Lab 9

Emily Gentles Weiyi Liu Jack McCarthy Qinzhe Wang

10/28/2021

Introduction

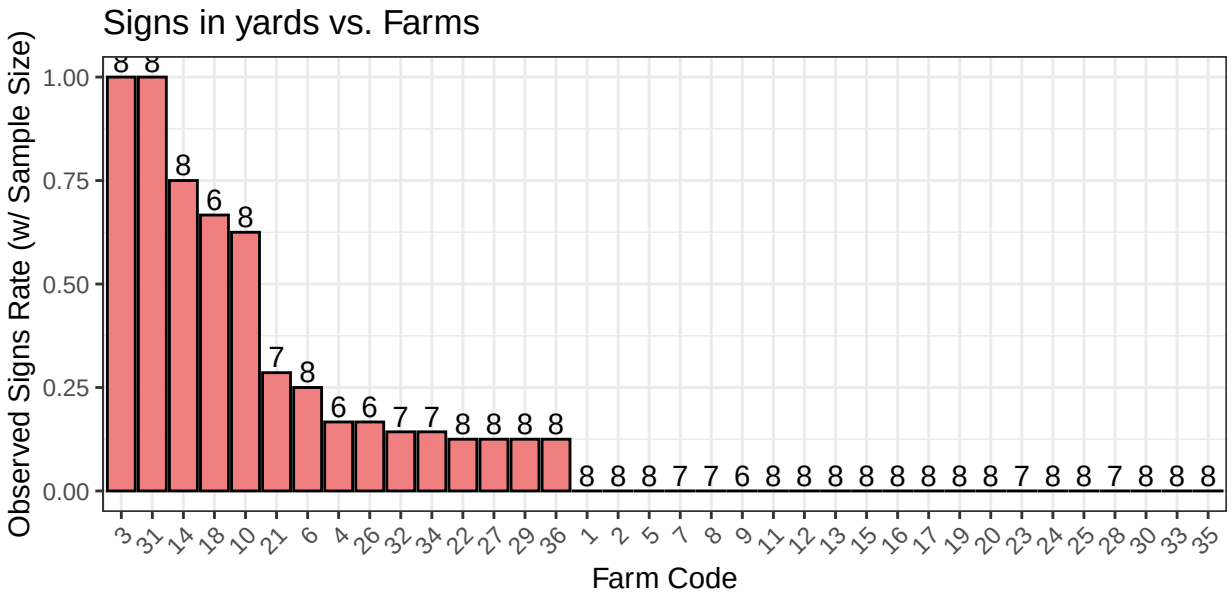
Management of badgers on Welsh and British farms is a disputed topic. Some people agree that culling badger populations are necessary to control the spread of bovine tuberculosis while others argue that badgers are not the primary cause of the spread of disease and that culling is inhumane.

This lab aims to examine the factors related to the presence of badger activity in the farmyard and the correlation over time in badger activity and farm-specific heterogeneity in the tendency to have badger activity. Each farm was observed up to eight times (once per season for a two-year period).

EDA

Reponse: Signs_in_yard

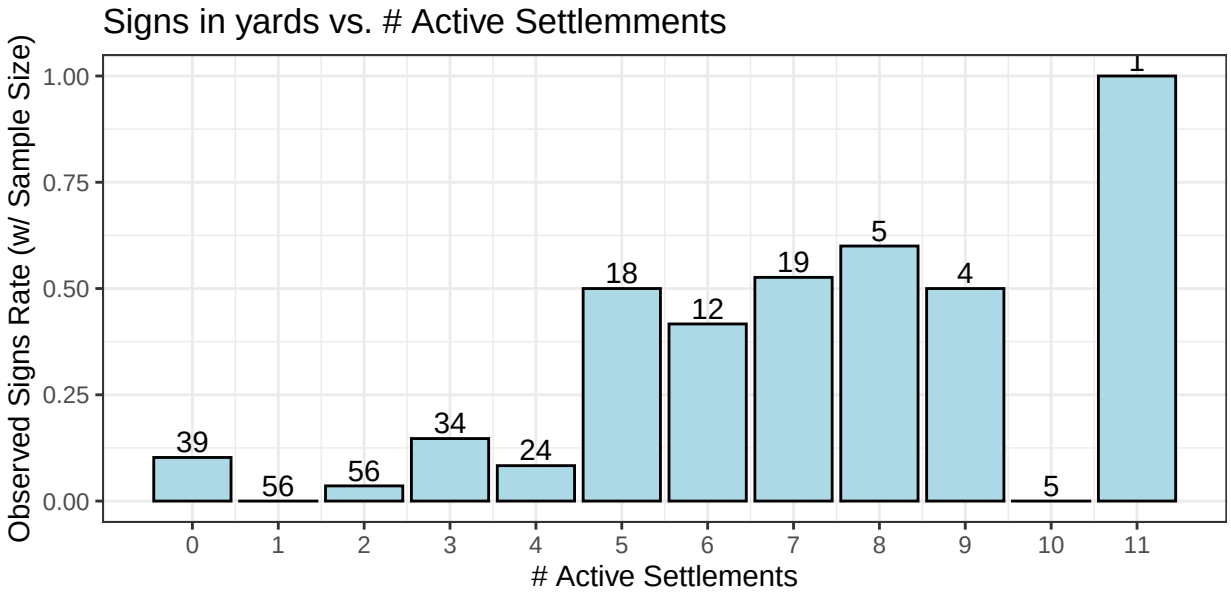
First, we look at the distribution of the response. We basically inspect the mean `signs_in_yard` value for each farm to get a "rate" of badger activity being observed. It seems like the signs rates vary widely across farms, as displayed in Figure 1. Therefore, we should expect to see heterogeneity in the presence of badger activity across farms.



Now we can look at the signs rates for some of the candidate predictors.

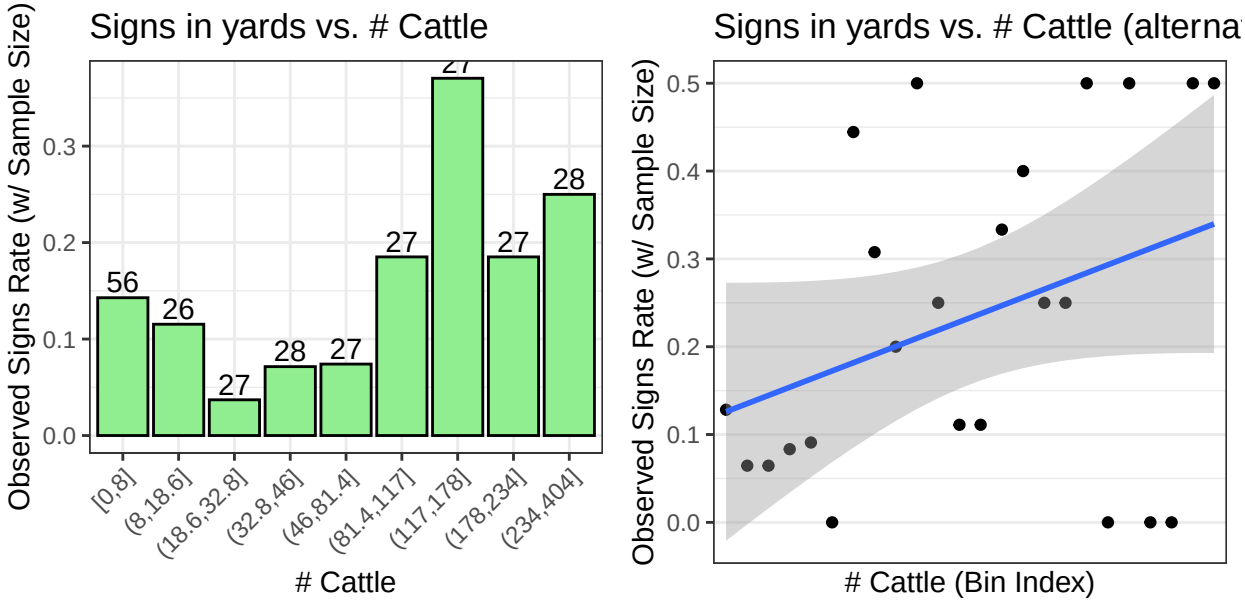
Number of active badger homes in nearby fields (no_active_setts_in_fields)

First, we look at the presence of badger activity against each value of the number of active badger homes in nearby fields. From the figure below, there seems to be a positive relationship between the number of active badger homes and the rate of badger signs in the yard.



Number of cattle on the farm (no_cattle_in_buildings_yard)

We may also look at the relationship between the number of cattle on the farm and the signs rate. To do so, we will bin the number of cattle into 10% quantile increments and chart the rate for each bin. Again, there seems to be a slight positive relationship between the number of cattle on the farm and the rate of badger sign observation, as shown in the plot on the left. The plot on the right is an alternative view of this relationship.



We also analyzed the relationship between the presence of badger activity and other predictors in number (`accessible_feed_store_present` and `no_setts_in_fields`) with boxplots. The results can be found in the appendix.

Indicator variables

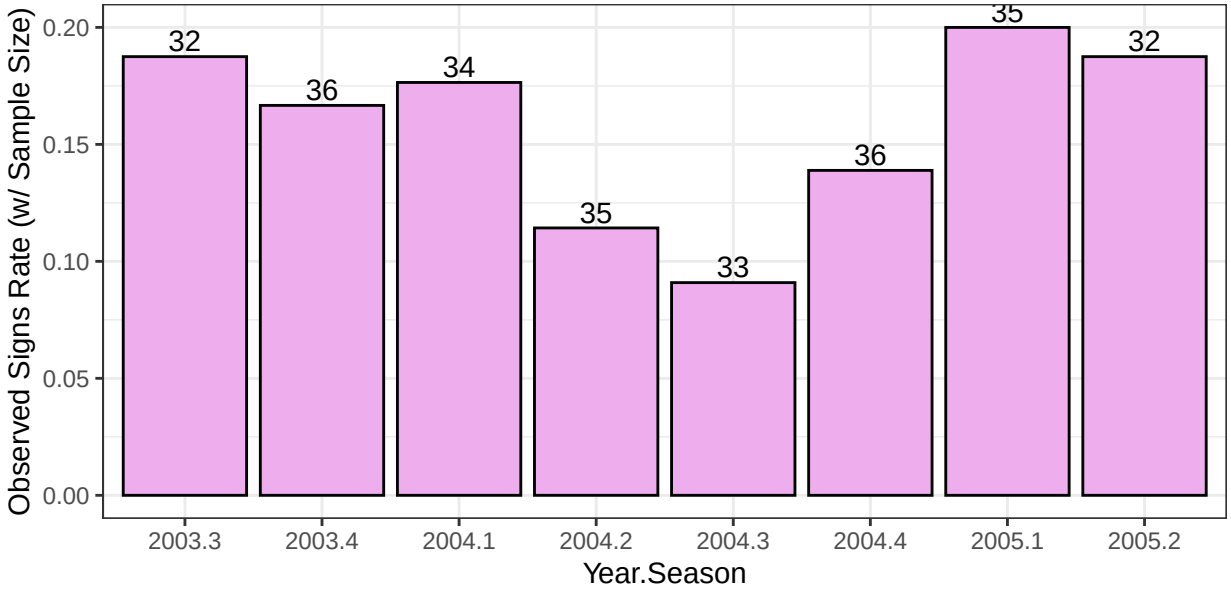
We also have several indicator variables to consider. We can look at the accuracy in predicting the response using each indicator as a sole predictor to investigate these. As displayed in the table below, `proteinblocks`, `sugarbeet`, and `vegetables` provide high accuracy in predicting the presence of badger activity in the farmyard. In addition, to avoid the accuracy paradox, we also plotted the distribution of the presence of badger activities against each indicator variable (see appendix).

Table 1: Accuracy for predicting Signs in Yard

Indicator	Accuracy
accessible_cattle_house_present	0.3590
accessible_feed_present	0.2491
grass_silage	0.6154
cereal_silage	0.6740
hay_straw	0.3516
cereal_grains	0.5348
concentrates	0.5568
proteinblocks	**0.8022**
sugarbeet	**0.7656**
vegetables	**0.7802**
molasses	0.6703

Year and Season

Finally, we may look at the signs rate of the presence of badger activity for each year and season combination. 2003 and 2005 seem to have similar rates, but 2004 may have an appreciably lower rate of badger sign observations.



Model

Model Selection

Our initial model only incorporates the random intercepts for farm, `farm_code_numeric` based on our research question. Then we did forward step-wise model selection with potential predictors found in the EDA part using BIC score, which considers both the likelihood and the model complexity and gives a more general sense of model performance while also being consistent.

Table 2: Forward model selection

Model	BIC
(1|farm_code_numeric)	174.1727
(1|farm_code_numeric) + no_active_setts_in_fields	169.5484
(1|farm_code_numeric) + no_active_setts_in_fields + no_cattle_in_buildings_yard	172.9292
(1|farm_code_numeric) + no_active_setts_in_fields + proteinblocks	174.4175
(1|farm_code_numeric) + no_active_setts_in_fields + sugarbeet	174.5213
(1|farm_code_numeric) + no_active_setts_in_fields + vegetables	173.4542
(1|farm_code_numeric) + no_active_setts_in_fields + year	176.4772
(1|farm_code_numeric) + no_active_setts_in_fields + season	185.5121

We decided only to add the fixed effect `no_active_setts_in_fields` to the initial model from the table above.

Final Model

Our final model is

$$Logit(E(Y_{ij}|b_i)) = X_{ij}^T\boldsymbol{\beta} + b_i$$

where

$$b_i \sim N(0, \sigma^2)$$

Here Y_{ij} represents the presence of badgers at time i on farm j, and X_{ij}'s has two columns, one for the intercept, and another for the `no_setts_in_fields` predictor.

Conclusion

Q1: What factors are relate to presence of badger activity in the farmyard?

Table 3: Estimates of fixed effects

	Estimate	Std. Error	lower 95%	upper 95%
(Intercept)	-4.8705	0.9297	-6.6927	-3.0482
no_active_setts_in_fields	0.4939	0.1541	0.1919	0.7960

Our final model can be interpreted in the following way:

- **Intercept**: For a fixed farm, having no active badger homes in nearby fields, has odds of $e^{-4.8705} \approx 0.0077$ of the presence of badger activity.

- **no_active_setts_in_fields**: For a fixed farm, for each unit increase in the number of active badger homes in nearby fields, the farm will have $e^{0.4939} = 1.6387$ times (a 63.87% increase) the odds of the presence of badger activity.

Therefore, the number of active badger homes in nearby fields is related to the presence of badger activity in the farmyard.

Q2: Estimate the correlation over time in badger activity

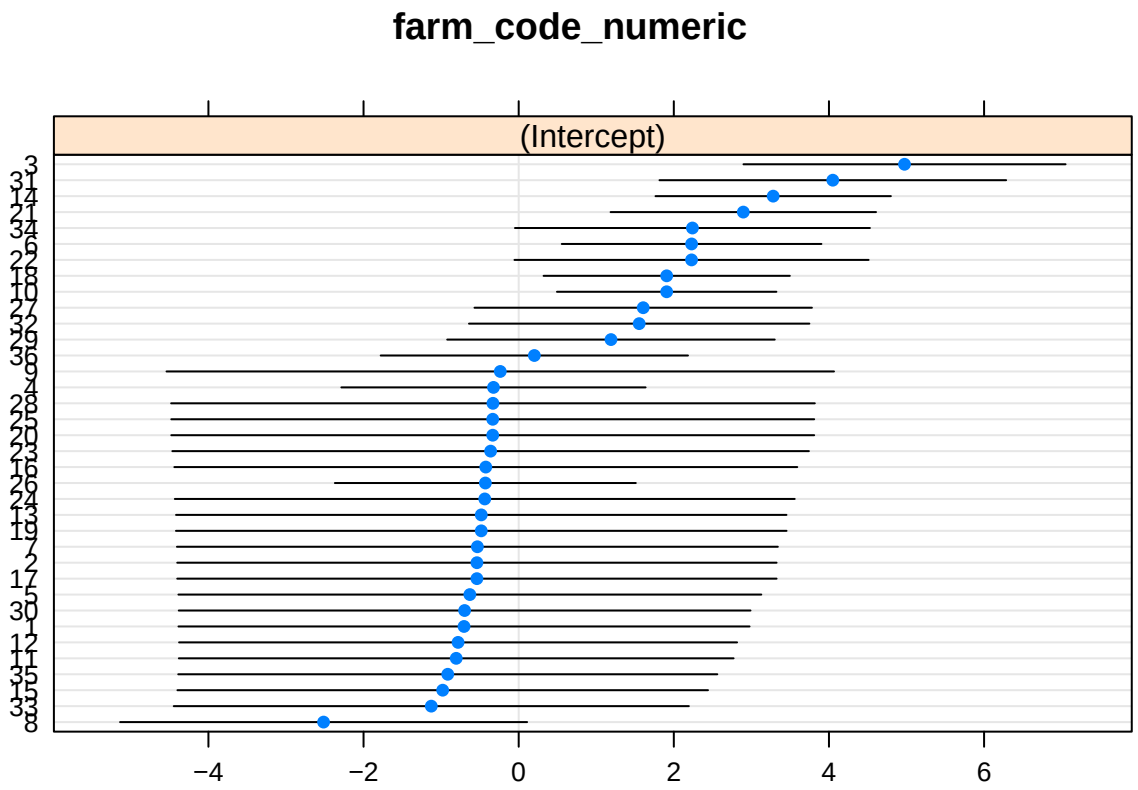
Table 4: Estimates of random effects

Group	Variance
farm_code_numeric	5.9575

$$ICC = \frac{\sigma^2}{\sigma^2 + \frac{\pi^2}{3}} = 0.6442$$

Since the observations for each farm are taken over a period of time, the ICC of this model represents the correlation over time. The estimated interclass correlation is 0.6442, indicating a relatively high correlation over time in badger activity.

Q3: Farm-specific heterogeneity in the tendency to have badger activity



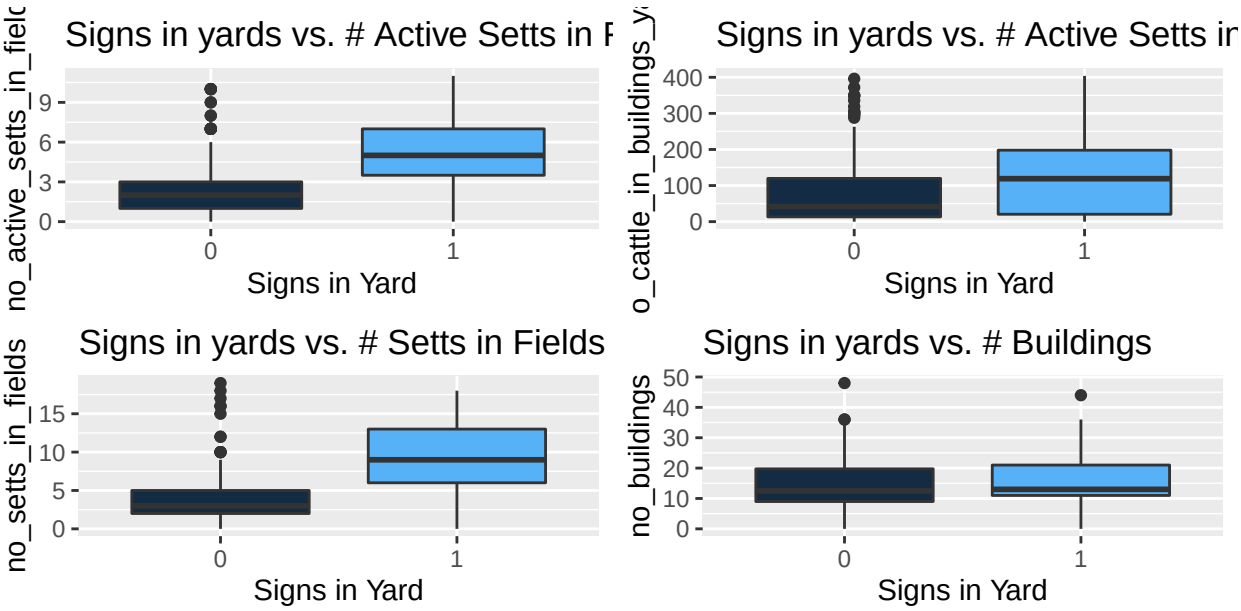
As displayed in the random intercepts plot, we did observe farm-specific heterogeneity in the tendency to have badger activity. On the baseline condition (no active badger homes in nearby fields), the odds of the presence of

badger activity in the farmyard ranges from $e^{-2.5153} = 0.0808$ (farm 8) to $e^{4.9743} = 144.6475$ (farm 3). However, we also saw the confidence intervals being quite wide.

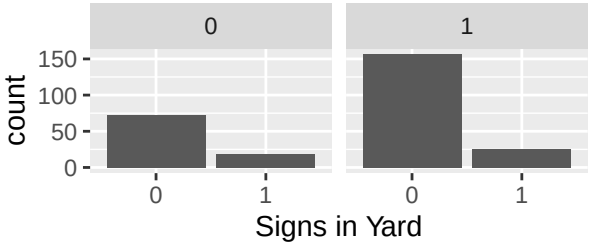
Appendix

Additional plots

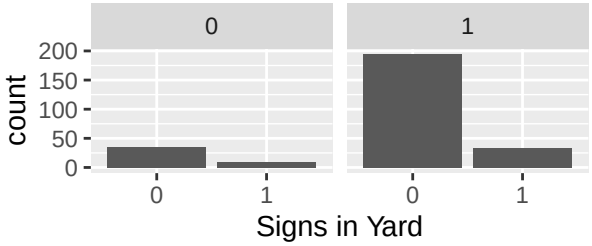
EDA



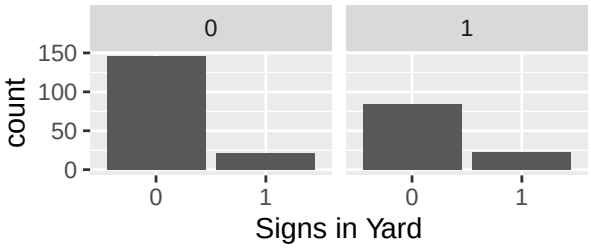
ngs in Yard vs. accessible_cattle_house_



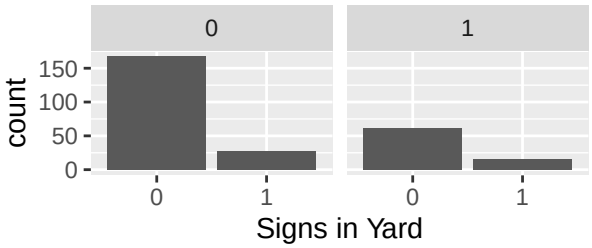
Sings in Yard vs. accessible_feed_pres



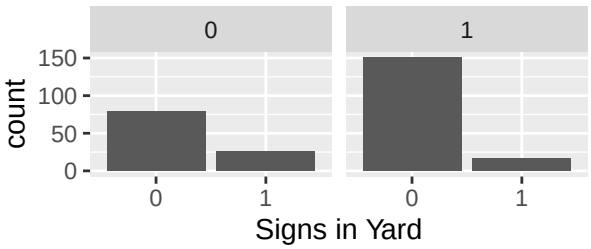
Sings in Yard vs. grass_silage



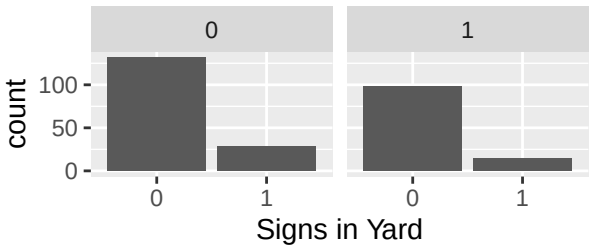
Sings in Yard vs. cereal_silage



Sings in Yard vs. hay_straw



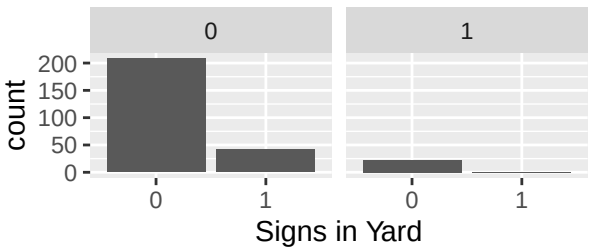
Sings in Yard vs. cereal_grains



Sings in Yard vs. concentrates

Sings in Yard vs. proteinblocks

Sings in Yard vs. sugarbeet



Sings in Yard vs. vegetables

Sings in Yard vs. molasses

Conclusion

Q3: Farm-specific heterogeneity in the tendency to have badger activity

Table 5: Estimated random intercepts

grpvar	term	grp	condval	condsd
farm_code_numeric	(Intercept)	3	4.9743	1.0594
farm_code_numeric	(Intercept)	31	4.0496	1.1401
farm_code_numeric	(Intercept)	14	3.2811	0.7741
farm_code_numeric	(Intercept)	21	2.8958	0.8730
farm_code_numeric	(Intercept)	34	2.2400	1.1674
farm_code_numeric	(Intercept)	6	2.2292	0.8539
farm_code_numeric	(Intercept)	22	2.2285	1.1648
farm_code_numeric	(Intercept)	18	1.9082	0.8099
farm_code_numeric	(Intercept)	10	1.9079	0.7217
farm_code_numeric	(Intercept)	27	1.6052	1.1098
farm_code_numeric	(Intercept)	32	1.5535	1.1195
farm_code_numeric	(Intercept)	29	1.1899	1.0773
farm_code_numeric	(Intercept)	36	0.2017	1.0107
farm_code_numeric	(Intercept)	9	-0.2375	2.1953
farm_code_numeric	(Intercept)	4	-0.3252	1.0007
farm_code_numeric	(Intercept)	28	-0.3316	2.1171
farm_code_numeric	(Intercept)	20	-0.3347	2.1146
farm_code_numeric	(Intercept)	25	-0.3347	2.1146
farm_code_numeric	(Intercept)	23	-0.3618	2.0936
farm_code_numeric	(Intercept)	16	-0.4241	2.0487
farm_code_numeric	(Intercept)	26	-0.4298	0.9897
farm_code_numeric	(Intercept)	24	-0.4373	2.0388
farm_code_numeric	(Intercept)	13	-0.4826	2.0077
farm_code_numeric	(Intercept)	19	-0.4826	2.0082
farm_code_numeric	(Intercept)	7	-0.5324	1.9763
farm_code_numeric	(Intercept)	2	-0.5387	1.9713
farm_code_numeric	(Intercept)	17	-0.5387	1.9713
farm_code_numeric	(Intercept)	5	-0.6296	1.9172
farm_code_numeric	(Intercept)	30	-0.6970	1.8805
farm_code_numeric	(Intercept)	1	-0.7048	1.8784
farm_code_numeric	(Intercept)	12	-0.7816	1.8352
farm_code_numeric	(Intercept)	11	-0.8054	1.8245
farm_code_numeric	(Intercept)	35	-0.9144	1.7730
farm_code_numeric	(Intercept)	15	-0.9796	1.7451
farm_code_numeric	(Intercept)	33	-1.1267	1.6941
farm_code_numeric	(Intercept)	8	-2.5153	1.3386

Codes

```
library(tidyverse)
library(lme4)
library(janitor)
library(reshape2)
library(kableExtra)
library(caret)
library(lattice)
library(gridExtra)
library(patchwork)

knitr::opts_chunk$set(
  message=FALSE,
  warning=FALSE
)
```

```
data <- read.table('BadgersFarmSurveysNoNA.txt', header=TRUE) %>%
  janitor::clean_names() %>%
  rename(no_cattle_in_buildings_yard=no_cattle_in_buidlings_yard)
```

```
data %>%
  group_by(farm_code_numeric) %>%
  summarise(
    n=n(),
    signs_rate=mean(signs_in_yard)
  ) %>%
  ggplot(aes(x=reorder(farm_code_numeric, -signs_rate), y=signs_rate)) +
    geom_bar(stat='identity', color='black', fill='lightcoral') +
    geom_text(aes(label=n), vjust=-0.25) +
    labs(x='Farm Code', y='Observed Signs Rate (w/ Sample Size)') +
    ggtitle("Signs in yards vs. Farms") +
    scale_x_discrete(guide=guide_axis(angle=45)) +
    theme_bw()
```

```
data %>%
  group_by(no_active_setts_in_fields) %>%
  summarise(
    n=n(),
    signs_rate=mean(signs_in_yard)
  ) %>%
  ggplot(aes(x=no_active_setts_in_fields, y=signs_rate)) +
    geom_bar(stat='identity', color='black', fill='lightblue') +
    geom_text(aes(label=n), vjust=-0.25) +
    scale_x_continuous(
      labels=as.character(unique(data$no_active_setts_in_fields)),
      breaks=unique(data$no_active_setts_in_fields)
    ) +
    ggtitle("Signs in yards vs. # Active Settlemments") +
    labs(x='# Active Settlements', y='Observed Signs Rate (w/ Sample Size)') +
    theme_bw()
```

```r
p1 <- data %>%
  mutate(cattle_bin=cut(
      no_cattle_in_buildings_yard,
      breaks=quantile(no_cattle_in_buildings_yard, (1:10)/10),
      include.lowest=TRUE
    )
  ) %>%
  group_by(cattle_bin) %>%
  summarise(
    n=n(),
    signs_rate=mean(signs_in_yard)
  ) %>%
  ggplot(aes(x=cattle_bin, y=signs_rate)) +
    geom_bar(stat='identity', color='black', fill='lightgreen') +
    geom_text(aes(label=n), vjust=-0.25) +
    scale_x_discrete(guide=guide_axis(angle=45)) +
    ggtitle("Signs in yards vs. # Cattle") +
    labs(x='# Cattle', y='Observed Signs Rate (w/ Sample Size)') +
    theme_bw()


p2 <- data %>%
  mutate(cattle_bin=cut_interval(no_cattle_in_buildings_yard, n=25)) %>%
  group_by(cattle_bin) %>%
  summarise(
    n=n(),
    signs_rate=mean(signs_in_yard)
  ) %>%
  ggplot(aes(x=1:length(unique(cattle_bin)), y=signs_rate)) +
    geom_point() +
    geom_smooth(method='lm') +
    scale_x_discrete(guide=guide_axis(angle=45)) +
    ggtitle("Signs in yards vs. # Cattle (alternative)") +
    labs(x='# Cattle (Bin Index)', y='Observed Signs Rate (w/ Sample Size)') +
    theme_bw()

grid.arrange(p1,p2,nrow=1)
```

```r
p1 <- ggplot(data, aes(x = as.factor(signs_in_yard), y = no_active_setts_in_fields, fill = signs_in_
  geom_boxplot() +
  theme(legend.position = "none") +
  ggtitle("Signs in yards vs. # Active Setts in Fields") +
  labs(x = "Signs in Yard")

p2 <- ggplot(data, aes(x = as.factor(signs_in_yard), y = no_cattle_in_buildings_yard, fill = signs_i
  geom_boxplot() +
  theme(legend.position = "none") +
  ggtitle("Signs in yards vs. # Active Setts in Buildings Yard") +
  labs(x = "Signs in Yard")

p3 <- ggplot(data, aes(x = as.factor(signs_in_yard), y = no_setts_in_fields, fill = signs_in_yard))
```

```
  geom_boxplot() +
  theme(legend.position = "none") +
  ggtitle("Signs in yards vs. # Setts in Fields") +
  labs(x = "Signs in Yard")


p4 <- ggplot(data, aes(x = as.factor(signs_in_yard), y = no_buildings, fill = signs_in_yard)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  ggtitle("Signs in yards vs. # Buildings") +
  labs(x = "Signs in Yard")

grid.arrange(p1, p2, p3, p4, nrow =2)
```

```
data %>%
  select(accessible_cattle_house_present:molasses) %>%
  sapply(., function(x) mean(x==data$signs_in_yard)) %>%
  enframe(name='Indicator', value='Accuracy') %>%
  kable(digits=4, caption = "Accuracy for predicting Signs in Yard") %>%
  kable_classic(full_width=FALSE, latex_options='HOLD_position') %>%
  row_spec(c(8:10), bold=TRUE)
```

```
## maybe add the plots to appendix
a <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("accessible_cattle_house_present", scales = "fixed") +
  ggtitle("Sings in Yard vs. accessible_cattle_house_present") +
  theme(plot.title = element_text(hjust = 0.5))

b <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("accessible_feed_present", scales = "fixed") +
  ggtitle("Sings in Yard vs. accessible_feed_present") +
  theme(plot.title = element_text(hjust = 0.5))

c <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("grass_silage", scales = "fixed") +
  ggtitle("Sings in Yard vs. grass_silage") +
  theme(plot.title = element_text(hjust = 0.5))

d <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
```

```
  facet_wrap("cereal_silage", scales = "fixed") +
  ggtitle("Sings in Yard vs. cereal_silage") +
  theme(plot.title = element_text(hjust = 0.5))

e <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("hay_straw", scales = "fixed") +
  ggtitle("Sings in Yard vs. hay_straw") +
  theme(plot.title = element_text(hjust = 0.5))

f <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("cereal_grains", scales = "fixed") +
  ggtitle("Sings in Yard vs. cereal_grains") +
  theme(plot.title = element_text(hjust = 0.5))

g <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("concentrates", scales = "fixed") +
  ggtitle("Sings in Yard vs. concentrates") +
  theme(plot.title = element_text(hjust = 0.5))

h <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("proteinblocks", scales = "fixed") +
  ggtitle("Sings in Yard vs. proteinblocks") +
  theme(plot.title = element_text(hjust = 0.5))

i <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("sugarbeet", scales = "fixed") +
  ggtitle("Sings in Yard vs. sugarbeet") +
  theme(plot.title = element_text(hjust = 0.5))

j <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("vegetables", scales = "fixed") +
  ggtitle("Sings in Yard vs. vegetables") +
  theme(plot.title = element_text(hjust = 0.5))
```

```r
k <- ggplot(data, aes(x = as.factor(signs_in_yard))) +
  geom_bar() +
  theme(legend.position = "none") +
  labs(x = "Signs in Yard") +
  facet_wrap("molasses", scales = "fixed") +
  ggtitle("Sings in Yard vs. molasses") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(a, b, c, d, nrow =2)

grid.arrange(e, f, g, h, nrow =2)

grid.arrange(j, h, k, nrow =2)
```

```r
data %>%
  mutate(year_season=paste0(year, '.', season)) %>%
  group_by(year_season) %>%
  summarise(
    n=n(),
    signs_rate=mean(signs_in_yard)
  ) %>%
  ggplot(aes(x=year_season, y=signs_rate)) +
    geom_bar(stat='identity', color='black', fill='plum2') +
    geom_text(aes(label=n), vjust=-0.25) +
    labs(x='Year.Season', y='Observed Signs Rate (w/ Sample Size)') +
    theme_bw()
```

```r
# make relevant variables into factors
model_data <- data
model_data[, c(1, 2, 10:21)] <- lapply(data[, c(1, 2, 10:21)] , factor)

# stepwise model selection
m0 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric),
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m0) # 174.1727

m1 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m1) # 169.5484

m2 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + no_cattle_in_buildings_yard
  family='binomial',
```

```
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m2) # 172.9292

m3 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + proteinblocks,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m3) # 174.4175

m4 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + sugarbeet,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m4) # 174.5285

m5 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + vegetables,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m5) # 173.4542

m6 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + year,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m6) # 176.4669

m7 <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields + season,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
BIC(m7) # 185.5121
```

```
model <- c("(1|farm_code_numeric)",
           "(1|farm_code_numeric) + no_active_setts_in_fields",
           "(1|farm_code_numeric) + no_active_setts_in_fields + no_cattle_in_buildings_yard",
           "(1|farm_code_numeric) + no_active_setts_in_fields + proteinblocks",
           "(1|farm_code_numeric) + no_active_setts_in_fields + sugarbeet",
           "(1|farm_code_numeric) + no_active_setts_in_fields + vegetables",
```

```
          "(1|farm_code_numeric) + no_active_setts_in_fields + year",
          "(1|farm_code_numeric) + no_active_setts_in_fields + season")

BIC_score <- sapply(c(m0, m1, m2, m3, m4, m5, m6, m7), BIC)

data.frame("Model" = model, 'BIC' = BIC_score) %>%
  kable(caption = "Forward model selection") %>%
    kable_styling(latex_options = c("HOLD_position","striped"))
```

```
model_final <- glmer(
  signs_in_yard ~ (1 | farm_code_numeric) + no_active_setts_in_fields,
  family='binomial',
  control=glmerControl(optimizer = "bobyqa"),
  data=model_data
)
```

```
summary(model_final)$coefficients %>%
  as.data.frame() %>%
  mutate("lower 95%" = Estimate - 1.96 * `Std. Error`,
         "upper 95%" = Estimate + 1.96 * `Std. Error`)  %>%
  select(c("Estimate", "Std. Error", "lower 95%", "upper 95%")) %>%
  kable(caption = "Estimates of fixed effects", digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))
```

```
data.frame(summary(model_final)$varcor) %>%
  mutate(Group = grp,
         Variance = vcov) %>%
  select(c("Group", "Variance")) %>%
  kable(caption = "Estimates of random effects", digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))

sigma2hat <- 5.9575
icc <- sigma2hat / (sigma2hat + pi^2/3)
```

```
dotplot(ranef(model_final))$farm_code_numeric
```

```
ranef(model_final, condVar = TRUE) %>%
  as.data.frame() %>%
  filter(grpvar == "farm_code_numeric") %>%
  arrange(desc(condval)) %>%
  kable(caption = "Estimated random intercepts", longtable = T, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position"))
```