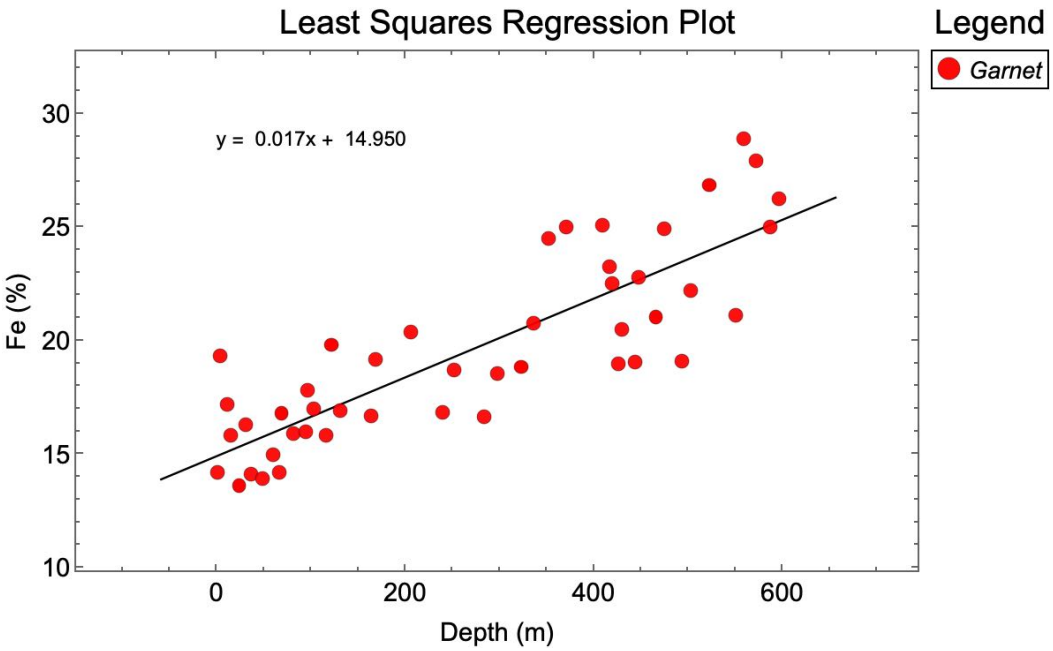
Data Analysis & Statistics for Earth Scientists
Nanjing University, Spring 2026

Lab 2 Assignment w/ Answers

1. The Garnets dataset (Garnets.dat, Garnets.csv) contains data that quantify changes in the abundance of iron in garnet crystals collected from a core drilled into the metamorphic halo surrounding an igneous intrusion.

    1. Plot these data. (10 points)



    2. Select a linear regression model that will allow the concentration of iron (Fe) to be predicted as a function of depth into the metamorphic halo. (10 points)

        Ordinary Least Squares Linear Regression (Bivariate)

        a. Justify your selection. (20 points)

        Since the point of the exercise is to predict Fe concentration in terms of core depth there needs to be a distinction between the variables and the regression needs to minimize variation about only one (the dependent variable). Thus the depth variable is the independent variable and Fe concentration the dependent variable.

    3. List the equation of the regression line for the model you have selected. (10 points)

$$y = 0.017x + 14.95$$

        a. Use this equation to predict the Fe concentration at a depth of 300 meters. (20 points)

$$y = (0.017 * 300) + 14.95 y = 20.05$$
$$y = 20.05$$

    4. Use an ANOVA *F* test to estimate the significance of the regression model.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	526.600	1	526.600	106.883
Deviation	216.800	44	4.926	
Total	743.300	45		

a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)

$$p = 0.00$$

b. Provide an interpretation of the ANOVA test result in terms of the degree to which the regression model can be regarded as constituting an accurate prediction. (20 points)

The regression model explains 70.85% of the total data sum of squares with only 29.17% explained via deviations from the regression model. Given the size of the dataset the resulting F-ratio suggests the null hypothesis of equivalence between explained and unexplained variance can be rejected ar the α = 0.05 confidence level.

c. Estimate the 95% confidence interval for the regression result. (20 points)

Regression slope: 0.020 (lower 95% limit) - 0.015 (upper 95% limit)
Regression y-intercept: 14.19 (lower 95% limit) - 15.71 (upper 95% limit)

1. Estimate the range of variation in Fe concentration values that would be expected at depth of 300 meters. (30 points)

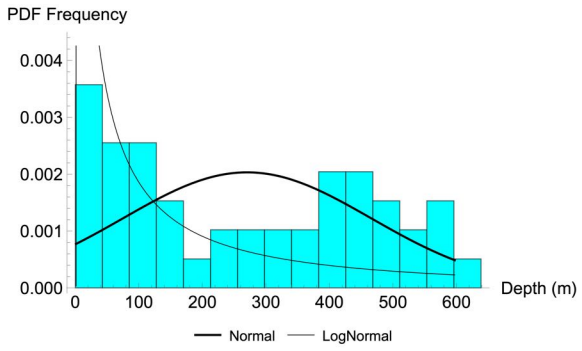
$$y = (0.020 * 300) + 14.19$$
$$y = 20.19$$

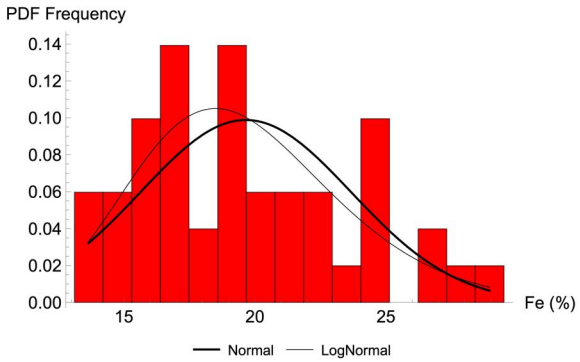
$$y = (0.015 * 300) + 15.71$$
$$y = 20.21$$

5. Do these data conform to the assumptions of an ANOVA test?

Probably not.

a. Justify your answer. (20 points)





Normal Distribution Tests

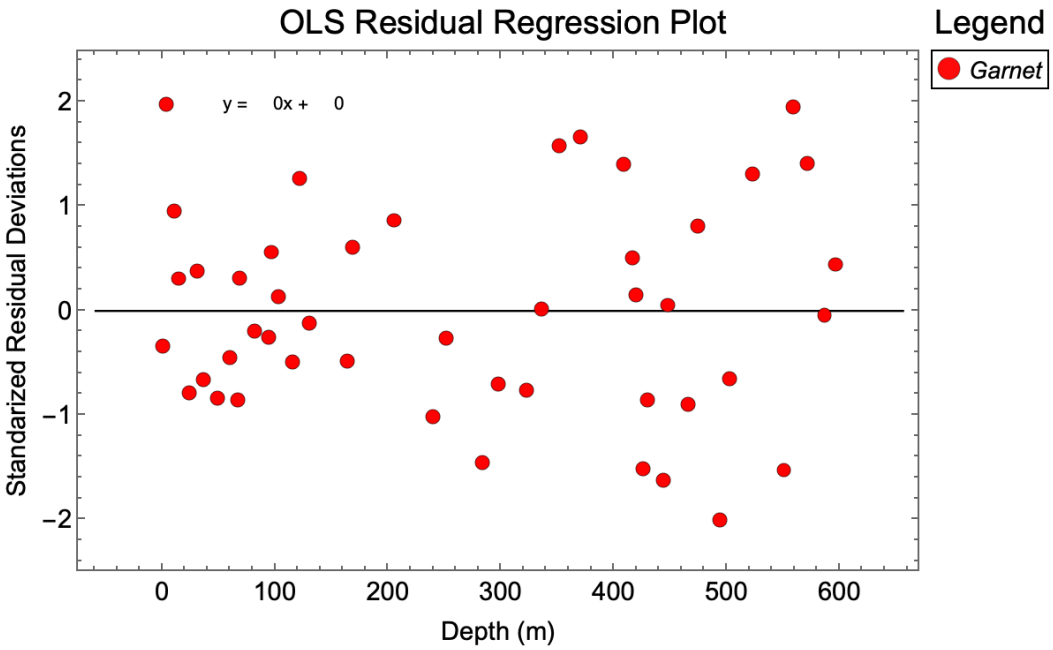
Variable	Mean	Std. Dev.	Anderson–Darling A^2	Crammér– von Mises ω^2	Pearson χ^2
Depth (m)	269.220	194.910	0.1948	0.2139	0.1077
Fe (%)	19.621	4.020	0.4976	0.4770	0.2787

Neither dataset conforms to a normal distribution especially well. In addition, Levene's test rejects the null hypothesis of equivalence between variances ($p = 1.58 \times 10^{-25}$).

There's no information regarding whether or not these garnets were sampled randomly along the length of the core. Most likely they were not.

However, this being said, the probability of value associated with the F-test is so low that a linear regression model does fit these data and the FE concentrations estimated as a result of this model are probably robust to deviations from ANOVA test.

6. Perform any additional test(s) you deem appropriate in order to confirm the validity of the regression model.
   a. Show all plots, secondary statistical tests, and results associated with these additional tests (if any are warranted). (20 points)



Residual Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	0	1	0	0
Deviation	17710.000	44	402.500	
Total	17710.000	45		

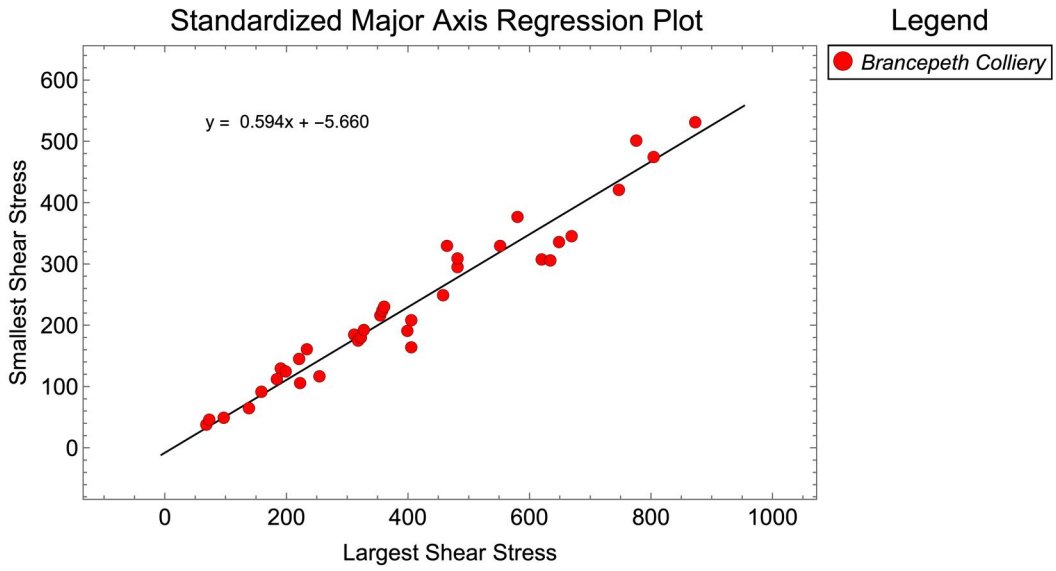
   b. Describe how the results of any additional tests (if any are warranted) either increased or decreased your confidence in your regression analysis. (20 points)

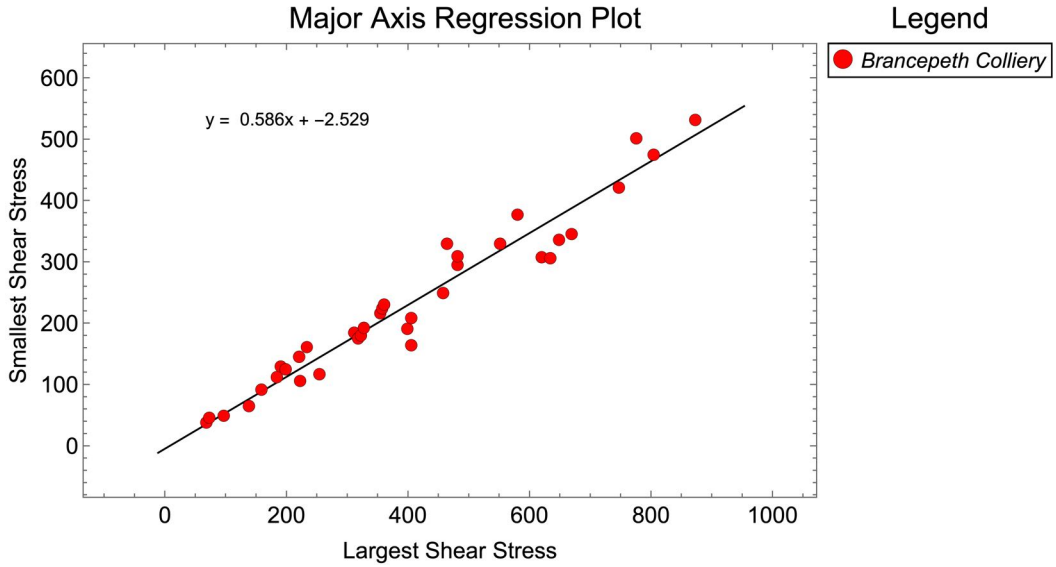
Analysis of the regression residuals show a bit of heteroscedascity, but the residual regression is clearly non-significant with no slope. Overall, this provides further evidence of the appropriateness of the OLS linear regression result.

2. On October 21 1966 a colliery spoil heap in the village of Aberfan, Wales collapsed after a period of locally heavy rainfall, burying part of the village that was located immediately downslope from the spoil heap. One-hundred and forty-four villagers died in this incident, including 116 children who attended the Pantglas Junior School in Aberfan that day. As a result, the physical and chemical conditions of colliery spoil heaps throughout Britain were investigated.

   Shear strength measurements were made on the Brancepeth Colliery spoil heap in County Durham (Taylor, 1973). The largest and smallest principal stresses collected from 36 spoil heap samples are listed in the Brancepeth Colliery dataset (Brancepeth Colliery.dat, Brancepeth Colliery.csv). Estimation of the principal stress angle is critical for the purpose of determining whether the spoil heap is prone to collapse under conditions like those that caused the collapse at Abferfan. The sine of the principle stress angle can be estimated as the slope of the "best fit" regression between these two variables.

1. Plot these data. (10 points)





2. Select a linear regression model that will allow the principle stress for this colliery spoil heap to be predicted. (10 points)

   Either Standardized Major Axis or Major Axis

   a. Justify your selection. (20 points)

   Because both variables are being used to estimate the principal stress there is no distinction between the variables and no prediction of one variable's value in terms of the other is involved. As a result, the regression model (either SMA or MA) must minimize the residuals of both variables from the regression model simultaneously. Because the ranges of the variables are quite close to one another the decision as to which regression model to choose is essentially a judgement call, with a slight preference for MA because the units for both variables are identical.

3. List the equation of the regression line for the model you have selected. (10 points)

   Standardized Major Axis

   $$y = 0.594x - 5.660$$

Major Axis
$y = 0.586x - 2.529$

a. Use this equation to predict the angle of the principle stress estimate (in degrees). (20 points)

SMA Regression: 36.4417°
MA Regression: 35.874°

4. Use an ANOVA *F* test to estimate the significance of the regression model.

Standardized Major Axis Trend Line ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Trend Line Residuals	582000.000	1	582000.000	2468.600
Deviation Residuals	8016.000	34	235.800	
Total	590000.000	35		

Major Axis Trend Line ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Trend Line Residuals	572400.000	1	572400.000	1106.470
Deviation Residuals	17590.000	34	517.300	
Total	590000.000	35		

a. Based on the purpose of this analysis decide what, in your opinion, an appropriate probability value would be for use in reporting your results.

$p = 0.99$

1. Justify your reasoning. (30 points)

Based on the disaster that struck Aberfan, Wales the danger to human life posed by these mine soil heaps is obvious. Accordingly, it was necessary to be very certain of the stability of similar spoil heaps in other parts of the UK. Indeed, since mining is widespread throughout the world, since mines are often sited in hilly or mountainous areas, since spoil heaps are often a by-product of mining operations and since human habitations are often sited near mines, the stability of mine spoil heaps represents a significant human health risk in a large number of communities, worldwide.

b. To two decimal places state the probability value associated with the ANOVA test result. (10 points)

$p = 0.00$ for both the SMS and MA regression of these data.

c. Provide an interpretation of the ANOVA test result in terms of the degree to which the regression model can be regarded as constituting an accurate prediction. (20 points)

In both cases (SMA and MA) the trend line accurately represents the overwhelming majority of the data with only a very small proportion left unaccounted for as residual variation.

d. Estimate the 95% confidence interval for the regression result. (20 points)

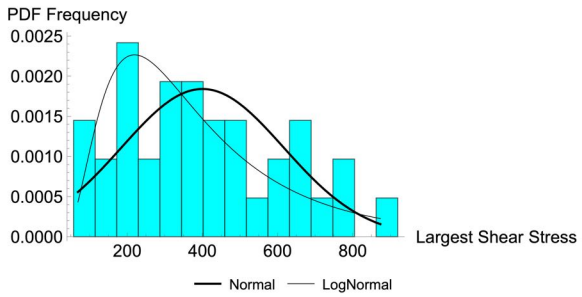
SMA Regression
Upper 99% Confidence Limit: $y = 0.662x - 32.670$
Lower 99% Confidence Limit: $y = 0.533x - 18.580$

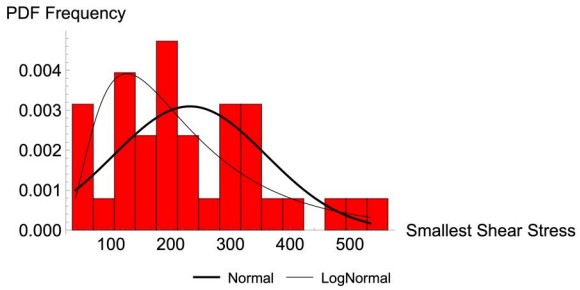
MA Regression
Upper 99% Confidence Limit: $y = 0.653x - 29.350$
Lower 99% Confidence Limit: $y = 0.522x - 22.800$

1. Estimate the range of principle stress angle variation that would be expected at this colliery. (30 points)

   SMA Regression: 32.208° to 41.453° → Principal Stress: 29.113° to 35.220°
   MA Regression: 31.467° to 40.768° → Principal Stress: 28.569° to 34.811°

5. Do these data conform to the assumptions of an ANOVA test?
   a. Justify your reasoning. (20 points)





Normal Distribution Tests

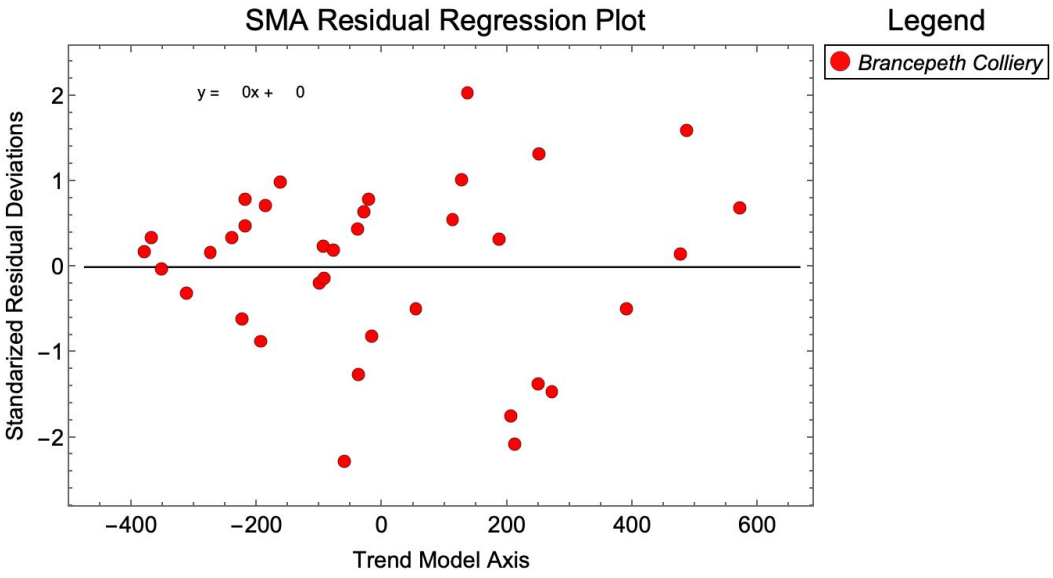
Variable	Mean	Std. Dev.	Anderson–Darling A^2	Crammér– von Mises ω^2	Pearson χ^2
Largest Shear Stress	398.320	215.580	0.7929	0.7488	0.1749
Smallest Shear Stress	230.870	128.020	0.7301	0.6798	0.7576

Both datasets conform to expectations of a normal distribution reasonably well. As usual, there no information regarding whether or not these stress assessments were sampled randomly. Most likely they were not.

However, the datasets do not conform to the assumption of variance homogeneity (Levene Test = 9.17, p = 0.003).

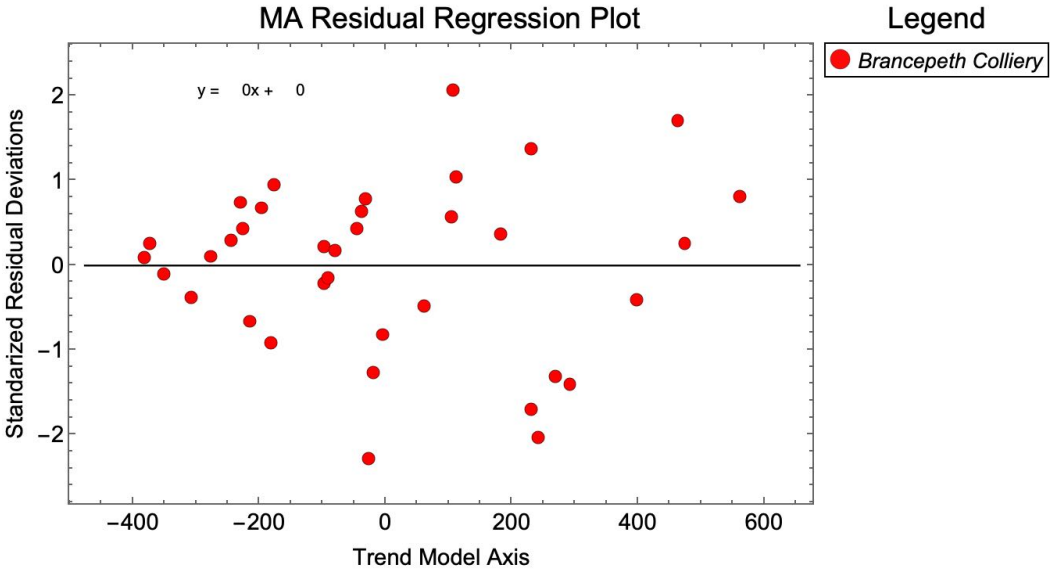
This being said, the probability of value associated with the F-test is so high it indicates that a linear regression model does fit these data well. Thus, the principal stress values estimated on the basis of this model are probably robust to deviations from ANOVA test assumptions.

6. Perform any additional test(s) you deem appropriate in order to confirm the validity of the regression model.
   a. Show all plots, secondary statistical tests, and results associated with these additional tests (if any are warranted). (20 points)



Residual Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Trend Line Residuals	0	1	0	0
Deviation Residuals	1.919×10^6	34	56440.000	
Total	1.919×10^6	35		



Residual Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Trend Line Residuals	0	1	0	0
Deviation Residuals	1.919×10^6	34	56440.000	
Total	1.919×10^6	35		

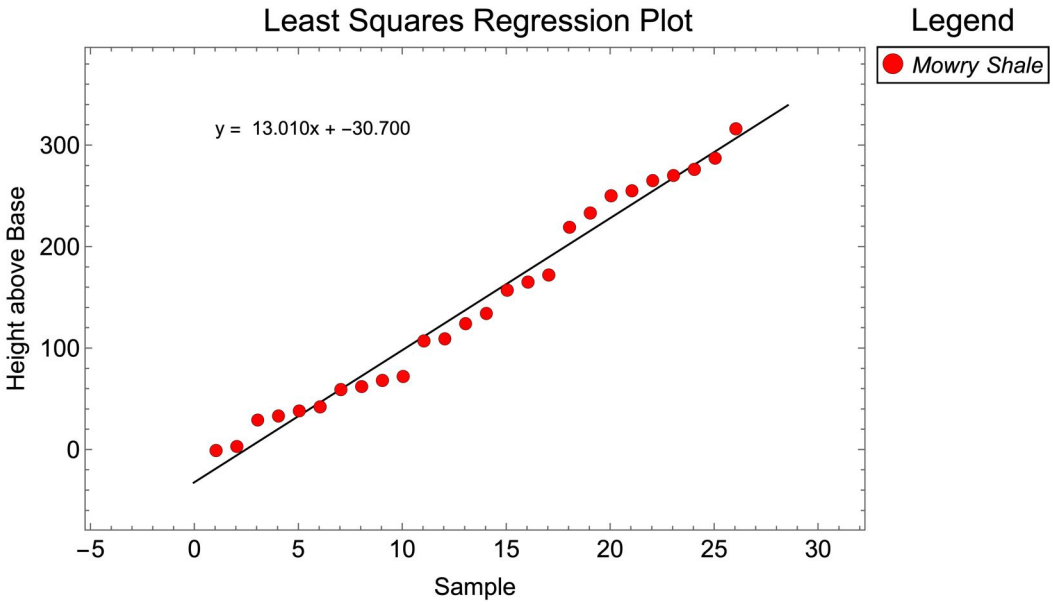
b. Describe how the results of any additional tests (if any are warranted) either increased or decreased your confidence in your regression analysis. (20 points)

Analysis of the regression residuals show a bit of heteroscedascity but the residual regression is clearly non-significant with no slope. Overall, this provides further evidence of the appropriateness of the SMA/MA linear regression models.

3. The Mowry Shale is a formation occurring in the US states of Colorado, Wyoming and Montana. This is a thick unit of black shale beds interspersed with numerous minor beds of commercially exploitable bentonite. Bentonite is almost entirely composed of the clay mineral montmorillonite, which is an alteration product of rhyolitic or andesitic volcanic ash.

   The Mowry Shale dataset (Mowry Shale.dat, Mowry Shale.csv) contains measurements of the thicknesses (in cm) and positions (in m) above the base of the Mowry Shale of a sequence of 26 bentonite beds. If it can be assumed that the marine shales of the Mowry Formation accumulated at a constant rate, it may be possible to determine the frequency of local volcanic eruptions. Test this hypothesis using linear regression analysis.

   1. Plot these data. (10 points)



   2. Select a linear regression model that will test the data for use in the estimation of eruption frequency. (10 points)

      Ordinary Least Squares Bivariate Regression

      a. Justify your selection. (20 points)

      Since the point of the analysis is to determine whether the spacing of bentonite beds in the Mowry Shale is sufficiently constant to enable eruption times to be estimated the analysis involves an assessment of the interval between beds. Assuming all eruptions are recorded by bentonite beds this is most easily accomplished by examining the relation between the rank orders of beds and the levels within the section. Since, in this case, we are predicting the level based on the bed rank orders the bed rank orders constitute the independent variation, and the bentonite bed levels, the dependent variable, of an OLS regression.

   3. List the equation of the regression line for the model you have selected. (10 points)

$$y = 13.01x - 30.70$$

Use this equation to predict the frequency of eruptions (if appropriate) assuming a standard range of rock accumulation rates for deep-sea shales (between 0.23 cm/yr and 4.6 cm/yr). (20 points)

$$13.01/0.23 = 56.57 \text{ years}$$

$$13.01/4.6 = 2.82 \text{ years}$$

Obviously, rock accumulation rates in the Cretaceous Interior Seaway were more toward the lower part of the accumulation range for modern deep-sea environments.

4. Use an ANOVA *F* test to estimate the significance of the regression model.

Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	247700.000	1	247700.000	1227.310
Deviation	4844.000	24	201.800	
Total	252600.000	25		

a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)

$$p = 0.0000$$

b. Provide an interpretation of the ANOVA test result in terms of the degree to which the regression model can be regarded as constituting an accurate prediction. (20 points)

For this dataset the regression model accounts for the overwhelming proportion of the observed sum-of-squares variance (98.06%), leaving only a very small residual error (1.94%) unaccounted for. Given the size of the dataset the resulting F-ratio suggest the null hypothesis of equivalence between explained and unexplained variance can be rejected ar the α = 0.05 confidence level.

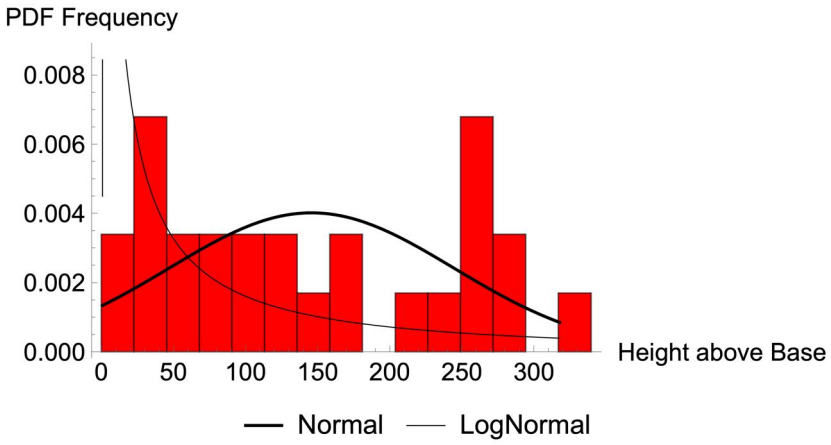
c. Estimate the 95% confidence interval for the regression result. (20 points)

OLS Regression
Upper 95% Confidence Limit: $y = 13.650x - 39.270$
Lower 95% Confidence Limit: $y = 12.380x - 22.120$

1. Estimate the range of variation in eruption frequency estimates that would be expected based on your analysis. (30 points)

Upper Confidence Limit: 59.35 - 2.97 years
Lower Confidence Limit: 53.83 - 2.69 years

5. Do these data conform to the assumptions of an ANOVA test?



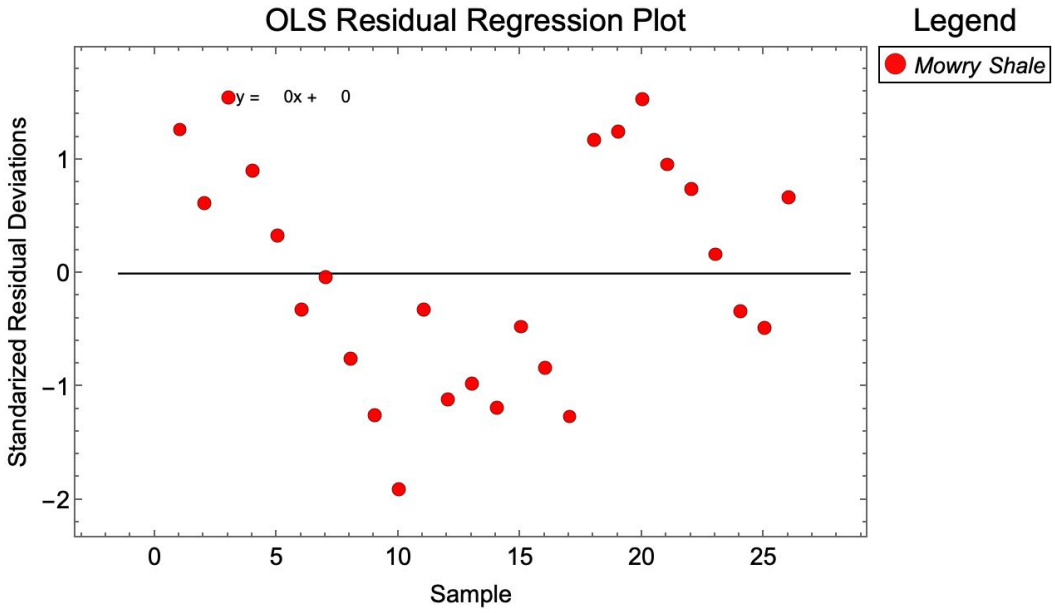
Normal Distribution Tests

Variable	Mean	Std. Dev.	Anderson–Darling A^2	Crammér– von Mises ω^2	Pearson χ^2
Height above Base	145.000	98.560	0.5249	0.5189	0.2067

a. Justify your reasoning. (20 points)

No, they do not. The data are not normally distributed and there is little evidence the data were, or could be, selected randomly from the population under consideration.

6. Perform any additional test(s) you deem appropriate in order to confirm the validity of the regression model.
   a. Show all plots, secondary statistical tests, and results associated with these additional tests (if any are warranted). (20 points)



Residual Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Trend Line Residuals	0	1	0	0
Deviation Residuals	1.919×10^6	34	56440.000	
Total	1.919×10^6	35		

b. Describe how the results of any additional tests (if any are warranted) either increased or decreased your confidence in your regression analysis. (20 points)

Strictly speaking the residual analysis suggests a single bivariate linear regression represents a suboptimal model for these data. Residual datapoints exhibit highly structured variation about the single regression line indicating systematic departures from its trend. Looking back at the original regression plot a (slightly) more accurate model could be obtained by subdividing the dataset into three different categories of eruption events at levels (samples) 11 and 18 and fitting separate OLS regressions to each of these three datasets. This alteration has substantial implications for the interpretation of the eruptions represented by these bentonite beds (see below). However, in terms of estimating their overall timing, it's doubtful the more complex, segmented regression model would have a substantial effect on that estimate.

7. If no significant linear trend is found list a few factors that many be responsible for this negative finding. (10 points)

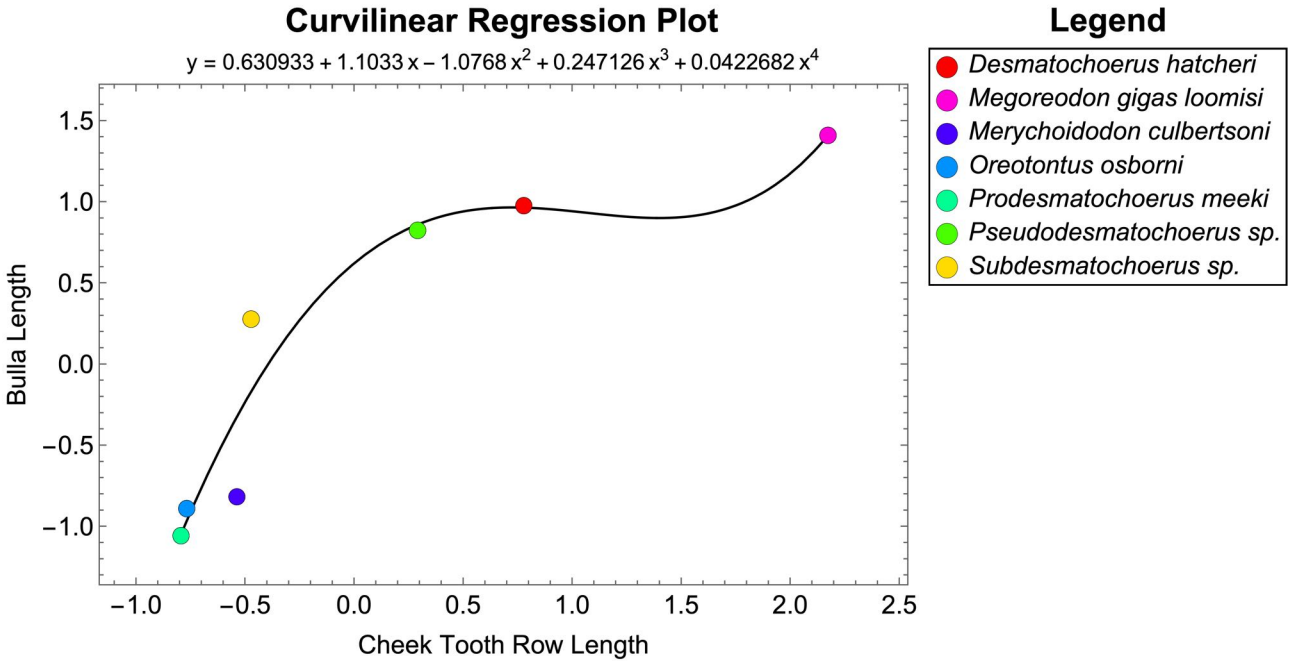
Since the residual analysis did find evidence against regarding the Mowry bentonite beds as evidence of a regular, clock-like, single source eruption interpretation, two alternative interpretations are indicated.

Interpretation 1: Each subordinate category of bentonite emplacement data represents the episodic eruption of a different volcano in the local area the timing of the eruptive episodes different slightly from those of other volcanoes in the area.

Interpretation 2: The bentonite beds record the eruption from a single volcanic center, but one whose episodic eruptive schedule was modified twice during the time interval represented by these beds. Such changes could reflect developmental stages in the magma chamber that was the ultimate source of the bentonite material.

4. Merycoidodontoids (known previously as “oreodonts”) are an extinct group of Cenozoic pig-like mammals whose skulls are commonly found vertebrate fossils in the American west. This was a primitive group of cud-chewing artiodactyls with short fasces and long, fang-like canine teeth. The Oreodont skulls dataset (Oreodont Skulls.dat, Oreodont Skulls.csv) contains a list of skull measurements collected from 72 specimens representing seven species. Fit a regression model to these data such that such that the groups are arranged in a systematic manner (with the groups as well-separated as possible) using any bivariate combination of variables and any appropriate linear regression model. Feel free to transform the variables if this improves the model fit.

   1. Plot these data. (10 points)



   2. Select a set of linear regression variables that will allow these these to be distinguished from one another to the maximum extent possible. (10 points)

      The best model I found was a fourth-order curvilinear (polynomial) OLS regression on the standardized variable contrast between the bivariate mean Cheek Tooth Row Length and Bulla Length for each taxonomic group.

      a. Justify your selection. (20 points)

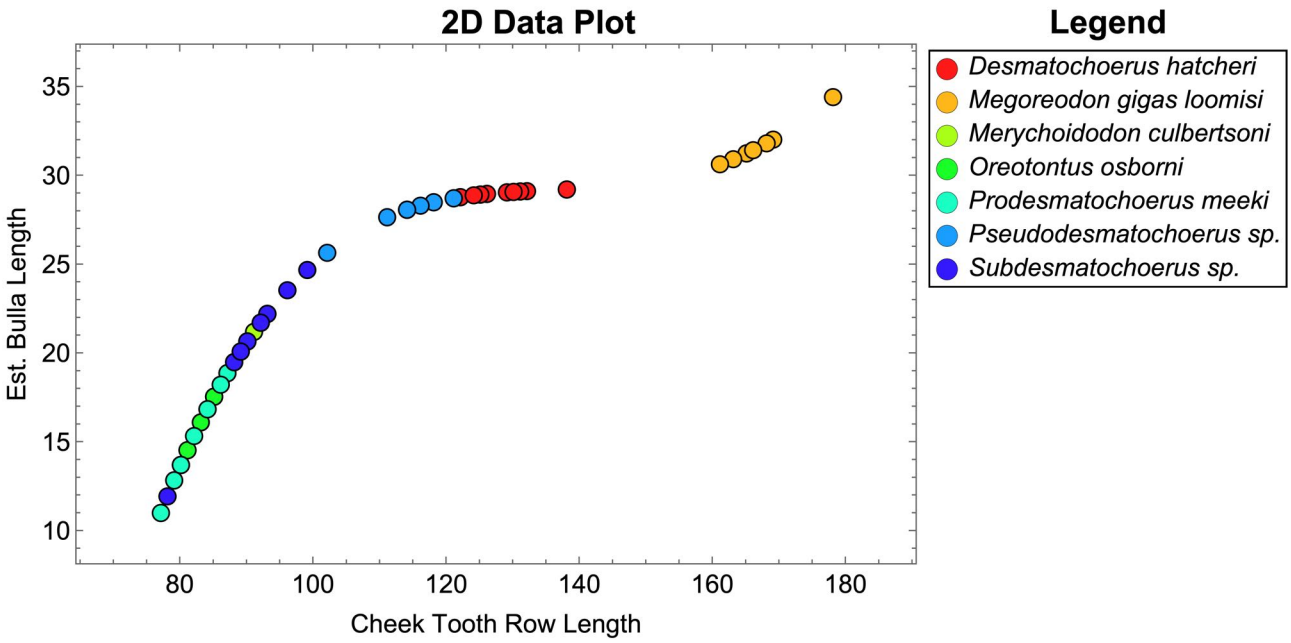
Standardization of the data is justified because, although both variables are lengths, the magnitudes of the values differ by an order of magnitude. Regression through the group means is justified because the distribution of bivariate means represents the best simple summary of the location of each group relative to the others in the bivariate space. Use of a curvilinear OLS regression to model the geometric relation between group means is justified because their trend lies along an obviously complex curve. Use of the contrast between cheek tooth length and bulla length is justified because these variables produce the most compact and mutually exclusive group dispersions among any bivariate ordination of these data. Use of the fourth-order polynomial function is justified because the use of this curve represents the simplest continuous function that results in localised group distinctions.

3. List the equation of the regression line for the model you have selected. (10 points)

$$y = 0.6309 + 1.1033x - 1.0768x^2 + 0.2471x^3 + 0.0423x^4$$

a. Use this equation to predict the taxonomic identities of the unknown skulls. (20 points)

In order to asses the accuracy of the model I randomly selected a skull from each species, recomputed the mode as above, using the remainder of the data, and used that model to assess the appropriateness of the estimated identifications. This strategy can be repeated *k* times to estimate how stable a linear model based on these variables is.



Oreodont Skulls (ID Test)

No.			**4**	**17**	**22**	**35**	**54**	**60**	**66**
Unknown			*D. hatcheri*	*M. gigas loomisi*	*M. culbertsoni*	*O. osborni*	*P. meeki*	*Pseudo.* sp.	*Subdesmat.* sp.
Cheek Tooth Row Length			123	174	86	87	85	117	100
Est. Bulla Length			28.87	33.27	18.27	18.91	17.59	28.44	25.07
Desmatochoerus hatcheri	128.56	29.11	5.57	45.63	43.92	42.79	45.06	11.58	28.85
Megoreodon gigas loomisi	166.88	31.63	43.97	7.31	81.98	80.88	83.07	49.98	67.20
Merychoidodon culbertsoni	91.00	17.80	33.86	84.43	5.02	4.15	6.00	28.09	11.57
Oreotontus osborni	83.75	16.75	41.08	91.75	2.71	3.90	1.50	35.25	18.25
Prodesmatochoerus meeki	83.10	16.30	41.83	92.47	3.50	4.70	2.30	36.01	19.04
Pseudodesmatochoerus sp.	113.71	28.00	9.33	60.52	29.37	28.21	30.54	3.32	14.02
Subdesmatochoerus sp.	91.56	24.89	31.69	82.87	8.65	7.52	9.82	25.69	8.44

Overall the curvilinear regression model appears to deliver correct identifications in c. 57 percent of cases. The species misidentified are all clustered in the lower end of the plot when species groups overlap strongly and mean values are clustered densely (see above).

4. Use an ANOVA *F* test to estimate the significance of the regression model.

Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F	Probability (%)
Regression	5.751	1	5.751	21.584	0.00560311
Deviation	1.332	5	0.266		
Total	7.083	6			

a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)

$$p = 0.0056$$

b. Provide an interpretation of the ANOVA test result in terms of the degree to which the regression model can be regarded as constituting an accurate prediction. (20 points)

The model captures a greater proportion of the observed data than would be expected under a random sampling from a population of normally distributed bivariate variable values.

c. Estimate the 95% confidence interval for the regression result. (20 points)

OLS Curvilinear (Polynomial) Regression

Upper 99% Confidence Limit: $y = 1.6206 + 3.1297x + 1.7876x^2 + 3.9101x^3 + 1.9378x^4$

Lower 99% Confidence Limit: $y = -0.3587 - 0.9231x - 3.9412x^2 - 3.4159x^3 - 1.8533x^4$

1. Determine whether the range of variation in these predicted values has an effect on the certainty of your identifications. (30 points)

The confidence intervals degrade the accuracy of identifications substantially. This is, no doubt, due to the decision to based the regression model on the bivariate group means, which dropped the sample size from 72 to 7, resulting is a corresponding increase in the confidence interval. However, in defense of this decision, basing the regression on the group means ensured the regression model adopted an orientation that passed through the group point clouds to a much greater extend that would have been the case had the full dataset been used. This strategy also corrected for variable group sample sizes.

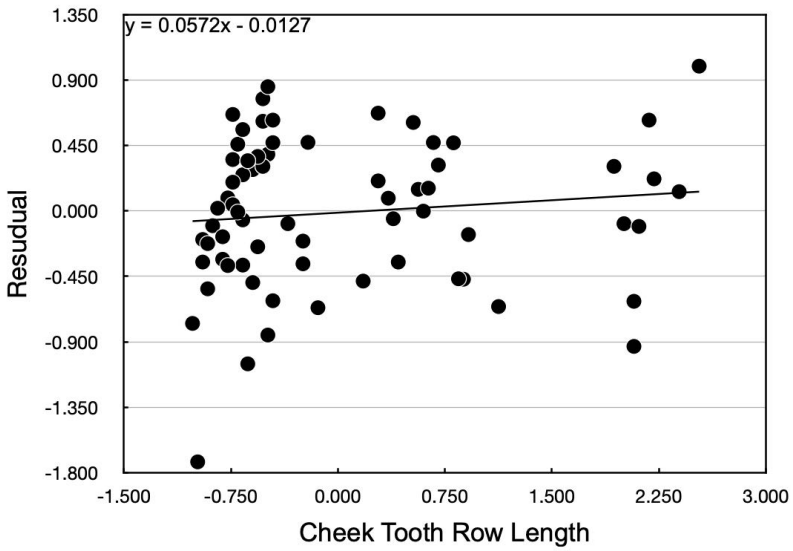
5. Do these data conform to the assumptions of an ANOVA test?
   a. Justify your reasoning. (20 points)

Unlikely for the data as a whole. However, the central limit theorem suggests the group means should conform to a normal distribution. Regardless, since there is no indication that the data constitute a truly random sample of the parent populations the strict assumptions of the ANOVA test should be regarded as having been violated for this sample.

6. Perform any additional test(s) you deem appropriate in order to confirm the validity of the regression model. (20 points)

Since the purpose of this analysis is merely to find a contrast between variables that provides the most consistent species identification the statistical validity of the regression is not really an issue for this analysis. The validity of the result can be assessed by the number of misidentifications produced by the model of the contrast selected. For this particular dataset the curvilinear regression was only able to manage a bit more than 50% correct post-hoc identifications.

However, for completeness a an analysis of the regression residuals can always be calculated.

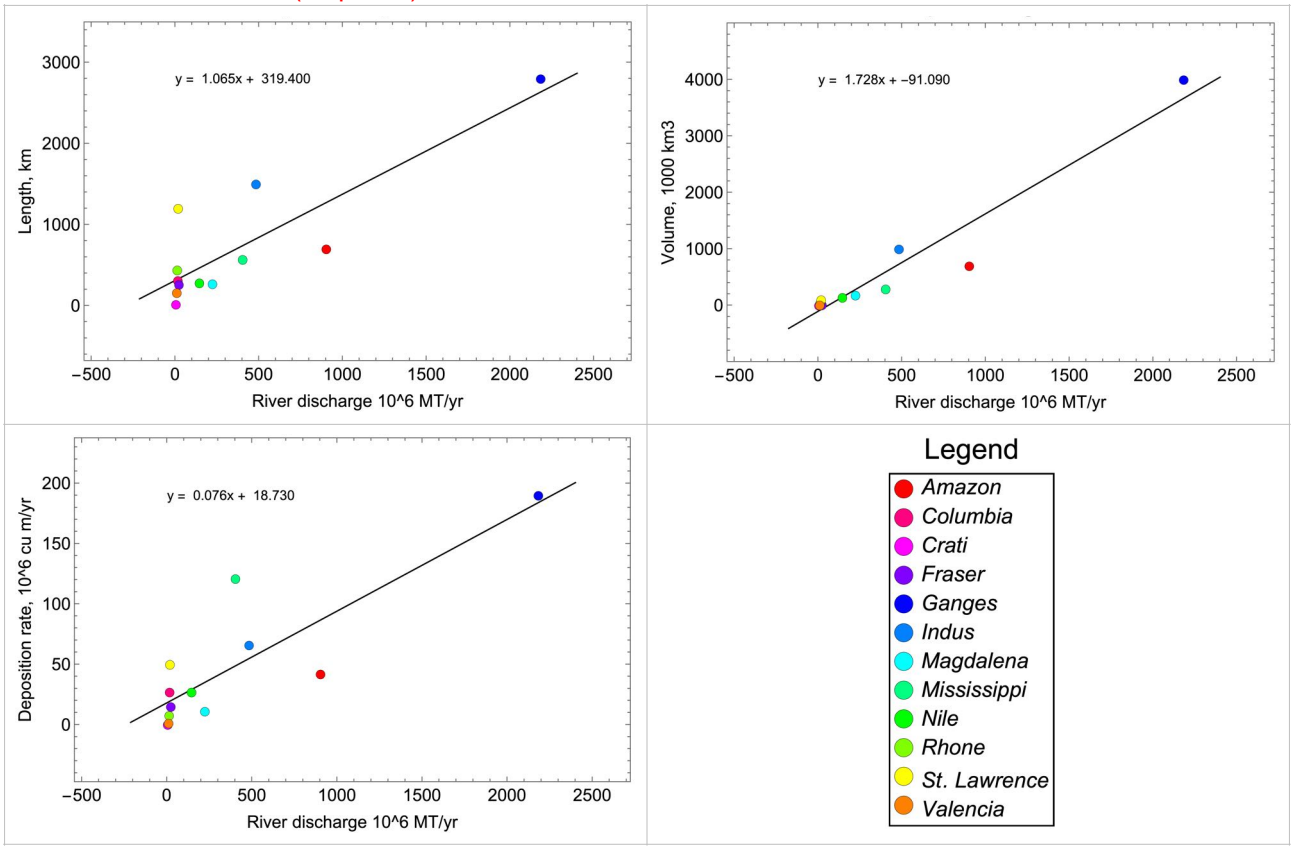


This result exhibits a great deal of structure in the residual values as well as a slight slope to the OLS regression model. As such, application of a curvilinear regression model to these data was, on the whole, only marginally successful in identifying differences among these variables sufficient to obtain reliable identifications.

5. Many major rivers are contiguous with extensive deep-se fans at their river mouths, the sizes of which presumably reflect the sediment loads that are transported by the river. However, submarine fans do exist that do not appear to be associated with a major current river system. These are probably relict fans whose rivers have either dried up due to climate change or been captured by a different drainage system.

   The submarine fans dataset (Submarine Fans.dat, Submarine Fans.csv) contains data for 12 submarine fans associated with extant drainage systems and 5 relict fans. Use multiple regression analysis to predict the relation between the river discharge variable and the other relevant physical variables included in these data.

   1. Plot these data. (10 points)



   2. Select a linear regression model that will allow the river discharge rate to be predicted as a function of the other relevant physical variables. (10 points)

      Ordinary least squares multiple regression

      a. Justify your selection. (20 points)

      Since the analysis is ultimately being carried out for the purpose of prediction the OLS model is preferred.

   3. List the equation of the regression line for the model you have selected. (10 points)

$$y = 101.753 - 0.203x_1 + 0.587x_2 + 1.942x_3$$

x_1 = Length
x_2 = Volume
x_3 = Depositional Rate

a. Use this equation to estimate the discharge rates for the missing rivers based on their relict fans. (50 points)

Fan	Length, km	Volume, 1000 km3	Deposition rate, 10^6 cu m/yr	River discharge 10^6 MT/yr
Cap Ferret	75	13	6.5	106.782
Delgada	350	40	3.6	61.174
La Jolla	40	12	0.6	101.842
Monterey	300	64	2.7	83.664
Navy	40	0.08	0.2	94.068

4. Use an ANOVA test to estimate the significance of your regression model.

Multiple Regression ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	4.17827×10^6	3	1.39276×10^6	46.6652
Deviation	238765.	8	29845.7	
Total	4.41703×10^6	11		
Probability (%)	0.00205523			

a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)

$$p = 0.00002$$

b. Provide an interpretation of the ANOVA test result in terms of the degree to which the regression model can be regarded as constituting an accurate prediction. (20 points)

The regression models does account for the overwhelming majority (94.59%) over the variation in the observed data. However, the sample size is quite small and the regression is being differentially influenced by a single prominent outlier (Ganges). If this outlier were removed from the dataset the regression model would be very different. Accordingly, this regression analysis, while useful in the context of the absence of an alternative estimation procedure, should be regarded as approximate at best.

c. Estimate the 95% confidence interval for the regression result. (20 points)

OLS Multiple Regression

Upper 95% Confidence Limit: $y = 112.85 + 0.283x_1 + 1.386x_2 + 3.360x_3$

Lower 95% Confidence Limit: $y = 90.655 - 0.689x_1 - 0.212x_2 + 1.931x_3$

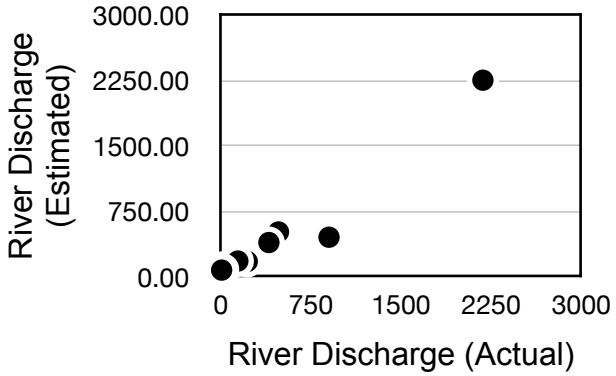
1. Estimate the range of variation relict river discharge rate values values that would be expected. (30 points)

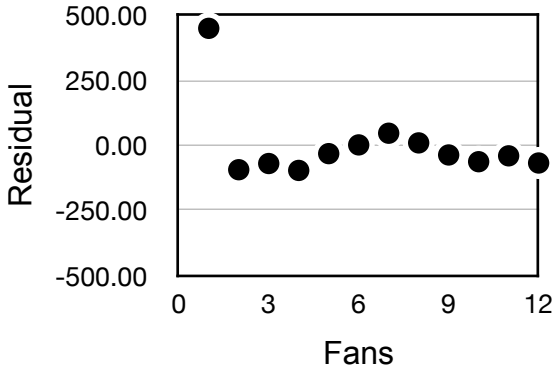
Fan	Length (km)	Volume (1000 km3)	Deposition rate (10^6 cu m/yr)	River discharge 10^6 MT/yr (Upper Limit)	River discharge 10^6 MT/yr (Lower Limit)
Cap Ferret	75	13	6.5	173.933	48.776
Delgada	350	40	3.6	279.436	-152.023
La Jolla	40	12	0.6	142.818	61.710
Monterey	300	64	2.7	295.526	-124.399
Navy	40	0.08	0.2	124.953	63.464

5. Do these data conform to the assumptions of an ANOVA test?
   a. Justify your reasoning. (20 points)

No. In addition to the presence of the outlier (see above) the sample size is quite small and quite possibly non-representative of all the rivers and associated submarine fans that have ever existed in earth history. Use of such a small number of modern submarine fans also more-or-less precludes random sampling.

6. Perform any additional test(s) you deem appropriate in order to confirm the validity of the regression model.
   a. Show all plots, secondary statistical tests, and results association with these additional tests (if any are warranted). (20 points)





   b. Describe how the results of any additional tests (if any are warranted) either increased or decreased your confidence in your regression analysis. (20 points)

Based on the ancillary analysis (shown above) confidence in the appropriateness of the multiple regression model has increased. For the submarine fans whose parameters are known the regression model provides a tolerably close prediction for all but one (Amazon). This is, perhaps explainable owing maturity and biotic productivity of the basin drained by that river. The Amazon also constitutes the only significant outlier in the plot of regression residual values.

Total = 1100 points