HW 9

No name needed

April 17

Rules:

This is an individual assignment, but you are welcome to discuss your answers with other students. Please document any discussions that were impactful on your answers to go with documenting any other outside of class resources you used.

Part I: Revisiting "Factors related to intra-tendinous morphology of Achilles tendon in runners"

- Ho K-Y, Baquet A, Chang Y-J, Chien L-C, Harty M, Bashford G, et al. (2019) Factors related to intra-tendinous morphology of Achilles tendon in runners. *PLoS ONE* 14(8): e0221183. https://doi.org/10.1371/journal.pone.0221183

The article is available at
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221183

We explored this data set in the initial labs this semester...

```
library(tidyverse)
data(TendonData) #Version of data set from catstats2
tendon <- TendonData

tendon <- tendon %>%
      rename(SubjectID = 'Subject ID',
             Sex = 'Sex  (M=1)',
             CurrPain = 'CurrPain (Y=1)',
             WaisttoHip = 'Waist to Hip',
             VISAA = 'VISA-A',
             Neovascularization = 'Neovascularization (Doppler) (Y=1)',
             HistoryPain = 'Hx Pain (Y=1)') %>%
      mutate_if(is.character, as.factor) %>%
      mutate(CurrPain = factor(CurrPain),
             Neovascularization = factor(Neovascularization),
             Sex = factor(Sex),
             HistoryPain = factor(HistoryPain))
tendonna1 <- tendon %>% drop_na(CSA)
tendonna2 <- tendon %>% drop_na()
tendonna2 <- tendonna2 %>% mutate(
  Sex = fct_recode(Sex,
                   "Female" = "0",
```

```
                    "Male" = "1"),
  Location = factor(substr(SubjectID, 1, 3))
  )
lm2 <- lm(PSFR ~ Sex, data = tendonna2)
summary(lm2)

##
## Call:
## lm(formula = PSFR ~ Sex, data = tendonna2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33828 -0.11924  0.00509  0.10822  0.36571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.02348    0.01996 101.394  < 2e-16
## SexMale     -0.10458    0.02582  -4.051 7.89e-05
##
## Residual standard error: 0.1621 on 162 degrees of freedom
## Multiple R-squared:  0.09198,    Adjusted R-squared:  0.08638
## F-statistic: 16.41 on 1 and 162 DF,  p-value: 7.893e-05

lm2 %>% tbl_regression(intercept = T) %>% add_global_p()
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	2.0	2.0, 2.1	<0.001
Sex			<0.001
Female	—	—	
Male	-0.10	-0.16, -0.05	

[1]CI = Confidence Interval

The prior work led to us having the following estimated model and size interpretation.

- $\hat{\mu}\{PSFR|Sex\} = 2.02 - 0.105I_{Sex=Male}$
- where $I_{Sex=Male}$ is 1 for a male observation and 0 otherwise

For two otherwise similar subjects but that differ on the sex of the subjects, the estimated mean PSFR level of male subjects is 0.10 mm^{-1} **lower than that of female subjects, controlled for subject-to-subject variation (95% CI from 0.05 to 0.16).**

There is strong evidence against the null hypothesis of no difference in the true mean PSFR between the sexes of the subjects ($t_{162} = -4.05$**, 2-sided p-value < 0.0001), so we would conclude there is a difference and keep it in the model.**

1) Repeat the previous analysis but with a linear mixed model that accounts for the repeated measures on the same subject. Fit the model and report a model `summary` and `tbl_regression(intercept = T)` on it. No discussion yet.

```
lmer <- lmer(PSFR ~ Sex + (1|SubjectID), data = tendonna2)
summary(lmer)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: PSFR ~ Sex + (1 | SubjectID)
##    Data: tendonna2
##
## REML criterion at convergence: -156.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.76668 -0.56075  0.01359  0.51430  2.08764
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  SubjectID (Intercept) 0.01567  0.1252
##  Residual              0.01081  0.1040
## Number of obs: 164, groups:  SubjectID, 82
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  2.02348    0.02527 80.00000  80.073  < 2e-16
## SexMale     -0.10458    0.03269 80.00000  -3.199  0.00198
##
## Correlation of Fixed Effects:
##         (Intr)
## SexMale -0.773

lmer %>% tbl_regression(intercept = T) %>% add_global_p()
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	2.0	2.0, 2.1	<0.001
Sex			0.001
Female	—	—	
Male	-0.10	-0.17, -0.04	

[1]CI = Confidence Interval

2) Revise the previous "size" sentence based on the two-level mixed model results.

- $\hat{\mu}\{PSFR|Sex\} = 2.02 - 0.105I_{Sex=Male}$
- where $I_{Sex=Male}$ is 1 for a male observation and 0 otherwise

For two otherwise similar subjects but that differ on the sex of the subjects, the estimated mean PSFR level of male subjects is 0.10 mm^{-1} lower than that of female subjects (95% CI from 0.04 to 0.17).

3) The evidence sentence for the previous `lm` is also provided above. Update it to reflect the two-level mixed model results.

There is strong evidence against the null hypothesis of no difference in the true mean PSFR between the sexes of the subjects ($t_{80} = -3.199$, 2-sided p-value = 0.00198), controlled for subject-to-subject variation, so we would conclude there is a difference and keep it in the model.

In the previous work with these data, we fit the following model:

```
lm2_all <- lm(PSFR ~ Age + Sex + BMI + WaisttoHip + YearsRunning + CurrPain +
                   HistoryPain + VISAA + CSA + Neovascularization + KneeWall
+
                   HeelRaise,
  data = tendonna2
)
Anova(lm2_all)
```

```
## Anova Table (Type II tests)
##
## Response: PSFR
##                     Sum Sq  Df F value   Pr(>F)
## Age                 0.0194   1  0.7304 0.394102
## Sex                 0.2360   1  8.8847 0.003353
## BMI                 0.0003   1  0.0105 0.918600
## WaisttoHip          0.0083   1  0.3112 0.577777
## YearsRunning        0.0026   1  0.0976 0.755219
## CurrPain            0.0817   1  3.0771 0.081431
## HistoryPain         0.0227   1  0.8535 0.357026
## VISAA               0.0157   1  0.5897 0.443741
## CSA                 0.0205   1  0.7716 0.381128
## Neovascularization 0.0632   1  2.3778 0.125163
## KneeWall            0.0018   1  0.0684 0.794009
## HeelRaise           0.0007   1  0.0263 0.871433
## Residuals           4.0103 151
```

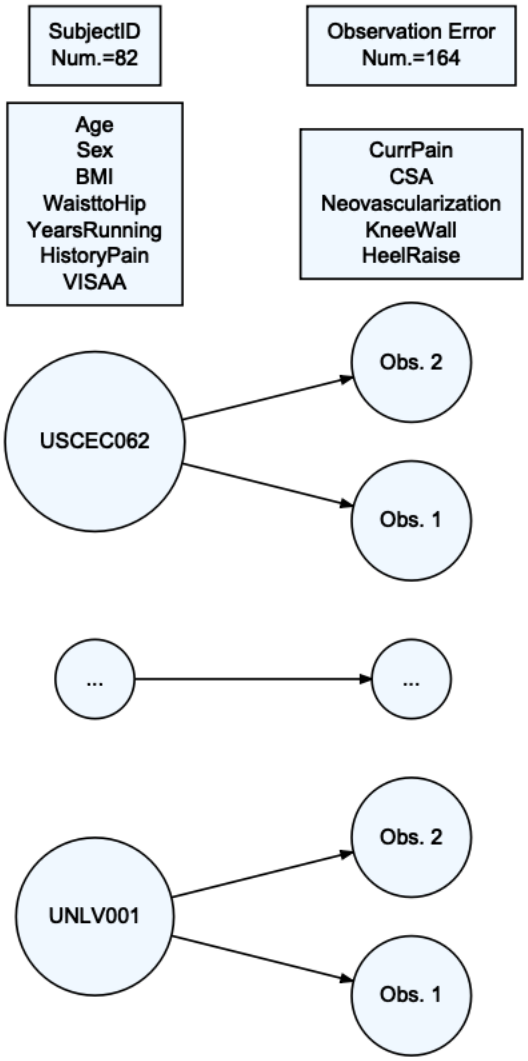
4) Fit the linear mixed model that accounts for the subject with the same suite of fixed effects as in `lm2_all` and make a `model_diagram`. Use the diagram to write a couple of sentences to report the levels of the hierarchy for each of the predictors (subject or leg in subject are the options).

The SubjeectID level has 82, which accounts for all of the individual subjects. Each of these 82 subjects went through two repeated observations based on side (L or R), which results in the 164 observations on the second level. Age, sex, BMI, WaisttoHip, years running, pain history, and VISAA are all variables arrigned at the subject level.

CurrPain, CSA, Neovascularization, KneeWall, and heelraise on the other hand are recorded per leg.

```
lmer_all <- lmer(PSFR ~ Age + Sex + BMI + WaisttoHip + YearsRunning +
CurrPain +
                   HistoryPain + VISAA + CSA + Neovascularization + KneeWall
+
                   HeelRaise + (1|SubjectID),
  data = tendonna2
)

model_diagram(lmer_all, heightVal = 800)
```



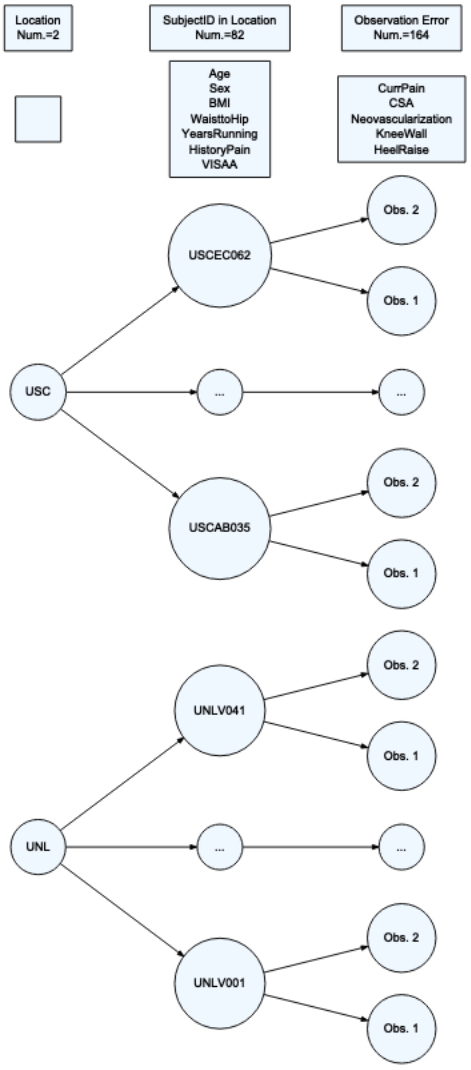
5) As we discussed in one of the labs, there is another issue here with two different locations where the tests were performed, with each subject being nested within

one of the two locations. The location information is available in the `Location` variable that is created in the provided code. Add `Location` to the model as a random effect with the same fixed effects as in the previous question and remake the `model_diagram`. Discuss why there are no `Location` level predictors in the model.

There are no location level predictors because none of the recorded information differs by site. If there were environmental factors that were used as predictors those may show up on a site level, but currently in the model there is nothing that differs at a site-specific level.

```
lmer_site <- lmer(PSFR ~ Age + Sex + BMI + WaisttoHip + YearsRunning +
CurrPain +
                    HistoryPain + VISAA + CSA + Neovascularization + KneeWall
+
                    HeelRaise + (1|Location/SubjectID),
  data = tendonna2
)

model_diagram(lmer_site, heightVal = 800)
```




```
summary(lmer_site)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: PSFR ~ Age + Sex + BMI + WaisttoHip + YearsRunning + CurrPain +
##     HistoryPain + VISAA + CSA + Neovascularization + KneeWall +
##     HeelRaise + (1 | Location/SubjectID)
##    Data: tendonna2
##
## REML criterion at convergence: -74.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7987 -0.5881  0.0273  0.5258  1.9845
##
```

```
## Random effects:
##  Groups              Name        Variance Std.Dev.
##  SubjectID:Location (Intercept) 0.015759 0.12553
##  Location           (Intercept) 0.002082 0.04563
##  Residual                       0.011038 0.10506
## Number of obs: 164, groups:  SubjectID:Location, 82; Location, 2
##
## Fixed effects:
##                       Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)          1.731e+00  3.556e-01  7.867e+01   4.869 5.69e-06
## Age                 -9.784e-04  1.825e-03  7.501e+01  -0.536   0.5934
## SexMale             -1.062e-01  4.377e-02  7.603e+01  -2.426   0.0176
## BMI                  1.559e-03  7.273e-03  7.367e+01   0.214   0.8308
## WaisttoHip           2.621e-01  4.067e-01  7.164e+01   0.645   0.5213
## YearsRunning        -6.050e-04  2.457e-03  7.133e+01  -0.246   0.8062
## CurrPain1           -2.122e-02  3.026e-02  1.223e+02  -0.701   0.4844
## HistoryPain1         1.721e-02  3.963e-02  7.155e+01   0.434   0.6654
## VISAA                9.071e-04  1.106e-03  7.666e+01   0.820   0.4147
## CSA                 -4.444e-04  7.260e-04  1.489e+02  -0.612   0.5414
## Neovascularization1  7.528e-02  5.144e-02  1.479e+02   1.463   0.1455
## KneeWall             1.992e-04  4.050e-03  1.070e+02   0.049   0.9609
## HeelRaise            6.225e-04  2.003e-03  1.154e+02   0.311   0.7565
```

6) Write a sentence that interprets the following estimated ICC:

Once we account for location and subjectID, the estimated correlation of the two grip strength measurements is 0.62. This indicates some correlation between two observations once we account for systematic changes across subjects and location.

```
(0.002082 + 0.015759)/(0.002082 + 0.015759 + 0.011038)
```

```
## [1] 0.6177845
```

7) Generate an ANOVA F-test table from the three-level mixed model. For `WaisttoHip` (ratio of the waist to hip measurements) and HeelRaise (number of heel raises), report the F-statistics, distributions under the null, and p-values. No other discussion - just extract those "numerical" results outside the code output.

WaisttoHip: $F_{1,72.041} = 0.4139, p - value = 0.52205$

HeelRaise: $F_{1,115.713} = 0.0932, p - value = 0.76064$

```
Anova(lmer_site, test.statistic = "F")
```

```
## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)
##
## Response: PSFR
##                        F Df  Df.res  Pr(>F)
## Age               0.2865  1  75.455 0.59401
## Sex               5.8170  1  76.383 0.01827
## BMI               0.0457  1  74.058 0.83136
```

```
## WaisttoHip          0.4139  1  72.041 0.52205
## YearsRunning        0.0604  1  71.738 0.80651
## CurrPain            0.4799  1 122.068 0.48979
## HistoryPain         0.1872  1  71.905 0.66658
## VISAA               0.6710  1  77.114 0.41523
## CSA                 0.3650  1 148.968 0.54667
## Neovascularization 2.0856  1 148.083 0.15081
## KneeWall            0.0020  1 105.702 0.96473
## HeelRaise           0.0932  1 115.713 0.76064
```

Part II: Moth log-latency mixed models

Continuing with Fabusova et al. (2024) and 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

```
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)

moth2 <- 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) #Slight change from prior HW to keep
more observations!

moth2 <- moth2 %>% mutate(Treatment = fct_recode(factor(Treatment),
                                                 "Cold Phosphor" = "1",
                                                 "Warm Phosphor" = "2",
                                                 "Cold RGB" = "3",
                                                 "Warm RGB" = "4"), #5
recoding not needed - controls dropped
                          logLatency = log(Latency),
                          Colour = factor(Colour),
                          Spectra = factor(Spectra))
```

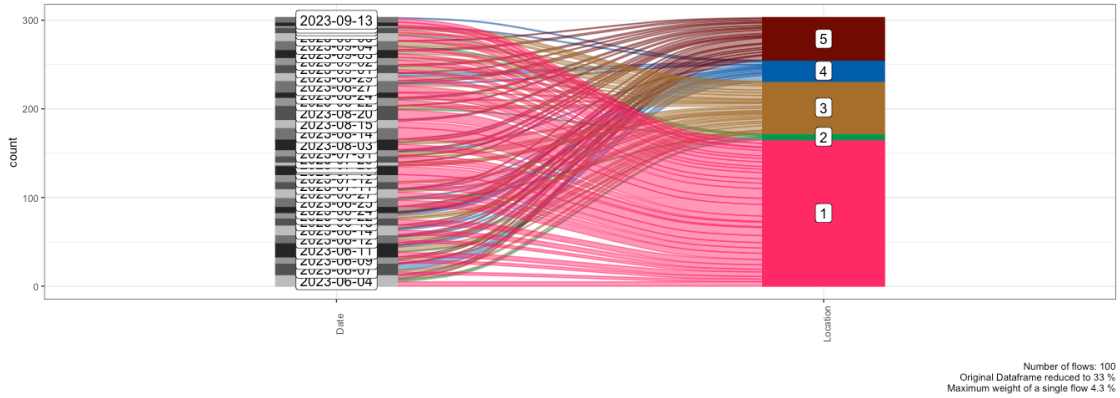
Usually we don't have such specific "genetic" (or familial?) information to group observations and I tend to worry more about time and space of data collection leading to similar observations that need to be built into the random effects in models. Since there are just four families of moths, it seems like accounting for that with a fixed effect might be reasonable, especially if the researchers wanted to directly compare characteristics between moth families (if the RQ had been "how do moth families differ on log-latency?").

By switching the random family effect to a fixed effect, we can account for the repeated measures on multiple moths collected on each sampling date at each location using a `Location and Date` set of random effects. This is more typical for what we would try to do in terms of using random effects to account for repeated measures and then still have some potential account for other systematic differences based on characteristics of interest across the subjects. But which way should we model the nesting? The following code tries the nesting in two directions:

```
tally(Date ~ Location, data = moth2)
```

```
##             Location
## Date          1  2  3  4  5
##   2023-06-04  5  2  1  0  5
##   2023-06-07  3  0  2  7  1
##   2023-06-09  3  0  1  0  3
##   2023-06-11  5  1  5  2  3
##   2023-06-12  3  0  3  1  2
##   2023-06-14  5  0  0  2  4
##   2023-06-18  6  0  2  0  0
##   2023-06-22  0  0  3  2  1
##   2023-06-24  0  0  0  1  6
##   2023-06-25  7  1  1  1  0
##   2023-06-27  6  0  2  1  1
##   2023-07-11  6  0  2  0  0
##   2023-07-12  8  0  0  0  0
##   2023-07-18  9  0  1  0  0
##   2023-07-19  0  0  1  0  0
##   2023-07-20  1  0  1  0  1
##   2023-07-29  5  0  1  0  1
##   2023-07-31  1  0  3  1  2
##   2023-08-03  6  0  2  0  4
##   2023-08-14 13  0  0  0  0
##   2023-08-15  7  0  0  0  2
##   2023-08-20 11  1  3  0  1
##   2023-08-22  6  0  1  0  2
##   2023-08-24  5  0  1  0  0
##   2023-08-27  7  0  2  1  3
##   2023-08-29  4  0  1  3  1
##   2023-09-01  3  0  4  0  2
##   2023-09-02  5  0  3  0  0
##   2023-09-03  4  0  3  0  2
##   2023-09-04  3  1  5  0  1
##   2023-09-06  6  1  2  0  0
##   2023-09-07  3  0  2  1  0
##   2023-09-08  1  0  1  0  0
##   2023-09-12  4  0  0  0  0
##   2023-09-13  4  0  0  1  0
```

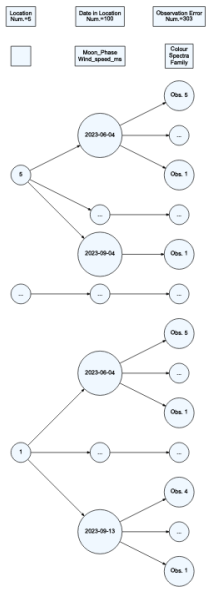
```
moth2 %>% dplyr::select(Date, Location) %>% mutate(Location =
factor(Location)) %>%
  alluvial_wide(fill_by = "last_variable")
```



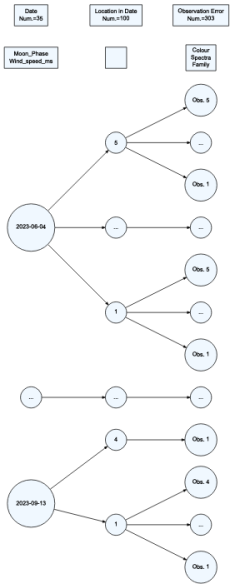

```
lmerLD <- lmer(logLatency ~ Moon_Phase + Colour + Spectra + Wind_speed_ms+
Family+ (1|Location/Date), data = moth2)

lmerDL <- lmer(logLatency ~ Moon_Phase + Colour + Spectra + Wind_speed_ms+
Family+ (1|Date/Location), data = moth2)

model_diagram(lmerLD, heightVal = 700)
```



```
model_diagram(lmerDL, heightVal = 700)
```



8) The study notes that they repeatedly visited the same five transects and caught and released some number of moths on each date they visited. The previous code fits the random effects two different ways. Which model, `lmerDL` or `lmerLD`, provides five locations and groups all the moths obtained on a given night within the location together in the diagram?

lmerLD, the first model diagram, provides five locations and groups all the moths obtained on a given night within the location together in the diagram. It starts with the 5 levels of location, then groups within each location by date.

9) Find the R-squareds for your selected model and write two sentences to interpret them. Make sure you are clear about the contents of the model being discussed in each sentence. Discuss what the results suggest about the aspects of the model and how well they explain the response.

The fixed effects of Moon_Phase, Colour, Spectra, Wind_speed_ms, and Family explain 13.1% of the variation in the log latency time. This suggests that while these variables contribute to the explanation of latency time, a large portion of the variation is still unaccounted for, potentially due to other factors not included in the model.The fixed effects with the random effects of location and date together explain 13.56% of the variation in log latency time. The small increase in variance explained by adding the random effects indicates that differences between locations and dates have a modest impact on latency time, and the model overall still leaves much of the variability unexplained.

```
r.squaredGLMM(lmerLD)

##             R2m       R2c
## [1,] 0.1308419 0.1355548
```

10) For your chosen model, `lmerLD` or `lmerDL`, write out the theoretical model for the response and define any aspects of that model, but you can leave the fixed effects represented in the model as μ_{ijk} and should not get in the x's and beta's part of it (that is very long here and something we can do another time). Do make sure you define the subscripts you are using and the distributions of the random parts of the model. To help you with that, we started writing out the model:

- $logLatency_{ijk} = \mu_{ijk} + u_{jk} + v_k + \epsilon_{ijk}$
- where $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ is the residual error term accounting for variation in individual observations, $u_{jk} \sim N\left(0, \sigma_{jk}^2\right)$ is the random intercept for date j nested within location k, $v_k \sim N(0, \sigma_k^2)$ is the random intercept for location k.
- and $i = 1, \ldots, n_{jk}$ (individual moths within date j and location k), $j = 1, \ldots, J_k$ (dates within location k), $k = 1, \ldots, K$ (locations).

```
summary(lmerLD)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: logLatency ~ Moon_Phase + Colour + Spectra + Wind_speed_ms +
##     Family + (1 | Location/Date)
##    Data: moth2
##
## REML criterion at convergence: 994.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.63413 -0.71301 -0.06347  0.77361  2.21870
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  Date:Location (Intercept) 0.008254 0.09085
##  Location      (Intercept) 0.000000 0.00000
##  Residual                  1.514009 1.23045
## Number of obs: 303, groups:  Date:Location, 100; Location, 5
##
## Fixed effects:
##                             Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)                  1.59648    0.46879  64.77715   3.406  0.00114
## Moon_PhaseFull_moon         -1.10965    0.51669  55.11403  -2.148  0.03616
## Moon_PhaseNew_moon           0.34417    0.38147  38.78105   0.902  0.37251
## Moon_PhaseThird_quarter     -0.17668    0.41717  53.44103  -0.424  0.67361
## Moon_PhaseWaning_crescent   -0.13369    0.37287  48.09076  -0.359  0.72151
## Moon_PhaseWaning_gibbous    -0.18345    0.34659  43.04986  -0.529  0.59931
## Moon_PhaseWaxing_crescent    0.08274    0.38972  47.69265   0.212  0.83277
## Moon_PhaseWaxing_gibbous     0.21014    0.36859  50.97713   0.570  0.57110
## Colourwarm                   0.28150    0.14752 284.85708   1.908  0.05736
## SpectraRGB                   0.19090    0.15141 287.26119   1.261  0.20841
```

```
## Wind_speed_ms                0.11052    0.08474  59.45308   1.304  0.19720
## FamilyErebidae               0.20891    0.23694 247.35734   0.882  0.37881
## FamilyGeometridae            0.06116    0.23360 284.01479   0.262  0.79364
## FamilyNoctuidae              0.53930    0.27538 266.04080   1.958  0.05123

## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
lmerLD %>% tbl_regression(intercept = T)
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	1.6	0.66, 2.5	0.001
Moon_Phase			
First_quarter	—	—	
Full_moon	-1.1	-2.1, -0.07	0.036
New_moon	0.34	-0.43, 1.1	0.4
Third_quarter	-0.18	-1.0, 0.66	0.7
Waning_crescent	-0.13	-0.88, 0.62	0.7
Waning_gibbous	-0.18	-0.88, 0.52	0.6
Waxing_crescent	0.08	-0.70, 0.87	0.8
Waxing_gibbous	0.21	-0.53, 0.95	0.6
Colour			
cold	—	—	
warm	0.28	-0.01, 0.57	0.057
Spectra			
LED	—	—	
RGB	0.19	-0.11, 0.49	0.2
Wind_speed_ms	0.11	-0.06, 0.28	0.2
Family			
Crambidae	—	—	
Erebidae	0.21	-0.26, 0.68	0.4
Geometridae	0.06	-0.40, 0.52	0.8
Noctuidae	0.54	0.00, 1.1	0.051

[1]CI = Confidence Interval

11) Note any resources used outside those provided during the class, impactful discussions with other students, or report NONE:

NONE