Weight Prediction with Linear Regression

Johnnie Oldfield

September 2019

1 Problem Definition

The goal of this project is to illustrate the relationship between a person's height and weight. Given an accurate data set of individual's personal health the model will try to predict someone's weight given their height. Height is not the only component to determine weight, there are others. The important ones includes sex and body mass index (BMI). However, I am going to ignore those factors for this project to keep it linear.

2 How linear regression can predict model

To accomplish the project's goal I will use linear regression analysis. While linear regression is a basic predictive analysis tool, it is also very common and effective. The general goal of using regression is to do two things. 1.) Use a set of predictor variables to predict an output. 2.) Illustrate whether or not which variables are playing a significant role on the output variables, and indicates the impact of each on the outcome variable[1]. This project will be using simple linear regression as there is one dependent and one independent variable. Here is the formula:

$$\hat{Y} = b_0 + b_1 x$$

$\hat{Y}$ is what the model is trying to replicate(weight), x is the independent variable(height), and b_0 & b_1 are the parameters for the model [2].

3 Data set

All of the data used is from the National Health Interview Survey (NHIS). The US Census Bureau has collected data for them for more than 50 years. The information was collected from almost 5,000 participants in 2007. Some of whom refused to answer some details. Details include, BMI, hours of sleep, weight, and height. Values from individuals who did not want to answer were set at extremes. For example, weight and height were set to 996, 998 and 96, 98 respectively. These entries are ignored.

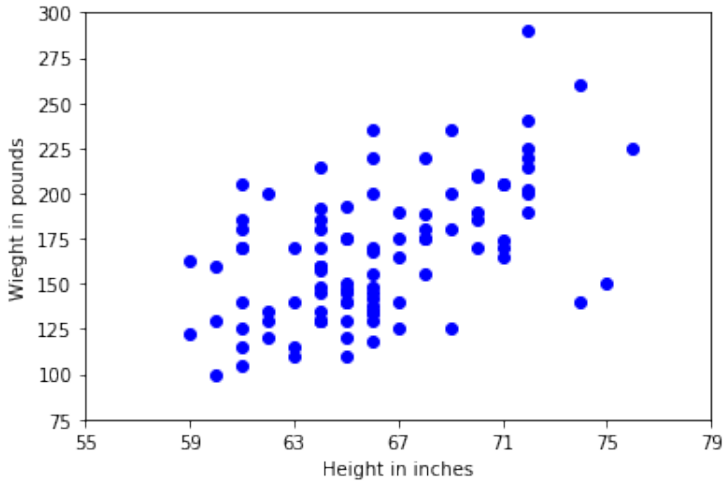

Figure 1: A sample of 100 individuals from the data set

4 Proposed Solution

Using the accurate data set from the NHIS, I am going to plot the data points and try to illustrate the positive relationship between a person's weight and their height. The linear regression model will predict weight based on height. To find the best fitting line for the data I am going to use the least squares method. The returned results are saved to be plotted later. I am going to compare different sample sizes; 5, 50, 100, 500, and 1000 persons.

5 Coded Solution

Listing 1: Data set manipulation

```
import pandas as pd
df = pd.read_csv('/content/drive/My Drive/NHIS 2007 data.csv')

# remove extremes from entries who did not answer survey
df = df[df.weight <= 900]
df = df[df.height <= 90]

# get weights and heights
X = df['height']
y = df['weight']

try:
  X = X.to_numpy(copy=True)
  y = y.to_numpy(copy=True)
except:
  print(type(X), type(y))
```

Listing 1 shows how I am getting the data from the data set using pandas. The data set extremes for height and weight are trimmed and then grabbed. However, they are saved as pandas data frames and need to be changed into numpy arrays. Which is the purpose of the try-except.

Listing 2: lin_regress function

```
def lin_regress(X, y, size):
  Y = lambda b0, b1, x: b0 + b1*x
  # reduce size
  X = np.resize(X, (size,))
  y = np.resize(y, (size,))
  # set domain
  dom = np.linspace(59, 76, 100)

  A = np.power(X[np.newaxis].T, [0, 1])
  b = np.linalg.lstsq(A, y, rcond=None)[0]
  vals = Y(b[0], b[1], dom)

  # plot
  fig, ax = plt.subplots(1,1)
  ax.plot(X, y, 'ob')
  ax.plot(dom, vals, '-r')
  ax.set_xlabel('Height in inches')
  ax.set_ylabel('Weight in pounds')
  print(stats.pearsonr(X, y))
  return vals
```

Listing 2 shows my line_regress function which takes in the data and a size. The data is reduced to size and put through the least squares function. Which gets b0 and b1 which are thrown into the lambda to create the model. The line is plotted and returned by the lin_regress function. I am using the stats.pearsonr function to print out the correlations of X and y.

6 Evaluation

From the beginning you can already see the positive relation ship between weight and height with just 5 data points, illustrated in Figure 2 (although not accurate). The high positive relationship is also confirmed by it's 0.79 correlation.

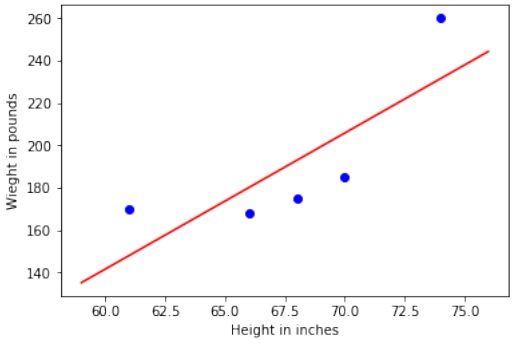

Figure 2: A sample of 5 individuals from the data set. Correlation: 0.79

After 50 data points each test shows the relationship get slightly weaker but still positive. With each test the model get increasingly more accurate and the correlation approaches 0.51. Which can be seen in Figure 3 and Figure 4

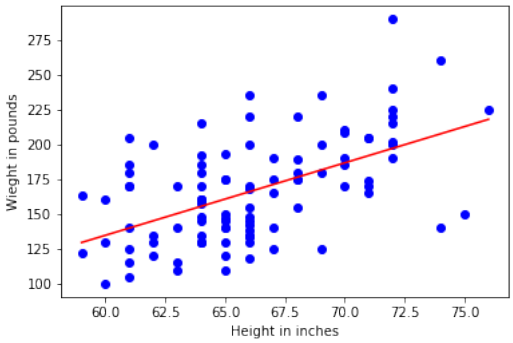

Figure 3: A sample of 100 individuals from the data set. Correlation: 0.54

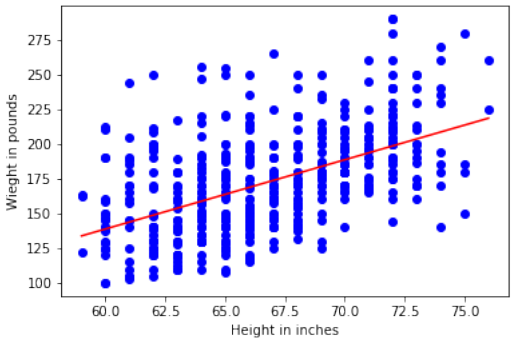

Figure 4: A sample of 500 individuals from the data set. Correlation: 0.52

Figure 5 is the final plot that shows each result from the different samples. They are put in the same plot to be easier to compare.

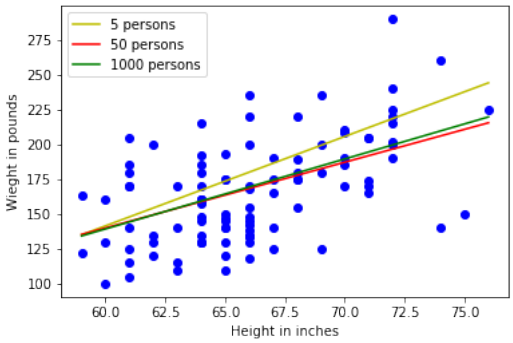

Figure 5: Each result from the different tests.

Table 1: Test sizes and correlations.

Tests	
Sample Size	Correlation
5	0.79
50	0.47
100	0.54
500	0.52
1000	0.51

7 Discussion

The project outcome went as expected for the most part. One point I found interesting was how the correlation greatly decreased to 0.47 from a sample size of 5 to 50. It then went up to 0.54 and slowly decreased to 0.51, see Table 1 above. The model did well to show the positive relationship between a person's height and weight. It intuitively makes sense that the taller the person the more they weigh. However, as mentioned in Section 3 there a other factors that play a role in a person's weight.

8 Future Work

My linear regression model is misleading since there are hidden factors affecting weight. Future work would be to acquire a more detailed data set that includes those factors (such as sex, BMI, exercise frequency, etc.) and then create a multi-variate model using a different form of regression that illustrates the influence of each factor. This would allow insight in not only how height, but how other variables as well.

References

[1] `https://www.statisticssolutions.com/what-is-linear-regression/`.

[2] Least squares regression line. 2014. `https://www.statisticshowto.datasciencecentral.com/least-squares-regression-line/`.