STATISTICSII PROJECT REPORT

Submitted by:

Chandrabali Karmakar

SGGW *Nr Albumu : 177639*

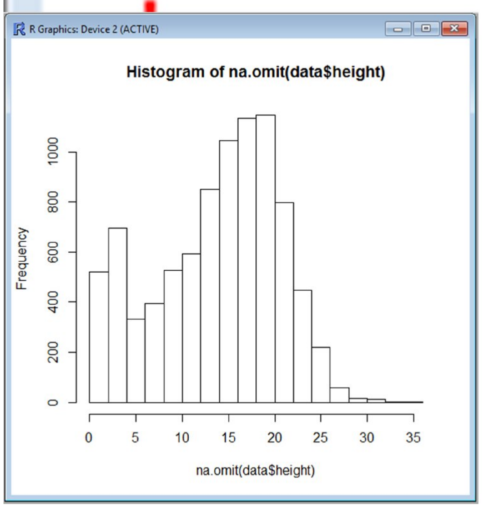
HNEE Matriculation No: 14209583

Contents

1) plot a histogram of *height* variable without *NA* records

Hist(na.omit(height))

```
> hist(na.omit(data$height))
```



2) calculate average height of all stands (not counting *NAs*) as well as average heights of stands by species

```
> summary(na.omit(height))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   10.00   15.00   14.27   19.00   35.00 
```

```
> vdata=cbind(height,species_cd)
> dfdata=data.frame(vdata)
> ref_dfdata=na.omit(dfdata)
> names(ref_dfdata)
[1] "height"     "species_cd"
> h=ref_dfdata$height
> s=ref_dfdata$species_cd
> result=tapply(h,s,mean)
> result
        1         2         3         4         5         6         7         8         9        10
12.296875 15.661017  6.942857 14.710037 19.142857  2.750000 12.435897  3.133333 26.333333 18.000000
       11        12        13        14        15        16        17        18        19        20
11.971429 16.000000 20.500000 11.702703 17.425220 13.000000 16.000000 14.161620 27.600000 21.000000
```

3) check if the average heights of *BRZ* (brzoza = birch) and *SO* (sosna = pine) species are the same

levels(species_cd)

1. "¦W"
2. "AK"
3. "BK"
4. "BRZ"
5. "DB"
6. "DB.B"
7. "DB.C"
8. "DB.S"
9. "DG"
10. "GB"
11. "JS"
12. "JW"
13. "LP"
14. "MD"
15. "OL"
16. "OL.S"
17. "OS"
18. "SO"
19. "TP"
20. "WZ"

So “BRZ” is the 4th in the list which shows mean as 14.710037 and “SO” is the 18th which gives 14.161620 as the average.

```
        1         2         3         4         5         6         7         8         9
.2.296875 15.661017  6.942857 14.710037 19.142857  2.750000 12.435897  3.133333 26.333333
       17        18        19        20
.6.000000 14.161620 27.600000 21.000000
```

4) check if the distribution of height for *SO* species follows the normal distribution

```
>  pine_heights = split(h,s)$'18'
> ks.test(pine_heights,"pnorm",mean(pine_heights),sd(pine_heights))

        One-sample Kolmogorov-Smirnov test

data:  pine_heights
D = 0.10243, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Is not a normal distribution.

[precise mean and standard variation are used so that the test is not done on a standard gaussian distribution.]

5) using two-way ANOVA - check if there is a signifficant difference in stand heights between species (species_cd) and age classes (0-20, 21-40, 41-60, 61-80, 81-100, 101-120, above 120 years, based on species_age variable). Use data only for *SO*, *BRZ* and *DB* (dąb = oak) species.

Using R

Step1: Getting the right data without “NA” values

```
> data=read.csv(file.choose())
> names(data)
 [1] "adress_forest"      "stand_struct_cd"    "sub_area"           "site_type_cd"
 [5] "storey_cd"          "density_cd"         "standdensity_index" "species_cd"
 [9] "species_rank_order" "part_cd"            "species_age"        "bhd"
[13] "height"             "basal_area"         "volume"             "stem_number"
> df = data.frame(data$species_cd,data$species_age,data$height)
```

```
> df_subset=subset(df,df$data.species_cd== "SO" | df$data.species_cd =="DB" |df$data.species_cd =="BRZ" )
> df_noNA = na.omit(df_subset)
> names(df_noNA)
[1] "data.species_cd"  "data.species_age" "data.height"
> 
```

Note that we had unequal sample sizes, so I put a little more effort to get sample with equal size for each species and age class.But the species “BRZ” does not have the “above 120” age class. However , the data satisfies the assumption of homogeneity of variance through Levene’s test (shown in SPSS result part).

Step 2: Creating categorical variable age_classes from the continuous variable “Species_age”

```
> df_noNA$age_classes<-ifelse(df_noNA$data.species_age<=20,rr2<-"0-20",
+    ifelse(df_noNA$data.species_age>20 & df_noNA$data.species_age <=40,rr2<-"21-40",
+          ifelse(df_noNA$data.species_age>40 & df_noNA$data.species_age <=60,rr2<-"41-60",
+                ifelse(df_noNA$data.species_age > 60 & df_noNA$data.species_age <= 80,rr2<- "61-80",
+                      ifelse(df_noNA$data.species_age > 80 & df_noNA$data.species_age <= 100,rr2<- "81-100",
+ ifelse(df_noNA$data.species_age > 100 & df_noNA$data.species_age <= 120,rr2<- "100-120",
+ ifelse(df_noNA$data.species_age > 120 ,rr2<- "Above 120",
+                             rr2<-"Not known")))))))
>
```

Step3:Performing the test.

Null hypothesis:

H0:Age has no significant effect on stand height
H0: Species type has no significant effect on stand height
H0: Age and species type interaction have no significant effect in stand height

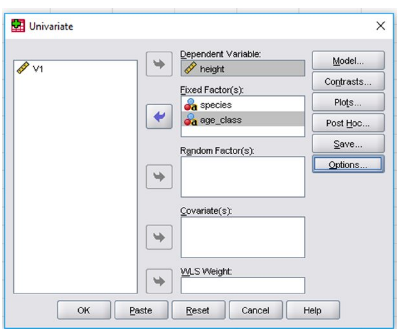
```
> names(df_noNA)
[1] "data.species_cd"  "data.species_age" "data.height"      "age_classes"
> mydata = data.frame(df_noNA$data.species_cd,df_noNA$age_classes,df_noNA$data.height)
> colnames(mydata)
[1] "df_noNA.data.species_cd" "df_noNA.age_classes"     "df_noNA.data.height"
> colnames(mydata) =c("species","age_class","height")
```

```
> names(mydata)
[1] "species"   "age_class" "height"
> summary(aov(mydata$height ~mydata$age_class + mydata$species +mydata$age_class * mydata$species))
                                 Df Sum Sq Mean Sq F value  Pr(>F)    
mydata$age_class                  5   3867   773.3 104.105 < 2e-16 ***
mydata$species                    2     81    40.7   5.483 0.00627 ** 
mydata$age_class:mydata$species   9    110    12.2   1.645 0.12079    
Residuals                        66    490     7.4                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

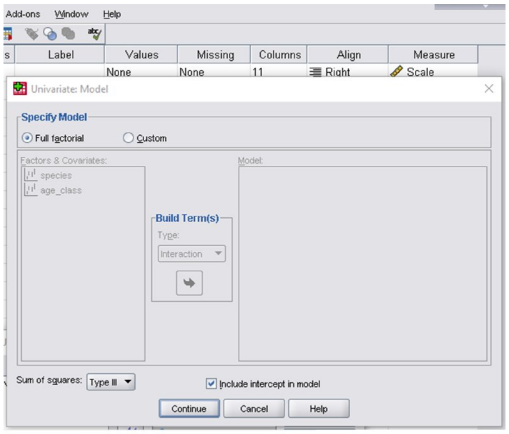
Discussion : Both age and species type have got impact on stand height. But the interaction between age and species have got no significant effect on stand height.
At the first age class, “BRZ” and “DB” have greater height than “SO”. At the second age class(21-40) “BRZ” displays the greatest height and “SO” the least.

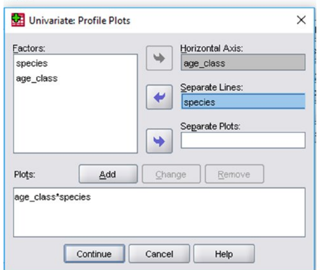
Using SPSS:.

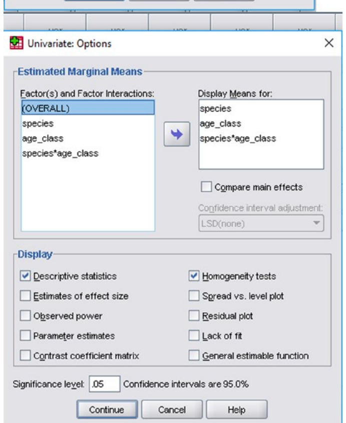
Step1:
Stating the dependent and factor variables:



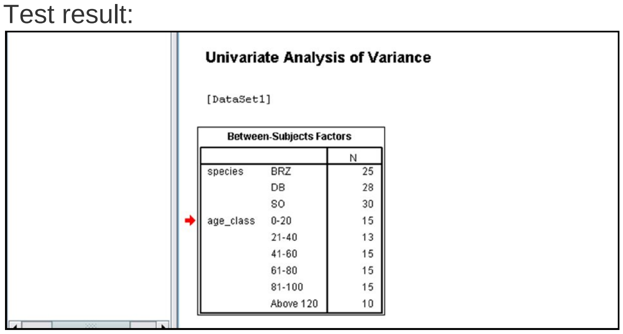
Step2:Specifying Model ,plots and other options







Test result:

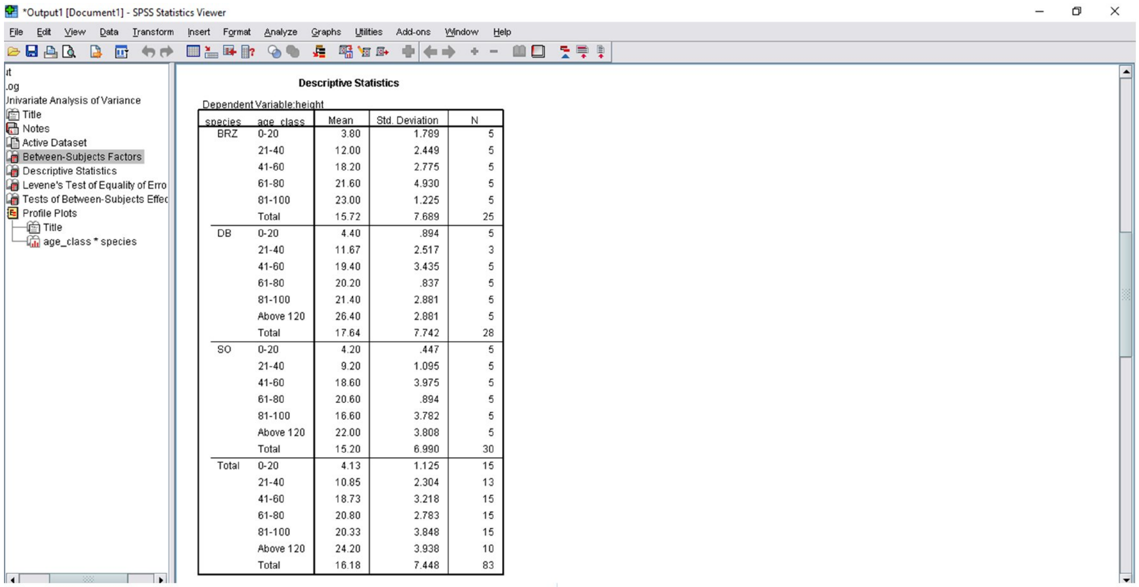


Univariate Analysis of Variance

[DataSet1]

Between-Subjects Factors

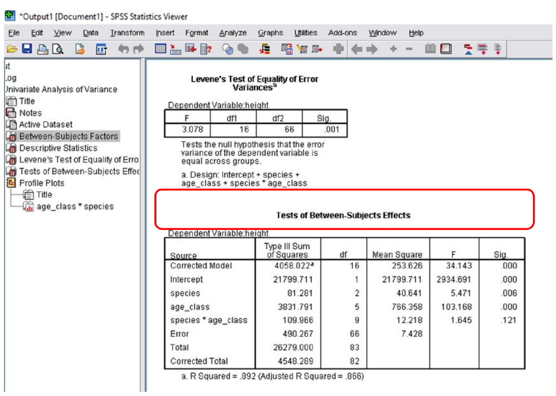
		N
species	BRZ	25
	DB	28
	SO	30
age_class	0-20	15
	21-40	13
	41-60	15
	61-80	15
	81-100	15
	Above 120	10



Descriptive Statistics

Dependent Variable:height

species	age_class	Mean	Std. Deviation	N
BRZ	0-20	3.80	1.789	5
	21-40	12.00	2.449	5
	41-60	18.20	2.775	5
	61-80	21.60	4.930	5
	81-100	23.00	1.225	5
	Total	15.72	7.689	25
DB	0-20	4.40	.894	5
	21-40	11.67	2.517	3
	41-60	19.40	3.435	5
	61-80	20.20	.837	5
	81-100	21.40	2.881	5
	Above 120	26.40	2.881	5
	Total	17.64	7.742	28
SO	0-20	4.20	.447	5
	21-40	9.20	1.095	5
	41-60	18.60	3.975	5
	61-80	20.60	.894	5
	81-100	16.60	3.782	5
	Above 120	22.00	3.808	5
	Total	15.20	6.990	30
Total	0-20	4.13	1.125	15
	21-40	10.85	2.304	13
	41-60	18.73	3.218	15
	61-80	20.80	2.783	15
	81-100	20.33	3.848	15
	Above 120	24.20	3.938	10
	Total	16.18	7.448	83



Levene's Test of Equality of Error Variances[a]

Dependent Variable:height

F	df1	df2	Sig.
3.078	16	66	.001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + species + age_class + species * age_class

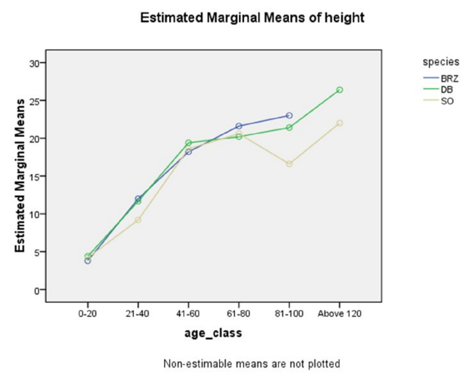
Tests of Between-Subjects Effects

Dependent Variable:height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4058.022[a]	16	253.626	34.143	.000
Intercept	21799.711	1	21799.711	2934.691	.000
species	81.281	2	40.641	5.471	.006
age_class	3831.791	5	766.358	103.168	.000
species * age_class	109.966	9	12.218	1.645	.121
Error	490.267	66	7.428		
Total	26279.000	83			
Corrected Total	4548.289	82			

a. R Squared = .892 (Adjusted R Squared = .866)

Profile Plots



Discussion : Both SPSS and R displays similar results that age and species have significant effect on stand height but not their interaction.

6) using data for *SO* species (without *NA* records) find two variables with a non-linear relationship (e.g. age <> height). Build a nonlinear model for this relationship (choose the best out of at least 3 candidates from example from "CurveCatalog.pdf" publication of from other source). Perform a detailed regression analysis as well as justify the choice of the final model. <u>Please do not limit yourself</u> to the automatic procedure in SPSS, but perform a real process of choosing the proper candidate functions.

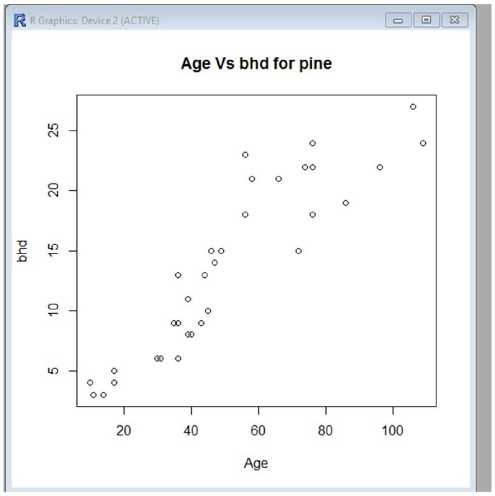
Using R

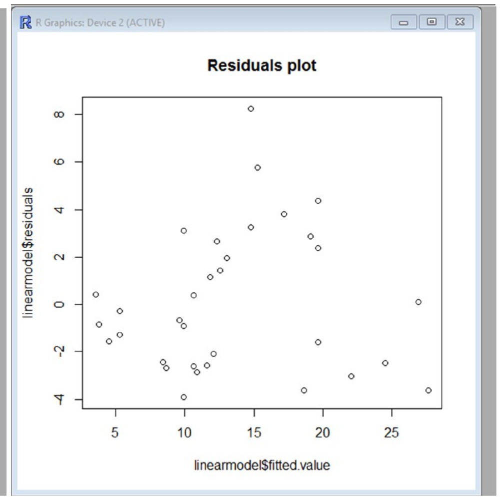
```
> data=read.csv(file.choose())
> names(data)
 [1] "adress_forest"      "stand_struct_cd"    "sub_area"
 [4] "site_type_cd"       "storey_cd"          "density_cd"
 [7] "standdensity_index" "species_cd"         "species_rank_order"
[10] "part_cd"            "species_age"        "bhd"
[13] "height"             "basal_area"         "volume"
[16] "stem_number"
> pine_data=subset(data,species_cd="SO")
> names(pine_data)
 [1] "adress_forest"      "stand_struct_cd"    "sub_area"
 [4] "site_type_cd"       "storey_cd"          "density_cd"
 [7] "standdensity_index" "species_cd"         "species_rank_order"
[10] "part_cd"            "species_age"        "bhd"
[13] "height"             "basal_area"         "volume"
[16] "stem_number"
```

```
> raw_data=cbind(pine_data$species_age,pine_data$bhd)
> df=data.frame(raw_data)
> df_noNA = na.omit(df)
> mydata=df_nona
Error: object 'df_nona' not found
> raw_data=cbind(pine_data$species_age,pine_data$bhd)
> df=data.frame(raw_data)
> df_noNA = na.omit(df)
> mydata=df_noNA
```

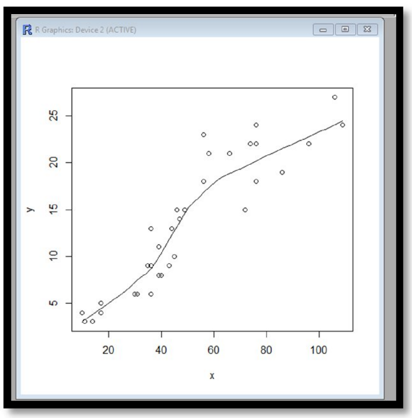
```
> number=(1:200)
> mysample=sample(number,40,replace=T)
> x=X1[mysample]
Error: object 'X1' not found
> x=mydata$X1[mysample]
> y=mydata$X2[mysample]
```

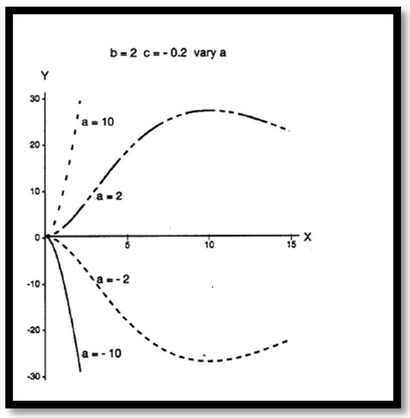
```
> plot(y~x,main="Age Vs bhd for pine" ,xlab="Age",ylab="bhd")
> linearmodel = lm(y~x)
> plot(linearmodel$residuals~linearmodel$fitted.value , main="Residuals plot")
>
```





The residuals are not evenly dispersed which says that I need a nonlinear model for this relationship.





The scatter plot seems to be an exponential function graph . But to ensure the correct model , I try a power function and a polynomial function and finally choose one among the three.

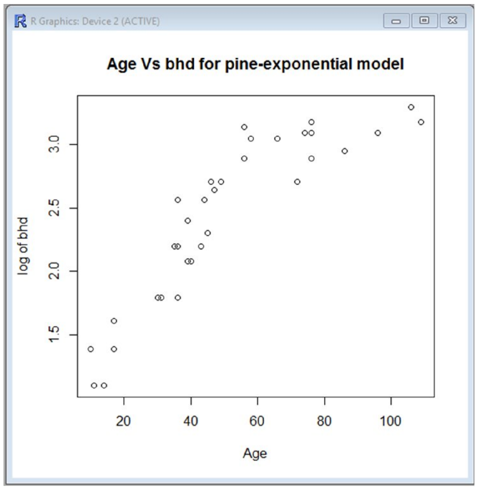
I. the exponential model:

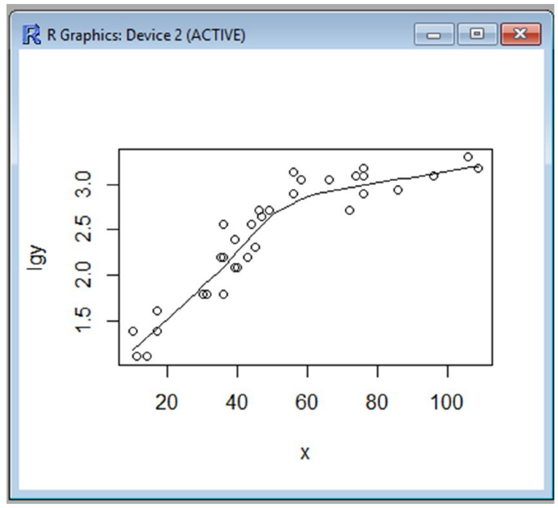
Y=ae^{bx}

A log transform is needed.

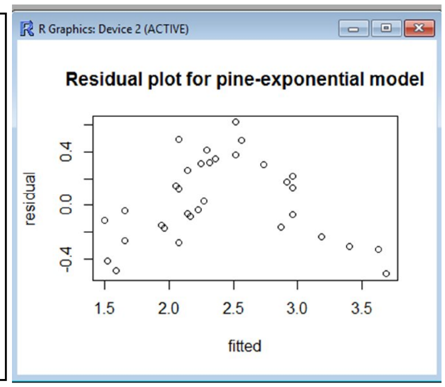
Linearized model and parameters: ln(Y) = b0 + b1X
a = eb0 b = b1

Description: The parameter **a** is the Y-intercept; the parameter **b** is the shape parameter of the curve.






```
Call:
lm(formula = lgy ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51307 -0.24399 -0.05115  0.27050  0.61923

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.274907   0.105315   12.11 1.31e-14 ***
x           0.022167   0.001853   11.96 1.86e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3254 on 38 degrees of freedom
Multiple R-squared:  0.7902,    Adjusted R-squared:  0.7847
F-statistic: 143.2 on 1 and 38 DF,  p-value: 1.864e-14
```

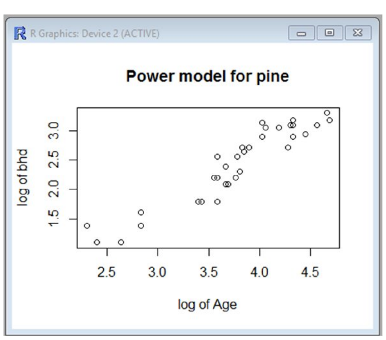


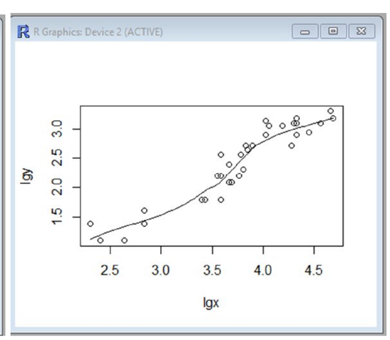
The p –values for intercept and slope are low, indicating that both of them are significantly different from zero.

II. The power model

Functional form: $Y = aX^b$

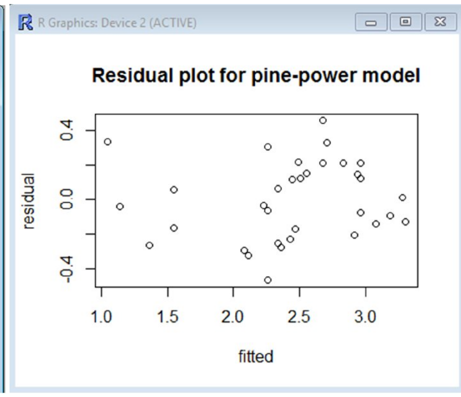
Linearized model and parameters: $\ln(Y) = b_0 + b_1 \ln(X)$
$a = eb_0 \quad b = b_1$





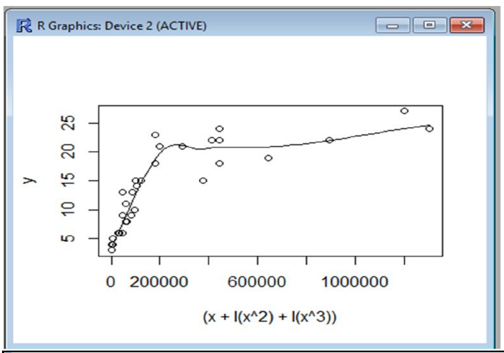
R Console

```
F-statistic: 143.2 on 1 and 38 DF,  p-value: 1.864e-14

> lgx= log(x)
> lmpow = lm(lgy~lgx)
> with(mydata,scatter.smooth(lgy~lgx))
> summary(mlpow)
Error in summary(mlpow) : object 'mlpow' not found
> summary(lmpow)

Call:
lm(formula = lgy ~ lgx)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46796 -0.17827 -0.03859  0.17006  0.45747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.13304    0.20228  -5.601    2e-06 ***
lgx          0.94677    0.05368  17.638   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2344 on 38 degrees of freedom
Multiple R-squared:  0.8912,    Adjusted R-squared:  0.8883
F-statistic: 311.1 on 1 and 38 DF,  p-value: < 2.2e-16
```

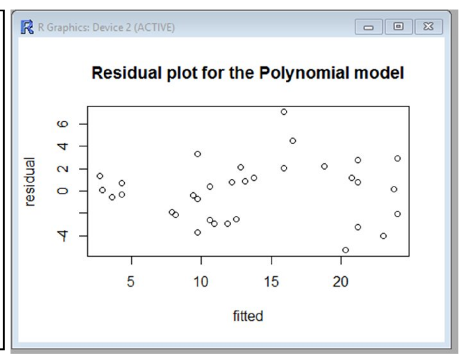


III The polynomial model

The graph may have a cubic term.
So I use a polynomial of degree 3.




```
> summary(lmpoly)

Call:
lm(formula = y ~ x + I(x^2) + I(x^3))

Residuals:
    Min      1Q  Median      3Q     Max
-5.3092 -2.2517  0.1427  1.3129  7.0730

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.676e-01  2.459e+00   0.394    0.696
x            1.334e-01  1.792e-01   0.745    0.461
I(x^2)       4.172e-03  3.633e-03   1.148    0.258
I(x^3)      -3.186e-05  2.092e-05  -1.523    0.136

Residual standard error: 2.977 on 36 degrees of freedom
Multiple R-squared:  0.8558,    Adjusted R-squared:  0.8438
F-statistic:  71.2 on 3 and 36 DF,  p-value: 3.311e-15
```



This model fails as the p-values are high meaning that the estimates are not significantly different from zero.

Dignostics:

Model comparison

Model	Fitted plot	Adjusted R Square	Residual plot
Exponential model		0.7843	The best
Power model		0.8883	Better than the linear model
Polynomial model	Fails		

So I can select the power model. The equation for power model would be:

Functional form: $Y = aX^b$

Linearized model and parameters: $\ln(Y) = b_0 + b_1 \ln(X)$

$a = e^{b0}$ $b = b_1$

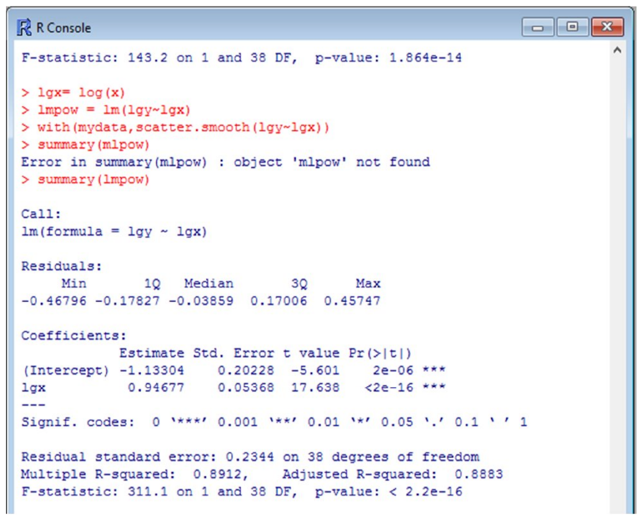
From R, we have got

b0 = -1.3304

b1= 0.94677

ln (Y) = -1.3304 + 0.94677 * ln (X)

⇨ (Raised to the power with base e)
⇨ Y = 0.26 * X $^{(0.94677)}$
⇨ BHD = 0.26 * (AGE) $^{(0.94677)}$



```
F-statistic: 143.2 on 1 and 38 DF,  p-value: 1.864e-14

> lgx= log(x)
> lmpow = lm(lgy~lgx)
> with(mydata,scatter.smooth(lgy~lgx))
> summary(mlpow)
Error in summary(mlpow) : object 'mlpow' not found
> summary(lmpow)

Call:
lm(formula = lgy ~ lgx)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46796 -0.17827 -0.03859  0.17006  0.45747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.13304    0.20228  -5.601    2e-06 ***
lgx          0.94677    0.05368  17.638   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2344 on 38 degrees of freedom
Multiple R-squared:  0.8912,    Adjusted R-squared:  0.8883
F-statistic: 311.1 on 1 and 38 DF,  p-value: < 2.2e-16
```

Example 1:

Age = 44

BHD = 0.26 * (44) $^{(0.94677)}$ = 9.35

Example 2:

Age = 39

BHD = 0.26 * (39) $^{(0.94677)}$ = 8.34

Using SPSS:

Now I check my result from R with SPSS:

Iteration History[b]

Iteration Number[a]	Residual Sum of Squares	Parameter	
		A	B
1.0	253487.400	.260	.947
1.1	95494.116	.339	.960
2.0	95494.116	.339	.960
2.1	95048.580	.340	.956
3.0	95048.580	.340	.956
3.1	95048.557	.340	.956
4.0	95048.557	.340	.956
4.1	95048.557	.340	.956

Derivatives are calculated numerically.

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
A	.340	.008	.325	.355
B	.956	.005	.946	.967

Correlations of Parameter Estimates

	A	B
A	1.000	-.995
B	-.995	1.000

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	2251273.443	2	1125636.721
Residual	95048.557	8785	10.819
Uncorrected Total	2346322.000	8787	
Corrected Total	612739.425	8786	

Dependent variable: bhd

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .845.

SPSS performed 8 iterations and found the equation as :

$BHD = 0.34 * (AGE)^{(0.956)}$

Discussion: Values I have retrieved with R are used as initial values in SPSS which seems to further improve the equation.