Take-Home Final

62 points

Yufu Yoshimura, Lydia Diehl, Nitasha Fazal, Grayson Tymchyshyn

Due May 11 at 11:00 PM in Gradescope

Dinosaur Clutch Volume

Directions

Answer each of the following questions in complete sentences. Any questions that require the use of `R` should include the code and output with your answer. Bold your answers for easy reading (**two asterisks at beginning of answer and two at the end**), equations do not need to be bolded. Points will be deducted for sentence fragments and egregious spelling/grammatical errors. Completed homeworks must be submitted as a PDF to Gradescope **with questions assigned to page numbers.**

This exam may be completed individually or in groups, though it is suggested you work and submit in groups (of no more than four people). If you choose to submit this exam individually, it is expected that **all work is your own** and there has been no collaboration with others. Everyone is encouraged to seek help from me if there are any questions about content or code.

Unauthorized collaboration or use of materials beyond those provided in this course will result in a zero on this exam and a report to the Dean of Students.

Background

Paleobiologists are interested in understanding factors that may be related to egg laying and care for archosaurs. Over four hundred species have been included in this study that measures that following variables.

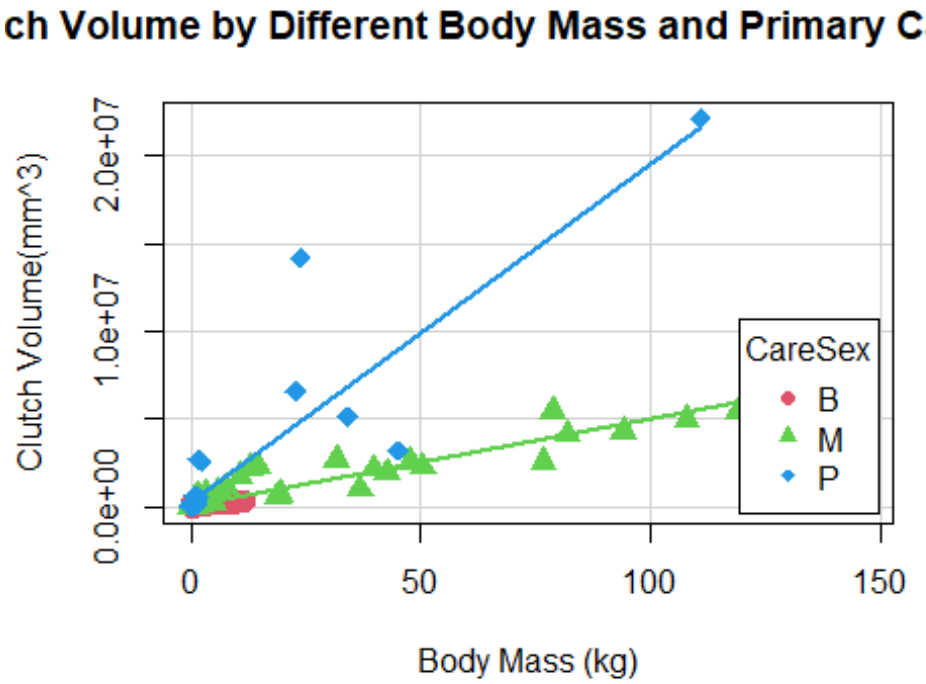
- `BodyMass`: Body mass (in kg)
- `ClutchVolume`: Volume of eggs in a single clutch (mm^3)
- `CareSex`: Sex of primary parent to care for eggs (B: biparental, M: maternal, P: paternal)
- `HigherTaxon`: Higher taxon classification the archosaur (A: Avian, C: Crocodilian)

We will use these data to estimate a model for clutch volume.

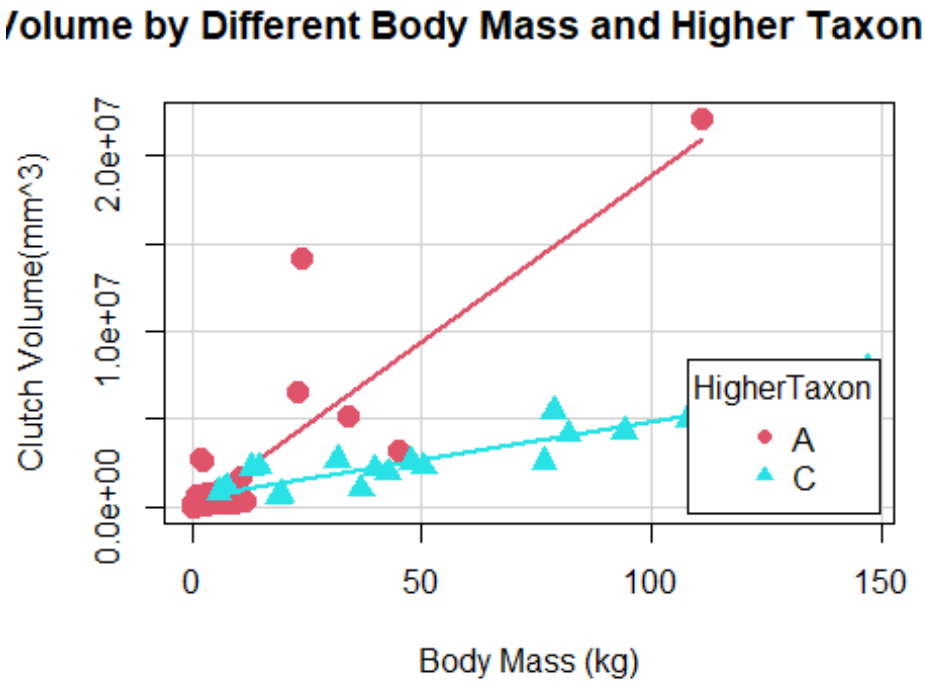
Data courtesy of: *Varricchio DJ, Moore JR, Erickson GM, Norell MA, Jackson FD, Borkowski JJ. Avian paternal care had dinosaur origin. Science. 2008 Dec 19;322(5909):1826-8. doi: 10.1126/science.1163245. Erratum in: Science. 2009 Aug 7;325(5941):676. PMID: 19095938.*

1. (10 pts) Create two summary graphics for these data that summarize the relationship between clutch volume and body mass: one with the sex of the primary care parent and the second with the higher taxon classification included. Provide appropriate titles, axis labels, and legends for each plot.

```
library(readr)
library(car)
library(mosaic)
library(psych)
dino<-read_csv("dino.csv")
scatterplot(ClutchVolume~BodyMass+CareSex, data=dino,
            regLine=TRUE,
            smooth=FALSE,
            col=c(10:12),
            pch=c(16,17,18),
            cex=rep(1.5,3),
            main="Clutch Volume by Different Body Mass and Primary Care
Parent.",
            xlab="Body Mass (kg)",
            ylab="Clutch Volume(mm^3)",
            legend=list(coords="bottomright"))
```



```
scatterplot(ClutchVolume~BodyMass+HigherTaxon, data=dino,
            regLine=TRUE,
            smooth=FALSE,
            col=c(2,5),
            pch=c(16,17),
            cex=rep(1.5,3),
            main="Clutch Volume by Different Body Mass and Higher Taxon
Classification.",
            xlab="Body Mass (kg)",
            ylab="Clutch Volume(mm^3)",
            legend=list(coords="bottomright"))
```



2. (6 pts) Discuss the potential for interactions between body mass and each categorical explanatory variable in this study. Reference specific features of your plots in your discussions.

Because the slope for the relationship between body mass and clutch volume varies between the various primary care parent groups and higher taxon groups, an interaction between both categorical variables and the explanatory should be considered.

3. (5 pts) Fit a multiple regression model with both two-way interactions discussed in the previous question, regardless of what the scatterplots suggested. Provide a model summary and identify which groups, if any, have been selected as baseline groups for your fitted model.

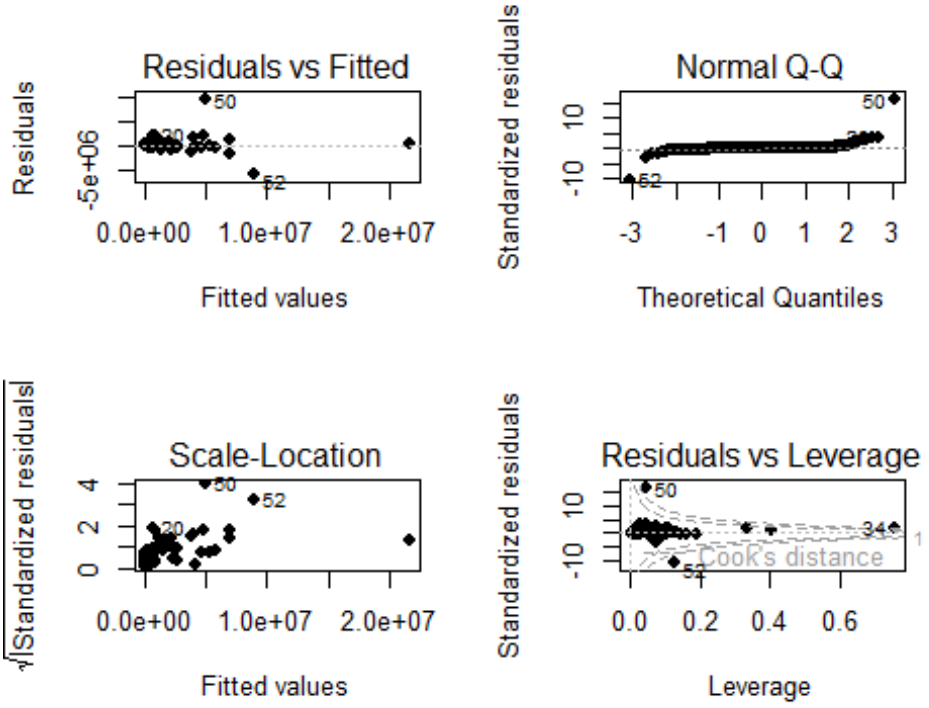
```
lmdino<-lm(ClutchVolume~BodyMass*CareSex+BodyMass*HigherTaxon, data=dino)
summary(lmdino)
```

```
##
## Call:
## lm(formula = ClutchVolume ~ BodyMass * CareSex + BodyMass * HigherTaxon,
##     data = dino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5773711   -42212   -12429    14341  9273363
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                51063      45473   1.123   0.2621
## BodyMass                   37070      21401   1.732   0.0840
## CareSexM                  -32966      66969  -0.492   0.6228
## CareSexP                  290524     112729   2.577   0.0103
## HigherTaxonC              512288     234608   2.184   0.0295
## BodyMass:CareSexM          95342      42538   2.241   0.0255
## BodyMass:CareSexP         154773      21946   7.052 7.22e-12
## BodyMass:HigherTaxonC     -88867      36918  -2.407   0.0165
##
## Residual standard error: 593400 on 423 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.8446
## F-statistic: 334.9 on 7 and 423 DF,  p-value: < 2.2e-16
```

Biparental is the baseline for primary care parent and Avian is the baseline for Higher taxonomy.

4. (10 pts) Create a set of residual diagnostics for these data. Assess all assumptions associated with this set of residual diagnostics, citing appropriate evidence as necessary.

```
par(mfrow=c(2,2))
plot(lmdino,pch=16,add.smooth=FALSE)
```



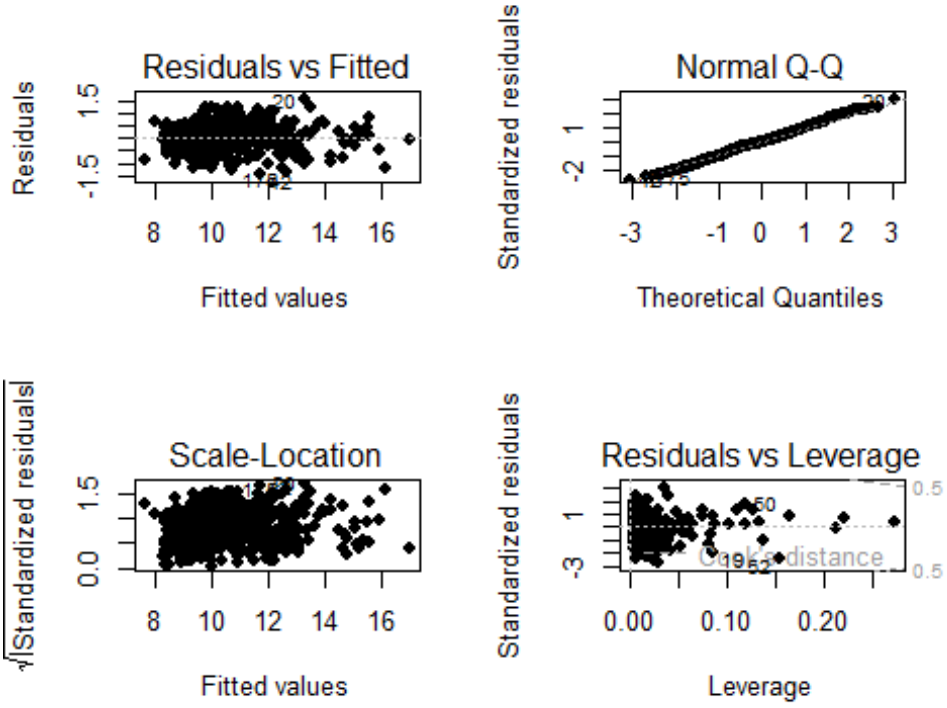
*There appear to be a few potential influential points which have a cook's distance over 1 (observations 34, 50, and 52). There is weak to moderate evidence against the assumption of linearity as there is random spread of residuals in the RvF plot. There is weak to moderate evidence equal variance as there a few observations with differing vertical spread, however these observations have been highlighted as influential points and may need to be removed. There is weak to moderate evidence against normality with a few large deviations from the line of normality, but the values with the greatest deviations are observations that are influential and may need to be removed from the model.**

5. (5 pts) Discuss the assumption of independence for these data, noting any potential violations or explaining why we do not have a violation (be specific about what you looked for to rule out violations).

There is moderate evidence against the assumption of independence due to possible selection bias as well as taxonomic clustering effects.

6. (6 pts) In the original paper, the researchers applied a log base 10 transformation to both quantitative variables. Apply a natural log transformation instead to both quantitative variables and refit the model with the interactions. Generate the set of residual diagnostics again and discuss any changes. Full assessments are not necessary, but you should briefly address each assumption.

```
dino$logBM<-log(dino$BodyMass)
dino$logCV<-log(dino$ClutchVolume)
lmlogdino<-lm(logCV~logBM*CareSex+logBM*HigherTaxon, data=dino)
par(mfrow=c(2,2))
plot(lmlogdino,pch=16,add.smooth=FALSE)
```



All of the assumptions look a lot better as there are no points with a Cook's distance greater than 0.5, there is little to no evidence against linearity and equal variance as all of the residuals look randomly spread with no visible patterns, and fairly constant vertical spread. There is little to no evidence against normality with little variance in residuals from the line of normality in the normal Q-Q plot.

7. (6 pts) Perform a formal hypothesis test for the interaction between log body mass and sex of primary parent in your fitted model. Provide the necessary output and a complete conclusion in context.

H_0: $\beta_{log(BodyMass):M} = \beta_{log(BodyMass):P} = 0$ H_a:At least one $\beta_k \neq 0$
where k = log(BodyMass):M, log(BodyMass):P

```
summary(lmlogdino)

##
## Call:
## lm(formula = logCV ~ logBM * CareSex + logBM * HigherTaxon, data = dino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4021 -0.3335 -0.0140  0.3180  1.5297
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.77331    0.05203 226.270  < 2e-16
## logBM               0.67638    0.01978  34.196  < 2e-16
```

```
## CareSexM            0.22654    0.08496   2.667  0.00796
## CareSexP            0.98196    0.10398   9.443  < 2e-16
## HigherTaxonC        0.35659    0.50243   0.710  0.47827
## logBM:CareSexM      0.08713    0.03028   2.878  0.00421
## logBM:CareSexP      0.22100    0.04757   4.645 4.54e-06
## logBM:HigherTaxonC -0.13641    0.13430  -1.016  0.31032
##
## Residual standard error: 0.5285 on 423 degrees of freedom
## Multiple R-squared:  0.9101, Adjusted R-squared:  0.9086
## F-statistic: 611.6 on 7 and 423 DF,  p-value: < 2.2e-16
```

```
Anova(lmlogdino)
```

```
## Anova Table (Type II tests)
##
## Response: logCV
##                   Sum Sq  Df   F value    Pr(>F)
## logBM             755.88   1 2706.1391 < 2.2e-16
## CareSex            18.99   2   33.9989 2.031e-14
## HigherTaxon         0.11   1    0.3983    0.5283
## logBM:CareSex       6.82   2   12.1999 7.060e-06
## logBM:HigherTaxon   0.29   1    1.0318    0.3103
## Residuals         118.15 423
```

There is a strong evidence that there is an interaction between logged Body mass and Sex of the primary care parent on the true mean logged clutch volume after accounting for the interaction between logged body mass and higher taxon in the model$\left(F_{2,423} = 12.1999, \text{p-value} < 0.001\right)$.

8. (6 pts) Now fit a purely additive model with all three explanatory variables (with the log transformed versions of the variables still). Provide a model summary and interpret the coefficient associated with body mass in context **on the original scale**. Hint: the usual backtransformation rules/language still apply to this model.

```
lmALD<-lm(logCV~logBM + CareSex + HigherTaxon, data=dino)
summary(lmALD)
```

```
##
## Call:
## lm(formula = logCV ~ logBM + CareSex + HigherTaxon, data = dino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50193 -0.32819 -0.02124  0.31773  1.71706
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.87623    0.04639 255.989  < 2e-16
## logBM         0.73202    0.01443  50.725  < 2e-16
## CareSexM      0.04996    0.05664   0.882    0.378
## CareSexP      0.78890    0.09866   7.996 1.22e-14
```

```
## HigherTaxonC  0.04730    0.15700   0.301    0.763
##
## Residual standard error: 0.542 on 426 degrees of freedom
## Multiple R-squared:  0.9048, Adjusted R-squared:  0.9039
## F-statistic:  1012 on 4 and 426 DF,  p-value: < 2.2e-16
```

```
2^0.73202
```

```
## [1] 1.660963
```

For each doubling of body mass(kg) , there is a 1.660963 multiplicative change in the estimated median clutch volume(mm^3) after accounting for the sex of the primary care parent and higher taxon group in the model.

9. (8 pts) Report the values of R^2 and adjusted R^2 for both the interaction model and the additive model (clearly label all four values). Discuss what we learn about the fitted models with these measures, addressing all four values through direct interpretation or comparison where appropriate.

Interaction Model:

$$R^2 = 0.9101$$

$$R^2_{adj} = 0.9086$$

Additive Model:

$$R^2 = 0.9048$$

$$R^2_{adj} = 0.9039$$

We see that in the interaction model 91.01% of the variation in logged clutch volume can be explained by the logged body mass, primary care parent sex, higher taxon group, and the interaction between logged body mass and the categorical variables. Whereas in the additive model, 90.48% of the variation in logged clutch volume can be explained by the logged body mass after accounting for primary care parent sex and higher taxon group. After adjusting for the complexity of each model, we still see that the interaction model is still doing a better job.