

# 7 Regressió generalitzada II

Mètodes empírics 2

23/05/2022

# Avui

- Cas d'estudis
- Generalització lineal Binomial/Bernoulli
- Temes avançats
- (Xarxes neuronals & ChatGPT)

# Zipf's Law of Abbreviation

---

Kanwal et al. (2017): Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Gesture*.

**Vols saber quins factors (no) afecten l'ús de la paraula curta a l'experiment de Kanwal et al.; aquestes són les dades:**

**<https://tinyurl.com/2s3p9s2z>**

- 1. Descriu les dades**
- 2. Descriu com penses que les variables es podrien relacionar**
- 3. Quins valors poden tenir les variables del teu interès?**

# Dades

##	pairnum	IP	trial	display	label
## 1	1	67.85.42.18	1	0	zop
## 2	1	67.85.42.18	2	3	zopudon
## 3	1	67.85.42.18	3	0	zop
## 4	1	67.85.42.18	4	0	zopekil
## 5	1	67.85.42.18	5	2	zopudon
## 6	1	67.85.42.18	6	1	zopekil

# Variables independents (predictors)

- trial: 1, 2, ..., 31, 32
- display: 0, 1, 2, 3

```
df$freq <- ifelse(df$display %in% c(0,1), 'freq', 'infreq')
```

- freq: infreq o freq

# Variable dependent (resultat)

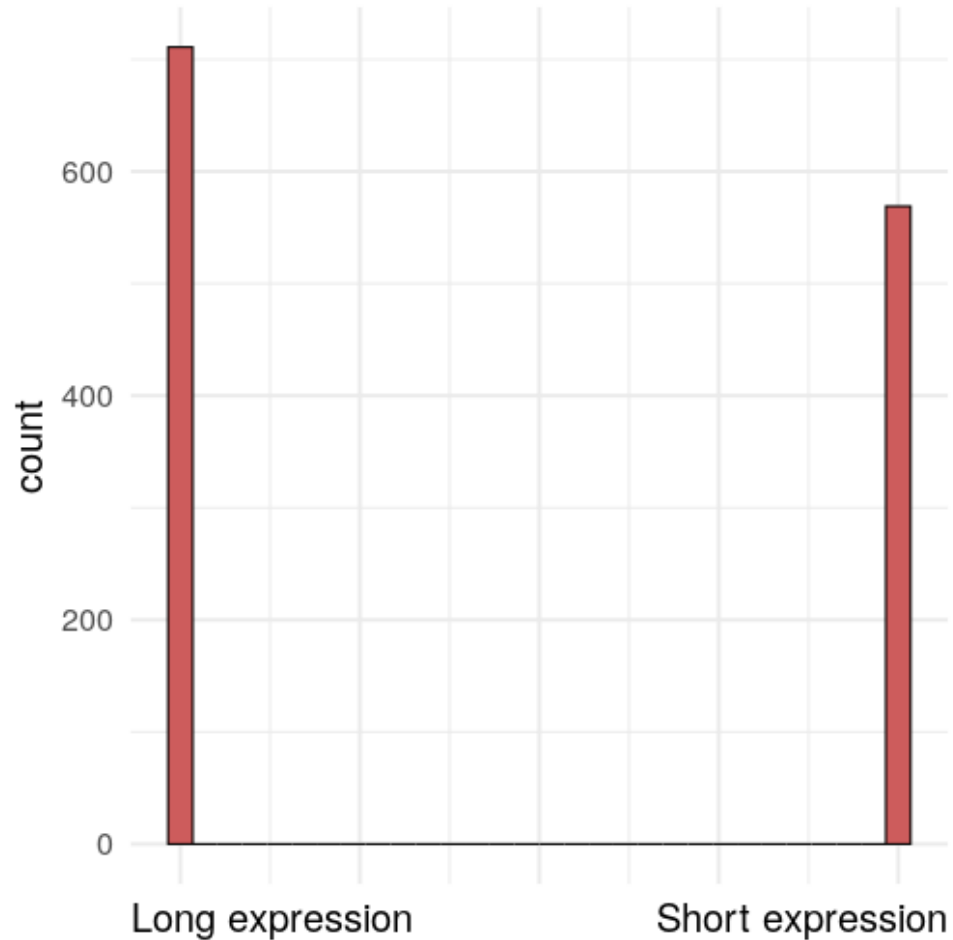
```
df$label[1:10]
```

```
## [1] zop      zopudon zop      zopekil zopudon zopekil zop      zop      zop
## Levels: zop zopekil zopudon
```

```
df$short <- ifelse(df$label == 'zop', 1, 0)
df$short[1:10]
```

```
## [1] 1 0 1 0 0 0 1 1 1 1
```

# Variable dependent (resultat)





# Model lineal generalitzat: Bernoulli / Binomial

# Sustainability

Model lineal generalitzat: Bernoulli / Binomial

$$y_i \sim \text{Bernoulli}(p_i)$$

$$p_i = f(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)$$

$$f(x) = \text{inverse logit}(x)$$

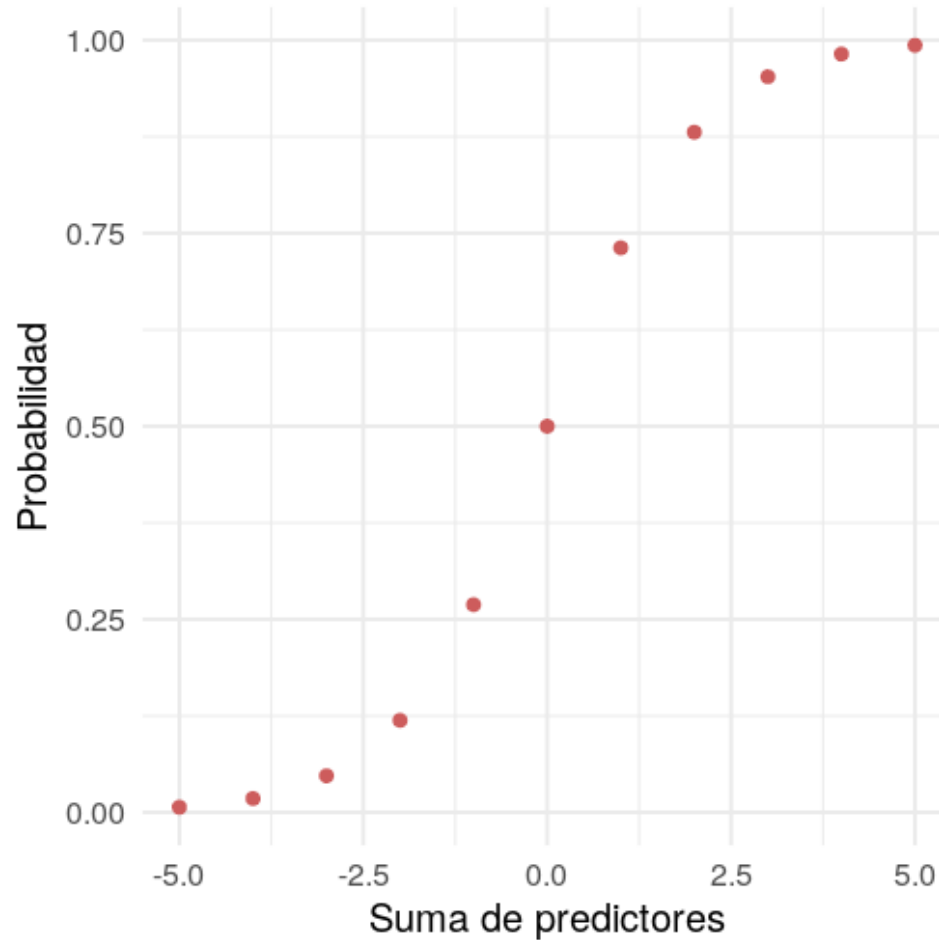
# Logit i el seu invers

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\text{inverse logit}(p) = \frac{\exp(p)}{1 + \exp(p)}$$

```
inv.logit <- function(x){  
  return(exp(x) / (1 + exp(x)) )  
}
```

# Espai invers logit



# Model lineal generalitzat: Bernoulli

$$y_i \sim \text{Bernoulli}(p_i)$$

$$p_i = \text{inv.logit}(\beta_0 + \beta_1 x_1)$$

Model lineal generalitzat: Bernoulli

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

# Regressió Binomial/Bernoulli (R)

```
glm(formula = short ~ 1 + freq,  
    data     = df,  
    family   = binomial(link = 'logit')  
)
```



# Zipf's Law of Abbreviation

---

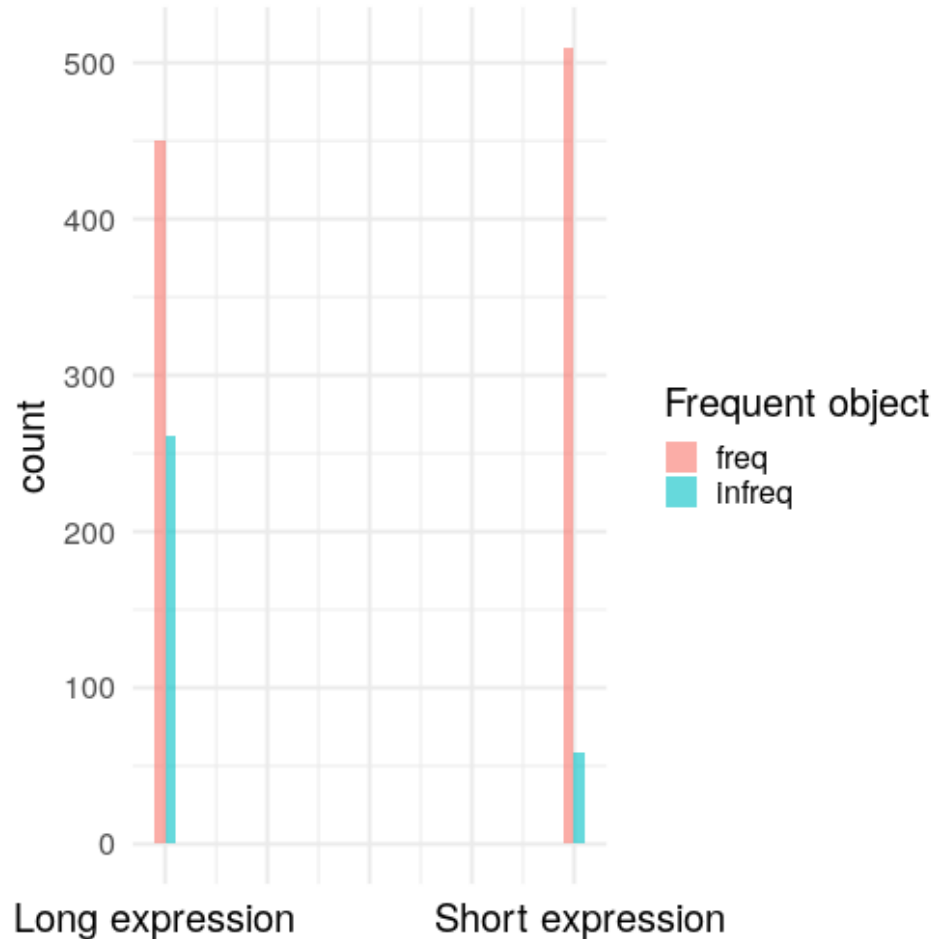
Kanwal et al. (2017): Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*.

# Dades

```
head(df)
```

##	pairnum	IP	trial	display	label	freq	short
## 1	1	67.85.42.18	1	0	zop	freq	1
## 2	1	67.85.42.18	2	3	zopudon	infreq	0
## 3	1	67.85.42.18	3	0	zop	freq	1
## 4	1	67.85.42.18	4	0	zopekil	freq	0
## 5	1	67.85.42.18	5	2	zopudon	infreq	0
## 6	1	67.85.42.18	6	1	zopekil	freq	0

# longitud d'expressió ~ freqüència d'objecte



# Model 1: Freqüència

```
zipf_freq <- glm(formula = short ~ 1 + freq,  
                  data     = df,  
                  family   = binomial(link = 'logit')  
                  )  
  
coef(zipf_freq)
```

```
## (Intercept)  freqinfreq  
##    0.1251631  -1.6121461
```

---

$$\text{logit}(p_i) \approx 0.13 - (\text{infrec} \times 1.612) \approx -1.482$$

# Model 1: Freqüència

$$\text{logit}(p_i) \approx 0.13 - (\text{infrec} \times 1.612) \approx -1.482$$

```
inv.logit(0.13 - 1.612)
```

```
## [1] 0.1851255
```

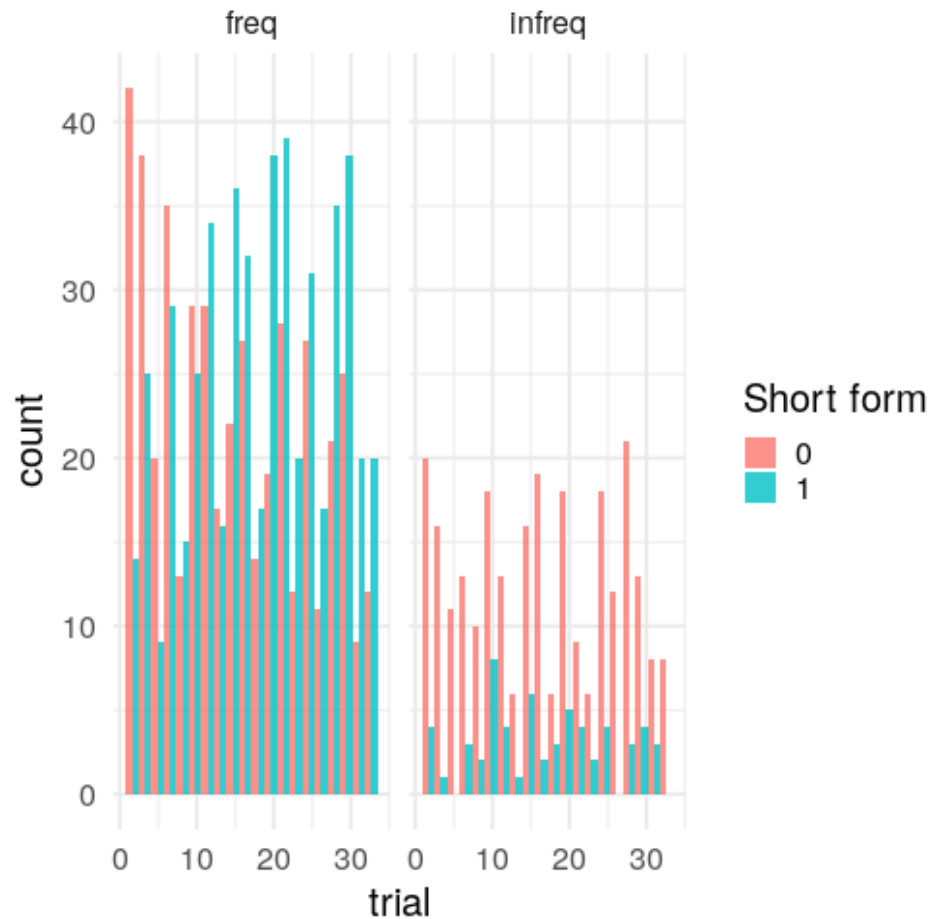
```
inv.logit(0.13)
```

```
## [1] 0.5324543
```

# summary(zipf\_freq)

```
##
## Call:
## glm(formula = short ~ 1 + freq, family = binomial(link = "logit"),
##      data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2310  -1.2310  -0.6384   1.1247   1.8389
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.12516    0.06468   1.935   0.053 .
## freqinfreq  -1.61215    0.15800 -10.204  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1758.7  on 1279  degrees of freedom
## Residual deviance: 1633.0  on 1278  degrees of freedom
## AIC: 1637
##
## Number of Fisher Scoring iterations: 4
```

# longitud d'expressió ~ torn



# Model 2: Torn

```
zipf_trial <- glm(formula = short ~ 1 + trial,  
                  data      = df,  
                  family    = binomial(link = 'logit')  
                  )  
coef(zipf_trial)
```

```
## (Intercept)      trial  
## -0.73645501  0.03085928
```



# Prediccions

```
inv.logit(-0.74 + (1 * 0.039))
```

```
## [1] 0.3315906
```

```
inv.logit(-0.74 + (20 * 0.039))
```

```
## [1] 0.5099987
```

```
inv.logit(-0.74 + (32 * 0.039))
```

```
## [1] 0.6243375
```

# summary(zipf\_freq)

```
##
## Call:
## glm(formula = short ~ 1 + trial, family = binomial(link = "logit"),
##      data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2857  -1.0889  -0.9189   1.2285   1.4879
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.736455   0.118639  -6.208 5.38e-10 ***
## trial        0.030859   0.006204   4.974 6.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1758.7  on 1279  degrees of freedom
## Residual deviance: 1733.5  on 1278  degrees of freedom
## AIC: 1737.5
##
## Number of Fisher Scoring iterations: 4
```

# Model 3: Tots dos predictors

```
zipf_trial_freq <- glm(formula = short ~ 1 + trial + freq,  
                        data      = df,  
                        family    = binomial(link = 'logit')  
                        )  
coef(zipf_trial_freq)
```

```
## (Intercept)      trial  freqinfreq  
## -0.42733747  0.03359338 -1.64081659
```

# Model 3: Tots dos predictors

```
##
## Call:
## glm(formula = short ~ 1 + trial + freq, family = binomial(link = "logit",
##      data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.462  -1.097  -0.562   1.089   2.077
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.427337   0.125352  -3.409 0.000652 ***
## trial        0.033593   0.006536   5.140 2.75e-07 ***
## freqinfreq  -1.640817   0.159698 -10.275 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1758.7  on 1279  degrees of freedom
## Residual deviance: 1606.0  on 1277  degrees of freedom
## AIC: 1612
##
```

# AICs

```
zipf_freq$aic
```

```
## [1] 1636.987
```

```
zipf_trial$aic
```

```
## [1] 1737.473
```

```
zipf_trial_freq$aic
```

```
## [1] 1611.984
```

# Temes avançats

# Interacciones

```
m4 <- glm(formula = short ~ 1 + trial + freq + trial:freq,  
          data      = df,  
          family    = binomial(link = 'logit')  
          )  
  
coef(m4)
```

##	(Intercept)	trial	freq	freq	trial:freq	freq
##	-0.53501091	0.04018210	-0.99080221		-0.03782272	

# Models hierarquics

```
library(lme4)

m5 <- glmer(formula = short ~ 1 + trial + freq +
              trial:freq +
              (1 | IP),
              data    = df,
              family   = binomial(link = 'logit')
              )
```



# Models hierarquics

```
fixef(m5)
```

##	(Intercept)	trial	freqinfreq	trial:freqinfreq
##	-1.19210355	0.07232261	-1.18527634	-0.07347172

```
head(ranef(m5)$IP)
```

##	(Intercept)
## 100.10.40.83	1.897175
## 100.2.122.157	1.737037
## 104.174.222.43	-1.743050
## 107.161.163.8	-2.364535
## 115.99.18.32	-1.490564
## 117.213.33.129	-2.768317

# Mes temes

- meta anàlisi
- GAMs
- Xarxes neuronals

# Propera sessió

- Lliurament part II de "Revisió per parells": 30/05
- 

- **Corpus**
- 

- Informe final: 25/06