# Variable selection

Does it matter from a machine learning perspective?

Filippo Biscarini
*Senior Scientist*
*CNR, Milan (Italy)*

# Traditional statistics

Two groups of pigs: diet A, diet B → which promotes growth better?

**weight = mean + diet + e**

!! is “diet” relevant/significant for growth?

| Diet A | Diet B |
|---|---|
| 90 kg | 89 kg |
| 88 kg | 82 kg |
| 92 kg | 79 kg |
| 87.5 kg | 83 kg |
| … | … |

# Traditional statistics

What about the sex of the pigs? And their age? Or breed? Interactions?

weight = mean + diet + sex + e

weight = mean + diet + sex + age + e

weight = mean + diet + sex + age + breed + e

weight = mean + diet + sex + age + breed + … + e

weight = mean + diet + sex + age + breed + age*breed + … + e

| Diet A | Diet B |
|---|---|
| 90 kg | 89 kg |
| 88 kg | 82 kg |
| 92 kg | 79 kg |
| 87.5 kg | 83 kg |
| … | … |

# What about machine learning?

In machine learning **the model learns on its own** which variables to use and how (not easily accessible by humans)

→ ”black box”

# What about machine learning?

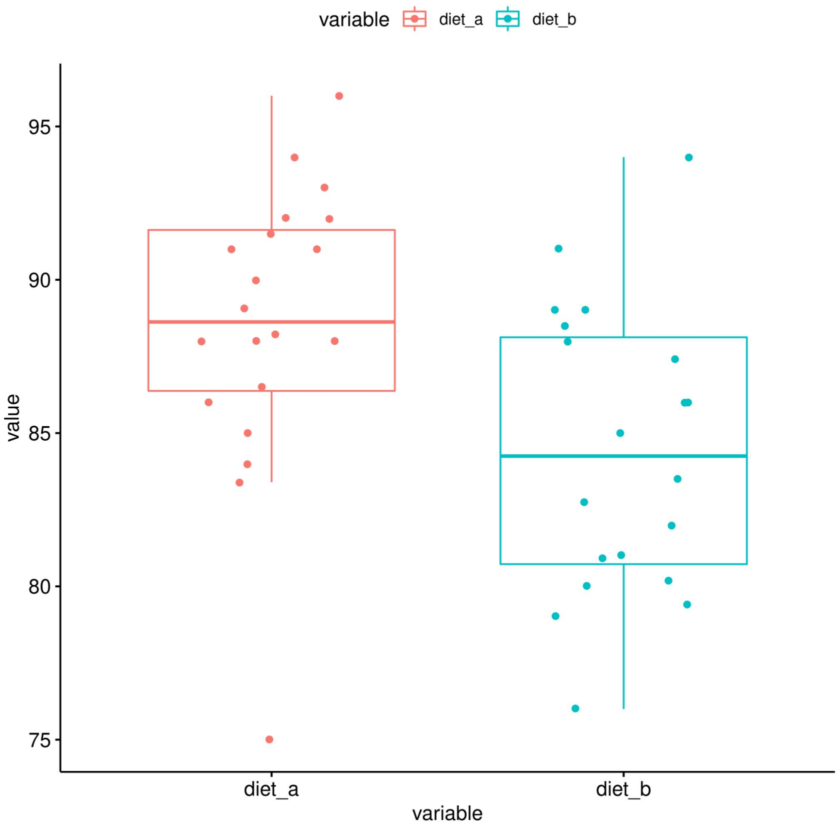
In machine learning **the model learns on its own** which variables to use and how (not easily accessible by humans) → ”black box”

- If the ML model is able to determine that burying ground quartz stuffed into the horn of a cow (which are said to harvest "cosmic forces in the soil") has no effect on the growth of pigs, this information is kept hidden in the model
- Sounds suboptimal, but **the interpretability of single variables loses sense as the number of variables** and their combinations and transformations **increases**
- And in modern statistics we usually have A LOT of variables!

# An illustration - ANOVA

| *term | SS | d.f. | F | p-value |
|---|---|---|---|---|
| intercept | 189.2 | 1 | 415.11 | 5.2e-37 |
| diet | 2.35 | 1 | 5.39 | 0.022 |
| residual | 4.79 | 100 | | |

*made up numbers!!

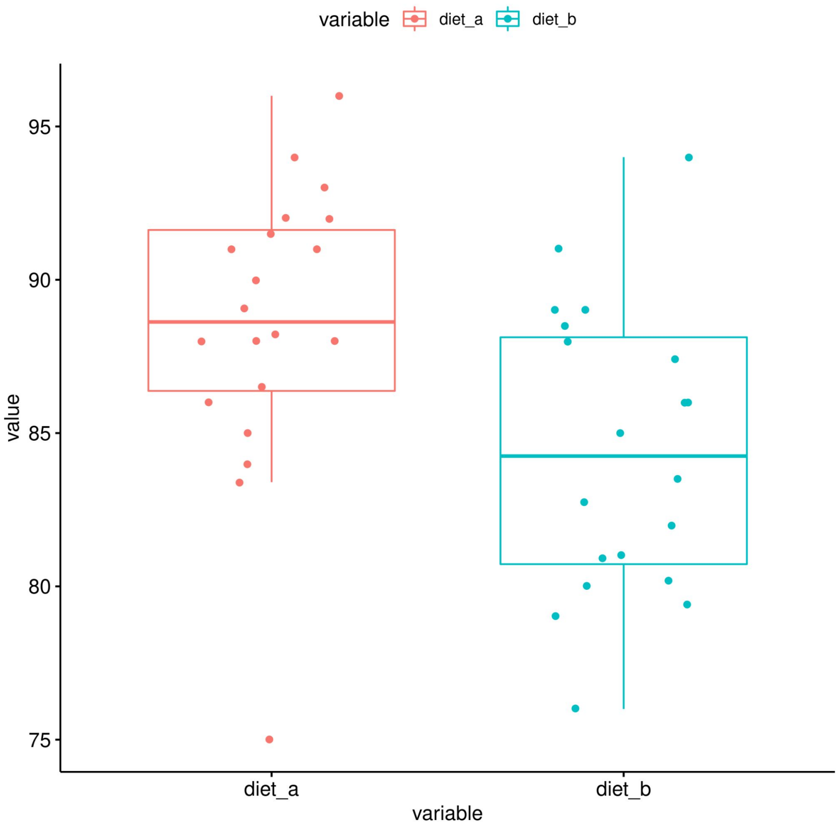


# An illustration - linear model

| *term | SS | d.f. | F | p-value |
|---|---|---|---|---|
| intercept | 189.2 | 1 | 415.11 | 5.2e-37 |
| diet | 2.35 | 1 | 5.39 | 0.022 |
| residual | 4.79 | 100 | | |

=

**weight = mean + b1*diet + e**



# An illustration - linear model

**weight = mean + b1*diet + e**

*mean = **80** kg
*b1 = **+2.75** kg [coding diet A = 1; diet B = 0]

<u>Interpretation</u>

- mean: average weight of pigs
- b1: average difference in weight between pigs fed with diet A and pigs fed with diet B

***made up numbers!!**

# An illustration - linear model

**weight = mean + b1*diet + b2*age + e**

mean = **80** kg
b1 = **+2.75** kg [coding diet A = 1; diet B = 0]
p-value = 0.006 → b2 = **+1.47** kg [coding age in years]

<u>Interpretation</u>

- mean: average weight of pigs
- b1: average difference in weight between pigs fed with diet A and pigs fed with diet B
- b2: average weight gain per year of age, keeping diet constant

**made up numbers!!**

# An illustration - linear model

Traditional statistics approach!

**weight = mean + b1*diet + b2*age + b3*motion_time + e**

The coefficient for “motion time” (b3) has a p-value of 0.29: we decide not to include motion time in the model

But what if the relationship between motion time and weight is not linear?
We can fit **polynomial terms**! (square, cube etc.) [ → this is still a linear model!]

# An illustration - linear model

$$weight = mean + b1*diet + b2*age + b3*motion\_time + b4*motion\_time^2 + b5*motion\_time^3 + e$$

The p-values for the polynomial terms are now 0.075, 0.051 and 0.032:

- **should we include these in the model?**

The coefficients for the polynomial terms are: -1.57, 0.24 and -0.03

- **how should we interpret these?**
- on average, we lose 1.57 kg per hour of motion, we gain 0.24 kg per hour-of-motion squared, and we lose 30 grams per hour-of-motion cubed

How to build and interpret the model becomes more and more confused

# The machine learning perspective

With many variables (but already with a handful of variables) it becomes a titanic task to decide which variables, combinations of variables and functions of variables include in the model
→ **let the model decide!**

The questions of variable selection and model interpretability become ill-posed
→ **predictions matter more than inference!**

Is this the end of the story? Can we really say nothing about why our model works (or does not work)?
→ **don't panic, we'll be able to crack the black box (at least partially)**

# The machine learning perspective

**variable selection ≠ data representation**

**variable selection ≠ feature engineering**

**How does the model decide which variables to use?** (hold on a little longer …)