Title: Regression Analysis Report **Date:** 19/12/2023

Course: Data Mining

Accepted by: Mehmet Zirek

1. Introduction

This report presents the results of a regression analysis conducted on a dataset containing information about mobile phones. The dataset was created by a collaboration of the whole class but, it was refined and fixed to be as good as possible for the regression since some missing data or inappropriate format was used.

The primary objective is to predict the price of a mobile phone based on various features. The analysis uses linear regression and includes statistical measures to evaluate the model's performance.

2. Data Description

The dataset includes the following attributes: Opinion, battery power, back camera, clock speed, dual sim, five G, internal memory, mobile depth, mobile weight, number of cores, height, width, ram, screen height, screen width, talk time. The target variable is price.

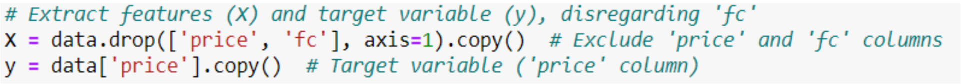
3. Methodology

First of all, to be able to run the file we need the following libraries:

- Pandas
- Matplotlib
- NumPy

Without them the code will not run.

The linear regression model is applied, excluding the 'fc' column from the features during analysis since it was a duplicated value, same as the back camera. Code below:



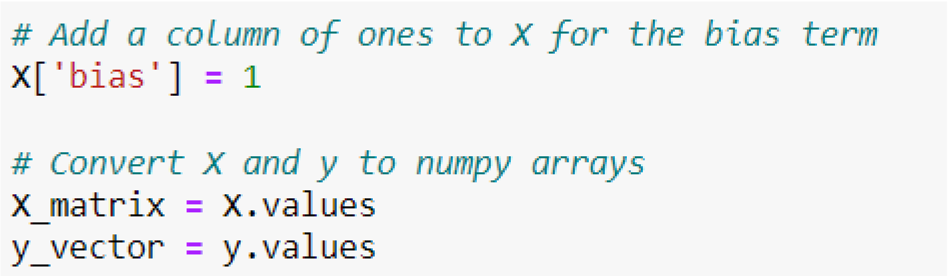
```
# Extract features (X) and target variable (y), disregarding 'fc'
X = data.drop(['price', 'fc'], axis=1).copy()  # Exclude 'price' and 'fc' columns
y = data['price'].copy()  # Target variable ('price' column)
```

The regression parameters are calculated using the normal equation, and predictions are made based on the obtained parameters.

4. Results

First, I converted the x and y variables mentioned above to NumPy arrays.

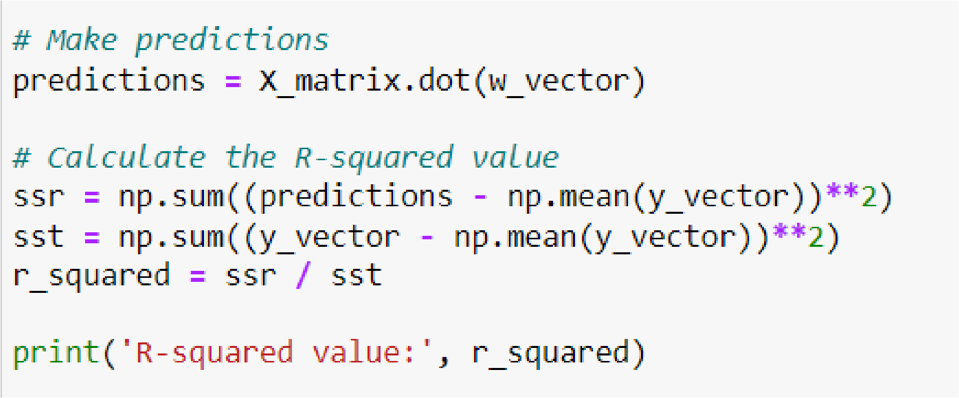
Then I calculated the regression parameters using the normal equation, using NumPy transpose and inverse functions as shown below.



```
# Add a column of ones to X for the bias term
X['bias'] = 1

# Convert X and y to numpy arrays
X_matrix = X.values
y_vector = y.values
```

I also found the R-squared (coefficient of determination) which concretely is the value in a regression analysis is a statistical measure that represents the proportion of the variance in the dependent variable (target variable) that is explained by the independent variables (features) in the model. In other words, it indicates the goodness of fit of the model. Code Below:



```
# Make predictions
predictions = X_matrix.dot(w_vector)

# Calculate the R-squared value
ssr = np.sum((predictions - np.mean(y_vector))**2)
sst = np.sum((y_vector - np.mean(y_vector))**2)
r_squared = ssr / sst

print('R-squared value:', r_squared)
```

5. Statistical Analysis

To give a clearer view of the evaluation, I also made some statistical analysis including:

- Residuals:

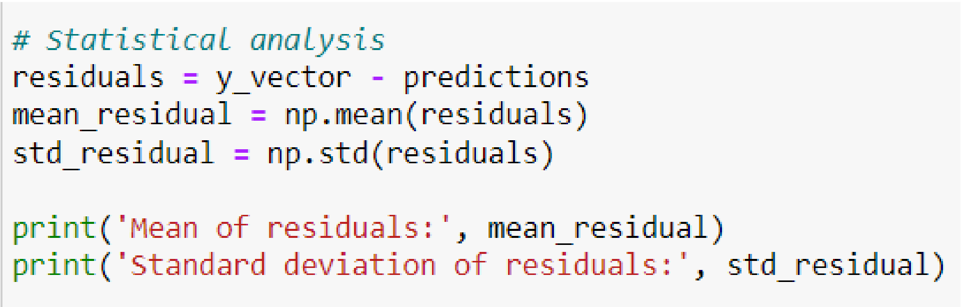
Represent the errors or the differences between the observed and predicted values.

- The mean of residuals:

Provides an indication of whether, on average, the model tends to overestimate or underestimate the actual values. A mean close to zero suggests that, on average, the model predictions are accurate.

- The standard deviation

Measures the spread or dispersion of the residuals around the mean. A lower standard deviation indicates that the residuals are generally close to the mean, while a higher standard deviation suggests more variability in the errors.

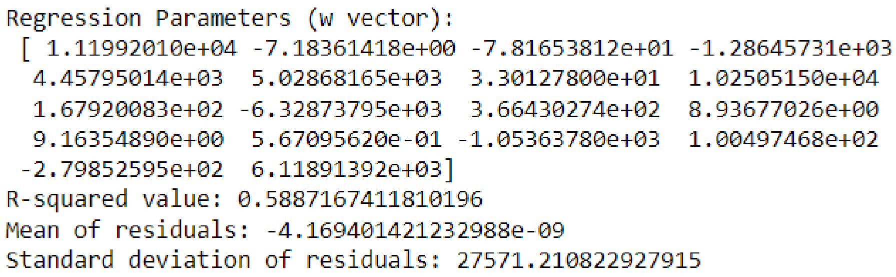


```python
# Statistical analysis
residuals = y_vector - predictions
mean_residual = np.mean(residuals)
std_residual = np.std(residuals)

print('Mean of residuals:', mean_residual)
print('Standard deviation of residuals:', std_residual)
```

6. Outputs

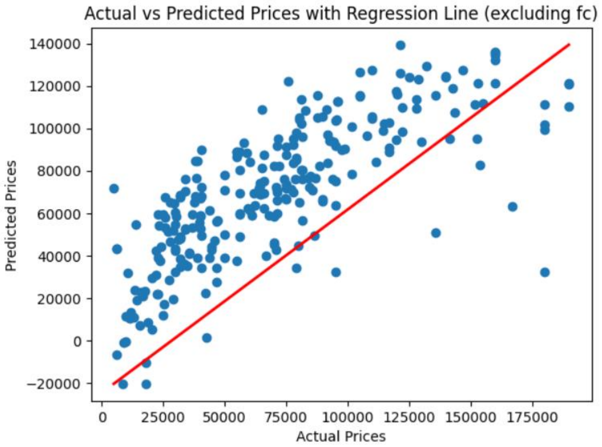
The following snippet will have the outputs, including the w-vector, and all the statistical analysis mentioned above:



```
Regression Parameters (w vector):
 [ 1.11992010e+04 -7.18361418e+00 -7.81653812e+01 -1.28645731e+03
  4.45795014e+03  5.02868165e+03  3.30127800e+01  1.02505150e+04
  1.67920083e+02 -6.32873795e+03  3.66430274e+02  8.93677026e+00
  9.16354890e+00  5.67095620e-01 -1.05363780e+03  1.00497468e+02
 -2.79852595e+02  6.11891392e+03]
R-squared value: 0.5887167411810196
Mean of residuals: -4.169401421232988e-09
Standard deviation of residuals: 27571.210822927915
```

7. Visualization

Created with the Matplotlib python library, the scatter plot below illustrates the relationship between actual and predicted prices, with a red regression line.



The linear regression model provides insights into the relationship between the selected features and the price of mobile phones. Further analysis and refinement of the model may be conducted to improve its predictive performance.

Thank you very much for your time!

This project will also be on Github on this link:

https://github.com/kjahaj/Market-Study-Machine-Learning

Worked by: Klei Jahaj