Bio220 - Lab Midterm 3 - Answers

2022-06-16

Contents

- Use mt3answer.R for your answers.
- Exam duration is 2 hours and consists of 3 questions.
- You should use 'primate' object for question #1 and #2 to perform necessary subsettings. If you cannot do the subsettings, you can use the objects already created to continue with the analysis. However, you will not get points for subsetting questions.
- For question #3, use 'newborn' object. No subsetting is required for this questions.

```
load('mt3data.rdata')
```

Question 1

'Primates' data set contain information about life history traits of several primate species. Column information of the data set is given below:

```
head(primates)
```

```
##             family          genus species mass.g gestation.mo newborn.g
## 585 Callitrichidae      Callimico goeldii 558.33         5.18     49.07
## 588 Callitrichidae     Callithrix pygmaea 116.83         4.60     16.16
## 591 Callitrichidae     Callithrix jacchus 309.00         4.90     27.12
## 592 Callitrichidae Leontopithecus rosalia 558.00         4.35     55.07
## 595 Callitrichidae       Saguinus bicolor 465.00         5.33   -999.00
## 597 Callitrichidae       Saguinus  mystax 526.13         5.00     46.90
##     weaning.mo wean.mass.g  AFR.mo max.life.mo litter.size litters.year
## 585       2.07         215   17.33         215        1.00         1.17
## 588       3.00          70   20.55         216        1.98      -999.00
## 591       5.42          97   16.33         144        2.24         2.00
## 592       4.30         165   25.01         338        1.95         1.75
## 595    -999.00        -999 -999.00        -999        1.50      -999.00
## 597       5.62        -999 -999.00        -999        1.89      -999.00
```

- family : taxonomical family
- genus : taxonomical genus
- species : taxonomical species
- mass.g : mass of the species in grams
- gestation.mo : gestation duration (gebelik süresi in turkish) in months
- newborn.g : newborn weight in grams
- other columns are not our interest.

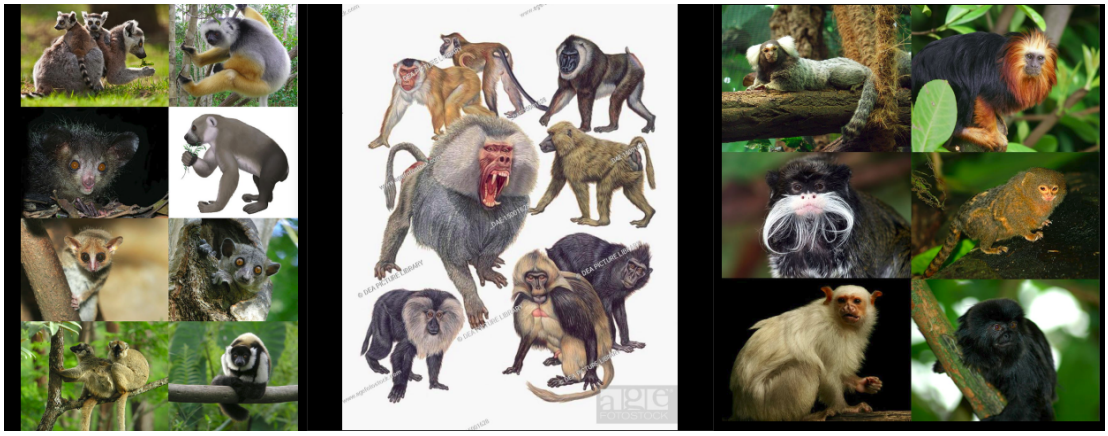


Figure 1: Lemuridae(left), Cercopithecidae(middle), Callitrichinae(right)

We want to compare gestation duration between Callitrichidae (marmosets, tamarins and lian tamarins family) and Cercopithecidae (old world monkeys). Specifically, we want to test if there is a significant mean gestation time difference between all the species in these two families.

- Subset the 'gestation.mo' as a vector for Callitrichidae family from the 'primate' data set and save it to an object called 'callit'. (1 pts) (no points if you subset from q2dat)

```
callit = primates[primates$family=='Callitrichidae','gestation.mo']
```

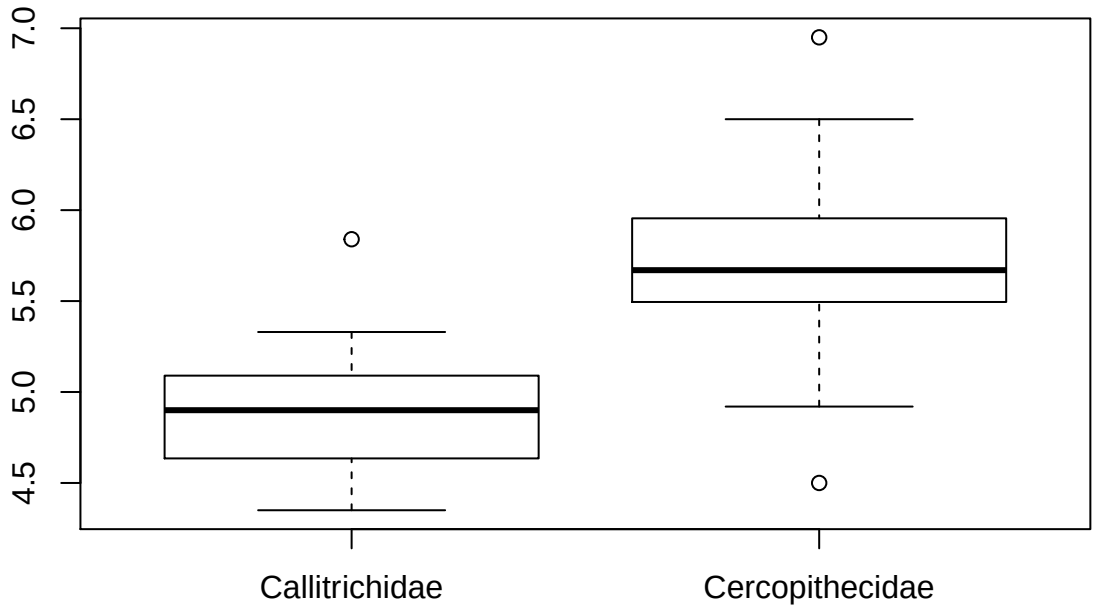
- Subset the 'gestation.mo' as a vector for Cercopithecidae family from the 'primate' data set and save it to an object called 'cercop'. (1 pts) (no points if you subset from q2dat)

```
cercop = primates[primates$family=='Cercopithecidae','gestation.mo']
```

(If you cannot subset and create the objects, use the objects with the same name from mt3data, no points will be given if you use those objects).

- Draw the boxplot of these two objects and give names to boxes. (no other plotting argument is needed) (2 pts)

```
boxplot(callit, cercop, names = c('Callitrichidae', 'Cercopithecidae'))
```



- Write null and alternative hypotheses. (2 pts)

```
# Ho: mean gestation time of callitrichidae and cercopithedae are the same
# Ha: mean gestation time of callitrichidae and cercopithedae are different
```

- Check the assumption of normality for the test you decided to use for both objects. (use only shapiro.test, you do not need to check normality using histogram or qqplot) (2 pts)

```
shapiro.test(callit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  callit
## W = 0.94078, p-value = 0.5297
```

```
shapiro.test(cercop)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cercop
## W = 0.95079, p-value = 0.08735
```

- According to the shapiro test result, comment on the normality of the two objects, separately. (example: p-value of shapiro test for object1 is <0.05, so the object1 is normally/nonnormally distributed, etc.) (2 pts)

```
# p-val for callit object is 0.53, -> normal
# p-val for cercop object is 0.087 and it's sample size is large (n=39) -> normal
```

- Perform the statistical test. (3 pts)

```
t.test(callit, cercop, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  callit and cercop
## t = -5.4766, df = 48, p-value = 1.563e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.08918 -0.50420
## sample estimates:
## mean of x mean of y
##  4.928182  5.724872
```

```
t.test(callit, cercop)
```

```
##
##  Welch Two Sample t-test
##
## data:  callit and cercop
## t = -5.6205, df = 16.712, p-value = 3.258e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0961415 -0.4972385
## sample estimates:
## mean of x mean of y
##  4.928182  5.724872
```

```
# Note: Question does not ask to check equal variances, so I will accept both as correct
```

- Comment on the test result. (example: p-value of the test result >0.05, so I reject/don't reject the null hypothesis. There is/There is no statistical difference between gestation time between the two families, etc.) (2 pts)

```
# Test result is highly significant, gestation times are different between these two
# families. (Cercopithecidae has higher gestation time)
```

Question 2

We want to compare mean gestation time between Callitrichidae, Cercopithecidae and Lemuridae families together in one statistical test. You have already subsetted the gestation times for Callitrichidae and Cercopithecidae families.

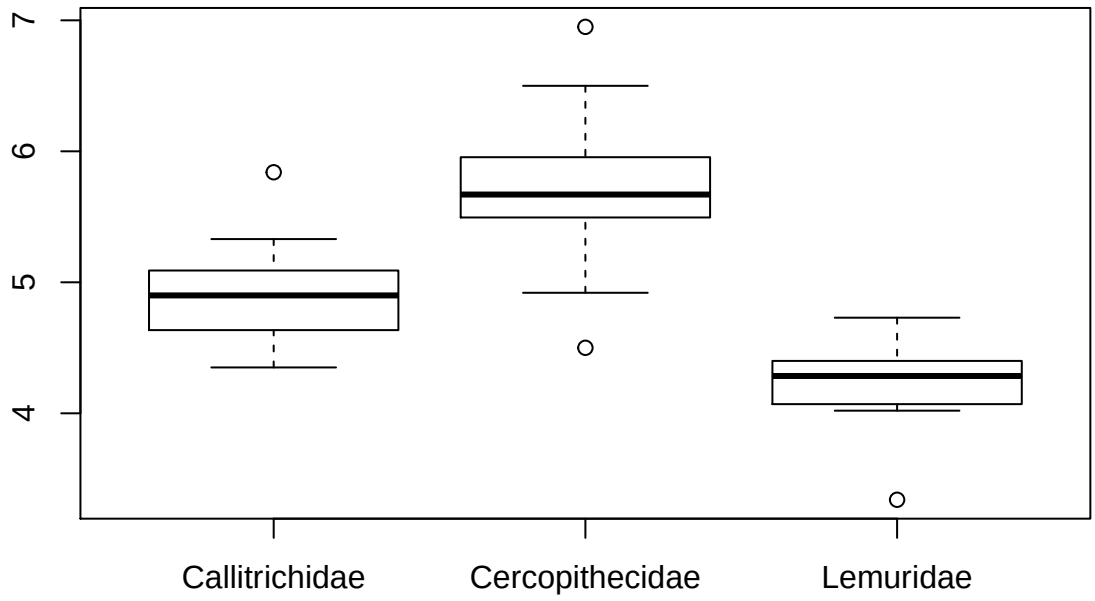
- Subset the 'gestation.mo' as a vector for Lemuridae family from the 'primate' data set and save it to an object called 'lemuri'. (1 pts)

(If you cannot subset and create the object, use the objects with the same name from mt3data, no points will be given if you use that object).

```
lemuri = primates[primates$family=='Lemuridae','gestation.mo']
```

- Draw the boxplot of these three objects and give names to boxes. (no other plotting argument is needed) (2 pts)

```
boxplot(callit, cercop, lemuri, names = c('Callitrichidae', 'Cercopithecidae', 'Lemuridae'))
```



- Write null and alternative hypotheses. (2 pts)

```
# Ho: three families have same mean gestation time
# Ha: at least one family has different mean gestation time
```

- Check the assumption of normality for the test you decided to use for 'lemuri' object only, as you have already checked the other two objects in the first question. (use only shapiro.test, you do not need to check normality using histogram or qqplot) (1 pts)

```
shapiro.test(lemuri)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lemuri
## W = 0.89165, p-value = 0.2424
```

- According to the shapiro test result, comment on the normality of the 'lemuri'. (example: p-value of shapiro test for object1 is <0.05, so the object1 is normally/nonnormally distributed, etc.) (1 pts)

```
# Shapiro test result is not significant (p-val = 0.24), so we can assume normality
```

- Note: assume homogeneity of variance between the groups.
- To perform the anova, you need to put these three object into a data frame with two columns. First column should contain all the gestation time data and the second column should contain

a character/factor that specify which family the gestation times in the first column belongs to. (5 pts bonus). If you cannot create this object, use the object called 'q2dat'. (no extra points).

```
datx = data.frame(gestation = c(callit, cercop, lemuri),
                  family = c(rep('callit', length(callit)),
                             rep('cercop', length(cercop)),
                             rep('lemuri', length(lemuri))))
head(datx)
```

```
##   gestation family
## 1      5.18 callit
## 2      4.60 callit
## 3      4.90 callit
## 4      4.35 callit
## 5      5.33 callit
## 6      5.00 callit
```

- Perform the statistical test. (3 pts)

```
q2aov = aov(gestation~family, dat=datx)
q2aov
```

```
## Call:
##    aov(formula = gestation ~ family, data = datx)
##
## Terms:
##                     family Residuals
## Sum of Squares  18.053953  9.870888
## Deg. of Freedom         2        55
##
## Residual standard error: 0.4236398
## Estimated effects may be unbalanced
```

```
anova(q2aov)
```

```
## Analysis of Variance Table
##
## Response: gestation
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## family     2 18.0540  9.0270  50.298 3.802e-13 ***
## Residuals 55  9.8709  0.1795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Comment on the test result. (example: p-value of the test result >0.05, so I reject/don't reject the null hypothesis. There is/ There is no statistical difference between gestation time between the three families, etc.) (2 pts)

```
# Anova test result is significant (p<0.05), we reject null hypothesis.
# There is difference in gestation time in at least one family.
```

- Perform a posthoc test if necessary and specify which families have different gestation time than the others (2 pts)

```
TukeyHSD(q2aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = gestation ~ family, data = datx)
##
## $family
##                       diff        lwr        upr     p adj
## cercop-callit  0.7966900  0.4483165  1.1450634 0.0000029
## lemuri-callit -0.7306818 -1.2048410 -0.2565226 0.0013783
## lemuri-cercop -1.5273718 -1.9234312 -1.1313124 0.0000000
```

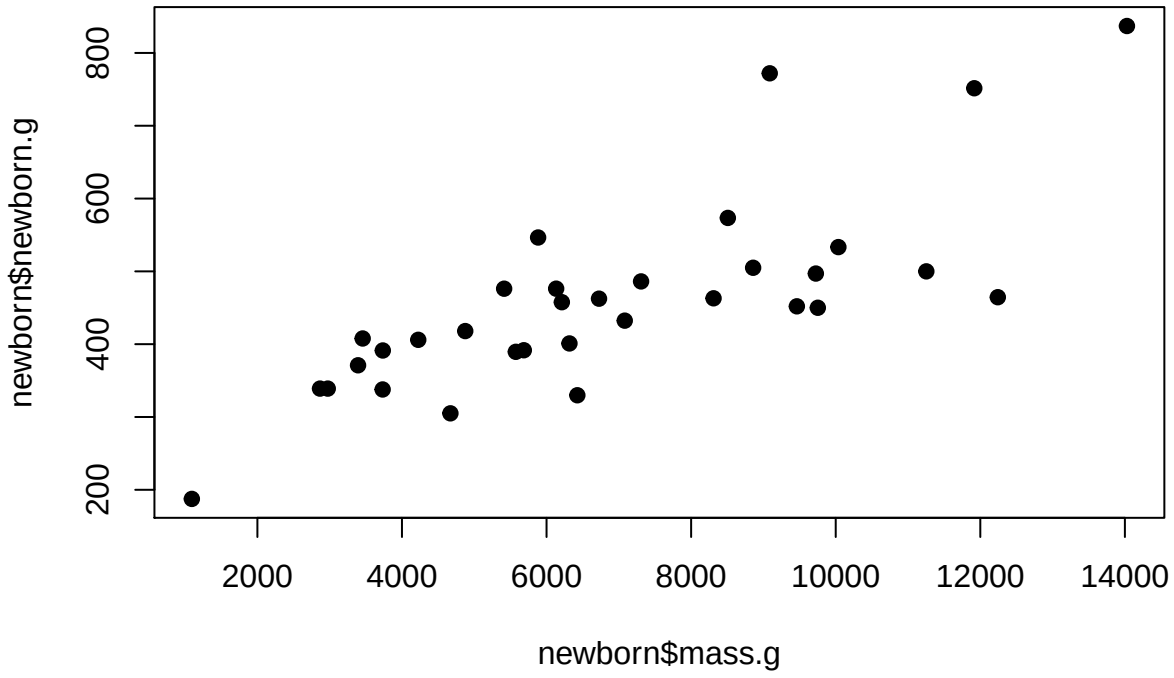
```
# All pairwise comparisons are significant (p<0.05).
# All families have different gestation times than the others
```

Question 3

We want to test if there is a relationship between 'mass.g' (mass of the parent) and 'newborn.g' (mass of the offspring) for Cercopithericade family. No subsetting required for this question. Use 'newborn' object.

- Draw scatterplot of the data such that parent mass on the x axis and newborn mass on the y axis. (2 pts)

```
plot(newborn$mass.g, newborn$newborn.g, pch=19)
```



- Assume that both variables have bivariate normal distribution and have linear relationship.
- Specify null and alternative hypotheses (2 pts) and perform a correlation analysis on the data (2 pts)

```
# Ho: correlation coefficient between newborn and mother mass is zero
# Ha: corr. coeff. is different than zero
```

```
cor.test(newborn$mass.g, newborn$newborn.g)
```

```
##
##  Pearson's product-moment correlation
##
## data:  newborn$mass.g and newborn$newborn.g
## t = 6.7469, df = 31, p-value = 1.495e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5821140 0.8812492
## sample estimates:
##       cor
## 0.7712855
```

- According to the test result, what kind of relationship do the mass.g and newborn.g have? (example: newborn mass increases/decreases as the parent mass increase/decrease) (2 pts)

```
# Correlation coefficient is positive (0.77), there is a positive relationship
# between newborn mass and parent mass
```

- Specify which variable is explanatory and which one is response (1 pts) and perform a regressin analysis (2 pts)

use only lm function, you do not need to check the significance of the slop

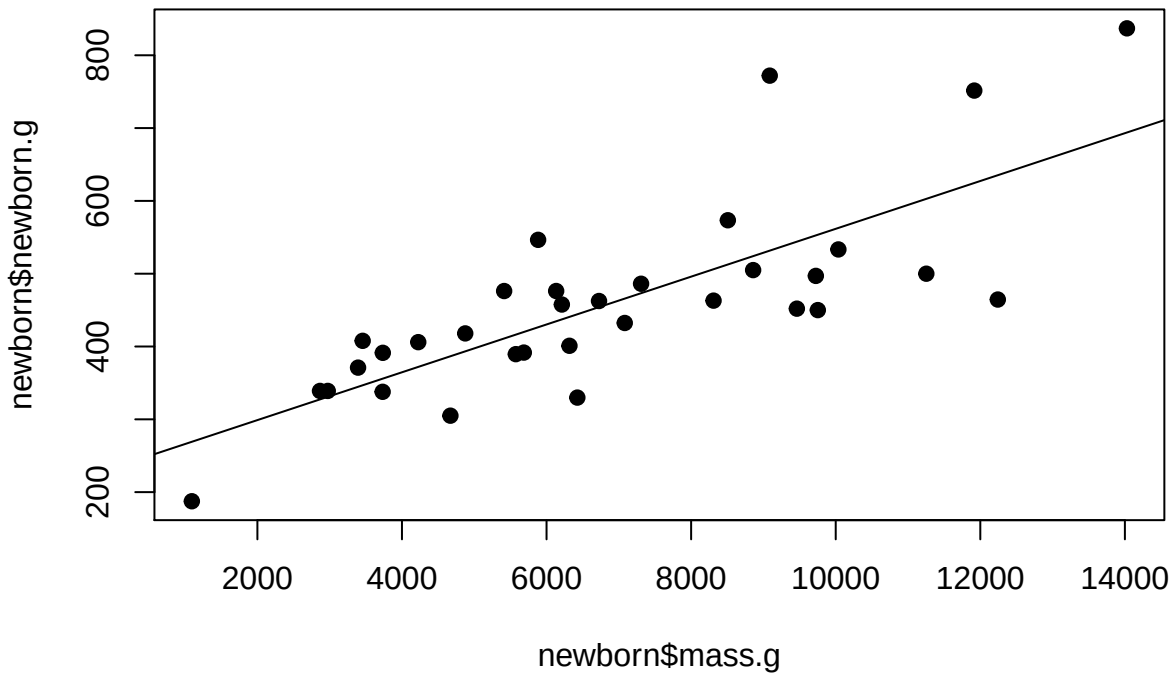
you do not need to check the assumptions of the regression analysis

```
# parent mass is explanatory variable
# newborn mass is response variable
lmx = lm(newborn.g~mass.g, data=newborn)
lmx
```

```
##
## Call:
## lm(formula = newborn.g ~ mass.g, data = newborn)
##
## Coefficients:
## (Intercept)       mass.g
##   233.26081      0.03284
```

- Add regression line onto the plot (1 pts)

```
plot(newborn$mass.g, newborn$newborn.g, pch=19)
abline(lmx)
```



- Using regressin analysis, what would be the average newborn mass if a newly discovered parent species has a mass of 1200 gram. (Use prediction with specifying confidence intervals) (2 pts).

```
predict(lmx, data.frame(mass.g = c(1200)), interval='prediction')
```

```
##        fit      lwr      upr
## 1 272.6672 88.34639 456.9881
```