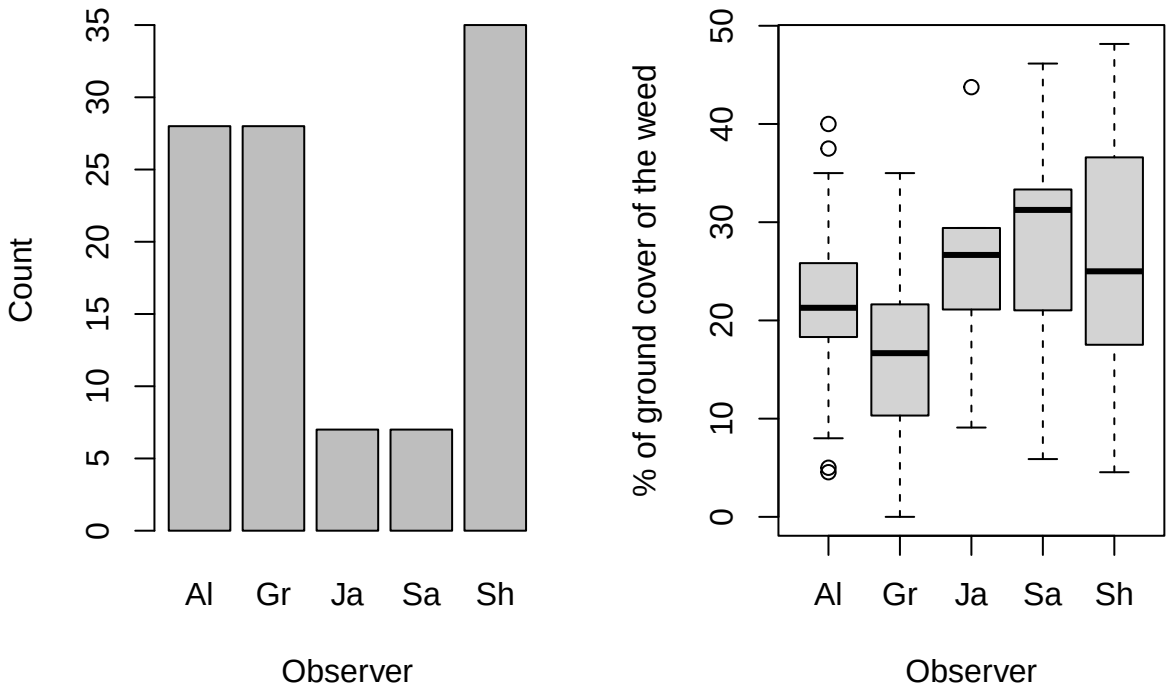
Assignment 5 STAT 315-463: Multivariable Statistical Methods and Applications

Lisa Lu 31088272

```
# Read in data file
initobs <- read.table("initobs.csv", header = TRUE, sep = ',', na.strings = "na")
```

a) Explain why observer should be included in a model as a random effect.

```
pardef <- par()
par(mfrow = c(1,2))
plot(as.factor(initobs$Observer), xlab="Observer", ylab="Count")
plot(as.factor(initobs$Observer),initobs$Bc/initobs$Steps * 100,
     xlab="Observer", ylab="% of ground cover of the weed")
```



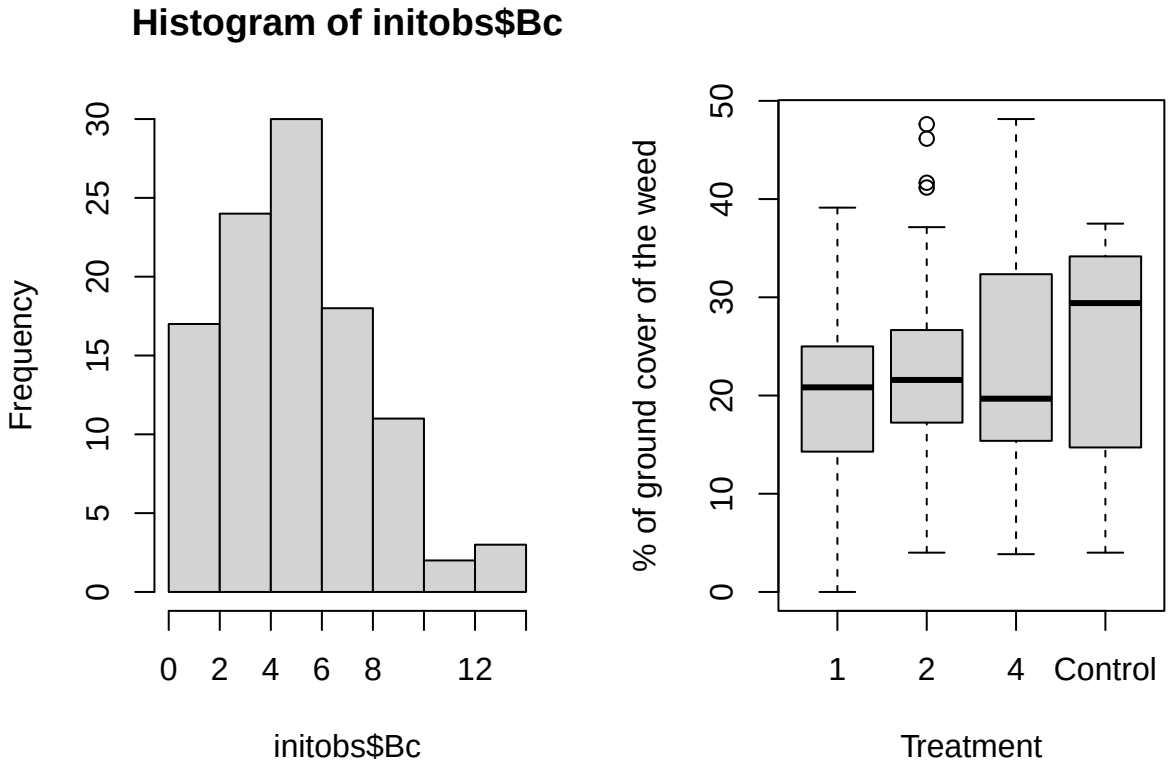
From the left plot above, we can see that the number of occurrence of each observer varies greatly. The right plot also suggested that the distribution of the percentage of ground covered by the weed differs from each observer. Therefore, the observers are the source of random variation, which should be taken into

consideration when choosing the model as it brings potential effects on the relationship between the control of the weed and the different treatments.

b) What distribution would be appropriate for these data and why?

The dependent variable here, namely, the occurrence of the weed (buttercup) are not continuous. In a more visualised way, from the histogram of the frequency of the buttercup shown in the left plot, we can see that it does not follow the normal distribution. Therefore it is not suitable to use normal linear regression models. Here, we would like to investigate the control of the weed by using different treatments. Each test has two potential outcomes: Has Bc or No Bc, and each test is exclusive or independent of one another. As shown in the right plot below, the data follows a discrete distribution. Therefore, binomial distribution should be considered as an appropriate way to fit the model.

```
pardef <- par()
par(mfrow = c(1,2))
hist(initobs$Bc)
plot(as.factor(initobs$Trt), initobs$Bc/initobs$Steps * 100,
     xlab="Treatment",ylab="% of ground cover of the weed")
```



c) Fit an appropriate random effects model to these data

```
library(lme4)
initobs$Plot <- as.factor(initobs$Plot)
model.0 <- glmer(cbind(Bc, NotBc) ~ Trt + (1|Observer),
            family = binomial, data = initobs)
model.1 <- glmer(cbind(Bc, NotBc) ~ Trt + (0 + Trt|Observer),
            family = binomial, data = initobs)
```

```
model.2<- glmer(cbind(Bc, NotBc) ~ Trt + (1|Observer) + (1|Farm) + (1|Plot),
            family = binomial, data = initobs)
model.3 <- glmer(cbind(Bc, NotBc) ~ Trt + Farm + (1|Observer),
            family = binomial, data = initobs)
model.4 <- glmer(cbind(Bc, NotBc) ~ Trt + Farm + Plot + (1|Observer),
            family = binomial, data = initobs)
anova(model.0, model.1, model.2,model.3, model.4)
```

```
## Data: initobs
## Models:
## model.0: cbind(Bc, NotBc) ~ Trt + (1 | Observer)
## model.2: cbind(Bc, NotBc) ~ Trt + (1 | Observer) + (1 | Farm) + (1 | Plot)
## model.3: cbind(Bc, NotBc) ~ Trt + Farm + (1 | Observer)
## model.4: cbind(Bc, NotBc) ~ Trt + Farm + Plot + (1 | Observer)
## model.1: cbind(Bc, NotBc) ~ Trt + (0 + Trt | Observer)
##         npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## model.0    5 490.26 503.53 -240.13   480.26
## model.2    7 481.80 500.37 -233.90   467.80 12.4660  2   0.001964 **
## model.3    9 477.48 501.36 -229.74   459.48  8.3185  2   0.015619 *
## model.4   12 469.85 501.69 -222.92   445.85 13.6313  3   0.003453 **
## model.1   14 506.86 544.01 -239.43   478.86  0.0000  2   1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Several models were fitted and compared in this section. My first thought was that Farm and Plot could work as the random effects. However, it seemed like taking these two variables as the fixed effect variables fit better models, which indicated that Farm and Plot variables along with the Treatment had a constant effect on the occurrence of the weed. I also tried the random effect models with random slopes and with both random intercepts and slopes. Nevertheless, these did not work out well. From the ANOVA output above, we can see that the model with **Trt**,**Farm**, and **Plot** as the predictor variables, and **Observer** as the random effect has the best fit.

```
summary(model.4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(Bc, NotBc) ~ Trt + Farm + Plot + (1 | Observer)
##    Data: initobs
##
##      AIC      BIC   logLik deviance df.resid
##    469.8    501.7   -222.9    445.8       93
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0735 -0.8022  0.0589  0.5238  3.7300
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Observer (Intercept) 0.045    0.2121
## Number of obs: 105, groups:  Observer, 5
##
```

```
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.57682    0.21374  -7.377 1.62e-13 ***
## Trt2         0.18289    0.19774   0.925  0.35502
## Trt4         0.49236    0.18796   2.620  0.00880 **
## TrtControl   0.49519    0.19142   2.587  0.00968 **
## FarmB        0.20066    0.20077   0.999  0.31757
## FarmF       -0.31418    0.20819  -1.509  0.13126
## FarmR       -0.39441    0.21333  -1.849  0.06448 .
## FarmS        0.05463    0.19537   0.280  0.77978
## Plot2        0.45813    0.18127   2.527  0.01149 *
## Plot3       -0.00930    0.17666  -0.053  0.95802
## Plot4        0.50351    0.18949   2.657  0.00788 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Trt2   Trt4   TrtCnt FarmB  FarmF  FarmR  FarmS  Plot2
## Trt2        -0.484
## Trt4        -0.507  0.552
## TrtControl -0.503  0.542  0.570
## FarmB       -0.523  0.009  0.004  0.015
## FarmF       -0.490 -0.004 -0.007  0.002  0.725
## FarmR       -0.480 -0.001 -0.005 -0.005  0.709  0.681
## FarmS       -0.522 -0.002 -0.002 -0.001  0.768  0.738  0.722
## Plot2        0.006 -0.518  0.000  0.000 -0.004 -0.015 -0.010 -0.003
## Plot3        0.006  0.001 -0.446  0.001  0.001 -0.010 -0.010 -0.011  0.000
## Plot4       -0.513  0.547  0.576  0.565  0.021  0.007  0.007  0.008  0.000
##             Plot3
## Trt2
## Trt4
## TrtControl
## FarmB
## FarmF
## FarmR
## FarmS
## Plot2
## Plot3
## Plot4        0.000
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 3 columns / coefficients
```

- a. Discuss the results of the analysis, include comments about the following:
  - i). Scaled residuals As we can see from the scaled residuals summary above, the median is closer to the 3rd Quartile, which means the distribution of the residuals is bit left-skewed.
  - ii). Random effects

  This section shows how much of the variation in having Bc is explained by the Observer variable, which in this case is 0.045. Because it is quite close to zero, it suggests a relatively small impact the Observer random effect has on explaining the variability of whether there is weed.

  - iii). Fixed effects

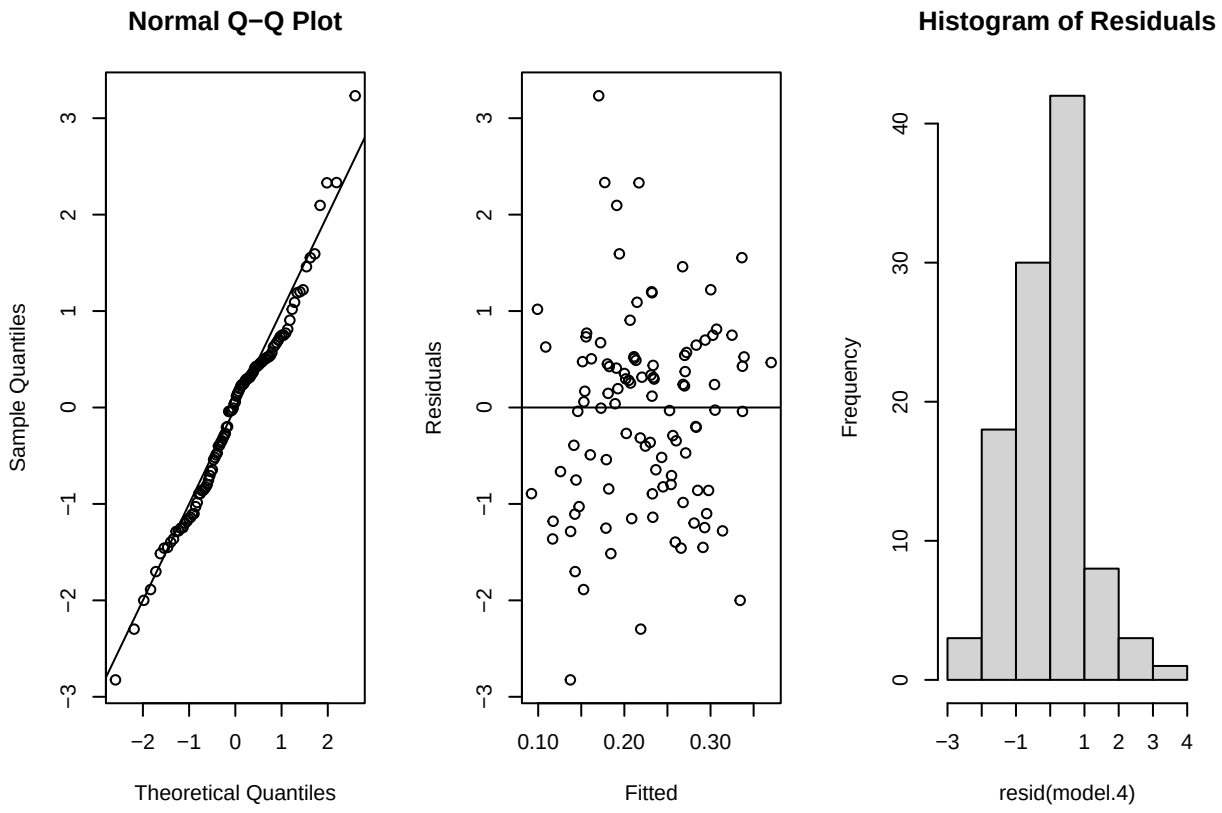
When fitting the model, the predictor variables were considered as factors. From the section of the fixed effects above, we can see that the Trt4, TrtControl, and Plot4 have the relatively more significant effects on the response, following by Plot2, and then FarmR. If we look closely at the estimated coefficients, except the FarmR, all the variables have the positive efficients, which means it is more likely to have the weed when the treatment is 4 or control, or when the plot is 2 or 4. However, the coefficient of FarmR is negative, which indicates on FarmR, it is less likely to have the weed.

- b. What would your overall conclusion be?

  In conclusion, it seems like with no treatment or with too many treatments both will not bring positive effect on preventing the weed from occurring. Also, there are differences among the farms and the plots on each farm. The random effects of observer does not have much significant impact on the occurrence of the weed.

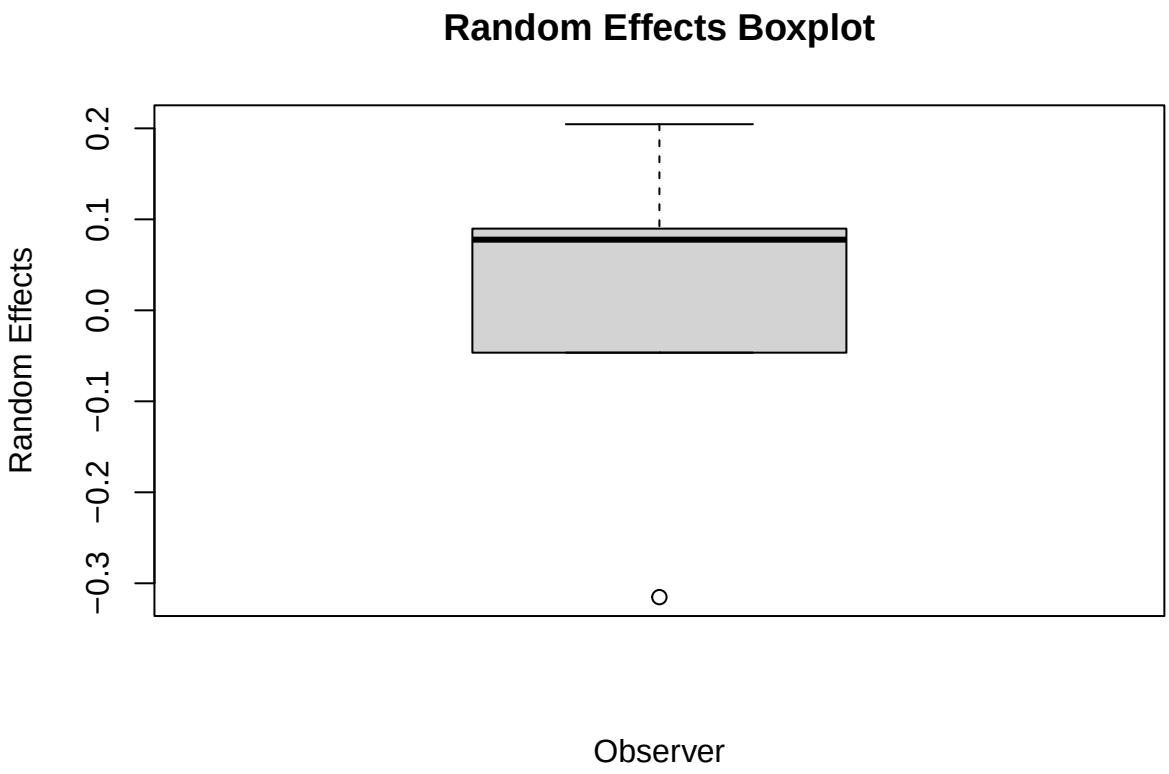
d) Draw graphs of the residuals and the random effects and comment on these.

```
# Create plots of the residuals
par(mfrow=c(1,3))
qqnorm(resid(model.4, main = "Residual Plot"))
abline(0,1)
plot(fitted(model.4), resid(model.4), xlab = "Fitted", ylab="Residuals")
abline(0,0)
hist(resid(model.4), main = "Histogram of Residuals")
```



The above three plots show the distribution of the residuals of the model. The Q-Q plot (the first plot) and the residual plot (the second plot) seem like there is no specific pattern in the distribution, but there are a few obvious outliers. From the last plot, namely the histogram of the residuals, we can clearly see that the distribution may be "heavy tailed".

```r
# Create plot of random effects
re1 <- ranef(model.4)$Observer
boxplot(re1, main="Random Effects Boxplot",
       xlab="Observer",ylab="Random Effects")
```



This boxplot suggests that the distribution of the random effects is heavily left-skewed and there is one outlier. This can be caused by the imbalanced occurrences of the observers. For example, from the two plots shown in question a), it is shown that Ja and Sa had the least occurrence, however, the proportion of the ground covered by the weed they reported was quite high in average. Because these two groups had higher values, this could cause the skewness of the random effects.