Exercises for “Multiple regression”

Timothée Bonnet

May 16, 2019

Find the content for today and previous workshops at https://github.com/timotheenivalis/RSB-R-Stats-Biology.

Contents

1 Multiple regression

* Exercise 1 Jumping

1. load jumpingdistance.csv. It contains jumping distances by people of different masses and heights.

2. Use plots and lm() to test whether mass increases or decreases jumping distance. Based on the classical mechanics what do you expect?

Answer of exercise 1

```
jumping <- read.csv(file = "jumpingdistance.csv")
```

A first approach suggests mass increases jumping distance:

```
summary(lm(jump ~ mass, data=jumping))
plot(mass, jump)
```

But that is incorrect and due to the correlation between mass and height:

```
summary(lm(jump ~ mass + height, data=jumping))
```

The direct (causal) effect of mass is negative, as revealed by a multiple regression. The NET effect of mass is positive, but conditional on height mass as a negative effect.

* Exercise 2 Babies

1. Load babies.csv

2. What drives change in number of babies born?

Answer of exercise 2

```r
babies <- read.csv("babies.csv")
summary(lm(babies_born ~ number_of_storks, data = babies))
```

```
##
## Call:
## lm(formula = babies_born ~ number_of_storks, data = babies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.723 -0.634 -0.286  0.572  2.302
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.4569     0.1747   82.74   <2e-16 ***
## number_of_storks   0.0886     0.0161    5.51    1e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 54 degrees of freedom
## Multiple R-squared:  0.36,Adjusted R-squared:  0.348
## F-statistic: 30.4 on 1 and 54 DF,  p-value: 1.02e-06
```

```r
summary(lm(babies_born ~ number_of_storks + year, data = babies))
```

```
##
## Call:
## lm(formula = babies_born ~ number_of_storks + year, data = babies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.871 -0.686 -0.178  0.769  2.124
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -72.3262    27.9141   -2.59    0.012 *
## number_of_storks   0.0196     0.0268    0.73    0.468
## year               0.0439     0.0141    3.11    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.953 on 53 degrees of freedom
## Multiple R-squared:  0.459,Adjusted R-squared:  0.438
## F-statistic: 22.5 on 2 and 53 DF,  p-value: 8.64e-08
```

```r
summary(lm(babies_born ~ year, data = babies))
```

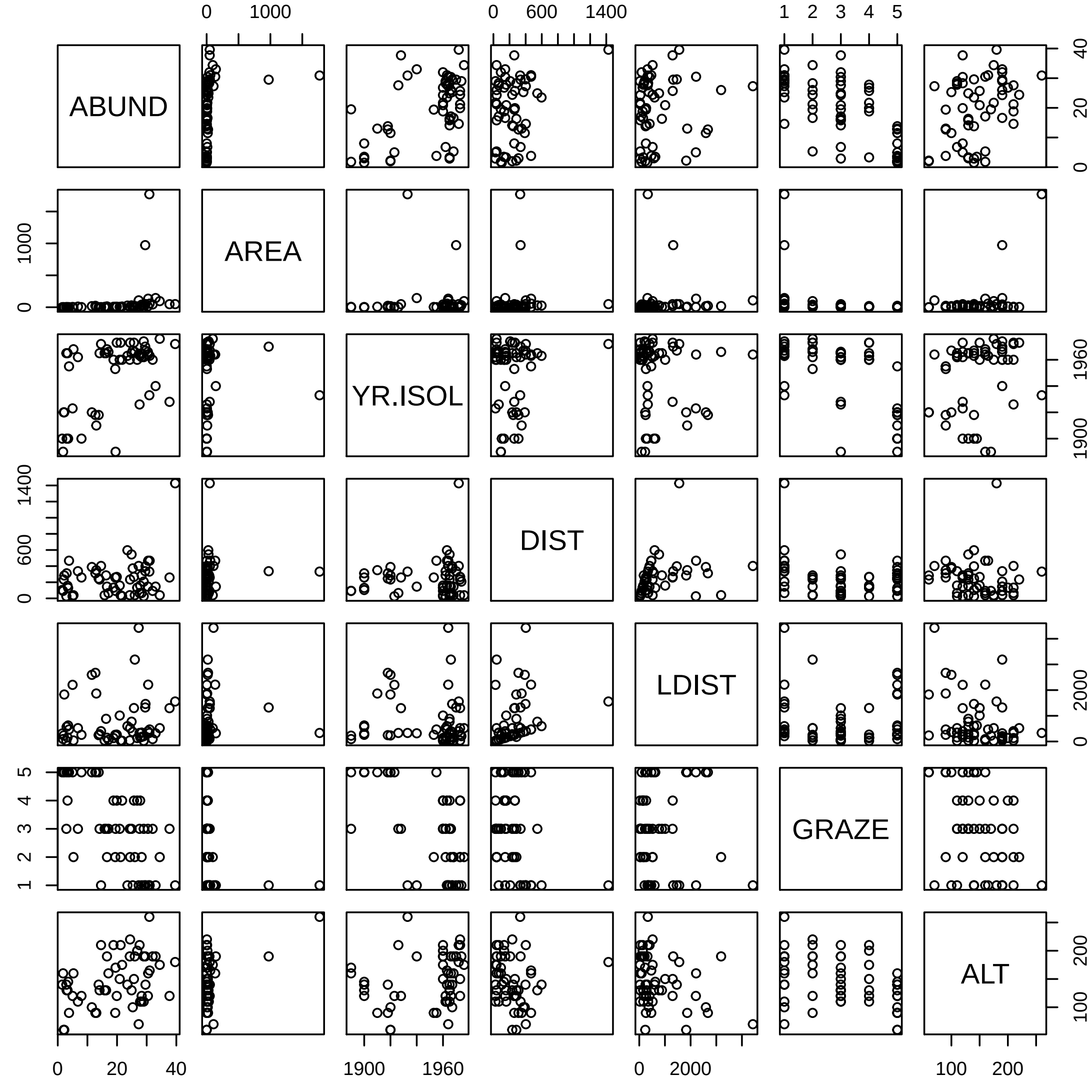
```
##
## Call:
## lm(formula = babies_born ~ year, data = babies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.922 -0.675 -0.208  0.723  2.133
##
```

** Exercise 3 Bird abundance

Loyn (1987) modeled the abundance of forest birds with six predictor variables (patch area, distance to nearest patch, distance to nearest larger patch, grazing intensity, altitude and years since the patch had been isolated). That is a classical example wrongly analyses in textbooks (they tend to say that the initial analysis was wrong because of correlations between predictors…however linear models do not make assumptions about correlations among predictors, as long as the correlations are not 1 or -1). Load the dataset `loyn.csv`. Think of a reasonable causal model that would predict bird abundance. Before rushing to fit models, look at the distributions of variables, some of them may benefit from a log-transformation (for practical convenience and for logic both!). Test it using an appropriate multiple regression. Also try a model containing all predictors. Compare your results to that of a series of simple regressions (one for each of your predictors). Try to understand the differences.

Answer of exercise 3

```r
birds <- read.csv("loyn.csv")
plot(birds)
```


ABUND
AREA
YR.ISOL
DIST
LDIST
GRAZE
ALT

```
summary(lm(ABUND ~ 1 + log(AREA) + YR.ISOL + ALT + log(LDIST) + as.factor(GRAZE), dat

##
## Call:
## lm(formula = ABUND ~ 1 + log(AREA) + YR.ISOL + ALT + log(LDIST) +
##     as.factor(GRAZE), data = birds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.80  -2.78  -0.37   2.78  11.21
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        36.7039   113.9498    0.32   0.7488
## log(AREA)           2.9642     0.6454    4.59  3.3e-05 ***
## YR.ISOL            -0.0125     0.0574   -0.22   0.8286
## ALT                 0.0103     0.0234    0.44   0.6618
## log(LDIST)          0.3972     0.8195    0.48   0.6302
## as.factor(GRAZE)2   0.3897     3.0117    0.13   0.8976
## as.factor(GRAZE)3  -0.0494     2.7716   -0.02   0.9859
## as.factor(GRAZE)4  -1.3218     3.1080   -0.43   0.6726
## as.factor(GRAZE)5 -12.5640     4.6693   -2.69   0.0098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.04 on 47 degrees of freedom
## Multiple R-squared:  0.729,Adjusted R-squared:  0.683
## F-statistic: 15.8 on 8 and 47 DF,  p-value: 5.12e-11
```

Suggests area as a strong and clear effect. Very strong grazing does have a negative effect. There is no clear statistical support for altitude and year of isolation.

Simple regressions of these two predictors show significant effects though:

```
summary(lm(ABUND ~ 1 + YR.ISOL , data = birds))

##
## Call:
## lm(formula = ABUND ~ 1 + YR.ISOL, data = birds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.835  -6.113   0.506   5.831  22.780
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -392.3208    96.2143   -4.08  0.00015 ***
## YR.ISOL        0.2112     0.0493    4.28  7.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.36 on 54 degrees of freedom
## Multiple R-squared:  0.253,Adjusted R-squared:  0.24
## F-statistic: 18.3 on 1 and 54 DF,  p-value: 7.68e-05

summary(lm(ABUND ~ 1 + ALT, data = birds))

##
## Call:
## lm(formula = ABUND ~ 1 + ALT, data = birds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.023  -7.562   0.006   8.573  20.683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5983     4.7209    1.19   0.2409
## ALT           0.0952     0.0310    3.07   0.0033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.99 on 54 degrees of freedom
## Multiple R-squared:  0.149,Adjusted R-squared:  0.133
## F-statistic: 9.45 on 1 and 54 DF,  p-value: 0.00332
```

That's probably due to their correlation with grazing pressure:

```
summary(lm(YR.ISOL ~ 1 + as.factor(GRAZE) , data = birds))

##
## Call:
## lm(formula = YR.ISOL ~ 1 + as.factor(GRAZE), data = birds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.67  -4.86   4.50   8.33  41.62
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1962.62       4.31  454.99  < 2e-16 ***
## as.factor(GRAZE)2     4.76       6.99    0.68     0.50
## as.factor(GRAZE)3    -8.95       5.89   -1.52     0.14
## as.factor(GRAZE)4     2.24       7.29    0.31     0.76
## as.factor(GRAZE)5   -49.23       6.10   -8.07  1.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.6 on 51 degrees of freedom
## Multiple R-squared:  0.657,Adjusted R-squared:  0.63
## F-statistic: 24.4 on 4 and 51 DF,  p-value: 2.46e-11

summary(lm(ALT ~ 1 + as.factor(GRAZE), data = birds))

##
## Call:
## lm(formula = ALT ~ 1 + as.factor(GRAZE), data = birds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91.92 -22.94   0.58  28.08  98.08
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         161.92      10.98   14.74   <2e-16 ***
## as.factor(GRAZE)2     7.45      17.80    0.42   0.6772
## as.factor(GRAZE)3   -15.92      15.01   -1.06   0.2937
## as.factor(GRAZE)4    -5.49      18.57   -0.30   0.7685
## as.factor(GRAZE)5   -50.77      15.53   -3.27   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.6 on 51 degrees of freedom
## Multiple R-squared:  0.232,Adjusted R-squared:  0.172
## F-statistic: 3.86 on 4 and 51 DF,  p-value: 0.00818
```

so they correlate with abundance, but the correlation is likely driven by a direct effect of grazing.