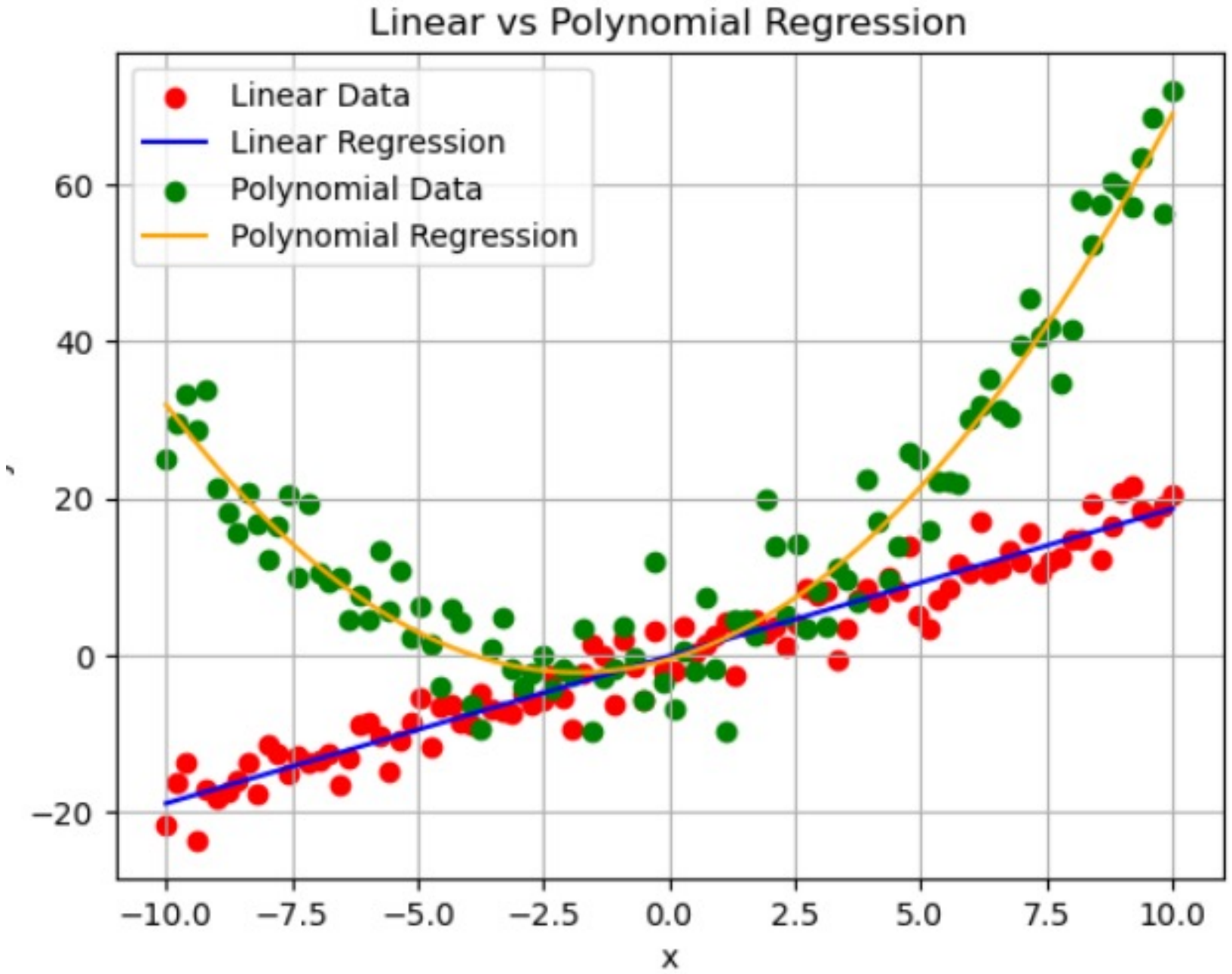
Difference Between Linear and Polynomial Regression

- **Linear Regression:**
- Models the relationship between a dependent variable and one or more independent variables using a straight line (linear equation).
- The equation is of the form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$

- Assumes that the relationship between the dependent and independent variables is linear.

- **Polynomial Regression:**
- A form of regression analysis where the relationship between the independent variable *x* and the dependent variable *y* is modeled as an n^{th} degree polynomial.
- The equation is of the form: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon.$

- Can model non-linear relationships, making it more flexible than linear regression for data with non-linear trends.
- **Key Difference:**
- The core difference lies in the capability of polynomial regression to fit a wide range of curvature by adjusting the degree of the polynomial, thereby providing a better fit for datasets with complex relationships.


Linear vs Polynomial Regression
Linear Data
Linear Regression
Polynomial Data
Polynomial Regression
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
x

Applications of Polynomial Regression

- **Predictive Modeling:**
- Used in areas where the relationship between variables is known to be non-linear, such as in economic growth models, where the growth rate might accelerate or decelerate over time.
- **Science and Engineering:**
- In physics and engineering, polynomial regression can model phenomena where changes in one variable cause non-linear effects on another, such as the relationship between stress and strain in materials under various conditions.
- **Financial Analysis:**
- Polynomial regression can be applied to forecast financial indicators, where the relationship between variables can be more complex than a simple linear trend, such as predicting stock prices based on historical performance and other economic factors.

Challenges and Overfitting

- **Understanding Complexity:**
- Choosing the right degree for the polynomial is crucial. Too low, and the model may not capture the true relationship (underfitting). Too high, and the model may become overly complex, capturing noise in the data as if it were a real pattern (overfitting).
- **Overfitting:**
- A significant challenge with polynomial regression is the risk of overfitting, especially with high-degree polynomials. Overfitting occurs when the model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- **Mitigation Strategies:**
- Cross-validation: Use cross-validation techniques to validate the choice of polynomial degree.
- Regularization: Techniques like Ridge or Lasso regression can help prevent overfitting by penalizing large coefficients.
- Model Selection Criteria: Use information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) to choose the appropriate model complexity.

See Lesson 3.pynb

- Cell Code
- `3.2 Polynomial Regression`

- And Lesson 3 Poly Reg example

Introduction to Regularization

- **What is Regularization?**
- Regularization is a technique used to prevent overfitting by discouraging overly complex models in regression. This is achieved by introducing a penalty term to the loss function used to fit the model.
- The key idea is to impose a cost on larger coefficients. Regularization methods shrink the regression coefficients by imposing a penalty on their size.
- **Why Use Regularization?**
- **Bias-Variance Trade-off:** Regularization helps manage the trade-off between bias and variance, aiming to minimize model error across different datasets.
- **Multicollinearity:** Reduces model variance by adding a penalty term to counteract the high variability caused by correlated predictors.
- **Feature Selection:** Can implicitly perform feature selection by shrinking coefficients of less important features to zero.

See Lesson 3.pynb

- Code cells
- `3.3 Ridge Regression`

- `3.4 Lasso Regression`

Introduction to Non-linear Regression

- Non-linear regression is a form of regression analysis in which observational data is modeled by a function that is a non-linear combination of the model parameters and depends on one or more independent variables.
- **Difference from Polynomial Regression:**
- While polynomial regression can model non-linear relationships, it is still a form of linear regression since it depends linearly on the model parameters. Non-linear regression models, however, can have parameters that are not linearly related to the outcome variable.
- Polynomial regression is a specific case of non-linear regression that uses polynomial functions of the independent variable, but non-linear regression encompasses a broader class of models.
- **Key Points:**
- Non-linear models are more flexible than linear or polynomial models, allowing for a better fit for complex data patterns.
- The complexity of non-linear models makes them more challenging to fit and require iterative numerical methods for parameter estimation.

Examples of Non-linear Models

- **Exponential Growth/Decay Model:**
- Used for modeling growth processes, radioactive decay, and more.
- Equation: $y = ae^{bx}$ where a and b are parameters
- **Logistic Growth Model:**
- Commonly used in biology for modeling population growth with a carrying capacity.
- Equation: $y = \frac{c}{1 + ae^{-bx}}$ where a, b, and c are parameters
- **Michaelis-Menten Kinetics:**
- Used in biochemistry to describe the rate of enzymatic reactions.
- Equation:
- $v = \frac{V_{max}[S]}{K_m + [S]}$ where V_{max} and K_m are parameters and $[S]$ is the substrate concentration

Fitting and Evaluating Non-linear Regression Models

- **Fitting Process:**
- Non-linear models often require initial guesses for the parameters and use iterative optimization techniques, such as gradient descent, Newton-Raphson, or the Levenberg-Marquardt algorithm, to find the best-fitting parameters.
- **Evaluation:**
- Similar to linear models, non-linear models can be evaluated using metrics like R-squared, Mean Squared Error (MSE), or Mean Absolute Error (MAE).
- Model selection criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can help compare different non-linear models.
- **Challenges:**
- Non-linear models can easily overfit the data, especially with many parameters.
- Ensuring convergence to the global minimum during optimization can be difficult.
- **Practical Considerations:**
- It's crucial to understand the underlying process that generated the data to choose an appropriate model.
- Visualization of both the fitted model and residuals can help assess fit quality and guide model refinement.

Mitigating Overfitting and Underfitting

- **Overfitting**
- **Definition:** Occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the model's performance on new data.
- **Solutions:**
  - **Cross-Validation:** Use techniques like k-fold cross-validation to ensure the model generalizes well.
  - **Regularization:** Apply L1 (Lasso), L2 (Ridge), or Elastic Net regularization to penalize large coefficients.
  - **Pruning:** For tree-based models, remove parts of the tree that do not provide additional information.
- **Underfitting:**
- **Definition:** Occurs when a model cannot capture the underlying trend of the data and performs poorly even on training data.
- **Solutions:**
  - **Feature Engineering:** Create new features or transform existing ones to better capture the data's structure.
  - **Model Complexity:** Increase the model complexity by adding more parameters or using a more sophisticated model.
  - **Reducing Regularization:** If using regularization, consider decreasing the regularization strength to allow more flexibility.

Interview Questions

1. **Scenario on Handling Multicollinearity**: Imagine you are working with a dataset intended to predict housing prices based on features like size, location, age of the property, and proximity to amenities. During your analysis, you discover significant multicollinearity between the size of the house and its age. Describe the steps you would take to address this issue. Which specific techniques or metrics would you use to confirm and mitigate multicollinearity to ensure the stability and interpretability of your model?
2. **Scenario on Model Evaluation Metrics**: You have developed a multiple linear regression model to forecast quarterly sales based on advertising spend, seasonal effects, and economic conditions. The model has an R-squared of 0.85, but your client is concerned about the reliability of predictions. Discuss how you would use MSE and RMSE in this scenario to evaluate model performance further. Explain the implications of these metrics and how they might influence your recommendations for model adjustments or client expectations.

MAIN POINTS

Machine Learning addresses simple, complex and very complex problems. Simple problems include Regression, Complex problems include doing Search (Auto driving car, Speech Recognition); and very complex problems include complex Decision Making

Science of Consciousness:

Scientific research on students practicing TM shows holistic improvement in intellectual performance, personality and individual differences and improved graduate academic performance.