I. Cover page

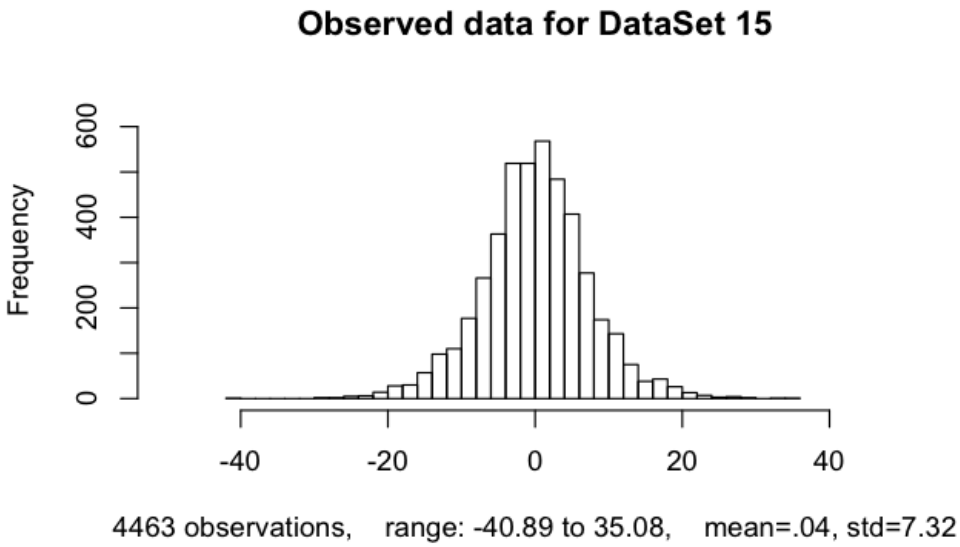
a. document: smde_ida_deliverable_1_diego_garcia-olano.pdf
b. SMDE Input Data Analysis Deliverable 1 – Data Set Assignment Id 15
c. Document Version Number: 1
d. Printing Date: November 10, 2013
e. Location of electronic version of file: none
f. Department & University: FIB, UPC

II. Description of process for fitting input distribution.

a. Statistical summary: graphical and moments

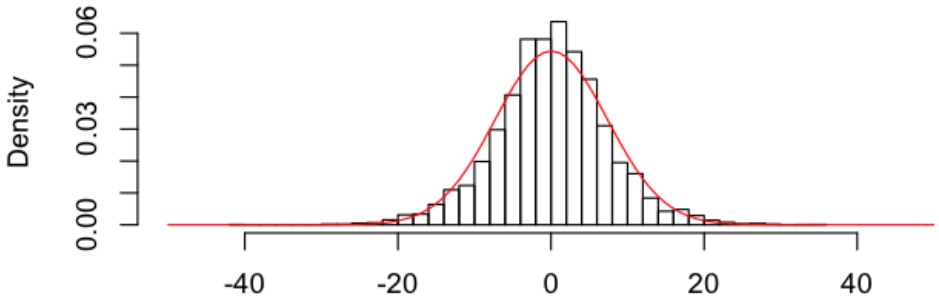
Using the R language, we first loaded the data file assigned into a variable d15, and then using summary() we observed some of our data moments, notably that our data spanned a range from -40.89 to 35.08 with a mean nero zero at .045. Using std() and length() , we then obtained the standard deviation, 7.329, and length of the dataset, 4463, respectively.

Then we decided to plot a histogram of our data using hist() and got the following histogram.



b. Selection of candidate distributions: matching theoretical to sample moments.

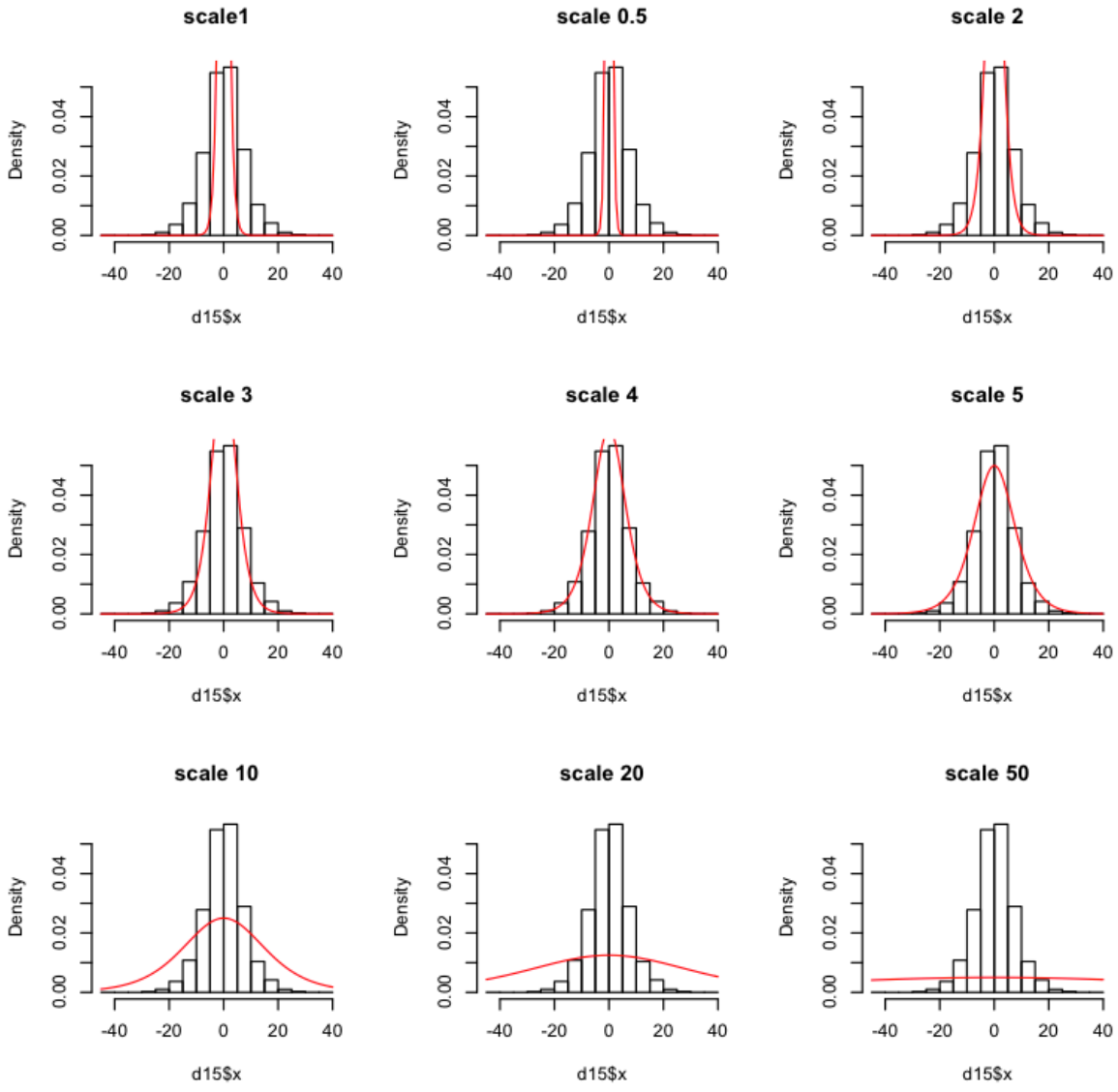
Based on first assessment of our histogram, we note the data is symmetrical with long tails and contains both postive and negative numbers. Looking at the data with head(), we also note that values appear to be of a continous nature. These factors lead us to the judgment that our data may come from a normal gaussian distribution with mean .04 and standard deviation of 7.32. As the data is centered near zero, we don't take any shift into account, and plot a curve of our theoretical distribution, using curve() and dnorm(), the density function of a normal distribution, against our observed data.



This initial fititng seems reasonable enough, so we use fitdistr(), the maximum likelihood function provided in the MASS library, to estimate model parameters we can then use to quickly test for goodness

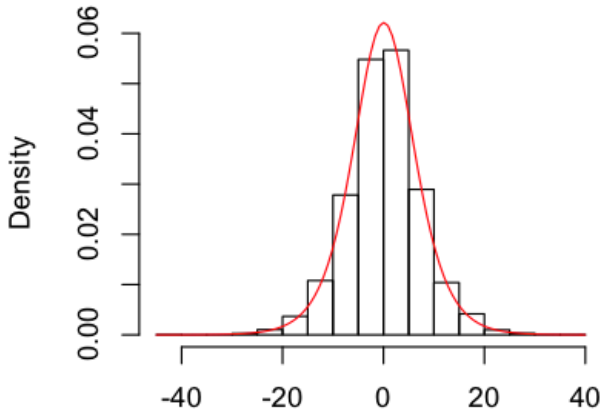
of fit with ks.test(), the Kolmogorov-Smirnoff Test. We obtain a very low p-value however and thus must reject the null hypothesis that our observed data came from the normal distribution.

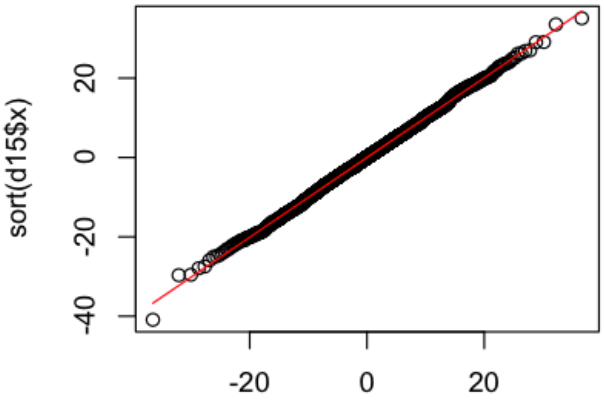
We then decide instead to look at another possible distribution from which our data could have been derived and again due to its symmetrical distribution shape, continuous nature and inclusion of negative values, we decide to evaluate it against a theoretical logistic distribution. This time however in addition to plotting a potential curve against our data, we also wish to calculate a linear regression of our model with respect to a theoretical logistic distribution whose location we'll first set to be the mean of our observed dataset and whose scale parameter we'll derive by plotting many QQplots of different scale values to see which visually appears to fit the best.



Its evident that scale=4 seems to be the best value along with the mean of 0.045 we held constant. To get better values and assess whether our visual findings are correct, we use fitdistr() with our dataset to see what model parameters it would return for a logistic distribution. This step gives us a location parameter of 0.062 and scale parameter of 4.02 which are both close to our approximations.

Then using cor(), we observe there is a high correlation between our observed data and our estimated logistic distribution, and plot the fitted values from our linear model against a probability plot of our data.



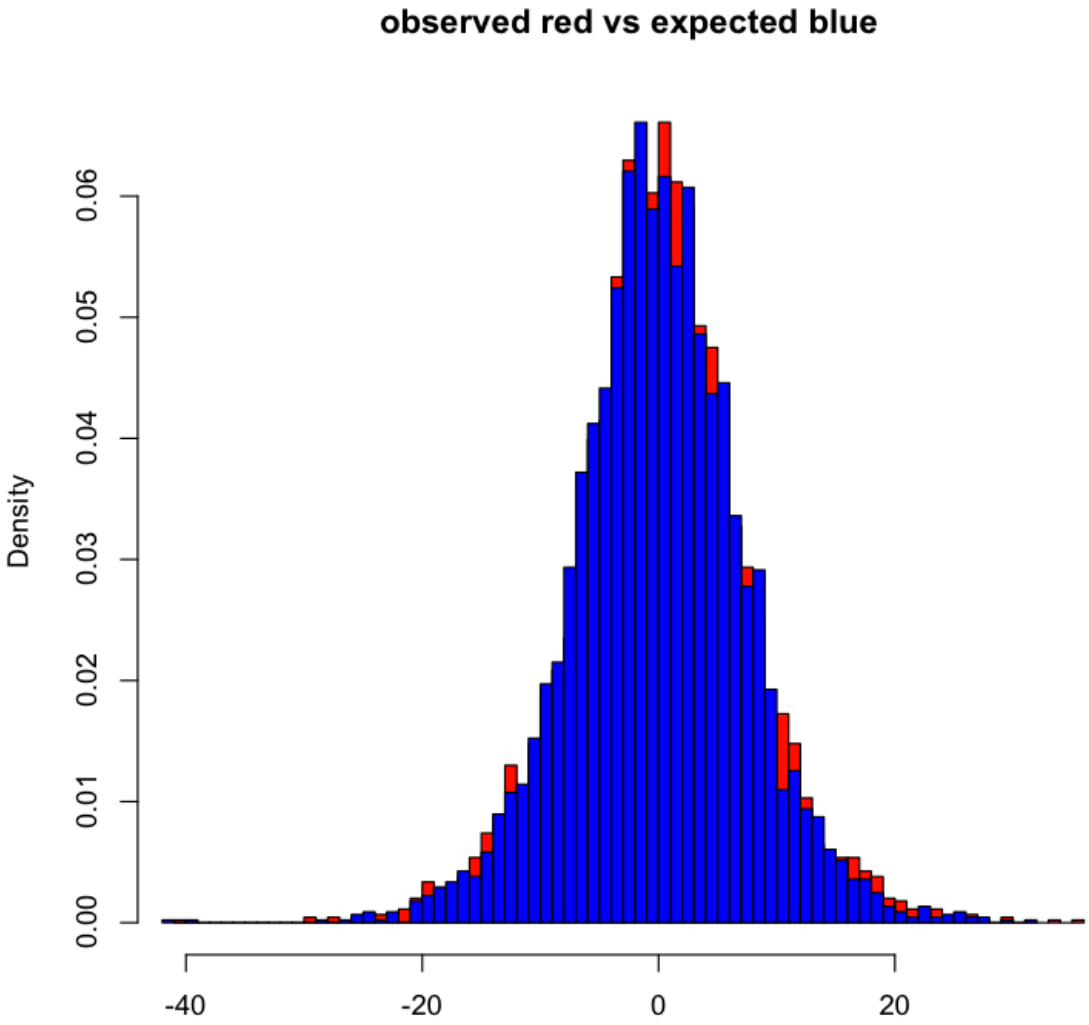


These plots are quite convincing and thus we test the goodness of fit of our theoretical logistic distribution using Kolmogorov-Smirnoff again and obtain a p-value of .91 and accept the null hypothesis with high confidence. Using R we attempt to optimize the values of the parameters we use to construct our theoretical logistic distribution by making a matrix composed of values from an upper and lower bounds for both location and scale parameters and test these to see which returns the greatest p-value. Similarly we could have constructed mainly different linear models by varying slightly the location and scale parameters for our logistic estimation and then observed which produced the greatest correlation. From our approach, we were able to generate a p-value of .96 using a location of .083 and a scale parameter of 4.025.

Now we run another goodness of fit test, the Chi-squared test, to validate further our findings that our dataset comes from a logistic distribution. This test however requires positive values so we shift our observed distribution to the right, thus fulling that requirement, by subtracting the minimum value of the distrubtion (-40.89) from every value in it. This shifting of our data however requires getting new parameters for our logistic estimated distribution which essentially just gives a shifted location parameter. We repeat the same method from above, using the model parameters returned from running fitdistr() on our shifted data to construct a theoretical distribution we test via ks.test() getting a p-value of .91 and then optimizing model parameters similarly as before to finally get a p-value of .96. With this new shifted estimated distribution with location and scale parameters of 40.97 and 4.025 respectively, we begin constructing the necessary components of the chi-squared test.

We construct quantiles of equal amounts of observations and thus having varying ranges for our observed data set, and then calculate X2 value which is just the squared value of the observed amounts per quantile minus those we would expect from a perfectly fitted distribution, ie equal amounts of observations per quantile, divided/normalized by the expected number of observations for the quantile. From this value and a degree of freedom we calculate from the total number of quantiles used, we obtain a p-value of 1 minus pchisq() equal to .4458. Similarly we could have used chisq.test(). Both methods are demonstrated in the accompanying R file.

We finally do a graphical assesment of our observed dataset against a randomally generated logistic distribution with the model parameters from our estimated distrubtion and obtain the following imposed histograms of red observed and blue expected distributions which seem very similar. A quick use of summary() and sd() on both our observed and expected distributions show they have similar moments.



red observed data with standard deviation 7.32

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-40.89000	-4.20800	0.11580	0.04535	4.45900	35.08000

blue randomly generated logistic model with standard deviation 7.19

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-37.8300	-4.1250	0.2480	0.2748	4.5780	32.9700

III. Conclusions

Based on analysis, the observed dataset comes from a logistic distribution with a location parameter of 0.083 and a scale parameter of 4.025. The Kolmogorov-Smirnoff test gives us a p-value of .96 while the Chi squared gives us a lower, but still rather strong belief, a p-value of .446, sufficient enough to at the very least to not reject the null hypothesis that the dataset is logistically distributed.