Modeling

Robert C Cline Sr

2022-04-14

### OurCodingClub Modeling Tutorial

Format: *What is the effect of the predictor.variable on the response.variable?
temp.model <- lm(dependent/response.variable ~ predictor.variable)

- `skylark.m <- lm(abundance ~ treatment + farm.area, family = poisson, data = skylarks)`

- Abundance represents count

- zero-inflated data allows for zero-valued observations, for which Poisson family is suitable.
- Continuous data use lm, mixed-effects models
- Count data -
- Poisson: glm, glmm

- Proportion data
  - if more outcomes: chi-squared test
  - habitat selection (does a species utilize a type of habitat in greater proportion than its availability.

  - chi squared: differences in vegitation types between sites or over time
  - binomial: glm, glmm

Model structure

- Let the hypothesis guide you.

- what do you want to examine; what are the *confounding varibles* that influence the response?

E.g. `skylark.m <- lm(abundance ~ treatment + farm.area)`

Overfitting

- If your model has a lot of variables, it has a danger of *overfitting*

- The model will be super-tailored to this specific dataset.

Collinearity

- If variables are very correlated, they will both explain similar amounts of variation in the response variables. E.g. mixing elevation and air temp effect on tree height.

Practice with linear models

```
## Warning: package ’agridat’ was built under R version 4.1.3
```

```
##   rep row pos spacing    stock     gen yield trt
## 1  R1   2   1       6 Seedling Redspur  70.9 601
## 2  R1   2   2       6 Seedling  Golden 130.9 602
## 3  R1   2   8       6    MM111 Redspur 114.5 611
## 4  R1   2   7       6    MM111  Golden  90.5 612
## 5  R1   2   3       6    M0007 Redspur 151.8 671
## 6  R1   2   4       6    M0007  Golden 125.0 672
```

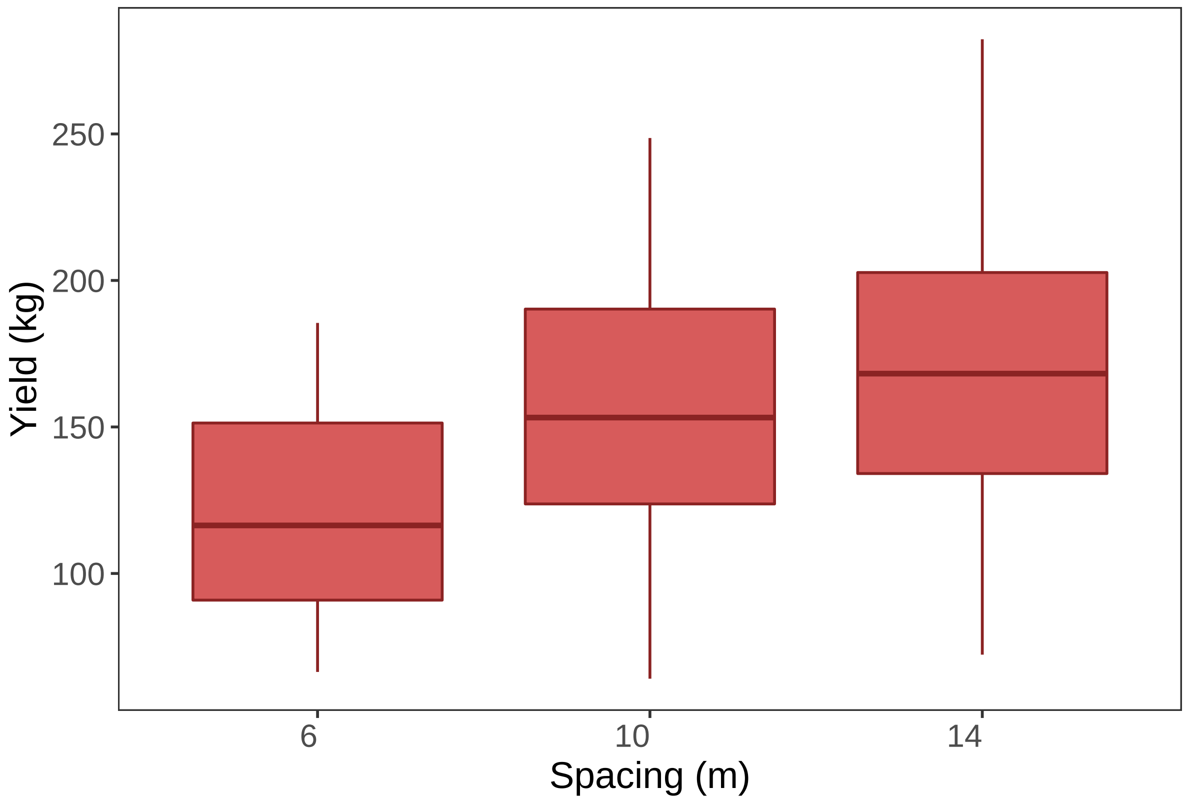
```
##  rep          row              pos            spacing        stock
##  R1:24   Min.   : 2.000   Min.   : 1.000   Min.   : 6   M0007   :30
##  R2:24   1st Qu.: 5.750   1st Qu.: 5.000   1st Qu.: 6   MM106   :30
##  R3:24   Median : 9.000   Median :10.000   Median :10   MM111   :30
##  R4:24   Mean   : 9.017   Mean   : 9.242   Mean   :10   Seedling:30
##  R5:24   3rd Qu.:13.000   3rd Qu.:14.000   3rd Qu.:14
##          Max.   :16.000   Max.   :17.000   Max.   :14
##
##       gen         yield           trt
##  Golden :60   Min.   : 64.1   Min.   : 601.0
##  Redspur:60   1st Qu.:108.2   1st Qu.: 668.8
##               Median :147.1   Median :1036.5
##               Mean   :145.4   Mean   :1036.5
##               3rd Qu.:176.5   3rd Qu.:1404.2
##               Max.   :282.3   Max.   :1472.0
##               NA’s   :28
```

Visualize the data

- Create *theme.clean*
- Check out the effect of spacing on apple yield.

- H0: The closer apple trees aare to other apple trees, the more they compete for resources
- Thus, the closer the trees are to each other, the less their yield.
- There are only three spacing distances, so make them a category.

```
## Warning: Removed 28 rows containing non-finite values (stat_boxplot).
```



Note that putting your entire ggplot code in brackets () creates the graph and then shows it in the plot viewer. If you don't have the brackets, you've only created the object, but will need to call it to visualise the plot.

From our boxplot, we can see that yield is pretty similar across the different spacing distances. Even though there is a trend towards higher yield at higher spacing, the range in the data across the categories almost completely overlap. From looking at this boxplot alone, one might think our hypothesis of higher yield at higher spacing is not supported. **Let's run a model to explicitly test this.**

```
## Warning: package 'sjPlot' was built under R version 4.1.2

## Install package "strengejacke" from GitHub ('devtools::install_github("strengejacke/strengejacke")')

## Warning: package 'sjmisc' was built under R version 4.1.2

##
## Attaching package: 'sjmisc'

## The following object is masked from 'package:purrr':
##
##     is_empty

## The following object is masked from 'package:tidyr':
##
##     replace_na

## The following object is masked from 'package:tibble':
##
##     add_case

##
## Attaching package: 'sjlabelled'
## The following object is masked from ’package:forcats’:
##
##     as_factor
```

```
## The following object is masked from ’package:dplyr’:
##
##     as_label
```

```
## The following object is masked from ’package:ggplot2’:
##
##     as_label
```

yield

Predictors

Estimates

CI

p

(Intercept)

120.57

105.90 – 135.23

<0.001

spacing2 [10]

35.92

13.92 – 57.93

0.002

spacing2 [14]

44.11

22.32 – 65.90

<0.001

Observations

92

R2 / R2 adjusted

0.174 / 0.156