Agenda

Feb 16

1. Worksheet notes
2. Introducing the multinomial distribution
3. Categorical outcome variables
4. Softmax link function
5. Interpreting coefficients
6. Multinomial logistic in R using ‘brms’

Worksheet notes

Referring to files

- When referring to files, make it something that will work on other computers
- Things like “C:/SomeFolder/data.csv” or “~/Downloads/data.csv” are likely to only work on your computer
- Instead, either get the data directly from the URL or put the data in your working directory and just use “data.csv”

Avoid interactive commands

- Some commands (e.g. “View()”) only work in certain interactive settings — avoid putting them into .Rmd files

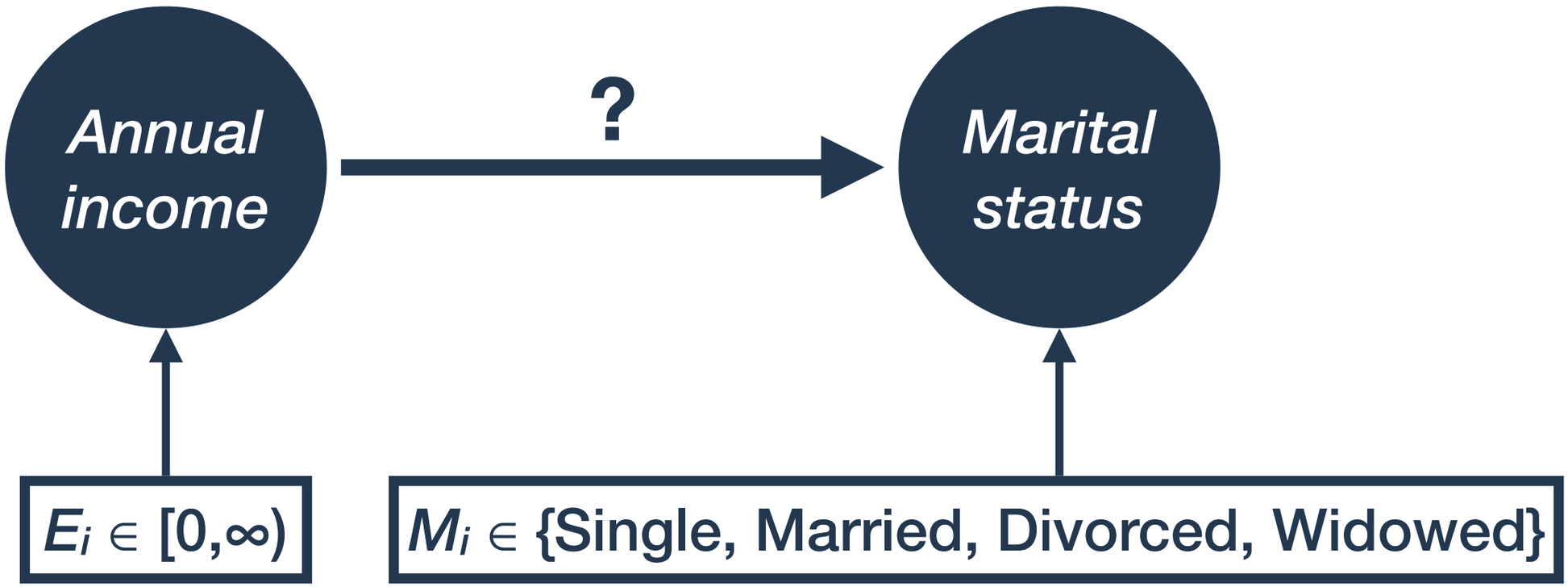
Minimize packages

- Only load the packages that you actually use in a script

Individual feedback

- Forthcoming!

Income and marital status



The problem	Outcome variable has multiple (>2) categories. Binomial and Poisson models won't work.
The solution	Use a multinomial outcome distribution (and a new link function) to account for the data.

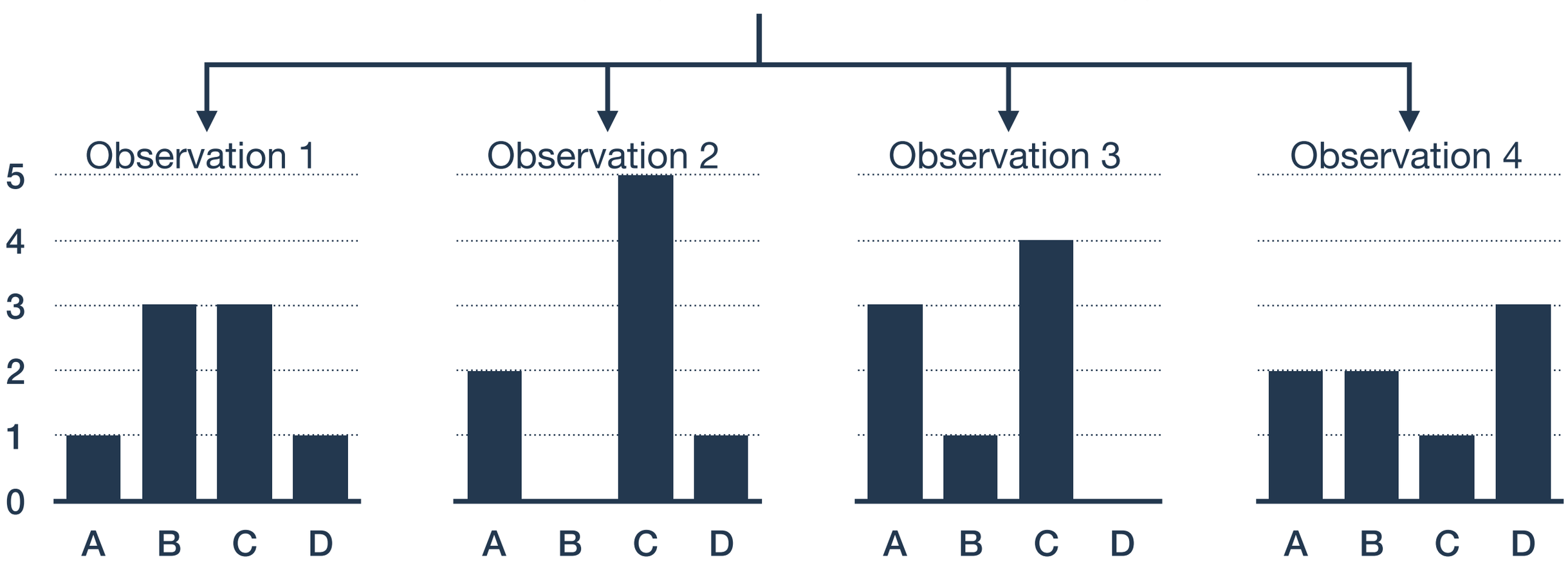
Multinomial distribution

$$\text{Multinom}\left(n, \left(p_1, \ldots, p_k\right)\right)$$

Multinomial distribution | Result of n trials, each of which can result in one of k outcomes with probability p_1, p_2, …, p_k.

Each 'observation' describes outcome of n trials:

$$\text{Multinom}\left(8, \left(0.20, 0.10, 0.45, 0.25\right)\right)$$



Multinomial distribution

Binomial, Bernoulli, and categorical distributions are special cases of the multinomial.

Binomial distribution	Bin(n, p) = Multinom(n, (1–p, p))
Bernoulli distribution	Bernoulli(p) = Multinom(1, (1–p, p))
Categorical distribution	Cat(p_1, p_2, …, p_k) = Multinom(1, (p_1, p_2, …, p_k))

	1 trial	>1 trials
2 categories	**Bernoulli**	**Binomial**
>2 categories	**Categorical**	**Multinomial**

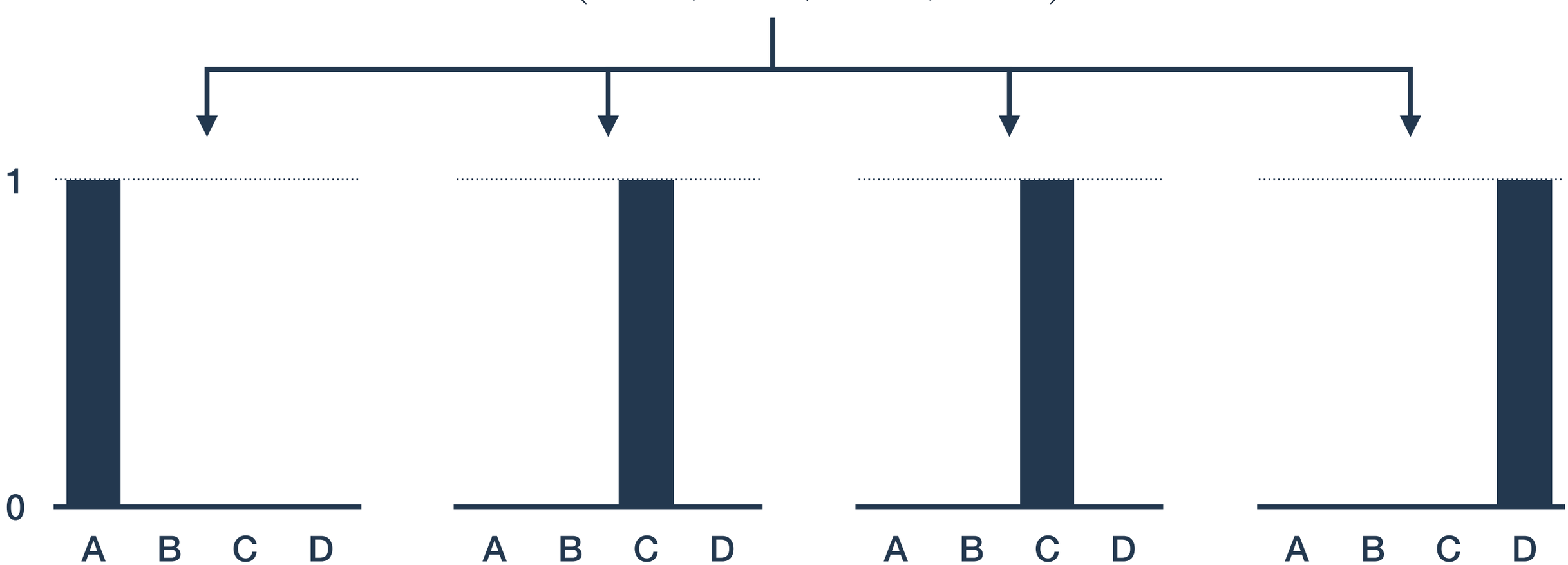
Categorical distribution

$$\text{Cat}\left(p_1, \ldots, p_k\right)$$

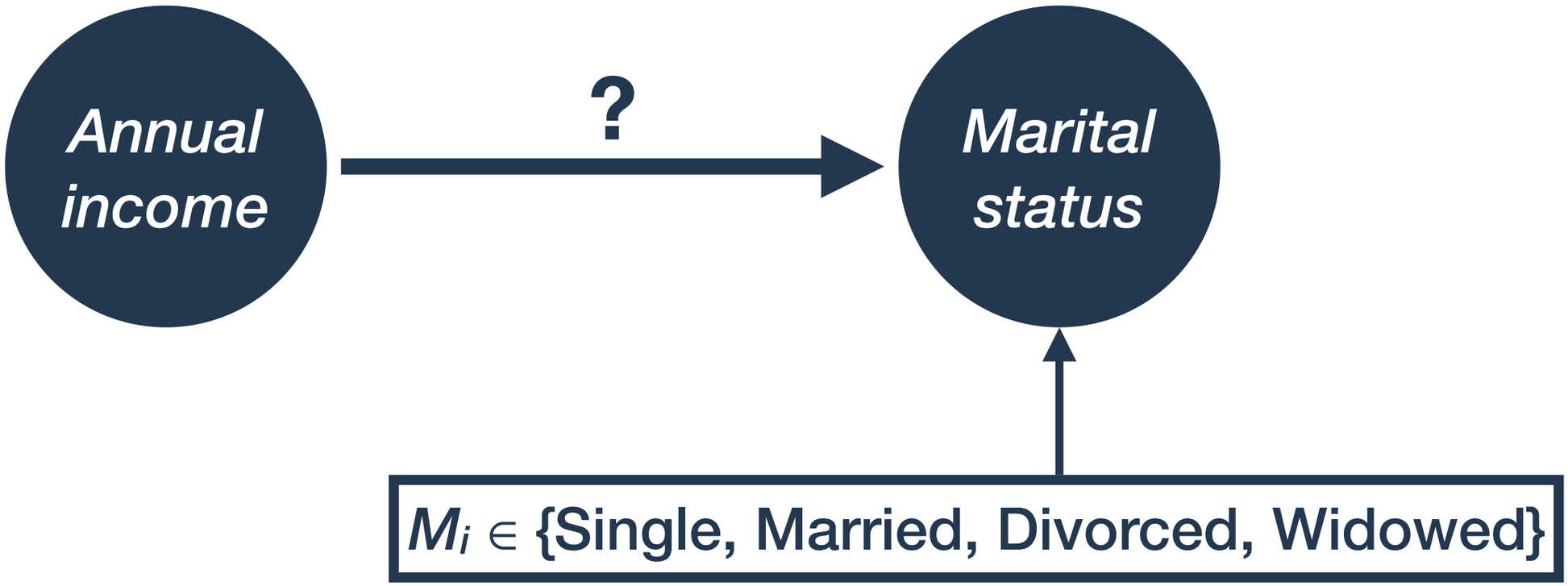
Categorical distribution | Multinomial distribution with just one trial

---

$$\text{Cat}\left(0.20, 0.10, 0.45, 0.25\right)$$



Categorical outcome



One category has to be the *reference* category.

$$s_s = 0$$

$$s_m = \alpha_m + \beta_m E_i$$

$$s_d = \alpha_d + \beta_d E_i$$

$$s_w = \alpha_w + \beta_w E_i$$

Each other category gets its own coefficient.

Softmax link function

$M_i \in$ {Single, Married, Divorced, Widowed}

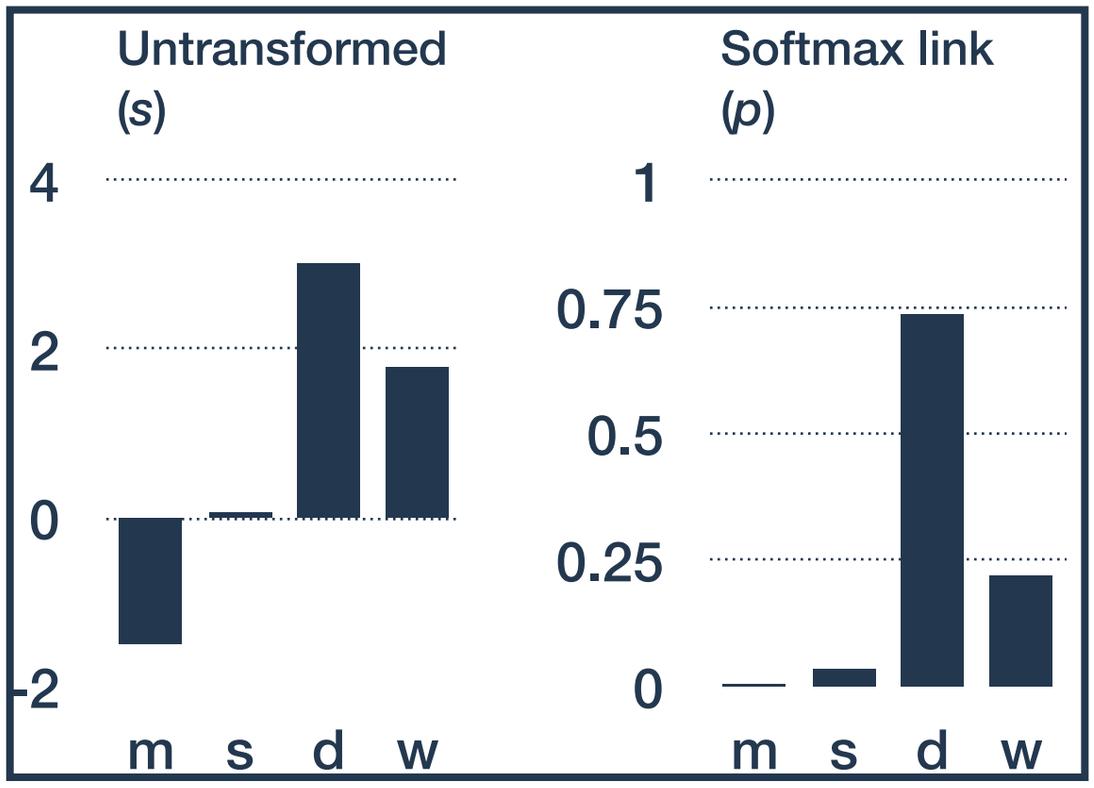
$$M_i \sim \text{Cat}\left(\text{softmax}(s_s, m_s, s_d, s_w)\right)$$

$$s_s = 0$$

$$s_m = \alpha_m + \beta_m E_i$$

$$s_d = \alpha_d + \beta_d E_i$$

$$s_w = \alpha_w + \beta_w E_i$$



Softmax is a multivariate generalization of inverse logit.

$$p_s = \text{softmax}(s_s) = \frac{\exp(s_s)}{\exp(s_s) + \exp(s_m) + \exp(s_d) + \exp(s_w)}$$

Multinomial logistic regression

Multinomial logistic (or categorical) regression model.

$$
\begin{aligned}
M_i &\sim \text{Cat}\left(\text{softmax}\left(s_{si}, s_{mi}, s_{di}, s_{wi}\right)\right) \\
s_{si} &= 0 \\
s_{mi} &= \alpha_m + \beta_m E_i \\
s_{di} &= \alpha_d + \beta_d E_i \\
s_{wi} &= \alpha_w + \beta_w E_i \\
\alpha_s, \alpha_d, \alpha_w &\sim \text{Norm}(0, 2) \\
\beta_s, \beta_d, \beta_w &\sim \text{Norm}(0, 3)
\end{aligned}
$$

Multinomial logistic regression

With two categories, the multinomial logistic model is the standard (binomial) logistic model.

$$M_i \sim \text{Cat}\left(\text{softmax}(s_{1i}, s_{2i})\right)$$

$$s_{1i} = 0$$

$$s_{2i} = \alpha + \beta E_i$$

$$\alpha \sim \text{Norm}(0, 1)$$

$$\beta \sim \text{Norm}(0, 3)$$

$$p_{2i} = \frac{\exp(s_{2i})}{1 + \exp(s_{2i})} = \text{logit}^{-1}(s_{2i})$$

Interpreting estimates

$$M_i \sim \text{Cat}\left(\text{softmax}(s_{mi}, s_{si}, s_{di}, s_{wi})\right)$$

$$s_{mi} = 0$$

$$s_{si} = \alpha_s + \beta_s E_i$$

$$s_{di} = \alpha_d + \beta_d E_i$$

$$s_{wi} = \alpha_w + \beta_w E_i$$

$$\alpha_s, \alpha_d, \alpha_w \sim \text{Norm}(0, 2)$$

$$\beta_s, \beta_d, \beta_w \sim \text{Norm}(0, 3)$$

	Mean	*90% credible interval*	
α_s	5.35	4.73	5.98
β_s	-0.59	-0.65	-0.53
α_d	0.57	-0.24	1.37
β_d	-0.18	-0.25	-0.10
α_w	1.94	0.89	2.98
β_w	-0.40	-0.50	-0.30

Interpreting estimates

	Mean	90% credible interval	
α_s	5.35	4.73	5.98
β_s	-0.59	-0.65	-0.53
α_d	0.57	-0.24	1.37
β_d	-0.18	-0.25	-0.10
α_w	1.94	0.89	2.98
β_w	-0.40	-0.50	-0.30

Estimation in R with brms

Estimating with brms

The `brm()` function allows you to use the model syntax from `lm()` and `glm()`

```
model <- marital_status ~ log_income

fit <- brm(model, data=d)
```

Priors are set using the `prior()` function

```
model <- marital_status ~ log_income
pr <- c(
  prior(normal(0,2), class='b'), # coefficients
  prior(normal(0,3), class='Intercept') # interc.
)
fit <- brm(model, data=d, prior=pr)
```