STAT446 ASSIGNMENT 3

(Chen Liang, Student ID: 46275313)

Introduction

Herbicides are the most active and rapidly developing agricultural chemicals in the world. The use of herbicides is increasing year by year, the proportion of which in various agriculture products tends to grow up.

After entering the twenty-first Century, population, food and environment are prominent problems that human being has to face. By 2030, the per capita cultivated land will be reduced to $0.08hm^2$. Therefore, increasing the unit output of crops is the only method to meet the food needs of mankind. Farmland grass damage is one of the most important factors leading to crop reduction, resulting in global agricultural losses of $95 each year. The use of herbicides not only effectively reduces the economic losses caused by weeds, but also improves the efficiency of weeding and saves labour. However, with the large-scale and extensive application of herbicides, the problem of herbicide resistance has become more and more prominent, and the speed of resistance formation is also accelerating. Therefore, studying herbicide is very important for the growth of crops producing.

This experiment focuses on the effectiveness of herbicides on Ranunculus Acris. A total of 18 paddock from nine farms participated in the experiment. The experimental data provides data of farms and paddies which each observation point belongs to, if the point is mown or not, herbicides used, and the coverage of Ranunculus and space area after the use of herbicides.

Part 1 Model for predicting Buttercuppc

Method

The main tool for this analysis was R Studio, which was a compiler based on R, providing many embedded functions and packages for data analysis and model evaluation.

1. Data processing

The data collected 864 observations for 6 variables, which were Farm, Paddock, Mow, Herbicide, Buttercuppc and BareGrnd (using dim() and names() function to check).

Summary () was conducted to display the statistics of each variable in dataset. is.na() was then used to figure out the row numbers of observations with missing value, which were needed to be removed to ensure that the model would not be affected by them. After the preliminary data cleaning, the scatterplot matrix (using ggpairs() function in R) was used to present the relationship of all the variables.

Since the fitting process was based on the Analysis of Variance Model, which required the corresponding variable was approximately normal distributed. If the data did not meet this requirement, it needed to be transformed. Therefore, geom_histogram() in package ggplot2 was used to plot a histogram to show the distribution of the responding variable, which was Buttercuppc.

After the corresponding variable was transformed, diagram of four predictors and the responding variable was drawn. There were many levels in variable Herbicide and Farm, so it was clearer to use point graph to show. For the variable Mow and Paddock which contained just two levels and the distribution of observations for each level was intensive, so boxplots were used.

2. Model Fitting

Since 'Buttercupcc' had already been transformed into approximate normal distributed, so first of all, 'Mow', 'Herbicide', their interaction, and the interaction of 'Farm' and 'Paddock' were treated as fixed effects which were added into a variance analysis model being built using aov () function. qqnorm () and shapiro.test () were then used to test the distribution of residuals. Shapiro.test () assumes that the residuals are normal distribution in the inspection process. If the p value is large, the null hypothesis should not be denied. Therefore, if the point of qqnorm () is close to a straight line and the p value of shapiro.test () is great enough, the assumption of the model is set up.

The lmer () function is executed to establish a linear mixed model. Among them, H'and M' are fixed effects, and'F'acts as random effects. And use summary () to get the relevant parameters and statistical data of the model.

Result

1. Data Processing

From summary, it could be seen that the data is well balanced because the counts were equal in each level for all the categorical predictors. At the same time, there were 48 missing value in total which were removed. The dataset then contained 816 observations for 6 variables.

Table 1 Summary of Buttercuppc

Farm	Paddock	Mow	Herbicide
Farm A : 96	Dry:432	Mow :432	Aminopyralid :108
Farm B : 96	Wet:432	No Mow:432	Aminopyralid+triclopyr:108
Farm C : 96			Flumetsulam :108
Farm D : 96			MCPA :108
Farm E : 96			MCPB :108
Farm F : 96			MCPB+bentazone :108
(Other):288			(Other) :216

Buttercuppc	BareGrnd
Min. : 0.000	Min. : 0.00
1st Qu.: 0.100	1st Qu.: 15.00
Median : 1.025	Median : 55.00
Mean : 4.204	Mean : 81.23
3rd Qu.: 4.250	3rd Qu.:115.00
Max. :65.000	Max. :525.00
NA's :48	NA's :48

It can be seen from Figure 1 that after removing the missing data, the data set was still balanced at all levels of the important predictors, Mow and Herbicide, and would not affect the model fitting.

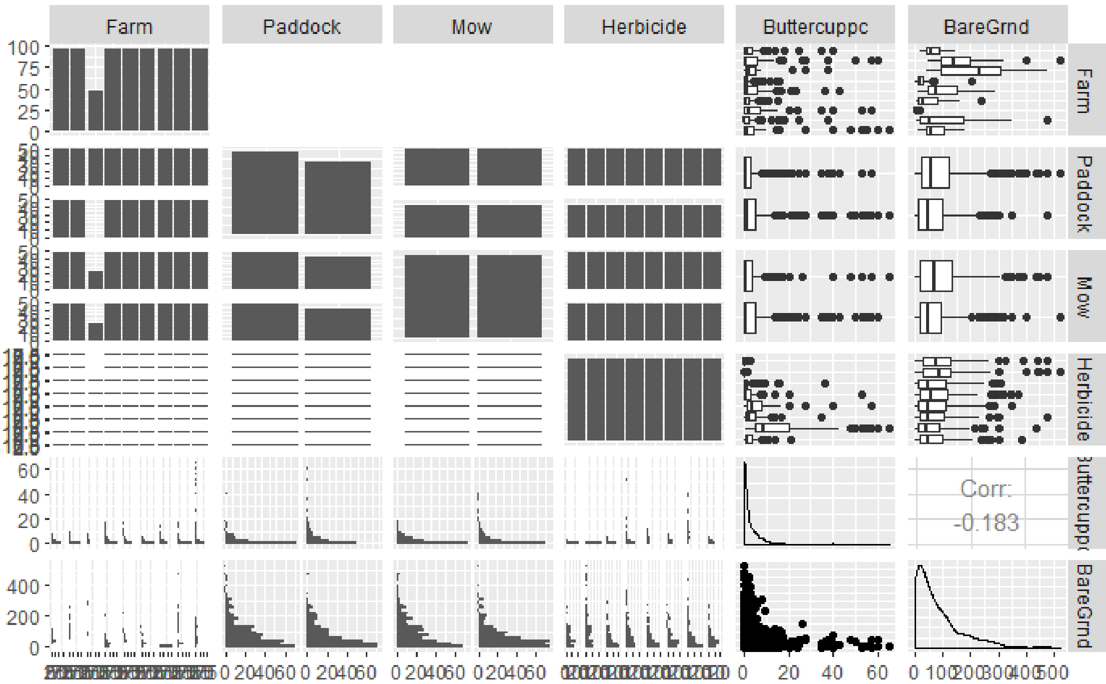


Figure 1

Figure 2 showed that the distribution of Buttercuppc was quite different when using different herbicide. For example, when ‘Aminopyralid+triclopyr’ or ‘Aminopyralid’ was used, there was an extremely low percentage of plot was covered by buttercup and the data was showing evidence of low variability. However, the percent cover of buttercup in each plot would be high and with high variability after using Nil.

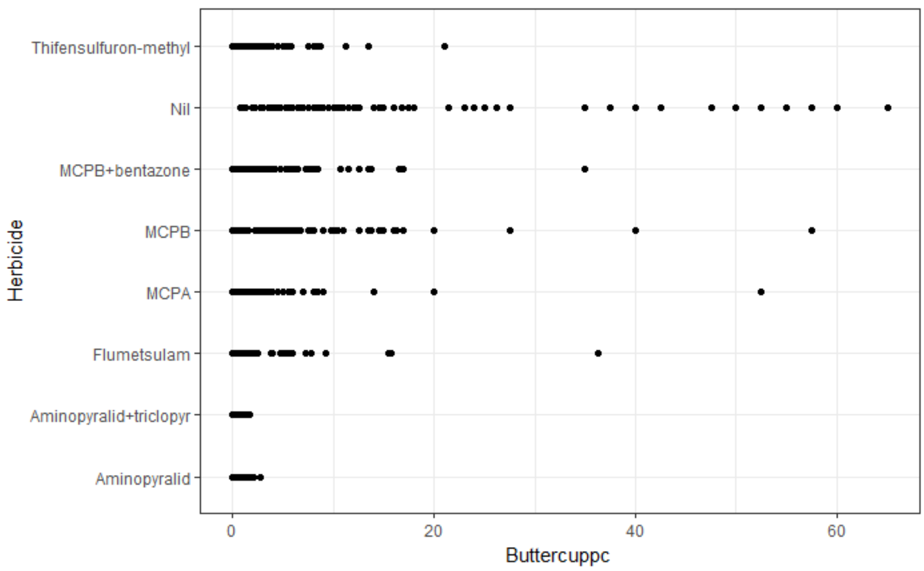


Figure 2 Distribution of Buttercuppc for each level of Herbicide

As could be seen from Figure 3, for 'Mow' and 'No Mow', the trend of distribution of Buttercup was similar, but when the plot was mown, Buttercup had a higher peak which was quite close to 0, representing the overall value was lower in this level.

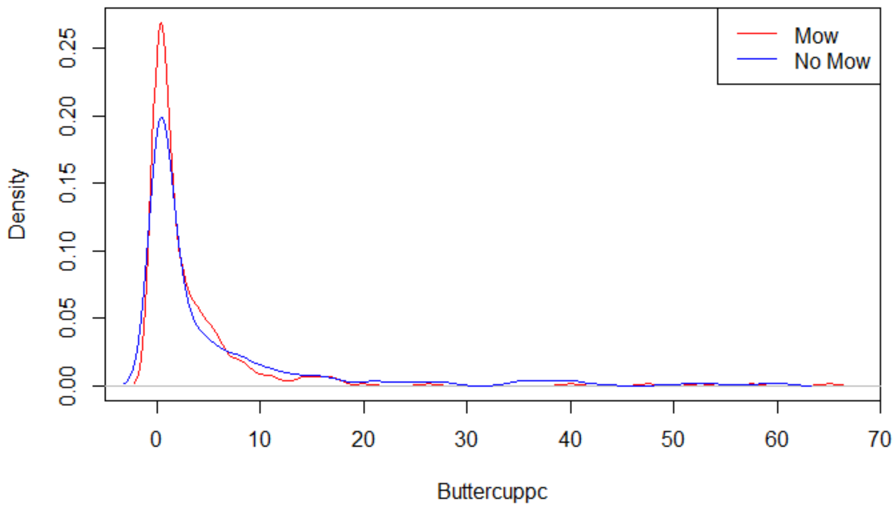


Figure 3 Distribution of Buttercuppc for Mow

Figure 4 showed that the distribution of responding variable in each level of interaction between Herbicide and Mow. As can be seen from the Figure, there was a clear difference between the distributions, and percent cover of buttercup in each plot was the highest when using Nil and without being mown. At the same time, the data which varied most was the level using Nil and being mown.

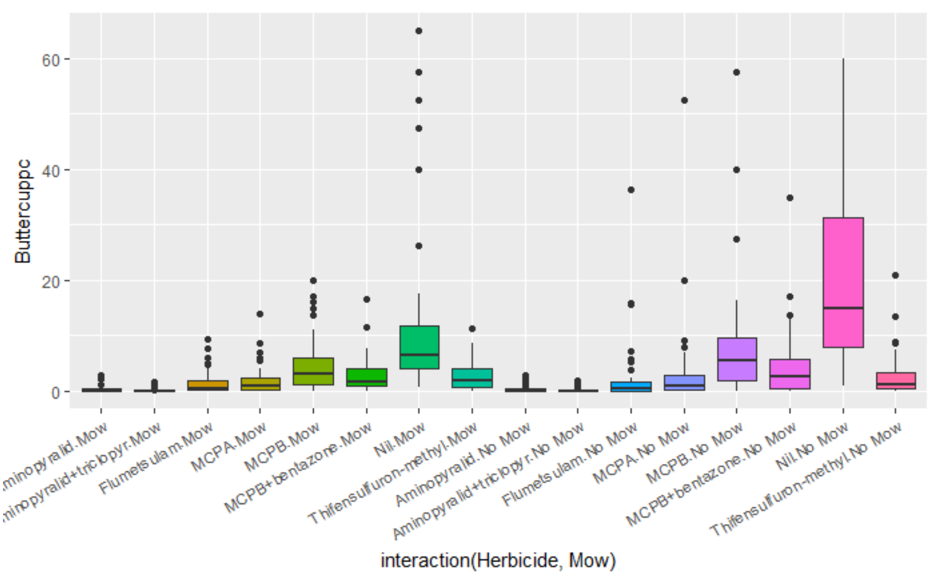


Figure 4 Distribution of Buttercuppc for each level of Interaction (Herbicide, Mow)

The differences of both variability and quartiles could be checked in Figure 5. The level 'Farm I. Wet' varied most and the level 'Farm B. Wet' had the highest median. Since the figure was showing an obvious evidence that the responding variable distributed differently for levels in interaction between Farm and Paddock, the interaction should be added into model fitting.

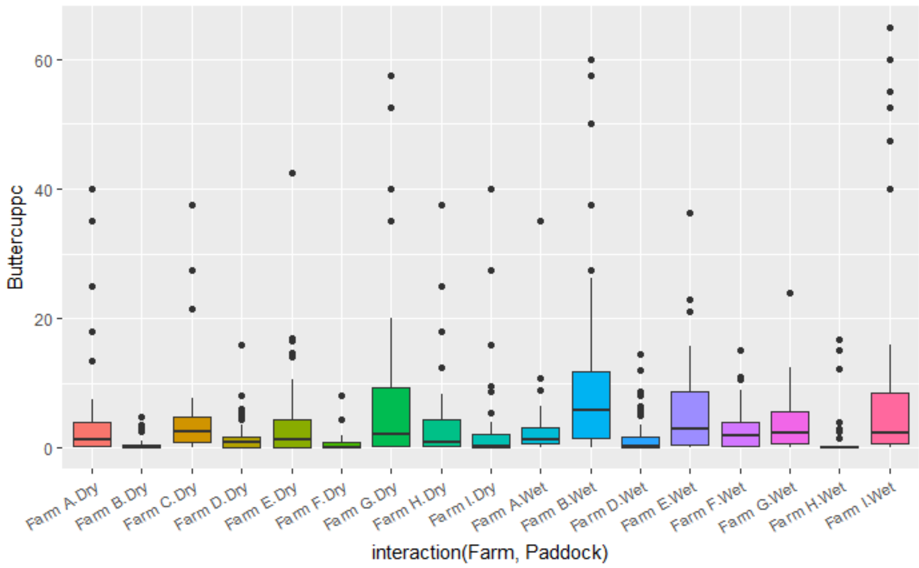
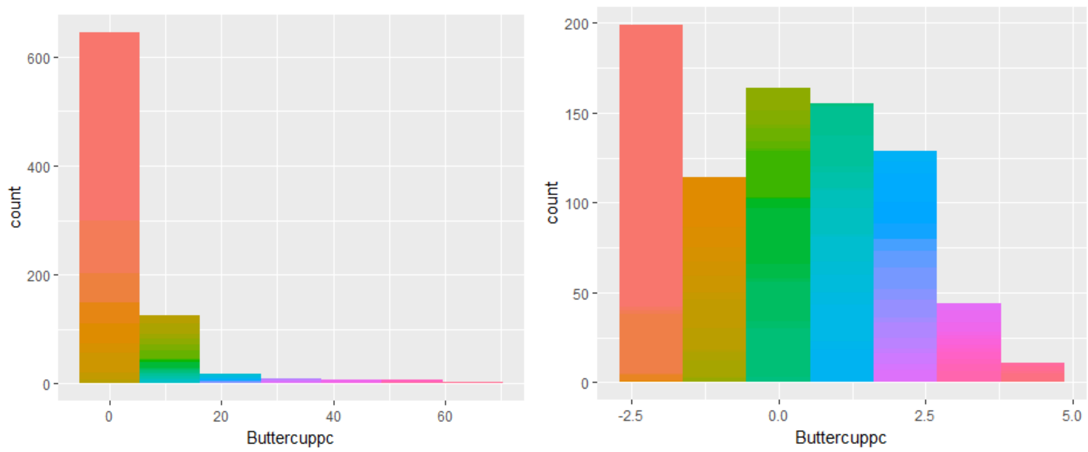


Figure 5 Distribution of Buttercuppc for each level of interaction (Farm, Paddock)

It could be seen from Figure 6 (a) that the original data of the responding variable was monotonous decreasing, and it displayed approximately normal after being transformed as shown in (b).



(a) **(b)**

Figure 6 Distribution of 'Buttercuppc' before and after being transformation

2. Model Fitting

Variance Analysis Model

Table 1 showed the summary of the variance analysis model. It could be figured out that there were 7 differences between the herbicides were significant ($p < 2e-16$). The number of significant differences in two interaction items were 16 and 7, the p value was 0.00532 and less than 2e-16 respectively. The difference between Mow and No Mow was significant as well ($p = 0.005$).

Table 2 Summary of Variance Analysis Model

	Df	Sum sq	Mean sq	p value
Herbicide	7	1254.2	179.17	< 2e-16
Mow	1	8.1	8.13	0.00344
Farm:Paddock	16	546.6	34.16	< 2e-16
Herbicide:Mow	7	16.1	2.3	0.01798
Residuals	784	740.6	0.94	

In the obtainied model, the level of HerbicideNil which had a coefficient of 3.39 could make the responding variable increase most. The level could make the responding variable decline most was Farm B.Dry which had a coefficient of -1.73q1.

Table 3 Coefficient of Variance Analysis Model

	Estimate		Estimate
(Intercept)	-1.23296511	inter_termFarm B.Dry	-1.7372
HerbicideAminopyralid+triclopyr	-0.41432581	inter_termFarm C.Dry	0.37951
HerbicideFlumetsulam	0.97800432	inter_termFarm D.Dry	-0.5769
HerbicideMCPA	1.35346575	inter_termFarm E.Dry	-0.1553
HerbicideMCPB	2.40946919	inter_termFarm F.Dry	-1.4787
HerbicideMCPB+bentazone	1.94594251	inter_termFarm G.Dry	0.53164
HerbicideNil	3.3915905	inter_termFarm H.Dry	-0.0934
HerbicideThifensulfuron-methyl	1.88631911	inter_termFarm I.Dry	-0.8951
MowNo Mow	0.20817045	inter_termFarm A.Wet	-0.0862
HerbicideAminopyralid+triclopyr:MowNo Mow	-0.02504201	inter_termFarm B.Wet	1.0621
HerbicideFlumetsulam:MowNo Mow	-0.16168083	inter_termFarm D.Wet	-0.8186
HerbicideMCPA:MowNo Mow	-0.1129004	inter_termFarm E.Wet	0.37337
HerbicideMCPB:MowNo Mow	0.24232849	inter_termFarm F.Wet	-0.0179
HerbicideMCPB+bentazone:MowNo Mow	-0.12553836	inter_termFarm G.Wet	0.32134
HerbicideNil:MowNo Mow	0.56296109	inter_termFarm H.Wet	-1.7655
HerbicideThifensulfuron-methyl:MowNo Mow	-0.44821798	inter_termFarm I.Wet	0.69969

The Q-Q plot showed that the residuals were close to normal. Therefore, the assumption of normally distributed random effects was reasonable.

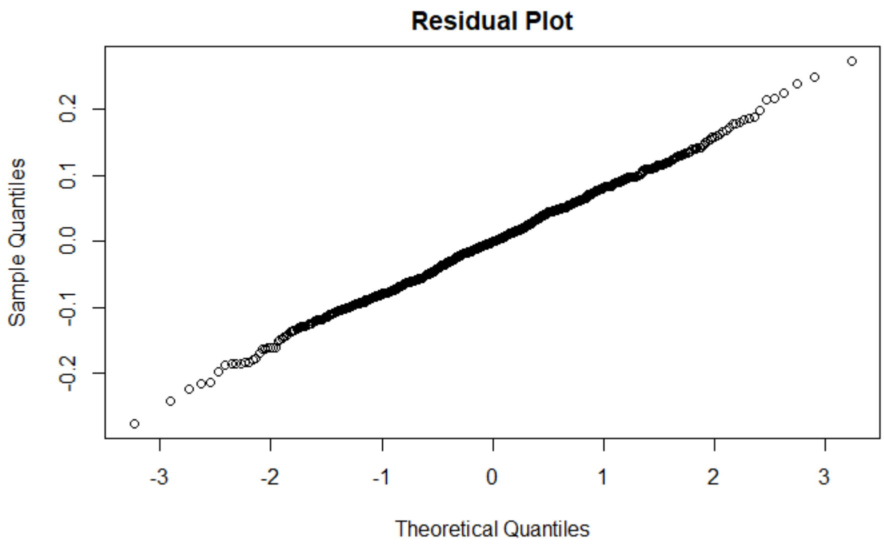


Figure 7 Q-Q plot for the Variance Analysis Model

Linear Mixed Model

The result of summary () function was shown in Table 4 which told that the residual variance was 0.94, which was the same as that of the variance analysis model. What is more, the fixed effect displayed the same coefficients as Table 3.

Table 4 Random Effect of Linear Mixed Model

Groups	Name	Variance	Std.Dev
inter_term	Intercept	0.692	0.8319
Residual		0.9447	0.9719

The result of ranef() offered the coefficients of levels in random item, which had some difference with that in Table 3 and led to some difference in intercept as well.

Table 5 Coefficients of Levels in Random Effect

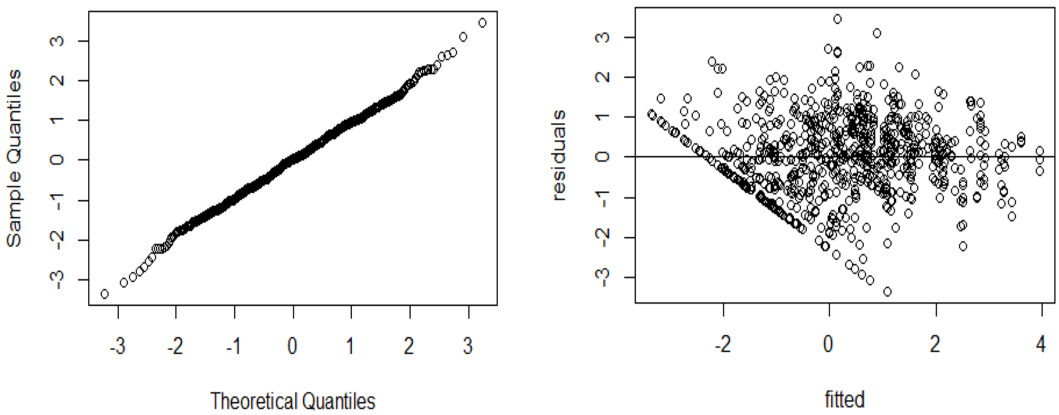
Estimate	Estimate
(Intercept) -1.48338	
Farm A.Dry 0.24349159	Farm I.Dry-0.62680949
Farm B.Dry-1.44567812	Farm A.Wet 0.15964227
Farm C.Dry 0.61250807	Farm B.Wet 1.27622279
Farm D.Dry-0.31748668	Farm D.Wet-0.55250729
Farm E.Dry 0.09253401	Farm E.Wet 0.60653361
Farm F.Dry-1.19427050	Farm F.Wet 0.22604425
Farm G.Dry 0.76043277	Farm G.Wet 0.55594533
Farm H.Dry 0.15271352	Farm H.Wet-1.47314901

The AIC value for the four models in Table 6 had evidence the interaction between Mow and Herbicide could be removed because both the complexity and AIC value would increase when it was added into the model. At the same time, the variable Mow had a relatively less significance as well since the AIC just went up slightly when it was removed. However, keeping Mow in the model could improve the interpretation of the model.

Table 6 AIC Value of Models

Model	AIC
`Buttercuppc ~ Mow*Herbicide + (1|inter_term)`	2381.092
`Buttercuppc ~ Mow+Herbicide + (1|inter_term)`	2375.943
`Buttercuppc ~ Mow + (1|inter_term)`	3124.081
`Buttercuppc ~ Herbicide + (1|inter_.erm)`	2378.884

Figure 8 (a) performed an approximately straight line which told that the residual was well normalised, that could be also checked in (b).



(a) **(b)**

Figure 8 Q-Q plot and Residualfor the Variance Analysis Model

Discussion

Part 2 Model for predicting BareGrnd

Method

1. Data Processing

2. Model Fitting and Diagnostics

Generalised Linear Model

When a series of continuous and categorical predictors are used to predict a counting variable, Poisson distribution should be conducted. In Buttercup dataset, the responding variable ' BaregGrnd' recorded the number of 10cm by 10cm squares of bare ground in each plot (10 metres by 2 metres). Therefore, glm () function with parameter 'family' being set into 'Poisson' was implemented to fit the generalised linear model for the data set.

In order to check the summary of the model, summary () and the variance analysis function anova() were both conducted.

Since it was necessary to diagnose the model assumption after the model was got, so QQ plot which was drew using qqnorm() and residual distribution were displayed.

Generalised Linear Mixed Model

In this study, herbicides were randomly used in the various Paddocks, and Paddock belonged to the nine farms, so the farm and its covariate Paddock could be randomly assigned to the model fitting process to get the generalized linear mixed model. The generalized linear hybrid model could separate the random parts from the generalized linear model, set different levels according to data, and decompose the residuals accordingly, thus it could lead the reduction of the proportion of the non-explanatory part of the total variation, and fit the data better.

QQ plot (conducted using qqnorm ()) and the plot of residuals and fitting were used for regression diagnosis of the generalized linear mixed model. If the regression assumption was reasonable, then in the QQ plot, the residuals should fell on a straight line of 45 degrees. In residual vs fitted plot, there should be no correlation between residual values and fitted values.

Result

1. Data Processing

Figure 1 showed that there were differences in the distribution for each level of BaregGrnd, all of them had a centralised distribution when BaregGrnd was less than 100, but some levels had more dispersive data, such as Aminopyralid+triclopyr.

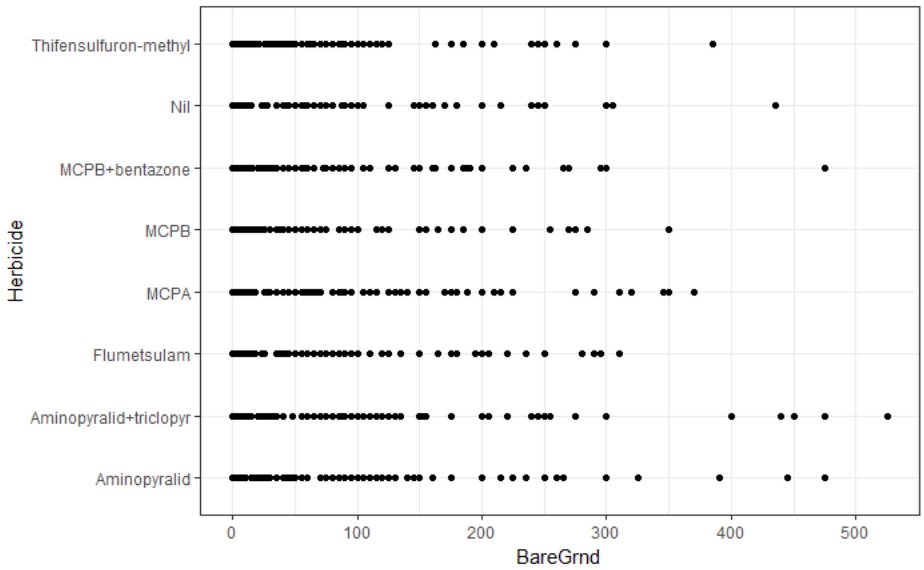


Figure 1 Distribution of BaregGrnd for each level of Herbicide

The density curve displayed in Figure 2 clearly showed the difference in distribution of BaregGrnd at different levels of Mow. Although the two distribution curves had similar range, when the plot was not Mown, the distribution was more concentrated and the mean was smaller.

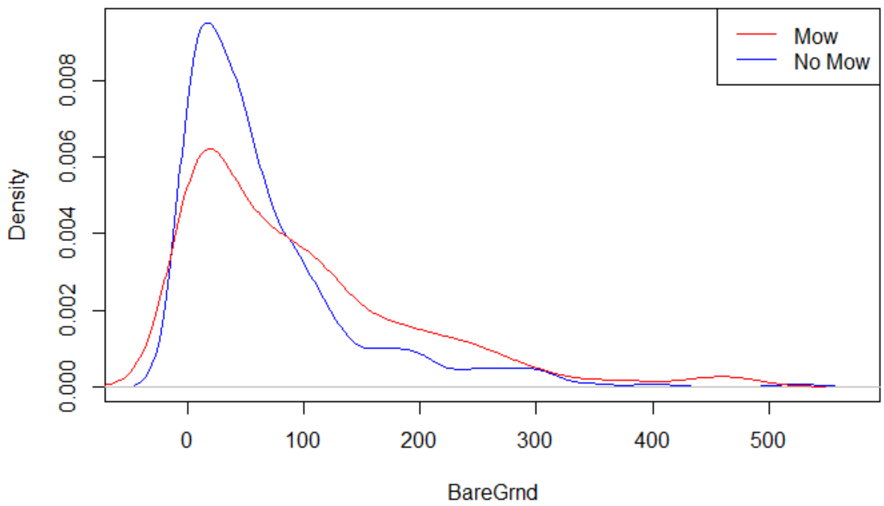


Figure 2 Distribution of BaregGrnd for Mow

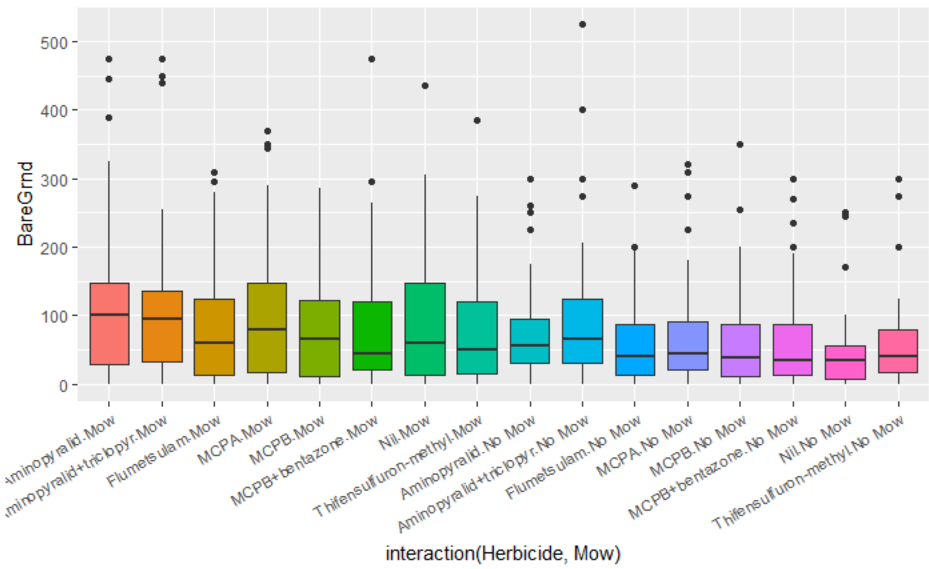


Figure 3 Distribution of BaregGrnd for each level of Interaction (Herbicide, Mow)

The boxplot which was shown as Figure 4 had evidence that the data distributed quite differently in each level of the interaction between Farm and Paddock. For example, the level Farm C. Dry had distribution with both high variability and great median. On the contrary, the data for Farm G was close to 0.

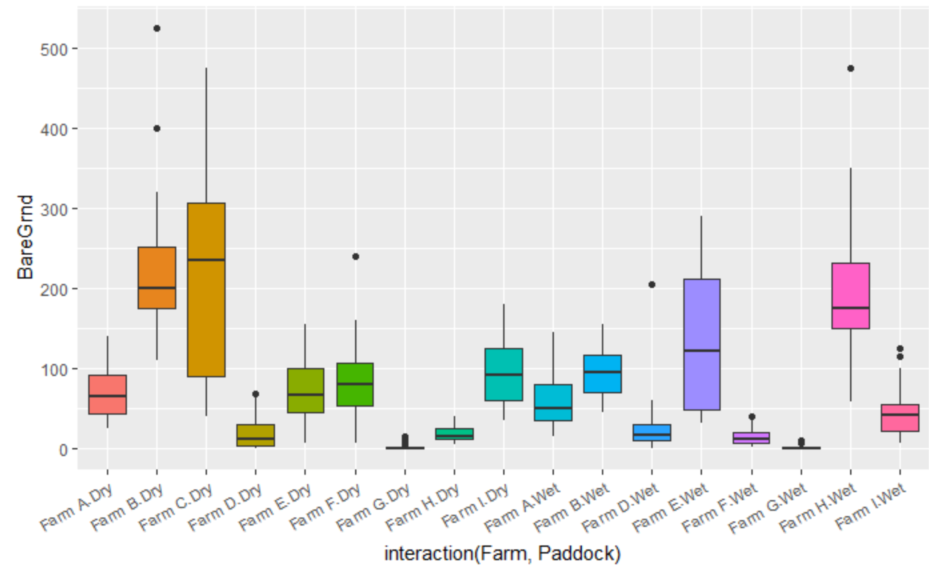


Figure 4 Distribution of BaregGrnd for each level of interaction (Farm, Paddock)

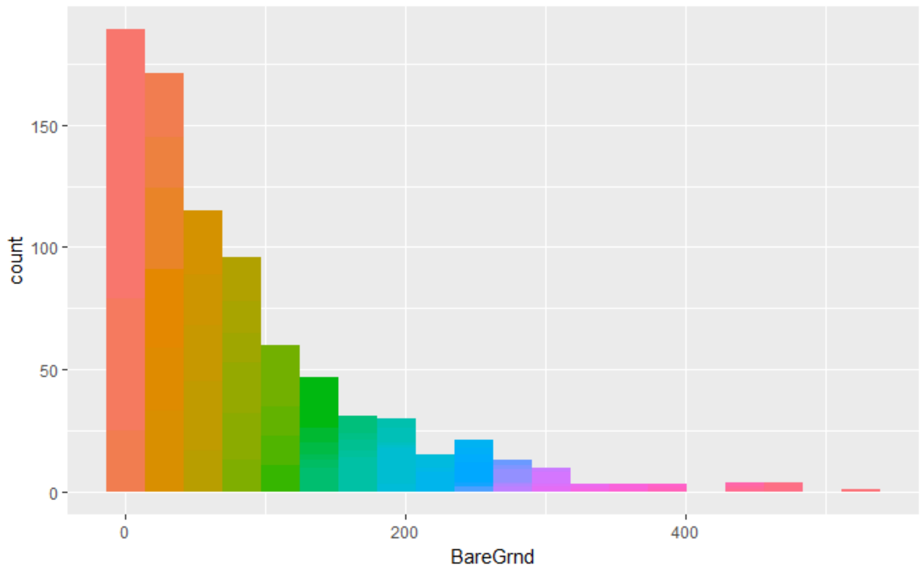


Figure 5 Distribution of BaregGrnd

2. Model Fitting

Generalised Linear Model

The result of anova() (shown in Table 2) showed that all the variable were significant in the model. However, from Table 1 which told the summary of the model, it could be known that each level contributes differently in the change of the corresponding variable. When the level was HerbicideMCPB+bentazone together with No Mow, BareFrnd increased most while it would decline most when the level was inter_term2Farm G and Dry.

Table 1 Coefficients of Generalised Linear Model

	Estimate Std. Error z value Pr(>|z|)
(Intercept)	4.63550 0.02108 219.941 < 2e-16 ***
HerbicideAminopyralid+triclopyr	0.02406 0.01816 1.325 0.18508
HerbicideFlumetsulam	-0.28219 0.01970 -14.326 < 2e-16 ***
HerbicideMCPA	-0.15290 0.01900 -8.046 8.59e-16 ***
HerbicideMCPB	-0.36133 0.02015 -17.928 < 2e-16 ***
HerbicideMCPB+bentazone	-0.31731 0.01990 -15.947 < 2e-16 ***
HerbicideNil	-0.30095 0.01980 -15.197 < 2e-16 ***
HerbicideThifensulfuron-methyl	-0.30728 0.01984 -15.488 < 2e-16 ***
MowNo Mow	-0.43329 0.02059 -21.041 < 2e-16 ***
inter_term2Farm B.Dry	1.13047 0.01967 57.471 < 2e-16 ***
inter_term2Farm C.Dry	1.14565 0.01963 58.350 < 2e-16 ***
inter_term2Farm D.Dry	-1.21841 0.03580 -34.034 < 2e-16 ***
inter_term2Farm E.Dry	0.02712 0.02402 1.129 0.25891
inter_term2Farm F.Dry	0.14487 0.02336 6.203 5.55e-10 ***
inter_term2Farm G.Dry	-3.76766 0.11380 -33.107 < 2e-16 ***

inter_term2Farm H.Dry	-1.31491	0.03717	-35.373	< 2e-16	***
inter_term2Farm I.Dry	0.26916	0.02271	11.850	< 2e-16	***
inter_term2Farm A.Wet	-0.21050	0.02556	-8.234	< 2e-16	***
inter_term2Farm B.Wet	0.28358	0.02264	12.523	< 2e-16	***
inter_term2Farm D.Wet	-1.04619	0.03354	-31.190	< 2e-16	***
inter_term2Farm E.Wet	0.62726	0.02118	29.613	< 2e-16	***
inter_term2Farm F.Wet	-1.56342	0.04110	-38.040	< 2e-16	***
inter_term2Farm G.Wet	-3.65977	0.10796	-33.898	< 2e-16	***
inter_term2Farm H.Wet	1.03075	0.01992	51.743	< 2e-16	***
inter_term2Farm I.Wet	-0.40781	0.02706	-15.071	< 2e-16	***
HerbicideAminopyralid+triclopyr:MowNo Mow	0.19439	0.02817	6.900	5.20e-12	***
HerbicideFlumetsulam:MowNo Mow	0.01949	0.03130	0.623	0.53344	
HerbicideMCPA:MowNo Mow	0.05059	0.03006	1.683	0.09232	.
HerbicideMCPB:MowNo Mow	0.12472	0.03146	3.965	7.34e-05	***
HerbicideMCPB+bentazone:MowNo Mow	0.16294	0.03088	5.277	1.31e-07	***
HerbicideNil:MowNo Mow	-0.24354	0.03306	-7.367	1.75e-13	***
HerbicideThifensulfuron-methyl:MowNo Mow	0.11809	0.03101	3.808	0.00014	***

Table 2 Variance Analysis of Generalised Linear Model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			815	68695	
Herbicide	7	1735	808	66960	< 2.2e-16 ***
Mow	1	2221	807	64740	< 2.2e-16 ***
inter_term2	16	51380	791	13360	< 2.2e-16 ***
Herbicide:Mow	7	235	784	13124	< 2.2e-16 ***

The distribution of residual (displayed in Figure 6) performed normally which justified that it was correct to conduct 'Poisson'.

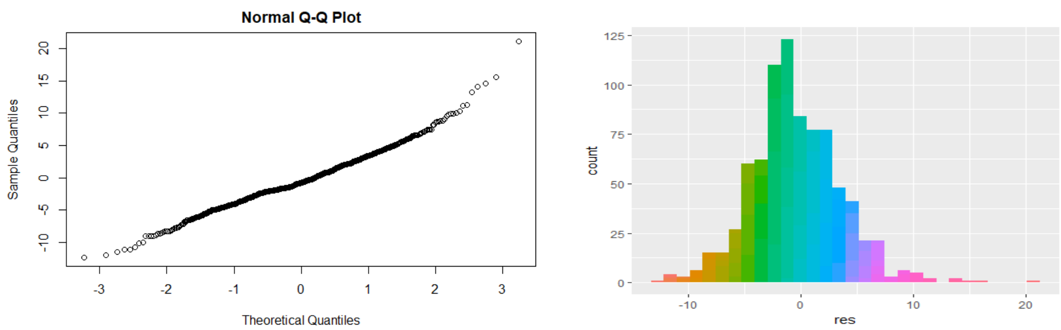


Figure 6 Distribution of Residual

Generalised Linear Mixed Model

The coefficients of variables in fixed effect were the same as those of generalised linear model. however, the coefficients of random effect which was shown in in Table 3 were slightly different.

Table 3 Coefficients of the Random Effect in Generalised Linear Mixed Model

	(Intercept)		(Intercept)
Farm A.Dry	0.50028802	Farm A.Wet	0.28980787
Farm B.Dry	1.63075184	Farm B.Wet	0.78385389
Farm C.Dry	1.64593533	Farm D.Wet	-0.54560835
Farm D.Dry	-0.71769650	Farm E.Wet	1.12753561
Farm E.Dry	0.52740840	Farm F.Wet	-1.06232884
Farm F.Dry	0.64515332	Farm G.Wet	-3.14189233
Farm G.Dry	-3.24715892	Farm H.Wet	1.53102831
Farm H.Dry	-0.81410681	Farm I.Wet	0.09253258
Farm I.Dry	0.76943359		

Table 4 Comparison of Generalised Linear Model and Generalised Linear Mixed Model

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
glmm1	17	17643	17723	-8804.7	17609			
glm1	32	17515	17665	-8725.3	17451	158.82	15	< 2.2e-16 ***