STA610 Lab 6 Team 4 Report

Cole Juracek Lauren Palazzo Lingyu Zhou Fan Zhu

2021-03-31

Introduction

Our data is drawn from the results of observational surveys of badger activities on 36 farms over three years. The aim of our analysis is to analyze factors associated with badger activity on farms, correlation over time of badger activity, and the level of variability in badger activity on the farms.

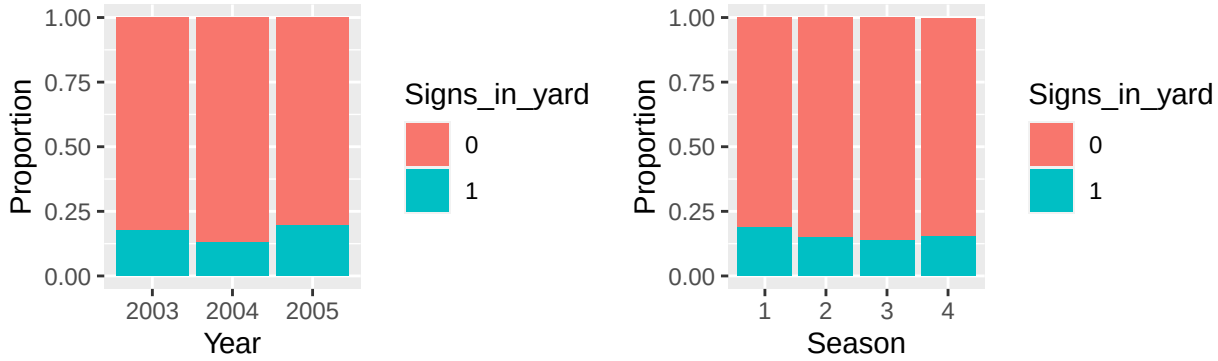
EDA

Response Variable: Signs_in_yard

Table 1: Frequency Table for Signs in Yard

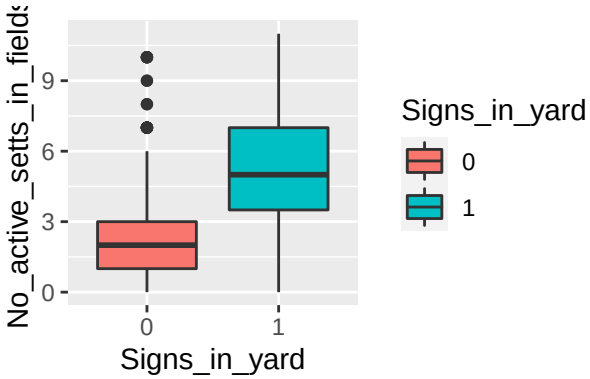
Signs in Yard	Freq
0	230
1	43

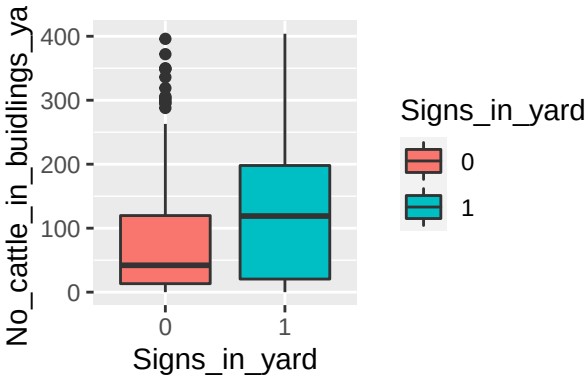
Year & Season



From the bar plot of Signs_in_yard by Year we can observe that the presence of badger activity in the farmyard in 2004 is lower than those in 2003 and 2005. From the bar plot of Signs_in_yard by Season we see that presence of badger activity is the highest in season 1 and the lowest in season 3. Even though badgers do not hibernate, they reduce activities during cold weathers so we will use a main effect of season in our model.

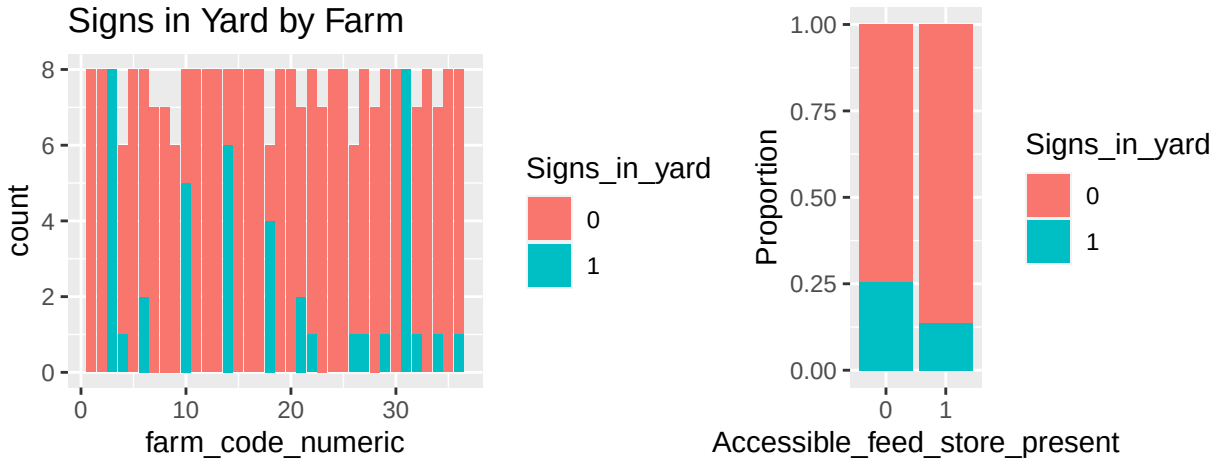
No_active_setts_in_fields & No_cattle_in_buidlings_yard





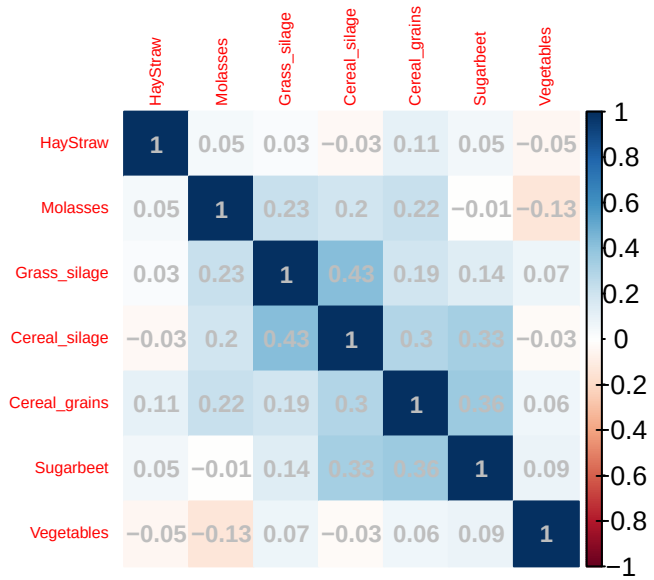
The median of No_active_setts_in_fields is much higher when there is badger activity in the farmyard than when there is no badger activity. The median of No_cattle_in_buidlings_yard is also much higher when there is badger activity in the farmyard than when there is no badger activity.

Farm



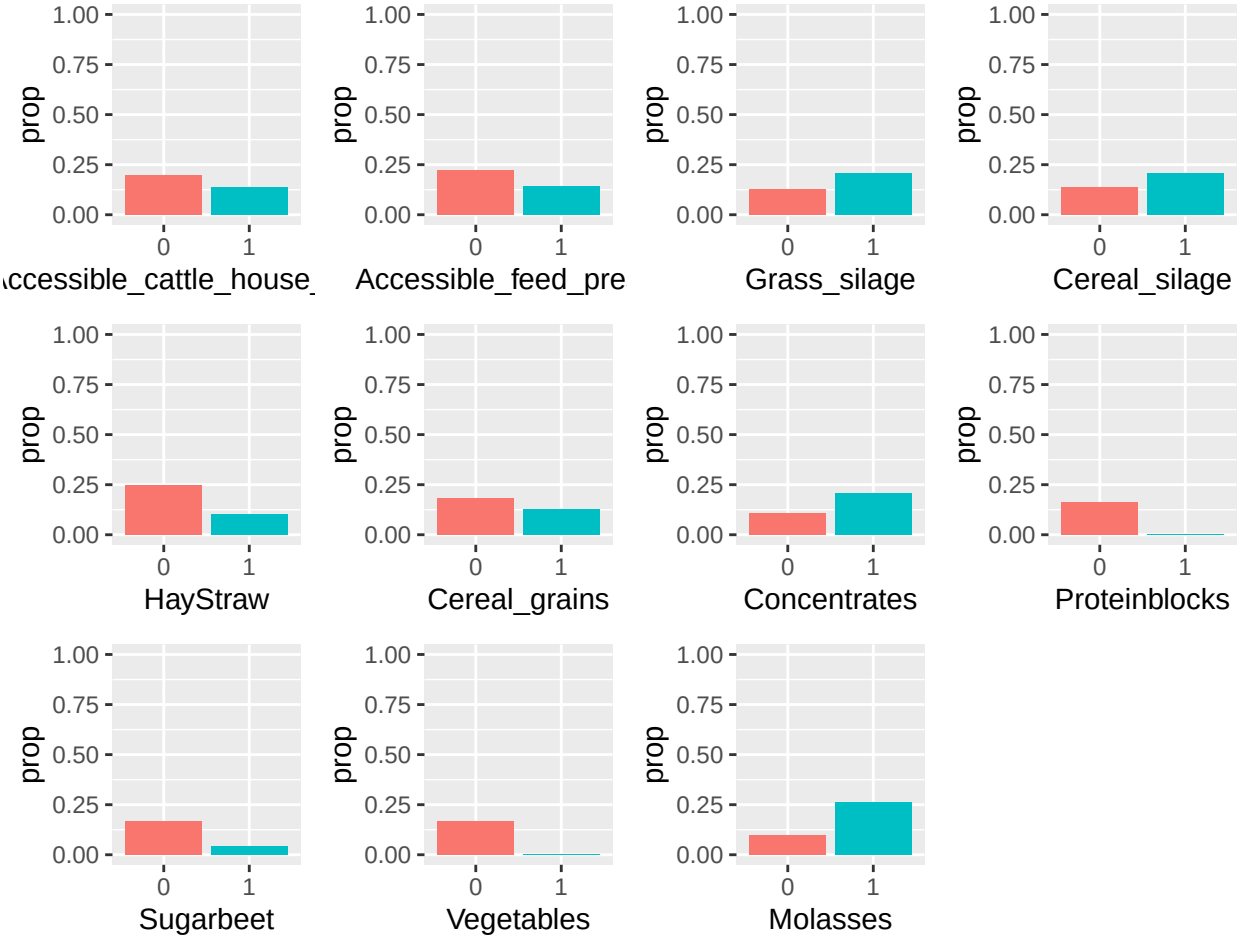
The patterns in the stacked bar chart of Signs_in_yard by farm indicate the incorporation of the random intercept by farm into our model. From the segmented bar chart for Accessible_feed_store_present we know that there are more presences of badger activities in the farmyard when there is Accessible_feed_store_present than when there is not.

Binary Variables



Cole EDA (remaining variables)

First, we want to check the relative proportion of badger activity for each binary variable:



Nearly all of the proportions look like they *could* be similar - especially due to the small sample size. We can formally test whether the proportions are different with a Chi-square test:

	p.val
Accessible_cattle_house_present	0.26
Accessible_feed_present	0.28
Grass_silage	0.10
Cereal_silage	0.24
HayStraw	0.00
Cereal_grains	0.29
Concentrates	0.03
Molasses	0.00

At a significance level of $\alpha = 0.05$, we reject the null hypotheses for the following

- HayStraw
- Concentrates
- Molasses

And conclude their proportions are not equal.

For the remaining variables, the assumptions of a Chi-sq test were violated (Proteinblocks, Sugarbeet, and Vegetables). At least one of the table values has an expected counts < 5. We can test these with a Fisher exact test for small sample sizes:

	p.val
Proteinblocks	0.223
Sugarbeet	0.143
Vegetables	0.084

The results of the Fisher test suggest we should not reject the null at the standard significance level. We do not conclude a meaningful association between these variables and the response.

Model

Model Selection From the correlation plot we can observe that there are some significant correlations between the binary variables. Therefore, we need to exclude some of them to prevent the collinearity issue. For instance, Cereal_silage and HayStraw cannot exist in the same model because of their significant correlation coefficient. We did not use random intercept for season because there are only four seasons. Interaction terms are removed because they would result in failures to converge and singularity issues. Based on these principles, we built multiple viable models by adding combinations of binary variables on a basic model $Signsinyard \sim Noactivesettsinfields + Nocattleinbuidlingsyard + (1|farmcodenumeric) + Season$.

Table 4: Model Comparisons with BIC

Model	Binary.Variables.Added.to.the.Basic.Model	BIC
Model4	HayStraw	193.18
Model5	Sugarbeet	193.59
Model6	Molasses	194.51
Model7	Grass silage	194.82

Model	Binary.Variables.Added.to.the.Basic.Model	BIC
Model8	Cereal silage	194.82
Model9	Cereal grains	194.83
Model10	Cereal silage+Cereal grains	200.43

From the model comparisons above we chose model 4 as our final model following the principle of parsimony.

Model Specification

$$y_{ij}|x_{ij} \sim Bernoulli(\pi_{ij})$$

* y_{ij} stands for response variable Signs_in_yard. * i stands for individuals, i = 1,. . . ,n. j stands for each farm, j = 1,. . . ,36.

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + b_{0j} + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4$$

$$b_{0j} \sim N(0, \sigma^2)$$

- β_{0j} stands for random intercept by farm.
- x_1 stands for number of active badger homes in nearby fields. (No_active_setts_in_fields).
- x_2 stands for number of cattle on the farm (No_cattle_in_buildings_yard).
- x_3 stands for Season.
- x_4 stands for HayStraw.

Results

Fixed Effects

Table 5: Factors Related to Badger Activity

	Estimate	Std.Error	CI - Low	CI - High
(Intercept)	-4.60	1.15	-6.86	-2.35
No_active_setts_in_fields	0.49	0.15	0.20	0.78
No_cattle_in_buidlings_yard	0.01	0.00	0.00	0.01
Season2	-0.23	0.74	-1.69	1.23
Season3	0.02	0.80	-1.55	1.59
Season4	-0.37	0.69	-1.72	0.99
HayStraw1	-0.78	0.60	-1.96	0.40

- *Intercept*: For a fixed farm (or across all farms), an observation, which has no number of active badger homes in nearby fields (no_active_setts_in_fields), no number of cattle on the farm (no_cattle_in_buildings_yard), Season being 1 and no HayStraw, has the odds of exp(-4.6049271) = 0.01 of badger activity being present in the farmyard.
- *Number of active badger homes in nearby fields (No_active_setts_in_fields)*: Controlling for other variables, for every one unit increase of number of active badger homes in nearby fields, the odds of badger activity being present in the farmyard increase by a multiplicative effect of exp(0.4878348) = 1.628786.

- *Number of cattle on the farm (No_cattle_in_buidlings_yard)*: Controlling for other variables, for every one unit increase of number of cattle on the farm, the odds of badger activity being present in the farmyard increase by a multiplicative effect of exp(0.0051664) = 1.00518.
- *Season* ("Season1" is the reference, other levels omitted for brevity):
  - Controlling for other variables, an observation in Season2 has exp(-0.2294778) = 0.7949486 times the odds of badger activity being present in the farmyard as an observation in Season1.
- *HayStraw* ("HayStraw0" is the reference):
  - Controlling for other variables, an observation with HayStraw1 has exp(-0.7805303) = 0.458163 times the odds of badger activity being present in the farmyard as an observation with HayStraw0.
- 95% confidence interval endpoints are also on the logit scale. For example, we can say that we are 95% confident that the odds estimate for HayStraw above is between exp(-1.9589250) = 0.14 and exp(0.3978645) = 1.49.

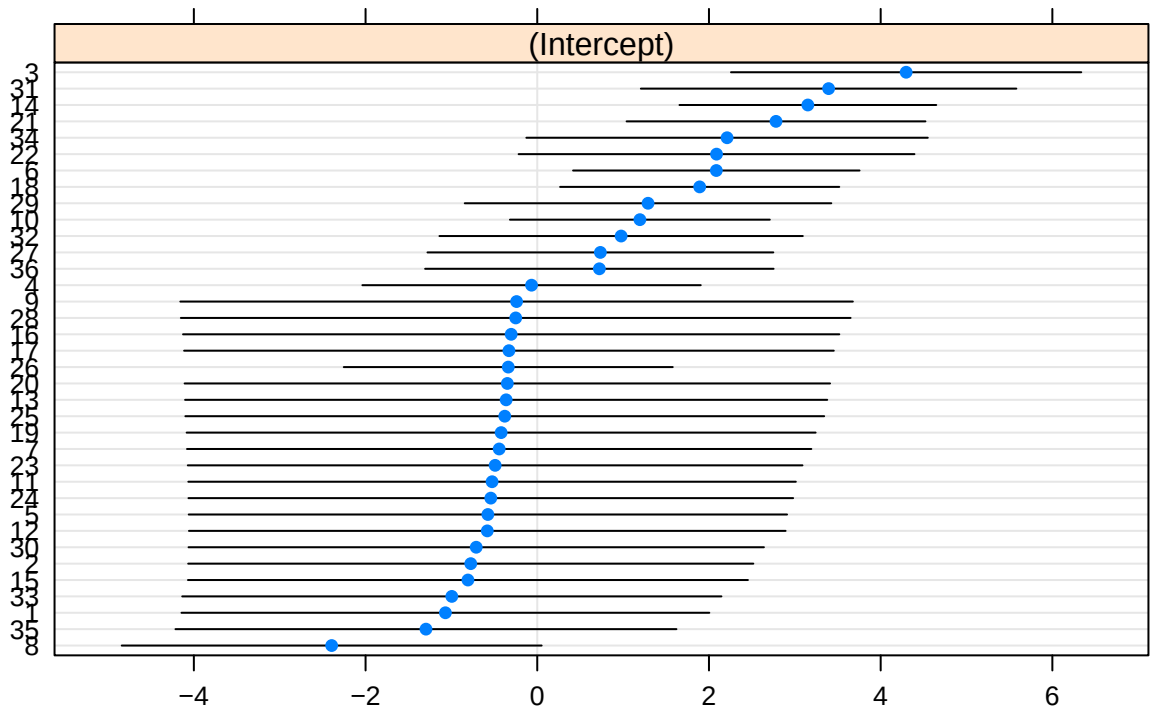
Random Effects

Table 6: Farm Variation

Group	Variance	Std.Dev.
Farm (Intercept)	4.94648457248147	2.22406937222773

The variance of the farm-level intercepts (i.e., the baseline log-odds of badger activity among the farms) is about 4.94, while the grand intercept is centered at about -4.60. In other words, the variance is about the same as the mean, which suggests a considerable amount of between-farm heterogeneity.

The adjusted ICC is 0.601, which indicates high similarity between values from the same farm. I.E., each farm is fairly homogeneous with respect to its observations of badger activity.



The odds of the random intercepts range from exp(4.29665538) = 73.45371 for farm with code 3 to exp(-2.39464632) = 0.09120493 for farm with code 8. We interpret this as follows: the estimated odds of badger activity being present in the farmyard is highest for farm with code 3 with zero active badger homes in nearby fields (no_active_setts_in_fields), zero cattle on the farm (no_cattle_in_buildings_yard), Season being 1 and no HayStraw. One interesting observation from the above graph is that farms with below-average intercepts appear to be generally somewhat more homogeneous (similar in value) than those with above-average intercepts.

With regard to correlation over time, the correlations of fixed effects in the model summary suggest how the coefficients of 'Season' tend to correlate with each other. For example, Season 1's respective correlations with Seasons 2, 3, and 4 are -0.444, -0.463, and -0.232. So, higher badger activity in Season 1 is expected to be associated with lower badger activity in the other three seasons.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, cache=TRUE, fig.height = 3)
df <- read.delim("BadgersFarmSurveysNoNA.txt", header = TRUE)
library(pacman)
pacman::p_load(tidyverse, lme4, gridExtra, grid, ggplot2, lattice, redres, stringr, influence.ME,
# devtools::install_github("goodekat/redres")
df$Signs_in_yard = factor(df$Signs_in_yard)
knitr::kable(table(df$Signs_in_yard),
             col.names = c("Signs in Yard","Freq"),
             caption = "Frequency Table for Signs in Yard")
df$Year = factor(df$Year)
df$Season = factor(df$Season)

bar_year <- ggplot(df,
       aes(x = Year,
           group = Signs_in_yard,
           fill = Signs_in_yard)) +
  geom_bar(position = "dodge")

bar_season <- ggplot(df,
       aes(x = Season,
           group = Signs_in_yard,
           fill = Signs_in_yard)) +
  geom_bar(position = "dodge")

grid.arrange(bar_year, bar_season, ncol=2)
bar_year <- ggplot(df,
       aes(x = Year,
           fill = Signs_in_yard)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")


bar_season <- ggplot(df,
       aes(x = Season,
           fill = Signs_in_yard)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")

grid.arrange(bar_year, bar_season, ncol=2)
hist_No_active_setts_in_fields <- ggplot(df, aes(x=No_active_setts_in_fields)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 20)+
 geom_density(alpha=.2, fill="red")

hist_No_cattle_in_buidlings_yard <- ggplot(df, aes(x=No_cattle_in_buidlings_yard)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 20)+
 geom_density(alpha=.2, fill="blue")
```

```r
grid.arrange(hist_No_active_setts_in_fields,hist_No_cattle_in_buidlings_yard, ncol=2)

# Both the distributions of No_active_setts_in_fields and No_cattle_in_buidlings_yard are extreme
box_No_active_setts_in_fields <- ggplot(data=df,
       mapping = aes(x = Signs_in_yard,
                     y = No_active_setts_in_fields,
                     fill = Signs_in_yard)) +
  geom_boxplot()

box_No_cattle_in_buidlings_yard <- ggplot(data=df,
       mapping = aes(x = Signs_in_yard,
                     y = No_cattle_in_buidlings_yard,
                     fill = Signs_in_yard)) +
  geom_boxplot()

grid.arrange(box_No_active_setts_in_fields,
             box_No_cattle_in_buidlings_yard,
             ncol=2)
df$Accessible_feed_store_present = factor(df$Accessible_feed_store_present)

bar_farm <- ggplot(df,
       aes(x = farm_code_numeric,
           fill = Signs_in_yard)) +
  geom_bar(position = "stack") +
  labs(title = "Signs in Yard by Farm")

bar_Accessible_feed_store_present <- ggplot(df,
       aes(x = Accessible_feed_store_present,
           fill = Signs_in_yard)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")

grid.arrange(bar_farm, bar_Accessible_feed_store_present,
             ncol=2,
             widths=c(3, 2))
corrplot(cor(df[,c(13,14,15,16,19,20,21)]),
         method = "color",
         addCoef.col="grey",
         order = "AOE",
         number.cex = 0.75,
         tl.cex = 0.5)
correlation::correlation(df[,c(10,13,14,15,16,19,20,21)])
table(df$Accessible_feed_store_present)
# For visualization, it helps to make these variables factors
df[, 11:21] <- apply(df[, 11:21], 2, factor)
plot_binary_data <- function(data, col) {
    var_df <- data %>% group_by_at(col) %>% summarise(prop = sum(Signs_in_yard == 1) / n())
    ggplot(var_df, aes_string(x=col, y='prop', fill=col)) + geom_col() + ylim(0, 1) + theme(legen
}
```

```
binary_plots <- lapply(names(df)[11:21], plot_binary_data, data=df)
grid.arrange(grobs=binary_plots, ncol = 4)
test_proportion <- function(col) {
    results <- chisq.test(table(df$Signs_in_yard, df[, col]))
    return(results$p.value)
}

good_ss_names <- c(11:17, 21)
results <- sapply(names(df)[good_ss_names], test_proportion)
kable(data.frame('p-val' = results), digits = 2)
results <- sapply(c('Proteinblocks', 'Sugarbeet', 'Vegetables'), function(col) {
    results <- fisher.test(table(df$Signs_in_yard, df[, col]))
    return(results$p.value)
})
kable(data.frame('p-val' = results), digits = 3)
mdl2 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl3 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl4 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl5 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl6 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

#anova(mdl4,mdl5,mdl6)


mdl7 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl8 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)

mdl9 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_c
              family = binomial(link="logit"),
              data = df)
```

```
mdl10 <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|farm_
               family = binomial(link="logit"),
               data = df)

#anova(mdl4, mdl7,mdl8,mdl9,mdl10)
mdl_final <- glmer(Signs_in_yard ~ No_active_setts_in_fields + No_cattle_in_buidlings_yard + (1|f
               family = binomial(link="logit"),
               data = df)
BIC_res <- sapply(c(mdl4, mdl5, mdl6, mdl7, mdl8, mdl9, mdl10), BIC)
re_results <- data.frame('Model' = c('Model4',
                                     'Model5',
                                     'Model6',
                                     'Model7',
                                     'Model8',
                                     'Model9',
                                     'Model10'),
                         'Binary Variables Added to the Basic Model' = c('HayStraw',
                                     'Sugarbeet',
                                     'Molasses',
                                     'Grass silage',
                                     'Cereal silage',
                                     'Cereal grains',
                                     'Cereal silage+Cereal grains'),
                         'BIC' = BIC_res)
knitr::kable(re_results, caption = 'Model Comparisons with BIC', digits=2)
coefs <- data.frame(coef(summary(mdl_final)))
# use normal distribution to approximate p-value
coefs$Pr...z.. <- 2 * (1 - pnorm(abs(coefs$z.value)))
coefs$`Low CI Endpt` <- coefs$Estimate - 1.96*coefs$Std..Error
coefs$`High CI Endpt` <- coefs$Estimate + 1.96*coefs$Std..Error
knitr::kable(coefs[, c(1, 2, 5, 6)], col.names = c("Estimate", "Std.Error", "CI - Low", "CI - Hig
re_dat = as.data.frame(VarCorr(mdl_final))
re_names = rbind("Farm (Intercept)")
re_results <- cbind(re_names, re_dat$vcov, re_dat$sdcor)
knitr::kable(re_results, col.names = c("Group", "Variance", "Std.Dev."), digits = 2, caption = "F
library(sjstats)
icc(mdl_final)
#Adjusted ICC: 0.601
#Conditional ICC: 0.484
dotplot(ranef(mdl_final, condVar=TRUE))$farm_code_numeric
summary(mdl_final)
```