

QML - Week 5

Regression models: the basics
Stefano Coretta

Word frequency and reaction times

What is the relationship between a word's lexical frequency and reaction times in a lexical decision task in Croatian?

- Data from [Lexical decision times for nouns from the Croatian Psycholinguistic Database](#).
- Lexical decision task (*Is it a real Croatian word?*)
- Reaction times.**
- Word frequency:** counts from the Croatian web Corpus *hrWaC*.

Which relationship?

Participate in an event

Event code: *** Join

Welcome back!
How would you like to log in?

Reaction times

Word frequency and RTs

Logged word frequency and RTs

Gaussian model of RT

$RT \sim Gaussian(\mu, \sigma)$

But we want to know what happens to RTs depending on the value of lexical frequency...

Then we let the mean μ vary by lexical frequency!

$RT \sim Gaussian(\mu, \sigma)$

$\mu = \beta_0 + \beta_1 \cdot \log f$

But what are those β_0 and β_1 ?

The equation of a line

$y = \beta_0 + \beta_1 \cdot x$

- Go to [Linear models illustrated](#).
- β_0 is the line **intercept**: the y value when x is zero.
- β_1 is the line **slope**: the change in y for each unit-increase of x .

Regression model

$RT \sim Gaussian(\mu, \sigma)$

$\mu = \beta_0 + \beta_1 \cdot \log f$ [Regression equation]

- A regression model** is a model based on the equation of a line.
- The model **estimates** β_0 (the intercept) and β_1 (the slope) from the data (i.e. the observed RT and log f values).
- β_0 , intercept
- Mean RT value when *logged frequency is 0* (i.e. when word frequency is 1; $\exp(0) = 1$).
- β_1 , slope
- Change in mean RT for each unit increase of log-frequency* (when log-frequency goes from x to $x+1$).

Model's posterior predictions

Participate in an event

Event code: *** Join

Word frequency and reaction times (bis)

What is the relationship between a word's lexical frequency and reaction times in a lexical decision task in Croatian?

- When log-frequency is 0, the mean RTs are between 1084 and 1129 ms at 95% confidence.
- For each unit increase of log-frequency, the mean RTs decrease by 43-48 ms, at 95% confidence.

Correlation in NOT causation

Be careful!

- Correlation** between two variables: they co-vary, i.e. they show a systematic association (their values tend to vary together in a consistent pattern).
- Spurious correlations:** two variables can look correlated because of bias from another variable.

- Number of plant names in a language vs. biodiversity of the region
- Languages in biodiverse regions have more words for plants.
- Mediator:** cultural reliance on plants.
- Language endangerment vs. economic development
- Higher economic development is associated with greater language endangerment.
- Confounder:** colonial history.
- Language prestige vs. government policy
- High prestige languages and officially supported languages each attract learners.

But it is if you use causal inference...

Causal inference

- Correlation *can* be interpreted causally if you adopt a **causal inference** approach.
- Learn about it in McElreath's textbook *Statistical Rethinking*. Also check [STeW](#).

Speaker notes