**Name:**

**Total: /100 points**

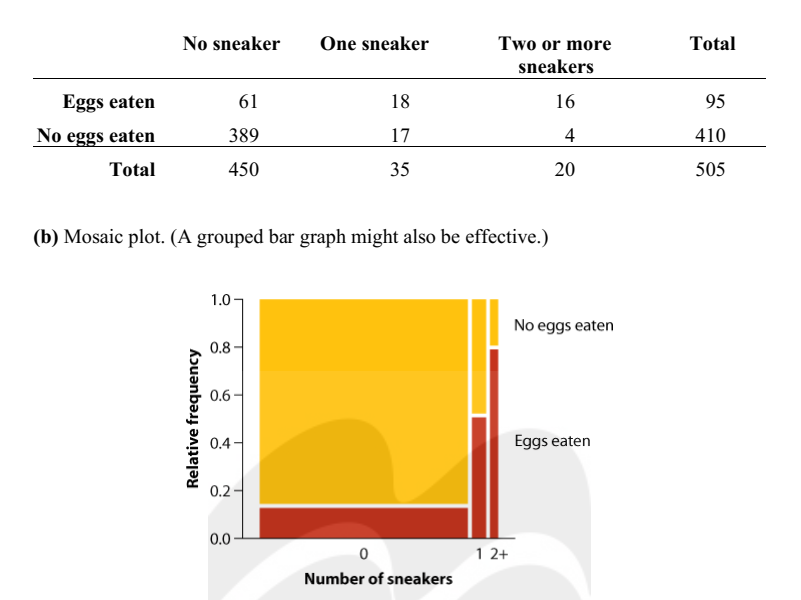
**Instructions:** Please complete the following questions and type your answers into this document. You may insert figures and/or screenshots of outputs from R where requested. Please complete questions requiring computations using R and submit your R code as a separate R file when you submit your assignment. Note that this is an open book exam so you may use any materials from the course, online, or in print to complete the answers. You **may** consult with others in the class about answers, however your answers must be formulated in your own words and you must hand in your own assignment. Please upload a completed version of your exam by **Tuesday, December 14th at 5:00 pm** in Blackboard.

1. **(10 Total; 2 pts each)** For each of the following scenarios, indicate the type of statistical test that you would choose to use.

|  |  |
| --- | --- |
| A single categorical independent variable with three categories and a continuous dependent variable that is non-normally distributed |  |
| A singe continuous dependent variable with values ranging from 0 to infinity and with higher variance among larger observations than smaller observations. Two continuous independent variables. |  |
| Comparison of two related means for normally distributed data |  |
| You would like to evaluate the effects of three independent variables (one categorical, 2 continuous) on a single categorical dependent variable with two categories |  |
| Measuring the interdependency between two continuous, normally distributed variables |  |

2. **(6 Total; 3 pts each)** When a courting male of the small Indonesian fish *Telmantherina sarasinorum* spawns with a female, other males sometimes sneak in and release sperm, too. The result is that not all of the female’s eggs are fertilized by the courting male. Gray et al. (2007) noticed that courting males occasionally cannibalize fertilized eggs immediately after spawning. Egg eating took place by 61 of 450 courting males who fathered the entire batch; the remaining 389 males did not cannibalize eggs. In contrast, 18 of 35 courting males ate eggs when a single sneaking male also participated in the spawning event. Finally, 16 of 20 males ate eggs when two more sneaking males were present. The raw data are available in the file “SneakerCannibalism.csv”.

1. Display the results in a table that best shows the association between cannibalism and the number of sneaking males. Identify the type of table you used.
2. Illustrate the same results using a graphical technique. Identify the type of graph you used.



A graph shows frequency distribution of similarity values between equivalent proteins in humans and puffer fish.
The horizontal axis is labeled Similarity percent, ranging from 0 to 100 with increments of 5. The vertical axis is labeled Number of proteins, ranging from 0 to 2700 with increments of 250. The approximate data are as follows. 6, 60; 11, 300; 16, 700; 21, 850; 26, 900; 31, 1000; 36, 1250; 41, 1359; 46, 1650; 51, 1750; 56, 1800; 61, 2000; 66, 2200; 71, 2200; 76, 2100; 81, 1900; 86, 1700; 91, 1300; 96, 950; 101, 500.3. **(4 Total; 1 pt each)** Examine the following figure, which displays the frequency distribution of similarity values (the percentage of amino acids that are the same) between equivalent (homologous) proteins in humans and pufferfish of the genus *Fugu* (data from Aparicio et al. 2002)

a) What type of graph is this?

Histogram

b) Identify the main flaw in the construction of the figure.

Bars should not have gaps between them

1. Describe the shape of the frequency distribution.

Skewed left

1. What is the mode of the frequency distribution?

70% similarity (presumably the interval number ‘70’ represents is 67.5-72.5)

4. **(6 Total; 2 pts each)** Hagan et al. (2011) estimated the home range sizes of four bumblebees (*Bombus*) by fitting them with tiny radio transmitters and tracking their positions by plane and ground surveys. The home range sizes for the bees (in ha) were: 27.69, 43.53, 0.25, and 1.37

a) What is the mean and standard error of the home range sizes?

b) Interpret the standard error of this estimate of the mean.

c) What would you recommend the researchers do next to reduce the standard error of their estimate of mean home range size?

5. **(3 pts)** Pitcher plants of the genus *Nepenthes* are typically carnivorous, obtaining a great deal of their nutrition from insects that become trapped in the pitcher, die, and decay. *N. lowii*, a pitcher plant from Borneo, produces a second type of pitcher that attracts tree shrews (*Tupaia montana*), which provide nutrients by defecating into the pitcher while they feed on a substance secreted by the plant. Based on measurements of 20 plants, Clarke et al. (2009) calculated a 95% confidence interval for the mean fraction of total leaf nitrogen in the plant species derived from tree shrews: 0.57 < μ < 1.0. Does this result imply that individual plants receive between 57% and 100% of their leaf nitrogen from tree shrews? Explain.

6. **(8 Total; 2 pts each)** Charles Darwin carried out an experiment to study whether seedlings from cross-fertilized plants tend to be superior to those from self-fertilized plants. He covered a number of plants with fine netting so that insects would be unable to fertilize them. He fertilized a number of flowers on each plant with their own pollen and he fertilized an equal number of flowers on the same plant with pollen from a distant plant. (He did not say how he decided which flowers received which treatments.) The seeds from the flowers were allowed to ripen and were set in sand to germinate. He placed two seedlings of the same age in a pot, one from a seed from a self-fertilized flower and one from a seed from a cross-fertilized flower. The data in the file “DarwinFlowers.csv” contain the heights of the plants at a certain point in time.

1. Examine the differences in height using a histogram

Chart, histogram

Description automatically generated

(b) State your biological and statistical hypotheses

H0: There is no treatment effect, i.e. the mean difference is zero; mu(Diff) = 0

H1: There is a treatment effect, i.e. the mean difference is NOT zero; mu(Diff) != 0

(c) Apply an appropriate test to determine whether there is a difference in heights between the pairs.

tt <- t.test(Diff)

t = 2.148, df = 14, p-value = 0.0497

tt$conf.int

0.003899165 5.229434169

# Alternatively, we can work with the original (paired)

# data and do a paired t-test:

t.test(Cross, Self, paired=TRUE)

# t = 2.148, df = 14, p-value = 0.0497

(d) State the statistical results and your biological interpretation of the results.

# Thus we find:

# P-value = 0.0497 (t-statistic = 2.148 with 14 df)

# 95% CI = (0.004, 5.229) for the mean difference

# (cross-fert. - self-fert.)

# The results constitute moderate evidence that there

# IS a treatment effect and that the cross-fertilized

# plants are, on average, taller.

7. **(12 Total)** The data file “LodgepolePineCones.csv” contains mean cone size (mass) of lodgepole pine in 16 sites in three types of environment in western North America (Edelaar and Benkman 2006). The three environments were islands of lodgepole pines in which pine squirrels were absent (an “island” refers to a patch of lodgepole pine surrounded by other habitat and separated from the large tracts of contiguous lodgepole pine forests), islands with squirrels present, and sites within the large areas of extensive lodgepole pines (“mainland”) that all have squirrels.

(a) (2) What test would be appropriate for comparing mean cone size among these three habitat types?

ANOVA

(b) (2) State assumptions of your test of choice. (Note the data set is too small to assess normality, so you may proceed as if the data are normally distributed).

-measurements in every group represent a random sample from the corresponding population

-the variable is normally distributed in each of k populations

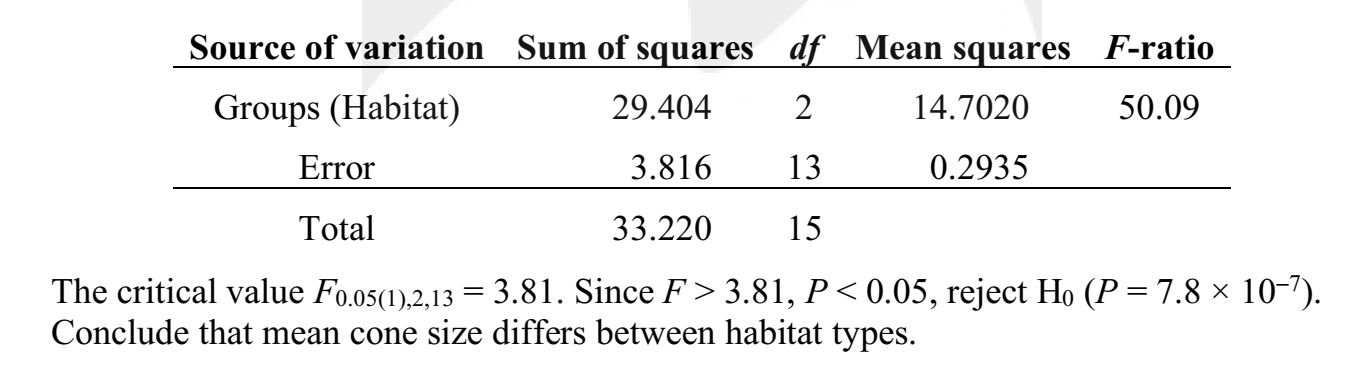
-the variance is the same in all k populations

(c) (2) State the statistical and biological hypotheses.

H0: Habitat types do not differ in mean cone size (µ1 = µ2 = µ3). HA: Habitat types

differ in mean cone size (at least one µi is different).

(d) (4) Conduct the test and interpret the results.



(e) (2) Determine which habitats differ in mean cone size using a post-hoc test.

> pineANOVA <- lm(conemass ~ habitat, data = Cones)

>

> # After running the linear model, we will now cast the results into an ANOVA table that we are familiar with

> anova(pineANOVA)

Analysis of Variance Table

Response: conemass

Df Sum Sq Mean Sq F value Pr(>F)

habitat 2 29.404 14.7020 50.085 7.787e-07 \*\*\*

Residuals 13 3.816 0.2935

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(aov(pineANOVA))

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = pineANOVA)

$habitat

diff lwr upr p adj

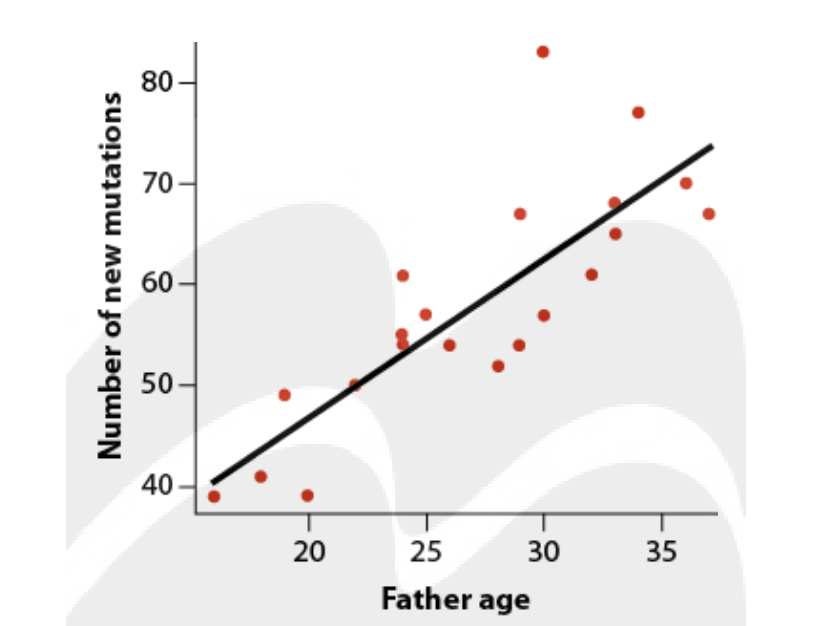
island.present-island.absent -2.82 -3.6862516 -1.9537484 0.0000028

mainland.present-island.absent -2.78 -3.6462516 -1.9137484 0.0000033

mainland.present-island.present 0.04 -0.8647703 0.9447703 0.9925198

8. **(14 Total)** Dads transmit many more new mutations than do mothers to their babies at conception. These mutations occur from copying errors during sperm production. There is increasing interest in the effect of father age on this process. As part of a larger study into the genetics of mental illness, Kong et al (2012) used complete-genome sequencing of 21 father-child pairs to tally the total number of new mutations inherited from each father (in this particular sample, all the offspring were affected with schizophrenia). These counts are listed in the data file “FatherAgeMutations.csv” along with the fathers’ ages at offspring conception. You may assume counts are approximately continuously distributed.

(a) (4) Graph the relationship between number of new mutations (Y) and father’s age (X). Add the regression line to your plot.



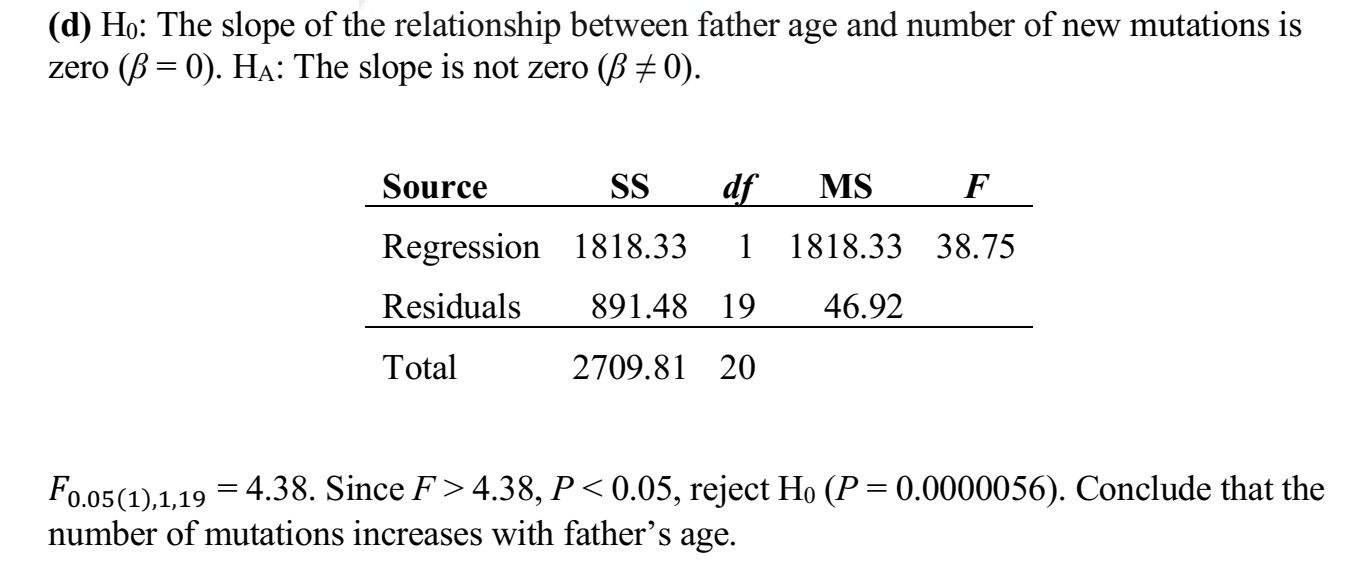
(b) (2) Based on these data, how rapidly does the number of new mutations increase with father’s age? Provide a standard error for your estimate.

B=1.583, SE=0.254

(c) (2) What is the predicted mean number of new mutations from fathers 36 years of age? How does this compare with the predicted number for fathers only 18 years old?

36 years: 72.19 mutations. 18 years: 43.70 mutations. Difference between 36 and 18 is about 28 new mutations.

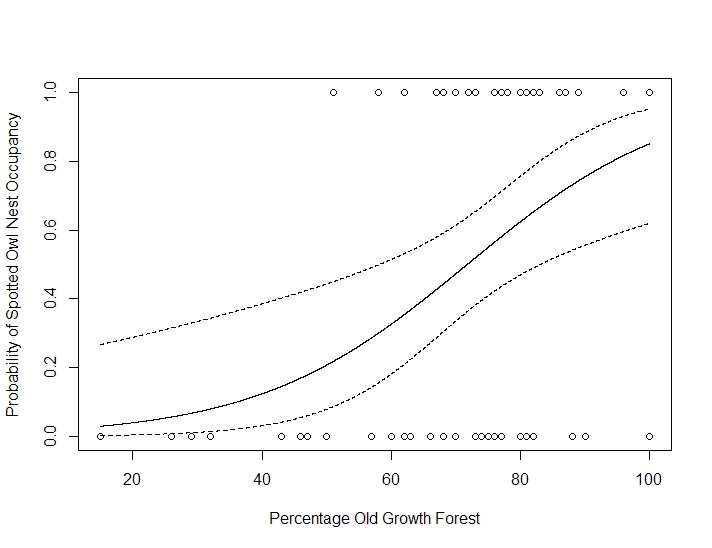
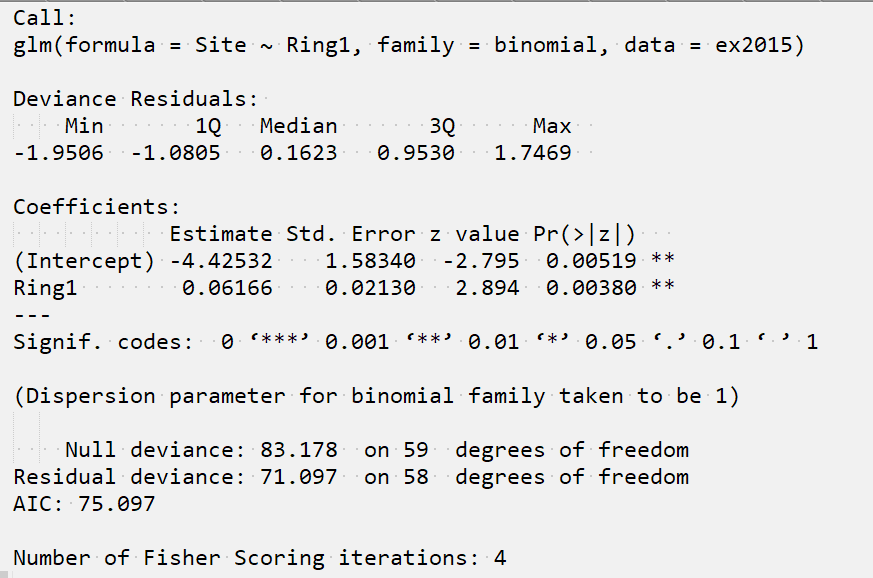
(d) (4) State your statistical and biological hypotheses. Use the ANOVA approach to test the null hypothesis of no relationship between father’s age and number of new mutations. Include an ANOVA table with your results.



(e) (2) What fraction of the variation among fathers in the number of new mutations is explained by father’s age?

R-squared = 0.67. 67% of the variance among fathers in the number of new mutations is explained by age.

9. **(9 Total)** A study examined the association between nesting locations of the Northern Spotted Owl and the availability of mature forests. Wildlife biologists identified 30 nest sites (Data from W.J. Ripple et al., “Old-growth and mature forests near spotted owl nests in Western Oregon, Journal of Wildlife Management 55: 316-18.) The researchers selected 30 other sites at random coordinates in the same forest. On the basis of aerial photography, the percentage of mature forest (older than 80 years) was measured in a ring with an outer radius of 0.91 km. Consider the following output from a logistic regression used to examine the relationship between whether spotted owl use a nest (occupancy = 1) or not (occupancy=0) as a function of percent old growth habitat in the ring. The resulting regression is graphed below the output.



1. (2) State the statistical and biological hypotheses being tested.

H0: The odds of a nest site being occupied are not related to the percentage old growth forest. (b=0)

H1: The odds of a site being occupied increase with percent old growth forest (H1: b!=0)

1. (2) What is your conclusion about the overall regression?

The small P-value implies that mature forest is significantly more available in the nesting site regions than in the control (random) regions.

1. (2) What is your interpretation of the effect that percent old growth habitat has on the probability of a site being occupied by spotted owls (again, be specific)?

The odds of a site being occupied increase by 1.06 times for every percent increase in old growth

1. (3) Imagine the researchers were also interested in evaluating whether a greater proportion of old growth surrounding a nest influences the stress hormone levels of nestlings. They visit each of the nests and measure cortisol in the feathers of all nestlings within each nest. Explain how this change in the sampling design would alter the modelling approach that would be appropriate for these data. Write a model (using word statements) to demonstrate the difference. Assume that cortisol is the response variable and %Old Growth Forest is the predictor variable of interest.

Cortisol = % old growth + 1|Nest

10. **(16 Total)** Both echolocation and flight are energetically costly. Zoologists therefore wondered whether the combined energy costs of echolocation and flight in bats was the sum of the flight energy costs and the at-rest energy echolocation costs, or whether the bats had developed a means of echolocation in flight that made the combined energy cost less than the sum. To answer this question, they examined the in-flight energy expenditure and body mass from 20 energy studies on three types of flying vertebrates: non-echolocating bats (Type 1), non-echolocating birds (Type 2), and echolocating bats (Type 3). They believed that if the combined energy expenditure for flight and echolocation were additive, the amount of energy expenditure (after accounting for body mass in g) would be greater for echolocating bats than for non-echolocating bats and non-echolocating birds. (Data from Speakman and Racey. 1991. Nature 350: 421-423.

(a) (4) Inspect the data in the file “Ecolocation.csv”. Apply a log transformation on the flight energy expenditure variable (Energy) and to the Mass variable (Mass) and show your histograms for the two variables below. (proceed as if log Energy data are normally distributed)

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

(b) (2) Write a word statement of the full model with an interaction term.

Energy = Type + Mass + Type\*Mass

(c) (4) Fit a model that includes an interaction term between logMass and Type. Show your output table below. What can you conclude about the interaction between logMass and Type?

Call:

lm(formula = logEnergy ~ logMass + Type + logMass \* Type, data = CASE1002)

Residuals:

Min 1Q Median 3Q Max

-0.25152 -0.12643 -0.00954 0.08124 0.32840

Coefficients:

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

(Intercept) -1.58084 0.29442 -5.369 9.90e-05 \*\*\*

logMass 0.83537 0.05532 15.100 4.66e-10 \*\*\*

Typeebats 0.11032 0.38474 0.287 0.779

Typenbats 1.37839 1.29524 1.064 0.305

logMass:Typeebats -0.03071 0.10283 -0.299 0.770

logMass:Typenbats -0.24559 0.21343 -1.151 0.269

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1899 on 14 degrees of freedom

Multiple R-squared: 0.9832, Adjusted R-squared: 0.9771

F-statistic: 163.4 on 5 and 14 DF, p-value: 6.696e-12

Interaction term is not significant.

1. (4) Fit a reduced model without an interaction term. Show your output table below. Compare the full model with the reduced model using an anova and interpret the results.

Call:

lm(formula = logEnergy ~ logMass + Type, data = CASE1002)

Residuals:

Min 1Q Median 3Q Max

-0.23224 -0.12199 -0.03637 0.12574 0.34457

Coefficients:

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

(Intercept) -1.47410 0.23902 -6.167 1.35e-05 \*\*\*

logMass 0.81496 0.04454 18.297 3.76e-12 \*\*\*

Typeebats -0.02360 0.15760 -0.150 0.883

Typenbats -0.10226 0.11418 -0.896 0.384

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.186 on 16 degrees of freedom

Multiple R-squared: 0.9815, Adjusted R-squared: 0.9781

F-statistic: 283.6 on 3 and 16 DF, p-value: 4.464e-14

> anova(energy2, energy)

Analysis of Variance Table

Model 1: logEnergy ~ logMass + Type

Model 2: logEnergy ~ logMass + Type + logMass \* Type

Res.Df RSS Df Sum of Sq F Pr(>F)

1 16 0.55332

2 14 0.50487 2 0.04845 0.6718 0.5265

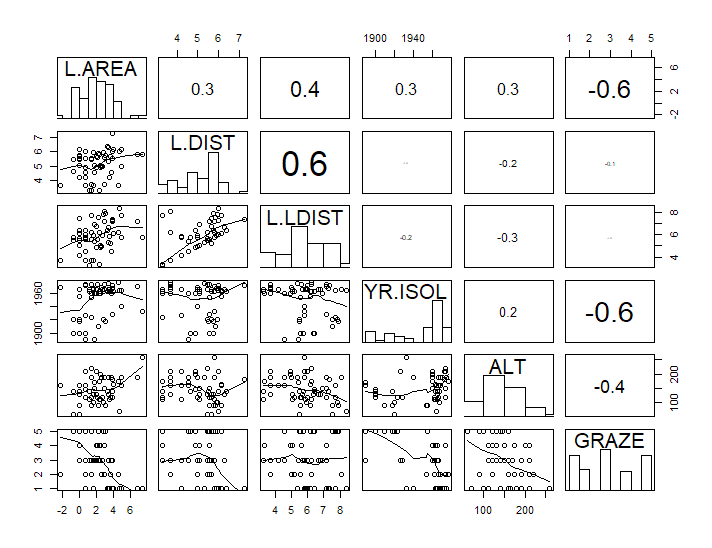
1. (2) What can you conclude about the effect of ‘Type’ on Energy?

No evidence of differences in energy expenditure among types of vetebrates after accounting for body mass.

11. **(4)** Forest bird densities were measured in 56 forest patches in south-eastern Victoria, Australia. The aim of the study was to relate bird densities (ABUND) to six habitat variables: size of the forest patch (AREA), distance to the nearest patch (DIST), distance to the nearest larger patch (LDIST), mean altitude of the patch (ALT), year of isolation by clearing (YR.ISOL), and an index of stock grazing history (GRAZE; 1= light, 5 = intensive). You are interested in running a multiple regression to predict bird densities from these variables. As one of your first steps in the analysis, check for multicollinearity. The data are in the file “Loyn.csv”. Are there any variables that you need to drop from the model?

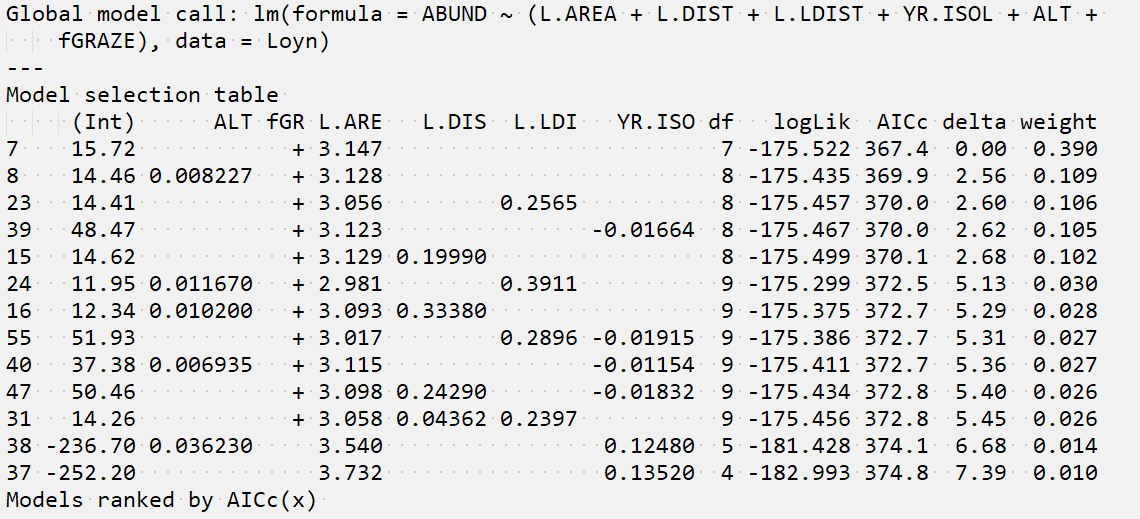
Yes. L.Dist and L.LDist are collinear (but they are the same variable; L.LDist is the log of L.Dist, so you wouldn’t include both in the same model)

L.AREA and Graze, YR.ISOL and GRAZE –possibly a concern, but not too bad



12. **(8 Total)** The following table shows the top 95% candidate model set for the multiple regression model predicting bird abundance using the variables described in Question 9. The variable GRAZE has been entered as a factor (fGRAZE in the model and fGR in the table). The variables AREA, DIST, and LDIST have been log transformed, as indicated by the L. pre-fix in the variable name.

(a) (4) What variables are included in the top model? Interpret the weight of the top model.



fGraze and L. Area are in the top model. There is a 39% chance that this model is the best model in the candidate model set.

(b) (4) Using the model output table below for the top model in the candidate model set, interpret the regression coefficients for fGRAZE5 and L.AREA.

Average bird abundance is 11.9 birds/unit area lower for graze level 5 compared with graze level 1.

Bird abundance increases by 3.1 birds/unit area for every unit increase in L.AREA.

