Correlation Structure

Lecture 07.3: Model Selection

Lauren Sullivan

Module: Linear, Nonlinear, and Mixed Effects Models

Readings

Required for class:

- NA

Optional:

- Prabhakaran, S. r-statistics.co - Model Selection Approaches
- Grace et al. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. Nature

Model Selection

Sometimes you have a lot of predictor variables and you want to find out which ones you should keep in the model that best predict your response variable. For example, you have a lot of environmental variables and you want to know what is affecting a biological factor.

- You have a lot of water variables that describe a stream and you want to know how that affects the invertebrate abundance.
- You have a lot of plant chemistry data and you want to know what variables predict protein content in seeds.

Model Selection

This often occurs when you do not have a hypothesis but you have a lot of data and you want to know what is significant.

Note: Some reviewers will not like this because it can feel like data dredging. But if it makes sense for your research question, you just need to justify it.

Stepwise Regression

Stepwise regression is a way to do model selection where you put in a bunch of additive variables into a linear model, and the `step()` function in the `stats` library uses backward selection (by default) to iteratively search through all variables and determine which ones have the most power.

The iterations occur by dropping one variable at a time and then all models are tested against each other with AIC. The variable that produces the smallest AIC score when dropped is then dropped for the next round and so on until there is no more significant drop in AIC.

Nutrient Network

Let's try an example with the Nutrient Network, examining how different plant factors (e.g. richness and productivity) respond to various environmental factors (e.g. temperature, preciptitation, nitrogen, carbon, soil fertility, etc.)

LETTER

doi:10.1038/nature16524

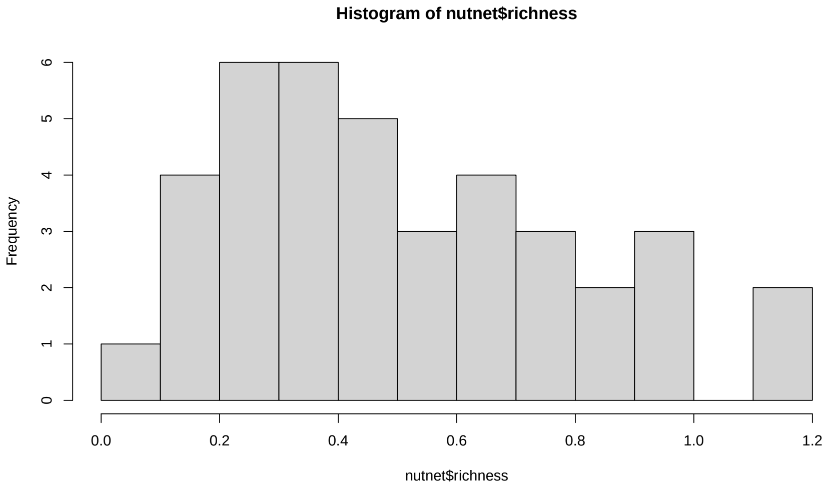
Integrative modelling reveals mechanisms linking productivity and plant species richness

James B. Grace[1], T. Michael Anderson[2], Eric W. Seabloom[3], Elizabeth T. Borer[3], Peter B. Adler[4], W. Stanley Harpole[5,6,7], Yann Hautier[8], Helmut Hillebrand[9], Eric M. Lind[3], Meelis Pärtel[10], Jonathan D. Bakker[11], Yvonne M. Buckley[12], Michael J. Crawley[13], Ellen I. Damschen[14], Kendi F. Davies[15], Philip A. Fay[16], Jennifer Firn[17], Daniel S. Gruner[18], Andy Hector[19], Johannes M. H. Knops[20], Andrew S. MacDougall[21], Brett A. Melbourne[15], John W. Morgan[22], John L. Orrock[14], Suzanne M. Prober[23] & Melinda D. Smith[24]

Plant Species Richness

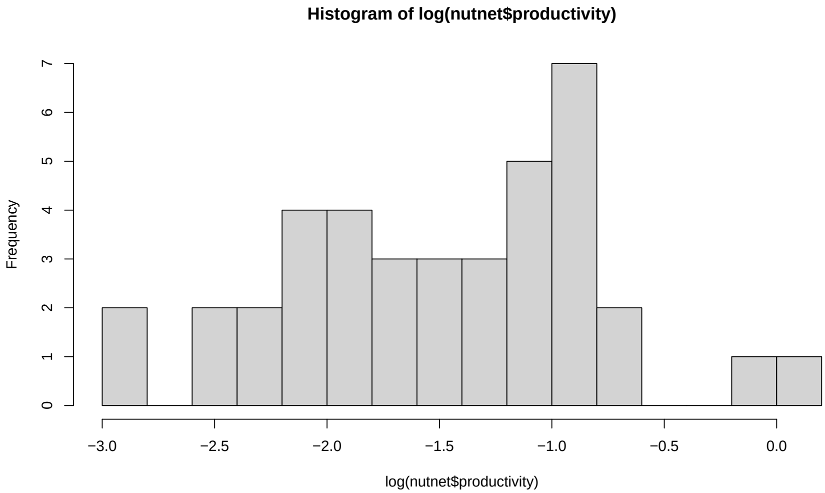
```r
nutnet <- read_csv("../data/nutnet.csv")

hist(nutnet$richness, breaks = 15)
```



Plant Productivity

```r
hist(log(nutnet$productivity), breaks = 15)
```



Stepwise Regression

```
richness <- nutnet[,-c(1,3)]
productivity <- nutnet[,-c(1:2)]

productivity[1:5,]
```

```
## # A tibble: 5 x 13
##   productivity SoilS~1 SoilFe~2 Climate sand.~3 silt.~4    ph  ln.p   ln.n
##          <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>  <dbl> <
## 1        0.441  -0.714 -0.00844  0.0506   0.757   0.159  5.56  4.19 0.395  2
## 2        0.369  -0.888  0.149   -0.0835   0.375   0.558  5.84  4.25 0.431  1
## 3        0.208  -0.715 -0.188    0.126    0.557   0.287  6.00  2.91 0.0554 0
## 4        0.167  -0.712 -0.344    0.122    0.72    0.181  6.82  2.86 0.0983 0
## 5        0.142  -0.795 -0.173   -0.0170   0.704   0.265  5.55  2.70 0.478  2
## # ... with 3 more variables: ln.k <dbl>, teperature <dbl>, preciptiation <db
## #   and abbreviated variable names 1: SoilSuitability, 2: SoilFertility,
## #   3: sand.prop, 4: silt.prop
```

Stepwise Regression

```
richness_lm <- lm(log(richness) ~ ., data = richness)
richness_selectedMod <- step(richness_lm)
```

```
summary(richness_selectedMod)
```

```
##
## Call:
## lm(formula = log(richness) ~ SoilFertility + Climate + teperature,
##     data = richness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13297 -0.28797  0.01088  0.29452  0.80449
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.0610     0.1388  -7.643 5.77e-09 ***
## SoilFertility  -2.6306     0.5216  -5.044 1.41e-05 ***
## Climate        -5.4161     1.0760  -5.033 1.45e-05 ***
## teperature      0.2507     0.1035   2.422   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4432 on 35 degrees of freedom
## Multiple R-squared:  0.5042, Adjusted R-squared:  0.4617
## F-statistic: 11.87 on 3 and 35 DF,  p-value: 1.634e-05
```

Stepwise Regression

```
# all_vifs <- car::vif(selectedMod)
# print(all_vifs)
#
# ?vif
```