MATH 338

EXAM 3

TUESDAY, JULY 25, 2017

Your name: _____

Your scores (to be filled in by Dr. Wynne):

Problem 1: ____/8

Problem 2: ____/16

Problem 3: ____/8

Problem 4: ____/12

Total: ____/44

You have 80 minutes to complete this exam. This exam is open book, open notes, open help menus, and open labs.

For full credit, include all R code (if using RStudio), graphs, and output. Save your answers as a .docx or .pdf file and upload the file to Titanium.

Problem 1. Upload the dinosaur-bone-lengths.csv file to Rguroo, or import it into RStudio. This file contains the bone lengths (in cm) from the most complete known adult skeleton of 89 different dinosaur species, according to the Open Dinosaur Project. Because some skeletons were missing bones, there is quite a bit of missing data in this file, but the software can deal with it. Assume these 89 dinosaurs are a nonrandom, but nevertheless representative, sample of all dinosaurs.

In Problems 1-3 we will investigate the relationship between the lengths of two arm/forelimb bones, the humerus and radius.

A) [4 pts] Create a scatterplot of radius length (response) vs. humerus length (predictor). Give every Clade its own color and/or shape. Don't forget to give the plot appropriate title/axis labels. Paste the resulting scatterplot, as well as any R code, below.

Rguroo hint: You can use the default colors and shapes; no need to mess around in the *Level Editor*.

RStudio hint: The ggplot2 library doesn't work well with more than 6 shapes, so just use `color =` and don't include a shape argument.

B) [1.5 pts] Based only on your scatterplot, do you believe humerus length and radius length to be positively correlated, negatively correlated, or uncorrelated? Justify your answer.

C) [1.5 pts] Do you believe there is a causal relationship between humerus length and radius length? Why or why not?

D) [1 pt] The data appear to follow a linear relationship, except for dinosaurs in one Clade. Which clade?

Problem 2. Remove from the data set the clade that was your answer to problem 1D, and save the new data set.

Rguroo hint: In the Data section, use the *Subset* function -> *Logical Expression*, then choose or type *Clade != 'answer to part 1D'* (including the single quotation marks around your answer) in the appropriate boxes. Then when the new data set comes up in *View*, save it.

RStudio hint: Load the dplyr library, then use the `filter()` command to subset the data and save the new data set using, for instance, `dinosaur_new <- dinosaur %>% filter(Clade != 'answer to part 1D')` (including the single quotation marks around your answer).

Use the new data set to answer the following parts:

A) [3 pts] Report the equation of the least-squares regression line that best fits the relationship between humerus length (predictor) and radius length (response). Include the Parameter Estimates (Coefficients:) table and any relevant R code below.

B) [1.5 pts] In the space below, paste the following three plots (along with any R code used to create them). Label which plot is which.

1. A scatter plot of radius length vs. humerus length, with the least-squares regression line included

2. The residual plot

3. The normal quantile (q-q) plot of the residuals

C) [2 pts] Do the diagnostic plots in part 2B suggest that we are okay to perform inference using this model? Justify your answer by checking each assumption of the model.

D) [3 pts] Regardless of your answer to part 2C, report and interpret a 95% confidence interval for the true slope of the linear relationship between radius length and humerus length. Include all relevant (code and) output below.

E) [1 pt] What is the correlation between radius length and humerus length?

RStudio hint: if you attempt to get this using the `cor()` command, you will have to include the argument `use = 'pairwise.complete.obs'` to get around the missing data.

F) [2 pts] Predict the radius length of a dinosaur whose humerus has length 400 cm. If you use Rguroo or RStudio to make the prediction, include all relevant (code and) output below. Otherwise, show your work.

G) [1.5 pt] Was the prediction in part 2D an example of interpolation or extrapolation? Justify your answer.

H) [2 pts] Would a dinosaur with a humerus length of 400 cm and a radius length of 250 cm have high influence, a high residual, both, or neither? Justify your answer.

Problem 3. In the data set you used in Problem 2, add two new variables, log_radius and log_humerus, that represent the natural logarithm of radius length and humerus length, and save the new data set.

Rguroo hint: Use the *Transform* function, and in the formula box, type log(Radius) to get the natural logarithm of of the radius length and log(Humerus) to get the natural logarithm of the humerus length.

RStudio hint: Use the **`mutate()`** command in the dplyr package; for instance, **`dinosaur_log_rh <- dinosaur_new %>% mutate(log_radius = log(Radius), log_humerus = log(Humerus))`**

A) [3 pts] Report the equation of the least-squares regression line that best fits the relationship between natural logarithm of humerus length (predictor) and natural logarithm of radius length (response). Include the Parameter Estimates (Coefficients:) table and any relevant R code below.

B) [1 pt] In this **new** model, what is the value of our estimate of σ, the population standard deviation of the residuals?

C) [4 pts] Suppose that the skeleton of a new dinosaur species is unearthed. The humerus is measured to be 600 cm long, but the radius is not found. Using this **new** model, report and interpret a 95% prediction interval for the radius length of the new dinosaur. Include all relevant (code and) output below.

Hints: the value for your predictor variable is actually log(600), or about 6.39693. The inverse function of the natural logarithm is exponentiation with base *e* (**`exp()`** command in R).

Problem 4. Using the original data set (dinosaur-bone-lengths), create a multiple linear regression model, with Femur as the response and Humerus, Tibia, and Scapula as predictors (in that order). Don't include any interaction effects, just the three predictors (like in Lab 9).

A) [3 pts] Paste the Parameter Estimates (Coefficients:) table below, and use it to write out the full least-squares regression equation for this model. Include any relevant code used to make the table.

B) [2 pts] Interpret the slope corresponding to the variable Tibia.

C) [1.5 pts] Overall, is this model significant at the 1% significance level? How do you know? Paste below any output from Rguroo/RStudio that supports your answer.

D) [2 pts] Is there evidence of collinearity? If so, paste below an appropriate table or plot and explain how the table/plot shows that collinearity exists. If not, paste below an appropriate table or plot and explain how the table/plot shows that it does not exist.

RStudio hint: You may want to use the `select()` function to get only the variables in the model; for instance, `dinosaur_femur <- dinosaur %>% select(Femur, Humerus, Tibia, Scapula)`

E) [2 pts] If you were to perform backward selection using this initial model, which variable would you remove first? Why would you remove that variable?

F) [1.5 pts] If you were to perform backward selection using this initial model, would the next model you created have a higher or lower (Multiple) R^2 value? Explain your answer.