ENVX2001 - Model selection

Learning Outcomes

In this lab, you will work towards achieving learning outcomes

Lab Objectives

In this lab, we will:

- []
- []

> **Tip**
>
> Please work on this exercise by creating your own R Markdown file.

Exercise 1: Model quality - multiple vs adjusted r^2

Data: *California_streamflow* spreadsheet

Import the "California_streamflow" sheet into R.

CODE

```r
# Load library if needed
library(readxl)
# Load data
stream_data <- read_xlsx("data/california_streamflow.xlsx", "streamflow")
```

in this exercise we will use the same data as last week. To jog your memory, the dataset contains 43 years of annual precipitation measurements (in mm) taken at (originally) 6 sites in the Owens Valley in California. Through model selection via partial F-test we have the final model as below:

CODE

```
fit <- lm(runoff_volume ~ rock_creek + pine_creek, data = stream_data)
```

We will now add a totally useless variable to the dataset. This variable is a random number generated from a normal distribution with mean 3 and standard deviation 2. We use the `set.seed()` function to make sure that everybody gets the same random values.

CODE

```
set.seed(100) # to make sure everybody gets the same results

# this generates the random number into the dataset
stream_data$random_no <- rnorm(n = nrow(stream_data), mean = 3, sd = 2)
```

We will see the impact of including a totally useless variable, such as this random variable, has on measures of model quality, r^2 and adjusted r^2 values.

Task: create two regression models:

1. `runoff_volume ~ rock_creek + pine_creek`
2. `runoff_volume ~ rock_creek + pine_creek + random_no`

Question 2

Compare each in terms of their multiple r^2 and adjusted r^2 values. Which performance measure (multiple r^2 or adj r^2) would you use to identify which predictors to use in your model?

Answer

There is only a small difference in the multiple r^2 and adj r^2 between the models, but one goes up and the other goes down. Since the random number is a totally useless value, this demonstrates that the adjusted r^2 is the better performance measure to use.

CODE

```
mod_rand1 <- lm(runoff_volume ~ rock_creek + pine_creek,data = stream_data)
summary(mod_rand1)
```

OUTPUT

```

Call:
lm(formula = runoff_volume ~ rock_creek + pine_creek, data = stream_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09832 -0.02350  0.01076  0.03291  0.08568 

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.35762    0.10547  31.835  < 2e-16 ***
rock_creek   0.44437    0.08925   4.979 1.26e-05 ***
pine_creek   0.21051    0.06861   3.068  0.00385 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04937 on 40 degrees of freedom
Multiple R-squared:  0.8749,	Adjusted R-squared:  0.8686
F-statistic: 139.8 on 2 and 40 DF,  p-value: < 2.2e-16
```

CODE

```r
mod_rand2 <- lm(runoff_volume ~ rock_creek + pine_creek + random_no ,data = stream_data)
summary(mod_rand2)
```

OUTPUT

```

Call:
lm(formula = runoff_volume ~ rock_creek + pine_creek + random_no,
    data = stream_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10327 -0.02401  0.01220  0.03179  0.08056

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.359356   0.106303  31.602  < 2e-16 ***
rock_creek  0.442795   0.089956   4.922  1.6e-05 ***
pine_creek  0.207234   0.069324   2.989  0.00482 **
random_no   0.003213   0.005077   0.633  0.53055
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04974 on 39 degrees of freedom
Multiple R-squared:  0.8761,	Adjusted R-squared:  0.8666
F-statistic: 91.96 on 3 and 39 DF,  p-value: < 2.2e-16
```

Exercise 2: Fish productivity

Fish communities were surveyed in lakes across a eutrophication gradient to investigate the relatioship between productivity and fish diversity.

The datasheet has thefollowing variables:

- `lake_id` Unique identifier for each lake
- `chla` Log-transformed Chlorophyll a
- `richness` Log-transformed species richness
- `evenness` Log-transformed Pielou's evenness

- `abundance` Log-transformed abundance per unit effort (NPUE)
- `biomass` Log-transformed biomass per unit effort (BPUE)
- `productivity` Log-transformed productivity proxy

CODE

```r
#Load library if necessary
library(tidyverse)

#Read in data
fish_data <- read_csv("data/fish_communities.csv")
```

OUTPUT

```
Rows: 39 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): lake_id
dbl (6): chla, richness, evenness, abundance, biomass, productivity

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

CODE

```r
#Have a look at the structure of the data
str(fish_data)
```

OUTPUT

```
spc_tbl_ [39 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ lake_id     : chr [1:39] "FTH" "XLH" "HGH" "LH" ...
 $ chla        : num [1:39] 4.61 4.53 3.84 3.71 2.43 ...
 $ richness    : num [1:39] 2.64 2.89 1.99 2.94 2.37 ...
 $ evenness    : num [1:39] -0.356 -0.67 -0.423 -0.405 -0.589 ...
 $ abundance   : num [1:39] 2.24 3.34 1.24 2.67 1.12 ...
 $ biomass     : num [1:39] 4.96 6.03 4.67 6.52 5.32 ...
 $ productivity: num [1:39] 3.47 4.47 2.71 4.47 3.24 ...
 - attr(*, "spec")=
  .. cols(
  ..   lake_id = col_character(),
  ..   chla = col_double(),
  ..   richness = col_double(),
  ..   evenness = col_double(),
  ..   abundance = col_double(),
  ..   biomass = col_double(),
  ..   productivity = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Explore the data on your own before making any models. (*Hint:* remember that we are only interested in the numeric variables for our linear models)

Question 1

Are there any obvious relationships between the response variable (productivity) and the other variables?

Answer

CODE

```
#correlation matrix
cor(fish_data[,-1])
```

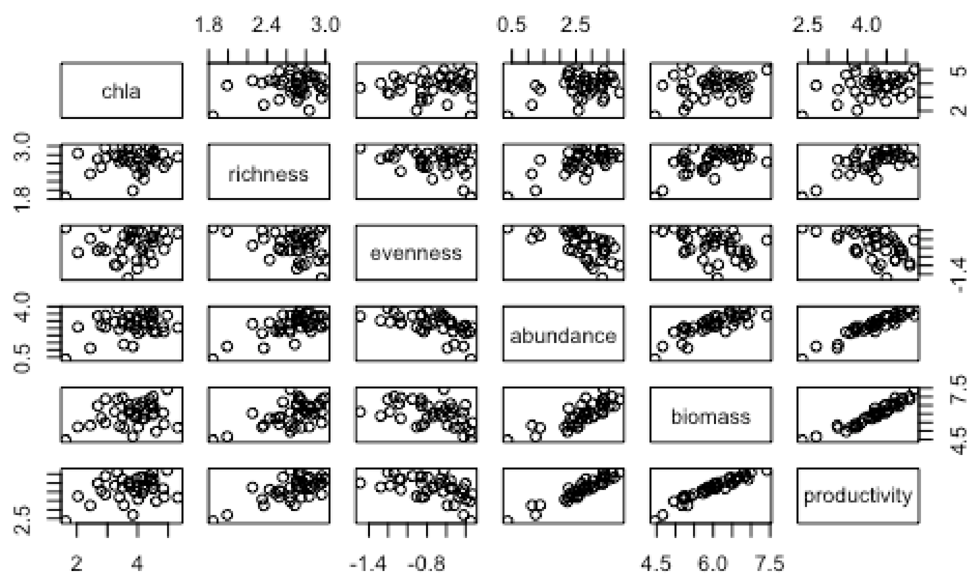
OUTPUT

```
                  chla   richness   evenness  abundance    biomass productivity
chla         1.0000000  0.2388845  0.0872381  0.3449158  0.2809369    0.3406440
richness     0.2388845  1.0000000 -0.3295238  0.6585895  0.5704803    0.6611413
evenness     0.0872381 -0.3295238  1.0000000 -0.5553243 -0.4396433   -0.5162017
abundance    0.3449158  0.6585895 -0.5553243  1.0000000  0.8089680    0.9379798
biomass      0.2809369  0.5704803 -0.4396433  0.8089680  1.0000000    0.9571907
productivity 0.3406440  0.6611413 -0.5162017  0.9379798  0.9571907    1.0000000
```

CODE

```
#Plot the variables

pairs(fish_data[,-1])
```



Productivity has a strong relationship with biomass and abundance

Question 2

Are there any other patterns or potential issues you can see from the exploratory plots?

Answer

Biomass is also highly correlated with abundance, so collinearity might be an issue.

Take home exercise:

Review

Attribution