

# How to build and interpret regression models

Marie Laure Delignette-Muller - VetAgro Sup - LBBE

2026-05-06 - *distributed under license CC BY-NC-ND*



# Introduction

## Definition of few terms we will use for regression

**The dependent variable = the outcome**

**Independent variables = input variables = regressors = predictors = explanatory variables**

- ▶ **Independent variable(s) of interest** (e.g. the treatment in a clinical trial)
- ▶ **Covariates = confounding variables or factors** = independent variables that may influence the outcome but are not of direct interest (the term factor is used for categorial variables)

## Definition of few terms we will use for risk assessment

**Risk factor = risk determinant** = something that increases the chance of developing a disease

*Definition found on <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/risk-factor>.*

Implicit **causal interpretation**: alteration of the factor  $\Rightarrow$  alteration of the outcome

**Potential leverage to reduce the risk if it is controllable.**

$\neq$

**Risk marker**

There is an **association between the marker and the outcome**, but an alteration of the marker will not necessarily affect the outcome.

## Main preliminary questions to build a model

- ▶ **What is the outcome ?** The choice of the type of model depends on the nature of the data.
- ▶ **What are the relevant input variables?** We should include **the main ones** (to avoid confusion bias) but **not too many** (to avoid a too strong uncertainty on coefficients that would make them useless), and **limiting the collinearity** between input variables.
- ▶ What is the **expected relationship between each input and the outcome?** Is the linearity assumption reasonable for quantitative input variables?
- ▶ What are the **potential interactions?** Which are the inputs that may have an interaction effect on the outcome?
- ▶ What is the **purpose of modeling: explanatory or predictive?** To identify **risk factors** or **risk markers?**

# Rationale for our presentation plan

1. **Description of the two most popular models:** their formulation, their hypotheses, the interpretation of their coefficients.
2. **Selection of input variables:** why it is necessary, the different strategies proposed and used for the selection of input variables.

Description of the two most popular models

## The linear model

# The linear model: formalization and interpretation

One **continuous outcome**  $Y$  and one or more **continuous** and/or **categorical input variables** coded by  $X_k$ ,

Each categorical variable with  $p$  categories is associated to  $p - 1$  dummy variables  $X_k$  coding for the membership of each observation to the  $p$  groups except the reference one.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \dots + \epsilon_i$$

with  $\epsilon_i \sim N(0, \sigma)$

Deterministic part: linear link

Stochastic part : Gaussian model

**Interpretation of the regression coefficients:**

- ▶ For **continuous** inputs:  $\beta_k$  estimates the **change in the outcome** corresponding to a **unit change in the input**
- ▶ For **categorical** inputs:  $\beta_k$  estimates the **difference of the mean** in group  $k$  to the mean in the reference group

## The linear model after log transformation of the outcome

$$\ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \dots + \epsilon_i$$

$$\Leftrightarrow Y_i = e^{\beta_0} \times e^{\beta_1 X_{1i}} \times e^{\beta_2 X_{2i}} \times \dots \times e^{\beta_k X_{ki}} \times \dots \times e^{\epsilon_i}$$

### Interpretation of the regression coefficients:

- ▶ For **continuous** inputs:  $e^{\beta_k}$  can be traduced as a **multiplicative effect on the outcome** corresponding to a **unit change in the input**
- ▶ For **categorical** inputs:  $e^{\beta_k}$  can be traduced as a **multiplicative effect** in group  $k$  by comparison to the reference group

## Go back to our tick example

```
dtot <- read.table("DATA/Milne1950.txt", header = TRUE)
str(dtot)
```

```
## 'data.frame':    100 obs. of  3 variables:
## $ rel_hum      : int  0 50 70 85 95 0 50 70 85 95 ...
## $ surv_time    : int  7 7 22 15 38 9 9 23 22 48 ...
## $ temperature: int  5 5 5 5 5 5 5 5 5 5 ...
```

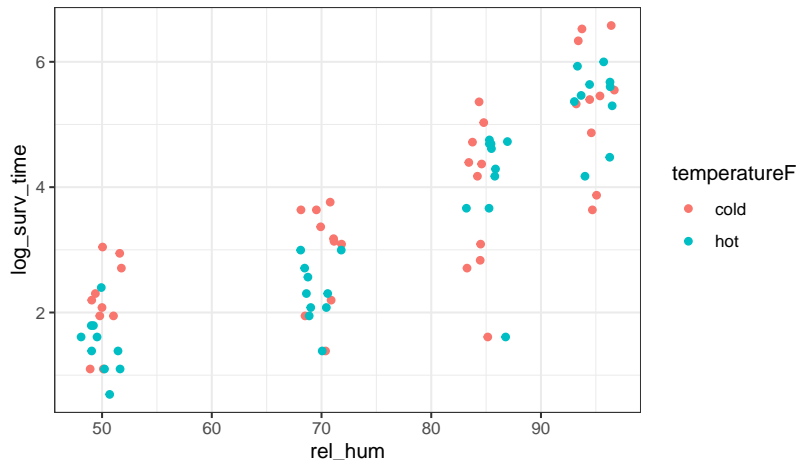
```
# replacement of 0% humidity by 10%
# as in the paper Wongnak et al. 2022
dtot$rel_hum[dtot$rel_hum == 0] <- 10

# add of the log transformed survival time
dtot$log_surv_time <- log(dtot$surv_time)
dtot$temperatureF <- as.factor(ifelse(dtot$temperature < 15,
                                     "cold", "hot"))

# Exclusion of the driest condition
dhum <- subset(dtot, rel_hum > 10)
```

## Plot of data

```
ggplot(data = dhum, aes(x = rel_hum, y = log_surv_time,  
col = temperatureF)) + geom_jitter(width = 2, height = 0)
```



## Fit of a model with the relative humidity as quantitative and the temperature as a categorical variable

```
(mancova <- lm(log_surv_time ~ rel_hum + temperatureF,  
               data = dhum))
```

```
##  
## Call:  
## lm(formula = log_surv_time ~ rel_hum + temperatureF, data = dhum)  
##  
## Coefficients:  
##      (Intercept)          rel_hum  temperatureFhot  
##      -2.1950          0.0768          -0.2452
```

```
# coefficients traduced in multiplicative factors  
exp(coef(mancova)[2:3])
```

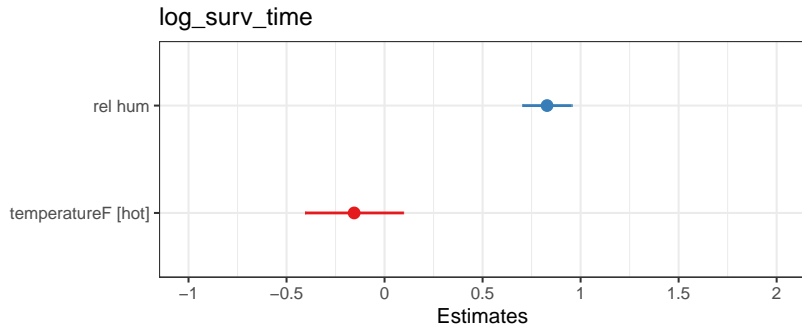
```
##          rel_hum  temperatureFhot  
##          1.080          0.783
```

## Plot of the coefficients as additive effects on $Y$ log scale

To use the same scale for all the coefficients, the  $\beta_k$  associated to continuous inputs may be multiplied by  $2 \times SD(X_k)$  as below.

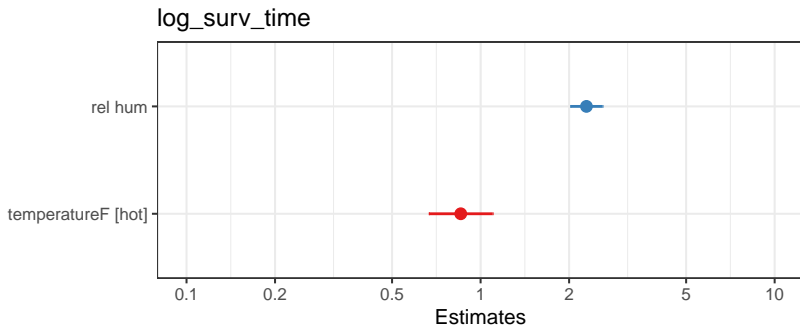
*Interpretation: outcome change for a change of 2 standard deviations of the input.*

```
plot_model(mancova, type = "std2")
```



# Plot of the coefficients as multiplicative effects on $Y$ raw scale

```
plot_model(mancova, type = "std2", transform = "exp")
```



## An example with many continuous and categorical input variables

Pankova *et al.* 2018. **Early weight gain after stopping smoking: a predictor of overall large weight gain?** A single-site retrospective cohort study. *BMJ open*, 8(12), e023987.

*A study based on 1050 patients who stopped smoking with a linear model linking a continuous outcome (relative change in weight 1 year after smoking cessation) with many various input variables.*

# Try to interpret the reported results



**Table 5** Multiple linear model for relative change in weight 1 year after smoking cessation

| Personal characteristics                  | Multiple model (n=765)   |          | Model with multiple imputation (n=772) |          |
|---|--------------------------|----------|--|----------|
|   | Beta (95% CI)            | P values | Beta (95% CI)                          | P values |
| Weight change in third month (%)          | 0.134 (-0.037 to 0.304)  | 0.124    | 0.141 (-0.027 to 0.31)                 | 0.101    |
| Female                                    | 0.804 (0.039 to 1.57)    | 0.039    | 0.781 (0.018 to 1.544)                 | 0.045    |
| Age at baseline visit (years)             | -0.005 (-0.034 to 0.024) | 0.736    | -0.006 (-0.034 to 0.023)               | 0.682    |
| BMI (kg/m <sup>2</sup> )                  | -0.209 (-0.3 to -0.118)  | 0.000    | -0.202 (-0.292 to -0.111)              | <0.001   |
| BDI score                                 |                          |          | 0.021 (-0.037 to 0.078)                | 0.467    |
| FTCD score                                | 0.004 (-0.211 to 0.218)  | 0.974    | 0 (-0.214 to 0.214)                    | 1.000    |
| Cigarettes per day                        | 0.033 (-0.018 to 0.084)  | 0.207    | 0.03 (-0.021 to 0.081)                 | 0.243    |
| Age at regular smoking initiation (years) | 0.05 (-0.046 to 0.145)   | 0.307    | 0.045 (-0.05 to 0.14)                  | 0.356    |
| Bupropion*                                | 0.011 (-1.337 to 1.359)  | 0.987    | -0.049 (-1.406 to 1.309)               | 0.944    |
| Varenicline*                              | -0.384 (-1.267 to 0.5)   | 0.395    | -0.349 (-1.225 to 0.527)               | 0.435    |
| Nicotine replacement therapy*             |                          | 0.010    | -1.068 (-1.879 to -0.257)              | 0.010    |
| Physical activity                         |                          |          |  |          |
| Regularly (more times weekly)             | -1.015 (-2.072 to 0.042) | 0.060    | -0.957 (-2.01 to 0.095)                | 0.075    |
| Weekly                                    | -0.137 (-1.185 to 0.911) | 0.798    | -0.1 (-1.146 to 0.946)                 | 0.851    |
| Irregularly                               | 0.402 (-0.657 to 1.462)  | 0.457    | 0.466 (-0.586 to 1.519)                | 0.385    |
| Never                                     | ref.                     |          | ref.                                   |          |
| Intercept                                 | 8.411 (4.993 to 11.828)  | <0.001   | 8.208 (4.795 to 11.621)                | <0.001   |

\*Using of specified pharmacotherapy (in monotherapy or in combination with other therapy).

BDI, Beck Depression Inventory; BMI, body mass index; FTCD, Fagerström Test of Cigarette Dependence.

## Extracts from the abstract

### Results

The regression coefficient per 1% rise in the first 3 months was +0.13% (95% CI -0.04% to 0.30%). In addition, lower body mass index was associated with greater weight gain, while using nicotine replacement therapy was associated with less weight gain at 1-year follow-up.

### Conclusions

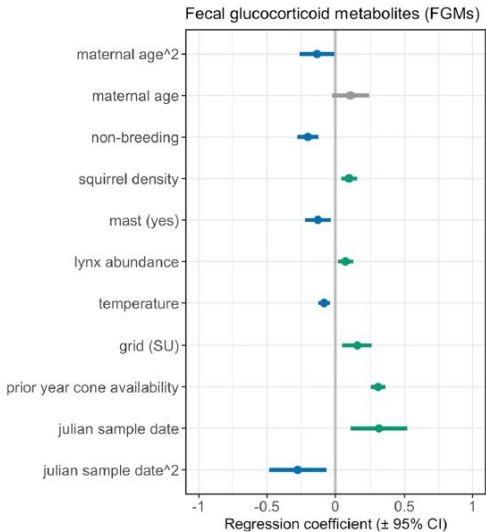
People who stop smoking and gain a larger amount of weight early after quitting are not more likely to gain excessively at 1 year.

What type of plot could help us to interpret those estimations ?

## Another similar example presented using a forest plot

Petrullo *et al.* 2022. The glucocorticoid response to environmental change is not specific to agents of natural selection in wild red squirrels. *Hormones and Behavior*, 146, 105262.

## Another similar example presented using a forest plot - Figure 5



## Another similar example presented using a forest plot - Figure 5 caption

**Fig. 5.** Female red squirrels exhibit a general, rather than specific, glucocorticoid response to environmental change. Forest plot of model estimates and associated confidence intervals corresponding to linear mixed-effects model testing the effects of the same potential ecological agents of selection from [Fig. 4](#) on concentrations of red squirrel fecal glucocorticoid metabolites (FGMs). The dataset contained 1298 FGM measures from 165 females across 7 years, and the model included maternal ID as a random factor (not shown, explained variance in FGM concentrations 0.03). Continuous fixed variables were standardized to a mean of zero and unit variance, and FGMs were log-transformed to achieve residual normality. We controlled for potential effects of intrinsic factors (i.e., maternal (linear and quadratic) age and reproductive status (breeding or non-breeding) on FGM concentrations. Squirrel density, mast (yes/no), and temperature were assessed the same as in selection models. We included lynx abundance as a continuous variable rather than a categorical variable of lynx-hare cycle as we expected squirrels to exhibit endocrine responses to the presence of lynx regardless of the abundance of hares. Nonsignificant effects are shown in gray; positive effects denoted in green, and negative effects in blue.

The logistic model for a binary outcome

# The logistic model for a binary outcome

One **binary outcome**  $Z$  and one or more **continuous** and/or **categorical input variables** coded by  $X_k$ ,

Stochastic part:  $Z_i \sim \text{Bernoulli}(p_i)$

Deterministic part:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \dots$$

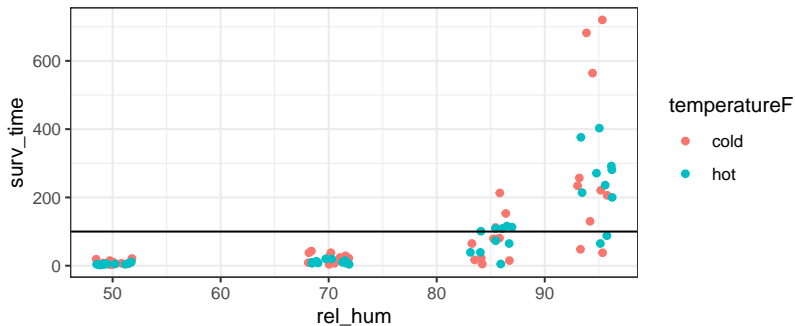
$$\Leftrightarrow \frac{p_i}{1-p_i} = \text{odd}_i = e^{\beta_0} \times e^{\beta_1 X_{1i}} \times e^{\beta_2 X_{2i}} \times \dots \times e^{\beta_k X_{ki}} \times \dots$$

**Interpretation of the regression coefficients:**

- ▶ For **continuous** inputs:  $e^{\beta_k}$  can be traduced as a **multiplicative effect on the odd** = odds ratio (OR) corresponding to a **unit change in the input**  $X_k$
- ▶ For **categorical** inputs:  $e^{\beta_k}$  can be traduced as the **odds ratio (OR) for group  $k$  compared to the reference group**

Take the example of tick survival by studying the survival or not after 100 days

```
ggplot(data = dhum, aes(x = rel_hum, y = surv_time,  
col = temperatureF)) + geom_jitter(width = 2, height = 0) +  
geom_hline(yintercept = 100)
```



## Fit of a GLM (logistic regression)

```
dhum$survabove100days <- as.factor(dhum$surv_time > 100)
(mlogis <- glm(survabove100days ~ rel_hum + temperatureF,
               family = "binomial", data = dhum))
```

```
##
```

```
## Call:  glm(formula = survabove100days ~ rel_hum + temperatureF, fami
##      data = dhum)
```

```
##
```

```
## Coefficients:
```

```
##      (Intercept)          rel_hum  temperatureFhot
##      -19.831          0.223          0.520
```

```
##
```

```
## Degrees of Freedom: 79 Total (i.e. Null);  77 Residual
```

```
## Null Deviance:          97.7
```

```
## Residual Deviance: 47.5  AIC: 53.5
```

## Estimated coefficients

```
# Coefficients
```

```
summary(mlogis)$coefficients
```

| ##                 | Estimate | Std. Error | z value | Pr(> z ) |
|--------------------|----------|------------|---------|----------|
| ## (Intercept)     | -19.831  | 5.1383     | -3.859  | 0.000114 |
| ## rel_hum         | 0.223    | 0.0576     | 3.866   | 0.000110 |
| ## temperatureFhot | 0.520    | 0.7280     | 0.715   | 0.474846 |

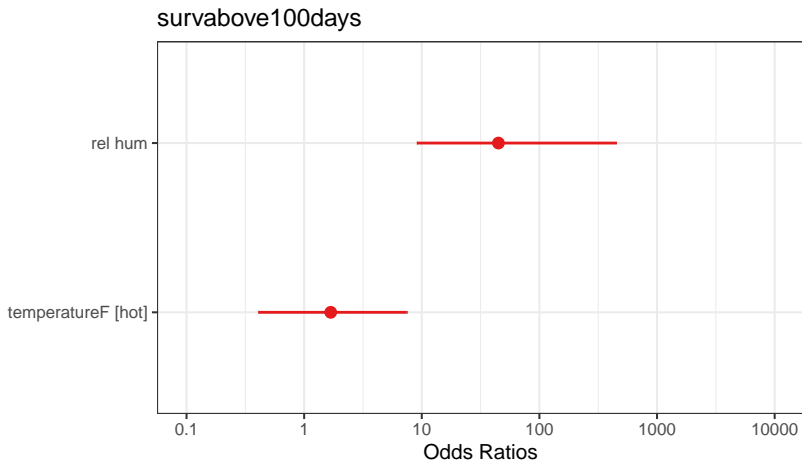
```
# Effects traduced in odds ratios
```

```
exp(summary(mlogis)$coefficients[2:3])
```

```
## [1] 1.25 1.68
```

## Plot of the coefficients after $1/\text{multiplication by } 2 \times SD(X_k)$ and $2/\text{exponential transformation}$

```
plot_model(mlogis, type = "std2")
```



# Now look at some logistic regression results in papers

## **Risk factors associated to leptospirosis**

Dung *et al.* 2022. A case–control study of agricultural and behavioral factors associated with leptospirosis in Vietnam. *BMC Infectious Diseases*, 22(1), 1-8.

T2 and 4

## **Risk factors associated to STEC positivity of farms**

Henry *et al.* 2017. British *Escherichia coli* O157 in Cattle Study (BECS): to determine the prevalence of *E. coli* O157 in herds with cattle destined for the food chain. *Epidemiology & Infection*, 145(15), 3168-3179. T2

Patterson *et al.* 2022. Risk factors of Shiga toxin-producing *Escherichia coli* in livestock raised on diversified small-scale farms in California. *Epidemiology & Infection*, 150. T4

## Dung et al. 2022 - risk factors associated to leptospirosis

Binary outcome = case of human leptospirosis or control

**Table 2** Multiple logistic regression analysis of occupations and agricultural practices

| Variable       | OR (95% CI)        | p     |
|----------------|--------------------|-------|
| Cultivating    | 2.83 (1.38–5.79)   | 0.004 |
| Animal farming | 8.26 (2.24–30.52)  | 0.002 |
| Pig owners     | 10.48 (5.05–21.73) | 0.000 |
| Cat owners     | 2.62 (1.49–4.61)   | 0.001 |

## Dung *et al.* 2022 - risk factors associated to leptospirosis

**Table 4** Multiple logistic regression analysis of behavioral factors

| Variable   | OR (95% CI)      | p     |
|--|------------------|-------|
| Hand washing after using toilet                            | 0.39 (0.23–0.68) | 0.001 |
| Hand washing after farming / gardening                     | 0.57 (0.38–0.86) | 0.007 |
| Bathing after farming, gardening, cattle / poultry contact | 0.33 (0.19–0.58) | 0.000 |
| Drinking unboiled water                                    | 1.72 (1.14–2.59) | 0.010 |

## Results

504 participants (252 cases and 252 controls) were identified. Cultivating (OR 2.83, CI 1.38–5.79), animal farming (OR 8.26, CI 2.24–30.52), pig owners (OR 10.48, CI 5.05–21.73), cat owners (OR 2.62, CI 1.49–4.61) and drinking unboiled water (OR 1.72, CI 1.14–2.59,  $p = 0.010$ ) were significantly associated with human leptospirosis in Vietnam. Hand washing after farming/ gardening (OR 0.57, CI 0.38–0.86,  $p = 0.007$ ) and bathing after farming, gardening, contact with cattle and poultry (OR 0.33, CI 0.19–0.58,  $p = 0.000$ ) were determined as protective factors for this disease.

# Henry *et al.* 2017 - risk factors associated to *E. coli* O157 positivity of farms

Binary outcome = farm classified as positive or negative

**Table 2** Final multivariable logistic model for Outcome 1<sup>a</sup> for the Scotland data

| Variable                                  | Value  | Coefficient | OR<br>[95% CI]            | <i>p</i> -value |
|---|--------|-------------|---------------------------|-----------------|
| Intercept                                 |        | -1.04       |                           | 0.040           |
| <i>breeding females brought on (BFBO)</i> | No     | Baseline    | 1.00                      |                 |
|   | Yes    | 1.71        | <b>5.52</b> [1.81–16.89]  | <b>0.003</b>    |
| <i>cattle 12–30 months</i>                |        | 0.009       | <b>1.009</b> [1.001–1.02] | <b>0.028</b>    |
| <i>bought other livestock</i>             | No     | Baseline    | 1.00                      |                 |
|   | Yes    | -1.45       | <b>0.23</b> [0.07–0.75]   | <b>0.015</b>    |
| <i>season</i>                             | Autumn | Baseline    | 1.00                      |                 |
|   | Winter | -0.53       | 0.59 [0.16–2.20]          | 0.433           |
|   | Spring | -2.33       | <b>0.10</b> [0.02–0.56]   | <b>0.009</b>    |
|   | Summer | -0.45       | 0.64 [0.18–2.30]          | 0.493           |

OR Odds ratio; CI Confidence Interval

Significant ( $p \leq 0.05$ ) OR and *p*-values are highlighted in bold text

McFadden's pseudo  $R^2 = 0.21$

<sup>a</sup> Farm classified as positive for *E. coli* O157

It is explained elsewhere in the paper that “cattle 12-30 months” is a quantitative input corresponding to the total number of cattle in this age class.

## Henry *et al.* 2017 - extracts from the abstract

### **Results**

In the England & Wales survey only one variable, herd size, was associated with the outcome farm-level *E. coli* O157 positive status. The odds increased for each additional animal in the herd. In the Scotland survey, as well as a measure of herd size (the number of cattle aged 12–30 months), having brought breeding females on to the farm in the last year also increased the odds, whereas farms sampled in spring were less likely to be positive compared to those sampled in autumn.

# Patterson *et al.* 2022 - risk factors associated to STEC positivity of farms

Binary outcome = farm classified as positive or negative

**Table 4.** Association between risk factors and the presence of Shiga toxin-producing *Escherichia coli* in faecal samples ( $n=502$ ) collected from 14<sup>3</sup> diversified small-scale farms in California between May 2015 and June 2016, as demonstrated by a generalised linear mixed model, with farm as a random effect

| Variable                                       | Variable category | Estimate         | OR   | 95% CI     | P-value |
|--|-------------------|------------------|------|------------|---------|
| Intercept                                      |                   | -2.57            |      |            | 0.001*  |
| Daily maximum temperature °C                   | <i>numeric</i>    | -0.06            | 0.95 | 0.91-0.98  | 0.003*  |
| Sample source species                          | <i>Swine</i>      | <i>Reference</i> |      |            |         |
|  | Goats             | 0.97             | 2.64 | 0.90-7.70  | 0.076   |
|  | Sheep             | 1.67             | 5.29 | 1.80-15.51 | 0.002*  |
|  | Cattle            | 1.53             | 4.61 | 1.64-12.96 | 0.004*  |
| Multiple species shared a barn                 | <i>No</i>         | <i>Reference</i> |      |            |         |
|  | Yes               | 1.83             | 6.23 | 1.84-21.15 | 0.003*  |
| Livestock were allowed contact with wild areas | <i>No</i>         | <i>Reference</i> |      |            |         |
|  | Yes               | 1.29             | 3.63 | 1.37-9.62  | 0.009*  |
| Number of years in operation                   | <i>6-30 years</i> | <i>Reference</i> |      |            |         |
|  | 1-5 years         | -0.98            | 0.38 | 0.13-1.11  | 0.076   |

<sup>a</sup>Two of the 16 participant surveys were not completed.

\*Indicates statistical significance with a  $P$ -value < 0.05.

## Results

Significant variables in the mixed-effect logistic regression model included daily maximum temperature (OR 0.95; CI95% 0.91–0.98), livestock sample source (cattle (OR 4.61; CI95% 1.64–12.96) and sheep (OR 5.29; CI95% 1.80–15.51)), multiple species sharing the same barn (OR 6.23; CI95% 1.84–21.15) and livestock having contact with wild areas (OR 3.63; CI95% 1.37–9.62).

Can we say that having multiple species sharing a same barn multiplies by 6.23 the risk for a farm to be STEC positive ?

## Remember, . . . , we should not confound odds ratios and risk ratios

Coefficients estimated by **logistic regression** are **odds ratios** between two conditions  $k$  and 0 (reference):  $OR = \frac{\frac{p_k}{1-p_k}}{\frac{p_0}{1-p_0}}$

A more intuitive statistics is risk ratio:  $RR = \frac{p_k}{p_0}$

Relation between OR and RR:  $OR = RR \times \frac{1-p_0}{1-p_k}$

OR is a **good approximation** of RR **only if**  $p_k$  and  $p_0$  are close to 0 (and/or  $p_k$  and  $p_0$  close together so  $RR$  close to one).

If this is not the case **the OR is exaggerating the RR.**

(If  $p_k > p_0$ ,  $RR > 1$  and  $1 - p_0 > 1 - p_k$  so  $\frac{1-p_0}{1-p_k} > 1$ )

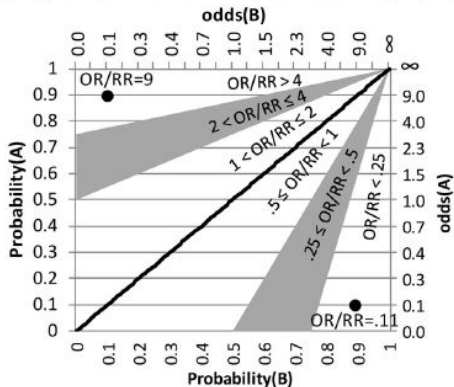
(If  $p_k < p_0$ ,  $RR < 1$  and  $1 - p_0 < 1 - p_k$  so  $\frac{1-p_0}{1-p_k} < 1$ )

**BE CAREFUL ! Interpretation of ORs as RRs is a very common error.**

# A reference that shows the impact of this common error

Sheldrick *et al.* 2017. Math matters: how misinterpretation of odds ratios and risk ratios may influence conclusions. *Academic Pediatrics*, 17(1), 1-3.

## C Degree to which OR overestimates RR [OR/RR]



**Figure 1.** Relationship between RR and OR. (A)  $RR = \text{probability A} / \text{probability B}$ . (B) OR. (C) Degree to which OR overestimates RR (OR/RR). RR indicates relative risk; OR, odds ratio.

## Selection of input variables

But why do we need to limit the number of input variables?

# Let us look at in-field data collected on Nile monitors

An example from Ciliberti *et al.* 2011.

Ciliberti *et al.* 2011. The Nile monitor (*Varanus niloticus*; Squamata: Varanidae) as a sentinel species for lead and cadmium contamination in sub-Saharan wetlands. *Science of the Total Environment*, 409(22), 4735-4745.

Nile monitors (large African lizards) were captured in different areas of Africa. The lead content in their kidneys was determined and different morphometric parameters were measured on these animals. We wish to build a **model describing the decimal logarithm of the lead content** ( $\log_{10} \text{Pb}$ ) as a function of the variables **sex** (*sex*), the **area of capture** (*site*), chosen to represent gradient of contamination level), the **fat somatic index** (FS), the **snout-vent length** (in  $\log_{10} \log_{10} \text{SVL}$ ) and the **body mass** (in  $\log_{10} \log_{10} \text{BM}$ ).

Fit the corresponding linear model (neglecting potential interactions here for sake of simplicity) and carefully look at the results.

## Importation of the data

```
dNM <- read.table("DATA/Nilemonitor.txt", header = TRUE,
                  stringsAsFactors = TRUE)
str(dNM)
```

```
## 'data.frame':    71 obs. of  6 variables:
## $ sex          : Factor w/ 2 levels "female","male": 2 2 2 2 2 2
## $ site         : Factor w/ 4 levels "dif","fla","nia",...: 4 4 4 4
## $ log10BM      : num  -0.108 0.373 0.25 0.334 0.491 ...
## $ log10SVL     : num  1.51 1.66 1.65 1.67 1.73 ...
## $ FS           : num  0.0526 0.0441 0.0826 0.0532 0.0529 ...
## $ log10Pb      : num  2.04 1.3 2.14 1.43 2.08 ...
```

```
xtabs(data = dNM, ~ sex + site)
```

```
##           site
## sex      dif fla nia nio
## female  17  5  3  5
## male    15  9  4 13
```

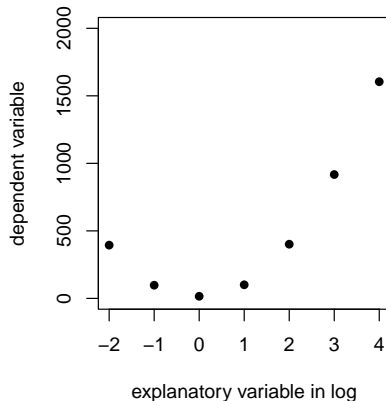
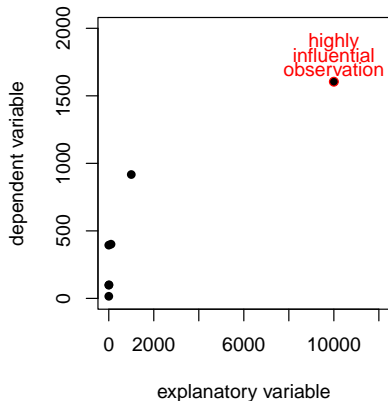
## Fit of the model

```
mNM <- lm(log10Pb ~ sex + site + log10BM + log10SVL + FS, data = dNM)
summary(mNM)$coefficients
```

| ##             | Estimate | Std. Error | t value | Pr(> t ) |
|----------------|----------|------------|---------|----------|
| ## (Intercept) | 2.5724   | 3.2317     | 0.7960  | 4.29e-01 |
| ## sexmale     | -0.2491  | 0.0651     | -3.8256 | 3.02e-04 |
| ## sitefla     | -0.6045  | 0.0698     | -8.6573 | 2.54e-12 |
| ## sitenia     | -0.1545  | 0.0988     | -1.5639 | 1.23e-01 |
| ## sitenio     | -0.5859  | 0.0652     | -8.9875 | 6.81e-13 |
| ## log10BM     | -0.1634  | 0.6681     | -0.2445 | 8.08e-01 |
| ## log10SVL    | 0.0926   | 2.0242     | 0.0457  | 9.64e-01 |
| ## FS          | 0.1603   | 1.5182     | 0.1056  | 9.16e-01 |

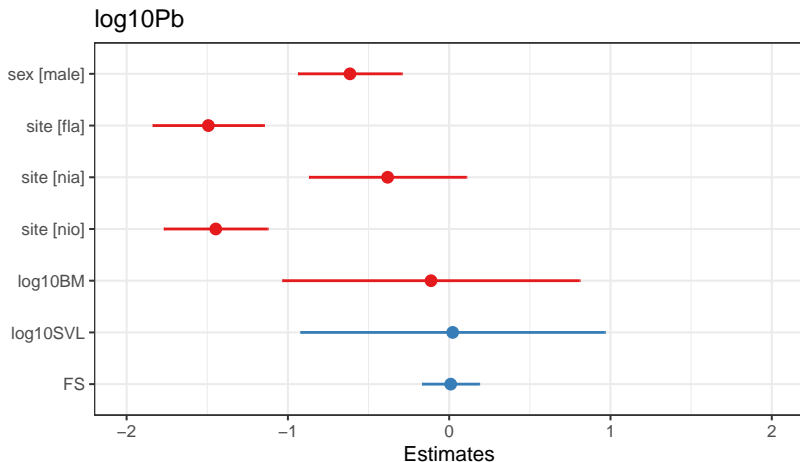
## By the way, why do you think explanatory variables are sometimes transformed into log ?

- ▶ To linearise a relationship, and/or
- ▶ to reach a better design, by preventing extreme values of  $x$ , associated to observations that could be highly influential.



# Forest plot of the coefficients

```
plot_model(mNM, type = "std2")
```

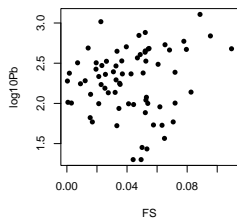
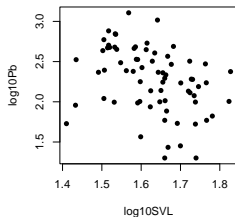
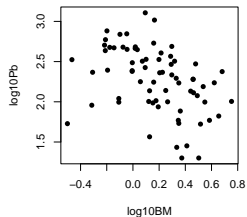


Carefully look at the estimated coefficients for  $\log_{10}BM$ ,  $\log_{10}SVL$  and  $FS$ .

- ▶ Do their sign correspond to what is expected ?
- ▶ To answer look at the bivariate correlations between each of those three model inputs and  $\log_{10}Pb$ .
- ▶ To find an explanation look at the pairwise correlations between those inputs.

## Bivariate correlations between $\log_{10}Pb$ and the three inputs.

```
par(mfrow = c(2,2)); par(mar = c(4, 4, 1, 1))  
plot(log10Pb ~ log10BM, data = dNM, pch = 16)  
plot(log10Pb ~ log10SVL, data = dNM, pch = 16)  
plot(log10Pb ~ FS, data = dNM, pch = 16)
```



## What are the expected signs of the three coefficients ?

```
cor(dNM$log10Pb, dNM$log10BM)
```

```
## [1] -0.415
```

```
cor(dNM$log10Pb, dNM$log10SVL)
```

```
## [1] -0.362
```

```
cor(dNM$log10Pb, dNM$FS)
```

```
## [1] 0.131
```

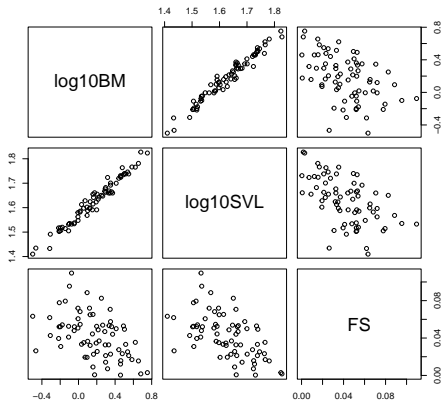
- ▶ negative for log10BM.
- ▶ negative for log10SVL. *Why isn't it negative ?*
- ▶ positive for FS.

```
coef(mNM)[6:8]
```

```
## log10BM log10SVL FS  
## -0.1634 0.0926 0.1603
```

## Pairwise correlations between those inputs

```
par(mar = c(1, 1, 1, 1)); pairs(dNM[, 3:5])
```



Collinearity between input variables (here log10BM and log10SVL) not only increases the uncertainty of the estimates, but also makes their interpretation meaningless.

## Conclusion on this rather simple example (in comparison to realistic examples in epidemiology)

- ▶ It is important to **prevent introducing collinear inputs** in a model.
- ▶ More generally we **need a strategy** to choose inputs to include in a model.

Comparison of models using information criteria

## Comparison of models using information criteria

- ▶ Various strategies are proposed, often based on a predictive perspective.
- ▶ Some are based on **hypothesis tests for comparing nested models**.
- ▶ Some are based on **information criteria** (the most popular being the Akaike's one : **AIC**). Based on their AIC values, **any number of models can be ranked, whatever they are nested or not**.

## Likelihood and deviance

The **likelihood** is generally expressed in log:

$$\logLik = \ln(\Pr(y | \theta))$$

with  $\theta$  the parameters of the model.

The **deviance** is twice the difference in loglikelihood between the fitted model and the saturated model (perfect model exactly describing the data)

$$Dev = -2\logLik(model) + 2\logLik(saturated\_model)$$

$$Dev = -2\logLik(model)$$

if we consider the likelihood of the saturated model equal to 1.

# Is the model with the smallest deviance always better ?

NO

**Including new input variables in a model always decreases the deviance,**

- ▶ but also **increases the complexity of the model,**
- ▶ and so **increases the uncertainty** on the parameter estimates,
- ▶ **decreases its robustness** to outliers,
- ▶ and so **decreases its ability to predict new data.**

So a compromise must be found: build **parsimonious models**, with just the right number of input variables to well fit the data.

## Aikake Information Criterion (AIC)

Information criteria were proposed to help us find a **balance between goodness-of-fit and complexity**, to build parsimonious models.

The most popular one is the **Akaike's information criterion (AIC)** in which the deviance is penalized by twice the number of estimated parameters  $p$  :

$$AIC = -2 \times \log Lik + 2 \times p$$

Given a set of models, the **one with the smaller AIC will be preferred**.

## Other popular information criteria

### The AIC corrected for small sample size

A correction for sample size ( $n$ ) is recommended when  $\frac{n}{p} < 40$ .

$$AICc = -2 \times \logLik + 2 \times p + \frac{2p(p+1)}{n-p-1}$$

### Bayesian Information Criterion

$$BIC = -2 \times \logLik + \ln(n) \times p$$

As its penalization for complexity is stronger, it tends to select simpler models than the AIC.

## Origin of the AIC

AIC was created to select the best models in a **predictive purpose**.

The penalization of the deviance in the AIC definition is here just to **correct for the underestimation of the error expected in prediction** with this model, as the deviance is calculated on the data used for fitting the model (and not on new data - external validation).

So the challenge of the selection of variables using the AIC is to select the smallest number of input variables that best predicts the outcome.

## Comparison of AIC values using R - *Ixodes ricinus* data

Let us take as an example the *Ixodes ricinus* data (Milne 1950) on which we introduced the linear model.

```
dtot <- read.table("DATA/Milne1950.txt", header = TRUE)
str(dtot)
```

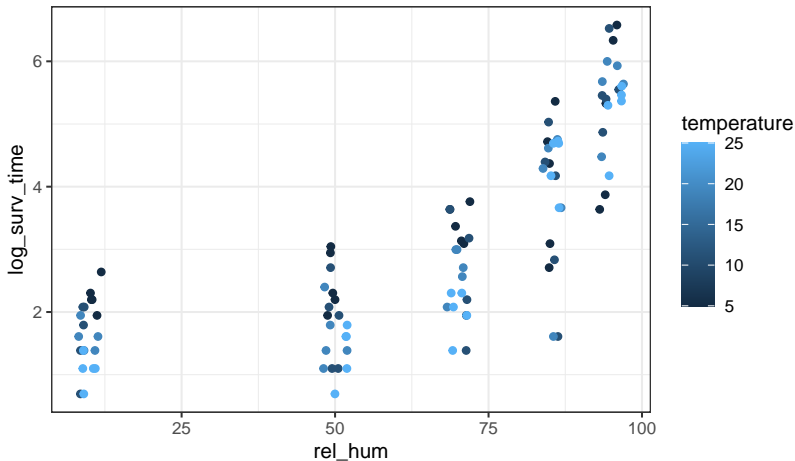
```
## 'data.frame':    100 obs. of  3 variables:
## $ rel_hum      : int  0 50 70 85 95 0 50 70 85 95 ...
## $ surv_time    : int  7 7 22 15 38 9 9 23 22 48 ...
## $ temperature: int  5 5 5 5 5 5 5 5 5 5 ...
```

```
# replacement of 0% humidity by 10%
# as in the paper Wongnak et al. 2022
dtot$rel_hum[dtot$rel_hum == 0] <- 10
```

```
# add of the log transformed survival time
dtot$log_surv_time <- log(dtot$surv_time)
```

## Plot of the data

```
ggplot(data = dtot, aes(x = rel_hum, y = log_surv_time,  
  col = temperature)) + geom_jitter(width = 2, height = 0)
```



## Comparison of several models

```
# null model
m0 <- lm(log_surv_time ~ 1, data = dtot)
# linear model
m1 <- lm(log_surv_time ~ rel_hum + temperature, data = dtot)
# quadratic model (response surface)
m2 <- lm(log_surv_time ~ rel_hum + I(rel_hum^2) +
  temperature + I(temperature^2) + rel_hum:temperature, data = d
AIC(m0, m1, m2)
```

```
##      df AIC
## m0   2 382
## m1   4 296
## m2   7 232
```

The more complex model clearly appears as the best one for the prediction of the survival rate. But is it too complex ?

## Estimates of the most complex model

```
summary(m2)$coefficients
```

| ##                     | Estimate  | Std. Error | t value | Pr(> t ) |
|------------------------|-----------|------------|---------|----------|
| ## (Intercept)         | 3.457218  | 0.506083   | 6.831   | 8.26e-10 |
| ## rel_hum             | -0.067772 | 0.011688   | -5.798  | 8.93e-08 |
| ## I(rel_hum^2)        | 0.000941  | 0.000100   | 9.403   | 3.42e-15 |
| ## temperature         | -0.113053 | 0.057605   | -1.963  | 5.27e-02 |
| ## I(temperature^2)    | 0.001338  | 0.001770   | 0.756   | 4.52e-01 |
| ## rel_hum:temperature | 0.000722  | 0.000324   | 2.228   | 2.83e-02 |

## Is it possible to compare all the possible models ?

If the number of potential predictors is low

(here 5 predictors, `rel_hum`, `temperature`, `rel_hum^2`, `temperature^2` and the interaction)

it remains possible (here  $2^5 = 32$  combinations)

but if it is large, it seems difficult

(e.g. with 10 potential predictors,  $2^{10} = 1024$  combinations).

## Stepwise algorithms to choose the best submodel

The R popular function `step()` proposes three stepwise methods to select the best model based on AIC:

- ▶ **backward elimination**: we start from the most complex considered model, and at each step we remove the predictor that best improves the fit, until the AIC cannot be reduced by removing a predictor.
- ▶ **forward selection**: we start from a minimal model (with the predictors you absolutely want to keep), and at each step we add the predictor that best improves the fit, until the AIC cannot be reduced by adding a predictor.
- ▶ **both** (sometimes called **bidirectional** or **two-way**): a combination of both algorithms, in which at each step predictors can be added or removed, until the AIC cannot be reduced by adding or removing a predictor.

## Backward elimination from model m2

```
step(m2, direction = "backward", trace = FALSE)
```

```
##  
## Call:  
## lm(formula = log_surv_time ~ rel_hum + I(rel_hum^2) + temperature +  
##     rel_hum:temperature, data = dtot)  
##  
## Coefficients:  
##           (Intercept)           rel_hum           I(rel_hum^2)  
##           3.233850           -0.067772           0.000941  
##           temperature  rel_hum:temperature  
##           -0.072927           0.000722
```

# Tracing of the backward elimination

```
step(m2, direction = "backward", trace = TRUE)
```

```
## Start: AIC=-53.5
## log_surv_time ~ rel_hum + I(rel_hum^2) + temperature + I(temperature^2) +
##   rel_hum:temperature
##
##              Df Sum of Sq  RSS   AIC
## - I(temperature^2)    1      0.3  52.3 -54.9
## <none>                    51.9 -53.5
## - rel_hum:temperature  1      2.7  54.7 -50.4
## - I(rel_hum^2)        1     48.8 100.8  10.8
##
## Step: AIC=-54.9
## log_surv_time ~ rel_hum + I(rel_hum^2) + temperature + rel_hum:temperature
##
##              Df Sum of Sq  RSS   AIC
## <none>                    52.3 -54.9
## - rel_hum:temperature  1      2.7  55.0 -51.8
## - I(rel_hum^2)        1     48.8 101.1   9.1

##
## Call:
## lm(formula = log_surv_time ~ rel_hum + I(rel_hum^2) + temperature +
##   rel_hum:temperature, data = dtot)
##
## Coefficients:
##      (Intercept)          rel_hum      I(rel_hum^2)
##          3.233850          -0.067772           0.000941
##      temperature rel_hum:temperature
##          -0.072927           0.000722
```

## Forward selection from model $m_0$ to model $m_2$

```
step(m0, scope = log_surv_time ~ rel_hum + I(rel_hum^2) +  
  temperature + I(temperature^2) + rel_hum:temperature,  
  direction = "forward", trace = FALSE)
```

```
##  
## Call:  
## lm(formula = log_surv_time ~ I(rel_hum^2) + rel_hum + temperature +  
##   rel_hum:temperature, data = dtot)  
##  
## Coefficients:  
##           (Intercept)           I(rel_hum^2)           rel_hum  
##           3.233850           0.000941           -0.067772  
##           temperature  rel_hum:temperature  
##           -0.072927           0.000722
```

On this example backward elimination and forward selection give the same result. But it is rarely the case when the number of potential predictors is large.

# Tracing the forward selection

```
step(m0, scope = log_surv_time ~ rel_hum + I(rel_hum^2) +  
  temperature + I(temperature^2) + rel_hum:temperature,  
  direction = "forward", trace = TRUE)
```

```
## Start: AIC=95.7  
## log_surv_time ~ 1  
##  
##  
##           Df Sum of Sq  RSS   AIC  
## + I(rel_hum^2)      1    179.9  75.4 -24.2  
## + rel_hum           1    146.9 108.4  12.1  
## <none>                255.3  95.7  
## + temperature      1         4.6 250.7  95.9  
## + I(temperature^2) 1         4.0 251.3  96.1  
##  
## Step: AIC=-24.2  
## log_surv_time ~ I(rel_hum^2)  
##  
##           Df Sum of Sq  RSS   AIC  
## + rel_hum           1    15.86  59.6 -45.8  
## + temperature      1         4.59  70.8 -28.5  
## + I(temperature^2) 1         4.02  71.4 -27.7  
## <none>                75.4 -24.2  
##  
## Step: AIC=-45.8  
## log_surv_time ~ I(rel_hum^2) + rel_hum  
##  
##           Df Sum of Sq  RSS   AIC  
## + temperature      1         4.59  55.0 -51.8  
## + I(temperature^2) 1         4.02  55.6 -50.8  
## <none>                59.6 -45.8  
##  
## Step: AIC=-51.8  
## log_surv_time ~ I(rel_hum^2) + rel_hum + temperature  
##  
##           Df Sum of Sq  RSS   AIC
```

## Bidirectional selection from model m0 to model m2

```
step(m0, scope = log_surv_time ~ rel_hum + I(rel_hum^2) +  
  temperature + I(temperature^2) + rel_hum:temperature,  
  direction = "both", trace = FALSE)
```

```
##
```

```
## Call:
```

```
## lm(formula = log_surv_time ~ I(rel_hum^2) + rel_hum + temperature +  
##   rel_hum:temperature, data = dtot)
```

```
##
```

```
## Coefficients:
```

```
##      (Intercept)      I(rel_hum^2)      rel_hum  
##      3.233850      0.000941      -0.067772  
##      temperature  rel_hum:temperature  
##      -0.072927      0.000722
```

The combination of both methods give the same result on this example.

# Tracing the bidirectional selection

```
step(m0, scope = log_surv_time ~ rel_hum + I(rel_hum^2) +  
  temperature + I(temperature^2) + rel_hum:temperature,  
  direction = "both", trace = TRUE)
```

```
## Start: AIC=95.7  
## log_surv_time ~ 1  
##  
##  
##           Df Sum of Sq  RSS   AIC  
## + I(rel_hum^2)      1    179.9  75.4 -24.2  
## + rel_hum           1    146.9 108.4  12.1  
## <none>                255.3  95.7  
## + temperature       1         4.6 250.7  95.9  
## + I(temperature^2)  1         4.0 251.3  96.1  
##  
## Step: AIC=-24.2  
## log_surv_time ~ I(rel_hum^2)  
##  
##           Df Sum of Sq  RSS   AIC  
## + rel_hum           1    15.9  59.6 -45.8  
## + temperature       1         4.6  70.8 -28.5  
## + I(temperature^2)  1         4.0  71.4 -27.7  
## <none>                75.4 -24.2  
## - I(rel_hum^2)       1    179.9 255.3  95.7  
##  
## Step: AIC=-45.8  
## log_surv_time ~ I(rel_hum^2) + rel_hum  
##  
##           Df Sum of Sq  RSS   AIC  
## + temperature       1         4.6  55.0 -51.8  
## + I(temperature^2)  1         4.0  55.6 -50.8  
## <none>                59.6 -45.8  
## - rel_hum           1    15.9  75.4 -24.2  
## - I(rel_hum^2)       1    48.8 108.4  12.1  
##  
## Step: AIC=-51.8
```

## Your turn to handle the `step()` function

Take the time to handle the `step()` function with the three options,

carefully looking at the outputs given when the argument `trace` is fixed to `TRUE`,

in order to be sure you well understand each algorithm.

You can use it on the `mNM` model (Nile monitor ex.)

## Comparison of backward elimination results from model mNM using AIC and BIC criteria

```
mNMAIC <- step(mNM, trace = FALSE)
coef(mNMAIC)[-1]
```

```
## sexmale sitefla sitenia sitenio log10BM
## -0.250 -0.605 -0.156 -0.586 -0.138
```

```
mNMBIC <- step(mNM, k = log(nrow(dNM)), trace = FALSE)
coef(mNMBIC)[-1]
```

```
## sexmale sitefla sitenia sitenio
## -0.273 -0.626 -0.187 -0.607
```

```
AIC(mNM, mNMAIC, mNMBIC)
```

```
##      df  AIC
## mNM   9 -10.8
## mNMAIC 7 -14.8
## mNMBIC 6 -14.7
```

## Stepwise methods based on other criteria using R

The three algorithms can be performed based on other criteria than AIC (other information criteria or tests for comparing nested models)

- ▶ It is possible to use of `step()` with the **BIC** by fixing the argument `k` to  $\ln(n)$ .
- ▶ To use **tests comparing nested models (p-values) on Gaussian linear models** one can use for example the R package `olsrr` (with `ols_step_forward_p()`, `ols_step_backward_p()` or `ols_step_both_p()` functions)
- ▶ To use the **AICc** one can use the R package `MuMIn`: its `dredge()` function uses AICc by default.

**Try to compare backward elimination from model `mNM` using AIC and BIC criteria**

## Limits of stepwise selection of input variables

- ▶ **What criterion** to choose ?
- ▶ **What algorithm** to choose (backward elimination, forward selection, both)? Even using one criterion, in the more general case the three approaches do not necessarily propose the same best model.
- ▶ Sometimes the **AIC difference between two models is rather small** (we often consider that an  $\Delta AIC < 2$  is not demonstrative). Is thus such an algorithm relevant ?
- ▶ The best model may contain non-significant coefficients.
- ▶ The best model may contain meaningless coefficients.
- ▶ The best model regarding the AIC does not necessarily respect the **conditions of use that must be carefully checked**.
- ▶ We must keep in mind that based on the AIC we will select the model that best approximates the data using a minimal number of input variables. But **is the model useful in an explanatory perspective ?**

What are the proposed strategies ?

# In “Data analysis using regression and multilevel/hierarchical models” - Gelman & Hill 2006)

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press. *more than 20600 Google Scholar citations in Nov. 2025*

1. Include **all variables that might be expected to be important** in predicting the outcome
2. Consider the **possibility of gathering some inputs** in one predictor, for example calculated as a score from several inputs.
3. Consider including **interactions** for inputs having large effects.
4. Select the predictors to remove following those rules:
  - ▶ exclude a predictor if its coefficient is non significant and has not the expected sign.
  - ▶ think hard about coefficients with significant but unexpected sign.
  - ▶ generally keep coefficient with expected sign even if non significant.
  - ▶ always keep a coefficient that is significant and with the expected sign.

*In other parts of the book the authors give various methods to check the final model, especially using predictive checking based on data simulation.*

# In “Applied logistic regression” (Hosmer & Lemeshow 2013)

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons. *more than 86400 Google Scholar citations in Nov. 2025*

1. **Bivariate analyses** - keep all variables associated with the outcome (**p-value < 0.25**)
2. **Multivariate analysis** with variables selected in step 1 **and all variables of known biological importance**. Do not recommend nor stepwise nor best subsets selection of variables: the *“analyst must be conscious that such methods can yield a biological implausible model or select irrelevant, or noise, variables”*.
3. For each coefficient, compare its values in **Steps 1 and 2**, and **eliminate predictors for which coefficients are of markedly different orders of magnitude**. Then compare the simplified model to the complete model using comparison of nested models. (*iterative process*). Try to **reintroduce in the model each variable not selected in step 1**.
4. Check the **conditions of use** (linearity for continuous inputs, appropriate categories for discrete variables)
5. Check for **interactions** among the variables in the model, adding each plausible interaction one at a time. And check the **goodness-of-fit** of the final model.

# In Bursac et al. 2008 (a highly quoted paper in **medicine**)

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, 3(1), 1-8. *4277 Google Scholar citations in Nov. 2025*

1. **Bivariate analyses** - keep all variables associated with the outcome (**p-value < 0.25**)
2. **Multivariate analysis** with variables selected in step 1.
3. **Eliminate predictors that are non significant (p-value > 0.1) and not a confounder** (assessed by checking that its elimination does not change the estimation of other coefficients by more than 15% or 20%) (*iterative process*)
4. Try to **reintroduce in the model each variable not selected in step 1** and keep it if its contribution is significant (**p-value < 0.1 or 0.15**). (*iterative process only concerning the reintroduction of variables*)

## In Harrison *et al.* 2018 (a highly quoted paper in ecology)

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., . . . & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794. 2247 Google Scholar citations in Nov. 2025

- ▶ Authors prevent the use of backward selection and hypothesis tests when the **number of input variables is large** and **recommend the ranking of competing models using AIC**. For balanced experimental designs with few inputs, they left the open question of which method to use between information criteria and tests.
- ▶ Authors recommend “**hard thinking about hypotheses**” underlying the different competing models instead of selection from all possible subsets, so starting from only a **handful of models with biological meaning and without collinear inputs**.
- ▶ They recommend the **ranking of models based on the AIC** (with correction if needed - AICc) to define a “top model set”, taking all the models with a  $\Delta AIC$  **from the best one less than 6**. and the **elimination of models that are more complex versions of nested models of others in the “top set”** as AIC is known to tends toward overly complex models.

*The final choice among the remaining models in the top set must be argued by the biologist. The authors recommend the use of data simulation (powerful but underused tool) to check the final model.*

What are the strategies used nowadays by authors to build models for identifying risk factors ?

## To have an idea, carefully look at the statistical part of some of the previous examples

Dung *et al.* 2022. A case–control study of agricultural and behavioral factors associated with leptospirosis in Vietnam. *BMC Infectious Diseases*, 22(1), 1-8.

Henry *et al.* 2017. British *Escherichia coli* O157 in Cattle Study (BECS): to determine the prevalence of *E. coli* O157 in herds with cattle destined for the food chain. *Epidemiology & Infection*, 145(15), 3168-3179.

Patterson *et al.* 2022. Risk factors of Shiga toxin-producing *Escherichia coli* in livestock raised on diversified small-scale farms in California. *Epidemiology & Infection*, 150.

- ▶ Bivariate analyses selecting input variables associated with the outcome (logistic regression with  $p - value < 0.05$  - threshold given only in the Supplementary Information)
- ▶ Multivariate analysis (logistic regression)
- ▶ Backward elimination (comparison of nested models with  $p - value < 0.05$ )

- ▶ Bivariate analyses selecting input variables (Potential Risk Factors - PFRs) associated with the outcome ( $\chi^2$  tests with  $p - value < 0.20$ )
- ▶ Multivariate analysis (logistic regression)
- ▶ Bidirectional (forward selection / backward elimination) stepwise method (comparison of nested models with  $p - value < 0.05$ )

## Patterson et al. 2022

- ▶ Elimination of some potential input variables from descriptive univariate analysis
- ▶ Bivariate analyses selecting input variables associated with the outcome ( $\chi^2$  tests with  $p - value < 0.05$  - see Table 3 of the paper)
- ▶ Multivariate analysis (logistic regression)
- ▶ Elimination of some input variables due to collinearity
- ▶ Bidirectional (two-way) stepwise method (using the AIC)
- ▶ Selection of top models among which a model is chosen

## Conclusion of this exploration ?

There seems to be  
a **great variability in used approaches**,  
and sometimes a **gap** between  
**what is recommended** and  
**what is really done.**

Model building in a risk assessment perspective

In which perspective is the model built, inference,  
prediction?

## In which perspective is the model built, inference, prediction?

- ▶ in a **predictive** perspective (the one for which information criteria were developed)? We want the model to predict the outcome (e.g. calculate clinical score to do a **pronostic**)
- ▶ in an **explanatory** perspective (**inference**)? For a better understanding of biological processes? We want to compare models based on different **competitive biological hypotheses**.

What are the perspectives for **risk management**? What are the main risk factors? (**explanatory**) How can I reduce the risk? (**prediction**, using **inputs that are the easiest to control**)

### Some references to go further

Tredennick *et al.* 2021. **A practical guide to selecting models for exploration, inference, and prediction in ecology**. *Ecology*, 102(6), e03336.

Mac Nally *et al.* 2018. **Model selection using information criteria, but is the “best” model any good?** *Journal of Applied Ecology*, 55(3), 1441-1444.

Is it easy to draw conclusions from models fitted on  
non-experimental data ?

## Is it easy to draw conclusions from models fitted on non-experimental data ?

Christenfeld *et al.* 2004. **Risk factors, confounding, and the illusion of statistical control.** Psychosomatic medicine, 66(6), 868-875.

Westfall & Yarkoni, 2016. **Statistically controlling for confounding constructs is harder than you think.** PloS one, 11(3), e0152719.

## A realistic “toy” example

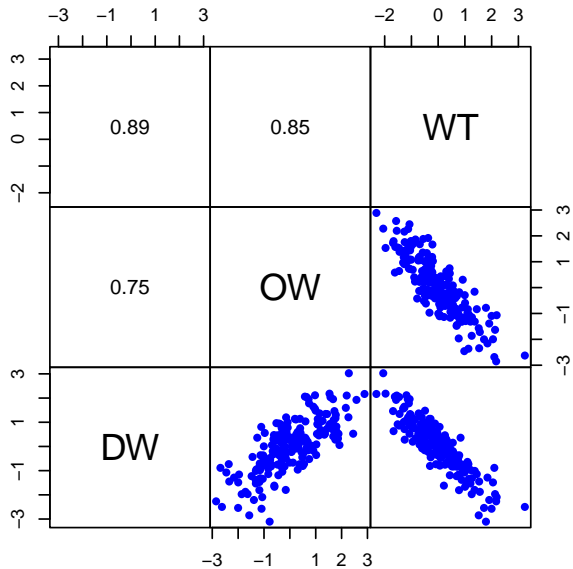
- ▶ **Dog weight** (standardized = centered and reduced):  $DW$
- ▶ **Owner’s weight** (standardized):  $OW$
- ▶ **Common daily walk time** (standardized):  $WT$

**With**  $WT \rightarrow DW$  **and**  $WT \rightarrow OW$  as causal relationships

Simulation of data from

- ▶  $WT \sim N(0, 1)$
- ▶  $OW \sim N(\alpha_{OW} + \beta_{OW} \times WT, \sigma_{OW})$
- ▶  $DW \sim N(\alpha_{DW} + \beta_{DW} \times WT, \sigma_{DW})$

## Visualization of simulated data



## Linear model: DW as a function of OW

```
mOW <- lm(DW ~ OW)
summary(mOW)$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.000514    0.0507   0.0101 9.92e-01
## OW          0.707517    0.0449  15.7572 8.64e-37
```

**Of course, if we do not take into account WT in the model, we highlight a correlation between DW and OW related to the common causal factor WT.**

## Linear model: DW as a function of OW and the confounding factor WT

```
mOWWT <- lm(DW ~ OW + WT)
summary(mOWWT)$coefficients
```

| ##             | Estimate | Std. Error | t value | Pr(> t ) |
|----------------|----------|------------|---------|----------|
| ## (Intercept) | 0.0168   | 0.0343     | 0.491   | 6.24e-01 |
| ## OW          | -0.0422  | 0.0574     | -0.735  | 4.63e-01 |
| ## WT          | -1.0584  | 0.0688     | -15.376 | 1.43e-35 |

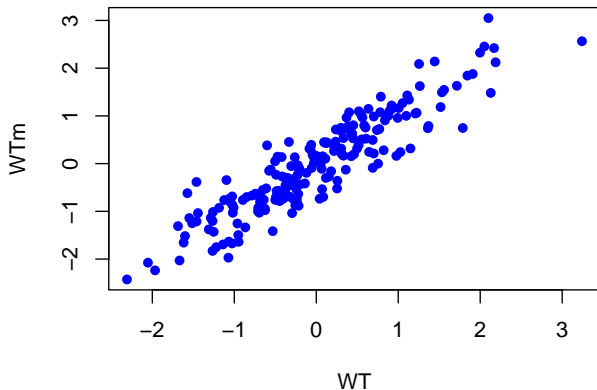
**Taking into account the confounding factor WT, as expected, we no longer show a significant effect of OW on DW.**

*So it works ?*

*Yes, but would it work in real life?*

But what if the available WT measurement is noisy  
(realistic in real life)?

$$WT_m \sim N(WT, \sigma_{WT_m})$$



## Linear model: DW as a function of OW and WTm

```
mOWWTm <- lm(DW ~ OW + WTm)
summary(mOWWTm)$coefficients
```

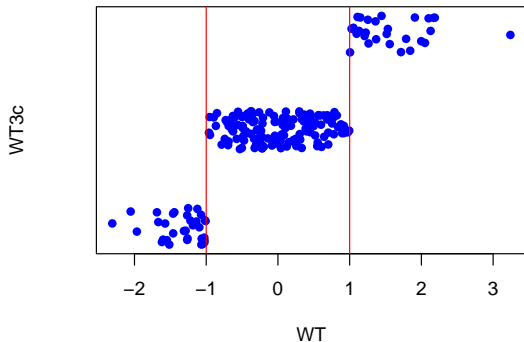
| ##             | Estimate | Std. Error | t value | Pr(> t ) |
|----------------|----------|------------|---------|----------|
| ## (Intercept) | 0.00398  | 0.0419     | 0.0949  | 9.24e-01 |
| ## OW          | 0.29128  | 0.0571     | 5.1014  | 7.90e-07 |
| ## WTm         | -0.61306 | 0.0639     | -9.6003 | 3.80e-18 |

**Even taking into account the confounding factor WTm (measured with some error), we show a significant effect of OW on DW.**

Another realistic case, if only a qualitative measure of WT is used (e.g. categorized in 3 classes)

WT transformed into a categorical variable (WT3c) with three classes:

$]-4; -1]$ ,  $]-1; 1]$ ,  $]1; 4]$ .



## Linear model: DW as a function of OW and WT3c

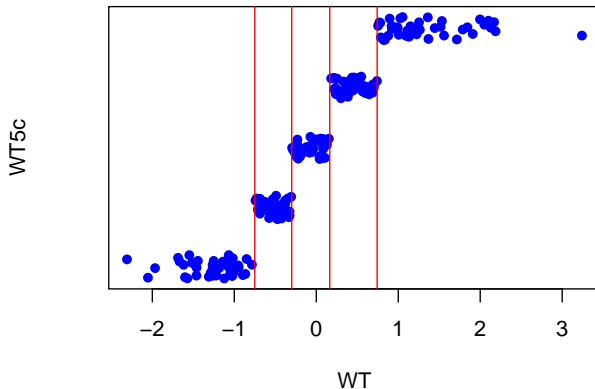
```
mOWWT3c <- lm(DW ~ OW + WT3c)
summary(mOWWT3c)$coefficients
```

| ##             | Estimate | Std. Error | t value | Pr(> t ) |
|----------------|----------|------------|---------|----------|
| ## (Intercept) | 0.942    | 0.1366     | 6.90    | 7.08e-11 |
| ## OW          | 0.324    | 0.0567     | 5.72    | 3.86e-08 |
| ## WT3c(-1,1]  | -0.909   | 0.1463     | -6.21   | 3.05e-09 |
| ## WT3c(1,4]   | -2.119   | 0.2317     | -9.14   | 7.76e-17 |

**We still show a significant effect of OW on DW!**

## And if the discretization of WT is less coarse and balanced

WT transformed into a categorical variable (WT5c) with five balanced classes, whose limits are defined by the quintiles.



## Linear model: DW as a function of OW and WT5c

```
mOWWT5c <- lm(DW ~ OW + WT5c)
summary(mOWWT5c)$coefficients
```

| ##                    | Estimate | Std. Error | t value | Pr(> t ) |
|-----------------------|----------|------------|---------|----------|
| ## (Intercept)        | 1.109    | 0.1247     | 8.90    | 4.08e-16 |
| ## OW                 | 0.141    | 0.0613     | 2.31    | 2.21e-02 |
| ## WT5c(-0.75,-0.299] | -0.678   | 0.1391     | -4.87   | 2.31e-06 |
| ## WT5c(-0.299,0.165] | -0.985   | 0.1506     | -6.54   | 5.22e-10 |
| ## WT5c(0.165,0.743]  | -1.446   | 0.1687     | -8.57   | 3.26e-15 |
| ## WT5c(0.743,4]      | -2.340   | 0.2121     | -11.03  | 2.83e-22 |

**We still show a significant effect of OW on DW!**

## Conclusion about the possibility of taking into account confounding factors in a linear model

A problem of this kind is very realistic and I let you imagine the consequences!

**Taking into account the potential confounding variables in a linear model is essential, but great caution is required when interpreting the results** of a linear model on **observational data** (i.e. with uncontrolled input variables).

Conclusion

## Conclusion

- ▶ Statistical modeling is a **powerful but not perfect** tool and should be handled with **great caution**.
- ▶ There is **no unique / best strategy to build a model**. Authors should be able to well describe and argue their own strategy in order to convince the reader it is well-founded.
- ▶ The question of the **end use of a model (explanatory / predictive)** is an underlying question that we should keep in mind while developing any type of model (also crucial in machine learning).