BME 313L: INTRODUCTION TO NUMERICAL METHODS IN BIOMEDICAL ENGINEERING

LAB REPORT

Lab #8: Linear/nonlinear regressions and least-squares

Last name, First name: Kirk, Andrew

EID: alk2488

Lab Section: Friday

Problem 1: Linear Regression Problem 14.12

On average, the surface area A of human beings is related to weight W and height H. Measurements on a number of individuals of height 180 cm and different weights (kg) give values of A (m^2) in the following table:

W (kg)	70	75	77	80	82	84	87	90
A (m^2)	2.10	2.12	2.15	2.20	2.22	2.23	2.26	2.30

(1) Show that a power law $A = aW^b$ fits these data reasonably well. Evaluate the constants *a* and *b*, and predict what the surface area is for a 95-kg person.

(2) Plot log(*A)* vs. log(*w)* not only for data from table, but also for least square fit in one figure. Use red and circle symbol for data from the table.

(3) Plot *A* vs. *W* not only for data from table, but also for $A = aW^b$ in one figure. Use red and circle symbol for data from the table.

(The following is your answer)

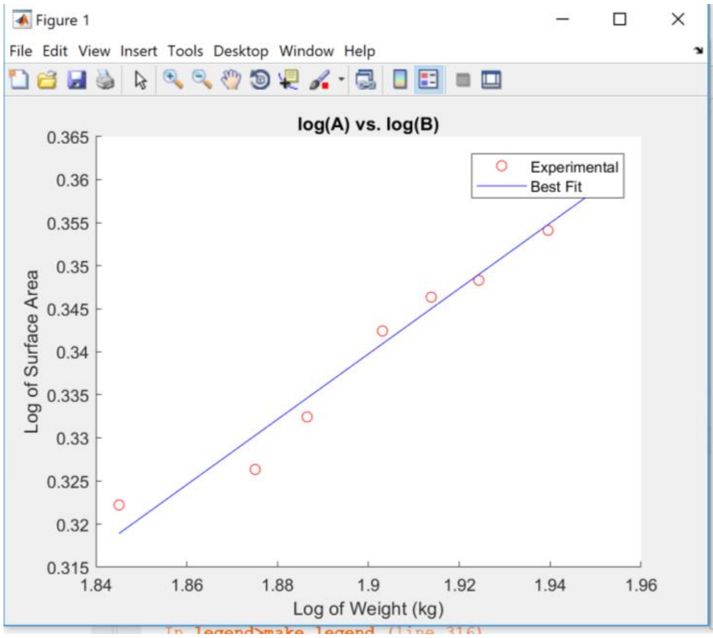
MATLAB code:

```
close all
clear all
clc

W = [70; 75; 77; 80; 82; 84; 87; 90]; % weights
A = [2.10; 2.12; 2.15; 2.20; 2.22; 2.23; 2.26; 2.30]; % Area
n = size(W, 1); % get number of data points

x = log10(W); %power linearization
y = log10(A);

sumx = sum(x); %calculate sums
sumy = sum(y);
sumxsqr = sum(x.*x);
sumxy = sum(x.*y);

a(1) = (n*sumxy-sumx*sumy)/(n*sumxsqr- sumx^2); % calculate coeffecients
a(2) = sumy/n-a(1)*sumx/n;
```

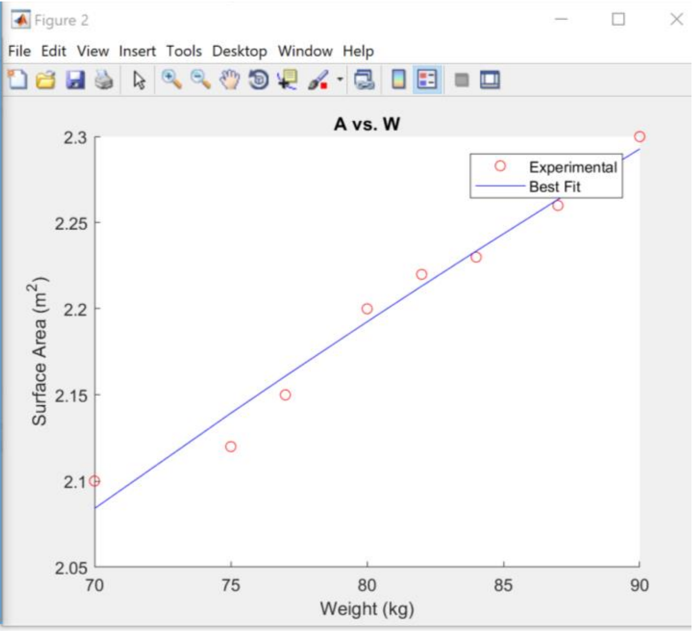
```
figure(1); clf;
scatter(x, y, 'r') % transformed data

Log = a(1)*x + a(2); % least square fit
figure(1); hold on;
plot(x, Log, 'b')
title('log(A) vs. log(B)')
xlabel('Log of Weight (kg)')
ylabel('Log of Surface Area')
legend({'Experimental','Best Fit','Location','Northwest'})

% Plotting the original data
figure(2); clf;
scatter(W, A, 'r')
% Plotting the Power Fit
PowerFit = (10^a(2))*W.^a(1);
figure(2); hold on;
plot(W, PowerFit, 'b')
title('A vs. W');
xlabel('Weight (kg)')
ylabel('Surface Area (m^2)')
legend({'Experimental','Best Fit','Location','Northwest'})
Surface_Area_for_95kg = (10^a(2))*95^a(1)
```

Results:

Surface_Area_for_95kg = 2.3404



```
In legend>make_legend (line 316)
```



Discussion:

The data is linearized to fit the power function. The slope is equal to a1 and the y intercept is a0. To do this the log of both sides of the equation is taken giving Log(y)=b*log(W) + log(a). From here a linear best fit line is calculated using the sums of x, x*y, and x^2. Once the b and a values are calculated the true best fit line for the experimental data is made. As seen from the graph, this method provides the best linear best fit line.

Problem 2: General Linear Least Squares

Environmental scientists and engineers dealing with the impacts of acid rain must determine the value of the ion product of water K_w as a function of temperature. Scientists have suggested the following equation to model this relationship:

$$-\log_{10} K_w = \frac{a}{T_a} + b\log_{10} T_a + cT_a + d$$

where T_a = absolute temperature (in K), and a, b, c, and d are parameters. The following data is observed:

K_w	$1.164\text{x}10^{-15}$	$2.950\text{x}10^{-15}$	$6.846\text{x}10^{-15}$	$1.467\text{x}10^{-14}$	$2.929\text{x}10^{-14}$
T_a	1	10	20	30	40

Answer the following questions: (100 word minimum for each question, 50 word minimum for discussing what you learned, what was reinforced)

(1) Is this equation equivalent to a general linear least squares model? If yes, what are the basis functions of this model?

(2) Implement an approach in MATLAB that uses the normal equations to estimate the parameters a, b, c and d. The usage of any pre-built regression functions is not permitted.

(3) Generate a plot of predicted ion products of water $\overline{K_w}$ versus temperature T_a (for T_a = 1, 10, 20, 30 and 40)

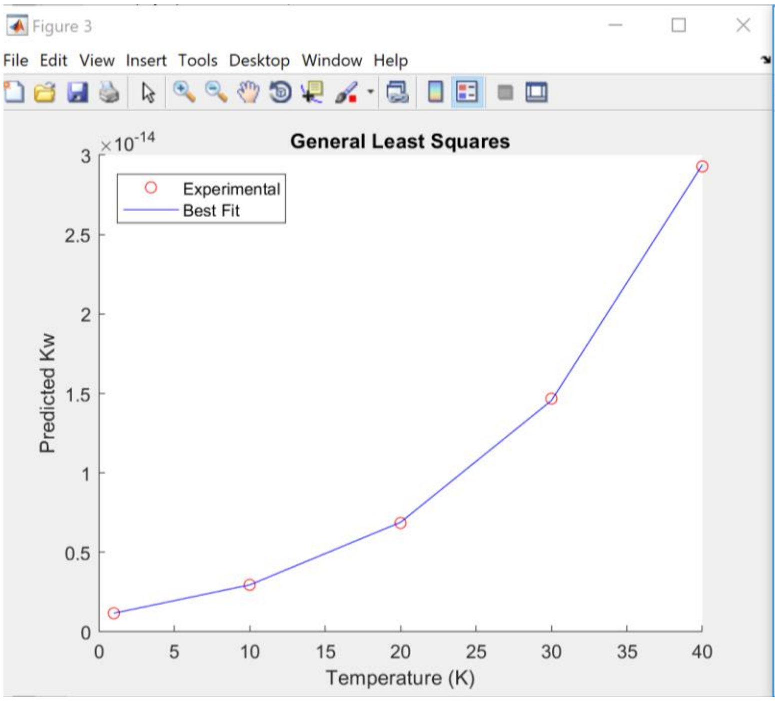
(4) Generate a plot of squared residuals $(K_w - \overline{K_w})^2$ versus temperature T_a (for T_a = 1, 10, 20, 30 and 40)

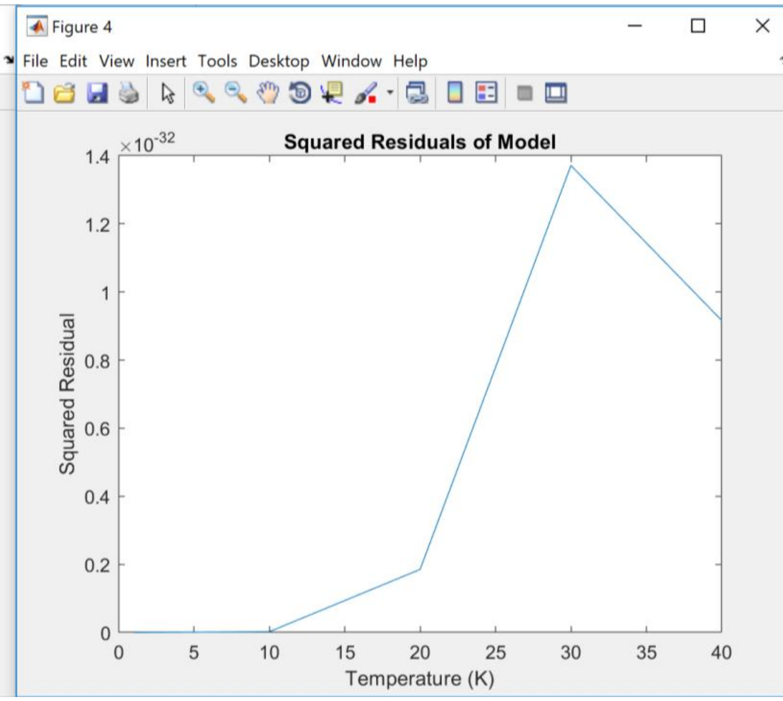
(The following is your answer)

MATLAB code:

```
close all
clear all
clc

Kw = [1.164e-15; 2.950e-15; 6.846e-15; 1.467e-14; 2.929e-14]; %create data matrices
Ta = [1; 10; 20; 30; 40];

Z(:,1) = ones(size(Kw));    %make Z matrix based on experimental datz
Z(:,2) = Ta;
Z(:,3) = log10(Ta);
Z(:,4) = Ta.^-1;

coeffecients = Z\(-1*log10(Kw)); %calculating Kfit
KFit = Z*coeffecients;
KFit = -1*KFit;
KFit = 10.^KFit;

figure(3); clf;
scatter(Ta,Kw,'r'); hold on;
plot(Ta, KFit,'b');
title('General Least Squares')
xlabel('Temperature (K)')
ylabel('Predicted Kw')
legend({'Experimental','Best Fit'},'Location','Northwest')

Sr = (Kw - KFit).^2; % find square residuals to plot
figure(4); clf;
plot(Ta, Sr);
title('Squared Residuals of Model')
xlabel('Temperature (K)')
ylabel('Squared Residual')
```

Results:





Discussion:

This is a general least square function because it does not follow the form of any of the other methods for determining regression. The basis functions for this equation are log10(Ta), Ta, 1/Ta, and

the constant basis function. All of these are put into the Z matrix and this is used to calculate the y values. This reinforced how to use the general least squares method for finding best fit when the relationship between points does not follow any obvious relationship that can be described with linear or polynomial regression. The residuals using this method are almost non-existent which means that the data fits very well to the best fit line.

Problem 3: General Least Squares

Mathematician Cristian Tomasetti and cancer geneticist Bert Vogelstein of Johns Hopkins University proposed a groundbreaking idea about tumorigenesis: two-thirds of adult cancers are caused by random mutation in the tissue cells during the ordinary process of stem cell division. In the other third, our genetic inheritance and lifestyles are the main factors. The research is reported in *Science* magazine in January 2015 (you can find this article in the Canvas). The report demonstrated the relationship between the number of stem cell division in the lifetime of a give tissue and the lifetime risk of cancer in that tissue, as shown in Fig. 1. This controversial paper has drawn many criticisms since its publication (also can be found in the Canvas). You can easily find more criticisms on line by typing the key words “cancer two-thirds bad luck”.

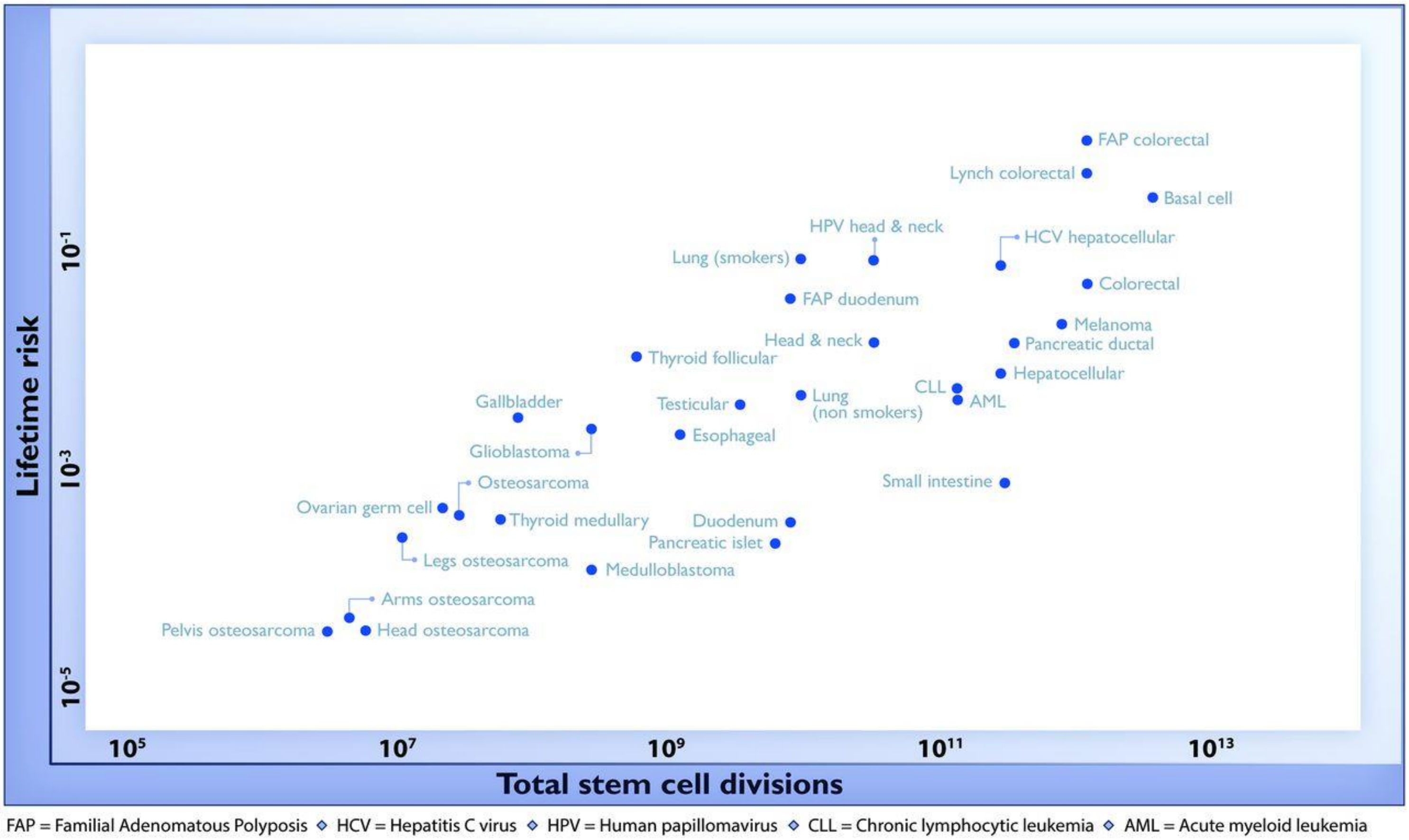


Fig. 1 The relationship between total stem cell divisions and lifetime risk of cancer.

In this problem, we are going to find out the correlation between stem cell divisions and lifetime risk of cancer by using linear least-squares regression.

(1) Please use the table to generate a plot like Fig. 1. (The x and y axes should be in log scale)

(2) Calculate the coefficient of determination (R^2) with linear least-squares regression. (Hint: it’s a linear relationship between total stem cell divisions and lifetime risk, so you have to conduct linearization of nonlinear

relationships first, and then fit the data points with linear regression.)

Cancer type	Lifetime cancer risk	Cumulative number of divisions of all stem cells per lifetime (lscd)
Acute myeloid leukemia	0.0041	1.299×10^{11}
Basal cell carcinoma	0.3	3.550×10^{12}
Chronic lymphocytic leukemia	0.0052	1.299×10^{11}
Colorectal adenocarcinoma	0.048	1.168×10^{12}
Colorectal adenocarcinoma with FAP	1	1.168×10^{12}
Colorectal adenocarcinoma with Lynch syndrome	0.5	1.168×10^{12}
Duodenum adenocarcinoma	0.0003	7.796×10^{9}
Duodenum adenocarcinoma with FAP	0.035	7.796×10^{9}
Esophageal squamous cell carcinoma	0.00194	1.203×10^{9}
Gallbladder non papillary adenocarcinoma	0.0028	7.840×10^{7}
Glioblastoma	0.00219	2.700×10^{8}
Head & neck squamous cell carcinoma	0.0138	3.186×10^{10}
Head & neck squamous cell carcinoma with HPV-16	0.07935	3.186×10^{10}
Hepatocellular carcinoma	0.0071	2.709×10^{11}
Hepatocellular carcinoma with HCV	0.071	2.709×10^{11}
Lung adenocarcinoma (nonsmokers)	0.0045	9.272×10^{9}
Lung adenocarcinoma (smokers)	0.081	9.272×10^{9}
Medulloblastoma	0.00011	2.720×10^{8}
Melanoma	0.0203	7.638×10^{11}
Osteosarcoma	0.00035	2.926×10^{7}
Osteosarcoma of the arms	0.00004	4.550×10^{6}
Osteosarcoma of the head	0.000030 2	6.020×10^{6}
Osteosarcoma of the legs	0.00022	1.113×10^{7}
Osteosarcoma of the pelvis	0.00003	3.150×10^{6}
Ovarian germ cell	0.00041	2.200×10^{7}
Pancreatic ductal adenocarcinoma	0.01359	3.428×10^{11}
Pancreatic endocrine (islet cell) carcinoma	0.00019	6.068×10^{9}
Small intestine adenocarcinoma	0.0007	2.922×10^{11}
Testicular germ cell cancer	0.0037	3.348×10^{9}
Thyroid papillary/follicular carcinoma	0.01026	5.850×10^{8}
Thyroid medullary carcinoma	0.00032	5.850×10^{7}

(The following is your answer)

MATLAB code:

```
close all
clear all
clc

divisions= [1.299e11; 3.550e12; 1.299e11; 1.168e12; 1.168e12; 1.168e12; 7.796e9; 7.796e9; 1.203e9; 7.840e7;
2.700e8; 3.186e10; 3.186e10; 2.709e11; 2.709e11; 9.272e9; 9.272e9; 2.720e8; 7.638e11; 2.926e7; 4.550e6;
6.020e6; 1.113e7; 3.150e6; 2.200e7; 3.428e11; 6.068e9; 2.922e11; 3.348e9; 5.850e8; 5.850e7];
risk = [0.0041; 0.3; 0.0052; 0.048; 1; 0.5; 0.0003; 0.035; 0.00194; 0.0028; 0.00219; 0.0138; 0.07935; 0.0071;
0.071; 0.0045; 0.081; 0.00011; 0.0203; 0.00035; 0.00004; 0.0000302; 0.00022; 0.00003; 0.00041; 0.01359;
0.00019; 0.0007; 0.0037; 0.01026; 0.00032];

n = size(divisions,1); % get number of data points

% Plotting data in log scale
figure(5); clf;
plot(log10(divisions),log10(risk), 'b.', 'MarkerSize', 15)
title('Cancer Risks vs. Stem Cell Divisions')
xlabel('Total stem cell divisions (log_1_0)')
ylabel('Lifetime risk (log_1_0)')

x = log10(divisions); %power linearization
y = log10(risk);

sumx = sum(x); %get sum values to calculate coeffecients and degree of fit
sumy = sum(y);
sumxsqr = sum(x.*x);
sumxy = sum(x.*y);
sumysqr = sum(y.*y);

a(1) = (n*sumxy-sumx*sumy)/(n*sumxsqr-sumx^2);    % calculating coefficients
a(2) = sumy/n-a(1)*sumx/n;

bestfit = a(2) + a(1)*log10(divisions);% Plot best fit
figure(5); hold on;
plot(log10(divisions),bestfit,'r','LineWidth',1); hold on;
legend({'Observed Data','Best Fit Line'},'Location','Northwest');

rsquared = ((n*sumxy-sumx*sumy)/sqrt(n*sumxsqr-sumx^2)/sqrt(n*sumysqr-sumy^2))^2 %coeffecient of
determination
```

Results:

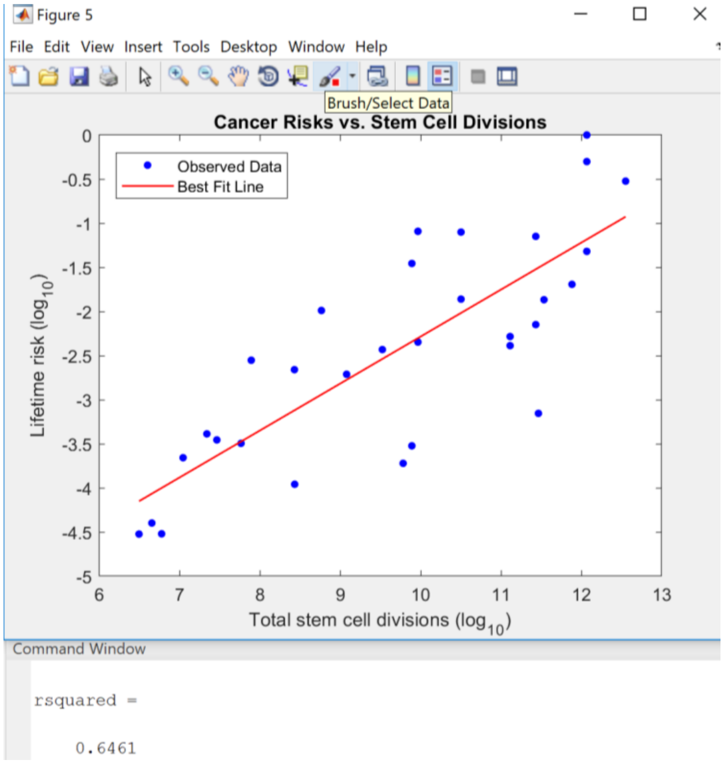


Figure 5
File Edit View Insert Tools Desktop Window Help
Brush/Select Data
Cancer Risks vs. Stem Cell Divisions
Observed Data
Best Fit Line
Lifetime risk ($\log_{10}$)
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-5
6
7
8
9
10
11
12
13
Total stem cell divisions ($\log_{10}$)
Command Window

```
rsquared =

    0.6461
```

Discussion:

<u>Things to discuss:</u>

1) The relationship between total stem cell divisions and lifetime risk of cancer is nonlinear. Please write down the procedure of linearization.

(2)The report arouses a heated discussion among the communities of cancer research. Do you agree with authors' viewpoint? Why or why not?

This is a power relationship because you can not have a negative risk for cancer and zero divisions represents zero risk for cancer. So to solve for the best fit you take the log of both sides which results in a general form of log(y) = log(a) + b*log(x). This is in the form of line where y=mx+b.

I do not agree with the author. Even thought the coefficient of determination is somewhat significant at 0.6461, when you look at the graph the points are all over and the line does not fit all of the points very well and there are a lot of outliers