Signature Assignment

Probaility and Statistics

Aaditya Pathak- 19757

Introduction:

The Kaggle dataset provides a unique perspective on the physical heights of US Presidents, offering historical information and a platform for statistical analyses. The study aims to examine a potential linear relationship between the order of the presidency and the height of the presidents, using linear regression techniques. The order of the presidency refers to the chronological sequence of each individual serving as President, starting from George Washington. The dependent variable, the height of the presidents, is influenced by factors like genetics, nutrition, and health standards. The findings could offer insights into the physical attributes of U.S. Presidents and contribute to the broader discussion on physical character.

Methodology: Linear Regression Model:

This analysis uses a linear regression model to model the relationship between a dependent variable and one or more independent variables. The model is defined by the equation $Y=\beta0+\beta1X+\epsilon$, where Y represents the height of presidents and X represents the presidential order. The intercept, B0, represents the estimated height when all independent variables are zero, B1 represents the slope of the regression line, and ϵ represents the error term.

Calculation Process in Linear Regression Analysis

The process of conducting linear regression analysis involves several key steps, from preparing the data to fitting the model and interpreting the results. Here's a detailed breakdown of each step in the context of our dataset on U.S. Presidents' heights.

1. **Data Preparation**

Selection of Variables:

Independent Variable (X): The 'order' column from the dataset is selected as the independent variable. It represents the chronological order of each president's term in office. This variable is used to predict or explain changes in the dependent variable.

Dependent Variable (Y): The 'height(cm)' column is chosen as the dependent variable. It represents the height of each president in centimeters. The goal is to see how this variable changes with the presidential order.

- **Data Structuring:**

The data is structured to align with the requirements of a linear regression model. Specifically, the independent variable X and the dependent variable Y are extracted from the dataset and reshaped if necessary (e.g., converting X into a 2D array for compatibility with certain libraries).

2. **Model Fitting**

Creating the Linear Regression Model:

A linear regression model is created using a statistical software or a programming library (like Python's Scikit-Learn). This model is designed to find the best-fit line through the data points.

- Estimating Coefficients:

The model estimates the coefficients$\beta0$ (intercept) and $\beta1$(slope). This is typically done using the least squares method, which minimizes the sum of the squares of the differences between the observed and predicted values.

- **Fitting the Model to Data:**

The model is then 'fitted' to our specific dataset. This involves the model 'learning' from our data by adjusting its parameters ($\beta0$) and $\beta1$ to minimize the difference between the predicted heights and the actual heights of the presidents.

3. **Predictions and Plotting**

- Making Predictions:

Using the fitted model, predictions are made for Y (height) based on X (presidential order). These predictions show what the model expects the height to be for each presidential order, based on the learned linear relationship.

- Visualization:

The results are visualized, typically in a scatter plot with the actual data points and a line representing the model's predictions. This plot helps in visually assessing how well the model fits the data.

The actual heights of the presidents are plotted as individual points.

The predicted heights, based on the model, are plotted as a line. This line represents the best-fit linear relationship as determined by the model.

- Analyzing the Plot:

By examining the plot, we can observe the distribution of data points around the regression line. A closer alignment of points to the line indicates a better fit and a stronger linear relationship.

Detailed Analysis of Results from the Linear Regression

After completing the linear regression analysis of the U.S. Presidents' heights dataset, we obtain several key results, including the model coefficients and various performance metrics. These results help us understand the nature of the relationship between the presidential order and the height of the presidents. Let's break down these results in detail:

1. Model Coefficients

- Intercept (β0): The intercept of the regression line was calculated to be approximately 175.47 cm. This value represents the model's prediction for the height of a president when the presidential order (independent variable X) is zero. While this scenario is not practically possible (as there is no 'zeroth' president), the intercept is a key part of the linear equation that helps position the regression line appropriately on the plot.
- Slope (β1): The slope was found to be approximately 0.1898 cm/order. This value indicates the change in the predicted height of the president for each unit increase in the presidential order. In simpler terms, for each successive president, the model predicts an average increase in height of about 0.1898 cm. This suggests a very gradual upward trend in the heights of presidents over time.

2. Performance Metrics

- Mean Squared Error (MSE): The MSE was calculated to be approximately 41.97. The MSE is a measure of the average of the squares of the errors, i.e., the average squared difference between the estimated values and the actual value. A lower MSE value indicates a better fit of the model to the data. In this context, the MSE provides an idea of the average deviation of the predicted presidential heights from the actual heights.
- R-squared (R^2): The R-squared (R^2) value of 0.127 indicates that 12.7% of the variation in presidential heights can be explained by the linear model based on the presidential order, indicating a relatively weak explanatory power of the model.

3. Interpretation of Results

- The positive slope suggests a minor upward trend in the heights of presidents over time. However, the magnitude of this slope is very small, implying that the change in height across presidencies is quite gradual and perhaps not particularly significant.
- The low R^2 value indicates that the presidential order does not strongly predict the height of the presidents. This means that while there might be a slight upward trend, the presidential order alone does not explain much of the variation in the presidents' heights. Other factors, likely genetic, environmental, or nutritional, might play a more significant role in determining height.
- The intercept and slope, taken together, provide a linear formula that can predict the height of a president based on their order of presidency. However, given the low R^2 value, predictions made using this model should be taken with caution.

Description of the code:

Importing libraries:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.linear_model impor
from sklearn.metrics import mea
```

- pandas: A library used for data manipulation and analysis. It's particularly well-suited for working with structured data (like CSV files).
- matplotlib.pyplot: A plotting library that provides a MATLAB-like interface for creating various types of graphs and plots.
- numpy: A library for numerical computing in Python, often used for working with arrays.
- LinearRegression: A class from the Scikit-learn library that implements linear regression.
- mean_squared_error, r2_score: Functions from Scikit-learn used to compute the mean squared error and R-squared score of the model, respectively.

Loading Datasets:

```
data = pd.read_csv('preside
data.head
```

This part of the code loads the dataset from a CSV file using pandas.

Preparing data for linear regression:

```
# Preparing the data for li
X = data['order'].values.re
Y = data['height(cm)'].valu
```

- X: Represents the independent variable (Presidential Order). The data is reshaped into a 2D array, which is a requirement for Scikit-learn's regression model.
- Y: Represents the dependent variable (Height of Presidents).

Creating and Fitting the Linear Regression Model

```
model = LinearRegression()
model.fit(X, Y)
```

A LinearRegression model is created and then fitted to the data (X and Y). This process involves finding the best-fit line that minimizes the sum of squared differences between observed and predicted values.

Making Predictions and Calculating Metrics

```
# Making predictions
Y_pred = model.predict(X)

# Calculating metrics
mse = mean_squared_error(Y, Y_pr
r2 = r2_score(Y, Y_pred)
```

- Y_pred: The model's predictions for the dependent variable based on the independent variable X.
- mse: The mean squared error of the model, calculated as the average of the squares of the differences between the actual and predicted values.
- r2: The R-squared score, representing the proportion of variance in the dependent variable that is predictable from the independent variable.

Plotting the Results

```
# Plotting the results
plt.figure(figsize=(10, 6))
plt.scatter(X, Y, color='blue', label=
plt.plot(X, Y_pred, color='red', label
plt.title('Linear Regression of Presid
plt.xlabel('Presidential Order')
plt.ylabel('Height (cm)')
plt.legend()
plt.show()
```

This section creates a scatter plot of the actual data points (Presidential Order vs. Height) and overlays the linear regression line (fitted line). The plot helps visualize how well the linear model fits the data.

Setting the Figure Size:

- plt.figure is a function from the matplotlib.pyplot library. It creates a new figure for plotting.
- figsize=(10, 6) sets the dimensions of the figure (width x height) in inches. Here, the figure is set to be 10 inches wide and 6 inches tall, providing enough space to clearly visualize the data and the regression line.

Plotting the Actual Data

- plt.scatter is used to create a scatter plot. It plots individual data points on the figure.
- X and Y are the data points to be plotted. X represents the presidential order (independent variable), and Y represents the height of the presidents (dependent variable).
- color='blue' sets the color of the data points. In this case, points are plotted in blue.
- label='Actual Data' provides a label for these data points, which will be used in the legend.

Plotting the Fitted Regression Line

- plt.plot is used to draw a line graph.
- X and Y_pred represent the points through which the line will be drawn. X is the same as before, and Y_pred is the array of predicted heights obtained from the linear regression model.
- color='red' sets the color of the line. The choice of red here helps distinguish the fitted line from the actual data points.
- label='Fitted Line' provides a label for the regression line, which will be displayed in the legend.

Adding Title and Labels

- plt.title sets the title of the plot.
- plt.xlabel and plt.ylabel set the labels for the x-axis and y-axis, respectively. These labels are important for understanding what each axis represents.

Adding a Legend and Displaying the Plot

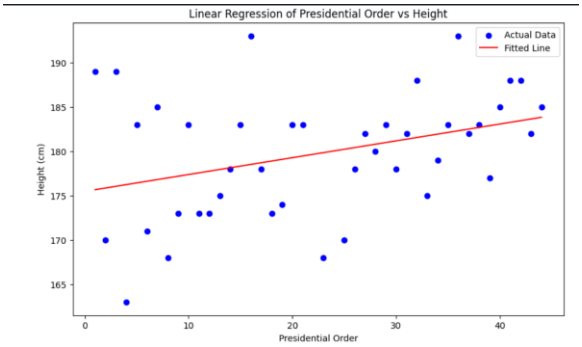
- plt.legend adds a legend to the plot, which helps in identifying the plotted data points and the regression line.
- plt.show displays the plot. This is typically the last line in a plotting sequence and renders the plot in the output.

Output:

```
# Outputting the model coefficien
print("Intercept (β0):", model.in
print("Slope (β1):", model.coef_[
print("Mean Squared Error (MSE):"
print("R-squared (R²):", r2)
```

- Finally, the code prints out the model's intercept (β0), slope (β1), mean squared error (MSE), and R-squared (R^2) score. These values provide insights into the nature of the linear relationship and the performance of the model.

Graph:



Scatter Plot Points (Blue):

- Each blue 'X' mark represents an actual data point in the dataset.

- The horizontal axis (X-axis) shows the presidential order, which is the sequence number of the president's term.
- The vertical axis (Y-axis) shows the height of the presidents in centimeters.
- These points are spread out across the plot, showing the variance in the heights of presidents across different orders.

Fitted Linear Regression Line (Red):

- The red line represents the best-fit line determined by the linear regression analysis.
- It indicates the trend that the linear regression model has found; in this case, it suggests a slight upward trend in the height of presidents as the presidential order increases.
- The slope of the line is positive, which means that on average, later presidents tend to be taller than earlier ones, according to this model.

Axes and Labels:

- The X-axis is labeled "Presidential Order," which indicates that presidents are arranged in the order in which they served.
- The Y-axis is labeled "Height (cm)," which represents the height of each president.
- The title of the graph, "Linear Regression of Presidential Order vs Height," clearly states the nature of the analysis being presented.

Legend:

- The legend in the upper-right corner distinguishes between the actual data points ("Actual Data") and the fitted regression line ("Fitted Line").
- It uses the same color coding (blue for actual data points and red for the fitted line) for easy reference.

Overall, the graph is designed to visually communicate the findings of the linear regression analysis, showing the relationship between the order in which presidents served and their height. The positive slope of the fitted line suggests a trend of increasing height, but as noted in the analysis, the trend is weak, given the considerable spread of the data points around the line.

Conclusion:

The linear regression analysis on the heights of U.S. Presidents relative to their order of presidency provides an interesting statistical exploration but also serves as a reminder of the complexities inherent in interpreting historical and biological data. The study underscores the necessity of careful analysis and the consideration of a multitude of factors when drawing conclusions from unique datasets.

References:

https://www.kaggle.com/datasets