Data Analysis & Statistics for Earth Scientists

Nanjing University, Spring 2026

Lab 2 Assignment

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.

   - Plot these data. (10 points)

   - 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)
     a. Justify your selection. (20 points)

   - List the equation of the regression line for the model you have selected. (10 points)
     a. Use this equation to predict the Fe concentration at a depth of 300 meters. (20 points)

   - Use an ANOVA *F* test to estimate the significance of the regression model.
     a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)
     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)
     c. Estimate the 95% confidence interval for the regression result. (20 points)
        1. Estimate the range of variation in Fe concentration values that would be expected at depth of 300 meters. (30 points)

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

   - 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)
     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)

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.

   - Plot these data. (10 points)

   - Select a linear regression model that will allow the principle stress for this colliery spoil heap to be predicted. (10 points)
     a. Justify your selection. (20 points)

   - List the equation of the regression line for the model you have selected. (10 points)
     a. Use this equation to predict the angle of the principle stress estimate (in degrees). (20 points)

   - Use an ANOVA *F* test to estimate the significance of the regression model.
     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.
        1. Justify your reasoning. (30 points)

b. To two decimal places state the probability value associated with the ANOVA test result. (10 points)
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)
d. Estimate the 95% confidence interval for the regression result. (20 points)
   1. Estimate the range of principle stress angle variation that would be expected at this colliery. (30 points)

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

- 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)
  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)

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.

- Plot these data. (10 points)

- Select a linear regression model that will test the data for use in the estimation of eruption frequency. (10 points)
  a. Justify your selection. (20 points)

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

- Use an ANOVA *F* test to estimate the significance of the regression model.
  a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)
  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)
  c. Estimate the 95% confidence interval for the regression result. (20 points)
     1. Estimate the range of variation in eruption frequency estimates that would be expected based on your analysis. (30 points)

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

- 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)
  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)

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

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.

   - Plot these data. (10 points)

   - 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)
     a. Justify your selection. (20 points)

   - List the equation of the regression line for the model you have selected. (10 points)
     a. Use this equation to predict the taxonomic identities of the unknown skulls. (20 points)

   - Use an ANOVA *F* test to estimate the significance of the regression model.
     a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)
     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)
     c. Estimate the 95% confidence interval for the regression result. (20 points)
        1. Determine whether the range of variation in these predicted values has an affect of the certainty of your identifications. (30 points)

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

   - 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)
     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)

5. Many major rives 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.

   - Plot these data. (10 points)

   - 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)
     a. Justify your selection. (20 points)

   - List the equation of the regression line for the model you have selected. (10 points)
     a. Use this equation to estimate the discharge rates for the missing rivers based on their relict fans. (50 points)

   - Use an ANOVA test to estimate the significance of your regression model.
     a. To two decimal places state the probability value associated with the ANOVA test result. (10 points)
     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)
     c. Estimate the 95% confidence interval for the regression result. (20 points)

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

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

- 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)

Total = 1150 points