APM 630 Regression Analysis
HW9 – Nonlinear Regression Models
Name: Ruta Basijokaite

Project Report

1. Introduction

The tree data was obtained containing 900 pairs of tree diameter at breast height and total height for western white pine. The variables in the study included:

(1) Dependent variable:
DBH – tree diameter at breast height (cm)

(2) Predictor variables:
HT – total height (m)

The three DBH was ranged from 0 to 220 cm (by 10cm).

The purpose of this study was to develop the "best" model for predicting the tree HT using observed DBH in a long-term forest growth and yield simulator. Five nonlinear asymptotic growth functions are considered as candidate models:

[1] Gompertz Function:

$$HT = a * e^{-b*e^{-c*DBH}}$$

[2] Lundqvist Function:

$$HT = a * e^{\left(-b*DBH^{-c}\right)}$$

[3] Modified Logistic Function:

$$HT = \frac{a}{1 + \left(\frac{1}{b}\right) * DBH^{-c}}$$

[4] Richards Function:

$$HT = a * \left[1 - e^{-b*DBH}\right]^{c}$$

[5] Weibull Function:

$$HT = a * \left[1 - e^{-b*DBH^{c}}\right]$$

where a, b, and c are regression coefficients to be estimated. Simulated "asymptote" was compared with observed big tree record (DBH=208 cm and HT=65m, according to the 1995 Idaho Champion Big Tree List) to determine the "best" prediction performance for large-sized trees.

Additional work was done to analyze iteration method affects the model comparing parameter estimates, asymptotic standard errors for the parameters, and correlations among the parameters for the three iteration methods.

2. Methods

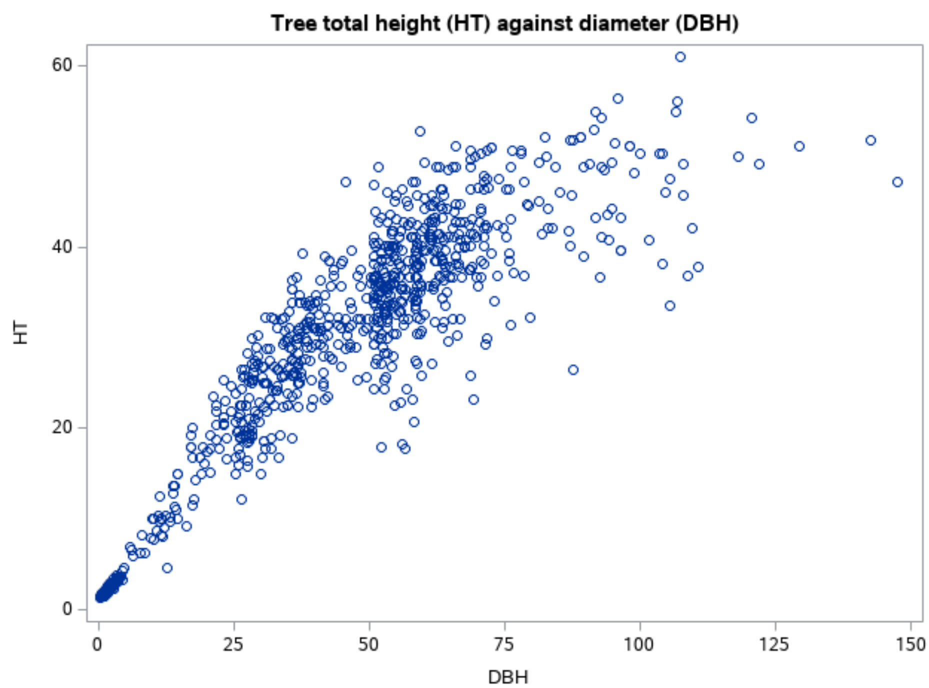
900 data points were collected to determine the relationship between total height and diameter at breast height. Descriptive statistics for all variables are summarized in Table 1.

Table 1. Summary of descriptive statistics of all variables

Variable	N	Mean	Median	Std Dev	Minimum	Maximum	Corrected SS	Uncorrected SS
HT	900	27.45	31.10	15.18	1.40	61.00	207093.41	885026.08
DBH	900	42.35	45.85	27.57	0.30	147.60	683562.88	2297495.98

Relationship between dependent and predictor variable is illustrated in Figure 1.

Figure 1. Scatter plot of total tree height against diameter



Widely used Ordinary Least Squares (OLS) models assumes that model coefficient structure is linear. However, in many natural sciences, nonlinear models may be more suitable to the data in the studies. In general, a nonlinear model is defined by at least one model coefficient being nonlinear in nature. Generally, nonlinear models fall into categories that are designed for specific situations. For example, five nonlinear functions presented below are often useful for modeling the “S-shaped” growth curves, in which Y is the response variable and X is the independent variable.

1) Gompertz Function:

$$Y = \alpha \cdot e^{-\beta e^{-\gamma X}} + \varepsilon \qquad [1]$$

where α is asymptotic model parameter; β and γ are model parameters controlling the shape of the curve. Gompertz function is non symmetric about its point of inflection.

2) Lundqvist (Koft) Function:

$$Y = \alpha \cdot e^{-\beta X^{-\gamma}} + \varepsilon \qquad [2]$$

where α is asymptotic model parameter; β is rate parameter; γ is the shape parameter of the curve.

3) Modified Logistic Function

$$Y = \frac{\alpha}{1+\frac{1}{\beta} \cdot X^{-\gamma}} + \varepsilon \qquad [3]$$

where α is asymptotic model parameter; β and γ are model parameters controlling the shape of the curve.

4) Richards Function

$$Y = \alpha \cdot (1 - e^{-\beta X})^{\gamma} + \varepsilon \qquad [4]$$

where α is asymptotic model parameter; β is rate parameter; γ is the shape parameter of the curve.

5) Weibull Function

$$Y = \alpha \cdot (1 - e^{-\beta X^{\gamma}}) + \varepsilon \qquad [5]$$

where α is asymptotic model parameter; β is rate parameter; γ is the shape parameter of the curve.

The numerical methods of fitting nonlinear models differ in the way of computing the $\Delta\hat{\theta}$ and in the rates of convergence to a solution and, in some cases, whether or not solution is obtained. There are numerous numerical methods that exist and no one method can be proclaimed as universally best and it may be desirable is some difficult cases to try more than one method. Some of the common methods for finding a solution to the nonlinear normal equation were analyzed in this study, which included:

- Gauss-Newton method. It used Taylor’s expansion of $f(X_i, \theta)$ about the starting values θ^0 to obtain a linear approximation of the model in the region near the starting values.
- Marquardt’s method. It is designed to capitalize on the best features of the Gauss-Newton and Steepest-Decent method.
- Newton method. It uses the second derivatives and solves the equation $\Delta\hat{\theta} = G^{-}X'e$

3. Results and Discussion

3.1 Three iteration methods (with Gompertz Function)

Results from three different iteration methods used are represented in Table 2.

Table 2a. Model result comparison using three different iteration methods

Estimation Summary	
Method	Gauss-Newton
Iterations	7
Subiterations	1
Average Subiterations	0.142857
R	4.788E-6
PPC(C)	4.071E-6
RPC(C)	0.000026
Object	1.344E-9
Objective	19426.36
Observations Read	900
Observations Used	899
Observations Missing	1

Estimation Summary	
Method	Marquardt
Iterations	6
Subiterations	6
Average Subiterations	1
R	4.003E-6
PPC(C)	3.864E-6
RPC(B)	0.000011
Object	9.48E-10
Objective	19426.36
Observations Read	900
Observations Used	899
Observations Missing	1

Estimation Summary	
Method	Newton
Iterations	6
R	1.497E-6
PPC(C)	1.379E-6
RPC(C)	0.000465
Object	5.847E-7
Objective	19426.36
Observations Read	900
Observations Used	899
Observations Missing	1

Three different iteration methods used produces very similar results. Two of the methods took 6 iterations, while one of them (Gauss-Newton) took 7. Parameter estimates produced by these methods (Table 2b) were equal producing MSE value equal to 21.6812. Approximate standard error values were identical for all three parameters for two of the methods (Marquardt and Gauss-Newton), while they were slightly higher for Newton. Approximate correlation matrix (Table 2c) indicated that none of the model parameters had high correlation (higher than 0.99). Correlation values were very similar in the first two methods and slightly higher in Newton method.

Table 2b. Parameter estimates and approximate standard error values approximated by three iteration methods. Top – Gauss-Newton, middle – Marquardt; bottom – Newton,

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	46.4205	0.5999	45.2431	47.5980
B	2.7996	0.0896	2.6238	2.9754
C	0.0460	0.00148	0.0431	0.0489

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	46.4205	0.5999	45.2431	47.5980
B	2.7996	0.0896	2.6238	2.9754
C	0.0460	0.00148	0.0431	0.0489

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	46.4205	0.6389	45.1666	47.6743
B	2.7996	0.0855	2.6317	2.9674
C	0.0460	0.00154	0.0430	0.0490

Table 2c. Approximate correlation matrixes produced by different iteration methods. Top – Gauss-Newton, middle – Marquardt; bottom – Newton.

Approximate Correlation Matrix			
	A	B	C
A	1.0000000	-0.3565133	-0.8299063
B	-0.3565133	1.0000000	0.7495395
C	-0.8299063	0.7495395	1.0000000

Approximate Correlation Matrix			
	A	B	C
A	1.0000000	-0.3565127	-0.8299062
B	-0.3565127	1.0000000	0.7495390
C	-0.8299062	0.7495390	1.0000000

Approximate Correlation Matrix			
	A	B	C
A	1.0000000	-0.3780364	-0.8552719
B	-0.3780364	1.0000000	0.7345284
C	-0.8552719	0.7345284	1.0000000

To estimate model parameter coefficients for five different functions described above, Gauss-Newton method was used.

3.2. Initial values

Nonlinear regression models for functions 1-5 were created to obtain initial values. Initial values obtained are represented in Table 3.

Table 3. Initial values (Values represented in the table were rounded as they were entered in SAS)

Function	a	b	c
Gompertz	65	2.669	0.0261
Lundqvist	120*	5*	0.4*
Modified Logistic	60*	0.01*	1*
Richards	65	0.01467	1
Weibull	65	0.02499	0.86106

*Indicate values that were changes due to non-convergence

3.3. Parameter estimates

3.3.1 Gompertz Function

Parameter estimates obtained from three iteration methods are represented in Table 4.

Table 4. Parameter estimates produced by iteration methods

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	46.4205	0.5999	45.2431	47.5980
B	2.7996	0.0896	2.6238	2.9754
C	0.0460	0.00148	0.0431	0.0489

Equation 1 was edited to fit the parameter values represented in Table 4 and can be expressed using the following equation:

$$\hat{Y} = 46.42 \cdot e^{-2.8e^{-0.046X}} \quad [6]$$

Model MSE was equal to 21.6812. Approximate standard errors for each model parameter estimate was below 1.

3.3.2 Lundqvist Function

Parameter estimates obtained for this function are represented in Table 5.

Table 5. Parameter estimates produced for Lundqvist function

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	138.6	24.7990	89.9740	187.3
B	6.9498	0.4770	6.0137	7.8858
C	0.4091	0.0491	0.3127	0.5055

Equation 2 was edited to fit the parameter values represented in Table 5 and can be expressed using the following equation:

$$\hat{Y} = 138.6 \cdot e^{-6.95X^{-0.41}} \quad [7]$$

Model MSE was equal to 21.3776. Approximate standard errors for β (or B in the table) and γ (or C in the table) were below 1, however α error estimate was much higher (equal to 24.8).

3.3.3 Modified Logistic Function

Parameter estimates obtained for this function are represented in Table 6.

Table 6. Parameter estimates produced for Modified Logistic function

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	68.2978	3.9123	60.6195	75.9761
B	0.00990	0.00147	0.00701	0.0128
C	1.1822	0.0645	1.0557	1.3088

Equation 3 was edited to fit the parameter values represented in Table 6 and can be expressed using the following equation:

$$\hat{Y} = \frac{68.3}{1 + \frac{1}{0.01} \cdot X^{-1.18}} \qquad [8]$$

Model MSE was equal to 20.8233. Approximate standard errors for β (or B in the table) and γ (or C in the table) were below 1, while α error estimate was higher equaling to 3.91.

3.3.4 Richards Function

Parameter estimates obtained for this function are represented in Table 7.

Table 7. Parameter estimates produced for Richards function

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	54.3887	1.7149	51.0231	57.7544
B	0.0210	0.00200	0.0171	0.0250
C	1.0690	0.0616	0.9482	1.1899

Equation 4 was edited to fit the parameter values represented in Table 7 and can be expressed using the following equation:

$$\hat{Y} = 54.39 \cdot (1 - e^{-0.02X})^{1.07} \qquad [9]$$

Model MSE was equal to 20.7093. Approximate standard errors for β and γ were below 1, while α error estimate was slightly higher than 1.

3.3.5 Weibull Function

Parameter estimates obtained for this function are represented in Table 8.

Table 8. Parameter estimates produced for Weibull function

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	53.5976	1.8960	49.8764	57.3188
B	0.0164	0.00189	0.0127	0.0201
C	1.0573	0.0421	0.9746	1.1400

Equation 5 was edited to fit the parameter values represented in Table 8 and can be expressed using the following equation:

$$\hat{Y} = 53.6 \cdot (1 - e^{-0.02X^{1.06}}) \qquad [10]$$

Model MSE was equal to 20.6981. Approximate standard errors for β and γ were below 1, while α error estimate was slightly higher than 1.

3.4. Simulation Results and Comparison

Simulation results can be found in Table 9.

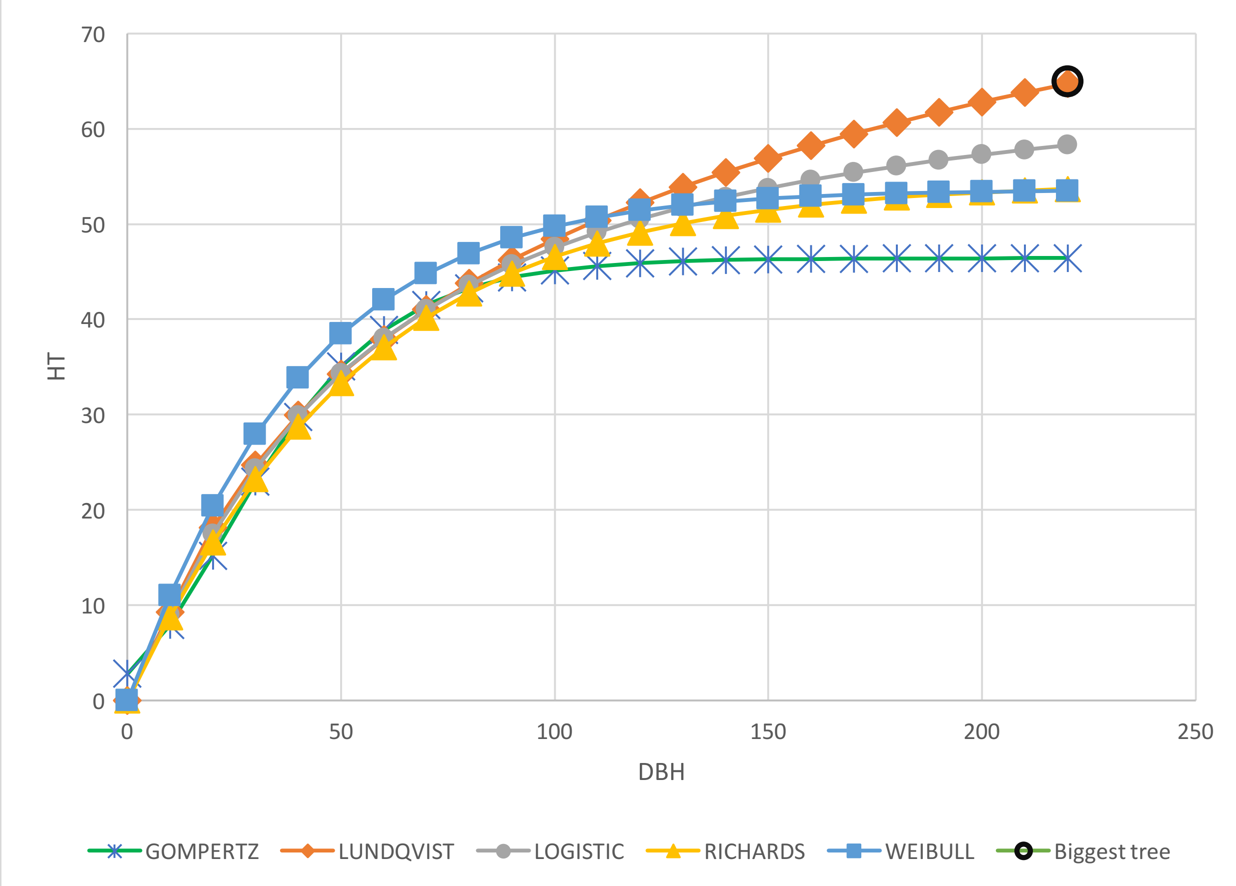
Table 9. Simulation results

Obs	DBH	GOMPERTZ	LUNDQVIST	LOGISTIC	RICHARDS	WEIBULL
1	0	2.8228	.	.	0	0
2	10	7.9259	9.2788	8.9786	8.7483	10.9973
3	20	15.2088	18.1125	17.4414	16.5912	20.3922
4	30	22.9498	24.7364	24.3306	23.2104	27.9199
5	40	29.7563	29.9644	29.8648	28.7259	33.846
6	50	35.0581	34.2578	34.3381	33.2947	38.4653
7	60	38.8813	37.8861	37.9952	37.0664	42.0419
8	70	41.5068	41.0184	41.0224	40.1734	44.7972
9	80	43.2547	43.7673	43.5585	42.729	46.9115
10	90	44.3959	46.2116	45.7069	44.8288	48.5285
11	100	45.1317	48.4081	47.5456	46.5528	49.7617
12	110	45.6025	50.3995	49.1338	47.9675	50.6999
13	120	45.9022	52.2183	50.517	49.1277	51.4121
14	130	46.0925	53.8902	51.731	50.079	51.9518
15	140	46.213	55.4354	52.8036	50.8587	52.3599
16	150	46.2892	56.8705	53.7572	51.4977	52.6682
17	160	46.3374	58.209	54.6098	52.0212	52.9006
18	170	46.3678	59.462	55.3761	52.4501	53.0756
19	180	46.3871	60.6391	56.0681	52.8014	53.2072
20	190	46.3992	61.7481	56.6957	53.0891	53.3061
21	200	46.4069	62.7958	57.2671	53.3248	53.3803
22	210	46.4117	63.7882	57.7894	53.5178	53.4359
23	220	46.4148	64.7303	58.2683	53.6758	53.4775

Simulation results are illustrated in Figure 2.

MSE values produces by five model functions varied slightly implying that all five model fits are similarly accurate. Gompertz function produced the highest MSE value equal to 21.6812, while Weibull MSE was the lowest and equal to 20.6981.

Figure 2. Simulation results for five different functions



Although overall fit of the five models is similar, Lundquist function is “best” at predicting tree height based on its diameter for taller trees as can be seen in Figure 2. Since accurately predicting tall trees is important in the field of forestry, Lundquist function can be concluded as the “best” model for predicting tree height.

4. Summary

Relationship between tree height and diameter at breast height was analyzed. Five nonlinear model functions were analyzed to develop a model predicting tree height based on its diameter. The main purpose was to determine which model function had the “best” prediction performance for the larger trees. After comparing three different iteration methods and five model functions, Lundquist function was concluded as the “best” model for predicting tree height for larger trees, as it was much more accurate at predicting height of larger trees.

5. SAS programs

```
*HW9* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW9/Tree.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW9DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  YMAX=65;
  *X=DBH;
  Y=HT;
  LNX=LOG(DBH);
  LOGISTIC=LOG((YMAX/Y)-1);
  GOMPERTZ=LOG(-LOG(Y/YMAX));
  RICHARDS=LOG(1-(Y/YMAX));
  WEIBULL=LOG(-LOG(1-(Y/YMAX)));
  LUNDQVIST=LOG(-LOG(Y/YMAX));
  SET HW9DATA;
RUN;
PROC SGPLOT DATA=ALL NOAUTOLEGEND;
   SCATTER X=DBH Y=HT;* / MARKERATTRS=(SYMBOL=CIRCLEFILLED COLOR=BLACK);
   TITLE'Tree total height (HT) against diameter (DBH)';
RUN;
PROC MEANS N MEAN MEDIAN STD MIN MAX CSS USS MAXDEC=2 DATA=ALL;
   VAR HT DBH;
   TITLE 'DESCRIPTIVE STATISTICS';
RUN;
*====ESTIMATING INITIAL VALUES OF MODEL PARAMETERS====;
PROC REG DATA=ALL;
  MODEL GOMPERTZ = DBH;
  MODEL LUNDQVIST = LNX;
  MODEL LOGISTIC = LNX;
  MODEL RICHARDS = DBH / NOINT;
  MODEL WEIBULL = LNX;
RUN;
TITLE'GOMPERTZ FUCTION';
*====DIFFERENT ITERATION METHODS FOR GOMPERTZ=====;
PROC NLIN DATA=ALL LIST;*METHOD=GAUSS MARQUARDT NEWTON;
  PARMS A=65 B=2.7 C=0.03;
  MODEL Y=A*EXP(-B*EXP(-C*DBH));
  OUTPUT OUT=OUT1 P=PREDICT R=RESIDUAL;
RUN;
ODS GRAPHICS ON;
PROC SGPLOT DATA=OUT1;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT1;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
PROC NLIN DATA=ALL METHOD=MARQUARDT;
  PARMS A=65 B=2.7 C=0.03;
  MODEL Y=A*EXP(-B*EXP(-C*DBH));
  OUTPUT OUT=OUT2 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT2;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
```

```
PROC SGPLOT DATA=OUT2;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
PROC NLIN DATA=ALL METHOD=NEWTON;
  PARMS A=65 B=2.7 C=0.03;
  MODEL Y=A*EXP(-B*EXP(-C*DBH));
  OUTPUT OUT=OUT3 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT3;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT3;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
*=====LUNDQVIST FUNCTION;
TITLE'LUNDQVIST FUCTION';
PROC NLIN DATA=ALL LIST;
  *PARMS A=65 B=4 C=0.5;
  PARMS A=120 B=5 C=0.4;
  MODEL Y=A*EXP(-B*DBH**(-C));
  OUTPUT OUT=OUT4 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT4;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT4;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
*=====MODIFIED LOGISTIC FUNCTION;
TITLE'MODIFIED LOGISTIC FUCTION';
PROC NLIN DATA=ALL LIST;
  *PARMS A=65 B=43.2 C=1;
  PARMS A=60 B=0.01 C=1.0;
  MODEL Y=A/(1+(1/B)*DBH**(-C));
  OUTPUT OUT=OUT5 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT5;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT5;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
*=====RICHARDS FUNCTION;
TITLE'RICHARDS FUCTION';
PROC NLIN DATA=ALL LIST;
  PARMS A=65 B=0.01 C=1;
  MODEL Y=A*(1-EXP(-B*DBH))**C;
  OUTPUT OUT=OUT6 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT6;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT6;
```

```
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
*=====WEIBULL FUNCTION;
TITLE'WEIBULL FUCTION';
PROC NLIN DATA=ALL LIST;
  PARMS A=65 B=0.02 C=0.9;
  MODEL Y=A*(1-EXP(-B*DBH**C));
  OUTPUT OUT=OUT7 P=PREDICT R=RESIDUAL;
RUN;
PROC SGPLOT DATA=OUT7;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT7;
  SCATTER X=DBH Y=PREDICT;
  REFLINE 0;
RUN;
DATA NEW;
  DO DBH = 0 TO 220 BY 10;
  GOMPERTZ=46.42*EXP(-2.8*EXP(-0.046*DBH));
  LUNDQVIST = 138.6*EXP(-6.95*DBH**(-0.41));
  LOGISTIC = 68.3/(1+(1/.01)*DBH**(-1.18));
  RICHARDS = 54.39*(1-EXP(-0.02*DBH))**1.07;
  WEIBULL = 53.6*(1-EXP(-0.02*DBH**1.06));
  OUTPUT;
  END;
  SET NEW;
RUN;
ODS GRAPHICS OFF;
PROC EXPORT DATA=NEW
            FILE=NEW
            DBMS=CSV REPLACE;
RUN;
%let _DATAOUT_MIME_TYPE=table/csv;
%let _DATAOUT_NAME=Simulation.csv;
```