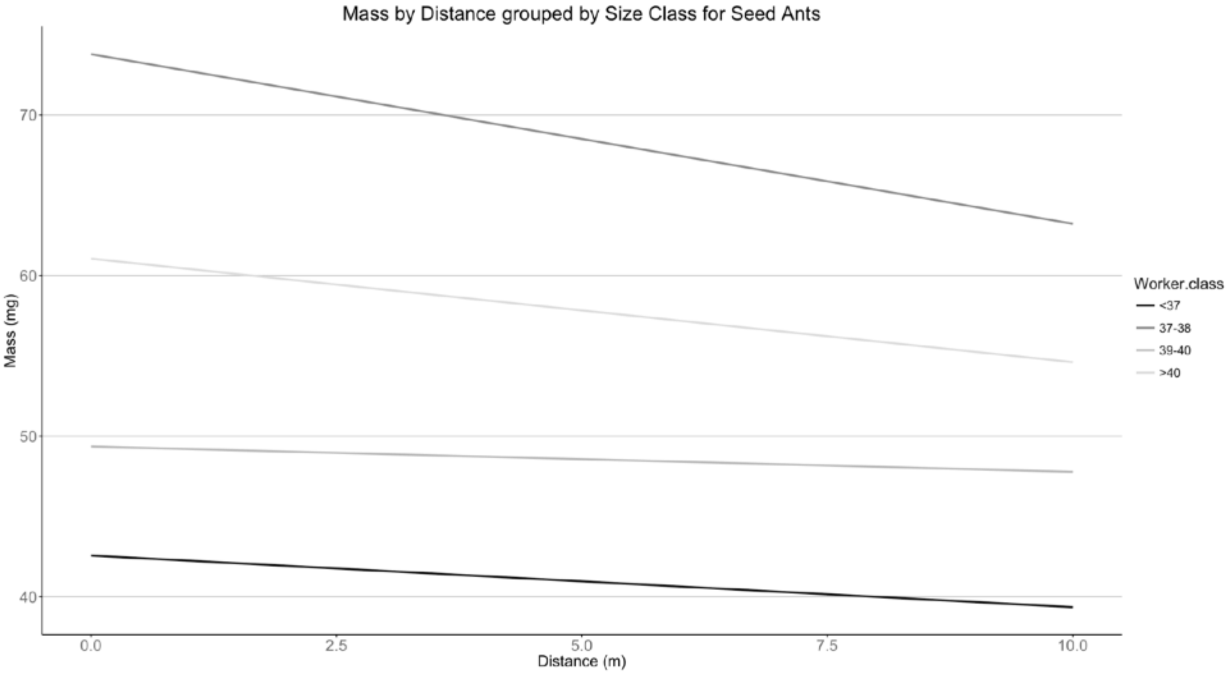
STA303
Assignment 3

by Justin SJ Lee

Let’s further investigate the ants data collected by Peter Nonacs from UCLA. The dataset and a data dictionary and some relevant background information can be found here. You’ll need to do some mild cleaning first. In particular, you should set appropriate levels for the size factor, and make Colony a factor as well. Distance you can leave as numeric.

1. We have been ignoring the fact that ants were from the same Colony. Let’s try to fix that by fitting a repeated measures model using lme() from the {nlme} package. For this part, we will use the Seed Ant data only. The question of interest is whether the relationship between Mass and Distance is the same for each size level.

(a) Present a plot showing how Mass varies by Distance, for each level of the size factor. Does it look like the relationship of Mass to Distance is the same for each size level?



It seems that while the general relationship between Mass and Distance is similar for the different size (worker) classes of ants, the slopes of the fitted lines differ from each other. However, their slopes are of the same direction (-). As Distance increases, the ants from all classes generally decrease in mass, but at different rates. We can conclude that while the relationship is not same for all levels of worker.class, the relationship is similar.

(b) Do a naïve ANCOVA analysis to answer the research question, assuming all observations are independent. Give the test statistic, df, p-value and a conclusion in English as to your findings. Present the ANOVA table (but not the summary output) in your solution.The question of interest is whether the relationship between Mass and Distance is the same for each size level.

H_o: The coefficient of Distance:Worker.class is equal to 0.
H_a: The coefficient of Distance:Worker.class is not equal to 0.

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distance	1	1759.186789	1759.186789	19.83933223	1.01E-05
Worker.class	3	53287.03661	17762.34554	200.315894	1.18E-88
Distance:Worker.class	3	972.6481447	324.2160482	3.656365508	0.012425765
Residuals	569	50454.18218	88.67167343	NA	NA

The test statistic is 3.656365508 with 3 numerator degrees of freedom and 574 denominator degrees of freedom. Since the p-value is 0.012425765, we can conclude that we have a significant interaction term with moderate evidence.

(c) Now fit a model using lme() with a random intercept for Colony (no nesting). Present the ANOVA output and answer the same research question with this model, giving your findings as above. Comment on any differences between the two approaches.

H_o: The coefficient of Distance:Worker.class is equal to 0.
H_a: The coefficient of Distance:Worker.class is not equal to 0.

Analysis of Variance Table

	numDF	denDF	F-value	p-value
(Intercept)	1	562	864.5757861	<.0001
Distance	1	562	44.44589912	6.26E-11
Worker.class	3	562	160.3362995	<.0001
Distance:Worker.class	3	562	4.457523692	0.004169234

The test statistic is 4.457523692 with numerator df of 3 and denominator df of 562. Since the p-value is 0.004169234, we can conclude that we have a significant interaction term with strong evidence. This approach is different from the previous model because it includes Colony as a random effect. This model is more accurate because it considers the fact that the ants are from different colonies which implies that they might have different behaviours and foraging strategies. In other words, this model considers the within group (colony) error and between group error (worker.class), while the model in (b) only considers the between group error(worker.class).

(d) Now fit the same model as the previous part, except considering Distance as a factor rather than a numerical covariate. State the findings as above, with ANOVA output. Which of the two LME models do you prefer? Why?

H_o: The coefficient of Distance:Worker.class is equal to 0.
H_α: The coefficient of Distance:Worker.class is not equal to 0.

Analysis of Variance Table

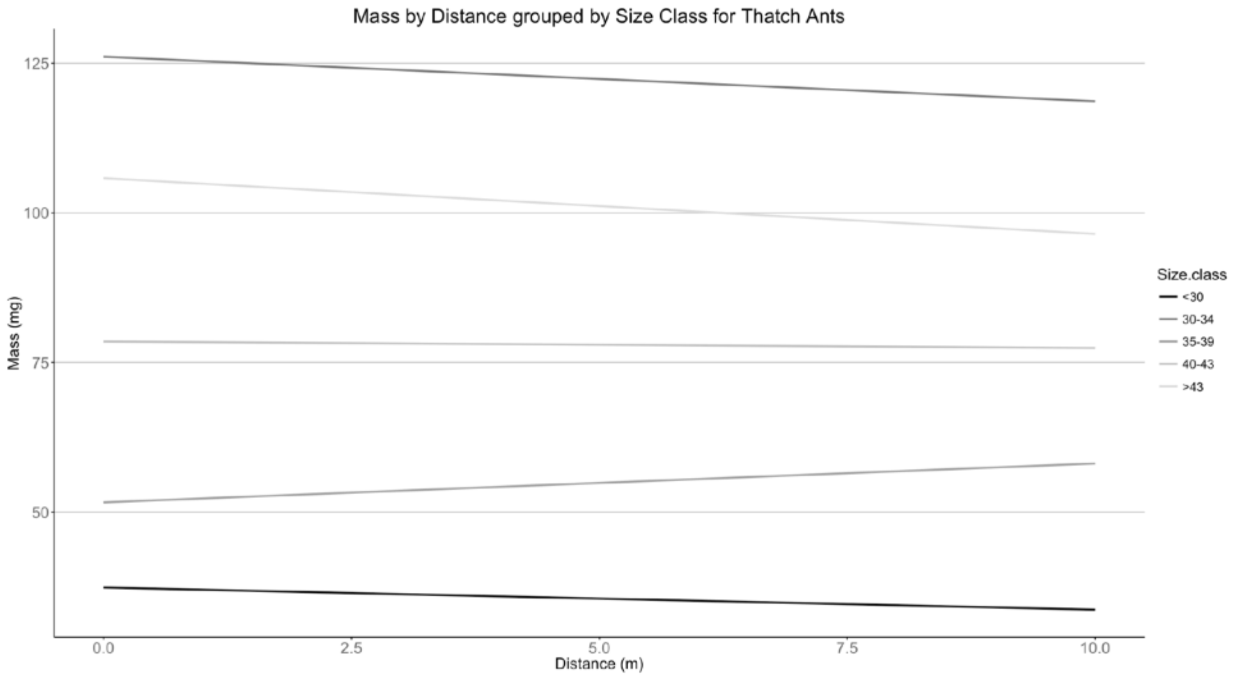
	numDF	denDF	F-value	p-value
(Intercept)	1	558	863.5946234	<.0001
Distance	2	558	22.08959013	5.85E-10
Worker.class	3	558	159.293132	<.0001
Distance:Worker.class	6	558	2.219349193	0.03986833

The test statistic is 2.219349193 with numerator df 6 and denominator df 558. Since the p-value is 0.0399, we can conclude that we have a significant interaction term with moderate evidence.

The AIC for the model fitted in (c) is 4098.813 and the AIC from this fit is 4077.903. Thus we can conclude that this model (d) is better than the model used in (c). In other words, the LME model in which Distance is treated as a factor has a lower AIC, thus it is the preferred model. This makes sense intuitively, since we measured distance in specific intervals, making distance a discrete variable. It is generally better practice to code discrete variables as categorial variables or factors rather than numeric variables.

2. Repeat the analysis from question 1. using the Thatch ants. If you get an error at any part, you can stop there but you must explain what caused the error.

(a) Present a plot showing how Mass varies by Distance, for each level of the size factor. Does it look like the relationship of Mass to Distance is the same for each size level?



From the plot above, we can say that the relationship between mass and distance for different size class(s) of ants is definitely not the same. While class: <30, 30-34, 40-43 and >43 exhibit a negative relationship between mass and distance, size class 35-39 seems to have a positive relationship. The slopes seem to be different for each line, meaning that the relationship between mass and distance change at different rates.

(b) Do a naive ANCOVA analysis to answer the research question, assuming all observations are independent. Give the test statistic, df, p-value and a conclusion in English as to your findings. Present the ANOVA table (but not the summary output) in your solution.The question of interest is whether the relationship between Mass and Distance is the same for each size level.

H_o: The coefficient of Distance:Size.class is equal to 0.
H_α: The coefficient of Distance:Size.class is not equal to 0.

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distance	1	453.8459992	453.8459992	1.798621934	0.180136176
Size.class	4	635212.2802	158803.07	629.3471474	2.88E-291
Distance:Size.class	4	3222.331118	805.5827796	3.19257823	0.012768359
Residuals	1185	299010.8699	252.329848	NA	NA

The test statistic is 3.19257823 with numerator degrees of freedom 4 and denominator degrees of freedom 1191. Since the p-value is 0.0128, we can conclude that we have a significant interaction term with moderate evidence.

(c) Now fit a model using lme() with a random intercept for Colony (no nesting). Present the ANOVA output and answer the same research question with this model, giving your findings as above. Comment on any differences between the two approaches.

Analysis of Variance Table

	numDF	denDF	F-value	p-value
(Intercept)	1	1175	11700.67903	<.0001
Distance	1	1175	1.766910507	0.18402292
Size.class	4	1175	642.602094	<.0001
Distance:Size.class	4	1175	2.764792856	0.026375767

H_o: The coefficient of Distance:Size.class is equal to 0.
H_α: The coefficient of Distance:Size.class is not equal to 0.

The test statistic is 2.764792856 with numerator df 4 and denominator df 1175. Since the p-value is 0.0264, we can conclude that we have a significant interaction term with moderate evidence. This approach is different from the previous model because it includes Colony as a random effect. This model is more accurate because it considers the fact that the ants are from different colonies which implies that they might have different behaviours and foraging strategies. In other words, this model considers the within group (colony) error and between group error (size.class), while the model in (b) only considers the between group error(size.class).

(d) Now fit the same model as the previous part, except considering Distance as a factor rather than a numerical covariate. State the findings as above, with ANOVA output. Which of the two LME models do you prefer? Why?

```
Error in MEEM(object, conLin, control$niterEM) : 
  Singularity in backsolve at level 0, block 1
```

The main problem with fitting a lme() with Distance as a factor is the way in which the lme() function treats empty cells. I've produced a copy of the table of frequencies by Distance and Size.class for thatch ants below. As you can see, [7m, <30] has a count of 0. When Distance is treated as a numerical variable, this is okay, but when Distance is set as a factor, this cell will be interpreted as NA by the lme() function. This is a clearly problem since you can't calculate the interaction effect for a NA value. This method seems to be causing the singularity error. I was able to confirm this by looking at the table of counts for seed ants, where we treated distance as a factor. As expected there were no cells with 0 counts and no errors occurred when lme() was implemented while distance was coded as a factor.

Distance\Size.class	<30	30-34	35-39	40-43	>43
0m	6	46	20	25	35
1m	13	65	22	82	92
4m	13	72	29	73	95
7m	**0 (NA)**	71	22	56	122
10m	3	61	11	45	116