Project Report

Name: Yifang Zhang

Email:yzhan377@syr.edu

1. Introduction

1.1 Background

A group of scientists made a research to figure out the relationship between the body weight and the brain weight of animals. Data on body weight and brain weight of 28 different kinds of animals were collected with two variables:

(1) One dependent variable:

Y = Brain weight (g)

(2) One independent variable:

X = Body weight (kg)

1.2 Purposes

The purposes of the study are:

(1) Compute Pearson and Spearman correlations between Y and X, LnY and LnX;

(2) Find the scatterplot of Y and X, LnY and LnX;

(3) Fit the OLS model $Y = \beta_0 + \beta_1 * X + \varepsilon$;

(4) For the model residuals, testing the normality, homogeneity of variance, possible outliers and autocorrelations;

(5) Fit the log-transformed model $Ln(Y) = \beta_0 + \beta_1 * Ln(X) + \varepsilon$;

(6) For the model residuals, testing the normality, homogeneity of variance, possible outliers and autocorrelations.

2. Data and Methods

2.1 Basic Data on Y and X

Data on body weight and brain weight of 28 different kinds of animals were collected with two variables Y and X defined above. The descriptive statistics of Y and X is shown in Table 1.

Table 1. Descriptive statistics of Y and X.

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
Y	28	574.52143	1335	137.00000	0.40000	5712
X	28	4278	16480	53.83000	0.02300	87000

Table 2 shows the Pearson correlation and Spearman correlation between the variables Y and X. From the Pearson correlation table, we can find that the linear correlation coefficient $\rho = -0.00534$, which means that Y and X are almost linearly uncorrelated. From the Spearman correlation table, we can find that the Spearman rank correlation coefficient $r_s = 0.71630$, which means the two variables Y and X are correlated without requiring that the relationship between Y and X is linear.

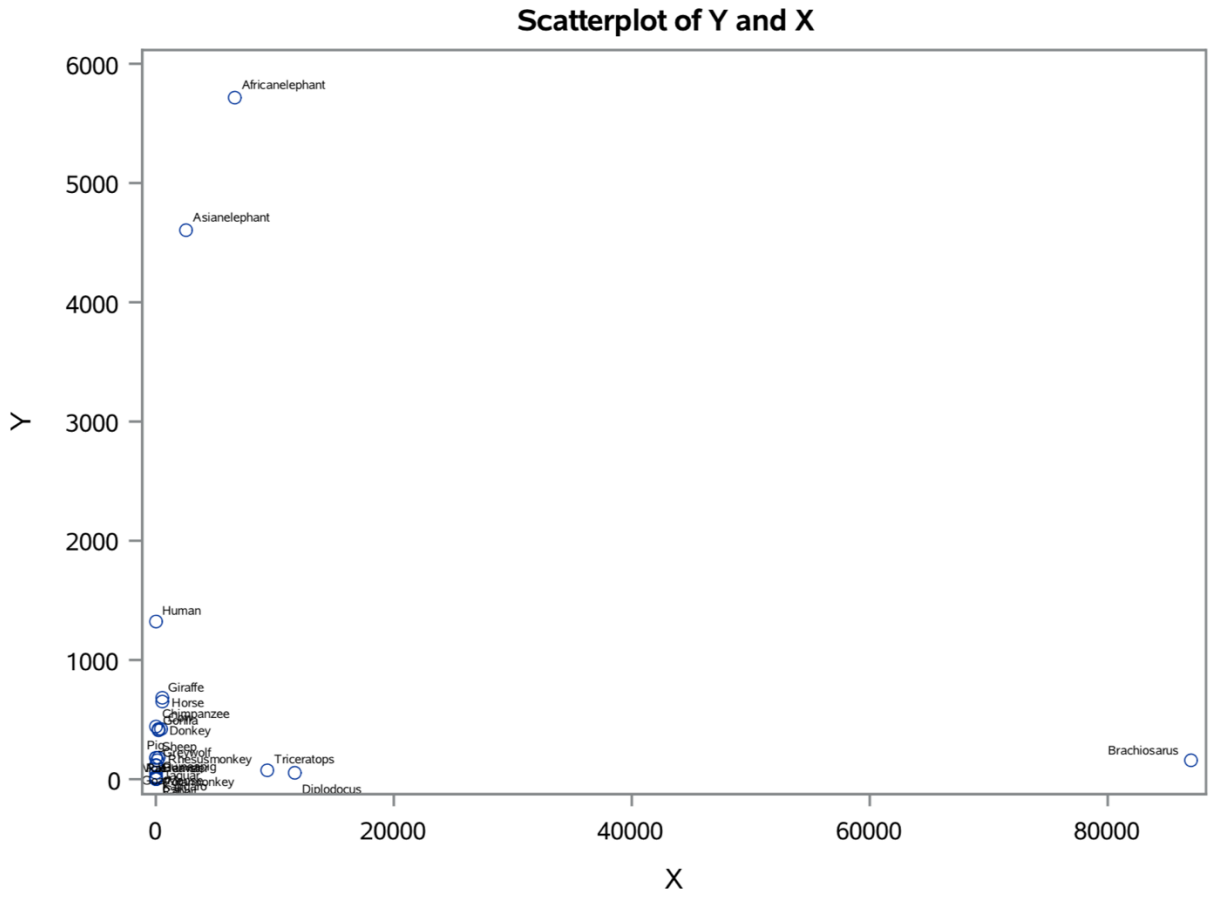
Table 2. Pearson and Spearman Correlations between Y and X.

Pearson Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	Y	X
Y	1.00000	-0.00534 0.9785
X	-0.00534 0.9785	1.00000

Spearman Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	Y	X
Y	1.00000	0.71630 <.0001
X	0.71630 <.0001	1.00000

Figure 1 draws the scatterplot of Y and X with data names. From the scatterplot we can double check that Y and X are linearly uncorrelated.

Figure 1. Scatterplot of Y and X



2.2. Methods

First, in this research, the OLS method was applied to fit the following linear regression model by using Statistical Analysis System (SAS):

$$Y = \beta_0 + \beta_1 * X + \varepsilon \qquad [1]$$

Where Y and X are defined as above in the background. β_0, β_1 are regression coefficients to be estimated, and ε is the model random error.

Second, we compute the natural log-transformation for both variables to get LnY and LnX and fit the log-transformation model by using Statistical Analysis System (SAS):

$$Ln(Y) = \beta_0 + \beta_1 * Ln(X) + \varepsilon \qquad [2]$$

Where LnY and LnX are computed by natural log-transformation. β_0, β_1 are regression coefficients to be estimated, and ε is the model random error.

Finally, for both model residuals, we test for: (1) normality by Shapiro-Wilk test since n<2000, (2) homogeneity of variance by White test, (3) possible outliers and (4) autocorrelations by Durbin-Watson statistic.

2.3 Basic Data on LnY and LnX

Table 3 exhibits the descriptive statistics of LnY and LnX, and table 4 gives the natural log-transformation for Y and X.

Table 3. Descriptive statistics of LnY and LnX.

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
LNY	28	4.42545	2.39928	4.91176	-0.91629	8.65032
LNX	28	3.77131	3.77066	3.98535	-3.77226	11.37366

Table 4. The natural log-transformation for Y and X

Obs	SPCS	X	Y	LNY	LNX
1	Beaver	1.35	8.1	2.09186	0.3001
2	Cow	465	423	6.04737	6.1420
3	Greywolf	36.33	119.5	4.78332	3.5926
4	Goat	27.66	115	4.74493	3.3200
5	Guineapig	1.04	5.5	1.70475	0.0392
6	Diplodocus	11700	50	3.91202	9.3673
7	Asianelephant	2547	4603	8.43446	7.8427
8	Donkey	187.1	419	6.03787	5.2316
9	Horse	521	655	6.48464	6.2558
10	Potarmonkey	10	115	4.74493	2.3026
11	Cat	3.3	25.6	3.24259	1.1939
12	Giraffe	529	680	6.52209	6.2710
13	Gorilla	207	406	6.00635	5.3327
14	Human	62	1320	7.18539	4.1271
15	Africanelephant	6654	5712	8.65032	8.8030
16	Triceratops	9400	70	4.24850	9.1485
17	Rhesusmonkey	6.8	179	5.18739	1.9169
18	Kangaro	35	56	4.02535	3.5553
19	Hamster	0.12	1	0.00000	-2.1203
20	Mouse	0.023	0.4	-0.91629	-3.7723
21	Rabbit	2.5	12.1	2.49321	0.9163
22	Sheep	55.5	175	5.16479	4.0164
23	Jaguar	100	157	5.05625	4.6052
24	Chimpanzee	52.16	440	6.08677	3.9543
25	Brachiosarus	87000	154.5	5.04019	11.3737
26	Rat	0.28	1.9	0.64185	-1.2730
27	Mole	0.122	3	1.09861	-2.1037
28	Pig	192	180	5.19296	5.2575

Table 5 shows the Pearson correlation and Spearman correlation between the variables LnY and LnX. From the Pearson correlation table, we can find that the linear correlation coefficient $\rho = 0.77949$, which means that Y and X have strong linear correlation. From the Spearman correlation table, we can find that the Spearman rank correlation coefficient $r_s = 0.71630$.

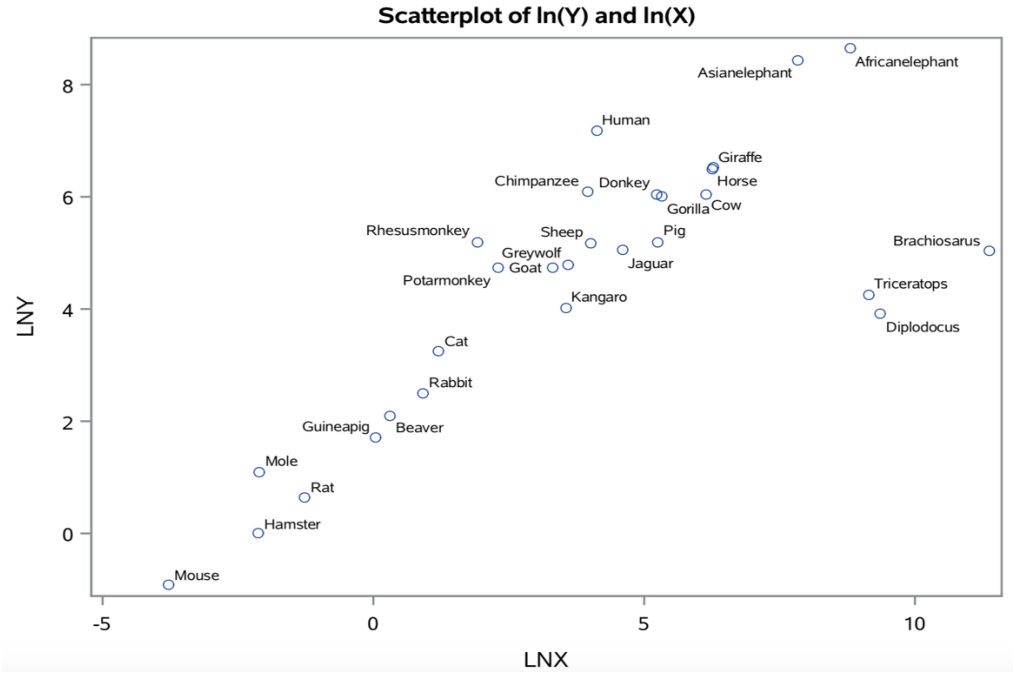
Table 5. Pearson and Spearman Correlations between LnY and LnX.

Pearson Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	LNY	LNX
LNY	1.00000	0.77949 <.0001
LNX	0.77949 <.0001	1.00000

Spearman Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	LNY	LNX
LNY	1.00000	0.71630 <.0001
LNX	0.71630 <.0001	1.00000

Figure 2 draws the scatterplot of LnY and LnX. From the scatterplot we can double check that LnY and LnX are linearly correlated.

Figure 2. Scatterplot of LnY and LnX.



3. Results and Discussion

3.1 The OLS model

Equation [1] was fit the data as follows:

$$Y = 576.37244 + (-0.0004326) *X \qquad [1]$$

The model R^2 was 0 and adjusted R^2 was -0.0384, indicating that none of the total variation in Y can be explained by the variable X by this model.

And the p-value of slope coefficient (β_1) is 0.9785 > 0.05, which means that it's not statistically significant. Table 6 gives the parameter estimates for the OLS model.

Table 6. Parameter Estimates of the OLS model.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	576.37244	265.91210	2.17	0.0395
X	1	-0.00043264	0.01589	-0.03	0.9785

3.2 The Log-transformation Model

Equation [2] was fit the log-transformation data as follows:

$$LnY = 2.5549 + 0.4960*LnX \qquad [2]$$

The model R^2 was 6076 and adjusted R^2 was 5925. The slope coefficients of X (β_1) is statistically significant at $\alpha = 0.05$ with the p-value < 0.001. Table 7 shows the parameter estimates of the log-transformation model.

Table 7. Parameter Estimates of the log-transformation model.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.55490	0.41314	6.18	<.0001
LNX	1	0.49599	0.07817	6.35	<.0001

3.3 Residual Analysis of the OLS model

(1) Residual plot and Student residual plot

The residual plot and student residual plot of the OLS model were respectively shown by Figure 3 and Figure 4.

Figure 3. The Residual Plot for the OLS model

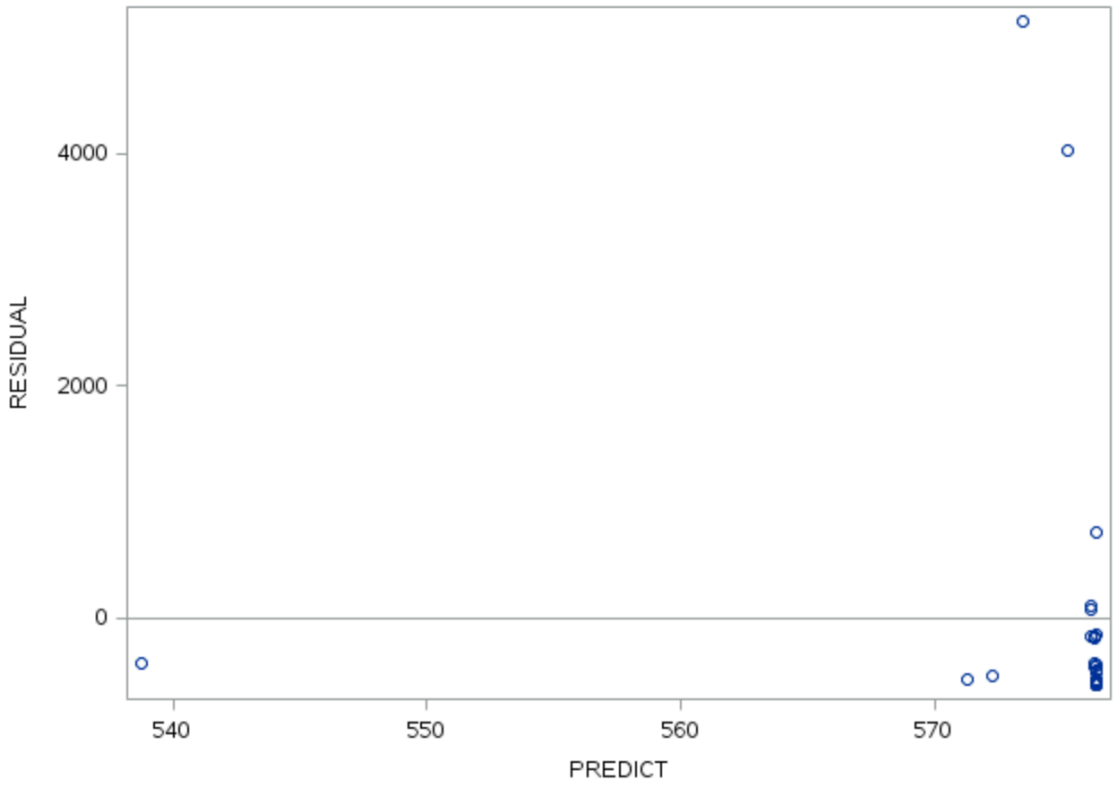
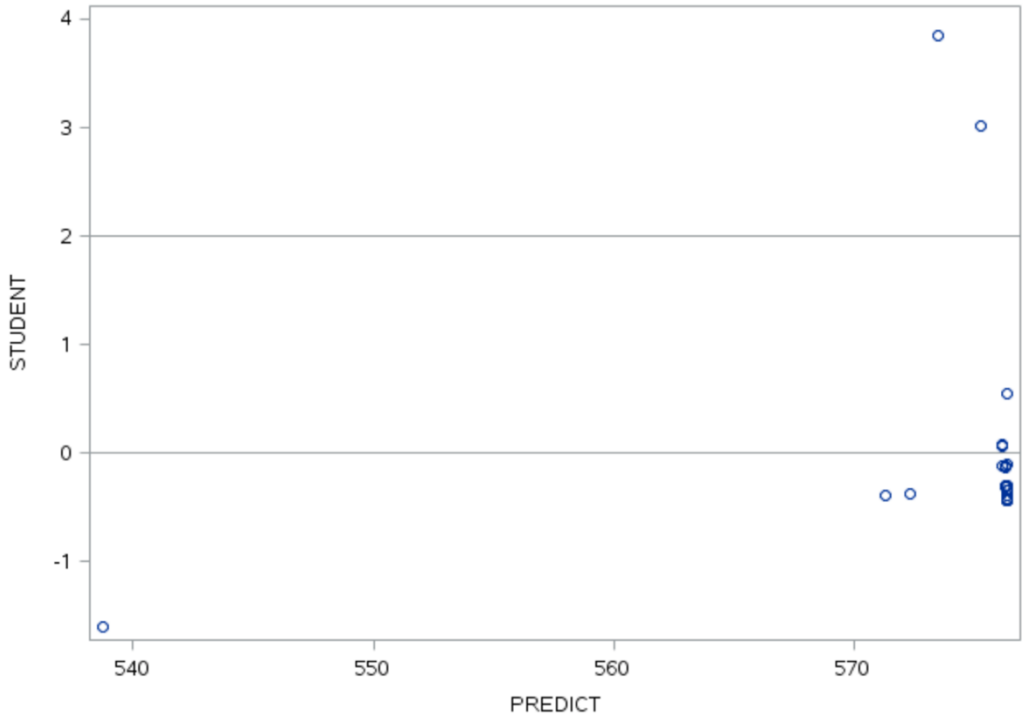


Figure 4. The Student Residual Plot for the OLS model



(2) Possible outliers

From Figure 4 we can find that there are plots (residuals) outside of (-2, +2). Thus, we can say that there are outliers.

(3) Normality Test

Table 8 shows the result of the normality test. Since the sample size is less than 2000, we compute the Shapiro-Wilk statistic to test the normality. The null hypothesis is that the values of the variable (the OLS model residuals) are a random sample from the normal distribution. Because the p-value < 0.001 <0.05, we can reject the hull hypothesis, which means that the model residuals are not normality.

Table 8. Test for Normality of the OLS model residuals.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.452258	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.362106	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.216081	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.243732	Pr > A-Sq	<0.0050

(4) Homogeneity of variance Test

Table 9 shows the result the homogeneity of variance test. We compute White test for homoscedasticity and the hull hypothesis is that the OLS model residuals are homoscedastic. Because the p-value = 0.4598 > 0.05, we cannot reject the null hypothesis, which indicates that the OLS model residuals are homoscedastic.

Table 9. Homogeneity of variance test of the OLS model residuals

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
2	1.55	0.4598

(5) Autocorrelations Test

Table 10 shows the result of the autocorrelation test. Durbin-Watson statistic is used to detect series autocorrelation (ρ) in the OLS model residuals. The null hypothesis H_0: ρ=0. Since the p-value of Pr < DW is 0.4024 > 0.05, and the p-value of Pr > DW is 0.5976 > 0.05, thus we cannot reject the null hypothesis, which means there is no series autocorrelation.

Table 10. Autocorrelations Test of the OLS model residuals

Durbin-Watson D	1.903
Pr < DW	0.4024
Pr > DW	0.5976
Number of Observations	28
1st Order Autocorrelation	0.043

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

3.4 Residual Analysis of the Log-transformation model

(6) Residual plot and Student residual plot

The residual plot and student residual plot of the Log-transformation model were respectively shown by Figure 5 and Figure 6.

Figure 5. The Residual Plot for the Log-transformation model

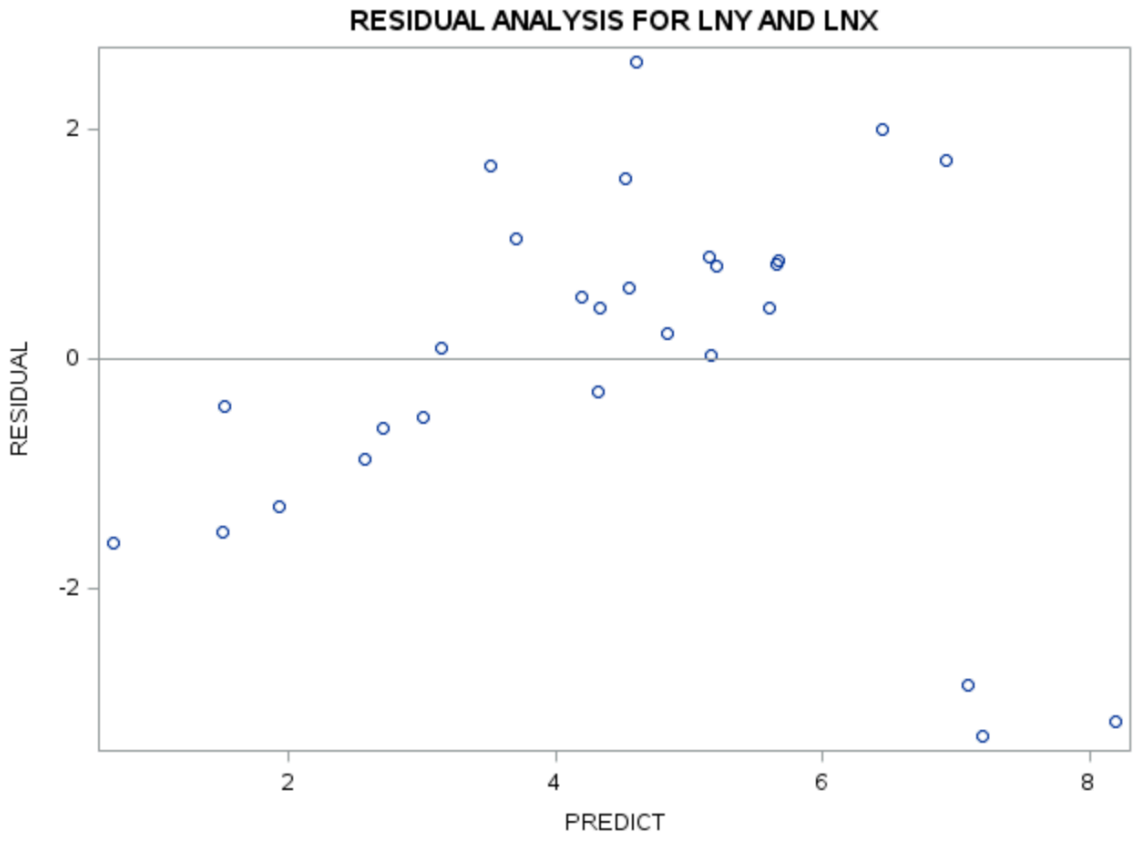
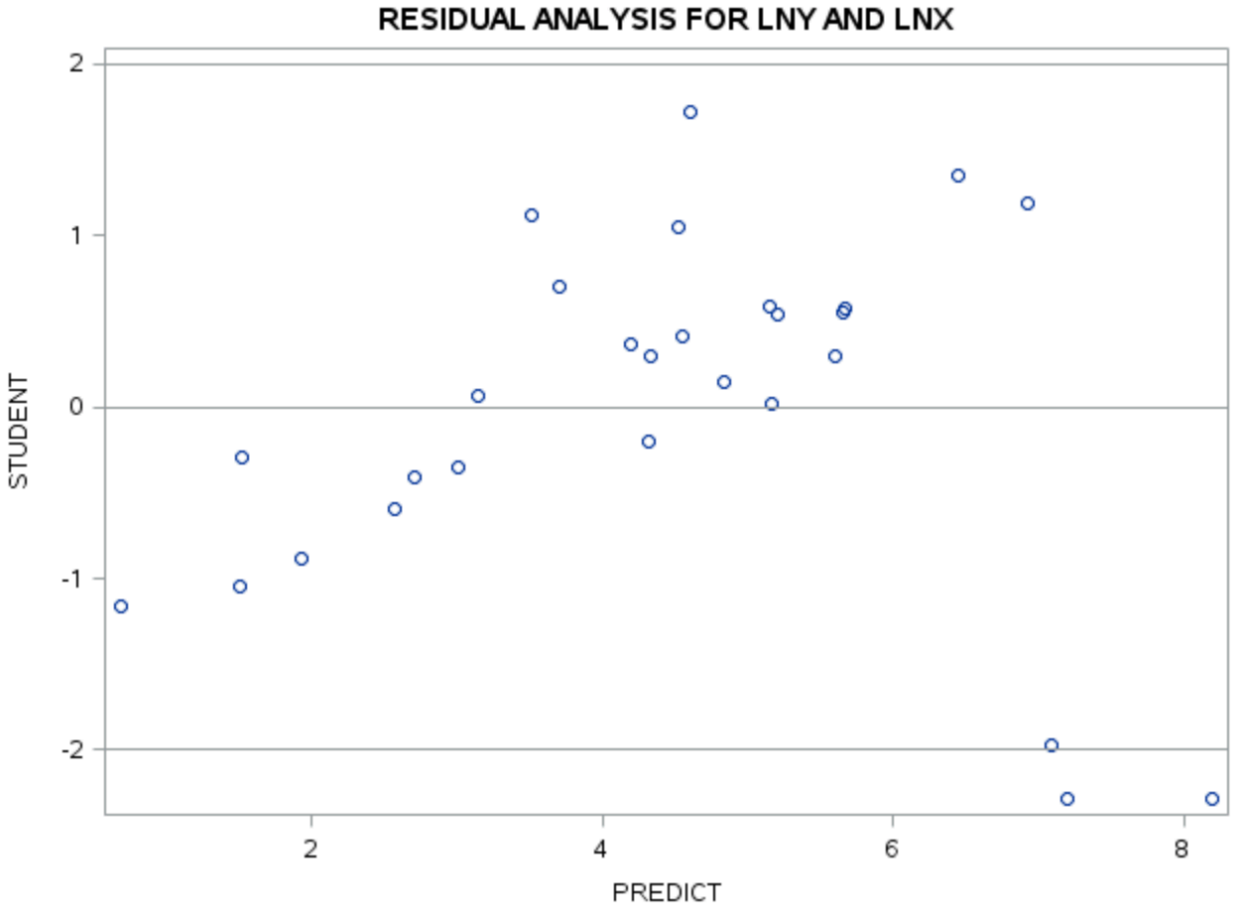


Figure 6. The Student Residual Plot for the Log-transformation model



(7) Possible outliers

From Figure 6 we can find that there are two plots (residuals) outside of (-2, +2). Thus, we can say that there are outliers.

(8) Normality Test

Table 11 shows the result of the normality test. Since the sample size is less than 2000, we compute the Shapiro-Wilk statistic to test the normality. The null hypothesis is that the values of the variable (the Log-transformation model residuals) are a random sample from the normal distribution.

Because the p-value = 0.1929 > 0.05, we cannot reject the hull hypothesis, which means that the model residuals are normality.

Table 11. Test for Normality of the Log-transformation model residuals.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.94954	Pr < W	0.1929
Kolmogorov-Smirnov	D	0.116687	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.074186	Pr > W-Sq	0.2411
Anderson-Darling	A-Sq	0.492211	Pr > A-Sq	0.2096

(9) Homogeneity of variance Test

Table 12 shows the result the homogeneity of variance test. We compute White test for homoscedasticity and the hull hypothesis is that the Log-transformation model residuals are homoscedastic. Because the p-value = 0.1355 > 0.05, we cannot reject the null hypothesis, which indicates that the Log-transformation model residuals are homoscedastic.

Table 12. Homogeneity of variance test of the Log-transformation model residuals

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
2	4.00	0.1355

(10) Autocorrelations Test

Table 12 shows the result of the autocorrelation test. Durbin-Watson statistic is used to detect series autocorrelation (ρ) in the Log-transformation model residuals. The null hypothesis H_0: ρ=0. Since the p-value of Pr < DW is 0.4748> 0.05, and the p-value of Pr > DW is 0.5252 > 0.05, thus we cannot reject the null hypothesis, which means there is no series autocorrelation.

Table 12. Autocorrelations Test of the Log-transformation model residuals

Durbin-Watson D	1.992
Pr < DW	0.4748
Pr > DW	0.5252
Number of Observations	28
1st Order Autocorrelation	0.001

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

4. Summary

In this research, we compute Pearson and Spearman correlations on the two variables Y (the brain weight / g) and X (the body weight / kg) and find that Y and X maybe correlated but not linearly correlated. Then we fit the OLS model and find that the model cannot explain the variation in Y.

Thus we use the log-transformation model to find the relationship between Y and X. By computing Pearson and Spearman correlations on the natural log-transformation variables LnY and LnX, we find they have linearly correlation. Then we fit the model and test both two model residuals for normality, possible outliers, homogeneity of variance and autocorrelation.

5. SAS programs

5.1 Programs on computing Y and X

```
OPTIONS NODATE LS=76 PS=45 PAGENO=1 NOLABEL;
PROC IMPORT
DATAFILE='/home/yifangz02120/sasuser.v94/Brain.xls'
OUT=BRAIN
DBMS=XLS
REPLACE;
RUN;
DATA ALL;
SET BRAIN;
RUN;
*===COMPUTE CORRELATIONS AND DRAW SCATTERPLOT===;
PROC CORR DATA=ALL PEARSON SPEARMAN;
VAR Y X;
RUN;
TITLE 'Scatterplot of Y and X';
PROC SGPLOT DATA=ALL;
SCATTER X=X Y=Y / DATALABEL=SPCS;
RUN;

*===FIT THE OLS WITH RESIDUAL===;
TITLE 'RESIDUAL ANALYSIS FOR Y AND X';
PROC REG DATA=ALL;
MODEL Y = X / SPEC DWPROB;
OUTPUT OUT=OUT1
P=PREDICT R=RESIDUAL STDR=STDR
STUDENT=STUDENT RSTUDENT=RSTUDENT H=H;
RUN;
PROC SGPLOT DATA=OUT1;
```

```
SCATTER X=PREDICT Y=RESIDUAL;
REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT1;
SCATTER X=PREDICT Y=STUDENT;
REFLINE 0;
REFLINE 2;
REFLINE -2;
RUN;
PROC PRINT DATA=OUT1;
VAR Y X PREDICT RESIDUAL STDR STUDENT RSTUDENT H;
RUN;
PROC UNIVARIATE DATA=OUT1 PLOT NORMAL;
VAR RESIDUAL;
RUN;
```

5.2 Programs on computing LnY and LnX

```
OPTIONS NODATE LS=76 PS=45 PAGENO=1 NOLABEL;
PROC IMPORT
DATAFILE='/home/yifangz02120/sasuser.v94/Brain.xls'
OUT=BRAIN
DBMS=XLS
REPLACE;
RUN;
DATA LNALL;
SET BRAIN;
LNY = LOG(Y);
LNX = LOG(X);
RUN;
TITLE 'The natural log-transformation for Y and X';
PROC PRINT DATA=LNALL;
```

```
RUN;
TITLE 'Correlations';
PROC CORR DATA=LNALL PEARSON SPEARMAN;
VAR LNY LNX;
RUN;
TITLE 'Scatterplot of ln(Y) and ln(X)';
PROC SGPLOT DATA=LNALL;
SCATTER X=LNX Y=LNY / DATALABEL=SPCS;
RUN;

*===FIT THE OLS WITH RESIDUAL===;
TITLE 'RESIDUAL ANALYSIS FOR LNY AND LNX';
PROC REG DATA=LNALL;
MODEL LNY = LNX / SPEC DWPROB;
OUTPUT OUT=OUT2
P=PREDICT R=RESIDUAL STDR=STDR
STUDENT=STUDENT RSTUDENT=RSTUDENT H=H;
RUN;
PROC SGPLOT DATA=OUT2;
SCATTER X=PREDICT Y=RESIDUAL;
REFLINE 0;
RUN;
PROC SGPLOT DATA=OUT2;
SCATTER X=PREDICT Y=STUDENT;
REFLINE 0;
REFLINE 2;
REFLINE -2;
RUN;
PROC PRINT DATA=OUT2;
VAR Y X LNY LNX PREDICT RESIDUAL STDR STUDENT RSTUDENT H;
RUN;
```

```
PROC UNIVARIATE DATA=OUT2 PLOT NORMAL;
VAR RESIDUAL;
HISTOGRAM RESIDUAL;
PROBPLOT RESIDUAL;
RUN;
```