Exercise 2 Report

Paddy Milner[1]

[1] *Department of Physics, University of Bristol*

17.1.2026

Word count: 1125

Abstract

In this exercise, a large amount of experimental data was manipulated and analysed in order to determine the correlations between various experimental properties, as well as to determine the strength of the linear relationship between fall height and fall time for sheres of different materials and sizes.

1 Introduction

In this exercise, a large set of experimental data was supplied for manipulation and analysis. We will first clean the data set, removing any incorrect or anomalous data points, followed by some initial analysis on the data. We will then determine the strength of the correlation between the different features of the experimental data, with a particular focus on how the various parameters affect the fall time. We will the calculate a linear regression using multiple methods, and test each method's effectiveness when compared both to the original data and to each other.

2 Theory and Methods

The sample data was first trimmed to remove any incorrect data, first by removing any lines with incorrect material names, then removing any lines containing non-numeric data, and finally by correcting any negative data, as all values should be positive given the units the data is given in.

Next, the data was filtered by material, and for each material a plot was made of fall time against height, with the data points being coloured based on radius.

In order to determine the correlation between each feature of the dataset, the pandas corr() function was used to generate a correlation matrix, which was then plotted, using a colour map to indicate the strength of the correlation between two features.

A linear correlation was then calculated from the data set, consisting of a coefficient for each feature to give the falling time. This linear fit was first analysed visually by plotting both the true values for falling time as well as the values the linear regresion predicted on the same axis, as well as a line of best fit based on the predicted data. The fit was then analysed quantitively by splitting the data set into a training set and a testing set. The linear fit was generated off the training data, and both the test data and the models prediction of the test data were plotted, and the R^2 values were compared to measure the accuracy of the linear regression. The residuals of the true data and the predicted data were then plotted against the radius.

Another linear regression model was then calculated, this time using the stochastic gradient descent mathod rather than the least squares regression method, as this avoids matrix inversion, which can reduce sompute time for data sets with a large number of parameters. The linear fit was first calculated based off of unscaled data, to confirm the necessity of scaling. The data was then scaled correctly and the linear regression was then recalculated. This model was compared to the model obtained from the least squares method, and finally the model was generated again using a different loss function to compare the differences.

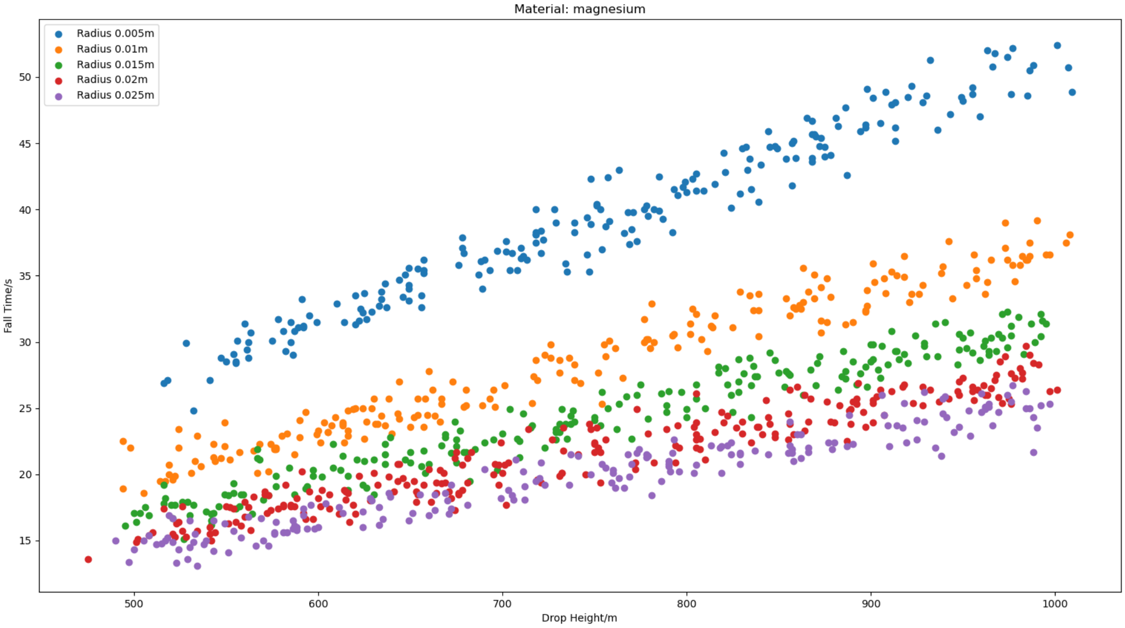

Figure 1: Plot of fall time against fall height for all magnesiusm data.

3 Results and discussion

The cleaned data functioned correctly, with all data being numeric and in a reasonable range, as confirmed by the calculated statistics. The plots of fall time against data appear as expected, with the data for each radius being mostly linear, with an example shown in figure 1. The correlation matrix, shown in figure 2 shows that pressure and temperature have negligable effect on fall time, height has a positive correlation with time, which is to be expected, and density, radius and mass all have significant negative correlations with fall time, which is also intuative, as a larger and therefore heavier sample will fall faster. The only other significant correlations are mass' correlations with density and radius, both of which are positive and are to be expected. After calculating the linear regression, the following coefficients were found for each of the dataset's features:

Feature	Coefficient
Density	-2.84×10^{-3}
Radius	-1060
Mass	34.7
Temperature	-0.0347
Height	0.0251

All of these values are to be expected, with temperature and pressure having small coefficients due to their weak correlation with fall time, radius and density having negative values due to their negative correlation, and height having a fairly large coefficient after taking into account the somewhat large mean value for height.

The data was then filtered to only contain one material, iron, and using our linear regression model the true values of time were plotted against the values the model predicted based off of the other features, shown in figure 3. The predicted data is somewhat close to the true values, though it tends to predict a consistently lower fall time,

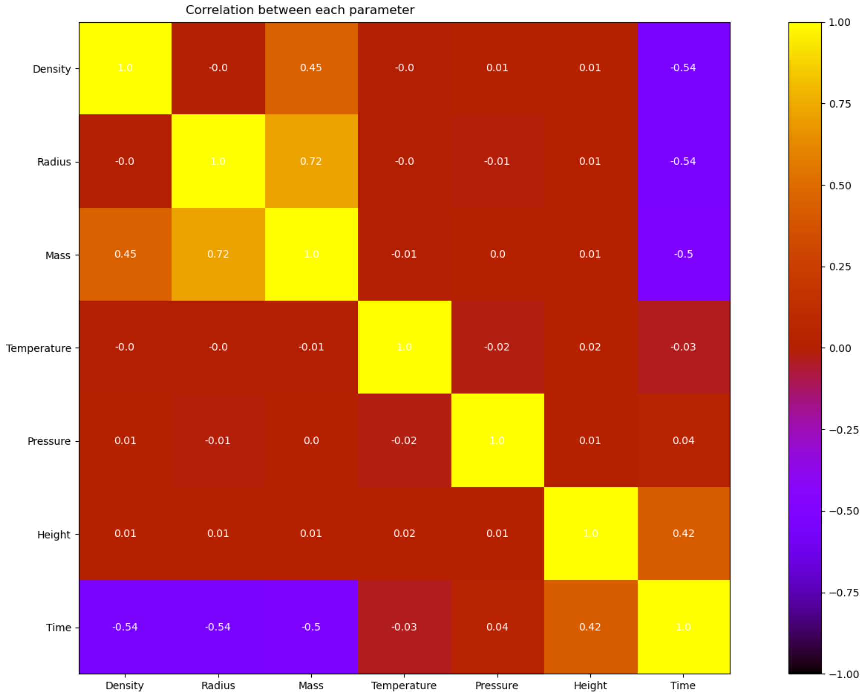

Figure 2: A correlation matrix showing the strength and direction of correlation between the variables in the dataset.

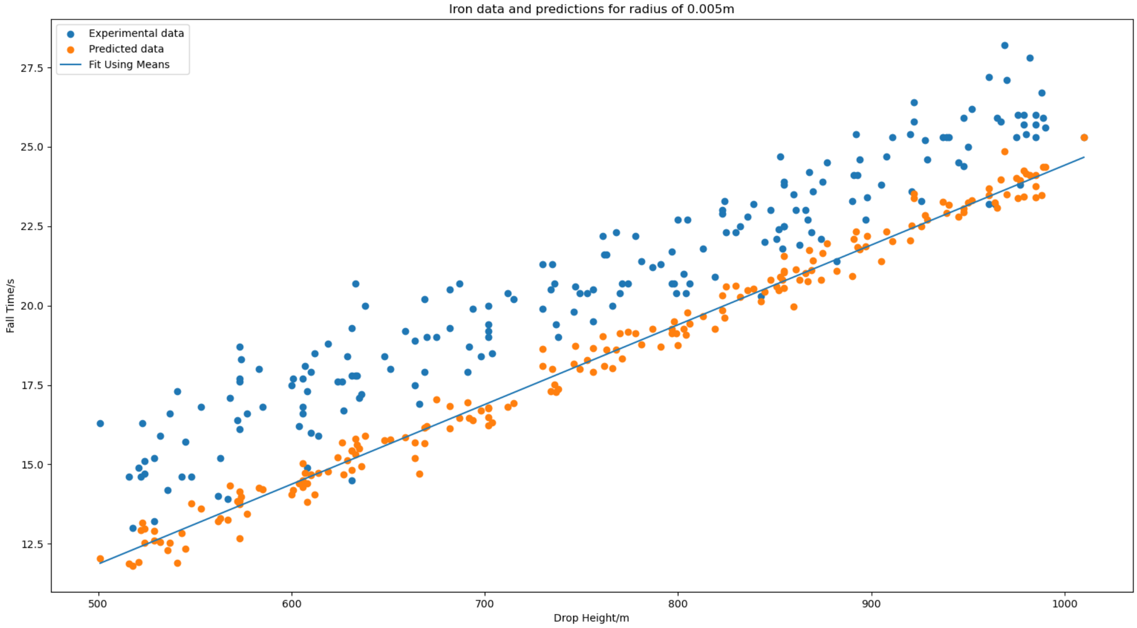

Figure 3: A plot of the true fall time data and the values predicted by the linear regression model.

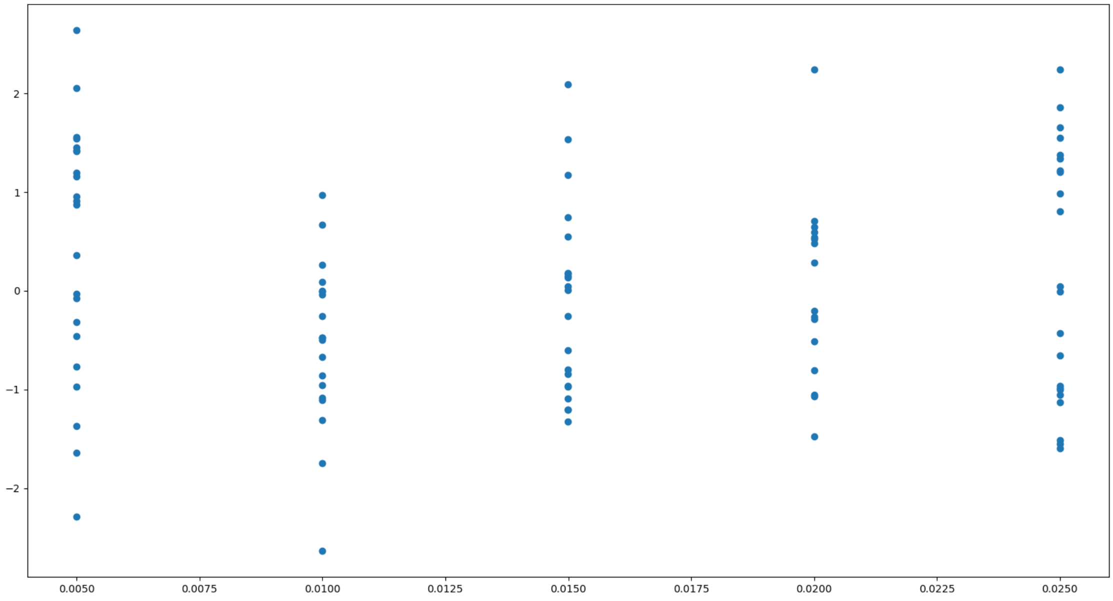

Figure 4: A plot of the difference between the true and predicted values against radius.

and follows a much stronger linear relationship This could likely be improved by a larger sample set, or a more focused one, perhaps only measureing one material or radius, as there is some varience between them. The dataset was then split up into training and test data, and the test data was plotted with the predicted values based off of a linear regression calculated from the training data. The data was split up by radius, and the R^2 value was vcalculated for each radius for both the true and predicted data. The average R^2 value for the true data was 0.812, whereas the average value for the test data was 0.991, again showing that the linear regression model has a tendency to predict a stronger linear relationship than the experimental data gives. The difference between the true values and predicted values were recorded and plot against radius, show in firgure 4. This shows that there is no strong correlation between model accuracy and radius. Though there is varience between the radii, it does not seem to follow any strong relationship.

A linear regression was then calculated using a stochastic gradient descent method. Initially, the model did not return good results, consistently giving an R^2 value on the order of -1×10^{35}. However, after the dataset features were scaled, the model improved greatly, giving an R^2 value of 0.847. From this, the coefficients of the scaled model were de-scaled, and compared to the least-squares method, which can be seen in the table below.

Feature	LS Coefficient	SGD Coefficient	Huber Coefficient
Density	-2.84×10^{-3}	-2.83×10^{-3}	-1.37×10^{-1}
Radius	-1060	-1056	-361
Mass	34.7	34.92	-6.56
Temperature	-0.0347	-0.0327	-0.0288
Height	0.0251	0.0251	0.0205

After scaling, the SGD model matches the least squares model very well, suggesting the SGD is a good alternative.

This method was retried using the Huber loss function, and the coefficients for this can also be seen in the table.

The coefficients for the Huber loss function vary significantly when compared to the other two methods, suggesting it is not a good fit for this dataset.

4 Conclusion

Overall, the data shows clear correlation between several measured features of the dataset, with the largest factor influencing the fall time being height, mass, radius and density. The data follows a clear linear relationship, which is quantifiable using a number of approaches, which all give results matching fairly well with the experimantel data, through there is still room for this to be improved.