ST 352

**Lab Assignment 5**

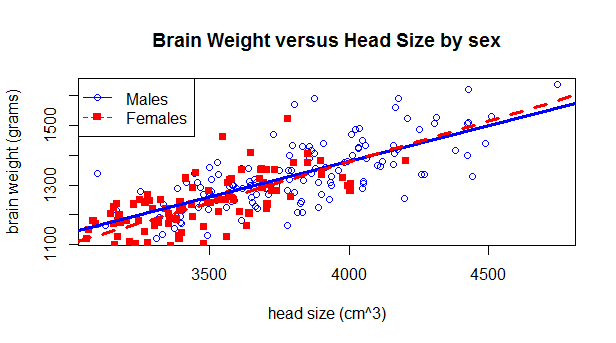
***24 points***

***Due 11:59 PM on Friday, November 8***

**Part I: 12 points total**

Use the graphs and output from Part I in the Lab 5 Notes (using the **brainhead**) data set to answer the questions 1 and 2.

1. ***(2 points)*** On pages 1 through 3 of the Lab 5 Notes, you constructed a scatterplot with different symbols for males and females, drew separate regression lines for the males and females, and included a legend. Include this scatterplot here.



***Deductions:***

***-1/2 point for no legend***

***-1/2 for not including regression lines***

***-1/2 for not using different symbols and/or colors for genders***

***-1/2 for no or improper labels***

***- partial points can be deducted (example: legend included but is not the right legend)***

2. ***(2 points)*** Based on the scatterplot from question 1, does it appear an interaction term is necessary to include in the model? Explain. (The answer to this question is based solely on the scatterplot and not on a hypothesis test.)

***Answers could go either way on this one. Without the regression lines, the relationship between head size and brain weight look very similar between males and females. When adding the regression lines, the lines aren’t perfectly parallel but are close. Look for support when grading this question.***

3. ***(4 points)*** Perform a hypothesis test to determine if the interaction term between *head size* and *gender* is significant. Make sure to 1) state the null and alternative hypotheses in words, 2) report the t-statistic with degrees of freedom, 3) report the p-value, and 4) state a conclusion in the context of the problem.

* ***Hypotheses: 1 point (1/2 point each)***
  + ***H0: the interaction term is not significant in the model***
  + ***HA: the interaction term is significant in the model***
* ***t-statistic (1/2 point), degrees of freedom (1/2 point), p-value (1/2 point):***
  + ***t-stat233 = -1.15, p-value = 0.2514***
* ***Conclusion: 1.5 points (1 point for correct adjective (“not enough” or “none”), ½ point for stating rest of conclusion. Note: if student says, “fail to reject null hypothesis” or “no evidence to say the alternative is true” and does not elaborate, take off 1 point.)***
  + ***There is not enough evidence to indicate the interaction term is significant in the model.***

4. ***(2 points)*** Regardless of your answer to question 3, leave the interaction term in the model. Write the least-squares regression equation. Define the terms in the equation in the context of the problem (i.e. what do , x1, and x2 represent in the context of the problem).

**= 286.0870 + 0.2728x1 +144.2157x2 -0.0354(x1\*x2)**

**where x1 = head size (cm3)**

**x2 = 1 for males and 0 for females**

**= predicted brain weight (grams)**

**1 point for equation**

**1 point for defining terms (Note: interaction term does not have to be defined since it is the product of x1 and x2, both of which have already been defined.)**

5. ***(2 points)*** Using the least-squares regression equation in question 4, interpret the coefficient of the interaction term in the context of the problem. Do so by using the template below.

***As head size increases by one 1 cm3, brain weight is predicted to decrease by 0.0354 grams more for males than females.***

***Also acceptable: As head size increases by one 1 cm3, brain weight is predicted to increase by 0.0354 grams more for females than males.***

***There may also be combinations may work: For example, “… brain weight is predicted to increase by 0.0354 grams less for males than females”, or “…brain weight is predicted to decrease by 0.0354 grams less for females than males” are also acceptable.***

***+1/2 for increasing “head size” by 1 cm3***

***+1/4 for using the response variable correctly***

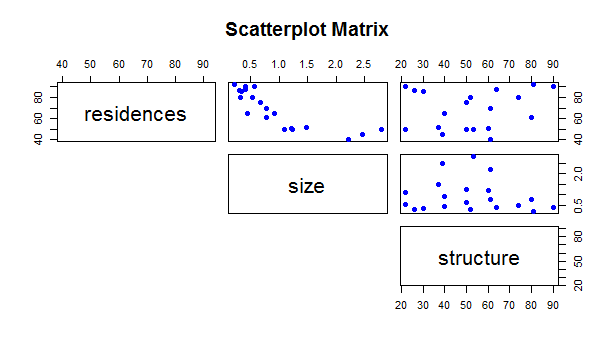
***+3/4 for including the idea of 0.0354 grams “more” for males than for females***

***+1/2 for using correct coefficient***

**Part II: 12 points total**

In Part II of the Lab 5 Notes, you obtained graphical displays to assess the conditions of the multiple linear regression model with all four explanatory variables.

6. ***(3 points)*** Include the scatterplot matrix of the quantitative explanatory variables and the correlation matrix here. Based on the scatterplot matrix and the correlation matrix, are any explanatory variables highly correlated? Briefly explain. (Your explanation should include support from both the scatterplot matrix and the correlation matrix.)



***There could be an argument that residences and size are highly correlated since 1) many of the points look “tightly-packed” in the scatterplot, and 2) the absolute value of the correlation coefficient is 0.83, which is getting close to the 0.9 mentioned in the “Steps for Multiple Regression” document. However, there are some points in the scatterplot that are not in the same cluster of “tightly-packed” points, and the absolute value of the correlation coefficient is still less than 0.9, so students may argue that the relationship is not strong enough to be considered “highly correlated.” None of the other explanatory variables are strongly related to each other. Look for support when grading.***

catch residences size access structure

catch 1.0000000 0.1491201 0.0428054 0.25614181 0.87534885

residences 0.1491201 1.0000000 -0.8286589 -0.39727354 0.16393700

size 0.0428054 -0.8286589 1.0000000 0.47569128 -0.11422505

access 0.2561418 -0.3972735 0.4756913 1.00000000 -0.06399828

structure 0.8753489 0.1639370 -0.1142250 -0.06399828 1.00000000

**Regardless of your answer to question 6, leave all quantitative explanatory variables in the model to assess the other conditions as the commission believes that residences and size may be important variables in predicting the seasonal catch.**

*Note: you are not asked to assess the outlier condition in this assignment. Even if you feel there is an outlier, leave it in to assess the remaining conditions.*

Go through the Lab 5 Notes and Lab 5 R script to obtain the scatterplot matrix that includes the response variable and *all* explanatory variables, the residual plot (residuals versus predicted values), and normal probability plot on the following models:

* Original scale (no transformations)
* Natural log transformation of *catch* only
* Natural log transformation of the *quantitative* explanatory variables only (ln(residences), ln(size), and ln(structure).
* Natural log transformation of all variables (except *access*)

On your own, assess the linearity, constant variation, and normality conditions on all four models.

7. ***(3 points)*** After assessing the conditions for all four models, which model best satisfies the conditions? Briefly explain your choice. While you may reference a number of plots form the different models when supporting your choice, **only include the scatterplot matrix, residual plot, and normality plot for the model you chose!**

***I have not included any graphs here. Look for support when grading. While there is not much difference in any of the models, the residual plot shows a little more curvature when the response variable is transformed (either catch only is transformed or catch and the quantitative variables are transformed). Between the original data and transforming only the quantitative explanatory variables, the conditions are quite similar. In such a situation, a strategy would be to choose the “simpler” model (i.e. the one with fewer transformations), which would be the original scale. Again, though, answers can vary – look for support.***

***“Support” should include discussion about the linearity, constant variation, and normality conditions. Their support should be what it is about the plots that led them to choose the model they chose.***

***+ 1 for discussing each condition***

Obtain the regression output from model that best satisfies the conditions. Use the output from that model to answer the questions 8 through 10.

***The answer to these questions should be based on output from the model that they chose in question 7. If students use a different model to answer these questions, take off 2 points (total, not per problem!). I will provide answers for what I considered the two choices of models: 1) original data, and 2) only quantitative explanatory variables transformed. I provided output for the other models below just in case a student chose one of these models.***

8. ***(2 points)*** Write the least-squares regression equation. Define the terms in the equation.

***Original scale: = - 2.7840 + 0.0268x1 + 0.5035x2 + 0.0511x3 + 0.7429x4***

***exp vars transformed: = -16.5388 + 2.3007ln(x1) + 0.5740ln(x2) + 2.4000ln(x3) + 0.6913x4***

***Where x1 = # lakeshore residences per square mile***

***x2 = size of lake in square miles***

***x3 = structure index***

***x4 = 1 if lake had public access, and 0 if lake did not have public access***

***+1 for correct equation (including have “ln” if transformed the explanatory variables)***

***+1 for defining the terms in the equation***

9. ***(2 points)*** Predict the seasonal catch of a 1.25 square mile lake with public access, had 70 lakeshore residences per square mile of lake area, and had a structure index of 45. Use the **predict()** command in R.

***Original data: 2.765 thousand (or 2,765)***

***Transformed explanatory variables model: 3.191 thousand (or 3,191)***

10. ***(2 points)*** Interpret the coefficient of *access* in the context of the problem. (If your model contained transformed *explanatory* variables, obtain a model without a transformation of the *explanatory* variables to answer this question.)

***For this one, students should use either the model with the original data (if they originally chose the model with the explanatory variables transformed), or the model with only the response variable transformed.***

***Original data:***

***For lakes the same size, same structure index, and same number of lakeshore residences, the seasonal bass caught is predicted to be 0.7429 thousand (or 742.9) more for public access lakes compared to non-public access lakes.***

***Model with catch transformed:***

***For lakes the same size, same structure index, and same number of lakeshore residences, the median seasonal bass caught is predicted to be e0.291726 = 1.3387 times more for public access lakes compared to non-public access lakes.***

***+1/2 for keeping the other variables the same***

***+1 for correctly using the coefficient in the interpretation (including back transforming the coefficient if using the model with catch transformed) AND INCLUDING THE WORKD “PREDICTED” OR “EXPECTED”– take off ¼ point for not including an indication that this is what we expect or predict to happen as opposed to what will happen.***

***+1/2 for indicating it would be more for public access lakes than non-public access lakes (including using the word “times” more if using the transformed model)***

**Other models:**

Call:

lm(formula = log(catch) ~ residences + size + structure + access,

data = fish)

Residuals:

Min 1Q Median 3Q Max

-0.31029 -0.12029 0.00497 0.14028 0.31527

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.130217 0.446701 -2.530 0.0231

residences 0.006585 0.005006 1.315 0.2081

size 0.194501 0.120898 1.609 0.1285

structure 0.022620 0.002487 9.095 1.72e-07

access 0.291726 0.110690 2.636 0.0187

Residual standard error: 0.2133 on 15 degrees of freedom

Multiple R-squared: 0.8663, Adjusted R-squared: 0.8306

F-statistic: 24.29 on 4 and 15 DF, p-value: 2.096e-06

Call:

lm(formula = log(catch) ~ log(residences) + log(size) + log(structure) +

access, data = fish)

Residuals:

Min 1Q Median 3Q Max

-0.30922 -0.05861 -0.00630 0.06515 0.33572

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.35568 1.54448 -4.115 0.000918

log(residences) 0.67181 0.37110 1.810 0.090324

log(size) 0.22036 0.14557 1.514 0.150856

log(structure) 1.11068 0.09442 11.763 5.67e-09

access 0.27524 0.09626 2.859 0.011941

Residual standard error: 0.173 on 15 degrees of freedom

Multiple R-squared: 0.9121, Adjusted R-squared: 0.8886

F-statistic: 38.89 on 4 and 15 DF, p-value: 9.465e-08