Linear Regression

Where 2 variables have a functional relationship

1

Linear regression

- Regression, in its simplest form, is the analysis of the functional relationship where one variable is dependent on or related to another.
  - i.e: the relationship between body weight and height
  - The analysis can be used to test hypotheses about the relationship between the variables and to make predictions about unknown outcomes.

2

- Linear regression is the simplest form of regression. It looks at the linear relationship between two variables

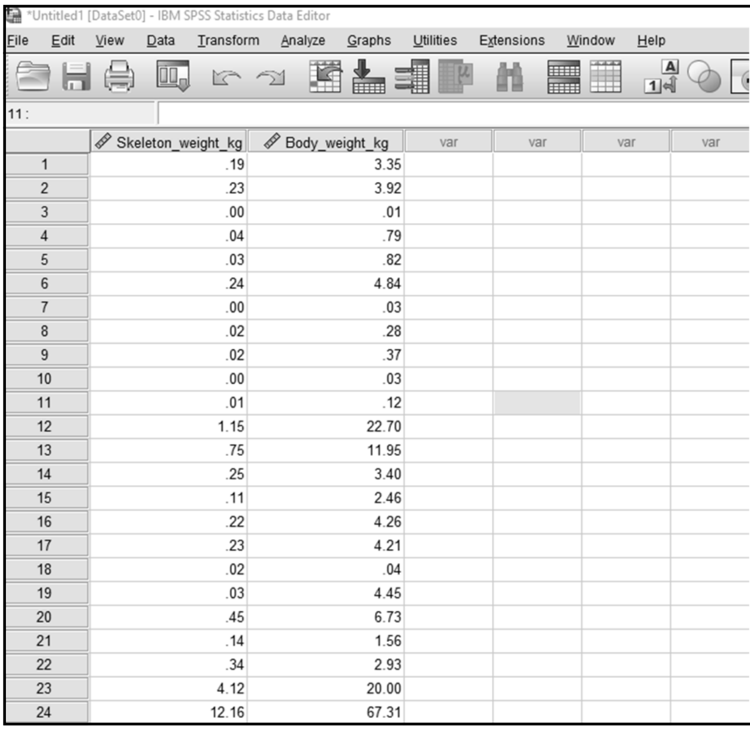
- The main assumption of linear regression is the fact that the relationship between the two variables is in fact linear (straight).
    - If the relationship is not linear it may be necessary to transform the data in one or both of the variables.

- Helps for the data to be normally distributed

3

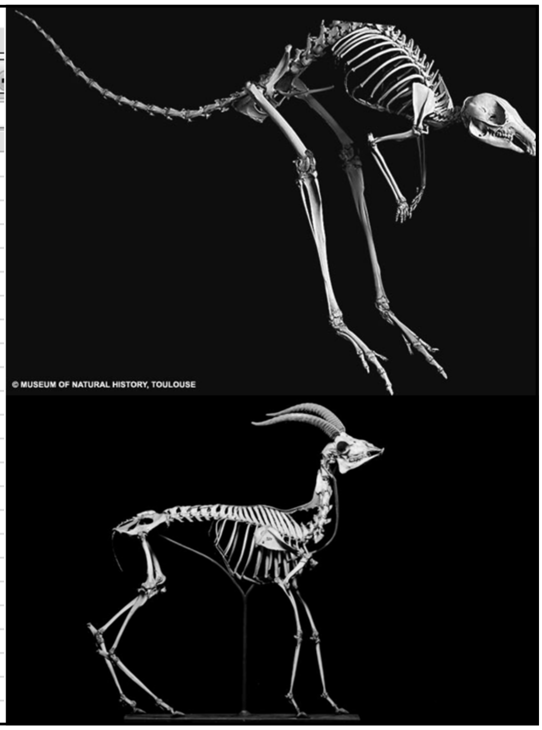
Regression example

- *In a study examining the relationship between skeletal weight and body weight of mammals a scientist wanted to determine if it was possible to predict skeletal weights from body weight data.*
- 24 different mammal species were examined in this study to produce a bi-variate data set (two variables for each individual)

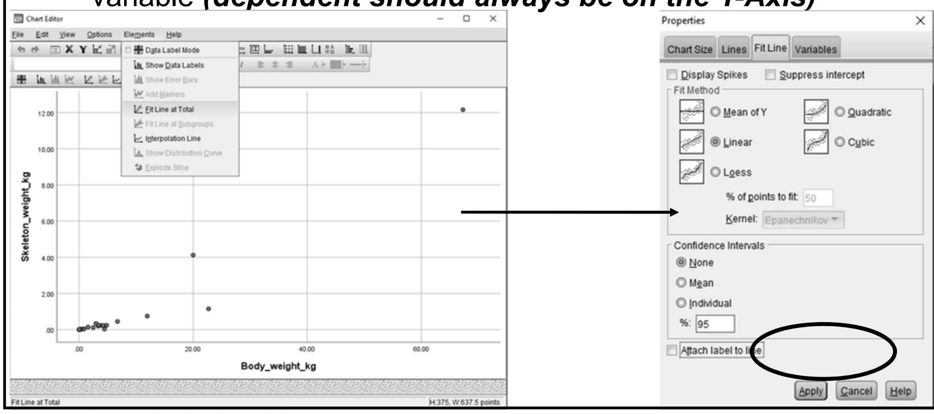


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

	Skeleton_weight_kg	Body_weight_kg	var	var	var	var
1	.19	3.35				
2	.23	3.92				
3	.00	.01				
4	.04	.79				
5	.03	.82				
6	.24	4.84				
7	.00	.03				
8	.02	.28				
9	.02	.37				
10	.00	.03				
11	.01	.12				
12	1.15	22.70				
13	.75	11.95				
14	.25	3.40				
15	.11	2.46				
16	.22	4.26				
17	.23	4.21				
18	.02	.04				
19	.03	4.45				
20	.45	6.73				
21	.14	1.56				
22	.34	2.93				
23	4.12	20.00				
24	12.16	67.31				

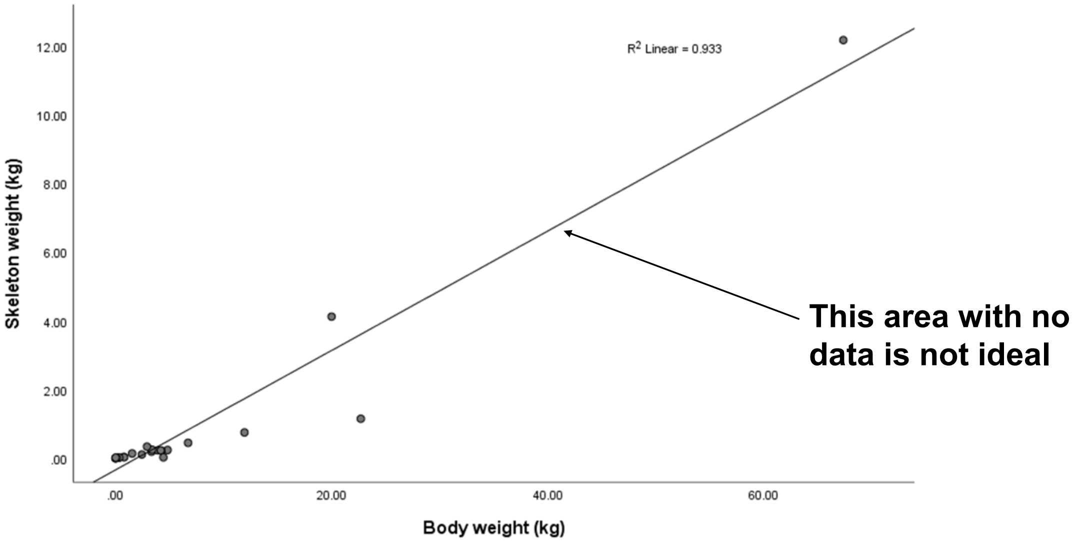


- The first step of any regression analysis should be to produce a scatter plot of the data (Called ‘eyeballing the data’).
- Production of a scatter plot allows you to determine if the data is reasonably linear.
  - We want to know if skeletal weight is dependent on body weight
  - Therefore, skeletal weight is called the dependent variable ***(dependent should always be on the Y-Axis)***



6

- This fit is reasonably linear (straight). It is not ideal, due to the uneven spread of the data (points).



7

- Now test the variables for normality
  - In this case neither variable is normal
  - P.S. Both variables go in the dependent section for normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Skeleton_weight_kg	.398	24	.000	.365	24	.000
Body_weight_kg	.351	24	.000	.500	24	.000

a. Lilliefors Significance Correction

8

Descriptives

			Statistic	Std. Error
Skeleton_weight_kg	Mean		.8647	.52032
	95% Confidence Interval for Mean	Lower Bound	-.2117	
		Upper Bound	1.9411	
	5% Trimmed Mean		.3596	
	Median		.1650	
	Variance		6.498	
	Std. Deviation		2.54904	
	Minimum		.00	
	Maximum		12.16	
	Range		12.16	
	Interquartile Range		.30	
	Skewness		4.201	.472
	Kurtosis		18.495	.918
Body_weight_kg	Mean		6.9400	2.89112
	95% Confidence Interval for Mean	Lower Bound	.9593	
		Upper Bound	12.9207	
	5% Trimmed Mean		4.3840	
	Median		3.1400	
	Variance		200.606	
	Std. Deviation		14.16355	
	Minimum		.01	
	Maximum		67.31	
	Range		67.30	
	Interquartile Range		4.44	
	Skewness		3.725	.472
	Kurtosis		15.358	.918

- Variances are greater than the mean, therefore log transform both variables.
- ***Note*** you can use one variable logged and another untransformed

9

- Data is now much closer to normal
  - Not perfect but considerably closer
  - Remember to check that data is still relatively linear

Tests of Normality

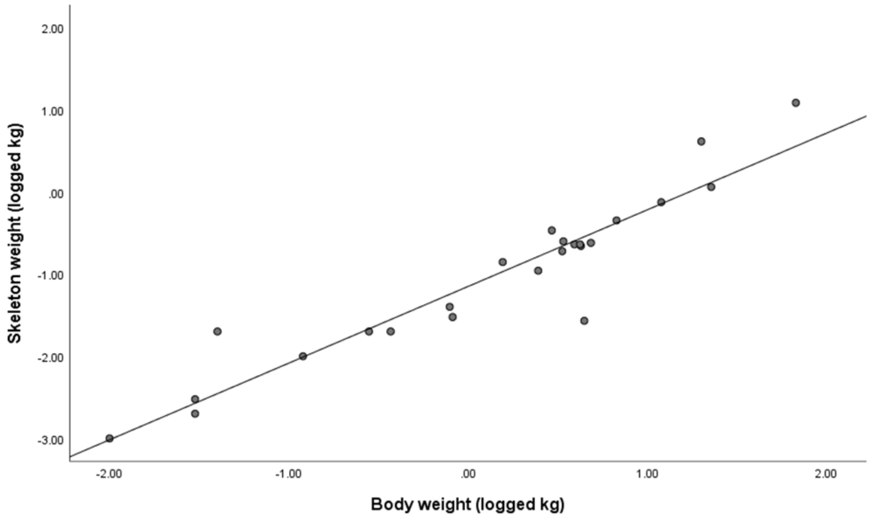
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Log_skeleton_kg	.121	24	.200*	.974	24	.768
Log_body_kg	.185	24	.032	.932	24	.110

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

10

- Relationship is still linear, but the data is now spread more evenly across the line.



11

Running a regression

- Having decided that your data is reasonably linear, and normal you can run the regression analysis.
- **Step one:** *Select the dependent and independent variables*
  - The dependent variable is usually the variable that is going to be predicted from the outcome of the analysis. It is in some way dependent on the other variable.

12

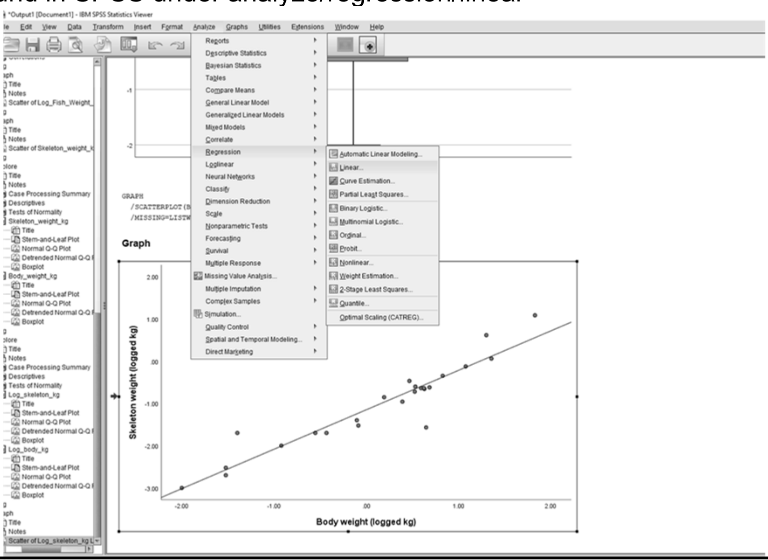
- In this case we want to predict skeleton mass from body weight data. Skeleton mass is completely dependent on body mass (i.e. you can not have a skeleton weight without having a body weight).

- <u>Note:</u> When producing scatter plots of this kind of data the <u>independent variable</u> should <u>always</u> be on the <u>x-axis</u>.

13

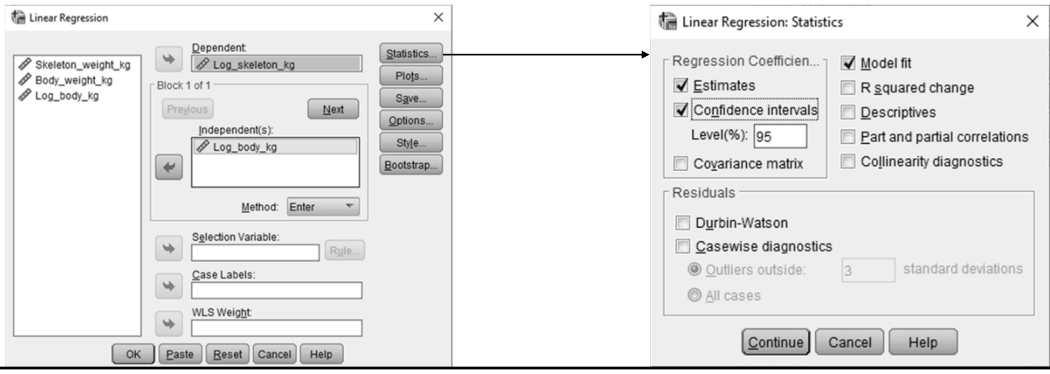
- **Step two:** *Select the test you are going to use*

- In this case we are using a linear regression analysis.
  - This can be found in SPSS under analyze/regression/linear regression.



14

- **Step 3**: *Insert all the information needed for the analysis*
- Dependent variable:- Skeleton mass (logged)
- Independent variable:- Body mass (logged)
- Click the statistics button:-
  - ♦ Select estimates
  - ♦ Select confidence intervals
- This will allow us to calculate predicted outcomes.



15

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.944[a]	.890	.885	.33673

a. Predictors: (Constant), Log_body_kg

Indicates the strength of the relationship between the dependent and the independent variable. In this case the regression model accounts for 88.5% of the variation in the data.

16

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.250	1	20.250	178.592	.000[b]
	Residual	2.495	22	.113		
	Total	22.745	23			

a. Dependent Variable: Log_skeleton_kg

b. Predictors: (Constant), Log_body_kg

Indicates whether the relationship is significant

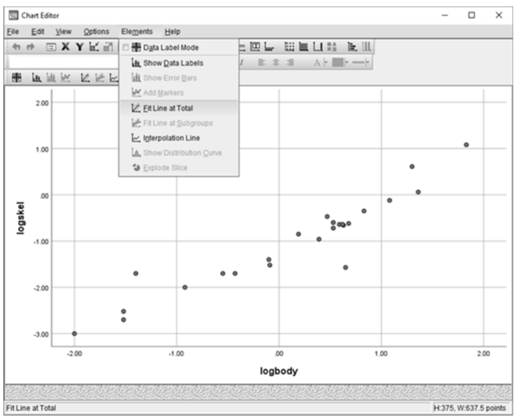
H_0:- Skeleton weight in mammals is not dependent on body weight

Skeleton weight in mammals is dependent on body weight (R^2=0.885, $F_{(1,22)}$=178.592, p<0.001)

17

Producing a regression plot

- *Select a scatter plot – click on simple*
  - x-axis should be your independent variable (body weight)
  - y-axis dependent variable
- *Produce the plot*

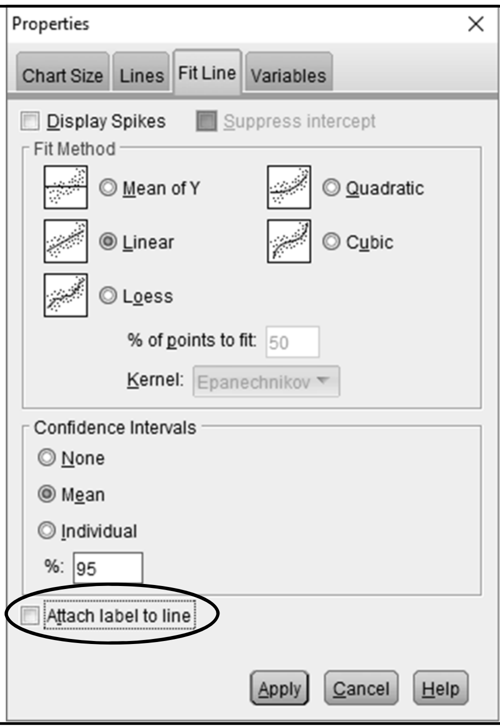


Click on the graph to edit it.

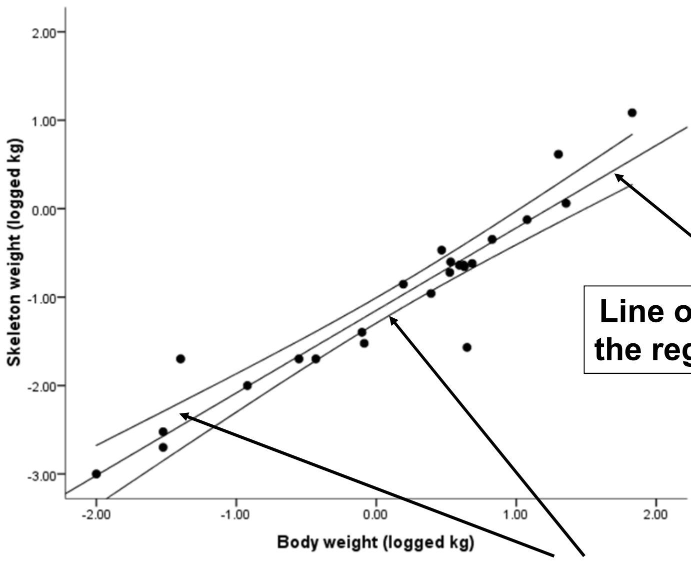
Then click on *Elements*

Then click on *Fit Line at Total*

18



19



Line of best 'fit' or the regression line

95% confidence intervals

20

Skeleton weight in mammals is dependent on body weight (R^2=0.885, $F_{(1,22)}$=178.592, p<0.001)(Figure 1).

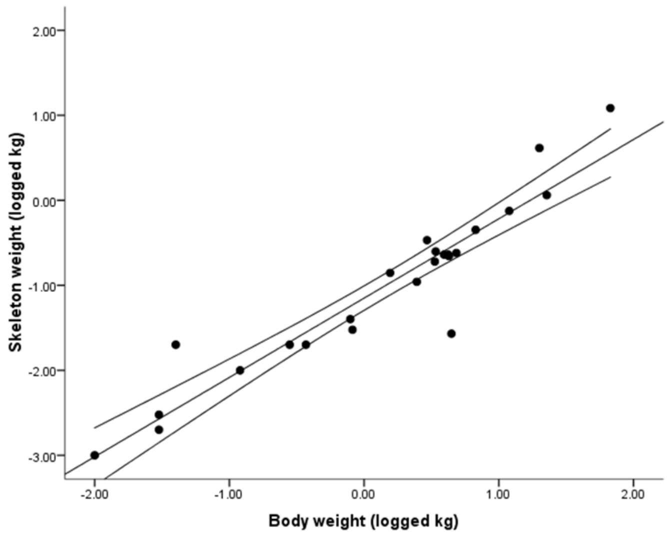


Figure 1. The relationship between body weight and skeleton weight in mammals (Line of best fit ± 95% CI)

21

Making predictions

- Regression models can be used to make predictions
  - Care should be taken when making predictions from regression models.
    - When your R^2 value is low (<0.60) you can not have high confidence in the models predictions
    - Do not make predictions outside the range of your data. You do not know how variables will relate to each other outside the range of the data you have.

22

- To make predictions from your regression model you need to use the coefficients output table
  - This table refers to the formula of the regression line
  - The information can be used to predict skeleton weight from body weights.
  - Remember the formula of a line is:-

Slope of the line

Y-intercept

Predicted skeleton weight

$$y = mx + c$$

Value of body weight for which a skeleton weight will be predicted

23

- Remember how you have transformed your data!
- e.g. Our formula will be
  - Log(Skeleton weight in kg) = m * log(body weight in kg) + c
  - All we need to know is what m and c are.
- Some times you will only have one axis transformed so make sure you look at your formula carefully

24

Represents c in the regression line formula

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-1.150	.069		-16.590	.000	-1.294	-1.007
	Log_body_kg	.933	.070	.944	13.364	.000	.788	1.078

a. Dependent Variable: Log_skeleton_kg

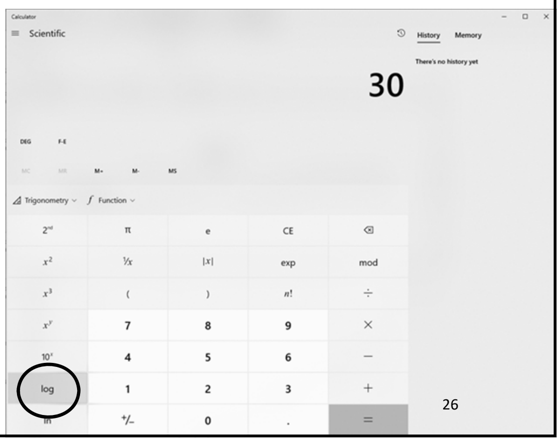
Represents m in the formula of the regression line

Used for estimating the upper and lower limits (95% conf) of a prediction

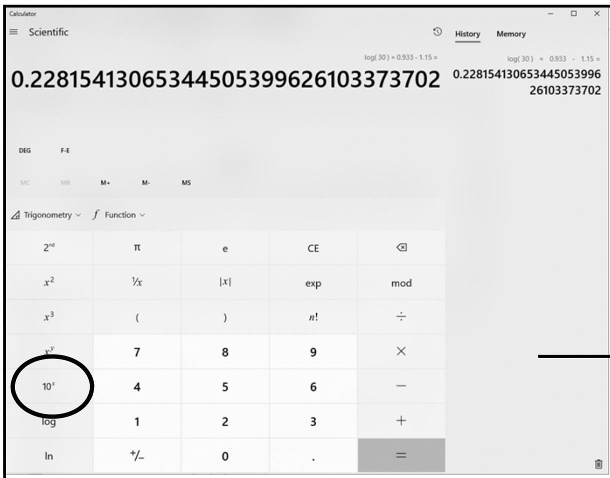
25

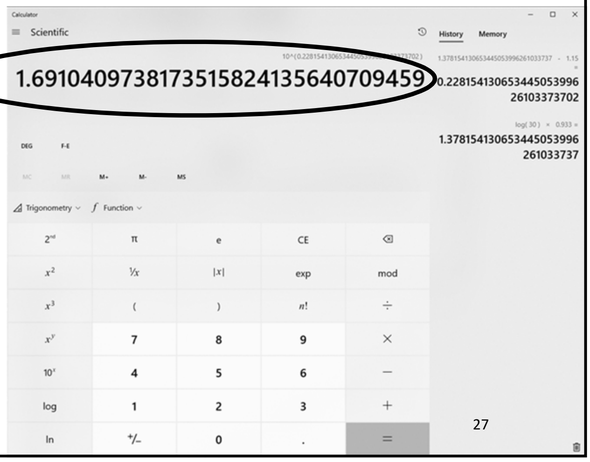
log (skeleton weight in KG) = 0.933*(log(body weight in KG)) - 1.150

- Example: An animal is captured with a body weight of 30. What is the predicted skeleton weight of the animal?
  - Step 1: Substitute the values for c,m & x into the formula of the line.
    - c= -1.150, m= 0.933, x= 30 (remember to log)
    - Log (Skeleton weight) = 0.933*log(30) - 1.150= 0.22815
    - Remember the output is still logged!!!!!



26





So ultimately a 30 kg animal has a skeleton weight of 1.69kg

27

- Step 2: Estimate the upper and lower confidence intervals
  - Upper = 1.078*log(30) -1.007 = 3.84
  - Lower = 0.788*log(30) -1.294 = 0.74
  - Remember these answers have been inverse logged
- The predicted skeleton weight of an animal weighing 30 is 1.69kg with a range of 0.74kg to 3.84kg (mean ± 95% CI)

28

- Be careful with uneven spreads of observations.
  - Outlying observations can significantly effect the estimation of a regression line
  - Transformation of data may fix uneven spreads of data
  - Log_{10} transformations are the most commonly used transformation but other transformations can be used where appropriate
- The skeleton/body weight regression had an uneven spread of data.
  - Log transforming both axis fixed this problem

29

Predicting from transformed data

- If you transform your data you need to change the formula accordingly
  - For example: You would need to log the body weight measurement before it went into the formula. The answer would also be logged so you need to inverse log that number to get a weight
  - In some cases you only transform one variable so this would mean one number would be transformed in the formula and one would not be.

30

Regression notes

- Check the data is linear
  - transform if it is not
- the R^2 value tells you how good your model is. The closer to 1 the better.
- Make sure the line is significant (ie: is $P<0.05$)
- Do not try to predict outside the limits of your data!!!!
- Remember if you transform you data, take this into account when developing your prediction formula.

31