Introduction

The goal of this project is to analyze a set of data in order to estimate the function, as well as the parameters of the function, that was used to generate the dependent variable value based on the value of the independent variable. Two sets of code are developed independently using R and Matlab, respectively, in order to ensure the correctness of the results.

Methodology

1. Preprocessing

After looking into the given data, it was found that both data sets of dependent variable and independent variable have missing entries, and the missing entries were not labeled with same ID. Therefore, in order to remove the missing entries in both data set, two different approaches were adopted to implement list wise deletion. Specifically, in the R environment, we removed the missing entries by list-wise deletion. The two original CSV files were uploaded into R, and named as "datax" and "datay". Since these two files share the column "ID", hence, we merged them into one data set that consisted of 971 entries by using command "data <- merge(datax, datay, by = c('ID'))". Additionally, the software omitted the rows with missing data. Therefore, the data set only has 971 entries, and also has ID number, x-value (denoted as x) and y-value (denoted as y). In Matlab environment, we first imported the data in Matlab and wrote a code to match the independent variable data set and the dependent variable data set, using their corresponding ID, to get the clean data with three columns: ID, IV and DV.

The raw data sets had total 984 independent variables with ID number ranging from 1 to 1000, and it also had total 984 dependent variables with ID number ranging from 1 to 1000. After removing 16 entries that missed either independent variable or dependent variable, 3 of which were missing from both x and y variables, we obtained clean data with total 971 entries and ID number ranging from 1 to 1000.

2. Data Analysis

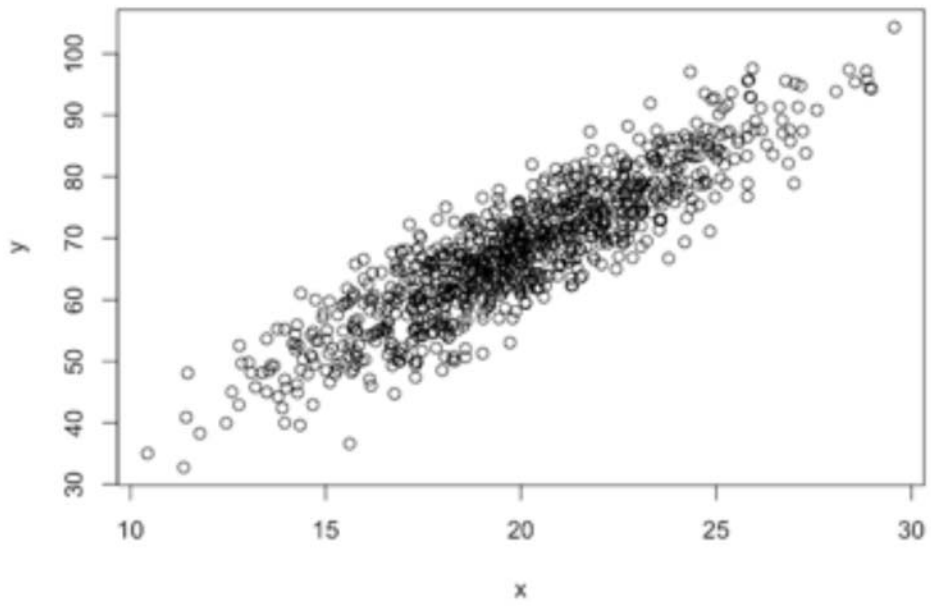
Data analysis was done independently in both R and Matlab environment.

(1). R: We used the lm command in R to create a linear model for x and y, and the summary command and anova command to find more statistic about the model. Specifically, by using the command “coefficient (model)”, we computed the values of the coefficients, the correlation between x and y by running the command “cor.test (model)”, and the ANOVA table by the command “anova (model)”. For graphing, we used the command “plot (x, y)” to get the plot graph of the model; for residuals graph, Q-Q graph, and the scale-location graph, we used the command ”plot (model)”. Furthermore, using the command “confint (model)”, we found the confidence interval. Also, in order to test whether our model is accurate, we decide to check is there any outlier in this model, we used to command ”outlier.test (model)” to check the outliers.
(2). Matlab: We adopted the linear model $Y = \beta_0+\beta_1X$ and in order to determine the coefficients we performed liner regression on the clean data. And then we obtained the fraction of variance explained, the confidence interval for slope and intercept . We also conducted the test of the null hypothesis that the slope was zero and computed the correlation coefficient between the independent and dependent variable. Additionally, we obtained the ANOVA table, Quantile-quantile plot and residuals plot.

The result we concluded using R and Matlab are the same.

Results

The fitted function of the model $Y = \beta_0+\beta_1X$ was DV = 4.815043 + 3.160205IV with 79.20% fraction of variance was explained. The 95% confidence interval for the slope was [3.0639, 3.2685] and the 95% confidence interval for the intercept was [2.7431, 6.8870]. Testing H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$ at the 0.01 significance level resulted in a test statistic of 60.751, which is much greater than the critical value of 2.58, therefore the null hypothesis of $\beta_1 = 0$ can be rejected. The analysis of variance table can be found below and the association between the independent variable and dependent variable is highly significant (p=0.000). Also, the coefficient correlation between the independent and dependent variable was 0.8897. For the outliers, the entry#193, #540 and #768 are the outliers in our model, which can be seen from the residuals & fitted graph, and scale-location graph.



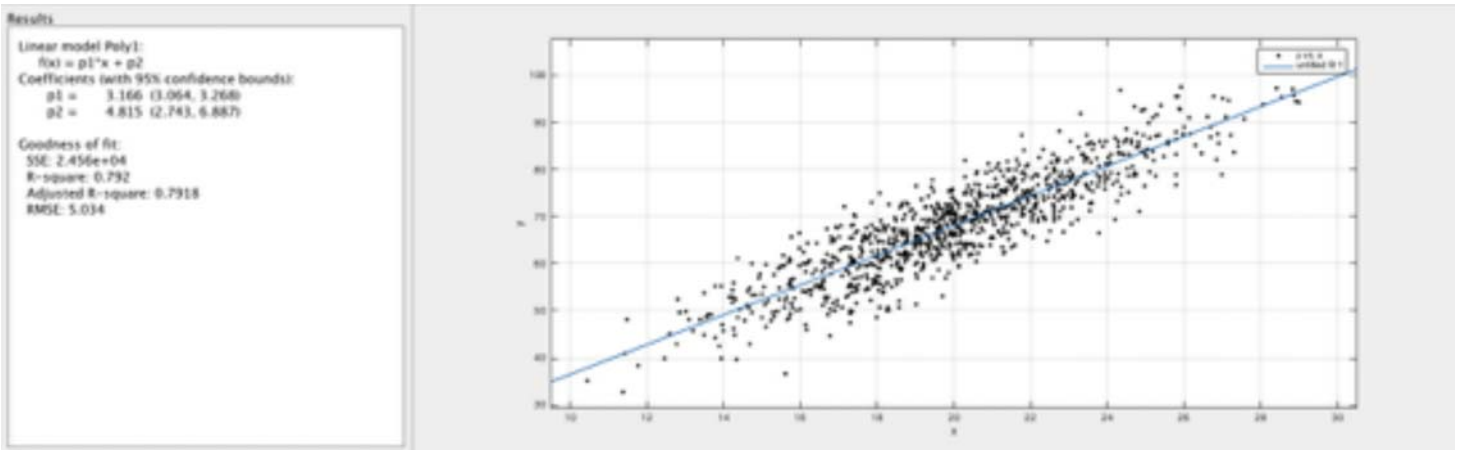
Analysis of Variance Table

(n=971)

	Sum of Square	DF	Mean of Square	F	P-Value
Residual	24559	969	25.345		
Model	93540	1	3540	3690.7	0.000
Total	118099	970			

Conclusion

The underlying function between independent and dependent variable can be explained by a linear model, as shown below



We found a statistically significant relationship between the independent and dependent variables. The independent variable explained 79.2% of the dependent variable. The significance of this model can be also seen from the plot graph and the residual graph.

Appendix

Matlab Code:

```
Mx = csvread('Group5x.csv');
My = csvread('Group5y.csv');

%Finding missing pairs and clean them from the original data
for i = 1:1000

  if ~isempty(find(My(:,1)==i)) & ~isempty(find(Mx(:,1)==i))
     allData(i,:) = [i, Mx(find(Mx(:,1)==i),2), My(find(My(:,1)==i),2)];

  end

end

cleanData(:,:) = allData(find(allData(:,1)~=0),:);

x = cleanData(:,2);
y = cleanData(:,3);
n = size(find(allData(:,1)~=0),1);

%Find the coefficients
meanx = sum(x)/n;
meany = sum(y)/n;

sxy = sum((x-meanx).*(y-meany));
sxx = sum((x-meanx).^2);

b1 = sxy/sxx;
b0 = meany-b1*meanx;

%the fitted function
yhat = b0+b1*x;


SSR = sum((yhat-meany).^2);
SST = sum((y-meany).^2);

SSE = sum((y-yhat).^2);

MSR = SSR/1;
MSE = SSE/(n-2);


%fraction of variance explained
gammaSquare = SSR/SST;
```

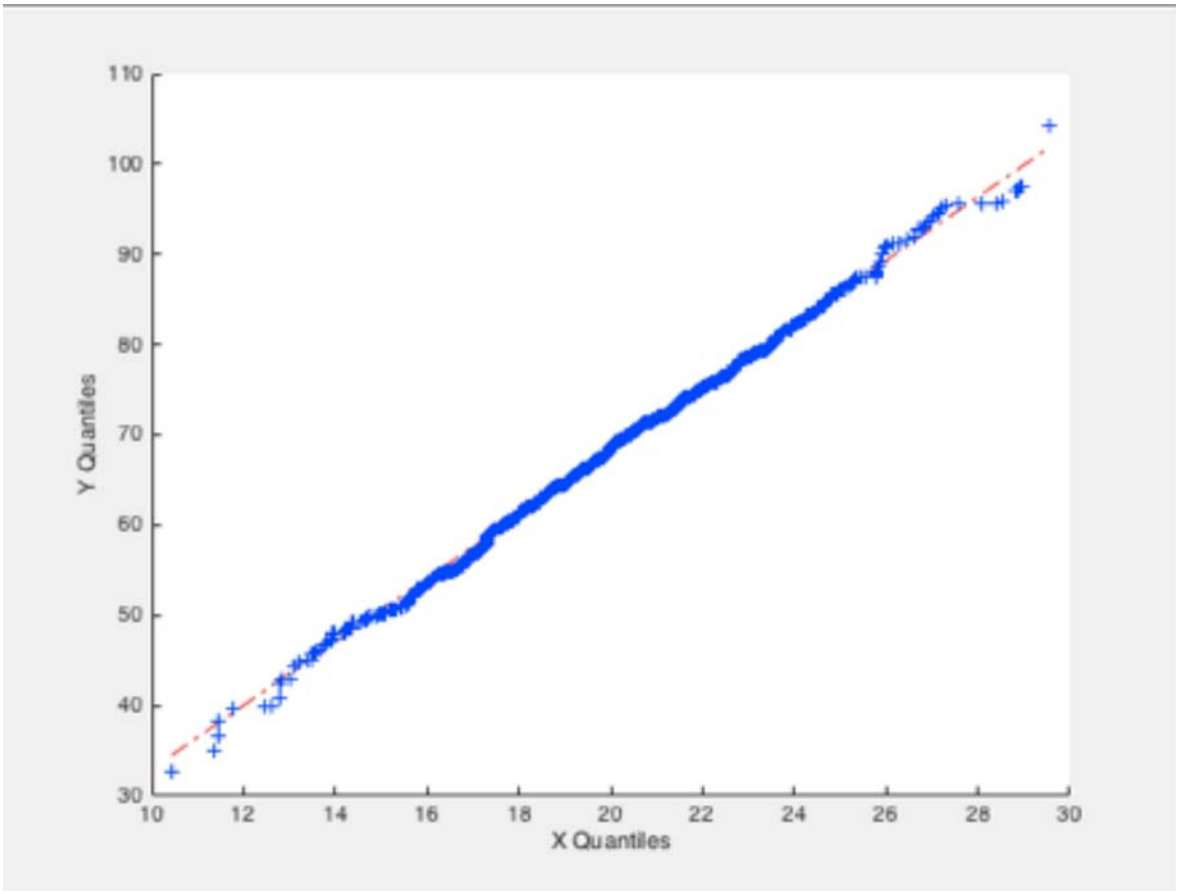
```
%confidence intervals for slope

se = sqrt(MSE);
ci = 0.95;
alpha = 1 - ci;

T_multiplier = tinv(1-alpha/2, n-2);

ci95s = T_multiplier*se/sqrt(sxx);

CI_slope = [b1 - ci95s, b1 + ci95s];

%confidence intervals for intercept

ci95i = T_multiplier*se*(sqrt((1/n)+meanx^2/sxx));

CI_intercept = [b0 - ci95i , b0 + ci95i];

%test of the null hypothesis that the slope was zero
%H0: b1 = 0
%H1: b1 ~= 0

t = (b1-0)/(se/sqrt(sxx));

if abs(t)>T_multiplier;
    fprintf('Reject H0 ');
else fprintf('Accept H0 ');
end

%test of Pearson product-moment correlation coefficient to measure the
%linear correlation between x and y
%H0: population correlation coefficient is equal to 0
%H1: population correlation coefficient is not equal to 0
%using CI = 95%
R = corrcoef(x,y);
R = R(2,1);

tR = R*sqrt(n-2)/sqrt(1-R^2);
if abs(tR)>T_multiplier;
    fprintf('Reject H0');
else fprintf('Accept H0');
end

qqplot(x,y);
title('')

%ANOVA
mdl = fitlm(x,y);
tbl = anova(mdl,'summary');
```

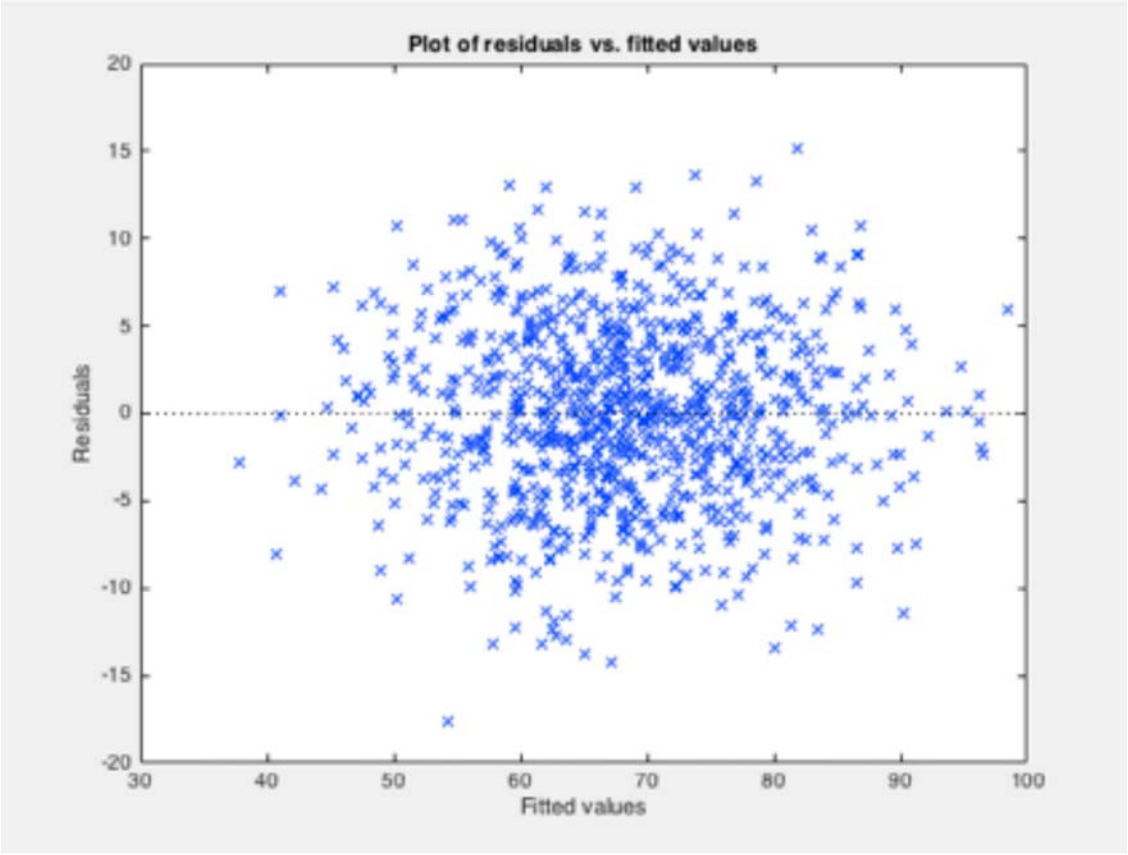
plotResiduals(mdl,'fitted');

R Code:

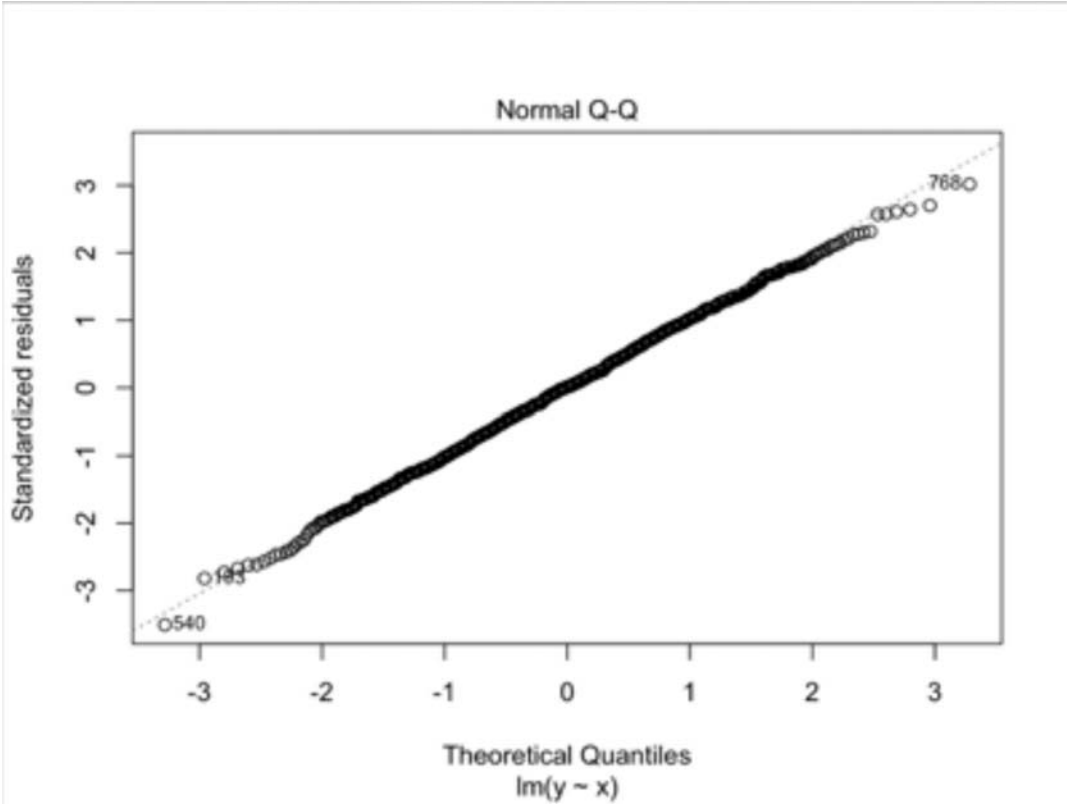
```r
data <- merge(datax, datay, by=c('ID'))
View(data)
x <- data$x
y <- data$y
plot(x,y)
model <- lm(y~x)
abline(model)
summary(model)
anova(model)
cor.test(x,y)
install.packages("car")
library(car)
coefficients(model)
confint(model)
outlier.test(model)
influence.measures(model)
plot(model)
```

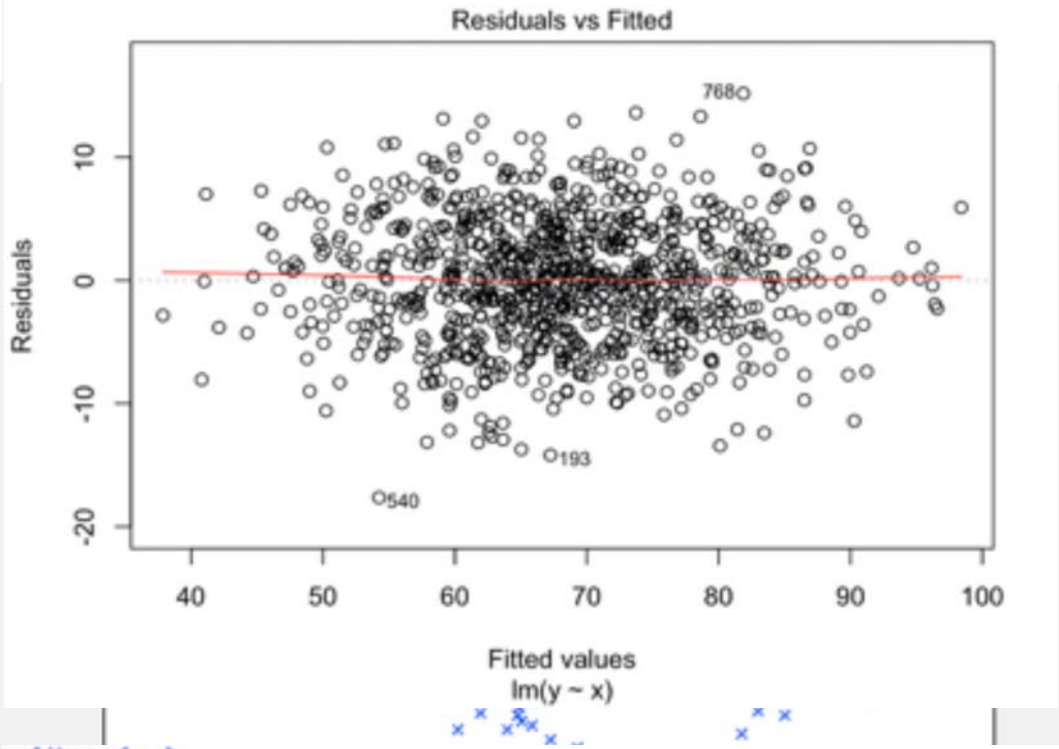
Matlab Plots:




Plot of residuals vs. fitted values
Residuals
Fitted values
20
15
10
5
0
-5
-10
-15
-20
30
40
50
60
70
80
90
100

R Plots:


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
768
183
540
3
2
1
0
-1
-2
-3




```
> library(car)
> coefficients(model)
(Intercept)           x
   4.815043    3.166205
> confint(model)
                2.5 %   97.5 %
(Intercept) 2.743124 6.886962
x           3.063929 3.268481
> outlier.test(model)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferonni p
540 -3.528292          0.0004379       0.4252
Warning message:
'outlier.test' is deprecated.
Use 'outlierTest' instead.
See help("Deprecated") and help("car-deprecated").



> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 60.751, df = 969, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8761117 0.9023596
sample estimates:
      cor
0.8899705

>
```

```
> plot(x,y)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-17.631  -3.376   0.054   3.539  15.165

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.81504    1.05580   4.561 5.75e-06 ***
x            3.16620    0.05212  60.751  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.034 on 969 degrees of freedom
Multiple R-squared:  0.792,     Adjusted R-squared:  0.7918
F-statistic:  3691 on 1 and 969 DF,  p-value: < 2.2e-16

>
```