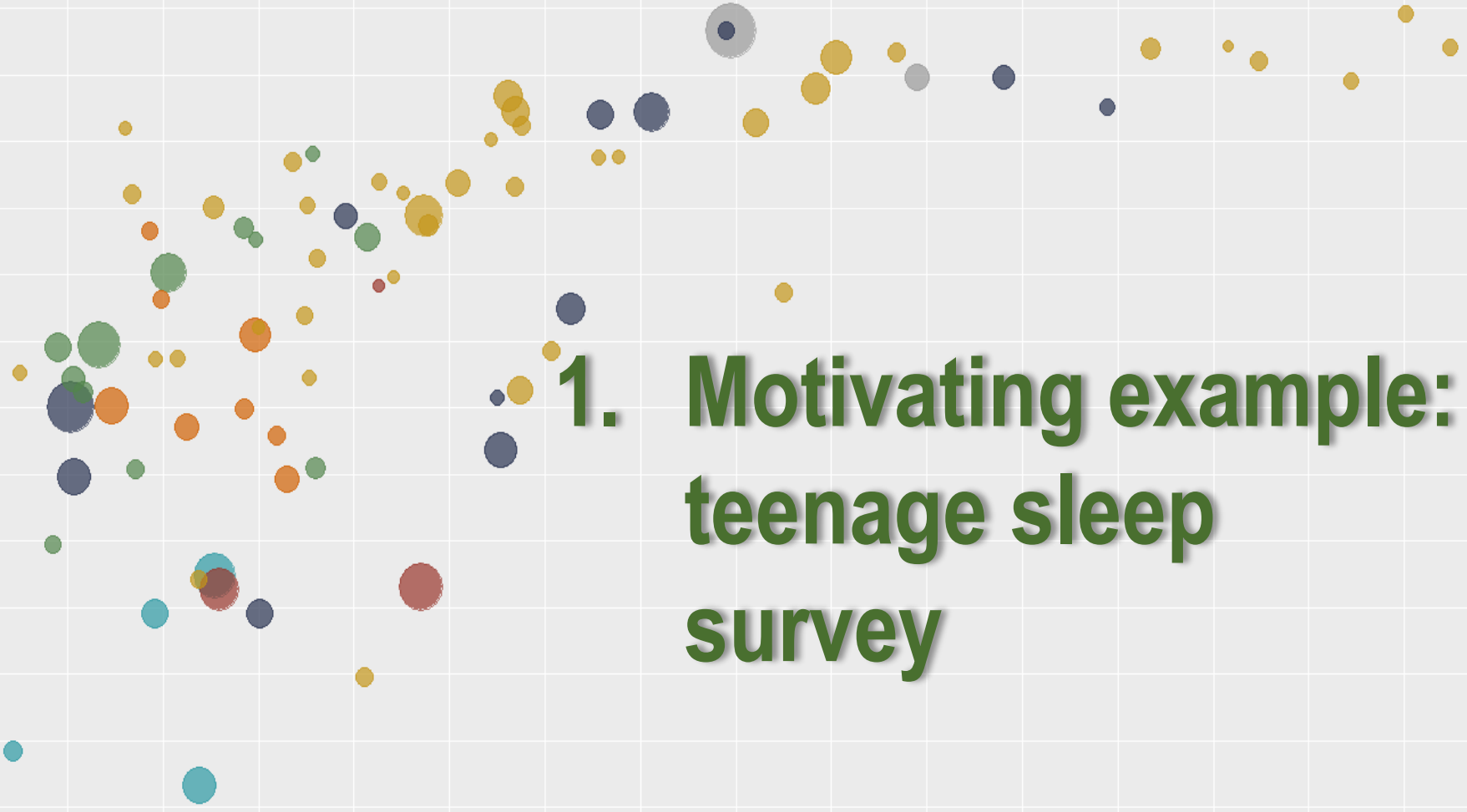
Quantitative analysis +

Chris Moreh, 2024

Week 2

Categories

Logistic regression and other generalised linear models



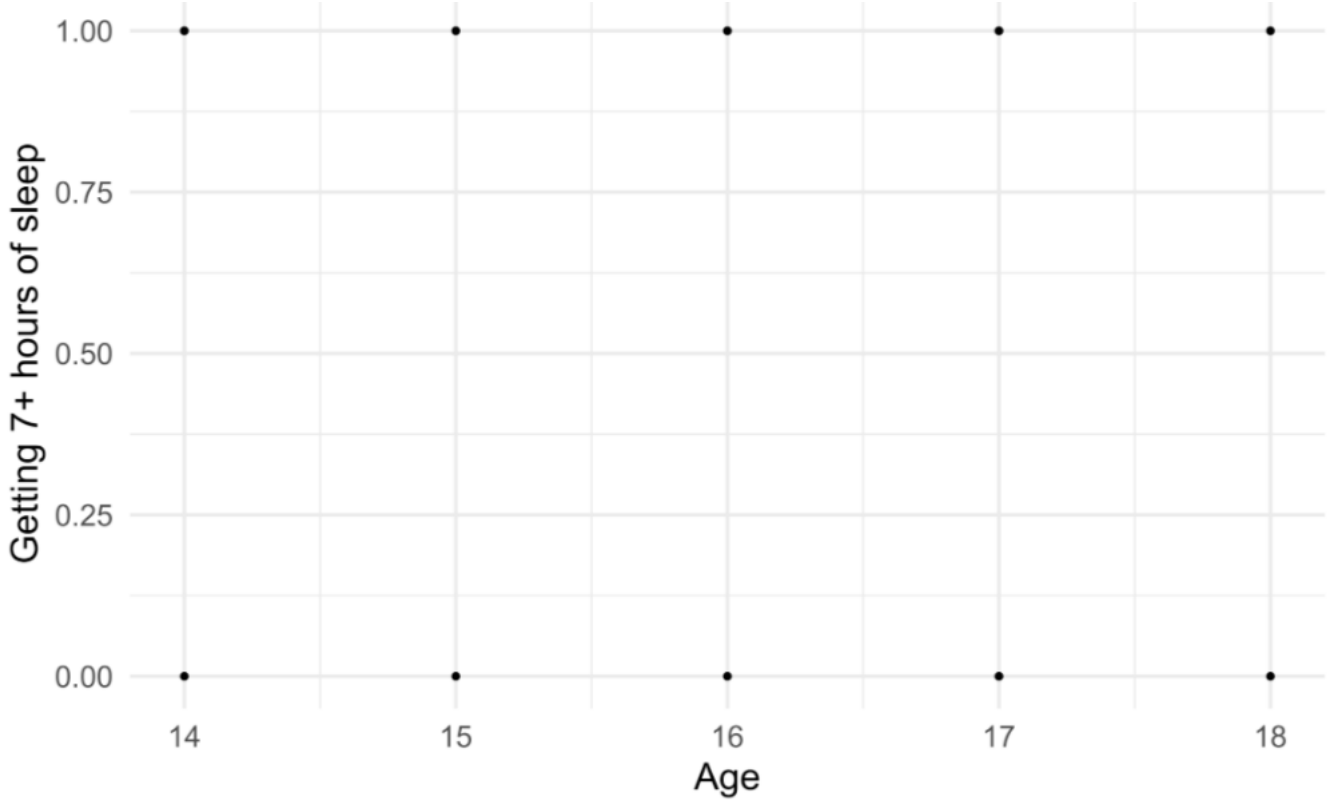
1. Motivating example: teenage sleep survey

Example

- Question: “Do teenagers get 7+ hours of sleep?”
- Students in grades 9 - 12 surveyed about health risk behaviours including whether they usually get 7 or more hours of sleep
- Sleep:
  - 1 = “Yes”
  - 0 = “No”

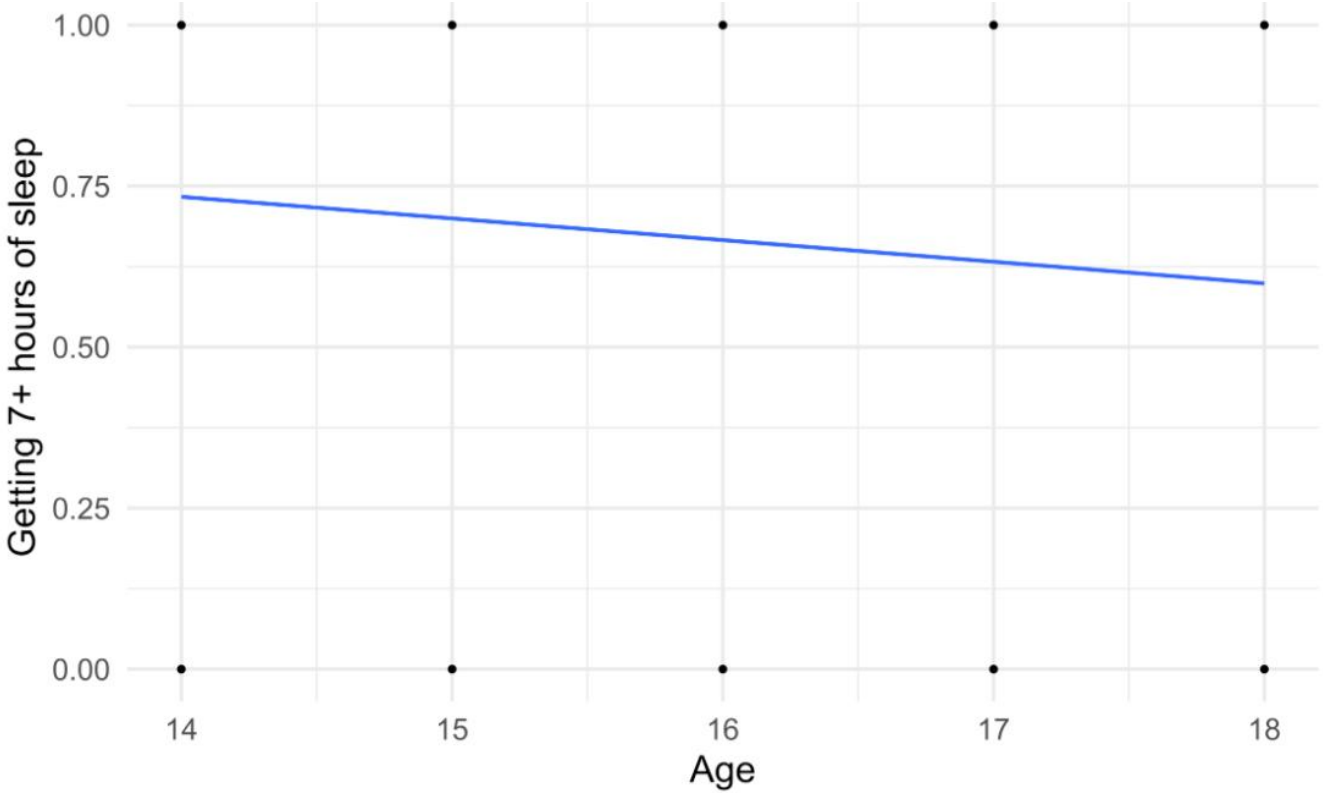
```
# A tibble: 446 × 6
     Age Sleep7 Sleep          SmokeLife SmokeDaily MarijuaEver
   <int>  <int> <fct>          <fct>     <fct>            <int>
 1    16      1 8 hours        Yes       Yes                  1
 2    17      0 5 hours        Yes       Yes                  1
 3    18      0 5 hours        Yes       Yes                  1
 4    17      1 7 hours        Yes       No                   1
 5    15      0 4 or less hours No        No                   0
 6    17      0 6 hours        No        No                   0
 7    17      1 7 hours        No        No                   0
 8    16      1 8 hours        Yes       No                   0
 9    16      1 8 hours        No        No                   0
10    18      0 4 or less hours Yes       Yes                  1
# … with 436 more rows
```

Example



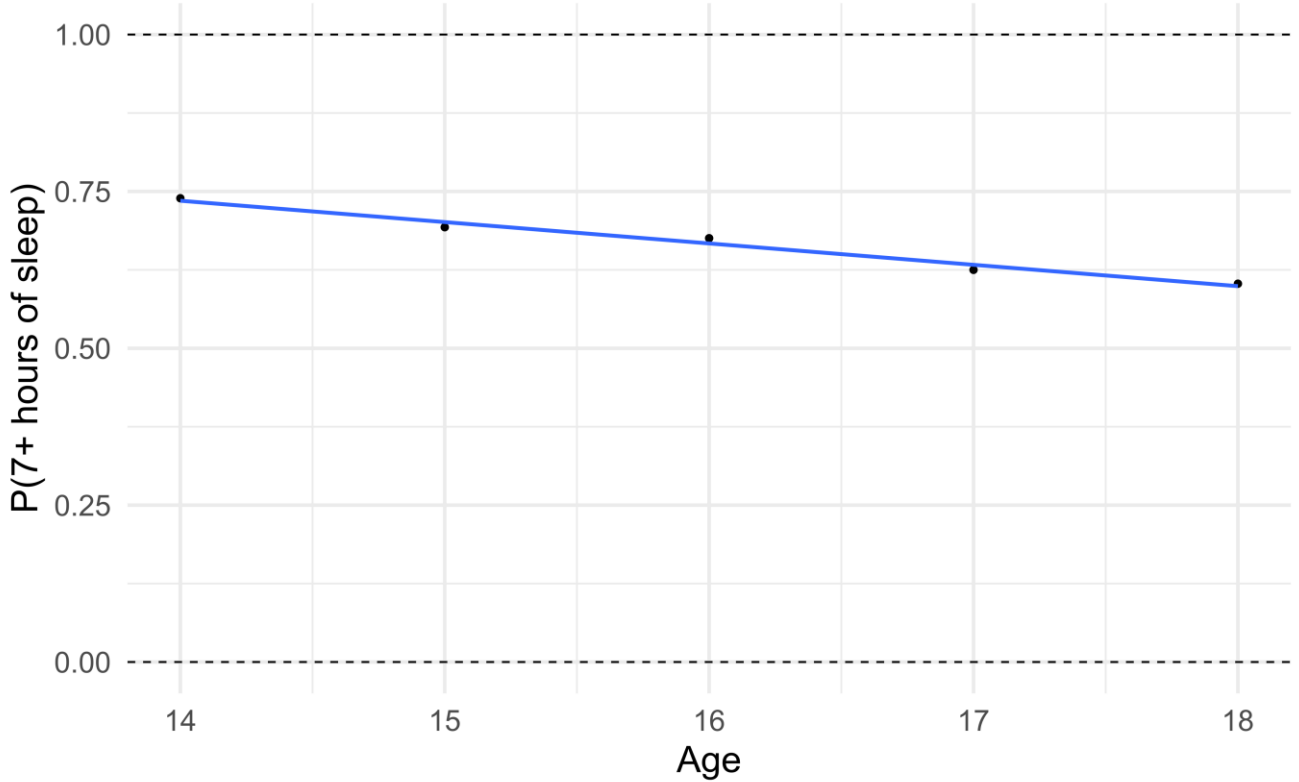
Example

- Outcome: 1 = yes, 0 = no



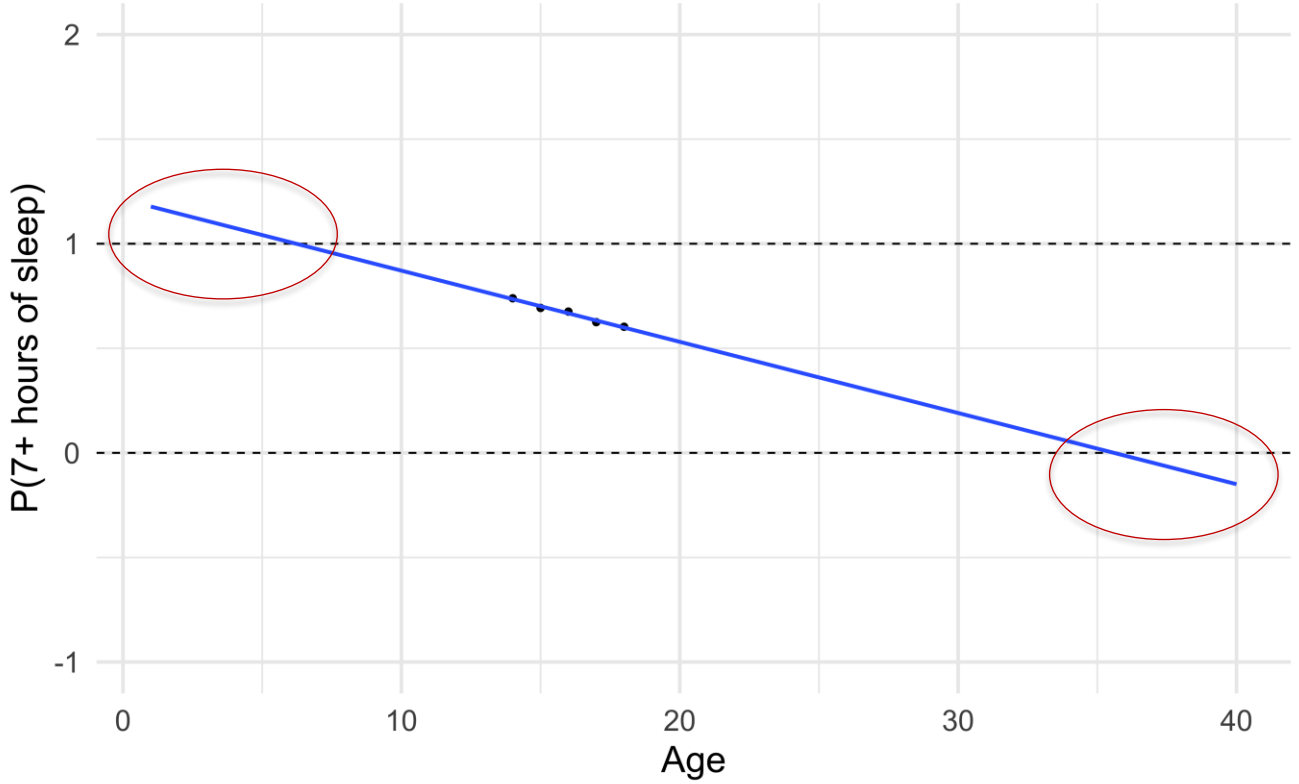
Example

- Outcome: Probability of getting 7+ hours of sleep



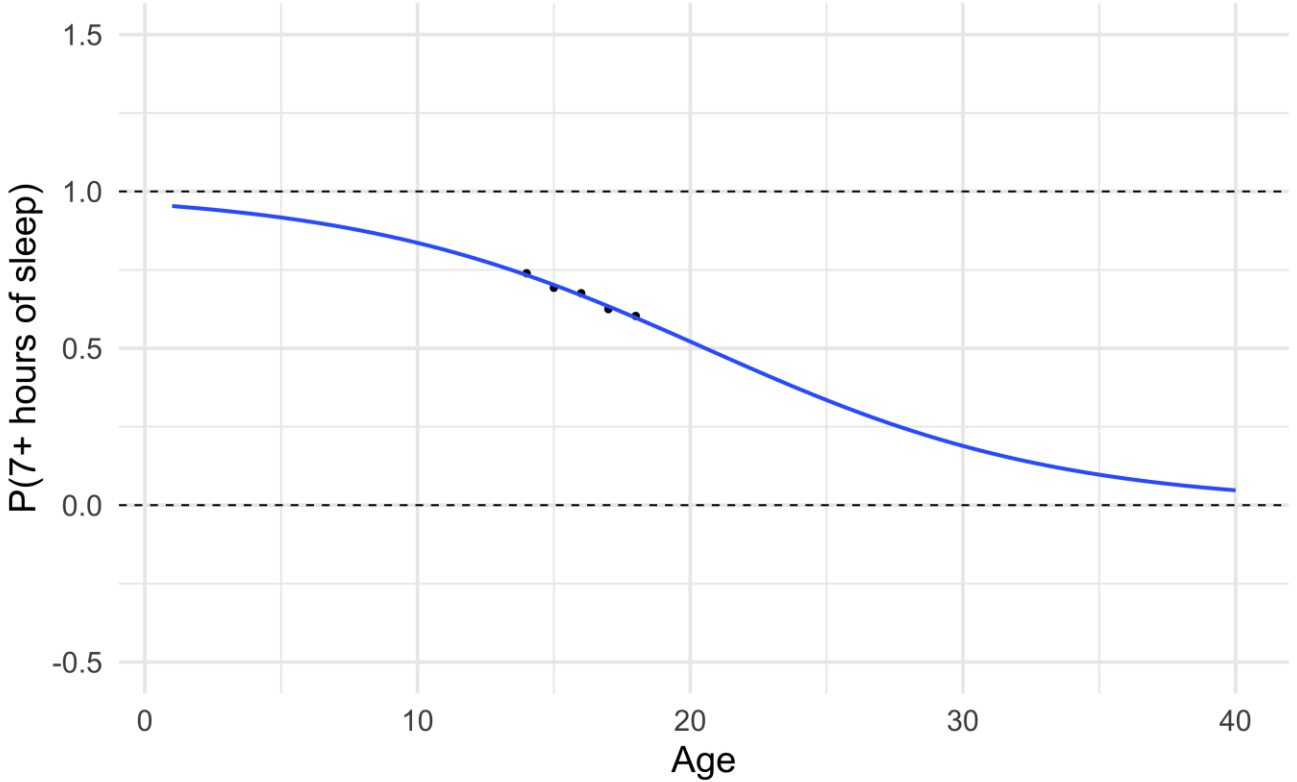
Example

- Outcome: Probability of getting 7+ hours of sleep (Linear model)



Example

- Outcome: Probability of getting 7+ hours of sleep (Logistic model)



Different types of models

Method	Outcome	Model
Linear regression	Quantitative	$Y = \beta_0 + \beta_1 X$
Linear regression (transform Y)	Quantitative	$\log(Y) = \beta_0 + \beta_1 X$
Logistic regression	Binary	$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

2. Logistic regression: the theory

What is logistic regression?

- Logistic regression is a generalized linear model where the outcome is a two-level categorical variable
- E.g “Trusting people” = 1, “Not trusting people” = 0
- The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i.
- E.g. Y_i will be used to represent whether person i is in the *trusting* category ($Y_i = 1$) or not ($Y_i = 0$.(
- The outcome, Y_i ,takes the value 1 with probability p_i and the value 0 with probability $1 - p_i$.
- Because each observation has a slightly different context, (e.g., different education level if we have education as an independent variable), the probability p_i will differ for each observation.
- It is this **probability** that we model in relation to the predictor variables

The logit transformation

- The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ is the value of variable 2 for observation i, and so on.
- We want to choose a **transformation** in the equation that makes practical and mathematical sense.

$$transformation(p_i) = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}$$

- For example, we want a transformation that makes the range of possibilities on the left hand side of the equation equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range.

The logit transformation

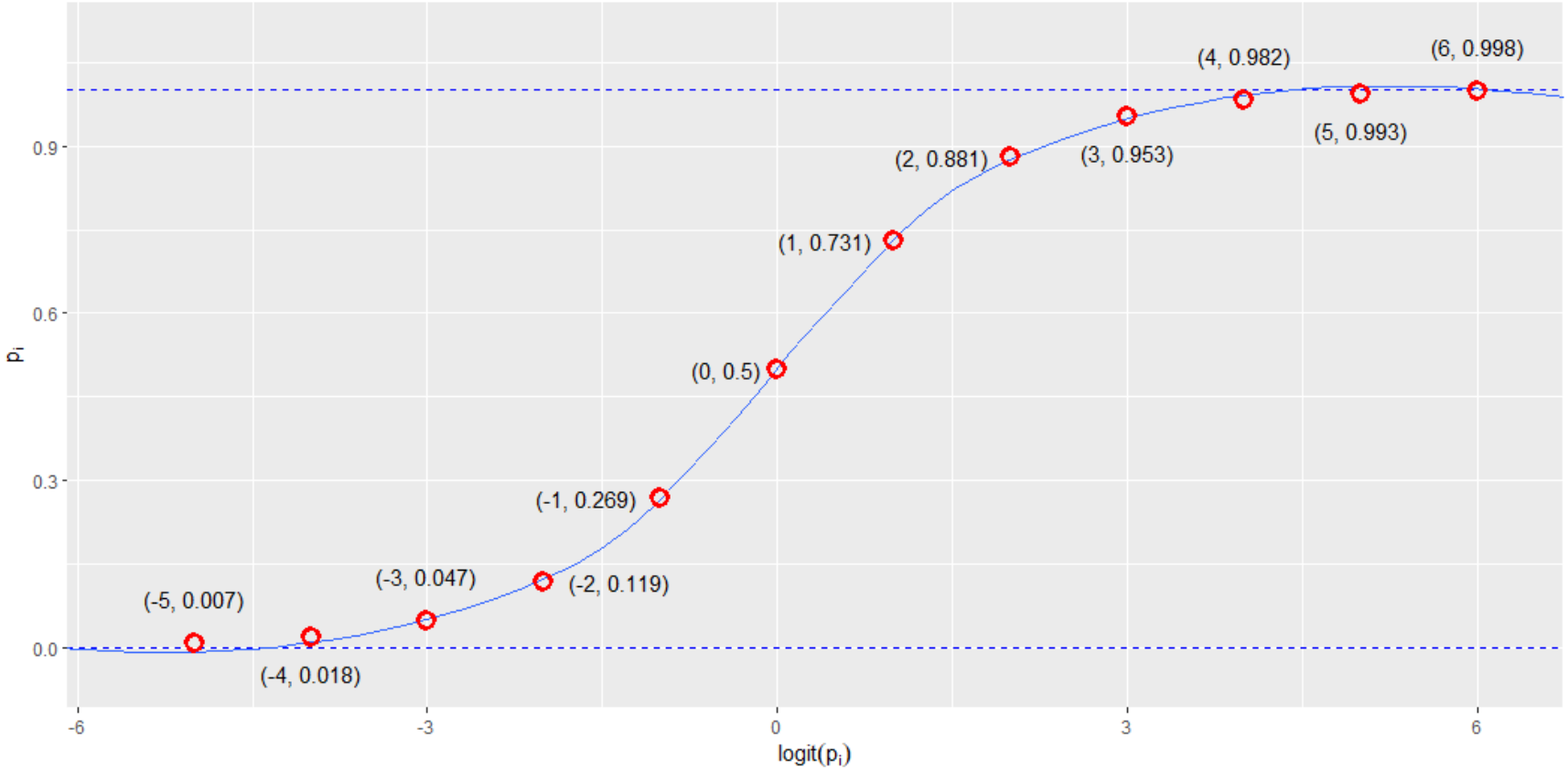
- A common transformation for p_i is the **logit transformation**, which may be written as:

$$logit(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right)$$

- We can rewrite the equation relating Y_i to its predictors using the logit transformation of p_i:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1x_{1,i} + b_2x_{2,i} + \cdots + b_kx_{k,i}$$

The logit transformation



Odds and *probabilities*

$Y = 1 : \text{ yes}, 0 : \text{ no}$

π: **probability** that $Y = 1$, i.e., $P(Y = 1)$

$\frac{\pi}{1-\pi}$: **odds** that $Y = 1$

$\log\left(\frac{\pi}{1-\pi}\right)$: **log odds**

Go from π to $\log\left(\frac{\pi}{1-\pi}\right)$ using the **logit transformation**

Odds and *probabilities*

$Y = 1 : \text{ yes}, 0 : \text{ no}$

π: **probability** that $Y = 1$, i.e

$\frac{\pi}{1-\pi}$: **odds** that $Y = 1$

$\log\left(\frac{\pi}{1-\pi}\right)$: **log odds**

Go from π to $\log\left(\frac{\pi}{1-\pi}\right)$ usin

probs	odds	log_odds
0.01	0.01	-4.60
0.05	0.05	-2.94
0.10	0.11	-2.20
0.20	0.25	-1.39
0.33	0.50	-0.69
0.40	0.67	-0.41
0.50	1.00	0.00
0.60	1.50	0.41
0.67	2.00	0.69
0.80	4.00	1.39
0.90	9.00	2.20
0.95	19.00	2.94
0.99	99.00	4.60

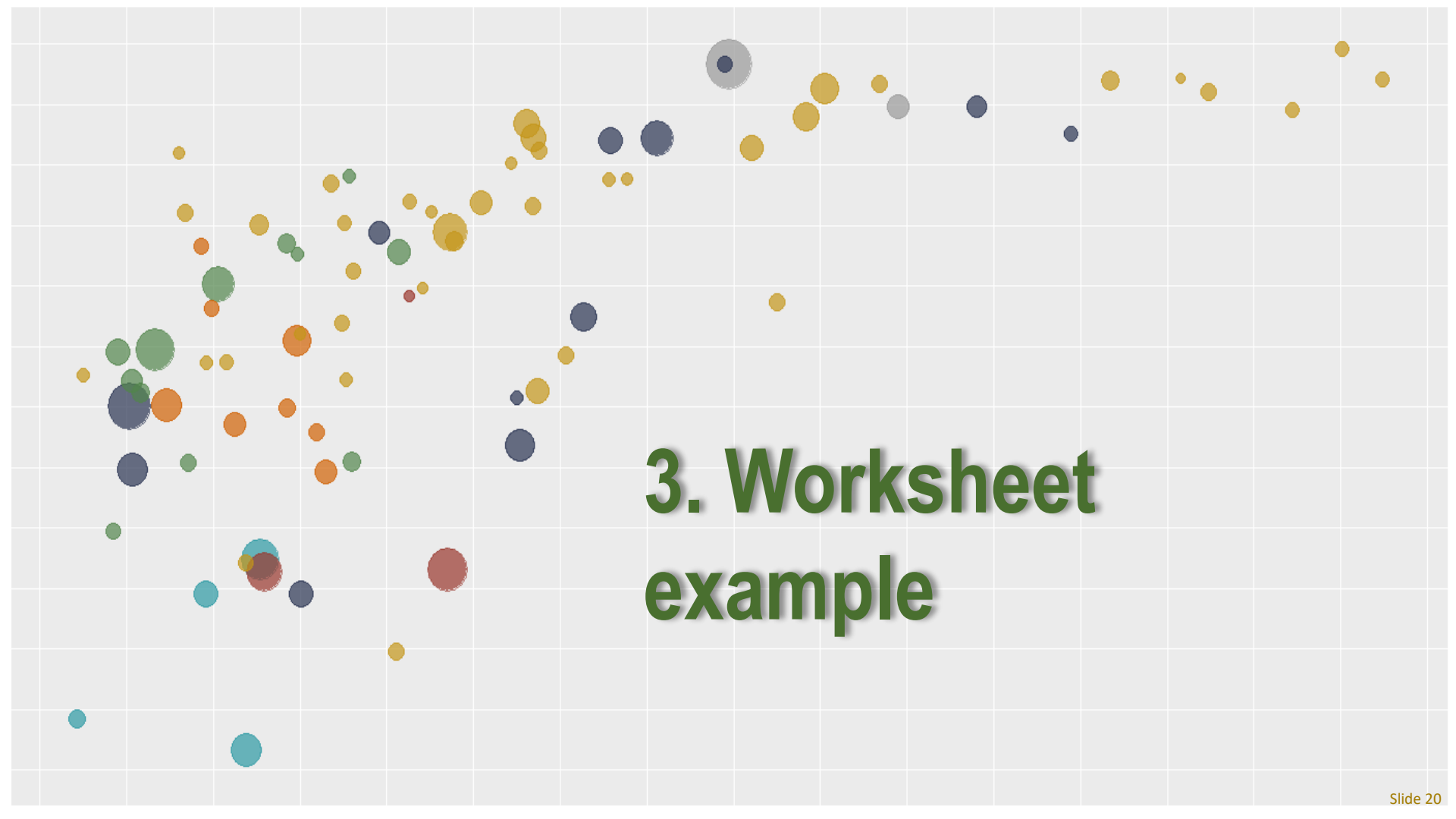
Odds

- Suppose there is **a 70% chance** it will rain tomorrow
- Probability it **will** rain is **p = 0.7**
- Probability it **won't** rain is **1 – p = 0.3**
- **Odds** it **will** rain are **7 to 3**, **7:3**, **0.7/0.3 ≈ 2.33**

Are teenagers getting enough sleep?

	Sleep7	n	p
	<int>	<int>	<dbl>
1	0	150	0.336
2	1	296	0.664

- P(7+ hours of sleep) = P(Y=1) = p = 0.664
- P(< 7 hours of sleep)=P(Y=0) = 1−p = 0.336
- P(odds of 7+ hours of sleep) = 0.664/0.336 = 1.976



3. Worksheet example

How does education relate to social trust?

- Application reading:

Wu, Cary. 2021. “Education and Social Trust in Global Perspective.” *Sociological Perspectives* 64 (6): 1166–86. https://doi.org/10.1177/0731121421990045.

- “there is a strong and positive relation between education level and trust”.
- However, “several studies have shown that education might yield differential impacts on trust in different societies. In Sweden, Sven Oskarsson et al. (2017) show that education has little impact on trust. In China, Cary Wu and Zhilei Shi (2020) suggest that education has a negative impact on people’s trust. Several cross-national studies have also shown that the education and trust association can vary from positive to negative depending on the specific institutional contexts …”

How does education relate to social trust?

- Methods:
  - "For the WVS, I use the standard survey item asking, **"Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" (Rosenberg 1956).** The variable is coded on a 0 to 1 scale, with **1 corresponding to high levels of trust**." (p. 1170)
  - "For the WVS, I measure **educational attainment** using respondents' highest education level attained with eight categories, namely, 1 = no formal education or inadequately completed elementary education, 2 = completed (compulsory) elementary education, 3 = incomplete secondary school: technical/vocational, 4 = complete secondary school: technical/vocational, 5 = incomplete secondary: university, 6 = complete secondary: university, 7 = some university without degree, and 8 = university with degree/higher education. In some analyses, I treat education as a categorical variable. To reduce the number of categories, I recode respondents' education into Primary, Secondary, Post-secondary, and Tertiary. This is also consistent with the most recent wave of the WVS coding" (p. 1171)
  - "I also **control for relevant demographic covariates** such as **gender**, **age**, **income**, **marital status**, and **occupational status** at the individual level" (p. 1172)

How does education relate to social trust?

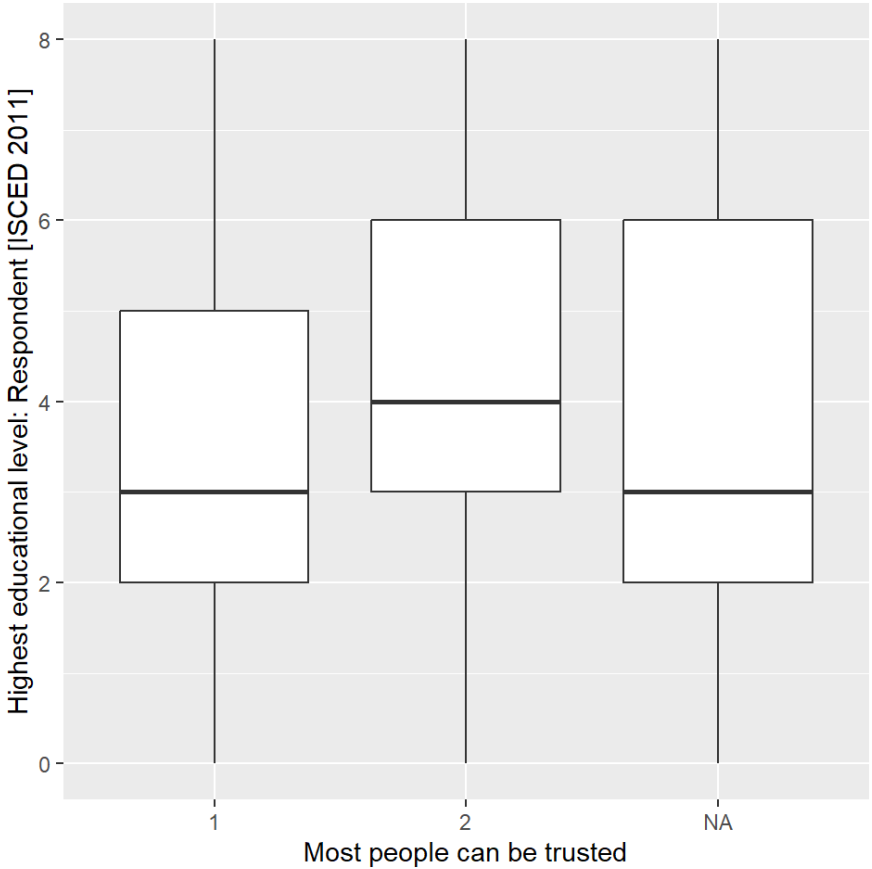
WVS7, selected variables (94278 rows and 7 variables, 7 shown)

ID	Name	Label	Type	Missings	Values	Value Labels	N
1	Q57	Most people can be trusted	categorical	1273 (1.4%)	1	Most people can be trusted	22552 (24.2%)
					2	Need to be very careful	70453 (75.8%)
2	Q275	Highest educational level: Respondent [ISCED 2011]	categorical	1012 (1.1%)	0	Early childhood education (ISCED 0) / no education	4708 (5.0%)
					1	Primary education (ISCED 1)	11101 (11.9%)
					2	Lower secondary education (ISCED 2)	14082 (15.1%)
					3	Upper secondary education (ISCED 3)	23880 (25.6%)
					4	Post-secondary non-tertiary education (ISCED 4)	8367 (9.0%)
					5	Short-cycle tertiary education (ISCED 5)	7818 (8.4%)
					6	Bachelor or equivalent (ISCED 6)	16133 (17.3%)
					7	Master or equivalent (ISCED 7)	6076 (6.5%)
					8	Doctoral or equivalent (ISCED 8)	1101 (1.2%)

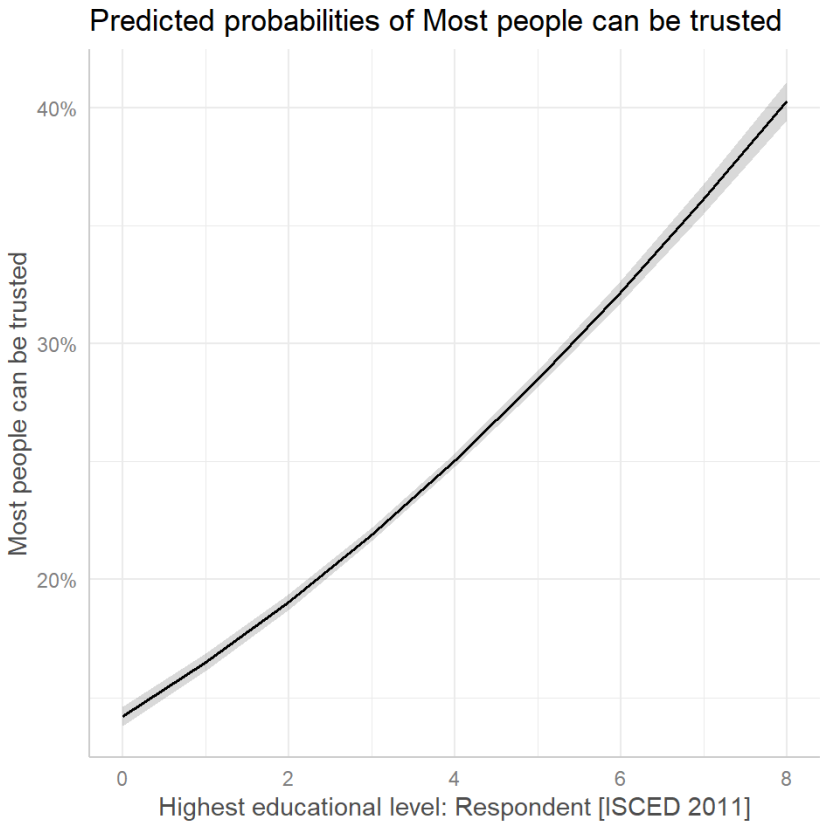
How does education relate to social trust?

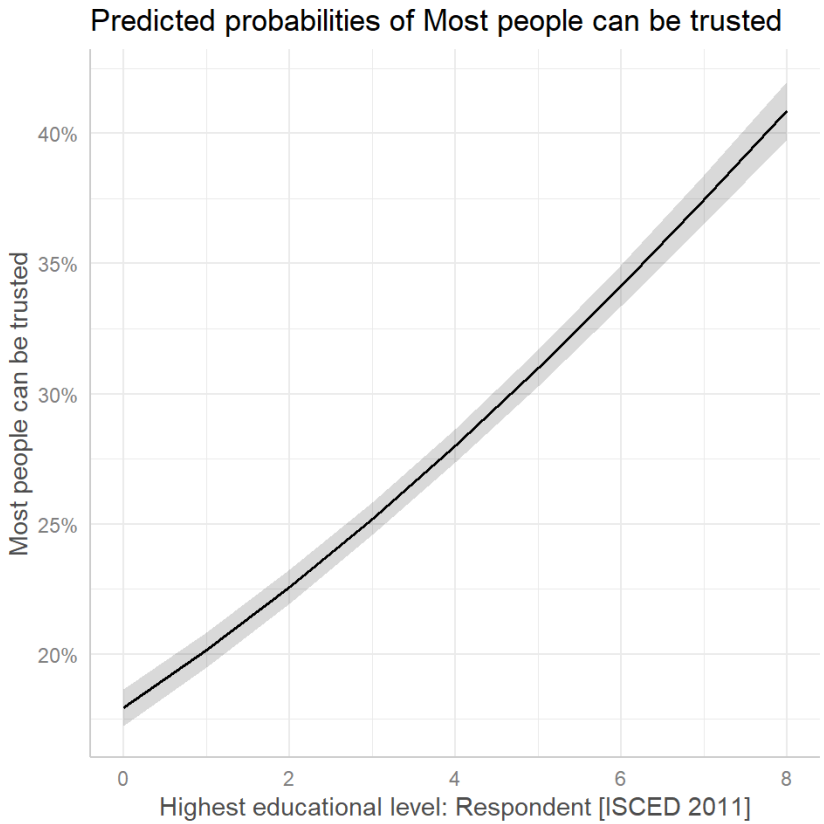
3	Q260	Sex	categorical	95 (0.1%)	1	Male	44403 (47.1%)
					2	Female	49780 (52.9%)
4	Q262	Age	numeric	510 (0.5%)	[16, 103]		93768
5	Q288	Scale of incomes	numeric	2928 (3.1%)	1	Lower step	7078 (7.7%)
					2	Second step	5344 (5.9%)
					3	Third step	10339 (11.3%)
					4	Fourth step	12847 (14.1%)
					5	Fifth step	21949 (24.0%)
					6	Sixth step	14108 (15.4%)
					7	Seventh step	10643 (11.7%)
					8	Eight step	5504 (6.0%)
					9	Nineth step	1706 (1.9%)
					10	Tenth step	1832 (2.0%)
6	Q273	Marital status	categorical	577 (0.6%)	1	Married	52028 (55.5%)
					2	Living together as married	7489 (8.0%)
					3	Divorced	4256 (4.5%)
					4	Separated	2082 (2.2%)
					5	Widowed	5427 (5.8%)
					6	Single	22419 (23.9%)
7	Q279	Employment status	categorical	1186 (1.3%)	1	Full time (30 hours a week or more)	34071 (36.6%)
					2	Part time (less than 30 hours a week)	7972 (8.6%)
					3	Self employed	13309 (14.3%)
					4	Retired/pensioned	11961 (12.8%)
					5	Homemaker not otherwise employed	12388 (13.3%)
					6	Student	5219 (5.6%)
					7	Unemployed	7073 (7.6%)
					8	Other	1099 (1.2%)

How does education relate to social trust?

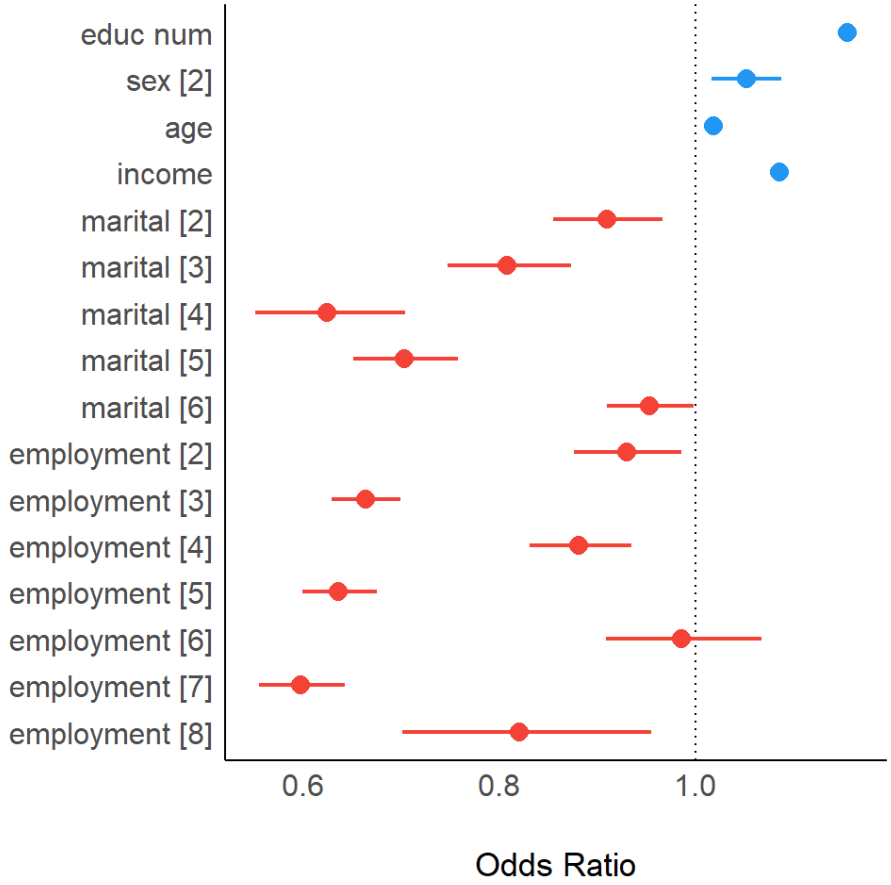


How does education relate to social trust?





How does education relate to social trust?



Good luck in the labs!