APM 630 Regression Analysis
HW5 – Residual Analysis
Name: Ruta Basijokaite

Project Report

1. Introduction

A study was conducted to investigate the relationship between body weight and brain weight. The variables in the study included:

Y = Brain weight (g)
X = Body weight (kg)

The purpose of the study is to (1) compute Pearson and Spearman correlations between variables, (2) compute natural log-transformation of both variables, (3) compute Pearson and Spearman correlations between lot-transformed variables, (4) fit OLS model to the data, (5) fit log-transformed model to the data, (6) test for normality, homogeneity of variance, possible outliers and autocorrelations for both models.

2. Methods

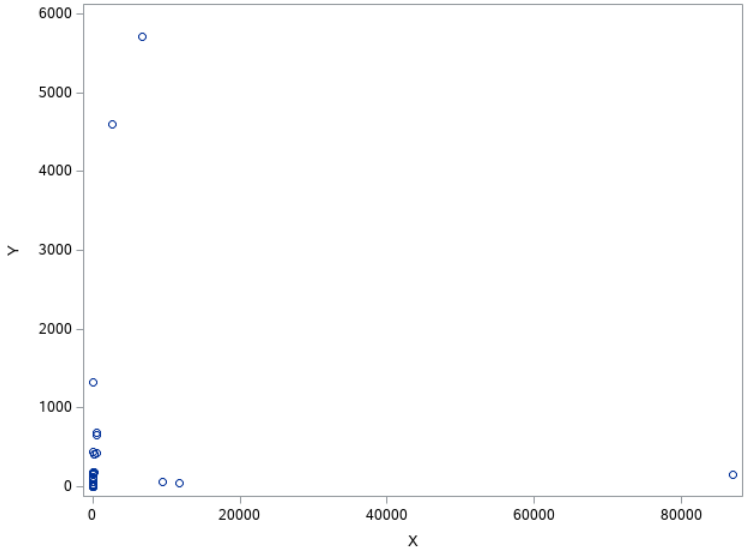
Body weight and brain weight data was collected from 28 animal species. Relationship between these variables were tested. Pearson and Spearman correlation coefficients are listed in Table 1.

Table 1. Pearson (left) and Spearman (right) correlation coefficients for both variables

Pearson Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0			Spearman Correlation Coefficients, N = 28 Prob > |r| under H0: Rho=0		
	Y	X		Y	X
Y	1.00000	-0.00534 0.9785	Y	1.00000	0.71630 <.0001
X	-0.00534 0.9785	1.00000	X	0.71630 <.0001	1.00000

Figure 1 illustrates the relationship between body weight (X) and brain weight (Y).

Figure 1. Relationship between body weight and brain weight.



Simple Ordinary Least Squares (OLS) model was created using Statistical Analysis System (SAS) to fit the data represented in the section 1 of this report. Model was developed based on the following linear regression formula:

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

where Y is brain weight (g), β_0 - β_1 are regression coefficients to be estimated, X is body weight (kg) and ε is the model error.

Variable transformation is often used to simplify the relationship between dependent variables and independent variables. Dependent variable can be transformed to overcome heterogeneous variance of model error terms. Also, dependent variable can be transformed to correct for non-normality. Many models non-linear in the coefficients can be linearized and re-expressed as a linear function of parameters, by appropriate transformations. In this case, both variables (Y and X) were natural log-transformed. After variable transformation, Pearson and Spearman correlation coefficients were computed (Table 2).

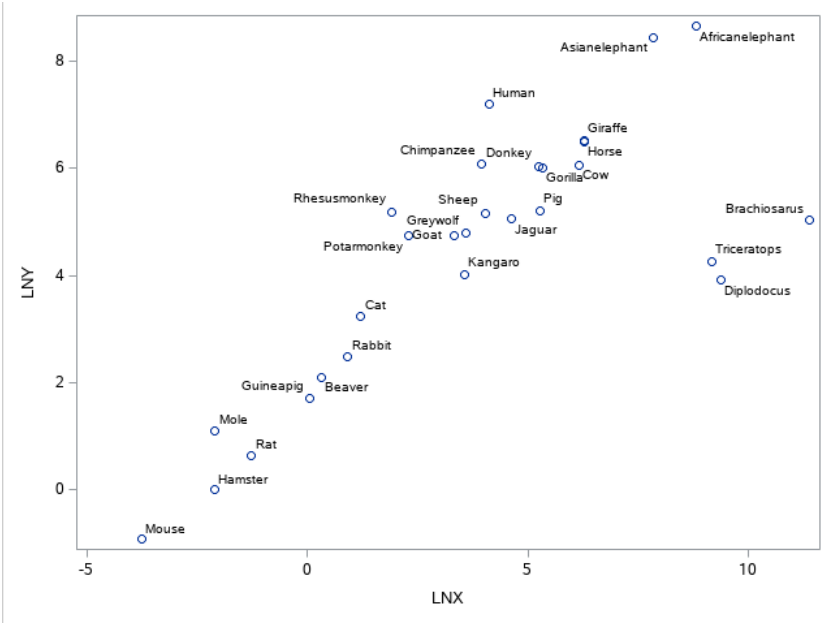
Table 2. Pearson (left) and Spearman (right) correlation coefficients for log-transformed variables

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 illustrates the relationship between natural log-transformed variables.

Figure 2. Relationship between log-transformed variables.



To fit log-transformed data, equation 1 was modified and log-transformed model was developed using the following formula:

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

where Y and X variables are natural log-transformed.

The OLS regression method is based on the assumption that the errors are additive, normally distributed, and independent with a mean zero and common variance σ^2. Some of the major problems in regression analysis relate to failures of basic assumptions: normality, common variance, independence of the errors, and outliers. Therefore, to efficiently evaluate linear regression models [1] and [2], model residuals were tested to ensure that basic OLS model assumptions are not violated.

- Non-normality. The assumption that the model residuals are normally distributed is not necessary for the estimation of regression coefficients and partitioning of total variation. Normality is needed only for hypothesis testing of significance. Commonly, variable transformations of dependent variable are a usual recourse to non-normality.
- Heterogeneous variance. The assumption of common variance implies that every observation on the dependent variable receives the same weight. The direct impact if heterogeneous variance is a loss of precision in the coefficient estimates.
- Correlated model residuals. Correlation among the residuals may arise from many sources and result in loss of precision in the coefficient estimates.
- Outliers. Since OLS gives equal weight to every observation, a single point far removed from the other data points can have almost as much influence on the regression results as all other points combined.

Tests that were performed to evaluate for normality, homogeneity of variance, possible outliers, and autocorrelations are summarized in Table 3.

Table 3. Tests that were performed to evaluate for normality, homogeneity of variance, possible outliers, and autocorrelations

Normality	• Shapiro-Wilk test • Kolmogorov-Smirnov test • Anderson-Darling test • Cramer-von Mises test
Homogeneity of variance	• White test
Correlated model residuals	• Durbin-Watson test
Outliers ($\overline{X} \pm 2 \cdot S$)	• Studentized Residuals • R-Student

3. Results and Discussion

3.1 Regression Model fitted with non-transformed data

Equation 1 was edited to the data using least-square method. Model can be represented using the following equation:

$\hat{Y}$ = 576.37244 – 0.00043264*X [3]

Model performed poorly, as relationship between variables is not linear (can be seen in Figure 1). Model produced R^2 value that was equal to 0 indicating non-linear relationship between variables. In addition, p-value analysis revealed that slope coefficient (β_1) was statistically insignificant at α=0.05 (Table 4).

Table 4. Estimated regression coefficients for linear 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 Regression Model fitted with log-transformed variables

Equation 2 was edited to fit log-transformed variables. Model can be represented using the following equation:

ln($\hat{Y}$) = 2.5549 – 0.49599*ln(X) [9]

Female model R^2 value was equal to 0.6076. MSE was equal to 2.34569. Slope coefficient was statistically significant at α=0.05 (Table 5).

Table 5. Estimated regression coefficients for log-transformed 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

Results from autocorrelation and homogeneity of variance tests for both models are represented in Table 6.

Table 6. Homogeneity of variance and autocorrelation tests for linear (left) and log-transformed (right) model

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

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

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

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

White test results revealed that residual variance is homogeneous in both model, since p-value produced by the test was higher than α=0.05 for both model. Therefore, null hypothesis (homogeneous σ^2) could not be rejected. Similarly, Durbin-Watson test produced results indicating that variables in both models are not autocorrelated, as p-values for testing positive and negative autocorrelations were higher than α=0.05.

Results from normality test are represented in Table 7 and Table 8.

Table 7. Departure from normality for linear (left) and log-transformed (right) model

Moments			
N	28	Sum Weights	28
Mean	0	Sum Observations	0
Std Deviation	1334.91015	Variance	1781985.11
Skewness	3.33535522	Kurtosis	10.6517854
Uncorrected SS	48113597.9	Corrected SS	48113597.9
Coeff Variation	.	Std Error Mean	252.274305

Moments			
N	28	Sum Weights	28
Mean	0	Sum Observations	0
Std Deviation	1.50293525	Variance	2.25881437
Skewness	-0.6774637	Kurtosis	0.11430048
Uncorrected SS	60.9879879	Corrected SS	60.9879879
Coeff Variation	.	Std Error Mean	0.28402806

Skewness results showed that both model distributions were skewed. For linear model, distribution was positively skewed (skewed to the right), where skewness value was equal to

3.34. Log-transformed model was skewed less, as its value was equal to -0.68. Log-transformed model distribution was slightly skewed to the left (skewness value was negative). Kurtosis values indicate that both models have distributions that are relatively peaked. Comparing the two kurtosis values, linear model distribution had a much bigger peak (kurtosis = 10.65) than log-transformed model (kurtosis=0.11), indicating that linear model is more departed from normality.

Table 8. Normality tests for linear (left) and log-transformed (right) model

Tests for Normality					Tests for Normality				
Test	Statistic		p Value		Test	Statistic		p Value	
Shapiro-Wilk	W	0.452258	Pr < W	<0.0001	Shapiro-Wilk	W	0.94954	Pr < W	0.1929
Kolmogorov-Smirnov	D	0.362106	Pr > D	<0.0100	Kolmogorov-Smirnov	D	0.116687	Pr > D	>0.1500
Cramer-von Mises	W-Sq	1.216081	Pr > W-Sq	<0.0050	Cramer-von Mises	W-Sq	0.074186	Pr > W-Sq	0.2411
Anderson-Darling	A-Sq	6.243732	Pr > A-Sq	<0.0050	Anderson-Darling	A-Sq	0.492211	Pr > A-Sq	0.2096

Normality tests results further confirmed the results represented in Table 7. All four tests that were used to test for model normality, revealed that linear model distribution is not normally distributed as p-values from all four tests were lower than α=0.05, rejecting the null hypothesis. On the other hand, log-transformed model normality test results indicated that distribution is normal (p-values > 0.05, failing to reject null hypothesis).

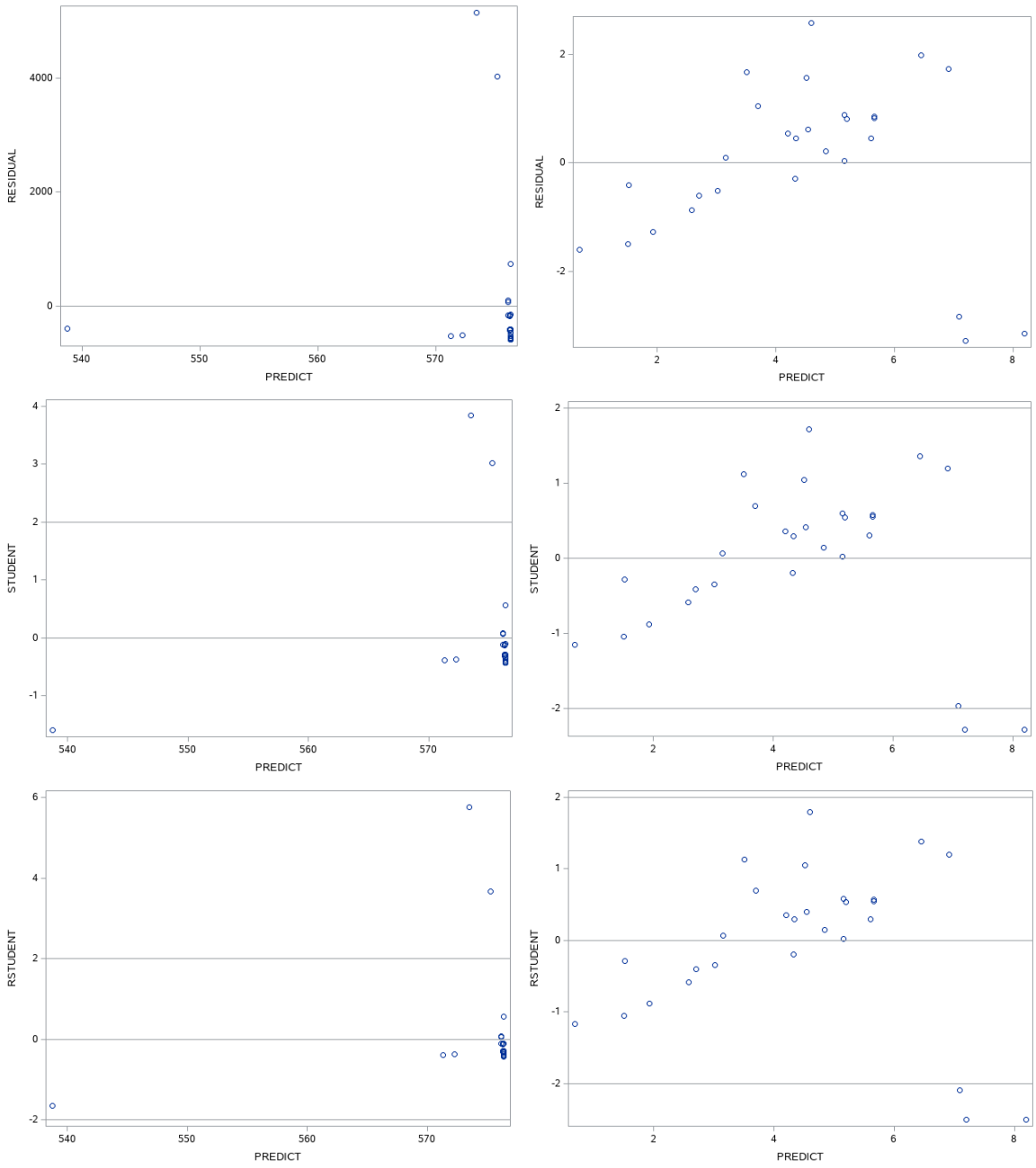
Studentized residuals and R-student values were computed and represented in Table 9.

Table 9. Studentized residuals and R-student values for linear (left) and log-transformed (right) model

Obs	Y	PREDICT	RESIDUAL	STDR	STUDENT	RSTUDENT	Obs	Y	PREDICT	RESIDUAL	STDR	STUDENT	RSTUDENT
1	8.1	576.372	-568.27	1334.10	-0.42596	-0.41915	1	8.1	2.70375	-0.61188	1.47929	-0.41363	-0.40694
2	423	576.171	-153.17	1334.45	-0.11478	-0.11258	2	423	5.60132	0.44606	1.49251	0.29886	0.29356
3	119.5	576.357	-456.86	1334.13	-0.34244	-0.33655	3	119.5	4.33683	0.44649	1.50390	0.29688	0.29161
4	115	576.360	-461.36	1334.12	-0.34582	-0.33988	4	115	4.20159	0.54334	1.50355	0.36137	0.35524
5	5.5	576.372	-570.87	1334.10	-0.42791	-0.42108	5	5.5	2.57435	-0.86960	1.47540	-0.58940	-0.58186
6	50	571.311	-521.31	1330.61	-0.39178	-0.38531	6	50	7.20105	-3.28903	1.43895	-2.28572	-2.50737
7	4603	575.271	4027.73	1335.54	3.01580	3.66746	7	4603	6.44482	1.98964	1.46991	1.35358	1.37669
8	419	576.291	-157.29	1334.24	-0.11789	-0.11563	8	419	5.14977	0.88811	1.49963	0.59222	0.58467
9	655	576.147	78.85	1334.49	0.05909	0.05794	9	655	5.65772	0.82692	1.49138	0.55447	0.54694
10	115	576.368	-461.37	1334.10	-0.34583	-0.33989	10	115	3.69697	1.04796	1.49958	0.69884	0.69180
11	25.6	576.371	-550.77	1334.10	-0.41284	-0.40616	11	25.6	3.14708	0.09552	1.49041	0.06409	0.06285
12	680	576.144	103.86	1334.50	0.07782	0.07632	12	680	5.66528	0.85682	1.49122	0.57458	0.56703
13	406	576.283	-170.28	1334.26	-0.12762	-0.12518	13	406	5.19990	0.80645	1.49901	0.53799	0.53051
14	1320	576.346	743.65	1334.15	0.55740	0.54987	14	1320	4.60193	2.58345	1.50371	1.71805	1.78932
15	5712	573.494	5138.51	1335.29	3.84822	5.75165	15	5712	6.92113	1.72920	1.45162	1.19122	1.20133
16	70	572.306	-502.31	1333.35	-0.37673	-0.37042	16	70	7.09249	-2.84399	1.44404	-1.96947	-2.09371
17	179	576.369	-397.37	1334.10	-0.29786	-0.29257	17	179	3.50568	1.68170	1.49697	1.12341	1.12934
18	56	576.357	-520.36	1334.12	-0.39004	-0.38359	18	56	4.31833	-0.29298	1.50387	-0.19482	-0.19117
19	1	576.372	-575.37	1334.10	-0.43128	-0.42443	19	1	1.50326	-1.50326	1.43172	-1.04997	-1.05213
20	0.4	576.372	-575.97	1334.10	-0.43173	-0.42487	20	0.4	0.68388	-1.60017	1.38355	-1.15657	-1.16446
21	12.1	576.371	-564.27	1334.10	-0.42296	-0.41618	21	12.1	3.00937	-0.51617	1.48732	-0.34705	-0.34110
22	175	576.348	-401.35	1334.14	-0.30083	-0.29550	22	175	4.54700	0.61778	1.50385	0.41080	0.40414
23	157	576.329	-419.33	1334.18	-0.31430	-0.30878	23	157	4.83904	0.21721	1.50255	0.14456	0.14181
24	440	576.350	-136.35	1334.14	-0.10220	-0.10024	24	440	4.51622	1.57056	1.50390	1.04432	1.04622
25	154.5	538.733	-384.23	240.19	-1.59973	-1.65208	25	154.5	8.19617	-3.15598	1.38158	-2.28433	-2.50546
26	1.9	576.372	-574.47	1334.10	-0.43061	-0.42376	26	1.9	1.92351	-1.28166	1.45136	-0.88308	-0.87921
27	3	576.372	-573.37	1334.10	-0.42978	-0.42294	27	3	1.51146	-0.41284	1.43213	-0.28827	-0.28313
28	180	576.289	-396.29	1334.25	-0.29701	-0.29174	28	180	5.16259	0.03037	1.49947	0.02025	0.01986

Outlier residual analysis is illustrated in Figure 3.

Figure 3. Residual analysis for linear (left) and log-transformed (right) models



Values outside -2 – 2 range were identified as outliers, since it was chosen to cover 95% of the observation with $\overline{X} \pm 2 \cdot S$ empirical rule. Studentized residual test revealed that there are 2 outlier in the linear model and 2 outliers in the log-transformed model. R-student test results indicated that there were 2 outliers in the linear model and 3 outliers in the log-transformed model.

4. Summary

Relationship between brain weight and body weight was analyzed comparing two linear regression models. First, linear regression model was created using non-transformed variables. Then, natural log-transformed regression model was created using log-transformed X and Y variables. Results revealed that log-transformed model produced much higher R^2 value of 0.6076 as compared to linear model. Although both models passed autocorrelation and residual homogeneity tests, only log-transformed model was proven to have normal distribution. Both model had two outliers as was identified by studentized residual test.

Further analysis needs to be performed on log-transformed model as its R^2 square value could be further improved. In order to determine more generalized trend, more data needs to be collected, especially from species that have higher body weight.

5. SAS programs

```
*HW5* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW5/Brain.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW5DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  SET HW5DATA;
RUN;
*Pearson and Spearman correlation coefficients;
PROC CORR PEARSON SPEARMAN;
  VAR Y X;
RUN;
*PROC PLOT;
*  PLOT Y*X='*';
*RUN;
ODS GRAPHICS ON; */ ATTRPRIORITY=NONE;
PROC SGPLOT;
  SCATTER X=X Y=Y; */ DATALABEL=SPCS;
  *TITLE 'Relationship between brain and body weight';
RUN;
DATA ALL;
  SET HW5DATA;
  LNY=LOG(Y);
  LNX=LOG(X);
RUN;
PROC CORR PEARSON SPEARMAN;
  VAR LNY LNX;
RUN;
PROC SGPLOT;
  SCATTER X=LNX Y=LNY / DATALABEL=SPCS;
  *TITLE 'Transformed relationship between brain and body weight';
RUN;
*Y VS X MODEL;
PROC REG DATA=ALL;
  MODEL Y=X / DWPROB SPEC;
  OUTPUT OUT=OUT1 P=PREDICT R=RESIDUAL STDR =STDR STUDENT=STUDENT RSTUDENT=RSTUDENT;
RUN;
PROC SGPLOT DATA=OUT1;
  SCATTER X=PREDICT Y=STUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
```

```
PROC SGPLOT DATA=OUT1;
  SCATTER X=PREDICT Y=RSTUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
PROC SGPLOT DATA=OUT1;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC PRINT DATA=OUT1;
  VAR Y PREDICT RESIDUAL STDR STUDENT RSTUDENT;
RUN;
PROC UNIVARIATE DATA=OUT1 PLOT NORMAL;
  VAR RESIDUAL;
RUN;
*LNY VS LNX MODEL;
PROC REG DATA=ALL;
  MODEL LNY=LNX / DWPROB SPEC;
  OUTPUT OUT=OUT2 P=PREDICT R=RESIDUAL STDR =STDR STUDENT=STUDENT RSTUDENT=RSTUDENT;
RUN;
PROC SGPLOT DATA=OUT2;
  SCATTER X=PREDICT Y=STUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
PROC SGPLOT DATA=OUT2;
  SCATTER X=PREDICT Y=RSTUDENT;
  REFLINE 0; REFLINE 2; REFLINE -2;
RUN;
PROC SGPLOT DATA=OUT2;
  SCATTER X=PREDICT Y=RESIDUAL;
  REFLINE 0;
RUN;
PROC PRINT DATA=OUT2;
  VAR Y PREDICT RESIDUAL STDR STUDENT RSTUDENT;
RUN;
PROC UNIVARIATE DATA=OUT2 PLOT NORMAL;
  VAR RESIDUAL;
RUN;
ODS GRAPHICS OFF;
```