Assignment 6

Robert Ivill 46012819

2023-05-18

Question 1: To prevent tree fungus, city parks treated trees with a chemical. To investigate whether the treatment is effective, seven parks that applied the treatment were chosen. At each park, four sites were chosen and the tree surface affected by fungus was measured (in m2) for two treated and two untreated trees.

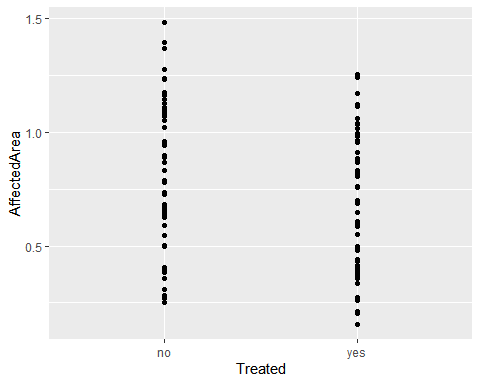
1. Load the data and explore it using str function and the graphs (use the code below). What do the graphs tell you?

After loading in the data we can see that the variables in the data frame are AffectedArea, Treated, Park and Site. By converting both the Park and Site variables to factors, various plots can be created.

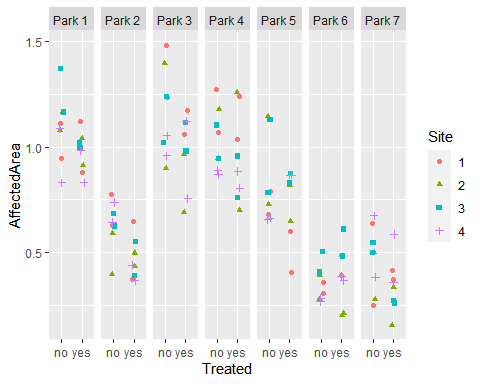
fungus\_df <- read.csv("/Users/robbi/Downloads/tree\_fungus.csv")  
str(fungus\_df)

## 'data.frame': 112 obs. of 5 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ AffectedArea: num 1.12 0.648 1.173 1.241 0.6 ...  
## $ Treated : chr "yes" "yes" "yes" "yes" ...  
## $ Park : int 1 2 3 4 5 6 7 1 2 3 ...  
## $ Site : int 1 1 1 1 1 1 1 1 1 1 ...

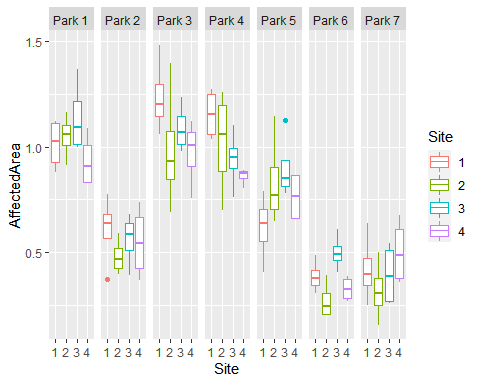
#This adds prefix "Park" to the Park variable for clear labels in the panels of the graphs  
fungus\_df$Park <- paste("Park", fungus\_df$Park)  
fungus\_df$Park <- as.factor(fungus\_df$Park)  
fungus\_df$Site <- as.factor(fungus\_df$Site)  
library(ggplot2)  
ggplot(fungus\_df, aes(x = Treated, y = AffectedArea)) + geom\_point()



ggplot(fungus\_df, aes(x = Treated,y = AffectedArea, color=Site, shape=Site))+  
geom\_jitter(width=0.1)+ facet\_grid(.~Park)



ggplot(fungus\_df, aes(x = Site,y = AffectedArea, color=Site, shape=Site))+  
geom\_boxplot()+ facet\_grid(.~Park)

 The first plot shows the relationship between the affected area variable and whether or not the tree was treated or not. We can see that trees that were treated had a lower affected area than those that did not get treated. The second plot shows the relationships between the affected area, the site and whether the tree was treated. It is also faceted to show the different parks. We can see that average affected area is different across different parks and that the trend of smaller affected areas for treated trees is universal across the various parks. It is difficult to see any relationships or trends when looking at the different sites. The third plot shows the distributions of affected areas across the different sites within each park. We can see that trends for different sites vary from park to park.

1. Which variables are random effects? Are the random effects crossed or nested? Explain why.

The random effect variables are the factors Park and Site because they show random variability across the model. The random effects are nested because each park has multiple sites and each site belongs to only one park.

1. Fit a linear mixed model with two random effects. Use the code below replacing response, fixed\_effect, random1, and random2 with corresponding variables. Answer the research question by fitting another model and significance testing (code not provided).

When creating the linear mixed model with two random effects, we replace ‘response’ with AffectedArea, ‘fixed\_effect’ with Treated, ‘random1’ with Park and ‘random2’ with Site. To answer the research question of ‘Does the treatment of tree fungus decrease the tree surface affected by fungus’ we have to fit another linear mixed model with only one intercept to test the null hypothesis that treatment has no effect on the affected area of the trees. We will call this fungus\_df.m0. Then we use the anova function to test for the significance of the fixed effects.

library(lme4)

## Warning: package 'lme4' was built under R version 4.2.3

## Loading required package: Matrix

fungus\_df.m1 <- lmer(AffectedArea ~ Treated + (1 | Park) + (1 |Park:Site),  
data=fungus\_df)  
fungus\_df.m0 <- lmer(AffectedArea ~ (1 | Park) + (1 |Park:Site), data=fungus\_df)  
anova(fungus\_df.m0,fungus\_df.m1)

## refitting model(s) with ML (instead of REML)

## Data: fungus\_df  
## Models:  
## fungus\_df.m0: AffectedArea ~ (1 | Park) + (1 | Park:Site)  
## fungus\_df.m1: AffectedArea ~ Treated + (1 | Park) + (1 | Park:Site)  
## npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)   
## fungus\_df.m0 4 -57.894 -47.020 32.947 -65.894   
## fungus\_df.m1 5 -69.900 -56.308 39.950 -79.900 14.006 1 0.0001822 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From the anova function, we can see that the p-value of the Treated variable in the alternative model is much less than 0.05 (0.0001822<0.05). Therefore, this suggests that the treatment variable is significant to the model and has an effect on the affected area of the tree fungus, which answers our research question.

1. Use summary and ranef functions to print the estimates. • What is the estimated average affected area for treated trees? • What is the average affected area for untreated trees in park 6 site 4?

summary(fungus\_df.m1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: AffectedArea ~ Treated + (1 | Park) + (1 | Park:Site)  
## Data: fungus\_df  
##   
## REML criterion at convergence: -71.9  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.03395 -0.70586 0.05594 0.58272 2.29425   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## Park:Site (Intercept) 0.005194 0.07207   
## Park (Intercept) 0.088266 0.29710   
## Residual 0.019483 0.13958   
## Number of obs: 112, groups: Park:Site, 28; Park, 7  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 0.79491 0.11464 6.934  
## Treatedyes -0.10237 0.02638 -3.881  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## Treatedyes -0.115

ranef(fungus\_df.m1)

## $`Park:Site`  
## (Intercept)  
## Park 1:1 -0.005818453  
## Park 1:2 0.012105214  
## Park 1:3 0.057594032  
## Park 1:4 -0.047333897  
## Park 2:1 0.027275957  
## Park 2:2 -0.038655287  
## Park 2:3 0.003900152  
## Park 2:4 -0.003654241  
## Park 3:1 0.090713984  
## Park 3:2 -0.038952035  
## Park 3:3 0.013016212  
## Park 3:4 -0.046035176  
## Park 4:1 0.086514752  
## Park 4:2 0.016661530  
## Park 4:3 -0.024392256  
## Park 4:4 -0.064449271  
## Park 5:1 -0.082496444  
## Park 5:2 0.028860624  
## Park 5:3 0.063649158  
## Park 5:4 -0.007951121  
## Park 6:1 0.002377306  
## Park 6:2 -0.056869045  
## Park 6:3 0.061149887  
## Park 6:4 -0.027989619  
## Park 7:1 0.001020943  
## Park 7:2 -0.051474877  
## Park 7:3 -0.012099566  
## Park 7:4 0.043331536  
##   
## $Park  
## (Intercept)  
## Park 1 0.28117272  
## Park 2 -0.18918433  
## Park 3 0.31848971  
## Park 4 0.24358298  
## Park 5 0.03504216  
## Park 6 -0.36247450  
## Park 7 -0.32662874  
##   
## with conditional variances for "Park:Site" "Park"

From the summary function output, we can see the estimate for the Treated coefficient is -0.10237. This number represents the difference of affected area of untreated and treated trees. As it is negative, this means that treated trees have lesser affected area than those untreated. The average affected area is calculated as the intercept estimate plus the Treatedyes estimate, which is 0.69254 m2. To find the average affected area of untreated trees in park 6 site 4, we take the intercept of untreated trees of 0.79491 and add the intercept of park 6 site 4 from the ranef output of -0.027989619. So the average affected are of untreated trees in park 6 site 4 is 0.76692 m2.

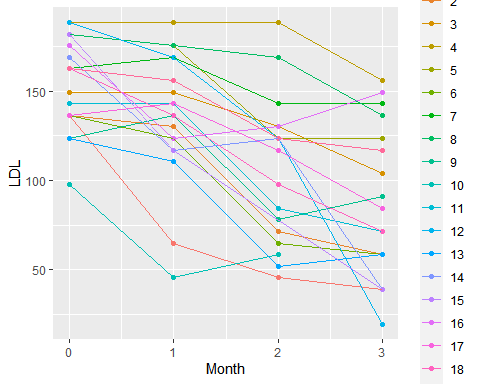
Question 2: The data file cholesterol.csv contains cholesterol level measurements (LDL) taken monthly from 19 patients who were treated for high cholesterol levels. Three patients have missing measurements. The research question is to estimate the average rate of cholesterol drop.

1. Load the data and check its structure. Convert integer variables that are categorical into factors (replace X with the necessary variable) and plot the graphs using the code below. Describe the trends you see in the data.

chol\_df <- read.csv("/Users/robbi/Downloads/cholesterol.csv")  
str(chol\_df)

## 'data.frame': 73 obs. of 4 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Patient: int 1 1 1 1 2 2 2 2 3 3 ...  
## $ Month : int 0 1 2 3 0 1 2 3 0 1 ...  
## $ LDL : num 136.5 65 45.5 39 136.5 ...

chol\_df$Patient <- as.factor(chol\_df$Patient)  
library(ggplot2)  
ggplot(aes(y=LDL, x=Month, group=Patient, color=Patient), data=chol\_df)+geom\_point()+geom\_line()

 We confirm that the variables in the data are Month, Patient and LDL. We see from the plot that each patient sees a decrease in cholestoral levels over the 3 months. However, some patients can see an increase in LDL over a month long period. The change in LDL seems to vary from patient to patient.

1. Fit three linear mixed models: • a random intercept model • a random slope model • a random slope and intercept model (with correlation between the slopes and intercepts) Use the code below replacing the response, fixed\_effect, and random\_effect with corresponding variables.

In the code below, we replace response with LDL, fixed\_effect with Month, and random\_effect with Patient.

library(lme4)  
chol.m1 <- lmer(LDL~1+Month+ (1|Patient), data=chol\_df)  
chol.m2 <- lmer(LDL~1+Month + (0+Month|Patient), data=chol\_df)  
chol.m3 <- lmer(LDL~1+Month + (1+Month|Patient), data=chol\_df)

1. Print summary for model chol.m3. What is the correlation of random effects? How would you interpret it?

summary(chol.m3)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: LDL ~ 1 + Month + (1 + Month | Patient)  
## Data: chol\_df  
##   
## REML criterion at convergence: 678.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.33827 -0.45538 0.02218 0.60047 1.55330   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr  
## Patient (Intercept) 605.3 24.60   
## Month 100.0 10.00 0.15  
## Residual 332.9 18.25   
## Number of obs: 73, groups: Patient, 19  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 156.243 6.706 23.30  
## Month -24.377 3.017 -8.08  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## Month -0.186

The summary output shows the estimated coefficients, standard errors and t-values for the fixed and random effects. The correlation of the random effects is 0.15. This represents the relationship between the random intercepts and slopes of the patients. As this value is positive, we can interpret this by saying that patients with greater intercepts tend to have a steeper slope of their LDL over the months.

1. Fit in a model with a random slope and a random intercept without correlation between the slopes and intercepts replacing the response, fixed\_effect, and random\_effect with corresponding variables in the code below.

In the code below, we replace response with LDL, fixed\_effect with Month, and random\_effect with Patient.

chol.m4 <- lmer(LDL~1+Month + (1|Patient) +  
(0+Month|Patient), data=chol\_df)

1. Use AIC function to calculate Akaike information criterion. Which model has the best fit?

AIC(chol.m1, chol.m2, chol.m3, chol.m4)

## df AIC  
## chol.m1 4 695.8332  
## chol.m2 4 705.5242  
## chol.m3 6 690.7198  
## chol.m4 5 688.8467

The output of the AIC function shows the Akaike information criterion for each model. The model with lowest AIC is said to have the best fit, meaning that the 4th model has the best fit for our data.

1. Print estimated fixed and random effects for the preferred model using fixef and ranef functions. • What is the estimated average cholesterol drop rate (average slope)? • Which patient has the fastest drop rate? To access random intercepts and slopes as vectors use ranef(best\_model)[[1]][1] and ranef(best\_model)[[1]][2] respectively. Replace best\_model with the corresponding model.

In the code below, we replace best\_model with chol.m4 as it has the best fit.

fixef(chol.m4)

## (Intercept) Month   
## 156.2410 -24.3372

ranef(chol.m4)

## $Patient  
## (Intercept) Month  
## 1 -33.017075 -7.377560  
## 2 -12.647043 -4.248545  
## 3 4.225020 5.839020  
## 4 37.112672 12.573597  
## 5 22.062088 3.752861  
## 6 -15.212719 -4.489679  
## 7 15.628444 11.364116  
## 8 28.922975 8.969942  
## 9 -16.300395 3.909952  
## 10 -52.584303 -3.336465  
## 11 -5.105394 -2.325198  
## 12 27.054647 -16.711217  
## 13 -25.786191 -3.054318  
## 14 4.327361 -8.321120  
## 15 5.249786 -12.941007  
## 16 6.972600 11.360302  
## 17 -4.120060 3.435312  
## 18 2.902401 -4.001691  
## 19 10.315187 5.601700  
##   
## with conditional variances for "Patient"

The estimated average cholesterol drop rate can be found from the estimate for Month in fixed effects, which is -24.3372. This means that for each month, the average cholesterol LDL drops by 24.3372 units. The patient with the fastest drop rate is the one with the steepest negative slope in the random effects estimates. This is patient 12 as it has a slope estimate of -16.711 in the random effects, meaning that their overall drop rate is 16.711+24.3372 = 41.0482.