HW 3

No names - just add upon submission in gradescope

Due 2/7/2025

# HW 3 Instructions:

* In groups of 2 to 4 (your choice of group members, but you must be in a group to complete the submission), complete the following questions. If you are having trouble identifying group members, post in the Water Cooler channel in MS Teams.

Read Fabusova et al. (2024) and use the provided data set (“moths.xlsx”) to answer the following questions.

* Fabusova M, Gaston KJ, Troscianko J. 2024 Pulsed artificial light at night alters moth flight behaviour. Biol. Lett. 20: 20240403. <https://doi.org/10.1098/rsbl.2024.0403>

The Latency (seconds) is the time to the moth starting to move after the treatment is applied or 10 seconds of dark for the control group. The treatment information is coded in three variables, Treatment with 5 levels (numbered unhelpfully as 1 to 5, shown with better labels once we figure out their meaning on the x-axis in Figure 2a) and also in Colour (cold or warm) and Spectra (LED or RGB - note that the LED level in this variable is the “Phosphor LED” and RGB level is the “RGB LED” in their Figure 2).

1. The following code reads in the data set from their published Excel file and removes the Sunset\_time that we will not use. What is the sample size of moths before any cleaning? Then use drop\_na to clean the data set to remove any missing observations across the following variables: Date, ID, Location, Treatment, Treatment\_y\_n, Colour, Spectra, Moon\_Phase, Temp, Wind\_speed\_ms, Humidity, Common\_name, Latin\_name, Family, Sub\_family, Latency, Erratic\_behaviour. What is the sample size after that cleaning?

* Hint: The list above is provided so you can directly copy it into drop\_na.
* **Sample size before cleaning is 393 observations. Sample size after cleaning is 302 observations.**

library(readxl)  
moth <- read\_excel("moths.xlsx",   
 sheet = "complete\_dataset\_MAIN FAMILIES.",   
 na = "NA")  
  
library(lubridate)  
moth <- moth %>%   
 mutate(Date = factor(ymd(Date))) %>%   
 dplyr::select(-Sunset\_time)  
  
dat <- moth %>% drop\_na(Date, ID, Location, Treatment, Treatment\_y\_n, Colour, Spectra, Moon\_Phase, Temp, Wind\_speed\_ms, Humidity, Common\_name, Latin\_name, Family, Sub\_family, Latency, Erratic\_behaviour)

1. Read their Methods section (mainly parts A and B). The Treatment, Colour, and Spectra variables tell an interesting story about the reason for some of the missingness in the original data set. Use tally(Spectra ~ Colour | Treatment, data = moth) to explore those variables and select the correct SINGLE response from these choices:
2. The Colour and Spectra variables had missing values entered in order to remove outliers in the latency responses.
3. The Colour and Spectra variables were missing for the level 5 of Treatment because the researchers failed to record what treatment level was applied to the moths in those rows.

**c) The value 5 for the Treatment variable is a control group of moths that were not subjected to any light treatments, so they entered a missing value for the two levels of Colour and Spectra.**

**Answer is bolded**

tally(Spectra ~ Colour | Treatment, data = dat)

## , , Treatment = 1  
##   
## Colour  
## Spectra cold warm  
## LED 88 0  
## RGB 0 0  
##   
## , , Treatment = 2  
##   
## Colour  
## Spectra cold warm  
## LED 0 71  
## RGB 0 0  
##   
## , , Treatment = 3  
##   
## Colour  
## Spectra cold warm  
## LED 0 0  
## RGB 70 0  
##   
## , , Treatment = 4  
##   
## Colour  
## Spectra cold warm  
## LED 0 0  
## RGB 0 73

1. Improve the factor levels of the Treatment variable in your cleaned data set to match the meaning of the levels using fct\_recode and also create a natural log-transformed version of the Latency variable for later use. No discussion, just code here.

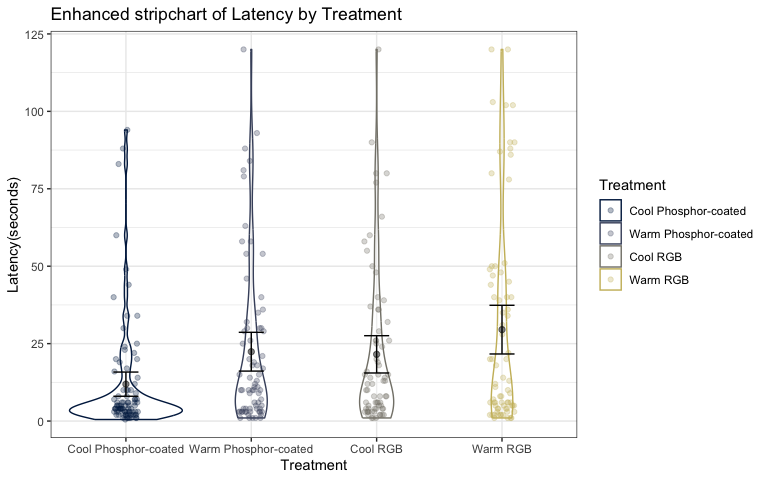
dat <- dat %>% mutate(Treatment = factor(Treatment),  
 Treatment = fct\_recode(Treatment,  
 "Cool Phosphor-coated" = "1",  
 "Warm Phosphor-coated" = "2",  
 "Cool RGB" = "3",  
 "Warm RGB" = "4"),  
 log\_Latency = log(Latency))

## Use the cleaned data set for all remaining questions on the HW:

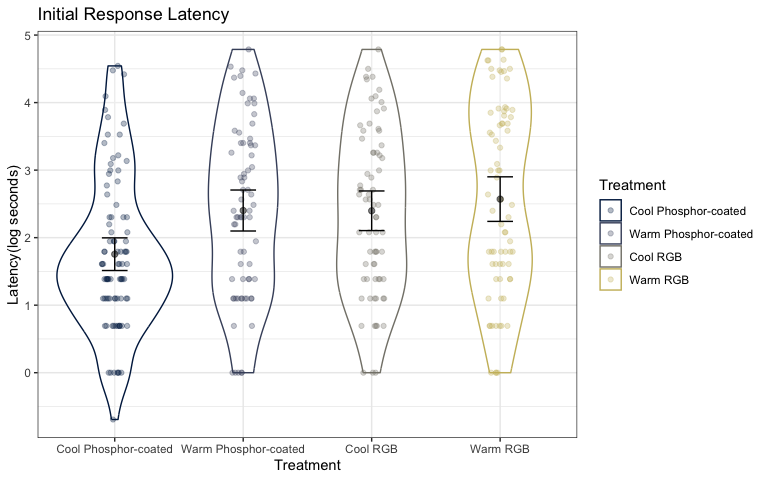
1. Using your improved labels for the Treatment variable, make a display of the Latency versus that grouping variable and then the logLatency versus that grouping variable. The second one should resemble their Figure 2b but be much more informative. What information does your *enhanced* (hint!) display provide that is not presented in their version of the same plot?

**The *enhanced* stripchart displays the individual observations which allows us to see clustering of points and possible data trends that we may not be able to see in the boxplot.**

enhanced\_stripchart(Latency ~ Treatment, data = dat) +   
 labs(y = "Latency(seconds)")



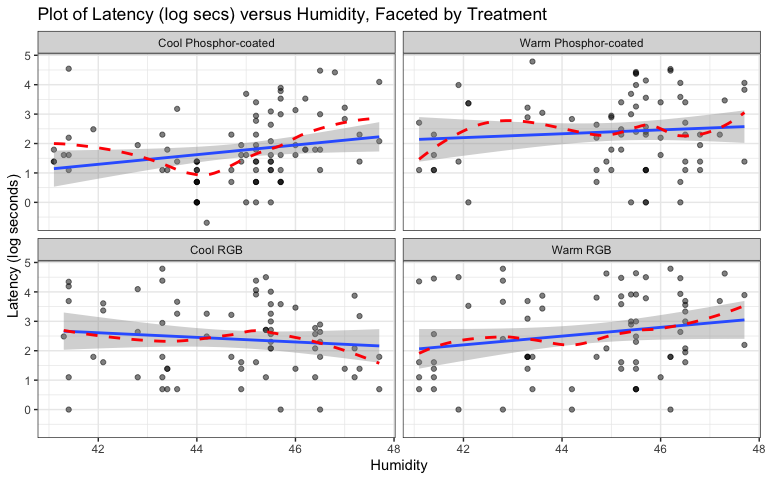
enhanced\_stripchart(log\_Latency ~ Treatment, data = dat) +   
 labs(y = "Latency(log seconds)",  
 title = "Initial Response Latency")



1. Make a scatterplot of the logLatency versus the Humidity faceted by Treatment with linear and nonparametric smoothing lines. Discuss the potential for an interaction between Humidity and Treatment on log-Latency based on the plot.

**The smoothing lines are not the same across all treatment groups, suggesting a possible interaction between Humidity and Treatment on log-latency. This is especially noticible in the cool light sources.**

dat %>% ggplot(mapping = aes(x = Humidity, y = log\_Latency)) +   
 geom\_point(alpha = 0.5) +   
 geom\_smooth(method = "lm") +   
 geom\_smooth(col = "red", lty = 2, se = F) +  
 facet\_wrap(~ Treatment) +   
 scale\_color\_viridis(discrete = TRUE) +  
 labs(x = "Humidity",   
 y = "Latency (log seconds)",  
 title = "Plot of Latency (log secs) versus Humidity, Faceted by Treatment")



1. Fit the interaction model with Humidity and Treatment for log-latency. Write out the estimated model. Make sure you define all your indicator variables and use the notation from this course.

moth\_model <- lm(log\_Latency ~ Humidity \* Treatment, data = dat)  
  
summary(moth\_model)

##   
## Call:  
## lm(formula = log\_Latency ~ Humidity \* Treatment, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8249 -0.9578 -0.1323 0.9886 3.3480   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -5.59197 3.85451 -1.451 0.1479  
## Humidity 0.16394 0.08596 1.907 0.0575  
## TreatmentWarm Phosphor-coated 5.03695 5.45739 0.923 0.3568  
## TreatmentCool RGB 11.49687 5.38314 2.136 0.0335  
## TreatmentWarm RGB 1.53140 5.19801 0.295 0.7685  
## Humidity:TreatmentWarm Phosphor-coated -0.09833 0.12135 -0.810 0.4184  
## Humidity:TreatmentCool RGB -0.24237 0.12016 -2.017 0.0446  
## Humidity:TreatmentWarm RGB -0.01491 0.11628 -0.128 0.8981  
##   
## Residual standard error: 1.255 on 294 degrees of freedom  
## Multiple R-squared: 0.09027, Adjusted R-squared: 0.06861   
## F-statistic: 4.168 on 7 and 294 DF, p-value: 0.0002154

where is 1 when treatment is Warm phosphor-coated LED and 0 if else, is 1 when treatment is Cool RGB LED and 0 if else, is 1 when treatment is Warm RGB LED and 0 if else.

1. The following work provides the simplified estimated models for two levels of the Treatment variable:

* Cold phosphor:
* Cold RGB:

Which of the following is the correct “size” interpretation (CIs are a topic for later - leave them as ellipses) for the 0.16 estimated slope coefficient for Humidity, assuming that this was the relative humidity they were using which is a percentage. Pick ONE - no discussion.

1. A one percent increase in relative humidity causes the estimated mean log-latency time to change by 0.16 log-seconds, controlled for treatment (95% CI of … to …).
2. For two otherwise similar moths but that had observed times that differed in relative humidity by 1 percent, the one measured during higher relative humidity had a higher estimated mean log-latency time by 0.16 log-seconds than the less humid observation, controlled for treatment (95% CI of … to …).
3. A one percent increase in relative humidity causes the estimated mean log-latency time to change by 0.16 log-seconds for an observation with a cold phosphorous light (95% CI of … to …).

**d) For two otherwise similar moths but that had observed times that differed in relative humidity by 1 percent, the one measured during higher relative humidity had a higher estimated mean log-latency time by 0.16 log-seconds than the less humid observation for an observation with a cold phosphorous light (95% CI of … to …).**

**Answer is bolded**

1. Drop the interaction from the model and re-fit it. Make an effects plot for the model. Write a size interpretation for the humidity slope coefficient from that model, making sure you find and include the 95% CI in parentheses in the size interpretation.

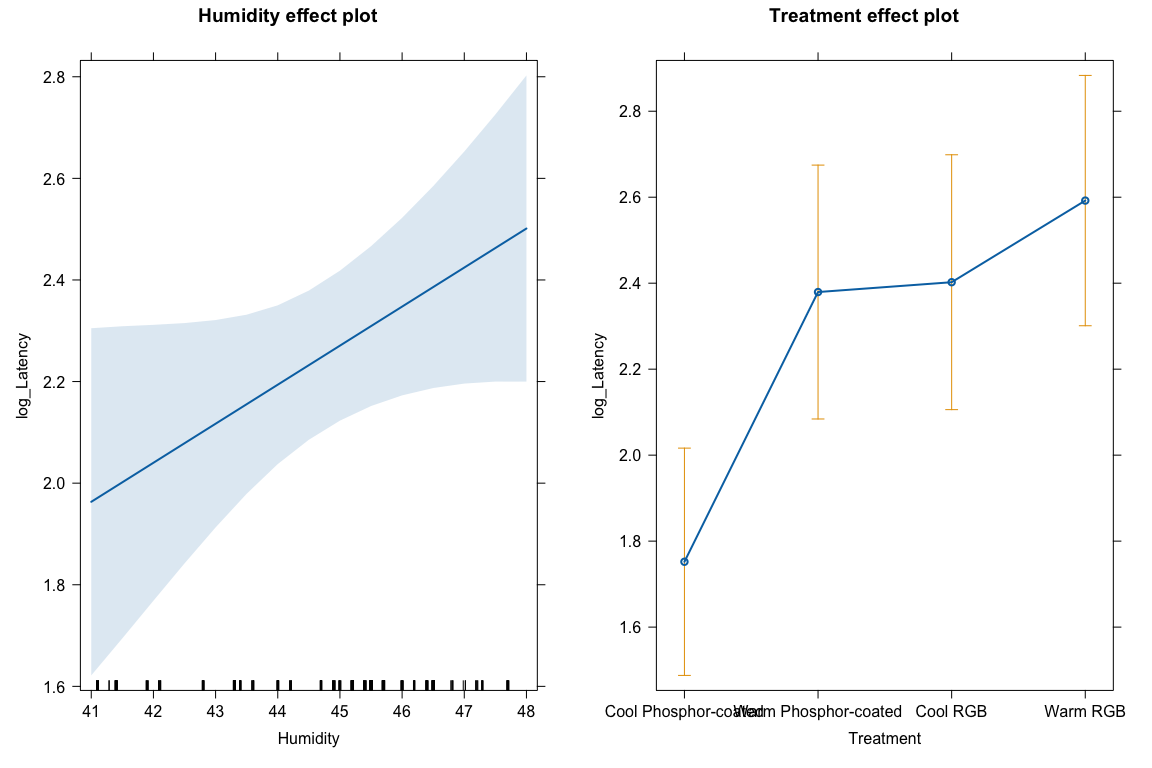
moth\_model2 <- lm(log\_Latency ~ Humidity + Treatment, data = dat)  
summary(moth\_model2)

##   
## Call:  
## lm(formula = log\_Latency ~ Humidity + Treatment, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.70176 -0.99937 -0.08427 0.98319 3.05061   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.68919 1.87851 -0.899 0.36926  
## Humidity 0.07686 0.04181 1.838 0.06702  
## TreatmentWarm Phosphor-coated 0.62741 0.20128 3.117 0.00200  
## TreatmentCool RGB 0.65030 0.20184 3.222 0.00142  
## TreatmentWarm RGB 0.84016 0.19992 4.202 3.5e-05  
##   
## Residual standard error: 1.26 on 297 degrees of freedom  
## Multiple R-squared: 0.07383, Adjusted R-squared: 0.06136   
## F-statistic: 5.919 on 4 and 297 DF, p-value: 0.0001353

confint(moth\_model2, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -5.386063998 2.0076750  
## Humidity -0.005424027 0.1591381  
## TreatmentWarm Phosphor-coated 0.231301550 1.0235259  
## TreatmentCool RGB 0.253084510 1.0475117  
## TreatmentWarm RGB 0.446713813 1.2336028

plot(allEffects(moth\_model2))



where is 1 when treatment is Warm phosphor-coated LED and 0 if else, is 1 when treatment is Cool RGB LED and 0 if else, is 1 when treatment is Warm RGB LED and 0 if else.

**For two otherwise similar moths but that had observed times that differed in relative humidity by 1 percent, the one measured during higher relative humidity had a higher estimated mean log-latency time by 0.08 log-seconds than the less humid observation, controlled for treatment (95% CI of -0.01 to 0.16).**

1. Document any collaborations or pertinent discussions with other students or resources outside of your group and the resources that I am providing and the *Sleuth* that you used to complete this assignment *or report that you did not have any*. If you used generative AI (chatGPT, Bard, etc.), report which question(s) and how/what you asked for.

* **NONE**