JMP® ACADEMIC CASE STUDY

JMP022: Archosaurs - The Relationship Between Body Size and Brain Size

Distribution, Regression, Transformation

Produced by:

Dr. DeWayne Derryberry
Idaho State University, Department of Mathematics

STATISTICAL DISCOVERY FROM SAS

Archosaurs: The Relationship Between Brain and Body Size

Distribution, Regression, Transformation

Key ideas

Distribution, Power Law Models, Logarithmic Transformations, Regression.

Background

Reptiles of the subclass Archosauria include the now extinct dinosaurs and pterosaurs (flying or gliding dinosaurs) as well as birds and crocodiles. For any subclass of animals, although individual species vary greatly in size, it has been found that there is a relatively consistent relationship between body size and brain size.

In "Relative Brain Size and Behavior of Archosaurian Reptiles" (Annual Review of Ecology and Systematics 1977, by James A. Hopson) the author cites a widely observed power law relationship:

$$Brain\ wgt = k \times (Body\ wgt)^{2/3}$$

Using logarithms, this can be transformed into a linear relationship:

$$Log(Brain\ wgt) = Log(k) + 0.67 \times Log(Body\ wgt)$$

It should be noted we would expect "smart" animals, such as the mammals, to differ from "dumb" animals, such as the reptiles, by having a linear relationship with a larger intercept. In other words, "*k*" would be larger for mammals than reptiles. Also, it does not matter which logarithmic transformation we use (common logs, natural logs or any others), but we will use the natural log.

Power law models are based on the relationship $y = ax^b$, which can be restated as a linear relationship $\log(y) = \log(a) + b \times \log(x)$ using a logarithmic transformation. Power law models represent scale invariant relationships, that is, relationships thought to hold in the same manner for very large and very small observed values. For example, in biology the relationship between habitat area and number of species fits a power law model quite well (see Case 1 and Exercise 22 from Chapter 8 of *The Statistical Sleuth, Second Edition*, Ramsey and Schafer, 2002). A Web search on power law models will result in dozens of examples, taken mostly from physics and engineering.

This is a typical example of a power law model. Although reptiles vary greatly in size, we would expect about the same *proportion* of their mass to be devoted to brain size. We expect the same for mammals, except we expect the brains of mammals to be bigger than the reptile of the same overall mass, hence the greater value for "*k*".

The Task

Determine whether a power law model fits the data. In other words:

- Does it make sense to take logarithmic transformations of body weight and brain weight in order to perform a linear regression?
- Are the assumptions of the linear regression met?
- Is our final model a good fit to the data?
- Does the slope of 2/3 actually seem to match the data?

Other questions of interest:

- How do we make predictions in a power law model?
- What do large positive and negative residuals mean in this context?

The Data Archosaurs.jmp

Using fossil records and numerous sources, the author was able to compile estimates of the body size and brain size for several Archosaurs (Exhibit 1).

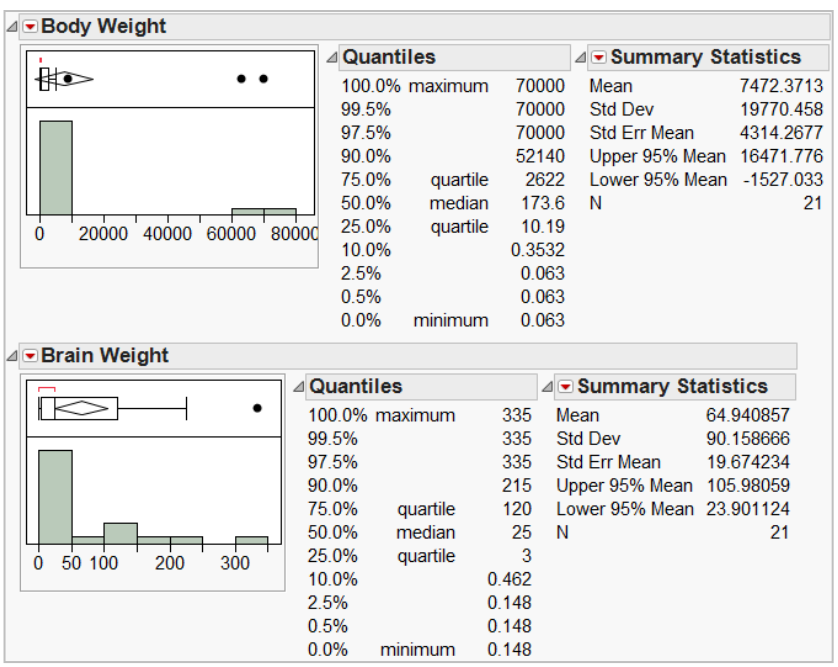
Type	The type of Archosaur.
Details	The specific name for that species.
Body Weight	Estimated body weight in kilograms.
Brain Weight	Estimated brain weight in grams.

It should be noted that there is quite a bit of error in these estimates. First, the author of the original article needed to use estimates based on fossils for most of the now extinct species. Second, the numbers themselves were estimated from a figure in that article.

Analysis

We start by exploring the distributions of Body Weight and Brain Weight. Exhibit 1 shows histograms and summary statistics for the two measures.

Exhibit 1 Distributions of Body Weight and Brain Weight



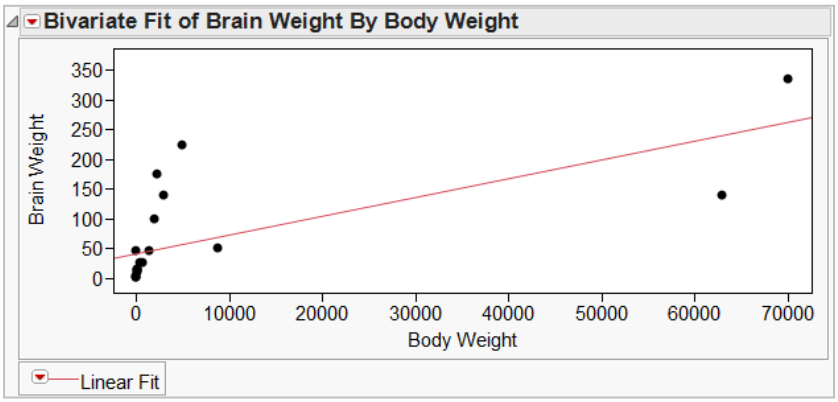
(Analyze > Distribution; for a horizontal layout select Stack under the top red triangle.)

The histograms of both brain weight and body weight are right-skewed. Since we are ultimately interested in modeling the relationship between these two variables, we might consider a transformation to make the distributions more normal. The logarithmic transformation often works well when data are right-skewed.

A second indication of when the logarithmic transformation makes sense to normalize data is when all the values are positive and cover several orders of magnitude (*The Statistical Sleuth, Second Edition*, Page 68). For example, the ratio of 75th to 25th quartile is quite revealing. For **Body Weight** this ratio is 2622/10.19 = 257.3, while this same ratio is 120/3 = 40 for **Brain Weight**. (Note: Comparing ratios to judge the span of orders of magnitude only makes sense for data restricted to be positive.)

It is obvious from Exhibit 2 that linear regression involving body weight and brain weight is not useful – a straight line does not fit the points very well.

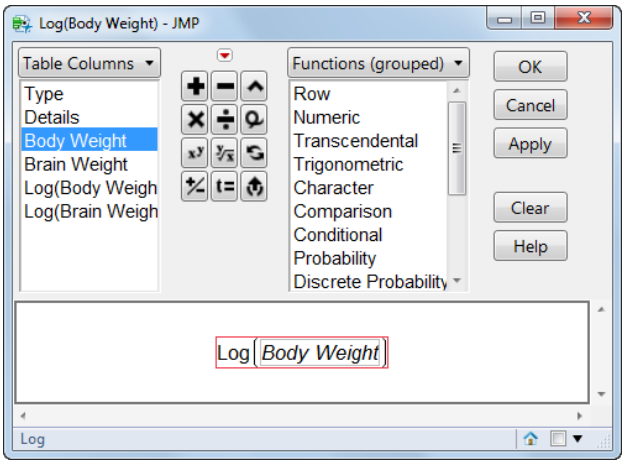
Exhibit 2 Regression with Body Weight and Brain Weight, Prior to Transformations



(Analyze > Fit Y by X; Use Brain Weight as Y, Response and Body Weight as X, Factor. Then select Fit Line from the top red triangle.)

To transform the two variables, we create two new columns, and apply a log transformation using the Formula Editor (Exhibit 3).

Exhibit 3 Transforming Body Weight



(Create a new column in the data table, and rename it Log(Body Weight). Right click on the column header, and select Formula to open the Formula Editor. To create the formula:

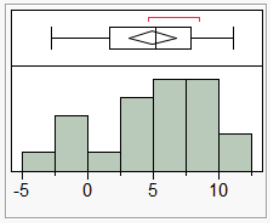
1. Click on Transcendental in the Functions list, and select Log.
2. Select Body Weight from the columns list.
3. Click OK.

Note: To create this transformed variable directly from the data table, right click on the variable and select New Formula Column > Transform > Log. Repeat for Log(Brain Weight).)

We now see that the extreme right-skewness has been removed from the variables (Exhibit 4). In addition, all observed values are roughly the same order of magnitude.

Exhibit 4 Distributions of Log(Body Weight) and Log(Brain Weight)

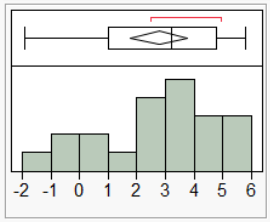
Log(Body Weight)



Quantiles		
100.0%	maximum	11.1563
99.5%		11.1563
97.5%		11.1563
90.0%		10.6549
75.0%	quartile	7.86074
50.0%	median	5.15675
25.0%	quartile	1.63326
10.0%		-1.0461
2.5%		-2.7646
0.5%		-2.7646
0.0%	minimum	-2.7646

Summary Statistics	
Mean	4.9463973
Std Dev	4.0443549
Std Err Mean	0.8825506
Upper 95% Mean	6.7873656
Lower 95% Mean	3.105429
N	21

Log(Brain Weight)



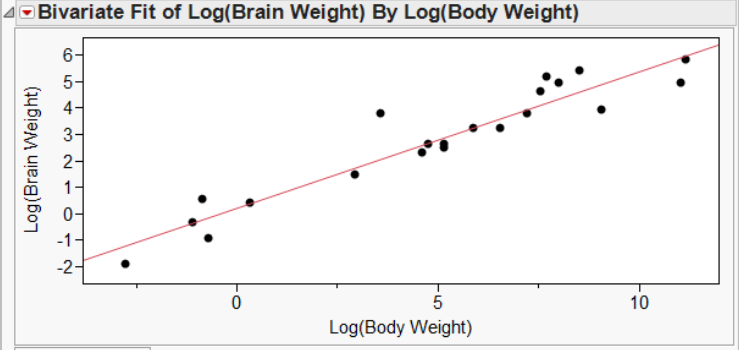
Quantiles		
100.0%	maximum	5.81413
99.5%		5.81413
97.5%		5.81413
90.0%		5.36584
75.0%	quartile	4.77341
50.0%	median	3.21888
25.0%	quartile	0.99462
10.0%		-0.8015
2.5%		-1.9105
0.5%		-1.9105
0.0%	minimum	-1.9105

Summary Statistics	
Mean	2.7684553
Std Dev	2.1966267
Std Err Mean	0.4793432
Upper 95% Mean	3.7683478
Lower 95% Mean	1.7685629
N	21

Now we fit a model using the transformed variables instead the original variables. A linear model seems to fit the transformed data well (Exhibit 5).

Exhibit 5 Regression with Log(Body Weight) and Log(Brain Weight), After Transformations

Bivariate Fit of Log(Brain Weight) By Log(Body Weight)



(Analyze > Fit Y by X; Use Log(Brain Weight) as Y, Response and Log(Body Weight) as X, Factor. Then select Fit Line from the top red triangle.)

The linear regression involving Log(Body Weight) and Log(Brain Weight) looks reasonable, but a more careful assesment of the model requires evaluating four key assumptions. This may lead us to discard the model as unacceptable, adopt the model with reservations, or adopt the model with confidence.

The assumptions we are concerned about are:

1. Observations are independent.
2. The distribution of the observations around the line is normal.
3. The spread is equal at all levels of the predictor.
4. The line fits the data.

1. Observations Are Independent

Observations are most often not independent when there are serial or clustering effects (*The Statistical Sleuth*, 62-63). Serial effects occur when observations are measured close in time or space. Clustering effects occur when two or more observed values are likley to be more alike than two randomly chosen observations. For example, two mice from the same litter would be expected to be more alike than two mice chosen at random. Given the way the data ere collected, neither of these effects seems relevant here. We'll assume that the measurements are independent.

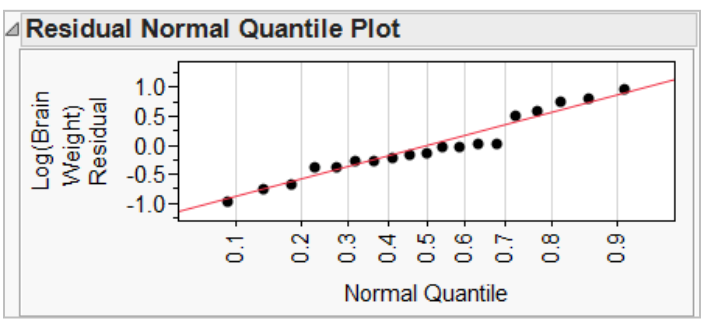
If the independence assumption were violated, this would usually be a very big problem. For example, while point estimates may not be greatly affected, hypothesis tests may produce misleading p-values, and confidence intervals for the estimates may not be the correct width (see Chapter 15 of *The Statistical Sleuth*). The impact can be quite large and in either direction (for example, confidence intervals can be too wide or too narrow).

2. The Distribution of the Observations Around the Line is Normal

We hope that the true errors (deviations from the true model) are normal. But we can't measure true errors – the residuals (deviations from the fitted model) are the closest thing we have. So we check to see if the residuals are normally distributed using a normal quantile plot.

All the residuals fall close to a straight line (Exhibit 6), indicating that the residuals are approximately normal.

Exhibit 6 Residual Normal Quantile Plot

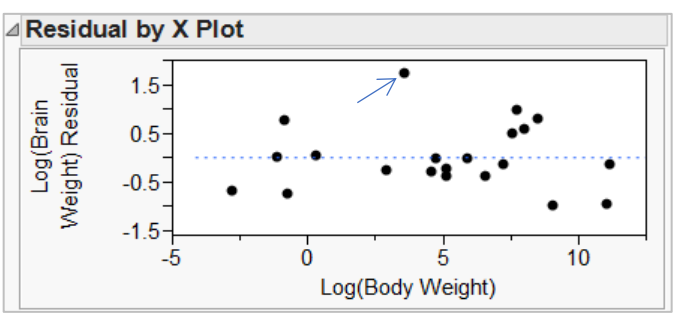


(Click on the red triangle next to Linear Fit (at the bottom of the regression plot shown in Exhibit 5), and select Plot Residuals to produce a series of residual plots.)

3. The Spread Is Equal at All Levels of the Predictor

A Residual by X Plot reveals how the residuals vary across levels of the predictor (X) variable. We would like to see the residuals falling completely at random, with points spread equally above and below the center line and with no obvious pattern. In Exhibit 7, there appears to be somewhat of a gap near the middle of the residual plot. This may be due to one observation, a mild outlier, in the middle of the plot. However, most of the points do indeed appear randomly scattered across the levels of the X variable. Although the residuals are not perfect, they are acceptable. (Note: We'll talk more about how to interpret residuals later.)

Exhibit 7 Residual by X Plot



4. The Line Fits the Data

The usual test for a linear model is:

Ho: a single mean fits the data
Ha: a line fits the data

For this test, we want a small p-value (and a large F statistic) as evidence that the line is modeling something.

A lack of fit test, on the other hand, tests:

Ho: a line fits the data
Ha: a more complex model fits the data

The nature of the "more complex" model would vary from circumstance to circumstance. For example, if there are replicated values (repeated X values), the more complex model may involve fitting a separate mean to each level of the replicated X values. When this is not possible, the more complex model might include a higher order polynomial (a curve). Or the complexity may lie in other variables that were not used in the model.

In any case, if a line is a reasonable model, the lack of fit test will produce a relatively large p-value (typically > 0.05 is considered sufficiently large). In our case, with a p-value of 0.1187 (Exhibit 8), we have little evidence that a line is not sufficient to fit the data. (Note: Sorry for the double negative language, but we cannot say we have evidence that the line fits the data – that would be affirming the null hypothesis. Furthermore, the p-value is not that big, so we don't want to fall in love with the line.)

Exhibit 8 Lack of Fit Test

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	18	9.3183358	0.517685	43.5717
Pure Error	1	0.0118812	0.011881	**Prob > F**
Total Error	19	9.3302170		0.1187

(Note: The Lack Of Fit results display by default only if the data set has replicated X values.)

Given what we have seen so far, a linear model is reasonable, although not perfect. A power law model does seem to fit the data.

In the article we have cited, the author claims, but does not test, the notion that the slope of the power law regression line should be 2/3. Our estimated slope is 0.5162 (Exhibit 9). Is this close to 2/3 or far from 2/3, based on the variation in the data?

Exhibit 9 Parameter Estimates

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.2150713	0.245176	0.88	0.3913
Log(Body Weight)	0.5162109	0.038744	13.32	<.0001*

It is easy to test this claim with a simple hypothesis test:

Ho: the slope is 2/3
Ha: the slope is different from 2/3

$$t = \frac{0.5162 - 0.6667}{0.038744} = -3.88!!$$

Yikes! The data does not support this hypothesis. In fact, a 95% confidence interval (with 21 – 2 = 19 degrees of freedom) is:

$$0.5162 \pm 2.093 \times 0.03874 = (0.435, 0.5973)$$

The entire confidence interval (shown again in Exhibit 10) is well below 2/3.

Exhibit 10 Parameter Estimates with 95% Confidence Intervals

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	0.2150713	0.245176	0.88	0.3913	-0.298088	0.7282302
Log(Body Weight)	0.5162109	0.038744	13.32	<.0001*	0.4351187	0.5973031

(Right-click over the Parameter Estimates table and select Columns > Lower 95% and Columns > Upper 95%.)

It would be interesting to examine other data sets claiming this value to see if a slope of 2/3 is reasonable for those data sets.

Other Questions of Interest

1. Predictions in the Power Law Model

Prediction in a power law model is similar to other models, except that we must take the logarithmic transformations into account. Suppose we want to estimate the average brain weight (in grams) of another species of archosaur with a body weight of about 20 kilograms.

For our data, the model is:

$$\text{Log}(Brain\ Weight) = 0.2151 + 0.5162 \times \text{Log}\ (Body\ weight)$$

We begin with:

$$\text{Log}(Brain\ Weight) = 0.2151 + 0.5162 \times \text{Log}(20) = 1.7615$$

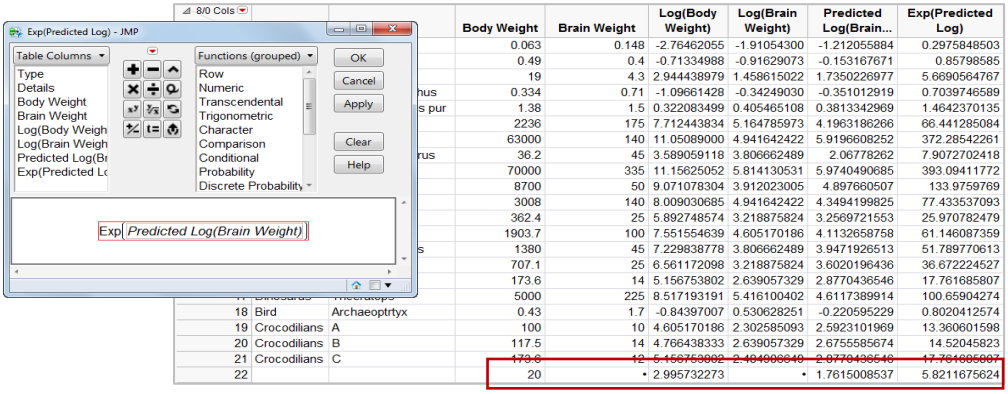
This is NOT a very useful answer! We want to know the brain weight, not the logarithm of the brain weight.

To find the predicted brain weight, we must use the inverse transformation. For the natural logarithmic function, the inverse is the exponential function:

$\exp(1.7615) = 5.82\ grams =$ predicted brain weight (don't forget the units).

These results are shown in Exhibit 11.

Exhibit 11 Predicting **Brain Weight**



			Body Weight	Brain Weight	Log(Body Weight)	Log(Brain Weight)	Predicted Log(Brain...	Exp(Predicted Log)
			0.063	0.148	-2.76462055	-1.91054300	-1.212055884	0.2975848503
			0.49	0.4	-0.71334988	-0.91629073	-0.153167671	0.85798585
			19	4.3	2.944438979	1.458615022	1.7350226977	5.6690564767
		hus	0.334	0.71	-1.09661428	-0.34249030	-0.351012919	0.7039746589
		s pur	1.38	1.5	0.322083499	0.405465108	0.3813342969	1.4642370135
			2236	175	7.712443834	5.164785973	4.1963186266	66.441285084
			63000	140	11.05089000	4.941642422	5.9196608252	372.28542261
		rus	36.2	45	3.589059118	3.806662489	2.06778262	7.9072702418
			70000	335	11.15625052	5.814130531	5.9740490685	393.09411772
			8700	50	9.071078304	3.912023005	4.897660507	133.9759769
			3008	140	8.009030685	4.941642422	4.3494199825	77.433537093
			362.4	25	5.892748574	3.218875824	3.2569721553	25.970782479
			1903.7	100	7.551554639	4.605170186	4.1132658758	61.146087359
		s	1380	45	7.229838778	3.806662489	3.9471926513	51.789770613
			707.1	25	6.561172098	3.218875824	3.6020196436	36.672224527
			173.6	14	5.156753802	2.639057329	2.8770436546	17.761685807
			5000	225	8.517193191	5.416100402	4.6117389914	100.65904274
18	Bird	Archaeoptrtyx	0.43	1.7	-0.84397007	0.530628251	-0.220595229	0.8020412574
19	Crocodilians	A	100	10	4.605170186	2.302585093	2.5923101969	13.360601598
20	Crocodilians	B	117.5	14	4.766438333	2.639057329	2.6755585674	14.52045823
21	Crocodilians	C	173.6	12	5.156753802	2.484906649	2.8770436546	17.761685807
22			20	•	2.995732273	•	1.7615008537	5.8211675624

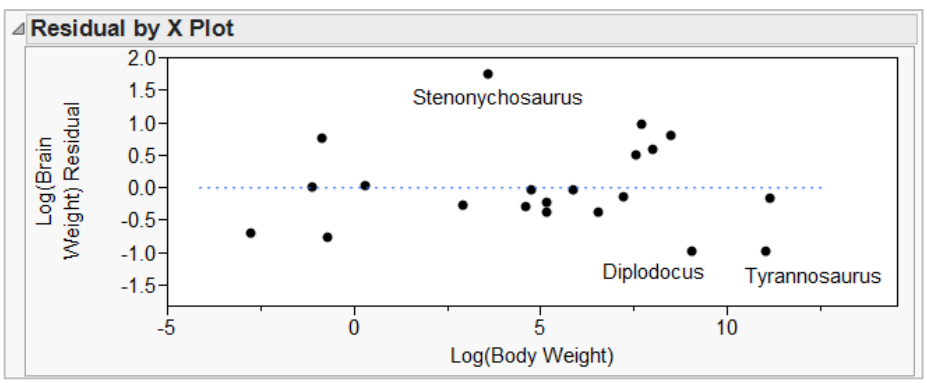
(In Fit Y by X, click on the red triangle for the Linear Fit and select Save Predicteds to save the fitted model to the data table. Then create a new column in the data table and enter the formula editor as shown earlier. Select Exp from the Transcendental Functions list, and then select Predicted Log(Brain Weight), and click OK. Finally, click to add a new row to the data table, and enter 20 in the Body Weight column.)

2. What Do Positive and Negative Residuals Mean?

In this context, a positive residual is a species whose brain is larger than would be predicted based on the model. A negative residual is a species whose brain is smaller than would be predicted by the model. We could speculate that the largest positive residuals might belong to some of the smartest species of archosaurs, and the negative residuals belong to some of the dumber species.

If this is the case, then the Stenonychosaurus is clearly the smartest archosaur (of this group), and the Diplodocus and Tyrannosaurus are the least brainy (see Exhibit 12).

Exhibit 12 Residual by X Plot with Labels



(To label points with the row number, select the points on the graph or in the data table. Then select Rows > Label. To permanently label points with the value of a column, right-click on the column name in the data table and select Label. To reposition labels on a graph, click and drag the label.)

In the exercise we will look at mammals. Any idea which species the model will indicate is the smartest?

Summary

Statistical Insights

A power law model did fit the data reasonably well. However, the slope suggested by the researcher does not seem correct based on this small data set.

Consider a logarithmic transformation for modeling involving right-skewed data. But keep in mind that the predicted response values will be in transformed units. The inverse transformation must be taken in order to produce predicted values in original units.

Always evaluate the quality of a model using residual plots and the lack of fit test before performing statistical inference. If model assumptions are violated, the model fit is not appropriate. For example, if there is curvature in the residuals, a quadratic model may be required.

Implications

A word about building consensus in science: It is the ability of the power law to explain a lot of similar data sets that gives it its power. Seeing this approach produce good results in many similar situations – replication – is how the model gains credibility. A similar analysis involving the brain and body weights of mammals is widely available (*The Practice of Statistics in the Life Sciences, Second Edition*, 2012, Moore and Baldi, Example 4.2; *The Statistical Sleuth, Second Edition*, 2002, Ramsey and Schafer, Case 9.2) and will be the basis for the exercises to follow.

JMP® Features and Hints

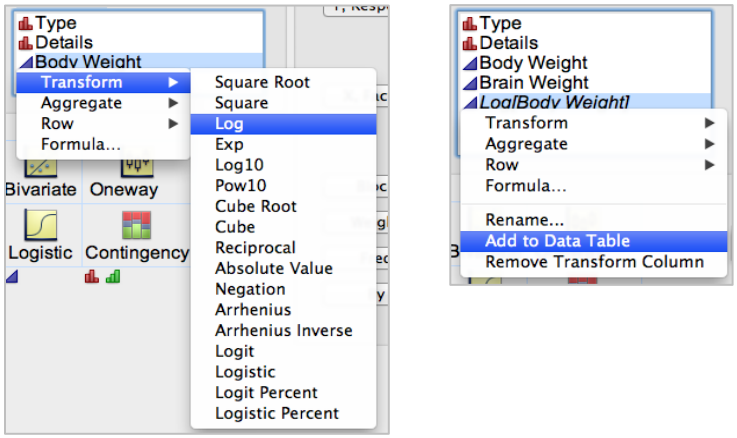
This case used the Distribution platform to display histograms and summary statistics and Fit Y by X to model the relationship between two variables. In Fit Y by X, when a line is fit a variety of statistical output is produced, including Parameter Estimates, the Summary of Fit, and an ANOVA table. If the data table has replicated predictor values, lack of fit test results will also be provided. To explore residuals, select “Plot Residuals” from the red triangle for the fitted model.

The formula editor was used to transform the variables. A number of transformations are available from the formula editor. In addition, a utility for transforming multiple variables is available in the Sample Data directory under the Help Menu (Open the Teaching Demonstrations outline under Teaching Resources (click on the gray icon), and select utilColumnTransformation.)

Note that in Fit Y by X, transformations are also available from the Fit Special option under the top red triangle. This bypasses the need to transform the variables prior to the analysis, and has the added advantage of automatically transforming saved predicted values to the original units.

Also note that as of JMP 11, variables can be dynamically transformed in any dialog window. Right-click on the variable in the column selection panel, select Transform and then select the transformation of interest (shown on the left in Exhibit 13). This creates a temporary variable. To save the transformed data to the data table, right-click on the transformed variable (shown in italics), and select Add to Data Table (on the right in Exhibit 13).

Exhibit 13 Dynamic Transformation of Variables



Exercises

From The Statistical Sleuth

Another example of a power law model involves 96 species of mammals (including humans). This also involves body weight in kilograms and brain weight in grams.

The data are available in Archosaur Exercise.jmp.

1. Does it make sense to take logarithmic transformations of **Body Weight** and **Brain Weight** in order to perform a linear regression?
2. Are the assumptions of the linear regression met?
3. Is our final model a good fit to the data?
4. Does the slope of 2/3 actually seem to match the data?
5. (Challenging) Do mammals appear to have generally larger brains than archosaurs, given their body size? In other words, is the entire line for the mammals higher than the entire line for the archosaurs?
6. If a species of mammals has a body size of about 20 kilograms, what would you estimate the brain weight to be in grams?
7. Based on the residuals, what would you expect the four smartest mammal species to be? Does this fit your intuition?

jmp STATISTICAL DISCOVERY FROM SAS

jmp.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2021, SAS Institute Inc. All rights reserved.