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

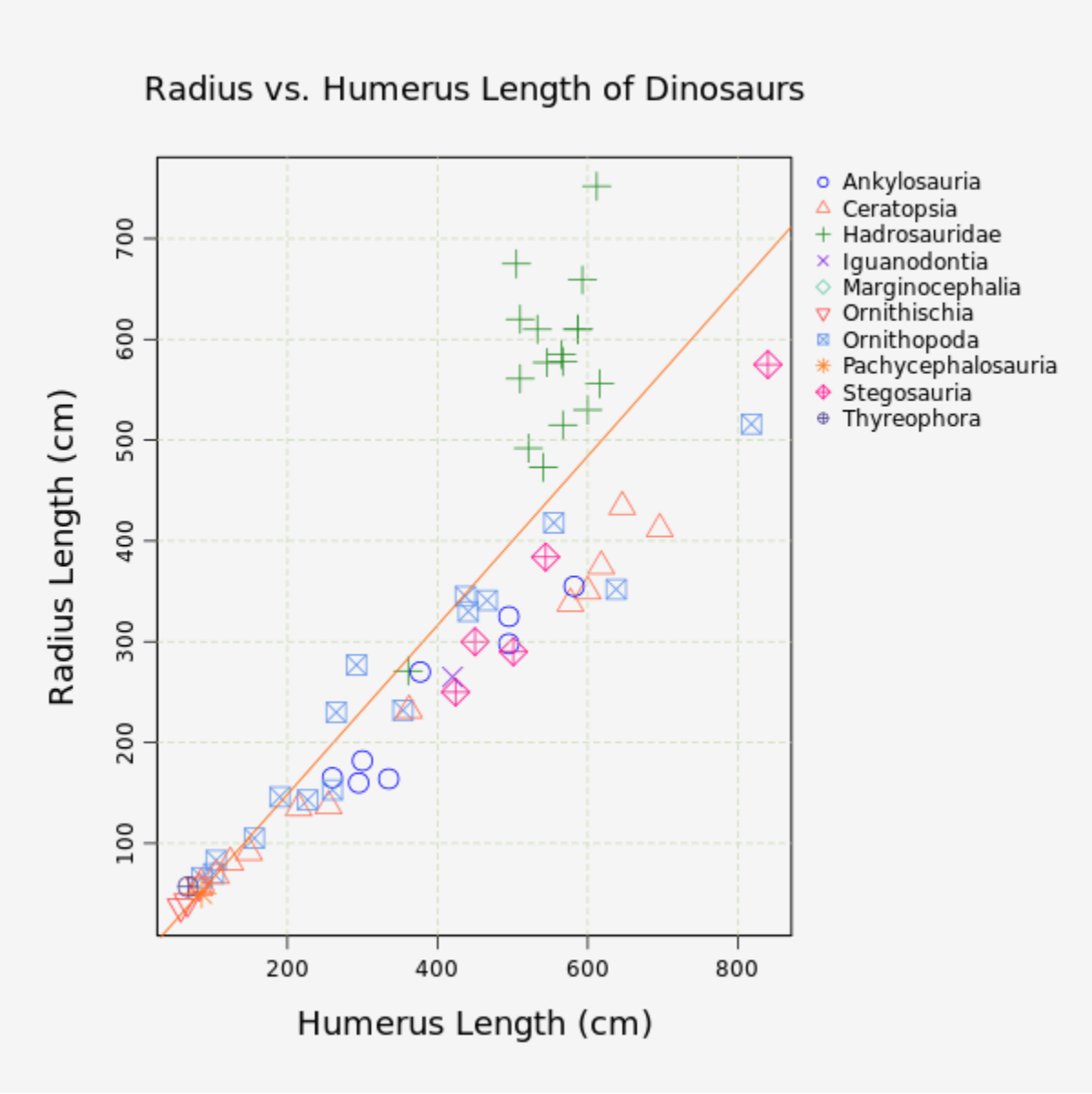
You have 80 minutes to complete this exam. This exam is open book, open notes, open help menus, and open labs.

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 (hint: you can just use the default colors/shapes). Don't forget to give the plot appropriate title/axis labels. Paste the resulting scatterplot, as well as any R code, below.



2 pt for scatterplot; 1 pt for correct axes and title labels; 1 pt for By Factor enabled

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.

0.5 pts positively correlated

1 pt as humerus length increases, radius length increases

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

0.5 pts no, not a causal relationship

1 pt there is not even a true explanatory-response relationship between these two variables, since it is not clear which bone length explains changes in the other bone length

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

1 pt Hadrosauridae

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 A'* (including the single quotation marks around your answer) in the appropriate boxes. Then when the new data set comes up in *View*, save it.

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.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	7.64857	8.28319	0.923384	0.360626
Humerus	0.629160	0.0203875	30.8600	2.17793e-32

Radius = 7.65 + 0.63 (Humerus)

1.5 pts parameter estimates table; 1.5 pts for the correct equation

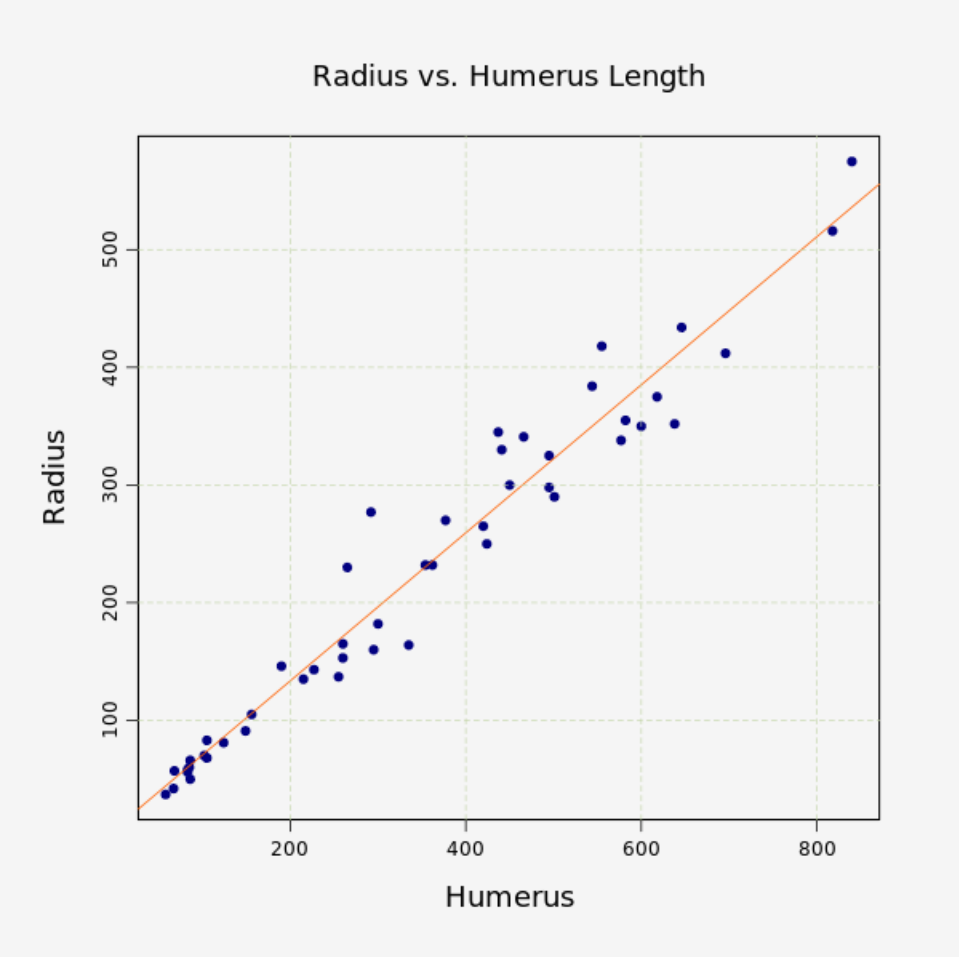
B) [1.5 pts] In the space below, paste the following three plots (along with any R code used to create them). Label which plot is which.

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

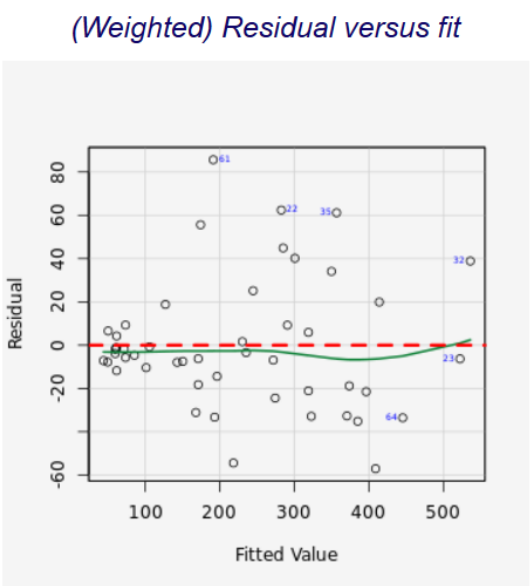
2. The residual plot

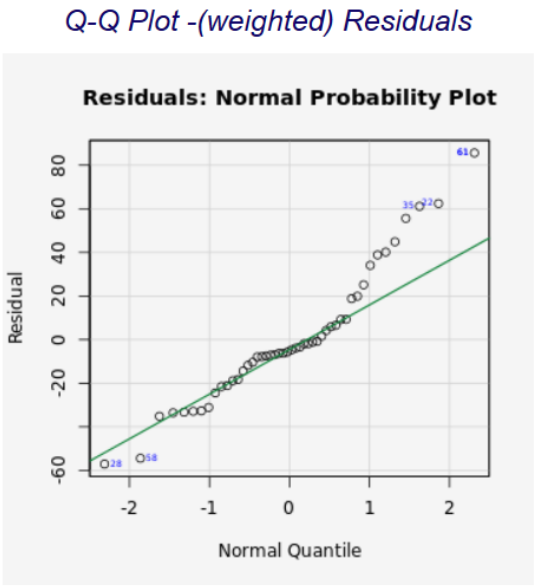
3. The normal quantile (q-q) plot of the residuals

0.5 pts per plot



Above: Scatter plot; Below left: Residual plot; Below right: Normal quantile plot





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.

0.5 pts a linear model is clearly appropriate

1.5 pts inference is not justified, because of one or more of the following reasons: nonzero mean of residuals (the green line never really gets above 0 until the very right of the graph), non-constant standard deviation (at very low fitted values all the residuals are very small), or non-normality of residuals (serious problems on the right side of the q-q plot). Full credit was given if all of these assumptions were checked, but none were considered problematic enough to make inference invalid.

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.

Parameter Confidence interval

Variable	Lower - 2.5 %	Upper - 97.5 %
(Intercept)	-9.02464	24.3218
Humerus	0.588121	0.670198

1 pt for the parameter confidence interval table above

1 pt the 95% confidence interval for slope is (0.588, 0.670)

1 pt We are 95% confident that for every 1 cm increase in radius length, the population mean humerus length increases by between 0.588 and 0.670 cm.

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

1 pt Since $R^2 = 0.953923$, and we have a positive relationship, $r = \text{sqrt}(R^2) = 0.977$

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.

1 pt plug in 400 cm correctly for humerus

1 pt radius length = 7.64857 + 0.629160(400) = 259.31 cm

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

0.5 pts interpolation

1 pt because 400 cm is between the minimum (34 cm) and maximum (840 cm) values of the predictor variable

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.

1 pt would not have a high influence because 400 cm is reasonably close to the mean of humerus length (324.91), so adding this point would not change the regression line all that much

1 pt would not have a high residual because the residual is 250 – 259.31 = -9.31, but according to the residual plot the residuals range from about -60 to 80 (-57.05 to 85.63 according to the diagnostics table)

Alternatively: full credit for pointing out that based on the scatter plot, the point (400, 250) more-or-less lies in the middle of the x-values and roughly along the regression line, so it would have neither a high influence nor a high residual

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 the radius length and log(Humerus) to get the natural logarithm of the humerus length.

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.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	-0.213593	0.135635	-1.57476	0.122164
log_humerus	0.963169	0.0240159	40.1055	1.94297e-37

log(radius) = -0.214 + 0.963(log(humerus))

1.5 pt for parameter estimates table; 1.5 pts for equation

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

(Adjusted) R-Squared

Residual Standard Error	DF	R-Squared	Adjusted R-squared
0.128102	46	0.972196	0.971592

1 pt our estimation of the population standard deviation of the residuals is s = 0.128

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() in R).

Diagnostics

Obs	log_radius	log_humerus	Predicted Values	Residuals	Std. Error Mean Predict	Lower Mean 2.5%	Upper Mean 97.5%	Lower Pred 2.5%	Upper Pred 97.5%
3	5.44674	5.86930	5.43953	0.00720320	0.0196285	5.40002	5.47904	5.17867	5.70040
4	5.85793	6.39693	5.94773	-0.0898007	0.0266985	5.89399	6.00148	5.68434	6.21113

1 pt You only needed to paste these first couple of rows since Obs 4 (the second row) gives us the prediction interval we want.

1 pt On the transformed scale, our prediction interval is (5.684, 6.211)

1 pt On the original scale, our prediction interval is (294.2, 498.3)

1 pt We are 95% confident that this new dinosaur will have a radius length between about 294 and 498 cm.

If you got (322.882, 447.407) as your interval then you used the wrong model and you got 3 pts if everything else was correct

If you got (362.85, 404.03) as your interval then you found a CI for the mean response instead of a prediction interval and you got 3 points if everything else was correct

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).

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.

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	-9.03088	24.6267	-0.366712	0.715632
Humerus	0.168409	0.184407	0.913243	0.366207
Tibia	0.386223	0.106293	3.63358	0.000740818
Scapula	0.711054	0.163641	4.34520	8.35153e-05

Femur = -9.03 + 0.168 (Humerus) + 0.386 (Tibia) + 0.711 (Scapula)

1.5 pts for table, 1.5 for equation

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

1 pt For every 1 cm increase in tibia length, we expect femur length to increase by 0.386 cm, …

1 pt … holding the values of Humerus and Scapula constant

C) [1.5 pts] Overall, is this model significant at the 1% significance level? Paste below any relevant tables or plots from the Rguroo/RStudio output that support your answer.

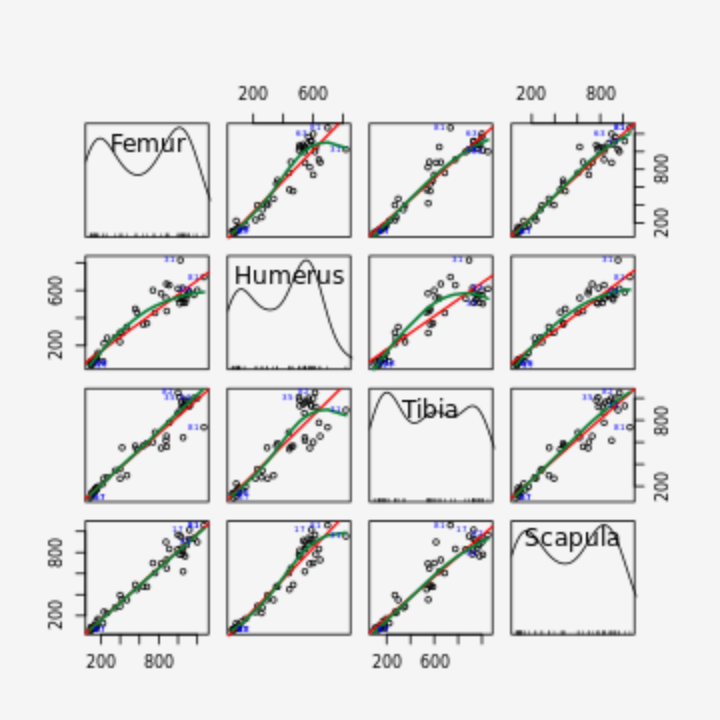
ANOVA Table

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Regression	3	7.21373e+06	2.40458e+06	430.510	4.40912e-32
Residual	43	240173	5585.41		
Total	46	7.45390e+06			

0.5 pts Yes

1 pt show the ANOVA table with p-value on the order of 10^{-32}

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.



1 pt show the scatterplot matrix

1 pt because there are clear linear relationships between Humerus-Tibia, Tibia-Scapula, and/or Humerus-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?

1 pt Humerus

1 pt Because it has the highest p-value, and therefore it's the least significant predictor

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.

0.5 pts lower R^2 value

1 pt either fit the next model and show that R^2 is lower, or explain that R^2 increases as variables are added to the model and therefore decreases as variables are removed from the model