Robert Sandel

Math 4753

Linear Regression Project

The purpose of this study was to determine whether there exists a predictable correlation between the measurements of ones’ head size and the weight of his or her brain therein. The subjects upon which the data were collected were all identified as relatively healthy individuals and were documented by the Middlesex Hospital in Oxford, United Kingdom in 1905. It should be noted, though, that of those deemed ‘healthy’ by the study, some included deaths by various diseases like cancer or Tuberculosis; however, these complicating conditions were assumed to have negligible effect on the parameters being measured. This assumption was derived through a comparison of the brain weight among subjects who died from wasting diseases, such as Tuberculosis, to those who died from accidental deaths or acute illness, and the results showed no significant difference in correlation. According to the initial study, the main goal of the investigation was to obtain a reconstruction formulae in turn to predict the approximate weight of the brain based on head size. In order to identify whether a correlation exists and to further quantify the strength of said correlation, a least squared regression model was employed, and the validity of the model was vigorously evaluated using techniques including an analysis of the residuals against the fitted, Cook’s distances, and a Shapiro-Wilks test of the normality. Based on the evidence, it is believed that a linear trend best fits the underlying trend between and individual’s head size and the weight of his or her brain.

The data was collected through the use of a variety of tools focused on measuring the length of the head, the diameter of the top of the skull, the height of the cranium, the horizontal circumference around the middle of the skull, and the varying accompanying deviations in the skull. A diagram of how the measurements were taken is shown in **Figure 1**. This resulted is measurements of the length, breadth, height, and circumference of the skull, which were all combined into a single variable described as ‘head size’, which was recorded as a volume in units of cubic centimeters. After the volume was measured according to these measurement techniques, the brain was removed from the cranial cavity and was weighed. The resulting brain weights were recorded in units of grams. Using the variables of skull volume in cubic centimeters and brain weight in grams, a least squares linear regression was generated attempting to correlate the weight of the brain to the relative size of the head. With this method of least squares regression, we assume that the data are independent and the errors are distributed normally with a mean of zero and a constant standard deviation.

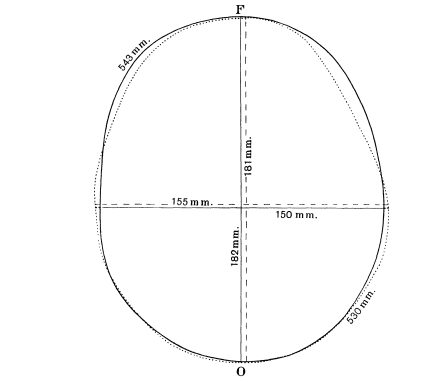


Figure 1 – Image of how head size measurements were taken

In order to gauge the general trend of the data, a trend scatter plot was generated using an f value of 0.5, and a relatively strong linear relationship can be tentatively seen below in **Figure 2.**  The middle line represents the running mean of the data, and the two dotted lines on the outskirts represent the average errors associated with .

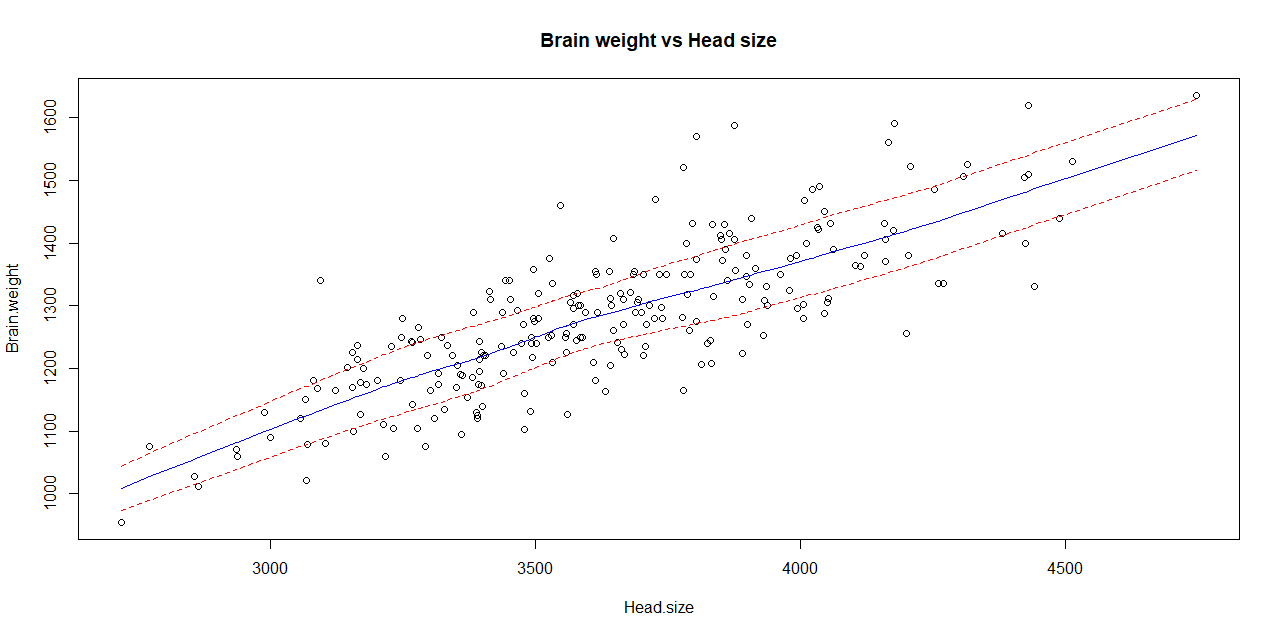


Figure 2 – Trendscatter plot for brain weight vs. head size. Head size is the independent variable, and brain weight is the dependent variable.

Using R, a linear model was generated having two distinct parameters of interest: , or the intercept, and , or the slope. The full generalized equation representing the liner regression is presented below, with being the independent variable and as the dependent variable for each . This equation includes representing the random error attempting to quantify the discrepancy between the real underlying trend and the generalized approximation quantified by this model.

Due to the assumptions behind the underlying data, the and variables are point estimators, and are changed to   and respectively. As such, the equation becomes:

In this equation now becomes, and the values and are determined using the derivation of least squares regression. Principle to this method is minimizing the sum of the squared residuals. This method aptly takes into account the deviations each data has to the trend being fit, but, for better or for worse, the data points falling within the extremes of deviation are exaggerated in regard to their effect on how the overall trend fits to the data. This means it is especially important to validate the data and to remove any clear outliers because they can greatly influence how the regression is fit. The sum of squared residuals, or SSE, is defined below:

The estimator for is determined by setting the derivatives of the sum of the squared residuals with respect to and equal to 0. The resulting definition for is given below using only values from the data itself.

Once the variable is calculated using the above equation, the variable is easily determined by plugging it, along with the means of the dependent and independent variables, into the basic definition of linear regression, as shown below.

Using R, the values for and were determine to be 0.26343 and 325.5742 respectively. This results in the least squares linear regression given below.

This regression model was then plotted against the data to quickly gauge how valid the regression is in regards to the data itself. As can be seen in the plot below (**Figure 3**), the trend appears to fit the data relatively well, especially with how the data tentatively appears to be equally and randomly distributed across this regression model. Following these quick evaluations of the plot, the preliminary indication is that the regression model defined above is a good fit for this data; however, further analysis will be employed to verify the randomness of the residuals about the regression and the validity of the overall model.

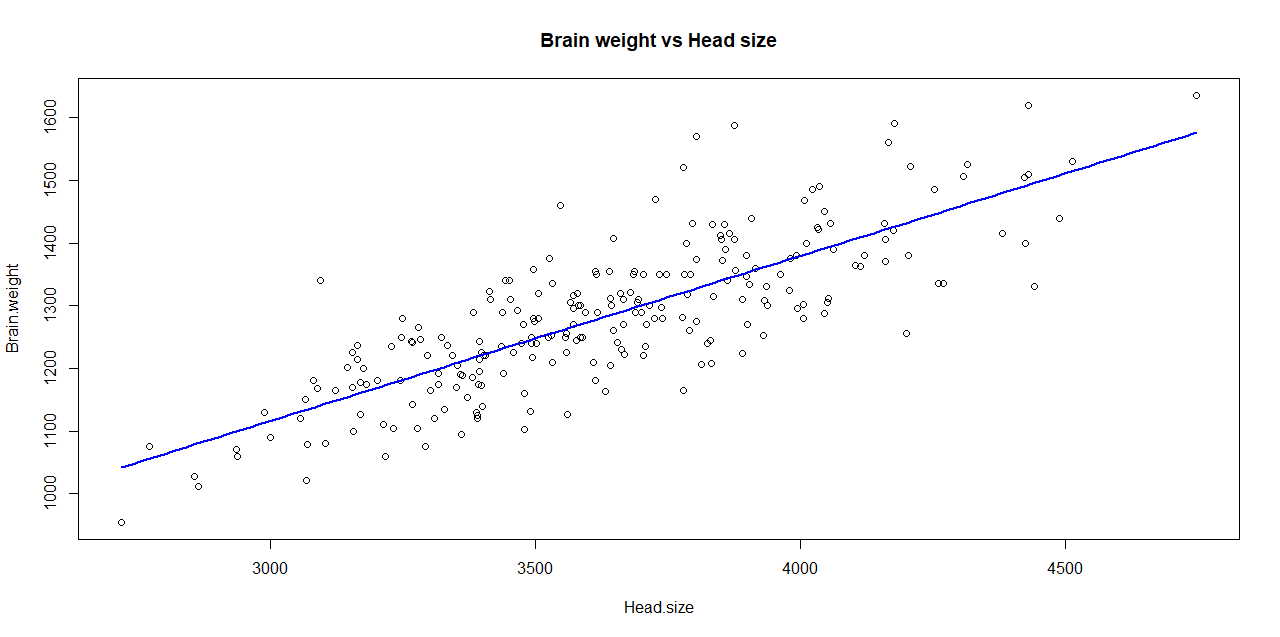


Figure 3 – Regression model superimposed over brain weight plotted against head size

The first step in validating the correlation is to determine whether the deviations between the observed data and the expected values and to verify that there is no trend in how the data is deviating from the expected. This is done by plotting the residuals against the expected values, or the fitted values, for each point of data. Ideally there should be no correlation between the two with a completely horizontal line representing complete independence. As can be seen in the **Figure 4**, shown below, the horizontality in the line, representing the running mean, indicates very little trend with the residuals across the model. This is further indication that the underlying trend is linear, but further analysis will still be employed to better validate the model.

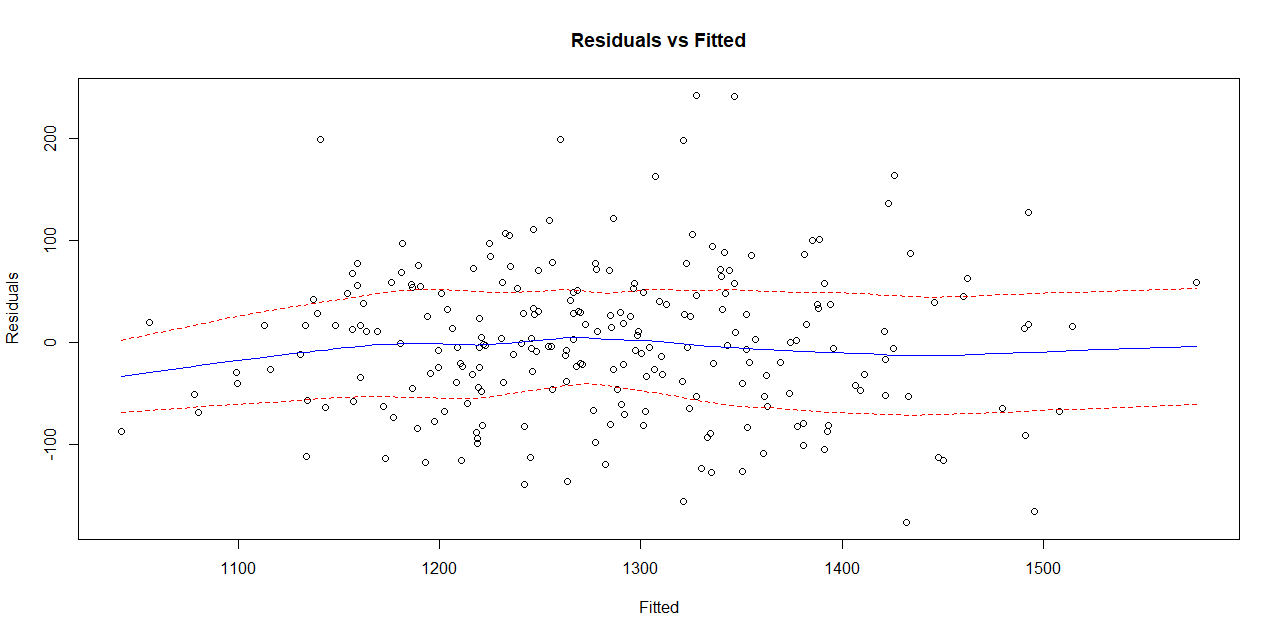


Figure 4 – Residuals versus fitted values from the trend line. The blue line represents the running mean across the data and the dashed red lines indicate the standard error.

To summarize the plotting parameters and the fitting of the regression, various plots in **Figure 5** below were generated. The first represents the regression plotted against the data itself. The second includes distances each data is from the expected at those points based on the regression. The third represents the distance the regression is from the mean, and the final plot depicts the distance each data is from the mean of the dependent variable, or in this case, the brain weight. It is important to note with this plot that the horizontal represents zero correlation, and a trend in the deviations from this line indicates correlation.

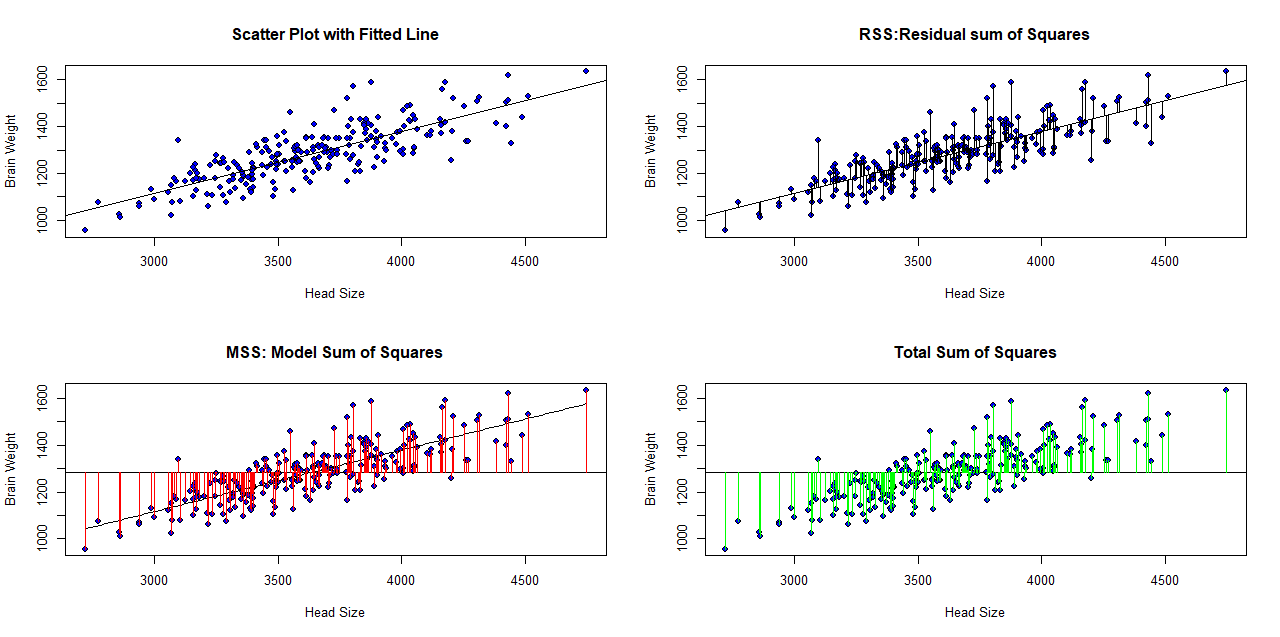


Figure 5 – Sum of squares. (Top left) Line of best fit, (top right) residual sum of squares, RSS, (bottom Left) model sum of squares, MSS, (bottom right) total sum of squares, TSS.

To further evaluate the validity of the regression, a Cook’s distance plot (**Figure 6**) was generated by taking the relative distances each data point is from the regression itself. This not only gives further indications as to possible underlying trends in the deviations from the expected, but it can very clearly show which data represent possible outliers and which have the strongest effect on the fit of the regression. It can be seen that only a few points lay in drastic exception from the expected.

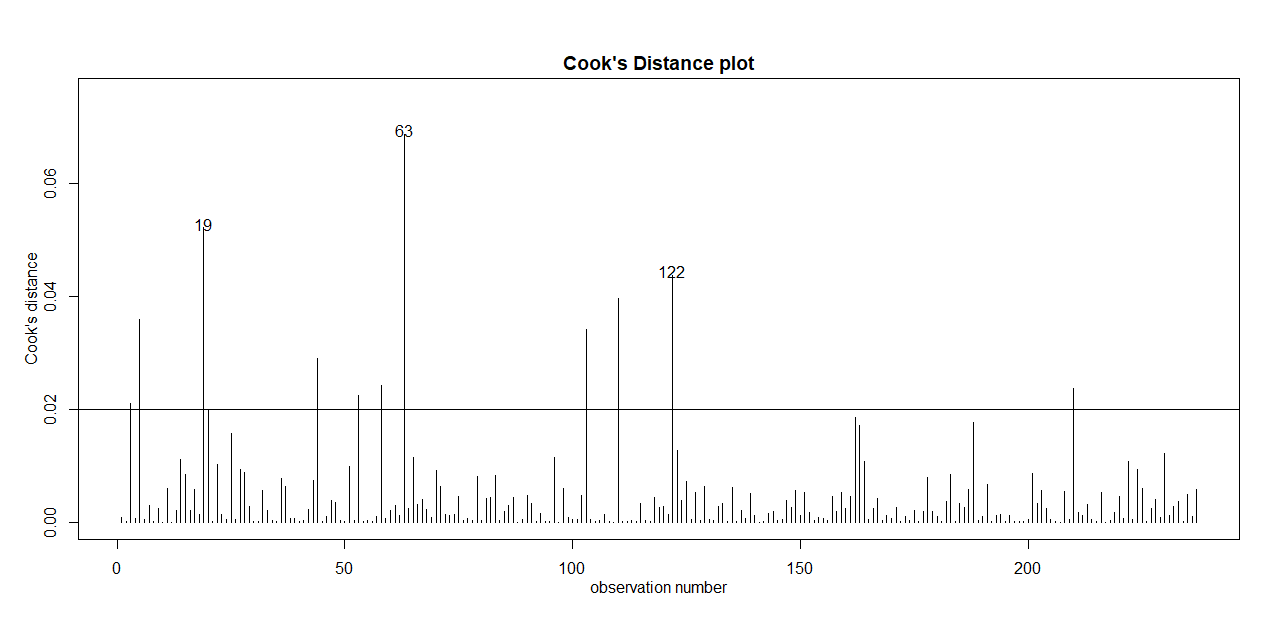


Figure 6 – Cook’s distance plot used to determine outliers from trend. A value of 0.02 was determined through 4/n in which n was the total number of observations

Although the regression appears to fit the observed data relatively well, and there appears to be no trend in the residuals vs. the fitted, a Shipiro-Wilks normality test was used to quantify the normality of the errors. The results are shown below in **Figure 7**, and it can be clearly seen that the data falls very strictly along the Q-Q plot with the only deviations existing on the extremities of the data set. Furthermore, the plot on the right shows a very clear normality, which is further supported by the fact that the p-value is 0.024, well below the 0.5 cutoff. This not only verifies the data is distributed randomly across the expected from this model, but that the correlation is statistically significant.

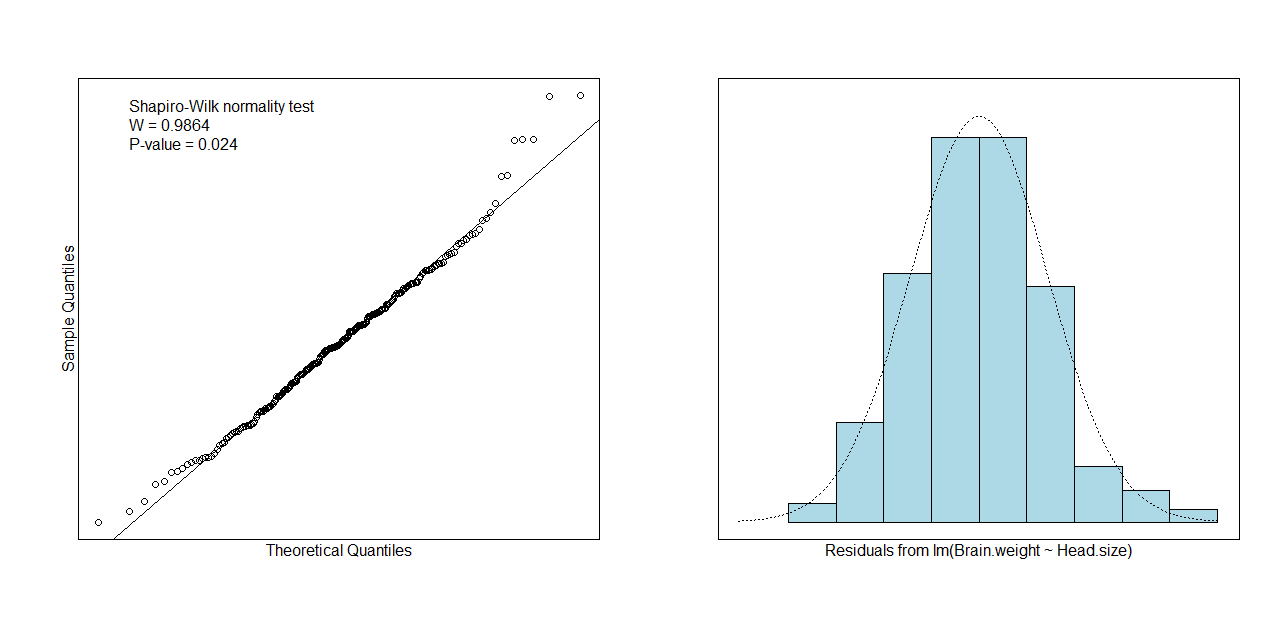


Figure 7 – (Left) Q-Q plot and (Right) Shapiro-Wilk Test to test for normality.

Using the regression model in R, 95% confidence intervals were generated for both and , and the results are shown below.

95 % C.I.lower 95 % C.I.upper

(Intercept) 232.7008 418.44609

Head.size 0.2380 0.28886

From these intervals we can reliably say the true and underlying mean for the slope, or , lies between 0.238 and 0.28886 cm3, and the value of the intercept , or the value of brain weight at zero head size, lies between 232.7008 and 418.44609 gm.

With the observed normality in the errors, the lack of trend with the residuals against the fitted, and the p-value being well below the 0.05 cutoff for a standard 95% confidence interval, the use of the initially determined estimators for the slope and intercept appears to be justified in predicting values of brain weight according to measurements of head size.